University of Nottingham



School of Mathematical Sciences

Empirical Likelihood in Euclidean and Some non-Euclidean Spaces

Charles Xi Yan

A thesis submitted to the University of Nottingham for the degree of

Doctor of Philosophy

2019

Abstract

Empirical likelihood is a non-parametric approach to statistical inference which bears some resemblance to parametric approaches and yet avoids making parametric assumptions. The aim of this thesis is to develop theoretical and computational aspects of empirical likelihood when applied to data from two different types of non-Euclidean space. The two spaces we focus on are the sphere, which is an important example of a manifold with non-Euclidean structure and is the appropriate sample space in directional statistics; and the 3-Spider, which is an example of a stratified manifold. The focus is mainly on developing empirical likelihood for a Fréchet mean in these two settings. Key achievements are: to prove versions of Wilks' theorem which open the way to the application of methods of inference including the construction of confidence regions and hypothesis testing; and the development of algorithms for performing the computations.

Acknowledgements

First I would like to thank my supervisors, Professor Andrew Wood and Professor Huiling Le, for their support in all aspects through all stages of my PhD.

Second I would like to thank my friend Luke Warren and Katie Severn for the encouragement and mental help, especially when I felt upset.

Last but not the least, I also would like to thank my family, Suping Sun, Xinmin Yan, Jie Yan and Eva Yan Wu.

CONTENTS

1 INTRODUCTION 1

1.1 Main Ideas of Empirical Likelihood 1

1.2 Literature Review 3

1.3 Structure and Content of the Thesis 5

2 BACKGROUND KNOWLEDGE 7

2.1 Empirical Likelihood 8

2.1.1 Empirical Likelihood for a Multivariate Mean 13

2.1.2 Bootstrap Calibration 14

2.2 Non-central Chi-squared Distribution 15

2.3 Local Alternatives 16

2.4 Differential Geometry Results 17

2.4.1 Tangent Space and Vector Field 17

2.4.2 Geodesics 18

2.4.3 log-map 19

2.5 Spherical Statistics Results 21

2.5.1 Mean Direction 22

2.5.2 Extrinsic and Intrinsic Mean 22

2.5.3 Local Alternatives on S^{k-1} 24

2.5.4 Useful Distributions on the Sphere 25

2.6 Model: 3-Spider 26

2.6.1 Folding Map on 3-Spider 27

2.6.2 Mean on $\mathcal{SP}_3$ 27

2.7 Phylogenetic Tree Data 32

2.8 Measure Theory Results 35

2.8.1 Glivenko-Cantelli Theorem 38

2.8.2 Stieltjes Integral 43
2.9 Matrix Algebra Results 45
2.9.1 Eigenvalue Property for Symmetric Matrices 45
2.9.2 Square Root of a Positive Definite Matrix 45
2.9.3 Matrix Calculus 46
2.9.4 Matrix Norm 47
2.10 Convex Optimization Results 49
2.10.1 Relative Interior 49
2.10.2 Separating Hyperplane Theorem 49
2.11 Stereographic Projection 50

3 EMPIRICAL LIKELIHOOD IN EUCLIDEAN SPACE 51
3.1 Introduction 51
3.2 Wilks' Theorem under Local Alternatives 52
3.3 Some Useful Lemmas 53
3.4 Proof of Theorem 3.1 69
3.5 Concavity of $l_{EL}(\mu)$ 73
3.6 Summary 78

4 EMPIRICAL LIKELIHOOD ON THE SPHERE 79
4.1 Introduction 79
4.2 Empirical Likelihood on $\mathcal{S}^{k-1}$ 80
4.2.1 Empirical Likelihood Ratio for Extrinsic Mean 80
4.3 Wilks' Theorem with Local Alternatives on $\mathcal{S}^{k-1}$ 86
4.3.1 Statement of Wilks' Theorem 86
4.3.2 Rotation on $\mathcal{S}^{k-1}$ 88
4.3.3 Some Useful Lemmas 90
4.3.4 Proof of Theorem 4.1 98
4.4 Empirical Likelihood Ratio Algorithm 102
4.5 Operating Characteristics in Hypothesis Testing 103
4.5.1 Probability of Type I Error 103
4.5.2 Probability of Type II Error 109

4.6 Confidence Region Construction 126
4.7 Summary 129

5 EMPIRICAL LIKELIHOOD ON 3-SPIDER 130
5.1 Introduction 130
5.2 Empirical Likelihood on the 3-Spider 131
5.2.1 Empirical Likelihood Ratio when b is off the Cone Point 132
5.2.2 Empirical Likelihood Ratio when b is on the Cone Point 133
5.3 Wilks' Theorem on 3-Spider 139
5.4 Two Useful Lemmas 141
5.5 Proof of Theorem 5.1 142
5.6 Algorithm: Empirical Likelihood Ratio on 3-Spider 145
5.7 Simulation 146
5.7.1 Non-Sticky Case 146
5.7.2 Partially-Sticky Case 147
5.7.3 Fully-Sticky Case 149
5.8 Phylogenetic Tree Data in 3-Spider 149
5.8.1 Phylogenetic Tree Data on 3-Spider 150
5.8.2 Confidence Region Construction 152
5.9 Numerical Realization 158
5.9.1 Probability of Type I Error 158
5.9.2 Probability of Type II Error 162
5.9.3 Application in Phylogenetic Tree Data 165
5.10 Summary 165

6 CONCLUSION AND FURTHER WORK 167
6.1 Wilks' Theorems and Concavity of log-Empirical Likelihood in $\mathbb{R}^k$ 167
6.2 Simulation Study 169
6.2.1 Confidence Region Construction 169
6.2.2 Operating Characteristics 169
6.3 Further Research 170

6.3.1 Intrinsic Distance on $\mathcal{S}^{k-1}$ 171
6.3.2 Local Alternatives in $\mathcal{SP}_3$ 171
6.3.3 Open Books and $\mathcal{Q}_5$ 172

1

INTRODUCTION

This chapter reviews the empirical likelihood method and discusses relevant background literature. Motivations for the developments in the thesis are explained and the structure and contents of the thesis are outlined.

1.1 MAIN IDEAS OF EMPIRICAL LIKELIHOOD

As a non-parametric statistical inference method, the study of empirical likelihood has drawn increasing attention to itself since late 1980s. Unlike parametric statistics, the biggest advantage of empirical likelihood is that no specific distribution family for the data is assumed. Advances in numerical computation make it easier to evaluate empirical likelihood in a variety of settings. As Owen (2001) stated, empirical likelihood can be regarded as a bootstrap method without resampling, and a likelihood method without parametric assumptions. Empirical likelihood has been applied in many fields including econometrics and biomedicine. Some applications of empirical likelihood have been investigated for survival analysis and time series.

Empirical likelihood is a method based on a likelihood function supported by the sample. However, how to define empirical likelihood on a specific space is the key of this method. The asymptotic distribution of the statistic, based on empirical likelihood, is used for statistical inference.

Wilks (1938) developed Wilks' theorem to offer an asymptotic distribution for the log-likelihood ratio statistic in parametric statistics. This theorem is widely used for statistical inference, such as outlining confidence regions and performing the likelihood-ratio test in parametric statistics. Owen (1988) defined empirical likelihood for the first time in Euclidean space. He proved that Wilks' theorem holds for empirical likelihood. For a hypothesis test with sequences of alternative converging to the null hypothesis at a rate of n^{-r}, where n is the sample size and r is determined by the alternative hypothesis, Sugiura (1973) found the asymptotic non-null distribution of the log-likelihood ratio statistic at alternatives. We refer to these sequences of alternatives as local alternatives in Section 2.3. Therefore, it is of interest to investigate the asymptotic distribution of Wilks' statistic for empirical likelihood evaluated at local alternatives.

Following the development of empirical likelihood and Wilks' theorem in Euclidean space, the question of how to build empirical likelihood and adapt Wilks' theorem in non-Euclidean spaces has received attention. Fisher et al. (1996) investigated confidence region construction for directional data by empirical likelihood algorithm. Throughout the thesis, we primarily concentrate on two non-Euclidean spaces, the sphere and the 3-Spider; we define the 3-Spider later in Section 2.6.

A possible approach to define empirical likelihood with spherical data is to transform the spherical data from the sphere to some chosen tangent space. In the tangent space, many methods for multivariate analysis are available, including the empirical likelihood method defined in the Euclidean space. As for local alternatives, we define them by so-called $\log_{\boldsymbol{p}}$ map, where $\boldsymbol{p}$ is an arbitrary point on the sphere; we refer to this later in Section 2.5. Then, the asymptotic non-null distribution can be studied by applying the result in the Euclidean space.

The 3-Spider is a complicated non-Euclidean space. In order to define empirical likelihood, the folding map and the Fréchet mean on this space are studied; see Hotz et al. (2013). The folding map connects the 3-Spider

with the Euclidean space. Therefore, the empirical likelihood method in the Euclidean space is applied to define empirical likelihood and develop Wilks' theorem for the 3-Spider.

1.2 LITERATURE REVIEW

Owen (1988) pioneered the work of empirical likelihood, building on earlier ideas of Kalbfleisch and Prentice (1980) and Thomas and Grunkemeier (1975). In frequentist statistics, the parametric likelihood function is widely used for constructing confidence regions and performing hypothesis tests. Owen (1988) showed that empirical likelihood, a form of non-parametric likelihood, can be used in similar fashion, but without making any parametric assumptions. One particularly interesting feature of empirical likelihood confidence regions is that their shape is determined by the sample and not by parametric assumptions.

Under mild regularity conditions, Wilks' theorem was first proposed and proved by Wilks (1938), showing that Wilks' statistic, which is based on the likelihood ratio function for the mean, has an asymptotic chi-squared distribution. Owen (1988) used the empirical likelihood ratio function to prove a non-parametric version of Wilks' theorem. In a later paper, Owen (1990) gave more detail of the proof of Wilks' theorem for the multivariate mean in the Euclidean space, and investigated confidence regions for the mean. In this paper, he also investigated the concavity of the log-empirical likelihood ratio function for the mean parameter. However, he did not present the full proof. More extensive accounts have been illustrated and discussed in the book by Owen (2001).

With Wilks' theorem, we can perform hypothesis tests. For arbitrary alternative hypotheses close to the null hypothesis, Sugiura (1973) found that the asymptotic alternative distribution may be degenerate. Local alternatives make up a sequence that the alternative hypotheses converge to the null hypothesis at a well-chosen rate of the sample size, say $-1/2$. This con-

cept was first included by Neyman (1937) for a local power analysis. Considering local alternatives, the Wilks' statistic has an asymptotic non-central chi-squared distribution.

The motivation for using empirical likelihood rather than parametric likelihood may be even greater for data from non-Euclidean spaces, because there is typically a shortage of convincing and natural parametric models in non-Euclidean spaces, especially when singularities are present. Thus, with empirical likelihood well-defined in these spaces, Wilks' theorem still guides the confidence region construction. Fréchet (1948) first defined the so-called Fréchet mean, giving a representative point of the centre for a cluster of points. It is determined by the choice of distance function. In the Euclidean space, the usual distance is Euclidean distance, so the Fréchet mean refers to the Euclidean mean. However, in a non-Euclidean space embedded in a Euclidean space, the Euclidean distance function on this space determines the extrinsic Fréchet mean.

Fisher et al. (1996) proposed two pivotal techniques for constructing confidence regions with directional data, one of which is the method of empirical likelihood. The advantage of empirical likelihood is that the log-empirical likelihood ratio statistic has an asymptotic chi-squared distribution, and it does not depend on any unknowns. Their simulation outcome suggests empirical likelihood performs best with bootstrap calibration in all methods they considered.

In hypothesis testing under local alternatives, resampling for bootstrap calibration needs to be performed under the null hypothesis. Amaral et al. (2007) presented an approach to transform directional data from the alternative hypothesis to the null hypothesis. With the transformation of the directional data, the asymptotic non-null distribution by empirical likelihood method is obtainable.

Hotz et al. (2013) developed the central limit theorem (CLT) on a non-Euclidean space, the open book. In this paper, the distance of any two arbitrary points is introduced, and hence the definition of the Fréchet mean.

They found three different cases for the location of the Fréchet mean, which lead to three different types of central limit theorems (CLTs).

1.3 STRUCTURE AND CONTENT OF THE THESIS

Below are the contents of all following chapters. The motivation for each chapter is also included.

Chapter 3 states and proves Wilks' theorem for empirical likelihood, considering local alternatives. As was proposed by Sugiura (1973), the parametric log-likelihood ratio statistic, under local alternatives, has an asymptotic non-central chi-squared distribution. Our focus is how to prove the theorem by using a non-parametric method, empirical likelihood, since the empirical likelihood method only requires very mild assumptions on the form of the distribution. This results in a non-parametric version of Wilks' theorem. It is also proved that the log-empirical likelihood is a concave function of the mean parameter. This result was stated by Owen (1990), but he did not provide a full proof.

Chapter 4 develops empirical likelihood on the sphere. As empirical likelihood is well-defined in the Euclidean space, we try to find an approach to connect spherical data with the Euclidean data to make empirical likelihood on the sphere easy to calculate. Considering local alternatives, Wilks' theorem for empirical likelihood is proved on the sphere. According to the theorem, the confidence region is constructed and operating characteristics of this confidence region procedure are examined.

Chapter 5 extends empirical likelihood to another non-Euclidean space, the 3-Spider. As is illustrated by Hotz et al. (2013), the Fréchet mean is used in the 3-Spider. In order to define empirical likelihood in 3-Spider, we use the folding map to send the 3-Spider data to the Euclidean space. There are three different cases for the location of the Fréchet mean, and these three different cases lead to three approaches to Wilks' theorem for the 3-Spider.

Similarly to Chapter 4, the confidence region is constructed and operating characteristics of this confidence region procedure are examined.

Chapter 6 presents conclusions and some ideas for the future work. In hypothesis testing, whether to use tabular quantiles or bootstrap calibrated quantiles is discussed, on the basis of numerical work in Chapter 4 and Chapter 5. Further problems are considered, including how to define sequences of local alternatives in the 3-Spider, and how to extend the empirical likelihood method to other non-Euclidean spaces.

2

BACKGROUND KNOWLEDGE

This chapter presents the background on the following topics: statistical inference, empirical likelihood, and notation for statistical analysis on non-Euclidean spaces used in the thesis. An extensive account of empirical likelihood can be found in the work by Owen (2001).

This chapter is organised as follows. Section 2.1 reviews the mathematical representation of empirical likelihood and one of its most important theoretical results, Wilks' theorem. In Section 2.2, the non-central chi-squared distribution is introduced. The non-central chi-squared distribution appears in Section 2.3, where we consider Wilks' theorem under a sequence of local alternatives. In Section 2.4 and Section 2.5, a review of relevant differential geometry and spherical statistics is presented. In Section 2.6, a non-Euclidean space, the 3-Spider, is considered. Some background for phylogenetic tree data is included in Section 2.7; the 3-Spider is relevant to this type of data. Background for measure theory, matrix algebra, and convex optimization that is required for various proofs throughout the thesis is reviewed in Section 2.8, Section 2.9 and Section 2.10 respectively. The method for graphical presentation of empirical likelihood against the mean direction parameter on the sphere is given in Section 2.11.

2.1 EMPIRICAL LIKELIHOOD

In statistics, non-parametric methods of inference are widely used when the parametric family is not assumed or given. Empirical Likelihood is one of the modern methods for inference based on a non-parametric likelihood ratio function defined through the data. Empirical likelihood was first introduced by Owen (1988, 1990). Hall and La Scala (1990) showed that an advantage of empirical likelihood is that the shape of confidence regions automatically reflects characteristics of the observed dataset. Furthermore, empirical likelihood behaves like a parametric likelihood in the sense that it satisfies Wilks' theorem proposed by Wilks (1938).

Let $\mathbb{1}_A(\boldsymbol{X})$ denote the indicator function for a set A, i.e.

$$\mathbb{1}_A(\boldsymbol{X}) = \begin{cases} 1, & \text{if } \boldsymbol{X} \in A; \\ 0, & \text{otherwise.} \end{cases}$$

There is another way to characterize the indicator function with respect to the set A:

$$\delta_{\boldsymbol{X}}(A) = \begin{cases} 1, & \text{if } \boldsymbol{X} \in A; \\ 0, & \text{otherwise.} \end{cases} \tag{2.1.1}$$

Note that $\mathbb{1}_A(\boldsymbol{X})$ is a function of $\boldsymbol{X}$, while $\delta_{\boldsymbol{X}}(A)$ is a function of sets. For convenience, throughout the thesis, A shall denote Borel set, while $\boldsymbol{X}$ here is a point where all the probability is concentrated.

Let $\boldsymbol{X_1}, \cdots, \boldsymbol{X_n}$ denote a random sample in $\mathbb{R}^k$ drawn from the distribution P_0. Owen (2001, Chapter 3.1) uses indicator functions to define the empirical distribution function, the empirical measure, and non-parametric likelihood as follows.

Definition 2.1.1 (Empirical Distribution) *Let $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$ be i.i.d. random vectors. The empirical distribution of $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$ is*

$$P_n(A) = \frac{1}{n} \sum_{i=1}^{n} \delta_{\boldsymbol{X}_i}(A). \tag{2.1.2}$$

In empirical likelihood, we build the model for the distribution P by setting

$$P(A) = \sum_{i=1}^{n} p_i \delta_{\boldsymbol{X}_i}(A), \tag{2.1.3}$$

where A is an arbitrary Borel-measurable subset of $\mathbb{R}^k$, and p_i is the weight of $\delta_{\boldsymbol{X}_i}(A)$ with

$$0 \leqslant p_i \leqslant 1,$$

and

$$\sum_{i=1}^{n} p_i = 1,$$

for $i = 1, 2, \cdots, n$. This model is built using values of observations in the sample.

For $i = 1, \cdots, n$, define the singleton set $A_i = \{\boldsymbol{X}_i\}$. Then, let

$$\begin{aligned} p_i &= P(A_i) \\ &= P(\{\boldsymbol{X}_i\}) \end{aligned} \tag{2.1.4}$$

be the probability mass of $\boldsymbol{X}_i$ in a sample from P.

For the distribution as in (2.1.3), we assume that the probabilities have already been summed over any ties. Here, $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ are tied, if $\boldsymbol{X}_i = \boldsymbol{X}_j$, for $i \neq j$. Otherwise, consider the ties in the data, and suppose the distinct value $\boldsymbol{X}_j$ arises $n_j \geqslant 1$ times in the sample, with probability p_j, $j = 1, \cdots, m$. Then the probability vector $\boldsymbol{p} = (p_1, \cdots, p_n)$ is redefined as $\boldsymbol{p}' = (p_1', \cdots, p_m')$, where m is the number of distinct values in the sample. Thus, we may assume that no ties are present. With a random sample from a continuous distribution, the sample points are distinct with probability 1.

Example 2.1.1 *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3$ be three untied observations with probability vector $\boldsymbol{p} = (p_1, p_2, p_3)$. However, if $\boldsymbol{X}_2$ and $\boldsymbol{X}_3$ are tied, then the probability vector becomes $\boldsymbol{p}' = (p_1', p_2')$ with*

$$\begin{aligned} p_1' &= p_1, \\ p_2' &= p_2 + p_3. \end{aligned}$$

Definition 2.1.2 (Non-parametric Likelihood) *The non-parametric likelihood of the distribution P is*

$$\begin{aligned} L(P) &= \prod_{i=1}^{n} P(A_i) \\ &= \prod_{i=1}^{n} p_i. \end{aligned} \tag{2.1.5}$$

Remark: Recalling a classical Lagrange multiplier optimization problem:

$$\begin{aligned} &\max_{p_1,\cdots,p_n} \prod_{i=1}^{n} p_i, \\ s.t. &\sum_{i=1}^{n} p_i = 1, \\ &0 < p_i < 1, i = 1, 2, \cdots, n. \end{aligned}$$

Let

$$H(p_1, \cdots, p_n, \gamma) = \prod_{i=1}^{n} p_i + \gamma \left(\sum_{i=1}^{n} p_i - 1 \right).$$

Now, we calculate the gradient:

$$\begin{aligned} \nabla_{p_1,\cdots,p_n,\gamma} H(p_1, \cdots, p_n, \gamma) &= \left(\frac{\partial H}{\partial p_1}, \cdots, \frac{\partial H}{\partial p_n}, \frac{\partial H}{\partial \gamma} \right)^\top \\ &= \begin{pmatrix} p_2 p_3 \cdots p_n + \gamma \\ p_1 p_3 \cdots p_n + \gamma \\ p_1 p_2 \cdots p_{n-1} + \gamma \end{pmatrix}, \end{aligned}$$

and therefore

$$\begin{aligned} &\nabla_{p_1,\cdots,p_n,\gamma} H(p_1, \cdots, p_n, \gamma) = \mathbf{0} \\ \iff &\begin{pmatrix} p_2 p_3 \cdots p_n + \gamma = 0 \\ p_1 p_3 \cdots p_n + \gamma = 0 \\ p_1 p_2 \cdots p_{n-1} + \gamma = 0 \end{pmatrix} \\ \iff &p_1 = p_2 = \cdots = p_n = \frac{1}{n}. \end{aligned}$$

The maximal value of $L(P)$ is given by $P = P_n$ such that $p_i = \frac{1}{n}$ for $i = 1, \cdots, n$.

Definition 2.1.3 (Empirical Likelihood Ratio) *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$ be i.i.d. random vectors in $\mathbb{R}^k$, $k \geqslant 1$, with common distribution P_0. Let $p_i \geqslant 0$ be as defined in (2.1.4), $i = 1, 2, \cdots, n$. The empirical likelihood ratio function for the mean is defined as*

$$\mathcal{R}(\boldsymbol{\mu}) = \max_{p_1, \cdots, p_n} \left\{ \prod_{i=1}^{n} np_i \Big| \sum_{i=1}^{n} p_i \boldsymbol{X}_i = \boldsymbol{\mu},\ \sum_{i=1}^{n} p_i = 1, p_i > 0, i = 1, \cdots n \right\}, \tag{2.1.6}$$

where $\boldsymbol{\mu}$ is the so-called mean parameter. Furthermore, the log-empirical likelihood ratio function for the mean parameter is defined as

$$\log \mathcal{R}(\boldsymbol{\mu}) = \max_{p_1, \cdots, p_n} \left\{ \sum_{i=1}^{n} \log np_i \Big| \sum_{i=1}^{n} p_i \boldsymbol{X}_i = \boldsymbol{\mu},\ \sum_{i=1}^{n} p_i = 1, p_i > 0, i = 1, \cdots n \right\}. \tag{2.1.7}$$

In Chapter 3, we state and prove Wilks' theorem under a sequence of local alternatives for the log-empirical likelihood (2.1.7) in the Euclidean setting. In Chapters 4 and 5, we state and prove analogous results in non-Euclidean settings. Note that the empirical likelihood ratio is derived from $L(P)/L(P_n)$, where $L(P)$ denotes the non-parametric likelihood function; see (2.1.5). Given the mean parameter $\boldsymbol{\mu}$, the empirical likelihood ratio has a corresponding maximum. Consider a non-trivial case that $0 < p_i < 1$, $i = 1, 2, \cdots, n$. Then solving (2.1.6) is equivalent to solving the system:

$$\begin{aligned} l(\boldsymbol{\mu}) &= \max_{p_1, \cdots, p_n} \sum_{i=1}^{n} \log np_i, \\ \text{subject to } & \sum_{i=1}^{n} p_i = 1,\ \sum_{i=1}^{n} p_i \boldsymbol{X}_i = \boldsymbol{\mu}, \\ & 0 < p_1 < 1,\ i = 1, 2, \cdots, n. \end{aligned} \tag{2.1.8}$$

Using the Lagrange multipliers, we set the objective function as

$$\begin{aligned} G(p_1, \cdots, p_n; \boldsymbol{\lambda}_0, \gamma) &= \sum_{i=1}^{n} \log np_i + \boldsymbol{\lambda}_0^\top \Big(\sum_{i=1}^{n} p_i \boldsymbol{X}_i - \boldsymbol{\mu} \Big) + \gamma \Big(\sum_{i=1}^{n} p_i - 1 \Big) \\ &= \sum_{i=1}^{n} \log np_i + \boldsymbol{\lambda}_0^\top \sum_{i=1}^{n} p_i (\boldsymbol{X}_i - \boldsymbol{\mu}) + \gamma \Big(\sum_{i=1}^{n} p_i - 1 \Big), \end{aligned}$$

where $\boldsymbol{\lambda}_0 \in \mathbb{R}^k$ and $\gamma \in \mathbb{R}$ are Lagrange Multipliers. Differentiating G with respect to $\boldsymbol{\lambda}_0$, γ, and p_i, $i = 1, 2, \cdots, n$ gives:

$$\begin{aligned} \frac{\partial G}{\partial \boldsymbol{\lambda}_0} &= \sum_{i=1}^{n} p_i \boldsymbol{X}_i - \boldsymbol{\mu} = \mathbf{0}, \\ \frac{\partial G}{\partial \gamma} &= \sum_{i=1}^{n} p_i - 1 = 0, \text{ and} \\ \frac{\partial G}{\partial p_i} &= \frac{1}{p_i} + \boldsymbol{\lambda}_0^\top (\boldsymbol{X}_i - \boldsymbol{\mu}) + \gamma = 0. \end{aligned} \tag{2.1.9}$$

By summing the third equality in (2.1.9),

$$\begin{aligned} \sum_{i=1}^{n} p_i \frac{\partial G}{\partial p_i} &= n + \boldsymbol{\lambda}_0 \sum_{i=1}^{n} p_i (\boldsymbol{X}_i - \boldsymbol{\mu}) + \gamma \sum_{i=1}^{n} p_i \\ &= n + \gamma \\ &= 0, \\ \Rightarrow \gamma &= -n. \end{aligned}$$

Finally, substituting $\gamma = -n$ into the third equality in (2.1.9), we have

$$p_i = \frac{1}{n(1 + \boldsymbol{\lambda}^\top (\boldsymbol{X}_i - \boldsymbol{\mu}))}, \tag{2.1.10}$$

where $\boldsymbol{\lambda} = -\dfrac{\boldsymbol{\lambda}_0}{n}$.

In summary, to calculate the log-empirical likelihood $l(\boldsymbol{\mu})$ against the mean parameter $\boldsymbol{\mu}$ as in (2.1.8), we need to find $\boldsymbol{\lambda}$ to solve

$$\sum_{i=1}^{n} p_i (\boldsymbol{X}_i - \boldsymbol{\mu}) = \mathbf{0},$$

where p_i is given by (2.1.10). A solution $\boldsymbol{\lambda}$ is finite if and only if $\boldsymbol{\mu}$ lies in the relative interior of the convex hull of $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$. We refer to the relative interior of a convex hull later in Section 2.10.

Remark: as is shown in the definition of convex hull in Section 2.10, $\boldsymbol{\mu}$ lying in the relative interior of the convex hull of $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$ is equivalent to finding a set of $\{p_i\}_{i=1}^n$, such that

$$\sum_{i=1}^{n} p_i = 1$$
$$0 < p_i < 1,\ i = 1, 2, \cdots, n.$$
$$\boldsymbol{\mu} = \sum_{i=1}^{n} p_i \boldsymbol{X}_i.$$

Thus, as p_i shown in (2.1.10), it is equivalent to having finite $\boldsymbol{\lambda}$.

On the other hand, when $\boldsymbol{\mu}$ lies outside the relative interior of the convex hull of $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$, it is equivalent to that there is no such a set of $\{p_i\}_{i=1}^n$, such that

$$\sum_{i=1}^{n} p_i = 1$$
$$0 \leqslant p_i \leqslant 1,\ i = 1, 2, \cdots, n.$$
$$\boldsymbol{\mu} = \sum_{i=1}^{n} p_i \boldsymbol{X}_i.$$

Thus, $\boldsymbol{\lambda}$ will not be defined, if p_i is shown in (2.1.10).

Lastly, when $\boldsymbol{\mu}$ lies on the boundary of the relative interior of the convex hull of $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$, in order to have

$$\boldsymbol{\mu} = \sum_{i=1}^{n} p_i \boldsymbol{X}_i,$$

where $\sum_{i=1}^{n} p_i = 1$ and $0 \leqslant p_i \leqslant 1$, for $i = 1, 2, \cdots, n$, there must have at least one $p_j = 0$, $j \in \{1, 2, \cdots, n\}$. It is equivalent to having $\|\boldsymbol{\lambda}\| = \infty$ in this case.

2.1.1 *Empirical Likelihood for a Multivariate Mean*

Corresponding to Definition 2.1.3, Owen (2001) showed that the confidence region is

$$C_{r,n} = \left\{ \sum_{i=1}^{n} p_i \boldsymbol{X}_i \middle| \prod_{i=1}^{n} n p_i \geqslant r, p_i \geqslant 0, \sum_{i=1}^{n} p_i = 1 \right\}$$

is a convex subset of $\mathbb{R}^k$, where r is a confidence level; see Theorem 2.1 below.

Substituting $\boldsymbol{\mu}_0$, the population mean of i.i.d. random vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$, into (2.1.7), we define the Wilks' statistic $-2\log\mathcal{R}(\boldsymbol{\mu}_0)$. Wilks' theorem proves that the Wilks' statistic has an asymptotic chi-squared distribution; see Owen (2001).

Theorem 2.1 (Wilks' Theorem) *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$ be independent random vectors in $\mathbb{R}^k$ with common distribution P_0 having mean $\boldsymbol{\mu}_0$ and positive semi-definite covariance matrix V_0 of rank $q > 0$. Then $C_{r,n}$ is a convex set and $-2\log\mathcal{R}(\boldsymbol{\mu}_0)$ converges in distribution to a chi-square random variable with q degrees of freedom, as $n \to \infty$.*

2.1.2 *Bootstrap Calibration*

In statistics, the bootstrap is an important way to assess the uncertainty in an estimator. Now, we introduce the bootstrap idea proposed by Efron and Tibshirani (1994). Let $\mathscr{X} = \{x_1, x_2, \cdots, x_n\}$ be a random sample from a distribution P. If the object of the study is an unknown parameter of the distribution, say $\theta = \theta(P)$, then we consider an estimate of the distribution first by empirical measure based on $\mathscr{X}$:

$$\hat{P}_n(A) = \frac{1}{n}\sum_{i=1}^{n}\delta_{\boldsymbol{X}_i}(A),$$

where set A and $\delta_{\boldsymbol{X}_i}(A)$, $i = 1, 2, \cdots, n$ are as defined in (2.1.1). Then the bootstrap estimate of the parameter θ is $\hat{\theta} = \theta(\hat{P}_n)$.

Now, we use the bootstrap to estimate quantiles of the distribution of $\hat{\theta}$. We call it bootstrap calibration for quantiles. For convenience, we use the notation $\hat{\theta}[\alpha]$ for the α-quantiles by bootstrap, where α is the nominal probability of type I error.

Let $\hat{\theta}$ denote the estimate based on the original sample and let $\hat{\theta}_1^*, \hat{\theta}_2^*, \cdots, \hat{\theta}_B^*$ denote estimators based on B bootstrap samples, selected randomly with

replacement from $\mathscr{X}$. Write $\hat{\theta}^*_{(1)} \leqslant \cdots \leqslant \hat{\theta}^*_{(B)}$ for the ordered values of $\hat{\theta}^*_1, \hat{\theta}^*_2, \cdots, \hat{\theta}^*_B$. Then we estimate $\hat{\theta}[j/(B+1)]$ by $\hat{\theta}^*_j$ and we estimate $\hat{\theta}[\alpha]$ for other values of α by linear interpolation.

Since Wilks' theorem for empirical likelihood gives the asymptotic distribution of the Wilks' statistic, we use bootstrap calibration to calculate quantiles of the log-empirical likelihood ratio for constructing confidence regions for the mean, and compare bootstrap calibrated confidence regions with confidence regions calculated by tabular quantiles in Section 4.5 and Section 5.9.

2.2 NON-CENTRAL CHI-SQUARED DISTRIBUTION

Johnson et al. (1994, Chapter 29) gave the definition of non-central chi-squared distribution by i.i.d. non-zero mean normal random variables. We shall give the definition as follows.

Definition 2.2.1 (Non-central Chi-squared Distribution) *If $U_1, U_2, \cdots, U_v$ are independent standard normal variables, and $\delta_1, \delta_2, \cdots, \delta_v$ are constants, then the distribution of*

$$V = \sum_{j=1}^{v} (U_j + \delta_j)^2$$

is called the non-central chi-squared distribution with v degrees of freedom and non-centrality parameter $A = \sum_{j=1}^{v} \delta_j^2$, and is written $\chi^2_v(A)$.

Note that the distribution of V depends on $\delta_1, \delta_2, \cdots, \delta_v$ only through the sum of their squares. Consider a sequence of independent k-dimensional real Gaussian random vectors $\{\boldsymbol{X}_i\}_{i=1}^n$, i.e.

$$\boldsymbol{X}_i \sim N_k(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i),$$

where for $i = 1, 2, \cdots, n$, $\boldsymbol{\Sigma}_i$ has full rank. Then, $\{\boldsymbol{\Sigma}_i^{-1/2}\boldsymbol{X}_i\}_{i=1}^n$ is a sequence of independent $N(\boldsymbol{\Sigma}_i^{-1/2}\boldsymbol{\mu}_i, \boldsymbol{I}_k)$ random vectors, where $\boldsymbol{I}_k$ is the k-dimensional identity matrix. According to Definition 2.2.1, we know that

$\sum_{i=1}^{n} \boldsymbol{X}_i^\top \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i$ is a non-central chi-squared random variable with the non-centrality parameter

$$\sum_{i=1}^{n} \boldsymbol{\mu}_i^\top \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\mu}_i,$$

and nk degrees of freedom.

2.3 LOCAL ALTERNATIVES

As is considered by Sugiura (1973), in parametric statistics for hypothesis tests with an alternative hypothesis close to the null hypothesis, if the alternative does not converge to the null hypothesis at some rate, the asymptotic non-null distribution evaluated under the sequence of alternatives may degenerate. A choice of the convergence rate is considered as $n^{-1/2}$ by Sugiura (1973). According to this idea, for the null hypothesis $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, we define the sequence of local alternatives $\{\boldsymbol{\theta}_n\}_{n \geqslant 1}$ as alternative hypotheses, where

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + \frac{\boldsymbol{\eta}}{\sqrt{n}}.$$

Local alternatives are widely used in local power analysis, which studies the power function of a hypothesis test in a neighborhood of the null hypothesis. As a useful device, this sequence was first used by Neyman (1937) for an asymptotic analysis of the power of a test.

Intuitively, local alternatives are quantities that converge to the parameter $\boldsymbol{\theta}_0$. Since the Euclidean space itself is isometric to its tangent space at any point, $\boldsymbol{\eta}/\sqrt{n}$ can be regarded as the image of $\log_{\boldsymbol{\theta}_0}$-map evaluated at local alternatives $\{\boldsymbol{\theta}_n\}_{n \geqslant 1}$; we refer to $\log_{\boldsymbol{\theta}_0}$-map later in Section 2.4.

2.4 DIFFERENTIAL GEOMETRY RESULTS

2.4.1 *Tangent Space and Vector Field*

In order to evaluate empirical likelihood against the mean direction on the sphere, we transform the random sample on $\mathcal{S}^{k-1}$ to a sample on a tangent space. Boothby (1986, Chapter 2.4 & 4.1) gives the definition of the tangent space at a point on a manifold. Now, a definition of the space of tangent vectors attached to a point in a manifold is given.

For an open set $U \subset \mathbb{R}^k$, $C^r = C^r(U)$ denotes the collection of functions $f : U \to \mathbb{R}$ that have continuous derivatives of order $1, 2, \cdots, r$ on U. As $U \subset \mathbb{R}^k$, it is noted that we refer to all derivatives as partial derivatives

$$\frac{\partial^\alpha f}{\partial^{\alpha_1} x_1 \partial^{\alpha_2} x_2 \cdots \partial^{\alpha_k} x_k}(y_1, y_2, \cdots, y_k),$$

where $\alpha = \sum_{i=1}^k \alpha_i$, $\alpha_i \geqslant 0$ and $(y_1, y_2, \cdots, y_k) \in U$. If $f \in C^r$ for all r, then we say f is smooth, or of class C^∞. We denote by $C^\infty(\boldsymbol{p})$ the collection of all C^∞ functions whose domain includes $\boldsymbol{p}$.

Definition 2.4.1 (Tangent Space at a Point of a Manifold $\mathcal{M}$) *Let $\mathcal{M}$ denote a C^∞ manifold of dimension k. and let $\boldsymbol{p} \in \mathcal{M}$. We define the tangent space $T_{\boldsymbol{p}}\mathcal{M}$ to $\mathcal{M}$ at $\boldsymbol{p}$ to be the set of all operators*

$$Y_{\boldsymbol{p}}, Z_{\boldsymbol{p}} : C^\infty(\boldsymbol{p}) \to \mathbb{R}$$

satisfying, for all $\alpha, \beta \in \mathbb{R}$, $f, g : \mathcal{M} \to \mathbb{R}$ and $f, g \in C^\infty(\boldsymbol{p})$, two conditions:

1. *Linearity:*

$$Y_{\boldsymbol{p}}(\alpha f + \beta g) = \alpha(Y_{\boldsymbol{p}} f) + \beta(Y_{\boldsymbol{p}} g) \tag{2.4.1}$$

2. *Leibniz rule:*

$$Y_{\boldsymbol{p}}(fg) = (Y_{\boldsymbol{p}} f)g(\boldsymbol{p}) + f(\boldsymbol{p})(Y_{\boldsymbol{p}} g) \tag{2.4.2}$$

with the vector space operations in $T_p\mathcal{M}$ defined by

$$(Y_p + Z_p)f = Y_p f + Z_p f,$$
$$(\alpha Y_p)f = \alpha(Y_p f).$$

A tangent vector to $\mathcal{M}$ at p is any $Y_p \in T_p\mathcal{M}$.

Now, we introduce the vector field, which is used to define geodesics.

Definition 2.4.2 (Vector Field) *A vector field $\boldsymbol{X}$ of a class C^r on $\mathcal{M}$ is a function assigning to each point $p \in \mathcal{M}$ a vector $\boldsymbol{X}_p \in T_p\mathcal{M}$ whose components in the frames of any local coordinates (U, φ) are functions of class C^r on the domain U of the coordinates.*

2.4.2 *Geodesics*

In Chapter 4, Wilks' theorem is proved on the sphere $\mathcal{S}^{k-1} \subset \mathbb{R}^k$. In Definition 2.5.2, we will define local alternatives on $\mathcal{S}^{k-1}$. The distance between two points on $\mathcal{S}^{k-1}$ is the length of a minimal geodesic between them. We shall introduce the geodesic as follows; see Boothby (1986, Chapter 7.5). Let $\boldsymbol{p}(t)$ be a t-parametrized curve on the differentiable manifold $\mathcal{M} \subset \mathbb{R}^k$, $a < t < b$, where $\boldsymbol{p}(a)$ and $\boldsymbol{p}(b)$ are two endpoints, and $a, b \in \mathbb{R}$. Let $\boldsymbol{Y}(t) = d\boldsymbol{p}(t)/dt$ be a vector field along the curve. Suppose $\boldsymbol{p}(t) \in C^2$. Then, $d\boldsymbol{Y}/dt$ is obtained by differentiating $\boldsymbol{Y}(t)$ along the curve in $\mathbb{R}^k$. However, $d\boldsymbol{Y}/dt$ may not be tangent to $\mathcal{M}$. Let $\pi : \mathbb{R}^k \to T_p\mathcal{M}$ be an orthogonal project. Thus, we define the covariant derivative, in order to project $d\boldsymbol{Y}/dt$ to a tangent vector.

Definition 2.4.3 (Covariant Derivative) *Let $\pi(d\boldsymbol{Y}/dt)$ be denoted by $D\boldsymbol{Y}/dt$, and we call it the covariant derivative of the tangent vector field $\boldsymbol{Y}$ on $\mathcal{M}$ along the curve $\boldsymbol{p}(t)$.*

Now, we define the geodesic, based on Definition 2.4.3.

Definition 2.4.4 (Geodesic) *The (parametrized) curve $\boldsymbol{p}(t)$ is said to be a geodesic if its velocity vector is constant (parallel), that is, if it satisfies the condition*

$$\frac{D}{dt}\frac{d\boldsymbol{p}}{dt} = 0,$$

the equation of a geodesic, for all $a < t < b$.

Note that the geodesic is not necessarily unique. Take $\mathcal{S}^{k-1}$ as an example:

1. If two endpoints are not antipodal, then those two points partition the unique great circle generated by these two points into two geodesics, and one of them has the minimal length. We call it the minimal geodesic.

2. If two endpoints are antipodal, then any great circle arc connecting those two points is a geodesic. All of those geodesics have the same curve length. They are all minimal geodesics.

2.4.3 log-*map*

A log-map is the inverse map of an exponential map; see Boothby (1986, Chapter 7.6). Those two maps are used to connect the manifold $\mathcal{M}$ and its tangent space $T_{\boldsymbol{\mu}}\mathcal{M}$ at $\boldsymbol{\mu} \in \mathcal{M}$. Since the tangent space is isometric to Euclidean space, a log-map will be an effective tool for transferring the empirical likelihood on the sphere to its tangent space, as we shall see later. This can then be used to develop Wilks' theorem. However, for a point $\boldsymbol{p} \in \mathcal{M}$, there possibly exists a point $\boldsymbol{q} \in \mathcal{M}$, such that there exist at least two distinct minimal geodesics from $\boldsymbol{q}$ to $\boldsymbol{p}$. For example, in $\mathcal{S}^{k-1}$, $\boldsymbol{q}$ is the antipodal point of $\boldsymbol{p}$. Thus, in order to define the following maps, we do not consider points like $\boldsymbol{q}$.

Definition 2.4.5 (Exponential Map at $\boldsymbol{u}$) *Let $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{M}$ and let $\boldsymbol{X_c} \in T_{\boldsymbol{u}}\mathcal{M}$ be a tangent vector, $\boldsymbol{c}(t)$, $t \in [0,1]$, be a parametrized curve on the surface of the manifold $\mathcal{M}$, such that*

$$\boldsymbol{c}(0) = \boldsymbol{u},$$
$$\boldsymbol{c}(1) = \boldsymbol{v}.$$

Then, the exponential map at $\boldsymbol{u}$ is defined as $exp_{\boldsymbol{u}}\boldsymbol{X_c} = \boldsymbol{v}$. That is, the image of the tangent vector $\boldsymbol{X_c}$ is defined to be the point on the unique geodesic determined by $\boldsymbol{X_c}$ where the curve is parametrized as $\boldsymbol{c}(1)$.

Correspondingly, as the inverse map of the exponential map, we give the definition of the log-map as follows.

Definition 2.4.6 (Logarithm Map) *Let $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{M}$. Then the* log*-map uniquely sends $\boldsymbol{v}$ to $\boldsymbol{V} \in T_{\boldsymbol{u}}M$, which is the tangent vector having the direction along the minimal geodesic from $\boldsymbol{u}$ to $\boldsymbol{v}$. We write,* $\log_{\boldsymbol{u}} \boldsymbol{v} = \boldsymbol{V}$.

Definition 2.4.4 implies that $\|d\boldsymbol{c}(t)/dt\|$ is constant along the geodesic $\boldsymbol{c}(t)$, and its length from $\boldsymbol{c}(0)$ to $\boldsymbol{c}(1)$ is:

$$L = \int_0^1 \left\| \frac{d\boldsymbol{c}(t)}{dt} \right\| dt = \int_0^1 \|\boldsymbol{X_c}\| dt = \|\boldsymbol{X_c}\|. \tag{2.4.3}$$

Thus, one property of the log-map and exp-map is that they preserve the distance. In other words, the Euclidean norm of the tangent vector in $T_{\boldsymbol{u}}\mathcal{M}$ equals the distance of the minimal geodesic from $\boldsymbol{u}$ to $\boldsymbol{v}$. Moreover, L in (2.4.3) can be used to define the spherical distance from $\boldsymbol{u}$ to $\boldsymbol{v}$.

Figure 2.4.1 and Figure 2.4.2 illustrate how the log-map is defined on $\mathcal{S}^2 \subset \mathbb{R}^k$. For $\boldsymbol{\mu}_0 \in \mathcal{S}^2$, the image of $\log_{\boldsymbol{\mu}_0}(\boldsymbol{X})$ is expressed as:

$$\boldsymbol{Y}(\boldsymbol{\mu_0}) = (\boldsymbol{I} - \boldsymbol{\mu}_0\boldsymbol{\mu}_0^\top)\boldsymbol{X}\frac{\beta}{\sin\beta},$$

where $\beta = \arccos \boldsymbol{X}^\top \boldsymbol{\mu_0}$, and $\sin\beta = \|(\boldsymbol{I} - \boldsymbol{\mu}_0\boldsymbol{\mu}_0^\top)\boldsymbol{X}\|$.

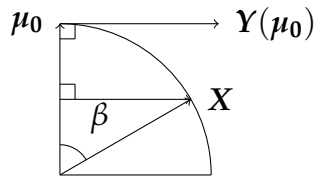


Figure 2.4.1: The log-map image of a random vector $\boldsymbol{X}$ on S^2, as $\boldsymbol{Y}(\boldsymbol{\mu_0})$, observed at the perspective facing to the plane spanned by $\boldsymbol{\mu_0}$ and $\boldsymbol{X}$. By the distance-preserve property of log-map, $\|\boldsymbol{Y}(\boldsymbol{\mu_0})\| = d_{S^2}(\boldsymbol{\mu_0}, \boldsymbol{X})$, where d_{S^2} denotes the spherical distance.

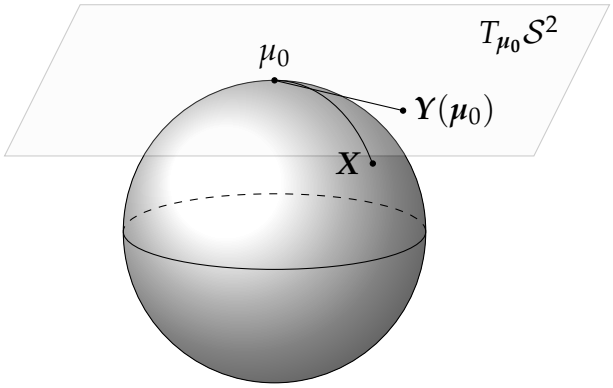


Figure 2.4.2: The log-map from $\mathcal{S}^2$ to $T_{\mu_0}\mathcal{S}^2$ as viewed in $\mathbb{R}^3$.

2.5 SPHERICAL STATISTICS RESULTS

The Wilks' statistic is evaluated either at the population mean as in Theorem 2.1 or local alternatives as in Theorem 3.1. The mean on $\mathcal{S}^{k-1}$ depends on the distance function $d(\cdot, \cdot)$; we refer to this later in (2.5.1) and (2.5.2). Fréchet (1948) defined the so-called Fréchet mean, which is induced by a distance $d(\cdot, \cdot)$. In particular, for the sphere, the Euclidean distance $d_E(\cdot, \cdot)$ and the arc length on the surface $d_{S^{k-1}}(\cdot, \cdot)$ determine two different Fréchet means, respectively labelled as the extrinsic mean and the intrinsic mean. Fisher et al. (1993, Chapter 3.2) gave the definition of the mean direction, which is related to the extrinsic mean on the sphere.

2.5.1 *Mean Direction*

We shall define the mean direction in $\mathcal{S}^2$, while the definition in $\mathcal{S}^k$, $k > 2$, is analogous. Assume $\mathcal{S}^2$ is embedded in $\mathbb{R}^3$, and consider a collection of points $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n \in \mathcal{S}^2$ with the origin at $\boldsymbol{O} = (0,0,0)$, where $\boldsymbol{X}_i$ is a unit vector with polar coordinate (θ_i, ϕ_i), and equivalently direction cosines $(x_1^i, x_2^i, x_3^i) = (\sin\theta_i \cos\phi_i, \sin\theta_i \sin\phi_i, \cos\theta_i)$ as a vector. It should be noted that the angle δ between two arbitrary vectors, say $\boldsymbol{Y}$ and $\boldsymbol{Z}$ is given by the inner product

$$\cos\delta = \boldsymbol{Y}^\top \boldsymbol{Z} = y_1 z_1 + y_2 z_2 + y_3 z_3,$$

where $\delta \in [0, \pi]$.

Definition 2.5.1 (Mean Direction) *For $i = 1, 2, \cdots, n$, $\boldsymbol{X}_i \in \mathcal{S}^2$ has direction cosines (x_1^i, x_2^i, x_3^i). Let*

$$S_1 = \sum_{i=1}^n x_1^i, \quad S_2 = \sum_{i=1}^n x_2^i, \quad S_3 = \sum_{i=1}^n x_3^i,$$

and let

$$R^2 = S_1^2 + S_2^2 + S_3^2.$$

The mean direction of $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$ is

$$\begin{aligned}(\hat{x}_1, \hat{x}_2, \hat{x}_3) =& (S_1/R, S_2/R, S_3/R) \\ =& (\sin\hat{\theta}\cos\hat{\phi}, \sin\hat{\theta}\sin\hat{\phi}, \cos\hat{\theta}),\end{aligned}$$

where $(\hat{\theta}, \hat{\phi})$ are the polar coordinates of the sample mean direction.

2.5.2 *Extrinsic and Intrinsic Mean*

Suppose that $\boldsymbol{Q}$ is a random vector on $\mathcal{M} \subset \mathbb{R}^k$. Then, the Fréchet mean of $\boldsymbol{Q}$ on the manifold $\mathcal{M}$ is

$$\boldsymbol{\mu}_0 = \underset{\boldsymbol{\mu} \in \mathcal{M}}{\operatorname{argmin}} \int_{\mathcal{M}} d(\boldsymbol{\mu}, \boldsymbol{Q})^2 dP(\boldsymbol{Q}),$$

where $P(\cdot)$ is the probability measure on the manifold $\mathcal{M}$, and $d(\cdot,\cdot)$ is the distance function on $\mathcal{M}$.

Correspondingly, for a random sample $\boldsymbol{Q}_1, \boldsymbol{Q}_2, \cdots, \boldsymbol{Q}_n$ with observed quantities $\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_n$, the sample Fréchet mean of $\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_n$ on the manifold $\mathcal{M}$ is

$$\boldsymbol{\mu}_s = \underset{\boldsymbol{\mu}\in\mathcal{M}}{\operatorname{argmin}} \sum_{i=1}^{n} d(\boldsymbol{\mu}, \boldsymbol{q}_i)^2.$$

When $d(\cdot,\cdot)$ is induced by the Euclidean distance $d_E(\cdot,\cdot)$ so $\mathcal{M} \subset \mathbb{R}^k$, Bhattacharya and Patrangenaru (2003, 2005) called the Fréchet mean the extrinsic mean. Thus for $\mathcal{M} = S^{k-1}$, the extrinsic mean direction of $\boldsymbol{Q}$ on $\mathcal{S}^{k-1}$ is

$$\boldsymbol{\mu}_0 = \underset{\boldsymbol{\mu}\in\mathcal{S}^{k-1}}{\operatorname{argmin}} \int_{\mathcal{S}^{k-1}} d_E(\boldsymbol{\mu}, \boldsymbol{Q})^2 dP(\boldsymbol{Q}). \tag{2.5.1}$$

Remark: for $\mathcal{S}^{k-1} \subset \mathbb{R}^k$, let

$$E[\boldsymbol{Q}] = \int_{\mathcal{S}^{k-1}} \boldsymbol{Q} dP(\boldsymbol{Q}),$$

and

$$\begin{aligned}\mathbf{0} =& (\boldsymbol{I} - \boldsymbol{\mu}_0\boldsymbol{\mu}_0^\top)\frac{\partial}{\partial\boldsymbol{\mu}}\Big|_{\boldsymbol{\mu}=\boldsymbol{\mu}_0} \int_{\mathcal{S}^{k-1}} d_E(\boldsymbol{\mu}, \boldsymbol{Q})^2 dP(\boldsymbol{Q}) \\ =& \int_{\mathcal{S}^{k-1}} (\boldsymbol{I} - \boldsymbol{\mu}_0\boldsymbol{\mu}_0^\top)\frac{\partial}{\partial\boldsymbol{\mu}} d_E(\boldsymbol{\mu}, \boldsymbol{Q})^2\Big|_{\boldsymbol{\mu}=\boldsymbol{\mu}_0} dP(\boldsymbol{Q}),\end{aligned}$$

where

$$(\boldsymbol{I} - \boldsymbol{\mu}_0\boldsymbol{\mu}_0^\top)\frac{\partial}{\partial\boldsymbol{\mu}} d_E(\boldsymbol{\mu}, \boldsymbol{Q})^2\Big|_{\boldsymbol{\mu}=\boldsymbol{\mu}_0}$$

is the projection of the gradient of the squared distance function onto the tangent space at $\boldsymbol{\mu}_0$. Then, $\boldsymbol{\mu}_0$ satisfies

$$\mathbf{0} = \int_{\mathcal{S}^{k-1}} (\boldsymbol{I} - \boldsymbol{\mu}_0\boldsymbol{\mu}_0^\top)(\boldsymbol{\mu}_0 - \boldsymbol{Q}) dP(\boldsymbol{Q}),$$

and hence,

$$\begin{aligned}\mathbf{0} &= (\boldsymbol{I} - \boldsymbol{\mu}_0\boldsymbol{\mu}_0^\top) \int_{\mathcal{S}^{k-1}} \boldsymbol{Q} dP(\boldsymbol{Q}) \\ &= E[\boldsymbol{Q}] - \boldsymbol{\mu}_0(\boldsymbol{\mu}_0^\top E[\boldsymbol{Q}]). \\ \Rightarrow \boldsymbol{\mu}_0 &= \frac{E[\boldsymbol{Q}]}{\|E[\boldsymbol{Q}]\|},\end{aligned}$$

the extrinsic mean direction. Similarly, the sample extrinsic Fréchet mean on a sphere turns out to be the sample mean direction.

For the intrinsic mean, if $d(\cdot,\cdot)$ is the metric on S^{k-1}, labelled as $d_{S^{k-1}}(\cdot,\cdot)$, which is the minimal geodesic length of two points on the surface, we find a point on the sphere as follows:

$$\boldsymbol{m}_0 = \underset{\boldsymbol{m}\in\mathcal{S}^{k-1}}{\operatorname{argmin}} \int_{\mathcal{S}^{k-1}} d_{S^{k-1}}(\boldsymbol{m},\boldsymbol{Q})^2 dP(\boldsymbol{Q}), \tag{2.5.2}$$

since $d_{\mathcal{S}^{k-1}}(\cdot,\cdot)$ is defined for points on $\mathcal{S}^{k-1}$. We call it the intrinsic mean.

2.5.3 *Local Alternatives on S^{k-1}*

By considering local alternatives, it is possible to derive a more general version of Wilks' theorem. Through the idea of the log-map, we define local alternatives on the sphere.

Definition 2.5.2 (Local Alternatives on Sphere) *Let $k \geqslant 2$ and let $\boldsymbol{\mu}_0$ denote the mean direction of a random vector $\boldsymbol{X} \in S^{k-1}$. Then for each $n \in \mathbb{N}$, there exists $\boldsymbol{\mu}_n \in S^{k-1}$, such that*

$$\log_{\boldsymbol{\mu}_0}(\boldsymbol{\mu}_n) = \frac{\boldsymbol{\eta}}{\sqrt{n}} \in T_{\boldsymbol{\mu}_0}\mathcal{S}^{k-1},$$

where $\boldsymbol{\eta}$ is a given k-dimensional constant vector in the tangent space at $\boldsymbol{\mu}_0$. We call $\{\boldsymbol{\mu}_n\}_{n\geqslant 1}$ a sequence of local alternatives.

Note that we write local alternatives as:

$$\boldsymbol{\mu}_n = \boldsymbol{\mu}_0 \cos\Big(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\Big) + \frac{\boldsymbol{\eta}}{\|\boldsymbol{\eta}\|}\sin\Big(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\Big),$$

where $\boldsymbol{\mu}_n$ is not antipodal of $\boldsymbol{\mu}_0$. This will be widely used in the proof of Theorem 4.1.

Remark: we define $\{\boldsymbol{\mu}_n\}_{n\geqslant 1}$ through the first order derivative to $\boldsymbol{\mu}_0$. In Euclidean space, it is obvious to see $\boldsymbol{\eta}/\sqrt{n}$ is the derivative between $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_n$. However, for the sphere, we need $T_{\boldsymbol{\mu}_0}\mathcal{S}^{k-1}$ to find the frist order derivative $\boldsymbol{\eta}/\sqrt{n}$ and then find $\boldsymbol{\mu}_n \in \mathcal{S}^{k-1}$, where $T_{\boldsymbol{\mu}_0}\mathcal{S}^{k-1} \subset \mathbb{R}^k$.

2.5.4 *Useful Distributions on the Sphere*

In Section 4.5 and 4.6, experiments are based on samples from von Mises-Fisher distributions and Kent distributions. First we introduce two concepts in spherical distributions. The unimodal distribution refers to a distribution with a clear peak (mode) or a most frequent value, while multimodal distributions have more than one peak. Now we introduce these two distribution in brief.

von Mises-Fisher Distribution

In Mardia and Jupp (2009), a random vector $\boldsymbol{X} \in \mathcal{S}^{p-1} \subset \mathbb{R}^p$, with $p-1$ dimensional von Mises-Fisher distribution $M_p(\boldsymbol{\mu}, \kappa)$, has the following density function:

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \kappa) = \left(\frac{\kappa}{2}\right)^{p/2-1} \frac{1}{\Gamma(p/2) I_{p/2-1}(\kappa)} \exp\{\kappa \boldsymbol{\mu}^\top \boldsymbol{x}\},$$

where $\kappa \geqslant 0$, $\|\boldsymbol{\mu}\| = 1$, and

$$I_q(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} \cos q\theta \exp\{\kappa \cos\theta\} d\theta.$$

It is noted that $\boldsymbol{\mu}$ is the mean direction and κ is called the concentration parameter. For $\kappa > 0$, there is a mode at the mean direction $\boldsymbol{\mu}$; for $\kappa = 0$, the distribution is uniform on $\mathcal{S}^{p-1}$. The larger κ is, the more concentrated the population is around the mean direction.

Kent Distribution

The Kent distribution, denoted by $FB_5(\kappa, \beta, \boldsymbol{\Gamma})$, was proposed by Kent (1982) for a random vector $\boldsymbol{X} \in \mathcal{S}^2 \subset \mathbb{R}^3$. It has the following density function:

$$f(\boldsymbol{x}; \kappa, \beta, \boldsymbol{\Gamma}) = c(\kappa, \beta)^{-1} \exp\{\kappa \gamma_1^\top \boldsymbol{x} + \beta[(\gamma_2^\top \boldsymbol{x})^2 - (\gamma_3^\top \boldsymbol{x})^2]\},$$

where $\kappa \geqslant 0$, $\beta \geqslant 0$ and a 3×3 orthogonal matrix $\boldsymbol{\Gamma} = \{\gamma_1, \gamma_2, \gamma_3\}$.

It is noted that κ is called the concentration parameter, as is explained in von Mises-Fisher distribution. The ovalness parameter β describe the ellipticity of the population. As is shown by Fisher et al. (1993), the greater β is,

the more the population is away from circular symmetry. This distribution has two main forms; if $\kappa/\beta \geqslant 2$, Kent distribution is unimodal; if $\kappa/\beta < 2$, the distribution is bimodal. For the matrix $\mathbf{\Gamma}$, γ_1 is the mean direction, γ_2 is the major axis and γ_3 is the minor axis.

In this thesis, we only consider the unimodal Kent distribution, i.e.

$$\kappa/\beta \geqslant 2.$$

2.6 MODEL: 3-SPIDER

The 3-Spider is regarded as a space comprising 3 disjoint rays, which share at a common endpoint; see Hotz et al. (2013).

Definition 2.6.1 (3-Spider) *The 3-Spider is a union of 3 rays, so-called legs, which only join together by their endpoints. As is shown in Figure 2.6.3, the endpoints of those three rays are glued at the point O, with each leg* $L_i = \mathbb{R}_{\geqslant 0}$, *and* $L_i^+ = \mathbb{R}_{>0}$, $i = 1, 2, 3$. *In particular, we call the joined endpoint O the cone point* L_0.

Let 3-Spider space be denoted by $\mathcal{SP}_3$. Note that the point where those three rays join together is a singularity point. This is explained in Section 5.2. Now, for any point $p \in \mathcal{SP}_3$, its coordinate is given as (p_0, k), where $k = 1, 2, 3$. The second coordinate shows the leg index p has and the first coordinate, $p_0 \geqslant 0$, gives the location on L_k.

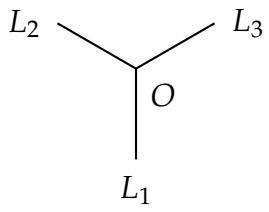


Figure 2.6.3: The sketch of 3-Spider, $\mathcal{SP}_3$.

2.6.1 *Folding Map on 3-Spider*

Similar to the intrinsic distance on the sphere, the distance on the 3-Spider $\mathcal{SP}_3$ is not completely Euclidean. However, the distance is still defined as the length of the shortest path. For points located on the same leg, the metric is Euclidean; for points respectively located on different legs, the shortest path goes through the cone point and we thus define the distance. In order to define the 3-Spider distance, we first define the so-called folding map; see Hotz et al. (2013, Definition 1.2):

Definition 2.6.2 (Folding Map) *Let $p = (p_0, k) \in \mathcal{SP}_3$ with $k \in \{1, 2, 3\}$. Then, for $j \in \{1, 2, 3\}$, define the folding map $F_j : \mathcal{SP}_3 \to \mathbb{R}$ by*

$$F_j p = F_j(p_0, k) = \begin{cases} p_0 & \text{if } k = j, \\ -p_0 & \text{if } k \neq j, \end{cases}$$

Now, we define the distance on $\mathcal{SP}_3$:

Definition 2.6.3 (Distance on $\mathcal{SP}_3$) *For $p = (p_0, j)$ and $q = (q_0, k)$, define the 3-Spider distance*

$$d_{\mathcal{SP}_3}(p, q) = \begin{cases} |p_0 - q_0| & \text{if } k = j, \\ p_0 + q_0 & \text{if } k \neq j, \end{cases}$$

2.6.2 *Mean on $\mathcal{SP}_3$*

The object of the research in Chapter 5 is about the statistical behaviour of the i.i.d. random variables on $\mathcal{SP}_3$. Suppose P_0 is a Borel probability measure on $\mathcal{SP}_3$. Throughout this thesis, we make the following assumptions: under the measure P_0, the squared integral satisfies

$$\int_{\mathcal{SP}_3} d_{\mathcal{SP}_3}(s, q)^2 dP_0(q) < \infty, \tag{2.6.1}$$

where s denotes the cone point L_0, and also the nondegeneracy satisfies

$$P_0(L_k^+) \in (0,1), \quad k = 1,2,3. \tag{2.6.2}$$

Note that the square integrability (2.6.1) implies:

$$\begin{aligned}\int_{\mathcal{SP}_3} d^2_{\mathcal{SP}_3}(s,q)dP_0(q) &= \int_{\mathcal{SP}_3} d^2_E(0,F_kq)dP_0(q)\\ &= \int_{\mathcal{SP}_3} (F_kq)^2 dP_0(q)\\ &< \infty.\end{aligned} \tag{2.6.3}$$

where $k = 1,2,3$.

The square integrability implies integrability, through Lyapunov's inequality, as is shown in Corollary 2.6.2. To prove this result, we first give Hölder inequality for random variables, as shown in Chung (2001). Consider a random variable X with distribution P, i.e. $X \sim P$, the expectation of X is

$$E_X[X] = \int x dP(x).$$

Corollary 2.6.1 (Hölder's Inequality for Random Variables) *Suppose X, Y are two random variables, and $p, q > 1$ such that*

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Then,

$$|E_{X,Y}[XY]| \leqslant E_{X,Y}[|XY|] \leqslant \left\{E_X[|X|^p]\right\}^{1/p}\left\{E_Y[|Y|^q]\right\}^{1/q},$$

and

$$\{E_{X,Y}[|X+Y|^p]\}^{1/p} \leqslant E_X[|X|^p]^{1/p} + E_Y[|Y|^p]^{1/p}.$$

Based on Corollary 2.6.1, we give an application of Hölder's inequality for random variables, Lyapunov's inequality.

Corollary 2.6.2 (Lyapunov's Inequality) *Suppose* $Y = 1$ *with probability* 1. *Then, for* $1 < p < \infty$

$$E_X[|X|] \leqslant \left\{E_X[|X|^p]\right\}^{1/p}.$$

Moreover, for $1 < r < p$,

$$E_X[|X|^r] \leqslant \left\{E_X[|X|^{pr}]\right\}^{1/p}.$$

To prove Lyapunov's inequality, let $s = pr$. Then,

$$E_X[|X|^r] \leqslant \left\{E_X[|X|^s]\right\}^{r/s},$$

and thus

$$\left\{E_X[|X|^r]\right\}^{1/r} \leqslant \left\{E_X[|X|^s]\right\}^{1/s},$$

for $1 < r < s < \infty$.

Corollary 2.6.2 (Lyapunov's inequality) now implies that:

$$\begin{aligned}\int_{\mathcal{SP}_3} d_{\mathcal{SP}_3}(s,q)dP_0(q) &= \int_{\mathcal{SP}_3} |F_k q|dP_0(q) \\ &\leqslant \Big(\int_{\mathcal{SP}_3} (F_k q)^2 dP_0(q)\Big)^{1/2} \\ &< \infty,\end{aligned} \tag{2.6.4}$$

where $k = 1, 2, 3$.

That is, under our assumption (2.6.3), the expectation of $d_{\mathcal{SP}_3}(s,q)$ is bounded under the measure P_0. Note that the assumption (2.6.3) is used to define the Fréchet mean on $\mathcal{SP}_3$.

Due to the singularity of the cone point on $\mathcal{SP}_3$, the folding map is applied to define the distance on $\mathcal{SP}_3$. Thus, recalling the intrinsic mean (2.5.2) on $\mathcal{S}^{k-1}$, we give the Fréchet mean on $\mathcal{SP}_3$:

Definition 2.6.4 (Fréchet Mean on 3-Spider) *Let* Ω *be a sample space. Let* $q_1, q_2, \cdots, q_n : \Omega \to \mathcal{SP}_3$ *be i.i.d. random variables with Borel probability measure* P_0. *Under the squared integrability (2.6.1) and the nondegeneracy (2.6.2) assumptions, the population Fréchet mean* $\bar{b}$ *on* $\mathcal{SP}_3$ *is defined as*

$$\bar{b} = \underset{p \in \mathcal{SP}_3}{\operatorname{argmin}} \int_{\mathcal{SP}_3} d_{\mathcal{SP}_3}(p,q)^2 dP_0(q).$$

Correspondingly, the sample Fréchet mean is given by

$$b_n = \underset{p \in \mathcal{SP}_3}{\operatorname{argmin}} \sum_{i=1}^{n} d_{\mathcal{SP}_3}(p, q_i)^2.$$

Hotz et al. (2013) shows that there are 3 different cases for the location of the Fréchet mean on $\mathcal{SP}_3$, according to the behaviour of the first moment. We give the definition of the first moment as follows.

Definition 2.6.5 (First Moment under the Folding Map) *Let $q = (q_0, k)$, $q_0 \in \mathbb{R}_{\geqslant 0}$ and $k \in \{1, 2, 3\}$. The first moment of the measure P_0 on the kth leg L_k, $k = 1, 2, 3$, is the real number*

$$m_k = \int_{\mathbb{R}} q_0 dP_k(q) = \int_{\mathcal{SP}_3} (F_k q) dP_0(q),$$

where $P_k = P_0 \circ F_k^{-1}$ is a Borel probability measure on $\mathbb{R}$, the image space. Further, define the sample first moment

$$\hat{m}_k = \frac{1}{n}(\hat{S}_k - \sum_{i \neq k} \hat{S}_i), \tag{2.6.5}$$

where $\hat{S}_k = \sum_{i:q_i \in L_k} |F_k q_i|$.

The equality (2.6.5) implies that

$$\hat{m}_k = \frac{1}{n} \sum_{i=1}^{n} F_k q_i.$$

From (2.6.4), we have

$$m_k = \int_{\mathcal{SP}_3} F_k q dP_0(q) \leqslant \int_{\mathcal{SP}_3} |F_k q| dP_0(q) < \infty.$$

Similarly, we also define

$$S_k = \int_{L_k} (F_k q) dP_0(q),$$

for $k = 1, 2, 3$.

Example 2.6.1 (The Fréchet Mean and The First Moments on $\mathcal{SP}_3$) *Consider three points respectively on each leg of the 3-Spider: $p_k \in L_k$, for $k = 1, 2, 3$ and $s \in L_0$. Initially, let these three points have the same distances to the cone point, as*

shown in Figure 2.6.4 (left). Then, $S_k < \sum_{i \neq k} S_i$, $m_k < 0$, $k = 1, 2, 3$, and it is not hard to see that the cone point is the Fréchet mean. Next as in Figure 2.6.4 (centre), move p_1 gradually away from L_0, until $d_{\mathcal{SP}_3}(s, p_1) = d_{\mathcal{SP}_3}(s, p_2) + d_{\mathcal{SP}_3}(s, p_3)$. Now, we have $S_1 = S_2 + S_3$, while $m_1 = 0$, $m_2 < 0$ and $m_3 < 0$. However, the Fréchet mean is still on the cone point. As Figure 2.6.4 (right) shows, we keep moving p_1 on L_1 downwards. Now, we find $S_1 > S_2 + S_3$, while $m_1 > 0$, $m_2 < 0$ and $m_3 < 0$. Hence, the Fréchet mean is strictly off the cone point.

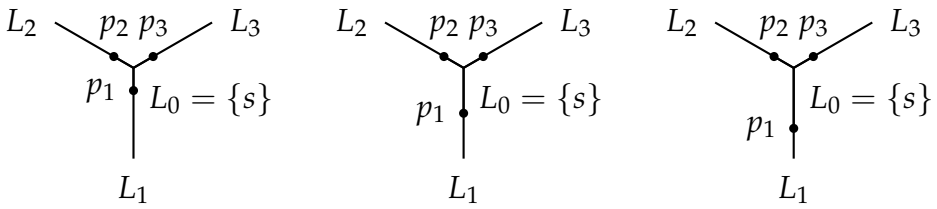


Figure 2.6.4: The Fréchet mean on $\mathcal{SP}_3$.

As illustrated above, the first moment on each leg varies when moving that point. Thus, Hotz et al. (2013) summarized the behaviour of the first moment m_k, $k = 1, 2, 3$, as follows.

Theorem 2.2 *Suppose integrability (2.6.4) and nondegeneracy (2.6.2) hold. Then, either $m_k < 0$ for all indices $k \in \{1, 2, 3\}$, or there is exactly one index $j \in \{1, 2, 3\}$ such that $m_j \geqslant 0$, in which case either:*

1. *$m_j > 0$, or*
2. *$m_j = 0$.*

With Theorem 2.2, we know the behaviour of m_k, $k = 1, 2, 3$, and thus we name these three different cases:

Definition 2.6.6 *Under integrability (2.6.4) and nondegeneracy (2.6.2), we say that the mean of the measure P on $\mathcal{SP}_3$ is either:*

1. ***Sticky** if $m_k < 0$ for all indices $k \in \{1, 2, 3\}$, or*

2. ***Non-Sticky*** *if* $m_j > 0$ *for some (unique)* $j \in \{1, 2, 3\}$, *or*

3. ***Partially-Sticky*** *if* $m_j = 0$ *for some (unique)* $j \in \{1, 2, 3\}$.

Hotz et al. (2013) presented the asymptotic behaviour of the sample Fréchet mean b_n by the Law of Large Number (LLN). For the sample space Ω and for $\omega \in \Omega$, denote the random sample by $\{q_1(\omega), \cdots, q_n(\omega)\}$, and the sample Fréchet mean by $b_n(\omega)$.

Lemma 2.6.1 (Strong Law of Large Numbers) *Under the square integrability (2.6.1),*

$$\lim_{n\to\infty} \boldsymbol{b}_n(\omega) = \bar{\boldsymbol{b}},$$

with probability 1, *where* $\bar{\boldsymbol{b}}$ *is unique, and is the population Fréchet mean of the probability measure P.*

Based on Lemma 2.6.1, Hotz et al. (2013) proposed the Law of Large Number on $\mathcal{SP}_3$.

Theorem 2.3 (Sticky LLN) *Suppose the nondegeneracy assumption (2.6.2) holds.*

1. *If the moment* m_k *satisfies* $m_k < 0$, *then there is a random integer* $N^*(\omega)$ *such that* $b_n(\omega) \notin L_k^+$ *for all* $n \geqslant N^*(\omega)$ *holds almost surely. Furthermore,* $\bar{b} \notin L_k^+$.

2. *If the moment* m_j *satisfies* $m_j > 0$, *then there is a random integer* $N^*(\omega)$ *such that* $b_n(\omega) \in L_j^+$ *for all* $n \geqslant N^*(\omega)$ *holds almost surely. Furthermore,* $\bar{b} \in L_j^+$.

3. *If the moment* m_j *satisfies* $m_j = 0$, *then there is a random integer* $N^*(\omega)$ *such that* $b_n(\omega) \in L_j^+$ *for all* $n \geqslant N^*(\omega)$ *holds almost surely. Furthermore,* $\bar{b} \in L_0$.

2.7 PHYLOGENETIC TREE DATA

In biology, the evolutionary ancestry of a set of tips is usually presented by a phylogenetic tree. These tips are typically referred to species, but can

also be referred to organisms, genes, or populations; see Baum and Smith (2013). It is easy to understand that the ancestry in biology is a biological relation that links parents, or more generally ancestors, and their children, or descendants. In modern biology, the phylogenetic tree is a useful tool to visualize evolution for descendants from common ancestors.

Hall (2011, Chapter 1) includes the definition of the phylogenetic tree. A phylogenetic tree consists of nodes and branches. The branches are the ancestry sequences through the nodes, i.e. the segments connecting those nodes. The nodes at the tips of branches stand for the the taxa that exist and that we can actually examine. There are two types of nodes, internal nodes and external nodes. The internal nodes represent the ancestral taxa, whose properties can only be inferred from the existing taxa. As is shown in Figure 2.7.5, taking Subfigure (a) as an example, we observe 7 nodes $0, 1, \cdots, 6$, with 4 external nodes $2, 4, 5, 6$ and 3 internal nodes $0, 1, 3$, where the node 0 is labelled as the root of the tree. There are two types: internal edges and external edges. An internal edge connects two internal taxa, while an external edge connects an internal taxon and an external taxon. Taking subfigure (a) in Figure 2.7.5 as an example again, edges a, c are internal edges, while edge b is an external edge. Note that the numbers labelled on the edges refer to the number of changes in a particular sequence along the branch, and we call those numbers the "lengths" of the edges. If the exact numbers are not presented, the relative lengths of the edges are drawn in proportion to the number of changes along the branch.

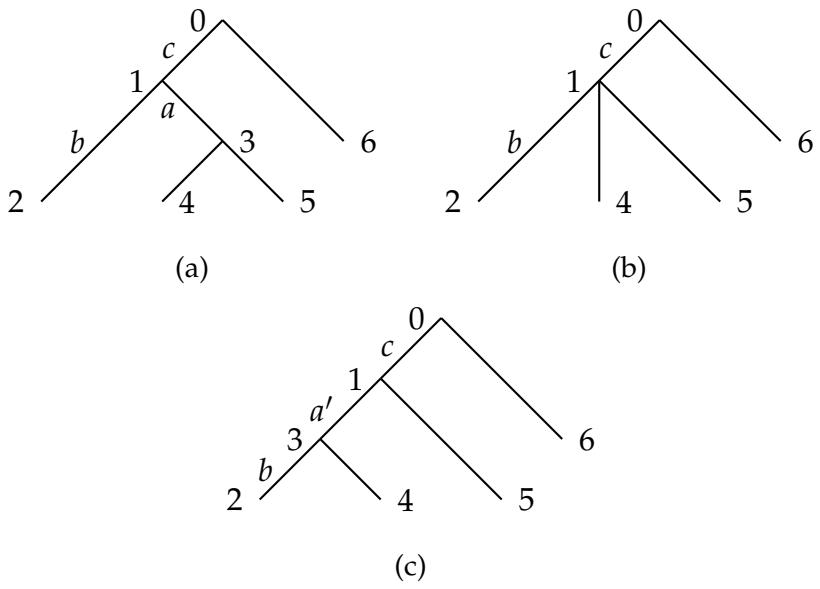


Figure 2.7.5: Subfigure (a): an example of a phylogenetic tree on tip set $\{2,4,5,6\}$ with root 0 and inner nodes $1,3$. "a" and "c" are two internal edges, and "b" is an external edge. Subfigure (b) and (c): shrink the internal edge a to 0, and then expand the internal edge a' from 0.

Phylogenetic tree can be written in the form of data to show such relations. As is shown in Figure 2.7.6, the tree can be written as

$$(S_2 : b, S_4 : d) : c, S_3 : a)$$

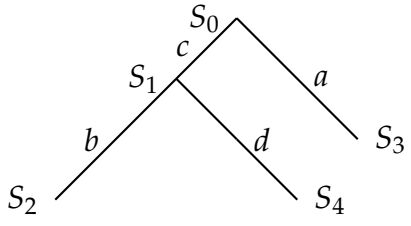


Figure 2.7.6: An example of expressing phylogenetic tree in the form of data.

In the explanation of phylogenetic trees, Baum and Smith (2013) use the following example. Suppose there were N cards in a row with their suit

noted. We use suit changing to refer to the biological mutation in genes or species. Now, M out of N cards changed their suit and the rest $N - M$ were changed. Consider the average change for each card. From this, Baum and Smith (2013) introduce evolutionary distance: the average number of changes of suit per card. The numbers in edges of the phylogenetic tree are labelled with evolutionary distances among generations. Intuitively, every inner nodes can be regarded as the most recent common ancestor of the corresponding descendents. For example, as shown in Subfigure (a) of Figure 2.7.5, node 3 is the most recent common ancestor of node 4 and 5. For internal edges, they give a proportion of the most recent common ancestor evolve to one of next descendent species with a feature different from another descendent branch. Take Subfigure (a) of Figure 2.7.5 as an example again, if 0 is a certain species of rodent animals, then c can be regarded as a percentage of the species 0 evolving to a species with a fuzzy tail 1 while species 6 does not have fuzzy tail. Thus, for a specific set of evolutionary distances on edges, the phylogenetic tree is uniquely determined. As Figure 2.7.5 shows the property of tree structure changing and the internal edge length changing for phylogenetic tree data, we can apply the phylogenetic tree dataset to $\mathcal{SP}_3$; we will refer to it later in Section 5.8.

2.8 MEASURE THEORY RESULTS

In the proof of Theorem 3.1, a version of Wilks' theorem for empirical likelihood in $\mathbb{R}^k$ under local alternatives, we need to apply the dominated convergence theorem, as stated by Chung (2001, Chapter 3.2). Some useful lemmas included by Owen (2001, Chapter 11) are also adapted to the context of this thesis. Note that in these lemmas and Theorem 3.1, we desire the norms of some key terms to be negligible so that the leading term will asymptotically decide the distribution of the Wilks' statistic. We use stochastic orders to express these terms; see Bishop et al. (2007). Now, we use the notation $(\Omega, \mathcal{B}, P)$ as a probability space for introducing some measure theory results.

If the object of the study is a single random variable, then $\Omega = \mathbb{R}$; otherwise, for a random vector, the probability space is $(\mathbb{R}^k, \sigma(\mathcal{B}^k), P)$, where $\mathcal{B}^k = \{(A_1, A_2, \cdots, A_k) : A_j \in \mathcal{B}, j = 1, 2, \cdots, k\}$ and $\sigma(\mathcal{B}^k)$ is the σ-algebra generated on $\mathcal{B}^k$. However, other choices of Ω are possible, e.g. $\Omega = \mathcal{S}^{k-1}$. First, the frequently-used types of convergences throughout the thesis are shown; see Chung (2001, Chapter 4). Note that a random variable X is a map $X : \Omega \to \mathbb{R}$.

Definition 2.8.1 (Convergence Almost Surely) *Let $\{X_n\}_{n\geqslant 1}$ be a sequence of random variables. $\{X_n\}_{n\geqslant 1}$ converges to a random variable X almost surely, if and only if there exits a set $N \in \Omega$ with $P(N) = 0$ such that*

$$\forall \omega \in N^c, \lim_{n\to\infty} X_n(\omega) = X(\omega) \in (-\infty, \infty).$$

Definition 2.8.2 (Convergence in Probability) *The sequence of random variables $\{X_n\}_{n\geqslant 1}$ converges to a random variable X in probability, if and only if for all $\epsilon > 0$,*

$$\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0.$$

Definition 2.8.3 (Convergence in Distribution) *The sequence of random variables $\{X_n\}_{n\geqslant 1}$ converges to a random variable X in distribution, if and only if the corresponding distribution functions $\{F_n\}_{n\geqslant 1}$ converge to the distribution function F of X, i.e.*

$$F_n(x) \to F(x),$$

as $n \to \infty$ for all $x \in \mathbb{R}$, where x is a continuity point of F.

We use the following notation for the convergences above:

1. Convergence Almost Surely. $X_n \xrightarrow{a.s.} X$, as $n \to \infty$.
2. Convergence in Probability. $X_n \xrightarrow{P} X$, as $n \to \infty$.
3. Convergence in Distribution. $X_n \xrightarrow{D} X$, as $n \to \infty$.

Theorem 2.4 (Dominated Convergence Theorem) *X, Y are random variables, and Λ is a set in $\mathcal{B}$. For a sequence of random variables $\{X_n\}_{n\geqslant 1}$, if $\lim_{n\to\infty} X_n = X$ a.s. on $\Lambda \in \mathcal{B}$ and $\forall n$: $|X_n| \leqslant Y$ a.s. on Λ, with $\int_\Lambda Y dP(\boldsymbol{x}) < \infty$, then*

$$\lim_{n\to\infty} \int_\Lambda X_n dP(\boldsymbol{x}) = \int_\Lambda X dP(\boldsymbol{x}) = \int_\Lambda \lim_{n\to\infty} X_n dP(\boldsymbol{x}).$$

Definition 2.8.4 ($X_n = o_p(b_n)$) *Let $\{X_n\}_{n\geqslant 1}$ be a sequence of random variables and $\{b_n\}_{n\geqslant 1}$ be a sequence of strictly positive real numbers. We say:*

1. *$X_n = o_p(1)$ if for every $\epsilon > 0$,*

$$\lim_{n\to\infty} P\{|X_n| \leqslant \epsilon\} = 1.$$

2. *$X_n = o_p(b_n)$ if*

$$X_n / b_n = o_p(1),$$

or equivalently, $X_n = b_n o_p(1)$.

Definition 2.8.5 ($X_n = O_p(b_n)$) *Let $\{X_n\}_{n\geqslant 1}$ be a sequence of random variables and $\{b_n\}_{n\geqslant 1}$ be a sequence of strictly positive real numbers. We say:*

1. *$X_n = O_p(1)$ if for every $\eta > 0$, there exists a constant $K(\eta)$ and an integer $N(\eta)$ such that if $n \geqslant N(\eta)$, then*

$$P\{|X_n| \leqslant K(\eta)\} \geqslant 1 - \eta.$$

2. *$X_n = O_p(b_n)$ if*

$$X_n / b_n = O_p(1),$$

or equivalently, $X_n = b_n O_p(1)$.

From the above definitions, we easily get the following result about the convergence in distribution.

Lemma 2.8.1 *Suppose $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$. Then, $\|\boldsymbol{X}_n\| = O_p(1)$.*

Let $\boldsymbol{\mu}$ be the population mean of i.i.d. random vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$, and

$$\overline{\boldsymbol{X}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i.$$

Then, the Central Limit Theorem (CLT) states $\sqrt{n}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}) \xrightarrow{D} \boldsymbol{Z}$, where $\boldsymbol{Z}$ is a normal random vector. Then, by the above lemma, $\|\sqrt{n}(\overline{\boldsymbol{X}} - \boldsymbol{\mu})\| = O_p(1)$. Equivalently, $\|\overline{\boldsymbol{X}} - \boldsymbol{\mu}\| = O_p(n^{-1/2})$. This will be useful later when establishing bounds of some key terms.

Now, in the proof of Theorem 3.1, Theorem 4.1 and related results, we will adapt the following lemmas in Owen (2001) to the context of the proof.

Lemma 2.8.2 *Let $Y_1, Y_2, \cdots, Y_n$ be i.i.d. random variables and $E(Y_i^2) < \infty$. Let $Z_n = \max_{1 \leqslant i \leqslant n} |Y_i|$. Then $Z_n = o_p(n^{1/2})$.*

Lemma 2.8.3 *Let $Y_i, Y_2, \cdots, Y_n$ be i.i.d. random variables and suppose that $E(Y_i^2) < \infty$. Then*

$$\frac{1}{n}\sum_{i=1}^{n} |Y_i|^3 = o_p(n^{1/2}).$$

Lemma 2.8.4 *Let P_0 be a distribution on $\mathbb{R}^k$ with mean $\boldsymbol{\mu_0}$ and covariance matrix $\boldsymbol{V}_0$ of full rank k. Let Θ be the set of unit vectors in $\mathbb{R}^k$. Then for $\boldsymbol{X}$ with distribution P_0, i.e. $\boldsymbol{X} \sim P_0$,*

$$\inf_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} P_0\Big(\boldsymbol{\theta}^\top (\boldsymbol{X} - \boldsymbol{\mu}_0) > 0\Big) > 0.$$

2.8.1 *Glivenko-Cantelli Theorem*

Glivenko-Cantelli theorem for the uniform convergence ranging over a class of functions is one of the important conclusions in empirical distribution theory as reported by van der Vaart (1998, Chapter 19.2) and van der Vaart and Wellner (1996, Chapter 2.8).

For the following definition and theorem, first let $\{\boldsymbol{X}_n\}_{n \geqslant 1}$ and $\{\boldsymbol{X}\}$ be real random vectors.

Now, we define

$$\begin{aligned} P_n f &= \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{X}_i), \text{ and} \\ P_0 f &= \int f dP_0. \end{aligned} \tag{2.8.1}$$

Recall that $\mathcal{P}$ is a class of probability measure on $\left(\mathbb{R}^k, \sigma(\mathcal{B}^k)\right)$. Following van der Vaart and Wellner (1996, Chapter 2.8), we have

Definition 2.8.6 (Glivenko-Cantelli Uniformly in $P \in \mathcal{P}$) *A class $\mathcal{F}$ of measurable functions on a measurable space $\left(\mathbb{R}^k, \sigma(\mathcal{B}^k)\right)$ is said to be Glivenko-Cantelli uniformly in $P \in \mathcal{P}$ for a given class $\mathcal{P}$ of probability measures on $\left(\mathbb{R}^k, \sigma(\mathcal{B}^k)\right)$ if*

$$\sup_{P\in\mathcal{P}} P\Big(\sup_{m\geqslant n} \sup_{f\in\mathcal{F}} |P_m f - Pf| > \epsilon \Big) \to 0,$$

for every $\epsilon > 0$ as $n \to \infty$.

Note that

$$\left\{ \sup_{m\geqslant n} \sup_{f\in\mathcal{F}} |P_m f - Pf| > \epsilon \right\}$$

is assumed to be P-measurable. Definition 2.8.6 is used for Theorem 2.5. Now, we introduce notation used in Glivenko-Cantelli theorem. Suppose that Q is a probability measure, and $r \in \mathbb{N}$. Then, the $L_r(Q)$-norm is given by

$$\|f\|_{Q,r} = (Q|f|^r)^{1/r}. \tag{2.8.2}$$

Then, we define the $L_r(Q)$-ball of radius ϵ centred at some $f \in \mathcal{F}$ by

$$\{g \in \mathcal{F} : \|g - f\|_{Q,r} < \epsilon\} = \{g \in \mathcal{F} : (Q|g - f|^r)^{1/r} < \epsilon\}.$$

Definition 2.8.7 (Covering Number) *Let $\mathcal{F}$ be a class of measurable functions, let Q be a probability measure, $r \in \mathbb{N}$, and let $\epsilon > 0$. Then, the covering number*

$$N(\epsilon, \mathcal{F}, L_r(Q))$$

is the minimal number of $L_r(Q)$-balls of radius ϵ needed to cover the set $\mathcal{F}$.

The Glivenko-Cantelli theorem has been established in many settings. For example, see Pollard (2012, Chapter 2.4) and van der Vaart (1998, Chapter 19.2). Now, we introduce the version shown by van der Vaart and Wellner (1996). First, we call F an envelope function of the class $\mathcal{F}$, if $|f(\boldsymbol{x})| \leqslant F(\boldsymbol{x})$, for all $\boldsymbol{x}$ and $f \in \mathcal{F}$.

Theorem 2.5 (Glivenko-Cantelli Uniformly in P) *Let $\mathcal{F}$ be a P-measurable class of functions on a measurable space for every probability measure P in a class $\mathcal{P}$, and let $\|\cdot\|_{Q,1}$ be defined as in (2.8.2). Suppose that, for some measurable envelope function F and a fixed number M,*

$$\lim_{M\to\infty} \sup_{P\in\mathcal{P}} PF\{F > M\} = 0,$$

$$\sup_{Q\in\mathcal{Q}_n} \log N\Big(\epsilon\|F\|_{Q,1}, \mathcal{F}, L_1(Q)\Big) = o(n), \text{ for every } \epsilon > 0,$$

where the supremum is taken over the set $\mathcal{Q}_n$ of all discrete probability measures with atoms of size integer multiples of $1/n$. Then $\mathcal{F}$ is Glivenko-Cantelli uniformly in $P \in \mathcal{P}$.

As van der Vaart and Wellner (1996) pointed out, VC classes of sets or functions satisfy the inequality

$$\sup_{Q\in\mathcal{Q}_n} \log N\Big(\epsilon\|F\|_{Q,1}, \mathcal{F}, L_1(Q)\Big) < \infty, \tag{2.8.3}$$

for the supremum taken over all probability measures Q; we show it for more detail in Lemma 2.8.5. However, we require that VC classes of sets or functions are suitably measurable. As Bolthausen et al. (2002, Page 401) explained, for the condition "suitably measurable" to hold, it suffices that for all f in the class there exists a countable collection $\mathcal{G}$ of functions, such that each $f \in \mathcal{F}$ is the pointwise limit of a sequence $\{g_m\}_{m\geqslant 1}$ in $\mathcal{G}$.

Vapnik and Chervonenkis (1971) proposed the concept of VC class. Recalling Definition 2.8.7, we let $N(\epsilon\|F\|_{Q,1}, \mathcal{F}, L_1(Q))$ denote the minimal number of $L_1(Q)$-ball of radius $\epsilon\|F\|_{Q,1}$ covering the class $\mathcal{F}$. The $L_1(Q)$-balls of radius $\epsilon\|F\|_{Q,1}$ centred at some $f \in \mathcal{F}$ is

$$\{g \in \mathcal{F} : \|g - f\|_{Q,1} < \epsilon\|F\|_{Q,1}\} = \{g \in \mathcal{F} : Q|g - f| < \epsilon QF\},$$

To find a finite covering number, van der Vaart (1998, Chapter 19.15) and van der Vaart and Wellner (1996, Chapter 2.8) presented that if the class $\mathcal{F}$ is any VC class of functions, then the condition (2.8.3) holds for the supremum taking over all probability measures over Q. Now we introduce VC class as follows.

For a sample space $\mathscr{X}$ and a collection $\mathcal{C}$ of subsets of $\mathscr{X}$, we say $\mathcal{C}$ picks out a certain subset A of the finite set $\{x_1, x_2, \cdots, x_n\} \subset \mathscr{X}$, if $A = \{x_1, x_2, \cdots, x_n\} \cap C$ for some $C \in \mathcal{C}$. Using this concept, we say $\mathcal{C}$ shatters $\{x_1, x_2, \cdots, x_n\}$ if $\mathcal{C}$ picks out each of its 2^n subsets. Then, we use the smallest n for which no set of size n is shattered by $\mathcal{C}$ for the VC index $V(\mathcal{C})$ of $\mathcal{C}$. Now, we define a VC-class of sets as follows.

Definition 2.8.8 (VC Class of Sets) *A collection $\mathcal{C}$ of measurable sets is called a VC class if its index $V(\mathcal{C})$ is finite.*

Note that, for a set with number of elements no less than $V(\mathcal{C})$, a VC class of sets picks out less than 2^n subsets.

Example 2.8.1 (VC Class of Sets) *Let $\mathscr{X} = \mathbb{R}$, and the finite set be $\{x_1, x_2\}$. Assume $a < x_1 < b < x_2 < c < d$. Now, let $\mathcal{C} = \Big\{(a,b), (c,d), (b,c), (a,d)\Big\}$. Since*

$$\begin{aligned}
\{x_1, x_2\} \cap (b,c) &= \{x_2\}, \\
\{x_1, x_2\} \cap (a,b) &= \{x_1\}, \\
\{x_1, x_2\} \cap (c,d) &= \emptyset, \\
\{x_1, x_2\} \cap (a,d) &= \{x_1, x_2\},
\end{aligned}$$

we say $\mathcal{C}$ shatters $\{x_1, x_2\}$. However, if we add one more point, say x_3, such that $a < x_1 < b < x_2 < c < x_3 < d$, $\mathcal{C}$ cannot shatter $\{x_1, x_2, x_3\}$. Then, the VC index of $\mathcal{C}$ is 2.

In the proof of Lemma 3.3.2, we aim to prove

$$\sup_{D \in \mathcal{D}} \Big| P_n\Big(D(\boldsymbol{\theta}, \boldsymbol{\mu})\Big) - P_0\Big(D(\boldsymbol{\theta}, \boldsymbol{\mu})\Big) \Big| \xrightarrow{P_0} 0,$$

where $D(\boldsymbol{\theta}, \boldsymbol{\mu})$ denotes the half-space of $\mathbb{R}^k$ characterised by $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$, $\mathcal{D}$ denotes the collection of the subsets $D(\boldsymbol{\theta}, \boldsymbol{\mu})$ characterised by $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$. Stengle and Yukich (1989) proposed a useful theorem to show that $\mathcal{D}$ we are working on actually is a VC class, so that Theorem 2.5 is applied.

Theorem 2.6 *Let $P(x_1, \cdots, x_m, y_1, \cdots, y_n, t_1, \cdots, t_p, u_1, \cdots, u_q) = P(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{t}, \boldsymbol{u})$ be a fixed real polynomial and T and U be fixed semialgebraic subsets of $\mathbb{R}^p$ and $\mathbb{R}^q$. Then the family of subsets of $\mathbb{R}^m$ of the form*

$$W_{\boldsymbol{a}} = \left\{ \boldsymbol{x} \middle| \sup_{\boldsymbol{t} \in T} \inf_{\boldsymbol{u} \in U} P(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{t}, \boldsymbol{u}) > 0 \right\}, \quad \boldsymbol{a} \in \mathbb{R}^n,$$

is a VC class.

Note that $\boldsymbol{t} \in T$ and $\boldsymbol{u} \in U$ make Theorem 2.6 more flexible in different situation. However, in Lemma 3.3.2, we fix $\boldsymbol{t} = \boldsymbol{t}_0$ and $\boldsymbol{u} = \boldsymbol{u}_0$, and thus we do not even need to consider what is called "semialgebraic sets". Therefore, we simplify Theorem 2.6 and give a corollary of it as follows:

Corollary 2.8.1 *Let $P(x_1, \cdots, x_m, y_1, \cdots, y_n) = P(\boldsymbol{x}, \boldsymbol{y})$ be a fixed real polynomial. Then the family of subsets of $\mathbb{R}^m$ of the form*

$$W_{\boldsymbol{a}} = \left\{ \boldsymbol{x} \middle| P(\boldsymbol{x}, \boldsymbol{a}) > 0 \right\}, \quad \boldsymbol{a} \in \mathbb{R}^n,$$

is a VC class.

Next we define VC-class of functions as follows.

Definition 2.8.9 (VC Class of Functions) *For a collection $\mathcal{F}$ of functions, if the collection of all subgraphs $\{(x, t) : f(x) < t\}$, $f \in \mathcal{F}$, forms a VC class of sets in $\mathcal{X} \times \mathbb{R}$, then $\mathcal{F}$ is a VC class of functions.*

Example 2.8.2 (VC Class of Functions) *Let $\theta \in \mathbb{R}$, and now consider an indicator function*

$$\mathbb{1}_\theta(x) = \begin{cases} 1, & x > \theta; \\ 0, & \textit{otherwise.} \end{cases}$$

Let $\mathcal{F} = \{\mathbb{1}_\theta(x)\}$. Then, it is easy to see the corresponding VC index is 1.

As van der Vaart (1998) noted, a collection of sets $\mathcal{C}$ is a VC class of sets if and only if the collection of corresponding indicator functions 1_C forms a VC class of functions. for $C \in \mathcal{C}$. Now, we introduce the lemma that makes (2.8.3) hold.

Lemma 2.8.5 (Upper Bound for Covering Number) *There exists a universal constant K such that for any VC class $\mathcal{F}$ of functions, any $r \geqslant 1$ and $0 < \epsilon < 1$,*

$$\sup_Q N(\epsilon \|F\|_{Q,1}, \mathcal{F}, L_1(Q)) \leqslant KV(\mathcal{F})(16e)^{V(\mathcal{F})} \left(\frac{1}{\epsilon}\right)^{r(V(\mathcal{F})-1)}.$$

It should be noted that the existence of an envelope is a sufficient condition, but there do exist other variance of this type of result in van der Vaart and Wellner (1996) which do not involve the existence of an envelope. The use of an envelope is convenient here because the functions are indicator functions and so an envelope exists.

2.8.2 *Stieltjes Integral*

Stieltjes integral is widely used in statistics. It allows us to integrate with respect to a function, and the function even does not need to be smooth; see Jeffreys et al. (1999, Chapter 1). Consequently, a unified notation can be used for discrete and continuous distributions.

Let the interval be partitioned by $x_1, x_2, \cdots, x_n$, where $a \leqslant x_1 < x_2 < \cdots < x_n = b$, and let real-valued bounded function $f(x)$ be defined on $[a, b]$. For the real-valued function $g(x)$, suppose

$$\sup_{x_1, x_2, \cdots, x_n} \sum_{i=1}^{n-1} |g(x_{i+1}) - g(x_i)|$$

is bounded, where the supremum is taken over all partitions of $[a, b]$.

For $i = 1, 2, \cdots, n-1$, choose an arbitrary $\psi_i \in [x_i, x_{i+1}]$. Let Δ denote the largest width of the interval. Write

$$S_n = \sum_{i=1}^{n-1} f(\psi_i)\{g(x_{i+1}) - g(x_i)\}.$$

If S_n tends to a unique limit, as $\Delta \to 0$, this limit is defined as a Stieltjes integral and denoted by

$$\int_a^b f(x)dg(x).$$

The function $g(x)$ can be a step function; see Dresher (1961), e.g.

$$g(x) = I_\alpha(x) = \begin{cases} 1 & x \geqslant \alpha, \\ 0 & x < \alpha. \end{cases}$$

and if $f(x)$ is continuous at α, then

$$\int_a^b f(x)dI_\alpha(x) = f(\alpha).$$

If $x_j < \alpha \leqslant x_{j+1}$, then for $i = 1, 2, \cdots, n-1$,

$$g(x_{i+1}) - g(x_i) = 1, \quad \text{for } i = j,$$
$$g(x_{i+1}) - g(x_i) = 0, \quad \text{for } i \neq j.$$

Hence,

$$S_n = f(\psi_j),$$

where $x_j \leqslant \psi_j \leqslant x_{j+1}$. By the continuity of $f(x)$ at $x = \alpha$, we have

$$\lim_{\substack{n\to\infty \\ \Delta\to 0}} f(\psi_j) = f(\alpha).$$

Therefore,

$$\lim_{\substack{n\to\infty \\ \Delta\to 0}} S_n = f(\alpha),$$

independent of the choice of ψ_i, $i = 1, 2, \cdots, n-1$. Thus,

$$\int_a^b f(x)dI_\alpha(x) = f(\alpha).$$

Now, consider a step function

$$g(x) = g_1 I_{\alpha_1}(x) + g_2 I_{\alpha_2}(x) + \cdots + g_n I_{\alpha_n}(x),$$

where $g_i \geqslant 0$, $a \leqslant \alpha_i \leqslant b$, $i = 1, 2, \cdots, n$, and $\sum_{i=1}^n g_i = 1$. If $f(x)$ is continuous at $\alpha_1, \alpha_2, \cdots, \alpha_n$, then

$$\int_a^b f(x)dg(x) = g_1 f(\alpha_1) + g_2 f(\alpha_2) + \cdots + g_n f(\alpha_n).$$

2.9 MATRIX ALGEBRA RESULTS

2.9.1 *Eigenvalue Property for Symmetric Matrices*

For real symmetric $k \times k$ matrices $\boldsymbol{S}$, it is possible to decompose them using eigenvalues and eigenvectors via the spectral theorem; see Strang (1993, Chapter 6.4).

Recall that a symmetric matrix $\boldsymbol{S}$ has k real eigenvalues λ_i, $i = 1, 2, \cdots, k$, and k orthonormal eigenvectors $\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_k$.

Theorem 2.7 (Spectral Theorem) *Every real symmetric $k \times k$ matrix $\boldsymbol{S}$ has the factorization*

$$\boldsymbol{S} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{\top}, \tag{2.9.1}$$

where $\lambda_1, \cdots, \lambda_k$ are k real eigenvalues, $\boldsymbol{q}_1, \cdots, \boldsymbol{q}_k$ are orthonormal eigenvectors, $\boldsymbol{\Lambda} = diag\{\lambda_1, \cdots, \lambda_k\}$, $\boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_k]$ and $\boldsymbol{Q}^{\top} = \boldsymbol{Q}^{-1}$.

Remark: Since $\boldsymbol{Q}$ is orthogonal, we can rewrite (2.9.1) as

$$\boldsymbol{\Lambda} = \boldsymbol{Q}^{-1}\boldsymbol{S}\boldsymbol{Q} = \boldsymbol{Q}^{\top}\boldsymbol{S}\boldsymbol{Q}.$$

where $\lambda_i = \boldsymbol{q}_i^{\top}\boldsymbol{S}\boldsymbol{q}_i$, for $i = 1, 2, \cdots, k$.

2.9.2 *Square Root of a Positive Definite Matrix*

Assume that the matrix $\boldsymbol{\Sigma} \in \mathcal{M}_{k\times k}(\mathbb{R})$ is symmetric and positive definite. Then, for example, by Mardia et al. (1979, Theorem A.6.4), we can write $\boldsymbol{\Sigma}$ as a spectral decomposition

$$\boldsymbol{\Sigma} = \boldsymbol{S}\boldsymbol{\Lambda}\boldsymbol{S}^{\top},$$

where $\boldsymbol{S}$ is orthogonal and $\boldsymbol{\Lambda} = diag\{\lambda_1, \cdots, \lambda_k\}$ is a diagonal matrix of the eigenvalues of $\boldsymbol{\Sigma}$.

It follows that

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{S}\boldsymbol{\Lambda}^{-1}\boldsymbol{S}^{\top},$$

where $\boldsymbol{\Lambda}^{-1} = diag\{1/\lambda_1, \cdots, 1/\lambda_k\}$ is the inverse matrix of $\boldsymbol{\Lambda}$.

Now, define the square root of $\boldsymbol{\Sigma}^{-1}$ as

$$\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{S}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{S}^{\top},$$

where $\boldsymbol{\Lambda}^{-1/2} = diag\{1/\sqrt{\lambda_1}, \cdots, 1/\sqrt{\lambda_k}\}$.

Remark: To verify that $\boldsymbol{\Sigma}^{-1/2}$ is a matrix square root of $\boldsymbol{\Sigma}^{-1}$, note that $\boldsymbol{\Sigma}^{-1/2}$:

$$\begin{aligned}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2} &= \boldsymbol{S}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{S}^{\top}\boldsymbol{S}\boldsymbol{\Lambda}^{-1/2}\boldsymbol{S}^{\top} \\ &= \boldsymbol{S}\boldsymbol{\Lambda}^{-1}\boldsymbol{S}^{\top} \\ &= \boldsymbol{\Sigma}^{-1}.\end{aligned}$$

2.9.3 *Matrix Calculus*

Assume $y : \mathbb{R} \to \mathbb{R}$ is a scalar function of $\boldsymbol{x} = (x_1, x_2 \cdots, x_n)^{\top} \in \mathbb{R}^n$, and define the first order derivatives of y as

$$\frac{\partial y}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix},$$

and

$$\frac{\partial y}{\partial \boldsymbol{x}^{\top}} = \left(\frac{\partial y}{\partial \boldsymbol{x}}\right)^{\top}.$$

Then its second order derivative is given by

$$\frac{\partial^2 y}{\partial \boldsymbol{x}\partial \boldsymbol{x}^{\top}} = \begin{bmatrix} \frac{\partial^2 y}{\partial x_1 \partial x_1} & \frac{\partial^2 y}{\partial x_1 \partial x_2} & \frac{\partial^2 y}{\partial x_1 \partial x_3} & \cdots & \frac{\partial^2 y}{\partial x_1 \partial x_n} \\ \frac{\partial^2 y}{\partial x_2 \partial x_1} & \frac{\partial^2 y}{\partial x_2 \partial x_2} & \frac{\partial^2 y}{\partial x_2 \partial x_3} & \cdots & \frac{\partial^2 y}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 y}{\partial x_n \partial x_1} & \frac{\partial^2 y}{\partial x_n \partial x_2} & \frac{\partial^2 y}{\partial x_n \partial x_3} & \cdots & \frac{\partial^2 y}{\partial x_n \partial x_n} \end{bmatrix}.$$

If $\boldsymbol{z} = (z_1, z_2 \cdots, z_n)^\top$ is a vector function of $\boldsymbol{x} = (x_1, x_2 \cdots, x_n)^\top$, then define the first order derivative as

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}^\top} = \begin{bmatrix} \frac{\partial z_1}{\partial x_1} & \frac{\partial z_1}{\partial x_2} & \frac{\partial z_1}{\partial x_3} & \cdots & \frac{\partial z_1}{\partial x_n} \\ \frac{\partial z_2}{\partial x_1} & \frac{\partial z_2}{\partial x_2} & \frac{\partial z_2}{\partial x_3} & \cdots & \frac{\partial z_2}{\partial x_n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{\partial z_n}{\partial x_1} & \frac{\partial z_n}{\partial x_2} & \frac{\partial z_n}{\partial x_3} & \cdots & \frac{\partial z_n}{\partial x_n} \end{bmatrix}.$$

2.9.4 *Matrix Norm*

An important tool throughout the thesis is the Frobenius matrix norm. We recall the definition of the vector norm and the matrix norm; see Conway (2013, Chapter 1.1 & 3.2). Let $\mathcal{V}$ be a vector space over $\mathbb{R}$. A norm on $\mathcal{V}$ is a map $\|\cdot\| : \mathcal{V} \to \mathbb{R}$ such that

1. $\|\boldsymbol{x}\| \geqslant 0$ for all $\boldsymbol{x} \in \mathcal{V}$ and $\|\boldsymbol{x}\| = 0$ if and only if $\boldsymbol{x} = \boldsymbol{0}$;
2. $\|\alpha \boldsymbol{x}\| = |\alpha| \cdot \|\boldsymbol{x}\|$ for all $\alpha \in \mathbb{R}$, $\boldsymbol{x} \in \mathcal{V}$;
3. $\|\boldsymbol{x} + \boldsymbol{y}\| \leqslant \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{V}$.

Then, $(\mathcal{V}, \|\cdot\|)$ is a normed vector space. Note that $(\mathbb{R}^k, \|\cdot\|)$ forms a vector space with the norm $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$, for any $\boldsymbol{x} \in \mathbb{R}^k$, called Euclidean norm.

Now, we define a matrix norm, and then the Frobenius matrix norm. Note that the vector norm is a special case of Frobenius norm; see Horn and Johnson (2012, Chapter 5.2 & 5.6) and Meyer (2000, Chapter 5).

Definition 2.9.1 (General Matrix Norm) *A matrix norm is a function $\|\cdot\|$ from the set of all real matrices (of all finite orders) into $\mathbb{R}$ that satisfies the following properties.*

1. $\|\boldsymbol{A}\| \geqslant 0$ *and* $\|\boldsymbol{A}\| = 0$ *if and only if* $\boldsymbol{A} = \boldsymbol{0}$.
2. $\|\alpha \boldsymbol{A}\| = |\alpha| \cdot \|\boldsymbol{A}\|$ *for all scalars* α.

3. $\|\boldsymbol{A}+\boldsymbol{B}\| \leqslant \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$ *for matrices of the same size.*

4. $\|\boldsymbol{AB}\| \leqslant \|\boldsymbol{A}\| \cdot \|\boldsymbol{B}\|$.

Definition 2.9.2 (Frobenius Matrix Norm) *Let $\mathcal{M}_{m\times n}(\mathbb{R})$ denote the set of $m \times n$ matrices over $\mathbb{R}$. The Frobenius norm of $\boldsymbol{A} \in \mathcal{M}_{m\times n}(\mathbb{R})$ is defined by*

$$\|\boldsymbol{A}\|_F = \sum_{i,j} |a_{ij}|^2 = \sum_i \|\boldsymbol{A}_{i*}\|^2 = \sum_j \|\boldsymbol{A}_{*j}\|^2 = tr(\boldsymbol{A}^\top \boldsymbol{A}),$$

where a_{ij} is the element of $\boldsymbol{A}$ at row i and column j, $i = 1, \cdots, m$ and $j = 1, \cdots, n$; $\boldsymbol{A}_{i*}$ denotes the row vector in row i and $\boldsymbol{A}_{*j}$ denotes the column vector in column j.

Note that the Frobenius matrix norm $\|\cdot\|_F$ is compatible with the Euclidean vector norm $\|\cdot\|$ by the inequality that for any $\boldsymbol{x} \in \mathbb{R}^n$:

$$\|\boldsymbol{Ax}\|^2 = \sum_i |\boldsymbol{A}_{i*}\boldsymbol{x}|^2 \leqslant \sum_i \|\boldsymbol{A}_{i*}\|^2\|\boldsymbol{x}\|^2 = \|\boldsymbol{A}\|_F\|\boldsymbol{x}\|^2. \tag{2.9.2}$$

Thus, the compatibility condition (2.9.2) implies that

$$\begin{aligned}
\|\boldsymbol{AB}\|_F^2 &= \sum_j \|[\boldsymbol{AB}]_{*j}\|^2 \\
&= \sum_j \|\boldsymbol{AB}_{*j}\|^2 \\
&\leqslant \sum_j \|\boldsymbol{A}\|_F^2\|\boldsymbol{B}_{*j}\|^2 \\
&= \|\boldsymbol{A}\|_F^2 \sum_j \|\boldsymbol{B}_{*j}\|^2 \\
&= \|\boldsymbol{A}\|_F^2\|\boldsymbol{B}\|_F^2.
\end{aligned}$$

Throughout the thesis, we will always use the Frobenius matrix norm for matrices and we shall use the notation $\|\cdot\|$ for convenience.

2.10 CONVEX OPTIMIZATION RESULTS

2.10.1 *Relative Interior*

In Theorem 3.2, an assumption is based on the relative interior. Boyd and Vandenberghe (2004, Chapter 2.1) gives the definitions of the affine set, affine hull and the relative interior. Here we use these definitions to define the convex set, the convex hull, and the relative interior of a convex set.

Definition 2.10.1 (Convex Sets) *A set $C \subset \mathbb{R}^k$ is convex if for any two distinct points $\boldsymbol{x}_1, \boldsymbol{x}_2 \in C$ and real number $\alpha \in [0,1]$, we have*

$$\alpha \boldsymbol{x}_1 + (1-\alpha)\boldsymbol{x}_2 \in C.$$

Definition 2.10.2 (Convex Hull) *The set of all convex combinations of points in some sets $C \subset \mathbb{R}^k$ is called the convex hull of C, and denoted $conv(C)$:*

$$conv(C) = \left\{ \theta_1 \boldsymbol{x}_1 + \cdots + \theta_r \boldsymbol{x}_r | \boldsymbol{x}_1 \cdots \boldsymbol{x}_r \in C, \sum_{i=1}^{r} \theta_i = 1, \theta_1, \cdots, \theta_r \geqslant 0 \right\}.$$

Definition 2.10.3 (Relative Interior) *Define the relative interior of the set $C \subset \mathbb{R}^k$, denoted $relint(C)$, as its interior relative to $conv(C)$:*

$$relint(C) = \{\boldsymbol{x} \in C | B(\boldsymbol{x}, \delta) \cap conv(C) \subset C \text{ for some } \delta > 0\},$$

where $B(\boldsymbol{x}, \delta) = \{\boldsymbol{y} | \|\boldsymbol{y} - \boldsymbol{x}\| \leqslant \delta\}$, the closed ball of radius δ and center $\boldsymbol{x}$ in the Euclidean norm $\|\cdot\|$.

2.10.2 *Separating Hyperplane Theorem*

In the proof of Theorem 3.1, the separating hyperplane theorem, as proposed by Boyd and Vandenberghe (2004, Chapter 2.5), is also used.

Theorem 2.8 (Separating Hyperplane Theorem) *Suppose C and D are nonempty disjoint convex sets, i.e., $C \cap D = \emptyset$. Then there exist $\boldsymbol{a} \neq \boldsymbol{0}$ and b such that $\boldsymbol{a}^\top \boldsymbol{x} \leqslant b$ for all $\boldsymbol{x} \in C$ and $\boldsymbol{a}^\top \boldsymbol{x} \geqslant b$ for all $\boldsymbol{x} \in D$.*

We give a corollary of Theorem 2.8:

Corollary 2.10.1 *Let C be a closed convex set and $x_0 \notin C$. Then there exist a non-zero vector $\boldsymbol{a}$ and real numbers $c_1 < c_2$ such that*

$$\boldsymbol{a}^\top \boldsymbol{x} > c_2 > c_1 > \boldsymbol{a}^\top \boldsymbol{x}_0,$$

for all $\boldsymbol{x} \in C$.

2.11 STEREOGRAPHIC PROJECTION

In Section 4.4, the log-empirical likelihood ratio against the mean direction parameter on $\mathcal{S}^2$ is plotted. In order to present it in a reasonable way, the mean direction parameters are mapped to a plane through stereographic projection. First, we denote $\{(x, y, z) \in \mathbb{R}^3 : x = 1\}$ as $\{x = 1\}$ for convenience. Many different maps will map $\mathcal{S}^2$ onto $\{x = 1\}$ but with the point $(-1, 0, 0)$ removed. Stereographic projection involves associating a point $\boldsymbol{X} = (x, y, z) \in \mathcal{S}^2$ with a point $\boldsymbol{Y} \in \{x = 1\}$ by extending the straight ray passing through $(-1, 0, 0)$ and $\boldsymbol{X} \in \mathcal{S}^2$ to $\boldsymbol{Y}$ on the plane. As is shown in Figure 2.11.7, we can use the fact that $\triangle PRX \sim \triangle PQY$ to get

$$Y = \left(1, \frac{2y}{x+1}, \frac{2z}{x+1}\right),$$

where $\sim$ indicates that the two triangles are similar, and $x \in (-1, 1]$.

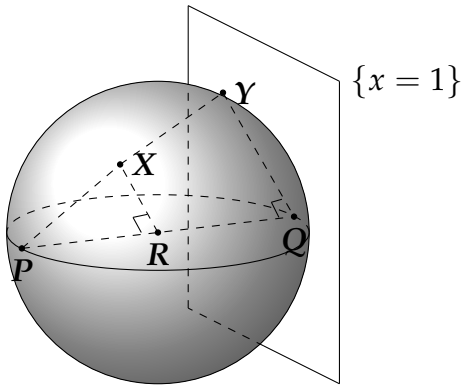


Figure 2.11.7: The stereographic projection from $\mathcal{S}^2$ to $\{x = 1\}$, where $P = (-1, 0, 0)$ is used as the point of projection. Note that $Q = (1, 0, 0)$, $R = (x, 0, 0)$.

3

EMPIRICAL LIKELIHOOD IN EUCLIDEAN SPACE

3.1 INTRODUCTION

Empirical likelihood (EL) methods in a Euclidean setting are studied in this chapter. Empirical likelihood, as a non-parametric method, is widely used when we know the sample but not the parametric family from which the sample was drawn. Wilks (1938) first proposed Wilks' theorem for the likelihood ratio statistic of parametric models. Owen (1988) first developed Wilks' theorem for empirical likelihood, stating that under the null hypothesis, Wilks' statistic for emipirical likelihood has an asymptotic chi-squared distribution, as in the parametric case. The proof of Wilks' theorem for empirical likelihood is presented by Owen (1990). An interesting extension of Wilks' theorem for empirical likelihood is to consider the behaviour of Wilks' statistic under local alternatives. As is shown by Sugiura (1973), local alternatives are sequences of alternative hypotheses that converge to the null hypothesis at a rate of n^{-r}. In this thesis, we choose $r = 1/2$. These local alternatives make the asymptotic distribution of Wilks' statistic for empirical likelihood a non-central chi-squared distribution with a non-centrality parameter associated with the definition of local alternatives. This chapter will mainly focus on deriving a general version of Wilks' theorem for empirical likelihood, where Wilks' statistic for empirical likelihood is evaluated under local alternatives. Then, Wilks' theorem proved by Owen (2001, Chapter 11)

is regarded as a special case of more general form of Wilks' theorem for empirical likelihood in this chapter.

According to Definition 2.1.3, the empirical likelihood ratio is defined as a function of the mean parameter. Knowing whether such a function is convex or concave is helpful in various applications of empirical likelihood. Owen (1990) mentioned the concavity of log-empirical likelihood for the mean parameter in Euclidean space but does not provide details. However, in this chapter, we give a full proof of the concavity of empirical likelihood with respect to the mean parameter.

The sections in this chapter are organized as follows. Section 3.2 gives the statement of Wilks' theorem for empirical likelihood, considering local alternatives. In Section 3.3, several useful lemmas are stated and used later in the proof of Wilks' theorem for empirical likelihood under local alternatives. Then the proof of Wilks' theorem for local alternatives is given in Section 3.4. In Section 3.5, it is proved that the log-empirical likelihood ratio for the mean parameter is a concave function. The results of previous sections are summarized in Section 3.6.

3.2 WILKS' THEOREM UNDER LOCAL ALTERNATIVES

Recall the empirical likelihood ratio $\mathcal{R}(\cdot)$ as in Definition 2.1.3, and the non-central chi-squared distribution as defined in Definition 2.2.1. Theorem 2.1 gives the asymptotic distribution of $-2\log\mathcal{R}(\boldsymbol{\mu}_0)$, where $\boldsymbol{\mu}_0$ is the population mean. Now, for sequences of local alternatives converging to the null hypothesis at the rate of $n^{-1/2}$, we develop Wilks' Theorem under local alternatives, i.e. finding the asymptotic distribution of $-2\log\mathcal{R}(\boldsymbol{\mu}_n)$, where $\{\boldsymbol{\mu}_n\}_{n\geqslant 1}$ denote local alternatives.

Theorem 3.1 (Wilks' Theorem under Local Alternatives) *Suppose that* $\boldsymbol{X}, \boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$ *are k-dimensional i.i.d. random vectors with common distribution* P_0*, and define*

$$\boldsymbol{\mu}_0 = \int_{\mathbb{R}^k} \boldsymbol{x} dP_0(\boldsymbol{x}).$$

Consider the sequence of local alternatives $\{\boldsymbol{\mu}_n\}_{n \geqslant 1}$*, such that* $\boldsymbol{\mu}_n = \boldsymbol{\mu}_0 + \boldsymbol{\eta}/\sqrt{n}$*, where* $\boldsymbol{\eta}$ *does not depend on n, and suppose the covariance matrix* $Cov(\boldsymbol{X}) = \boldsymbol{\Sigma}$ *has full rank k. Then* $-2\log \mathcal{R}(\boldsymbol{\mu}_n)$ *converges in distribution to a non-central chi-squared distribution, with non-centrality parameter* $\boldsymbol{\eta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\eta}$ *and degrees of freedom k.*

As Owen (2001) states in Theorem 2.1, the covariance matrix is not necessary to have full rank. The following corollary deals with this case.

Let $\boldsymbol{\Sigma}^+$ denote the Moore-Penrose inverse of $\boldsymbol{\Sigma}$; see for example Mardia et al. (1979) for the definition of Moore-Penrose inverse.

Corollary 3.2.1 *Under conditions in Theorem 3.1, if the covariance matrix has rank* $l < k$*,* $-2\log \mathcal{R}(\boldsymbol{\mu}_n)$ *converges in distribution to a non-central chi-squared distribution, with non-centrality parameter* $\boldsymbol{\eta}^\top \Sigma^+ \boldsymbol{\eta}$ *and degrees of freedom l.*

Owen (2001, Chapter 11) offers the full proof of Theorem 2.1. We will use this proof structure to prove Theorem 3.1. In the proof of Theorem 3.1, Lemma 2.8.2, Lemma 2.8.3 and Lemma 2.8.4 are adapted to the context with local alternatives. Although the structure of our proof is similar to the proof by Owen (2001), some new technical issues arise.

3.3 SOME USEFUL LEMMAS

Following the proof of Theorem 2.1 by Owen (2001, Chapter 11), we will expand Wilks' statistic for empirical likelihood evaluated at local alternatives via Taylor series, including a main term and some remainder terms. However, for the proof of Theorem 3.1, some modifications to Owen's proof are

needed, say calculating the bound of some terms, where $\boldsymbol{\mu}_0$ is replaced by $\boldsymbol{\mu}_n$.

Considering non-trivial cases, i.e. $0 < p_i < 1$, $i = 1, 2, \cdots, n$, the expression (2.1.10) implies that, as $n \to \infty$, a positive lower bound for the probability of $\{\boldsymbol{\theta}^\top(\boldsymbol{X} - \boldsymbol{\mu}) > 0\}$ under the empirical measure is required. Lemma 3.3.1, Lemma 3.3.2 and Lemma 3.3.3 deal with this problem. Lemma 3.3.4 to Lemma 3.3.8 show the bound calculation of crucial parameters in the proof of Theorem 3.1, so that the proof will be much shortened and more concise.

Lemma 3.3.1 *Let P_0 be a probability distribution on $\mathbb{R}^k$ with mean $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{V}_0$, where $\boldsymbol{V}_0$ is assumed to have full rank k. Consider the compact sets*

$$\Theta = \mathcal{S}^{k-1} = \{\boldsymbol{\theta} \in \mathbb{R}^k : \|\boldsymbol{\theta}\| = 1\} \subset \mathbb{R}^k, \tag{3.3.1}$$

and

$$\overline{B}_\delta(\boldsymbol{\mu}_0) = \{\boldsymbol{x} \in \mathbb{R}^k : \|\boldsymbol{x} - \boldsymbol{\mu}_0\| \leqslant \delta\} \subset \mathbb{R}^k, \tag{3.3.2}$$

where $\delta > 0$ is given. Let $\mathcal{D}$ denote a class of subsets of $\mathbb{R}^k$, i.e.

$$\begin{aligned} \mathcal{D} &= \{D(\boldsymbol{\theta}, \boldsymbol{\mu}) : \boldsymbol{\theta} \in \Theta, \boldsymbol{\mu} \in \overline{B}_\delta(\boldsymbol{\mu}_0)\}, \\ D(\boldsymbol{\theta}, \boldsymbol{\mu}) &= \{\boldsymbol{x} \in \mathbb{R}^k : \boldsymbol{\theta}^\top(\boldsymbol{x} - \boldsymbol{\mu}) > 0\}. \end{aligned} \tag{3.3.3}$$

Then,

$$\inf_{D \in \mathcal{D}} P_0(D) > 0. \tag{3.3.4}$$

Remark: sometimes we simply use the notation D to denote $D(\boldsymbol{\theta}, \boldsymbol{\mu})$, where $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\mu} \in \overline{B}_\delta(\boldsymbol{\mu}_0)$; otherwise, we use the notation $D(\cdot, \cdot)$ for $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$ specified.

PROOF. The proof uses the method of contradiction. If the inequality does not hold, there exists a sequence $\{\boldsymbol{\theta}_m, \boldsymbol{\mu}_m\}_{m \geqslant 1}$, such that

$$P_0\Big(D(\boldsymbol{\theta}_m, \boldsymbol{\mu}_m)\Big) < \frac{1}{m}. \tag{3.3.5}$$

The compactness of Θ and $\overline{B}_\delta(\boldsymbol{\mu}_0)$, and therefore the compactness of $\Theta \times \overline{B}_\delta(\boldsymbol{\mu}_0)$, implies there exists a convergent subsequence $\{\boldsymbol{\theta}_{m(j)}, \boldsymbol{\mu}_{m(j)}\}_{j\geqslant 1}$, with $m(j) \to \infty$ as $j \to \infty$, such that

$$\lim_{j\to\infty}\{\boldsymbol{\theta}_{m(j)}, \boldsymbol{\mu}_{m(j)}\} = \{\boldsymbol{\theta}^*, \boldsymbol{\mu}^*\}, \tag{3.3.6}$$

for some $\boldsymbol{\theta}^* \in \boldsymbol{\Theta}$ and $\boldsymbol{\mu}^* \in \overline{B}_\delta(\boldsymbol{\mu}_0)$. It follows that,

$$\mathbb{1}_{\{\boldsymbol{x}\in\mathbb{R}^k:(\boldsymbol{x}-\boldsymbol{\mu}_{m(j)})^\top\boldsymbol{\theta}_{m(j)}>0\}} = \mathbb{1}_{D(\boldsymbol{\theta}_{m(j)},\boldsymbol{\mu}_{m(j)})} \to \mathbb{1}_{\{\boldsymbol{x}\in\mathbb{R}^k:(\boldsymbol{x}-\boldsymbol{\mu}^*)^\top\boldsymbol{\theta}^*>0\}} = \mathbb{1}_{D(\boldsymbol{\theta}^*,\boldsymbol{\mu}^*)},$$

as $j \to \infty$.

The dominated convergence theorem (Theorem 2.4) implies that

$$\begin{aligned}
P_0\Big(D(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*)\Big) &= \int_{D(\boldsymbol{\theta}^*,\boldsymbol{\mu}^*)} \mathbb{1}_{D(\boldsymbol{\theta}^*,\boldsymbol{\mu}^*)} dP_0(\boldsymbol{x}) \\
&= \int_{D(\boldsymbol{\theta}^*,\boldsymbol{\mu}^*)} \lim_{j\to\infty} \mathbb{1}_{D(\boldsymbol{\theta}_{m(j)},\boldsymbol{\mu}_{m(j)})} dP_0(\boldsymbol{x}) \\
&= \lim_{j\to\infty} \int_{D(\boldsymbol{\theta}^*,\boldsymbol{\mu}^*)} \mathbb{1}_{D(\boldsymbol{\theta}_{m(j)},\boldsymbol{\mu}_{m(j)})} dP_0(\boldsymbol{x}) \\
&\leqslant \lim_{j\to\infty} \int_{\mathbb{R}^k} \mathbb{1}_{D(\boldsymbol{\theta}_{m(j)},\boldsymbol{\mu}_{m(j)})} dP_0(\boldsymbol{x}) \\
&= \lim_{j\to\infty} P_0\Big(D(\boldsymbol{\theta}_{m(j)}, \boldsymbol{\mu}_{m(j)})\Big) \\
&\leqslant \lim_{j\to\infty} \frac{1}{m(j)} \\
&= 0.
\end{aligned} \tag{3.3.7}$$

Let

$$\begin{aligned}
D_1(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*) &= \{\boldsymbol{x} \in \mathbb{R}^k : (\boldsymbol{x} - \boldsymbol{\mu}^*)^\top \boldsymbol{\theta}^* < 0\} \\
D_2(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*) &= \{\boldsymbol{x} \in \mathbb{R}^k : (\boldsymbol{x} - \boldsymbol{\mu}^*)^\top \boldsymbol{\theta}^* = 0\}.
\end{aligned}$$

Similarly as in (3.3.7), we have

$$P_0\Big(D_1(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*)\Big) = 0.$$

Since

$$D(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*) \cup D_1(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*) \cup D_2(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*) = \mathbb{R}^k,$$

and those three subsets are disjoint, then

$$P_0\Big(D_2(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*)\Big) = 1.$$

Thus,

$$Var\Big((\boldsymbol{X} - \boldsymbol{\mu}^*)^\top \boldsymbol{\theta}^*\Big) = 0.$$

However, this contradicts the assumption that the covariance matrix $\boldsymbol{V}_0$ has full rank k. Therefore, the inequality (3.3.4) is proved. □

Lemma 3.3.1 is stated under the distribution P_0. Following the proof of Theorem 2.1 by Owen (2001, Chapter 11), we need to adapt Lemma 3.3.1 under the empirical distribution P_n, so that there exists a set of $\{p_i\}_{i=1}^n$ such that $\sum_i^n p_i = 1$ and $0 < p_i < 1$ in probability, as $n \to \infty$. That is also equivalent to proving that the local alternative $\boldsymbol{\mu}_n$ is inside the convex hull of the random sample with probability 1, as $n \to \infty$, which is presented in Lemma 3.3.3.

Lemma 3.3.2 *Let P_0 be a probability measures on $\big(\mathbb{R}^k, \sigma(\mathcal{B}^k)\big)$. Assume that the covariance matrix $\boldsymbol{V}_0$ as defined in Lemma 3.3.1 has full rank. Let D denote the subset $D(\boldsymbol{\theta}, \boldsymbol{\mu}) \in \mathcal{D}$, for all $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\mu} \in \overline{B}_\delta(\boldsymbol{\mu}_0)$ as is defined in Lemma 3.3.1. Then there exists a constant $c > 0$ such that*

$$P_0\{\inf_{D \in \mathcal{D}} P_n(D) \geqslant c\} \to 1,$$

as $n \to \infty$.

PROOF. Recalling (2.8.1), letting

$$\delta_{\boldsymbol{x}}\left(D(\boldsymbol{\theta}, \boldsymbol{\mu})\right) = \begin{cases} 1, & \text{if } \boldsymbol{x} \in D(\boldsymbol{\theta}, \boldsymbol{\mu}), \\ 0, & \text{otherwise,} \end{cases}$$

define

$$\mathcal{F} = \{\delta_{\boldsymbol{x}}\big(D(\boldsymbol{\theta}, \boldsymbol{\mu})\big) : \boldsymbol{\theta} \in \Theta, \boldsymbol{\mu} \in \overline{B}_\delta(\boldsymbol{\mu}_0)\}.$$

Then, for $f \in \mathcal{F}$

$$P_n f = P_n\Big(D(\boldsymbol{\theta}, \boldsymbol{\mu})\Big), \text{ and}$$
$$P_0 f = P_0\Big(D(\boldsymbol{\theta}, \boldsymbol{\mu})\Big).$$

It follows that

$$\sup_{f \in \mathcal{F}} |P_n f - P_0 f| = \sup_{D \in \mathcal{D}} |P_n(D) - P_0(D)|.$$

For any $\boldsymbol{\theta} \in \Theta$ and $\boldsymbol{\mu} \in \overline{B}_\delta(\boldsymbol{\mu}_0)$, where Θ and $\overline{B}_\delta(\boldsymbol{\mu}_0)$ are defined by (3.3.1) and (3.3.2) respectively, $D(\boldsymbol{\theta}, \boldsymbol{\mu})$ defines a half-space in $\mathbb{R}^k$. Corollary 2.6 implies that $\mathcal{D}$ is a VC class of sets. Equivalently, $\mathcal{F}$ is a VC class of functions. Lemma 2.8.5 implies that for any $0 < \epsilon < 1$, and any $r \geqslant 1$, $\sup_Q N(\epsilon \|F\|_{Q,1}, \mathcal{F}, L_1(Q))$ has a finite upper bound. Now, in order to apply Theorem 2.5, we show $\mathcal{F}$ is suitably measurable. For any $D(\boldsymbol{\theta}, \boldsymbol{\mu})$, there exists a sequence of $\{(\boldsymbol{\theta}_m, \boldsymbol{\mu}_m)\}_{m \geqslant 1}$, such that $\{\boldsymbol{\theta}_m\}_{m \geqslant 1}$ and $\{\boldsymbol{\mu}_m\}_{m \geqslant 1}$ have all components rational and

$$\lim_{m \to \infty} \boldsymbol{\theta}_m = \boldsymbol{\theta},$$
$$\lim_{m \to \infty} \boldsymbol{\mu}_m = \boldsymbol{\mu}.$$

Then, by the denseness property of rational numbers, it is easy to see that

$$\delta_{\boldsymbol{x}}\big(D(\boldsymbol{\theta}_m, \boldsymbol{\mu}_m)\big) \to \delta_{\boldsymbol{x}}\big(D(\boldsymbol{\theta}, \boldsymbol{\mu})\big),$$

pointwise in $\boldsymbol{x}$, as $n \to \infty$.

Hence, for any $f \in \mathcal{F}$, each f is the pointwise limit of the corresponding sequence of functions $\{g_m\}_{m \geqslant 1}$, and then we find a collection of these $\{g_m\}_{m \geqslant 1}$, as $\mathcal{G}$. Now, $\mathcal{F}$ is suitably measurable, and we have the uniform convergence

$$P_0\Big(\sup_{D \in \mathcal{D}} |P_n(D) - P_0(D)| > \epsilon\Big) \to 0,$$

for every $\epsilon > 0$ as $n \to \infty$. It follows that

$$\sup_{D \in \mathcal{D}} |P_n(D) - P_0(D)| \xrightarrow{P_0} 0, \tag{3.3.8}$$

as $n \to \infty$.

Therefore, given any $\epsilon > 0$ and $c > 0$, there exists $N = N(\epsilon, c, P_0)$, such that for all $n \geqslant N$,

$$P_0\Big(\sup_{D\in\mathcal{D}} |P_0(D) - P_n(D)| > c\Big) \leqslant \epsilon,$$

or, equivalently,

$$P_0\Big(\sup_{D\in\mathcal{D}} |P_0(D) - P_n(D)| \leqslant c\Big) > 1 - \epsilon. \tag{3.3.9}$$

From Lemma 3.3.1, we can find $c > 0$ such that

$$\inf_{D\in\mathcal{D}} P_0(D) \geqslant 2c > 0.$$

Writing

$$\Big\{\omega \in \Omega | \inf_{D\in\mathcal{D}} P_n(D) \geqslant c\Big\}$$

as

$$\Big\{ \inf_{D\in\mathcal{D}} P_n(D) \geqslant c\Big\}$$

for convenience, since D is a set of $\boldsymbol{X}(\omega)$ as defined in (3.3.3). Then we have

$$\begin{aligned}
\Big\{ \inf_{D\in\mathcal{D}} P_n(D) \geqslant c\Big\} =& \Big\{ \inf_{D\in\mathcal{D}} \Big(P_0(D) + P_n(D) - P_0(D)\Big) \geqslant c\Big\} \\
\supseteq& \Big\{ \inf_{D\in\mathcal{D}} P_0(D) - \sup_{D\in\mathcal{D}} \Big(P_0(D) - P_n(D)\Big) \geqslant c\Big\} \\
\supseteq& \Big\{ \inf_{D\in\mathcal{D}} P_0(D) - \sup_{D\in\mathcal{D}} |P_0(D) - P_n(D)| \geqslant c\Big\} \\
\supseteq& \Big\{2c - \sup_{D\in\mathcal{D}} |P_0(D) - P_n(D)| \geqslant c\Big\} \\
=& \Big\{ \sup_{D\in\mathcal{D}} |P_n(D) - P_0(D)| \leqslant c\Big\}.
\end{aligned}$$

Therefore, for all $n \geqslant N$, the inequality (3.3.9) implies that

$$\begin{aligned}
P_0\Big\{ \inf_{D\in\mathcal{D}} P_n(D) \geqslant c\Big\} \geqslant& P_0\Big\{ \sup_{D\in\mathcal{D}} |P_n(D) - P_0(D)| \leqslant \epsilon\Big\} \\
\geqslant& 1 - \epsilon.
\end{aligned}$$

Then,

$$P_0\Big\{ \inf_{D\in\mathcal{D}} P_n(D) \geqslant c\Big\} \to 1, \tag{3.3.10}$$

as $n \to \infty$. □

Now, given local alternatives $\{\boldsymbol{\mu}_n\}_{n\geqslant 1}$ as

$$\boldsymbol{\mu}_n = \boldsymbol{\mu}_0 + \boldsymbol{\eta}/\sqrt{n}, \tag{3.3.11}$$

the results (3.3.4), (3.3.8) and (3.3.10) imply that for any $\boldsymbol{\theta} \in \Theta$,

$$\inf_{\boldsymbol{\theta}\in\Theta} P_0\Big(D(\boldsymbol{\theta}, \boldsymbol{\mu}_n)\Big) > 0,$$
$$\sup_{\boldsymbol{\theta}\in\Theta} \Big|P_0\Big(D(\boldsymbol{\theta}, \boldsymbol{\mu}_n)\Big) - P_n\Big(D(\boldsymbol{\theta}, \boldsymbol{\mu}_n)\Big)\Big| \xrightarrow{P} 0,$$
$$P_0\Big\{\inf_{\boldsymbol{\theta}\in\Theta} P_n\Big(D(\boldsymbol{\theta}, \boldsymbol{\mu}_n)\Big) > 0\Big\} \to 1,$$

as $n \to \infty$.

Note that we are interested in the behaviour of $\{\boldsymbol{\mu}_n\}_{n\geqslant 1}$, when $n \to \infty$, so we can always find $N(\delta) \in \mathbb{N}$, such that when $n \geqslant N(\delta)$,

$$\boldsymbol{\mu}_n = \boldsymbol{\mu}_0 + \frac{\boldsymbol{\eta}}{\sqrt{n}} \in \overline{B}_\delta(\boldsymbol{\mu}_0).$$

Lemma 3.3.3 *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$ denote i.i.d. random vectors in $\mathbb{R}^k$ with distribution P_0, and consider local alternatives as defined in (3.3.11). Then, $\boldsymbol{\mu}_n$ lies in the interior of the convex hull of $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$, with probability converging to 1, as $n \to \infty$, i.e.*

$$\lim_{n\to\infty} P_0\Big[\boldsymbol{\mu}_n \in int\Big(Conv(\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n)\Big)\Big] = 1,$$

where $int(\cdot)$ denotes the interior of a set, and

$$Conv(\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n) = \left\{\lambda_1\boldsymbol{X}_1 + \cdots, \lambda_n\boldsymbol{X}_n : \lambda_1, \cdots, \lambda_n \in [0,1], \sum_{i=1}^n \lambda_i = 1\right\}.$$

PROOF. Assume $\boldsymbol{\mu}_n$ is not in the interior of the convex hull of $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$. Recall Θ as defined in (3.3.1). Corollary 2.10.1 implies that for some $\boldsymbol{\theta} \in \Theta$ there exists some real numbers $c_2 > c_1$ such that

$$\boldsymbol{\theta}^\top \boldsymbol{X}_j > c_2 > c_1 > \boldsymbol{\theta}^\top \boldsymbol{\mu}_n,$$

for $j = 1, \cdots, n$. That is,

$$\boldsymbol{\theta}^\top (\boldsymbol{X}_j - \boldsymbol{\mu}_n) > c_2 - c_1 > 0,$$

for $j = 1, \cdots, n$.

Then, for $-\boldsymbol{\theta} \in \Theta$, we have

$$-\boldsymbol{\theta}^\top (\boldsymbol{X}_j - \boldsymbol{\mu}_n) < 0,$$

for $j = 1, \cdots, n$. Thus, we construct $\mathcal{D}'$, a class of subsets $D(\boldsymbol{\theta}, \boldsymbol{\mu}_n)$, where $\boldsymbol{\theta}$ satisfies

$$\boldsymbol{\theta}^\top (\boldsymbol{X}_j - \boldsymbol{\mu}_n) < 0,$$

for $j = 1, \cdots, n$. Lemma 3.3.2 implies that for the constant $c > 0$,

$$P_0 \left(\inf_{D \in \mathcal{D}'} P_n \left(D(\boldsymbol{\theta}, \boldsymbol{\mu}_n) \right) \geqslant c \right) \to 0,$$

as $n \to \infty$. Therefore, $\boldsymbol{\mu}_n$ is in the interior of the convex hull of $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$ with probability converging to 1. □

Lemma 3.3.4 will be used to bound the remainder term in the expansion of Wilks' statistic.

Lemma 3.3.4 *If $|Y| \leqslant c$, where $0 < c < 1$, then*

$$\log(1 + Y) = Y - \frac{1}{2} Y^2 + \xi,$$

where

$$|\xi| \leqslant B |Y|^3,$$

and

$$B = \frac{1}{3(1 - c)}.$$

PROOF. Since $0 < c < 1$, then by Taylor expansion,

$$\begin{aligned}
|\xi| =& \Big| \log(1+Y) - Y + \frac{1}{2}Y^2 \Big| \\
=& \Big| \frac{1}{3}Y^3 - \frac{1}{4}Y^4 + \frac{1}{5}Y^5 - \cdots \Big| \\
\leqslant& \Big| \frac{1}{3}Y^3 \Big| + \Big| \frac{1}{4}Y^4 \Big| + \Big| \frac{1}{5}Y^5 \Big| + \cdots \\
\leqslant& \frac{1}{3}|Y^3|(1 + |Y| + |Y|^2 + \cdots) \\
=& \frac{1}{3(1-|Y|)}|Y^3| \\
\leqslant& \frac{1}{3(1-c)}|Y|^3,
\end{aligned}$$

from which Lemma 3.3.4 follows. □

Lemma 3.3.5 *Let $\boldsymbol{X}, \boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$ denote k-dimensional i.i.d. random vectors in $\mathbb{R}^k$ with mean $\boldsymbol{\mu}_0$, and suppose*

$$\boldsymbol{\Sigma} = Cov(\boldsymbol{X}) \tag{3.3.12}$$

has full rank. For local alternatives $\{\boldsymbol{\mu}_n\}_{n \geqslant 1}$ as defined in (3.3.11), write

$$\boldsymbol{S}_n = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \boldsymbol{\mu}_n)(\boldsymbol{X}_i - \boldsymbol{\mu}_n)^\top. \tag{3.3.13}$$

Then,

$$\|\boldsymbol{S}_n^{-1} - \boldsymbol{\Sigma}^{-1}\| = o_p(1).$$

PROOF. Since $\boldsymbol{\Sigma}$ is assumed to have full rank k, rewrite $\boldsymbol{S}_n^{-1}$ as

$$\begin{aligned}
\boldsymbol{S}_n^{-1} &= (\boldsymbol{\Sigma} + \boldsymbol{S}_n - \boldsymbol{\Sigma})^{-1} \\
&= \boldsymbol{\Sigma}^{-1/2}\Big(\boldsymbol{I} + \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{S}_n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1/2}\Big)^{-1}\boldsymbol{\Sigma}^{-1/2},
\end{aligned} \tag{3.3.14}$$

and set

$$\boldsymbol{A} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{S}_n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1/2},$$

so that $\boldsymbol{A} = \boldsymbol{A}^\top$. We first obtain a bound for $\|\boldsymbol{S}_n - \boldsymbol{\Sigma}\|$. Now,

$$\begin{aligned}
\boldsymbol{S}_n &= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \boldsymbol{\mu}_n)(\boldsymbol{X}_i - \boldsymbol{\mu}_n)^\top \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big(\boldsymbol{X}_i - \boldsymbol{\mu}_0 - \frac{\boldsymbol{\eta}}{\sqrt{n}}\Big)\Big(\boldsymbol{X}_i - \boldsymbol{\mu}_0 - \frac{\boldsymbol{\eta}}{\sqrt{n}}\Big)^\top \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big\{(\boldsymbol{X}_i - \boldsymbol{\mu}_0)(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^\top - (\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0)\frac{\boldsymbol{\eta}^\top}{\sqrt{n}} - \frac{\boldsymbol{\eta}}{\sqrt{n}}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top + \frac{\boldsymbol{\eta}\boldsymbol{\eta}^\top}{n}\Big\} \\
&= \boldsymbol{S} - (\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0)\frac{\boldsymbol{\eta}^\top}{\sqrt{n}} - \frac{\boldsymbol{\eta}}{\sqrt{n}}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0)^\top + \frac{\boldsymbol{\eta}\boldsymbol{\eta}^\top}{n},
\end{aligned}$$

where

$$\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X_i} - \boldsymbol{\mu_0})(\boldsymbol{X_i} - \boldsymbol{\mu_0})^\top.$$

The weak LLN implies that

$$\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \boldsymbol{\mu}_0)(\boldsymbol{X}_i - \boldsymbol{\mu}_0)^\top \xrightarrow{P} \boldsymbol{\Sigma}, \tag{3.3.15}$$

and Lemma 2.8.1 implies that

$$\|\overline{\boldsymbol{X}} - \boldsymbol{\mu_0}\| = O_p(n^{-1/2}).$$

Since $\boldsymbol{\eta}$ is a constant vector,

$$\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}} = O_p(n^{-1/2}).$$

It follows that

$$\|\boldsymbol{S}_n - \boldsymbol{S}\| = O_p(n^{-1}),$$

and consequently (3.3.15) implies that

$$\|\boldsymbol{S}_n - \boldsymbol{\Sigma}\| = o_p(1).$$

Hence,

$$\|\boldsymbol{A}\| = \|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{S}_n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1/2}\| = o_p(1).$$

Now, consider

$$(\boldsymbol{I} + \boldsymbol{A})^{-1} = \boldsymbol{I} - \boldsymbol{A} + \boldsymbol{A}^2 - \boldsymbol{A}^3 + \cdots$$

Then, we have

$$\begin{aligned}\|(\boldsymbol{I}+\boldsymbol{A})^{-1}-\boldsymbol{I}\| &\leqslant -\frac{\|\boldsymbol{A}\|}{1-\|\boldsymbol{A}\|}\\ &= o_p(1).\end{aligned}$$

Rewrite

$$\begin{aligned}\boldsymbol{S}_n^{-1}-\boldsymbol{\Sigma}^{-1} &= \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{I}+\boldsymbol{A})^{-1}\boldsymbol{\Sigma}^{-1/2}-\boldsymbol{\Sigma}^{-1/2}\boldsymbol{I}\boldsymbol{\Sigma}^{-1/2}\\ &= -\boldsymbol{\Sigma}^{-1/2}((\boldsymbol{I}+\boldsymbol{A})^{-1}\boldsymbol{A})\boldsymbol{\Sigma}^{-1/2}.\end{aligned}$$

Therefore,

$$\|\boldsymbol{S}_n^{-1}-\boldsymbol{\Sigma}^{-1}\| = o_p(1),$$

as required. □

Note that Lemma 3.3.5 also implies that

$$\|\boldsymbol{S}_n^{-1}\| = O_p(1).$$

Lemma 3.3.6 *Let local alternatives $\{\boldsymbol{\mu}_n\}_{n\geqslant 1}$ be defined as in (3.3.11), and let $\boldsymbol{\Sigma}$, $\boldsymbol{S}_n$ be defined as in Lemma 3.3.5. For $i = 1, 2, \cdots, n$, write*

$$Y_i = \boldsymbol{\lambda}^\top(\boldsymbol{X}_i-\boldsymbol{\mu}_n), \tag{3.3.16}$$

where $\boldsymbol{\lambda}$ is given in (2.1.10) and satisfies the condition

$$\sum_{i=1}^{n}\frac{\boldsymbol{X}_i-\boldsymbol{\mu}_n}{1+\boldsymbol{\lambda}^\top(\boldsymbol{X}_i-\boldsymbol{\mu}_n)} = \boldsymbol{0}_k. \tag{3.3.17}$$

Then,

$$\|\boldsymbol{\lambda}\| = O_p(n^{-1/2}). \tag{3.3.18}$$

PROOF. Following the proof structure by Owen (2001, Chapter 11), we define

$$\begin{aligned}Z_n^* &= \max_{1\leqslant i\leqslant n}\|\boldsymbol{X}_i-\boldsymbol{\mu}_n\|,\\ \boldsymbol{S}_1 &= \frac{1}{n}\sum_{i=1}^{n}\frac{(\boldsymbol{X_i}-\boldsymbol{\mu_n})(\boldsymbol{X_i}-\boldsymbol{\mu_n}^\top}{1+Y_i},\\ \boldsymbol{\lambda} &= \|\boldsymbol{\lambda}\|\boldsymbol{\theta},\end{aligned}$$

where $\boldsymbol{\theta} \in \Theta$ is a unit vector, and $\|\cdot\|$ is the Euclidean norm.

Substitute

$$\frac{1}{1+Y_i} = 1 - \frac{Y_i}{1+Y_i}$$

into (3.3.17) and then (3.3.17) gives

$$\begin{aligned}
0 &= \boldsymbol{\theta}^\top \sum_{i=1}^{n} \frac{\boldsymbol{X}_i - \boldsymbol{\mu}_n}{1 + \boldsymbol{\lambda}^\top (\boldsymbol{X}_i - \boldsymbol{\mu}_n)} \\
&= \boldsymbol{\theta}^\top \frac{1}{n} \sum_{i=1}^{n} \frac{\boldsymbol{X}_i - \boldsymbol{\mu}_n}{1+Y_i} \\
&= \boldsymbol{\theta}^\top \frac{1}{n} \sum_{i=1}^{n} \big(\boldsymbol{X}_i - \boldsymbol{\mu}_n\big)\big(1 - \frac{Y_i}{1+Y_i}\big) \\
&= \boldsymbol{\theta}^\top \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{X}_i - \boldsymbol{\mu}_n) - \boldsymbol{\theta}^\top \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{X}_i - \boldsymbol{\mu}_n) \frac{Y_i}{1+Y_i} \\
&= \boldsymbol{\theta}^\top (\overline{\boldsymbol{X}} - \boldsymbol{\mu}_n) - \boldsymbol{\theta}^\top \frac{1}{n} \sum_{i=1}^{n} \frac{(\boldsymbol{X}_i - \boldsymbol{\mu}_n)(\boldsymbol{X}_i - \boldsymbol{\mu}_n)^\top \boldsymbol{\lambda}}{1+Y_i}. \qquad (3.3.19)
\end{aligned}$$

Rewrite (3.3.19) as

$$\boldsymbol{\theta}^\top (\overline{\boldsymbol{X}} - \boldsymbol{\mu}_n) = \|\boldsymbol{\lambda}\| \boldsymbol{\theta}^\top \boldsymbol{S}_1 \boldsymbol{\theta}.$$

Lemma 3.3.2 and 3.3.3 imply that $\{\boldsymbol{\mu}_n\}_{n \geqslant 1}$ lie in the convex hull of $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$ with probability approaching 1, as $n \to \infty$. Consequently,

$$\min_{1 \leqslant i \leqslant n} (1 + Y_i) > 0,$$

with probability approaching 1, as $n \to \infty$. Then, for $\boldsymbol{S}_n$ as defined in (3.3.13), we have

$$\begin{aligned}
\|\boldsymbol{\lambda}\| \boldsymbol{\theta}^\top \boldsymbol{S}_n \boldsymbol{\theta} &= \|\boldsymbol{\lambda}\| \boldsymbol{\theta}^\top \frac{1}{n} \sum_{i=1}^{n} \frac{(\boldsymbol{X}_i - \boldsymbol{\mu}_n)(\boldsymbol{X}_i - \boldsymbol{\mu}_n)^\top}{1+Y_i} (1+Y_i) \boldsymbol{\theta} \\
&\leqslant \|\boldsymbol{\lambda}\| \boldsymbol{\theta}^\top \boldsymbol{S}_1 \boldsymbol{\theta} (1 + \max_{1 \leqslant i \leqslant n} Y_i) \\
&\leqslant \|\boldsymbol{\lambda}\| \boldsymbol{\theta}^\top \boldsymbol{S}_1 \boldsymbol{\theta} (1 + \|\boldsymbol{\lambda}\| Z_n^*) \\
&= \boldsymbol{\theta}^\top (\overline{\boldsymbol{X}} - \boldsymbol{\mu}_n)(1 + \|\boldsymbol{\lambda}\| Z_n^*). \qquad (3.3.20)
\end{aligned}$$

Rearranging (3.3.20) gives

$$\|\boldsymbol{\lambda}\| \Big(\boldsymbol{\theta}^\top \boldsymbol{S}_n \boldsymbol{\theta} - Z_n^* \boldsymbol{\theta}^\top (\overline{\boldsymbol{X}} - \boldsymbol{\mu}_n) \Big) \leqslant \boldsymbol{\theta}^\top (\overline{\boldsymbol{X}} - \boldsymbol{\mu}_n). \qquad (3.3.21)$$

Now, in order to determine $\|\boldsymbol{\lambda}\|$, we study $\boldsymbol{\theta}^\top \boldsymbol{S}_n \boldsymbol{\theta}$, Z_n^* and $\|\overline{\boldsymbol{X}} - \boldsymbol{\mu}_n\|$ first.

1. $\lambda_k + o_p(1) \leqslant \boldsymbol{\theta}^\top \boldsymbol{S_n} \boldsymbol{\theta} \leqslant \lambda_1 + o_p(1)$, where $\lambda_1 \geqslant \lambda_k > 0$ are respectively the largest and smallest eigenvalues of $\boldsymbol{\Sigma}$. This is because the result in Section 2.9.1 implies that

$$\lambda_k \leqslant \boldsymbol{\theta}^\top \boldsymbol{\Sigma} \boldsymbol{\theta} \leqslant \lambda_1.$$

By rewriting

$$\boldsymbol{\theta}^\top \boldsymbol{S}_n \boldsymbol{\theta} = \boldsymbol{\theta}^\top (\boldsymbol{S}_n - \boldsymbol{\Sigma}) \boldsymbol{\theta} + \boldsymbol{\theta}^\top \boldsymbol{\Sigma} \boldsymbol{\theta},$$

Lemma 3.3.5 implies that

$$\boldsymbol{\theta}^\top (\boldsymbol{S}_n - \boldsymbol{\Sigma}) \boldsymbol{\theta} = o_p(1),$$

so that

$$\lambda_k + o_p(1) \leqslant \boldsymbol{\theta}^\top \boldsymbol{S}_n \boldsymbol{\theta} \leqslant \lambda_1 + o_p(1). \tag{3.3.22}$$

2. $Z_n^* = O_p(n^{-1/2})$. This follows from the parts that

$$\begin{aligned} Z_n^* &= \max_{1 \leqslant i \leqslant n} \|\boldsymbol{X}_i - \boldsymbol{\mu}_n\| \\ &= \max_{1 \leqslant i \leqslant n} \|\boldsymbol{X}_i - \boldsymbol{\mu}_0 - \boldsymbol{\eta}/\sqrt{n}\| \\ &\leqslant \max_{1 \leqslant i \leqslant n} \|\boldsymbol{X}_i - \boldsymbol{\mu}_0\| + \|\boldsymbol{\eta}\|/\sqrt{n}, \end{aligned}$$

and that Lemma 2.8.2 implies that

$$\max_{1 \leqslant i \leqslant n} \|\boldsymbol{X}_i - \boldsymbol{\mu}_0\| = o_p(n^{1/2}).$$

Recall

$$\|\boldsymbol{\eta}\|/\sqrt{n} = O_p(n^{-1/2}).$$

Therefore,

$$Z_n^* = o_p(n^{1/2}). \tag{3.3.23}$$

3. The central limit theorem (CLT) and Lemma 2.8.1 imply that

$$\|\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0\| = O_p(n^{-1/2}).$$

Thus, we have

$$\begin{aligned}\|\overline{\boldsymbol{X}} - \boldsymbol{\mu}_n\| =& \|\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0 + \boldsymbol{\mu}_0 - \boldsymbol{\mu}_n\| \\ \leqslant& \|\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0\| + \|\boldsymbol{\mu}_n - \boldsymbol{\mu}_0\| \\ =& \|\overline{\boldsymbol{X}} - \boldsymbol{\mu}_0\| + \|\boldsymbol{\eta}/\sqrt{n}\| \\ =& O_p(n^{-1/2}).\end{aligned}$$

It follows that

$$|\boldsymbol{\theta}^\top(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_n)| = O_p(n^{-1/2}). \tag{3.3.24}$$

Now, returning to $\|\boldsymbol{\lambda}\|$, by substituting (3.3.22), (3.3.23) and (3.3.24) into (3.3.21), we have

$$\|\boldsymbol{\lambda}\|\Big(O_p(1) + o_p(1)\Big) = O_p(n^{-1/2}).$$

Consequently,

$$\|\boldsymbol{\lambda}\| = O_p(n^{-1/2}).$$

□

Lemma 3.3.7 *Let Y_i, $i = 1, \cdots, n$, be defined as in Lemma 3.3.6. Then,*

(*i*) $\max\limits_{1\leqslant i\leqslant n} |Y_i| = o_p(1)$, *and*

(*ii*) $\max\limits_{1\leqslant i\leqslant n} \dfrac{1}{|1 + Y_i|} = O_p(1)$.

PROOF. To prove (i), for $i = 1, \cdots, n$,

$$\begin{aligned}Y_i &= \boldsymbol{\lambda}^\top(\boldsymbol{X}_i - \boldsymbol{\mu}_n) \\ &= \|\boldsymbol{\lambda}\|\left(\boldsymbol{\theta}^\top(\boldsymbol{X}_i - \boldsymbol{\mu}_0) - \frac{\boldsymbol{\theta}^\top\boldsymbol{\eta}}{\sqrt{n}}\right),\end{aligned}$$

where $\boldsymbol{\theta}^\top(\boldsymbol{X}_i - \boldsymbol{\mu}_0), \cdots, \boldsymbol{\theta}^\top(\boldsymbol{X}_n - \boldsymbol{\mu}_0)$ are i.i.d random variables. Thus, Lemma 2.8.2 implies that

$$\max_{1\leqslant i\leqslant n} |\boldsymbol{\theta}^\top(\boldsymbol{X}_i - \boldsymbol{\mu}_0)| = o_p(n^{1/2}).$$

Since $\|\boldsymbol{\theta}\| = 1$, and $\boldsymbol{\eta}$ is a constant vector,

$$\left\| \frac{\boldsymbol{\theta}^\top \boldsymbol{\eta}}{\sqrt{n}} \right\| = O_p(n^{-1/2}).$$

Now, the equality (3.3.18) implies that

$$\max_{1\leqslant i\leqslant n} |Y_i| = O_p(n^{-1/2})o_p(n^{1/2}) + O_p(n^{-1/2})O_p(n^{-1/2}) = o_p(1).$$

To prove (ii), we know from Lemma 3.3.2 that for any $\epsilon > 0$,

$$1 + Y_i \geqslant \epsilon,\ i = 1, 2, \cdots, n.$$

That is,

$$\frac{1}{1+Y_i} \leqslant \frac{1}{\epsilon},\ i = 1, 2, \cdots, n.$$

Thus, we have

$$\max_{1\leqslant i\leqslant n} \frac{1}{|1+Y_i|} = \max_{1\leqslant i\leqslant n} \frac{1}{1+Y_i} \leqslant \frac{1}{\epsilon}.$$

Hence,

$$\max_{1\leqslant i\leqslant n} \frac{1}{|1+Y_i|} = O_p(1).$$

□

Lemma 3.3.8 *Let $\boldsymbol{\lambda}$ be as in (2.1.10). Let Y_i, $i = 1, 2, \cdots, n$, be defined as in Lemma 3.3.6, local alternatives $\{\boldsymbol{\mu}_n\}_{n\geqslant 1}$ be defined as in (3.3.11) and $\boldsymbol{\Sigma}$, $\boldsymbol{S}_n$ be defined as in Lemma 3.3.5. Then,*

$$\boldsymbol{\lambda} = \boldsymbol{S}_n^{-1}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_n) + \boldsymbol{\beta}, \tag{3.3.25}$$

where

$$\boldsymbol{\beta} = \frac{1}{n}\sum_{i=1}^{n} \frac{\boldsymbol{S}_n^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}_n)Y_i^2}{1+Y_i}, \tag{3.3.26}$$

and

$$\|\boldsymbol{\beta}\| = o_p(n^{-1/2}). \tag{3.3.27}$$

PROOF. As Lemma 3.3.7 states, $\max_{1\leqslant i\leqslant n}|Y_i| = o_p(1)$, and thus we can apply geometric series to the equality (3.3.17) to get

$$\begin{aligned}\mathbf{0} &= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{X}_i - \boldsymbol{\mu}_n)(1 - Y_i + \frac{Y_i^2}{1+Y_i}) \\ &= \overline{\boldsymbol{X}} - \boldsymbol{\mu}_n - \boldsymbol{S}_n\boldsymbol{\lambda} + \frac{1}{n}\sum_{i=1}^{n}\frac{(\boldsymbol{X_i} - \boldsymbol{\mu}_n)Y_i^2}{1+Y_i}.\end{aligned} \tag{3.3.28}$$

Hence,

$$\boldsymbol{\lambda} = \boldsymbol{S}_n^{-1}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_n) + \boldsymbol{\beta},$$

where

$$\boldsymbol{\beta} = \frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{S}_n^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}_n)Y_i^2}{1+Y_i}.$$

Lemma 3.3.5 and Lemma 3.3.7 give that

$$\|\boldsymbol{S}_n^{-1} - \boldsymbol{\Sigma}^{-1}\| = o_p(1),$$

and

$$\max_{1\leqslant i\leqslant n}\frac{1}{|1+Y_i|} = O_p(1).$$

Recalling the proof of Lemma 2.8.3, we similarly prove that

$$\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{X}_i - \boldsymbol{\mu}_n\|^3 \leqslant \frac{H}{n}\sum_{i=1}^{n}\|\boldsymbol{X}_i - \boldsymbol{\mu}_n\|^2,$$

where

$$\begin{aligned}H &= \max_{1\leqslant i\leqslant n}\|\boldsymbol{X}_i - \boldsymbol{\mu}_n\| \\ &= o_p(n^{1/2}).\end{aligned}$$

The law of large numbers shows that

$$\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{X}_i - \boldsymbol{\mu}_0\|^2 = O_p(1).$$

Then,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{X}_i-\boldsymbol{\mu}_n\|^2\\
=&\frac{1}{n}\sum_{i=1}^{n}\Big\{\|\boldsymbol{X}_i-\boldsymbol{\mu}_0\|^2-2\Big|\frac{\boldsymbol{\eta}^\top}{\sqrt{n}}(\boldsymbol{X}_i-\boldsymbol{\mu}_0)\Big|+\frac{\boldsymbol{\eta}^\top\boldsymbol{\eta}}{n}\Big\}\\
\leqslant&\frac{1}{n}\sum_{i=1}^{n}\Big\{\|\boldsymbol{X}_i-\boldsymbol{\mu}_0\|^2-2\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\cdot\max_{1\leqslant i\leqslant n}\|\boldsymbol{X}_i-\boldsymbol{\mu}_0\|+\frac{\boldsymbol{\eta}^\top\boldsymbol{\eta}}{n}\Big\}\\
=&O_p(1).
\end{aligned}$$

Hence, we have

$$\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{X}_i-\boldsymbol{\mu}_n\|^3=o_p(n^{1/2}).$$

Thus,

$$\begin{aligned}
\|\boldsymbol{\beta}\|&\leqslant\|\boldsymbol{S}_n^{-1}\|\cdot\|\boldsymbol{\lambda}\|^2\cdot\max_{1\leqslant i\leqslant n}\frac{1}{|1+Y_i|}\cdot\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{X}_i-\boldsymbol{\mu}_n\|^3\\
&=o_p(n^{-1/2}).
\end{aligned}$$

□

3.4 PROOF OF THEOREM 3.1

The proof will be in the following steps.

Step 1. Use Taylor series to derive expressions of several useful terms, such as $\boldsymbol{\lambda}$ (3.3.25) and $\boldsymbol{\beta}$ (3.3.26). Find the term bounds of $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$.

Step 2. Apply Taylor series to Wilks' statistic to prove the Wilks' theorem.

PROOF OF STEP 1 : For $i=1,2,\cdots,n$, the equality (2.1.10) gives an expression for p_i, in which the Lagrange multiplier $\boldsymbol{\lambda}$ satisfies

$$g(\boldsymbol{\lambda})=\frac{1}{n}\sum_{i=1}^{n}\frac{(\boldsymbol{X}_i-\boldsymbol{\mu}_n)}{1+\boldsymbol{\lambda}^\top(\boldsymbol{X}_i-\boldsymbol{\mu}_n)}=\boldsymbol{0}. \tag{3.4.1}$$

The purpose of Wilks' theorem for empirical likelihood is to see the asymptotic behaviour of the Wilks' statistic. Following the proof structure by

Owen (2001, Chapter 11), we expand (3.4.1) with respect to $\boldsymbol{\lambda}^\top(\boldsymbol{X_i} - \boldsymbol{\mu_n})$ to get an expression for $\boldsymbol{\lambda}$.

In this proof, we use notation as defined in (3.3.11), (3.3.12), (3.3.13) and (3.3.16).

Lemma 3.3.6 and Lemma 3.3.7 show that

$$\|\boldsymbol{\lambda}\| = O_p(n^{-1/2}),$$

and

$$\max_{1\leqslant i\leqslant n} |Y_i| = \max_{1\leqslant i\leqslant n} |\boldsymbol{\lambda}^\top(\boldsymbol{X}_i - \boldsymbol{\mu}_n)| = o_p(1).$$

Thus, we can apply geometric series to derive the expression of $\boldsymbol{\lambda}$, as is shown in Lemma 3.3.8.

We hope to see the asymptotic behaviour of the Wilks' statistic for empirical likelihood, which is mainly influenced by the leading term of the expanded Wilks' statistic. For non-leading terms, we hope their term bounds are small enough. Lemma 3.3.8 shows one non-leading term $\boldsymbol{\beta}$ with bound

$$\|\boldsymbol{\beta}\| = o_p(n^{-1/2}).$$

□

PROOF OF STEP 2 : Now, we investigate the asymptotic behaviour of Wilks' statistic for empirical likelihood. First, by Taylor Theorem,

$$\log(1 + Y_i) = Y_i - \frac{1}{2}Y_i^2 + \xi_i. \tag{3.4.2}$$

Lemma 3.3.7 states that

$$\max_{1\leqslant i\leqslant n} |Y_i| = o_p(1).$$

Thus, for all $\epsilon > 0$,

$$Pr(\max_{1\leqslant i\leqslant n} |Y_i| \geqslant \epsilon) \to 0,$$

as $n \to \infty$.

That is, for any $c > 0$,

$$Pr(\max_{1 \leqslant i \leqslant n} |Y_i| \leqslant c) \to 1,$$

as $n \to \infty$.

Lemma 3.3.4 implies that

$$P(|\xi_i| \leqslant B|Y_i|^3,\ 1 \leqslant i \leqslant n) \to 1,$$

as $n \to \infty$, and

$$B = \frac{1}{3(1-c)}.$$

In addition, Lemma 2.8.3 implies that

$$\begin{aligned} \sum_{i=1}^{n} |\xi_i| &\leqslant B\|\boldsymbol{\lambda}\|^3 \sum_{i=1}^{n} \|\boldsymbol{X}_i - \boldsymbol{\mu}_n\|^3 \\ &= O_p(n^{-1/2})^3 \cdot n \cdot o_p(n^{1/2}) \\ &= o_p(1). \end{aligned} \tag{3.4.3}$$

For convenience, we use the notation in (3.3.16) for Y_i, $i = 1, 2, \cdots, n$. Now, we apply (3.4.2) to Wilks' statistic to get

$$\begin{aligned} -2\log \mathcal{R}(\boldsymbol{\mu_n}) =& -2\sum_{i=1}^{n} \log(np_i) \\ =& -2\sum_{i}^{n} \log\left(n \cdot \frac{1}{n\Big(1 + \boldsymbol{\lambda}^\top(\boldsymbol{X}_i - \boldsymbol{\mu}_0)\Big)} \right) \\ =& 2\sum_{i=1}^{n} \log(1 + Y_i) \\ =& 2\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} Y_i^2 + 2\sum_{i=1}^{n} \xi_i \\ =& 2n\boldsymbol{\lambda}^\top(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_n) - n\boldsymbol{\lambda}^\top \boldsymbol{S}_n \boldsymbol{\lambda} + 2\sum_{i=1}^{n} \xi_i \\ =& n(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_n)^\top \boldsymbol{S}_n^{-1}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}_n) - n\boldsymbol{\beta}^\top \boldsymbol{S_n}\boldsymbol{\beta} + 2\sum_{i=1}^{n} \xi_i, \end{aligned} \tag{3.4.4}$$

where

$$\begin{aligned} |n\boldsymbol{\beta}^\top \boldsymbol{S}_n \boldsymbol{\beta}| &= n \cdot o_p(n^{-1/2}) O_p(1) o_p(n^{-1/2}) \\ &= o_p(1). \end{aligned}$$

Since

$$\begin{aligned}&\sqrt{n}(\overline{\boldsymbol{X}}-\boldsymbol{\mu}_n)\xrightarrow{D}N_k(-\boldsymbol{\eta},\boldsymbol{\Sigma})\\ \Leftrightarrow&\sqrt{n}\boldsymbol{\Sigma}^{-1/2}(\overline{\boldsymbol{X}}-\boldsymbol{\mu}_n)\xrightarrow{D}N_k(-\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta},\boldsymbol{I}_k),\end{aligned}$$

applying Lemma 3.3.5 to (3.4.4), we have

$$\begin{aligned}-2\log\mathcal{R}(\boldsymbol{\mu_n})=&n(\overline{\boldsymbol{X}}-\boldsymbol{\mu}_n)^\top\boldsymbol{S}_n^{-1}(\overline{\boldsymbol{X}}-\boldsymbol{\mu}_n)+R_1\\ =&n(\overline{\boldsymbol{X}}-\boldsymbol{\mu}_n)^\top\boldsymbol{\Sigma}^{-1}(\overline{\boldsymbol{X}}-\boldsymbol{\mu}_n)+\\ &n(\overline{\boldsymbol{X}}-\boldsymbol{\mu}_n)^\top(\boldsymbol{S}_n^{-1}-\boldsymbol{\Sigma}^{-1})(\overline{\boldsymbol{X}}-\boldsymbol{\mu}_n)+R_1\\ =&n(\overline{\boldsymbol{X}}-\boldsymbol{\mu}_n)^\top\boldsymbol{\Sigma}^{-1}(\overline{\boldsymbol{X}}-\boldsymbol{\mu}_n)+R_2\\ \xrightarrow{D}&\chi_k^2(\boldsymbol{\eta}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}).\end{aligned}$$

where

$$\begin{aligned}&R_1=-n\boldsymbol{\beta}^\top\boldsymbol{S}_n\boldsymbol{\beta}+2\sum_{i=1}^n\xi_i,\\ &R_2=n(\overline{\boldsymbol{X}}-\boldsymbol{\mu}_n)^\top(\boldsymbol{S}_n^{-1}-\boldsymbol{\Sigma}^{-1})(\overline{\boldsymbol{X}}-\boldsymbol{\mu}_n)+R_1,\\ &|R_1|=o_p(1),\\ &|R_2|=o_p(1).\end{aligned}$$

□

Note that Owen (2001) showed (3.4.3) but did not explain in detail as Lemma 3.3.4. Furthermore, Owen (1990) stated the result of Theorem 3.1 but did not prove it.

3.5 CONCAVITY OF $l_{EL}(\mu)$

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n \in \mathbb{R}^k$ be i.i.d. random vectors. We consider the log-empirical likelihood for a multivariate mean, which, at a candidate value $\boldsymbol{\mu} \in \mathbb{R}^k$, is defined as the solution to the following optimisation problem:

$$\begin{aligned} l_{EL}(\boldsymbol{\mu}) = \max_{p_1,\cdots,p_n} & \sum_{i=1}^{n} \log p_i \\ \text{subject to } & p_1 > 0, \cdots, p_n > 0; \\ & \sum_{i=1}^{n} p_i = 1; \\ \text{and } & \sum_{i=1}^{n} p_i \boldsymbol{X}_i = \boldsymbol{\mu}. \end{aligned} \tag{3.5.1}$$

In this section we prove that $l_{EL}(\boldsymbol{\mu})$ is a concave function of $\boldsymbol{\mu}$. As far as we are aware, Owen (1990) mentioned the concavity result but did not provide detailed justification.

Theorem 3.2 (Concavity of $l_{EL}(\boldsymbol{\mu})$) *Suppose that $\boldsymbol{X}, \boldsymbol{X}_1, \cdots, \boldsymbol{X}_n \in \mathbb{R}^k$ are i.i.d. random vectors from a continuous population with positive definite covariance matrix $\boldsymbol{\Sigma} = Cov(\boldsymbol{X})$. If the sample size $n \geqslant k+1$, then the log-empirical likelihood $l_{EL}(\boldsymbol{\mu})$ defined in (3.5.1) is concave. Specifically:*

1. *If $\boldsymbol{\mu}$ lies in the relative interior, as in Definition 2.10.3, of the convex hull of the sample $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$, then*

$$\frac{\partial^2 l_{EL}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}^\top}$$

   is non-negative definite;

2. *$l_{EL}(\boldsymbol{\mu}) = -\infty$ otherwise.*

Remark: the proof mainly focuses on the case that the dimension of the relative interior of the convex hull of $\boldsymbol{X}, \boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$ is k. Let V be the sub-

space spanned by $\boldsymbol{X}, \boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$. If $dim(V) = r < k$, there must exist a set of orthogonal bases of V, written as $\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_r$, such that

$$\boldsymbol{X}_i = \sum_{j=1}^{r} c_{ij}\boldsymbol{f}_j,\ i = 1, 2, \cdots, n;$$

$$\text{and} \quad \boldsymbol{\mu} = \sum_{j=1}^{r} \mu_j \boldsymbol{f}_j,$$

where $c_{ij}, i = 1, 2, \cdots, n; j = 1, 2, \cdots, r$, and $\mu_j, j = 1, 2, \cdots, r$, are new real number components of corresponding vectors. Thus, under this set of bases, we rewrite the covariance matrix and it will be positive definite.

PROOF. By Definition 2.1.3, the empirical likelihood ratio is a function of the mean parameter $\boldsymbol{\mu}$. Substitute (2.1.10) into $l_{EL}(\boldsymbol{\mu})$, and obtain

$$\begin{aligned} l_{EL}(\boldsymbol{\mu}) &= \sum_{i=1}^{n} \log\left(\frac{1}{1 + \boldsymbol{\lambda}^\top(\boldsymbol{X_i} - \boldsymbol{\mu})}\right) - n\log n \\ &= -\sum_{i=1}^{n} \log\Big(1 + \boldsymbol{\lambda}^\top(\boldsymbol{X_i} - \boldsymbol{\mu})\Big) - n\log n, \end{aligned}$$

where $\boldsymbol{\lambda}$ satisfies

$$\sum_{i=1}^{n} \frac{1}{1 + \boldsymbol{\lambda}^\top(\boldsymbol{X_i} - \boldsymbol{\mu})}(\boldsymbol{X}_i - \boldsymbol{\mu}) = \mathbf{0}_k.$$

Since $l_{EL}(\boldsymbol{\mu})$ is a smooth function of $\boldsymbol{\mu}$, where $\boldsymbol{\mu}$ lies in the relative interior of the convex hull of $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$, we differentiate $l_{EL}(\boldsymbol{\mu})$ with respect to $\boldsymbol{\mu}$ twice. The first order derivative is given by

$$\begin{aligned} \frac{\partial l_{EL}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} &= -\sum_{i=1}^{n} \frac{\partial \log\Big(1 + \boldsymbol{\lambda}^\top(\boldsymbol{X_i} - \boldsymbol{\mu})\Big)}{\partial \boldsymbol{\mu}} \\ &= -\sum_{i=1}^{n} \frac{1}{1 + \boldsymbol{\lambda}^\top(\boldsymbol{X_i} - \boldsymbol{\mu})}\left(-\boldsymbol{\lambda} + \frac{\partial \boldsymbol{\lambda}^\top}{\partial \boldsymbol{\mu}}(\boldsymbol{X_i} - \boldsymbol{\mu})\right) \\ &= \sum_{i=1}^{n} np_i\boldsymbol{\lambda} - \frac{\partial \boldsymbol{\lambda}^\top}{\partial \boldsymbol{\mu}} \sum_{i=1}^{n} np_i(\boldsymbol{X_i} - \boldsymbol{\mu}) \\ &= n\boldsymbol{\lambda}, \end{aligned}$$

because the second sum is the zero vector due to the constraint

$$\sum_{i=1}^{n} p_i \boldsymbol{X}_i = \boldsymbol{\mu}.$$

The second order derivative is given by

$$\left(\frac{\partial^2 l_{EL}(\boldsymbol{\mu})}{\partial \boldsymbol{\mu} \partial \boldsymbol{\mu}^\top}\right)_{k\times k} = n\frac{\partial \boldsymbol{\lambda}}{\partial \boldsymbol{\mu}^\top}.$$

Now, consider the constraint $\sum_{i=1}^n p_i = 1$ in (2.1.8), and differentiate it with respect to $\boldsymbol{\mu}$:

$$\begin{aligned}
\mathbf{0}_k =& \frac{\partial}{\partial \boldsymbol{\mu}}\left(\sum_{i=1}^n p_i - 1\right) \\
=& \frac{\partial}{\partial \boldsymbol{\mu}} \sum_{i=1}^n \frac{1}{n\Big(1+\boldsymbol{\lambda}^\top(\boldsymbol{X_i}-\boldsymbol{\mu})\Big)} \\
=& \frac{1}{n}\sum_{i=1}^n \frac{\dfrac{\partial \boldsymbol{\lambda}^\top}{\partial \boldsymbol{\mu}}(\boldsymbol{X_i}-\boldsymbol{\mu}) - \boldsymbol{\lambda}}{\Big(1+\boldsymbol{\lambda}^\top(\boldsymbol{X_i}-\boldsymbol{\mu})\Big)^2} \\
=& n\sum_{i=1}^n p_i^2 \left(\frac{\partial \boldsymbol{\lambda}^\top}{\partial \boldsymbol{\mu}}(\boldsymbol{X_i}-\boldsymbol{\mu}) - \boldsymbol{\lambda}\right).
\end{aligned} \tag{3.5.2}$$

Rearranging the equation (3.5.2) gives an expression for $\boldsymbol{\lambda}$ as follows,

$$\boldsymbol{\lambda} = \frac{\sum_{i=1}^n p_i^2 \dfrac{\partial \boldsymbol{\lambda}^\top}{\partial \boldsymbol{\mu}}(\boldsymbol{X_i}-\boldsymbol{\mu})}{\sum_{i=1}^n p_i^2}. \tag{3.5.3}$$

Considering the other constraint in (3.5.1) that

$$\sum_{i=1}^n p_i \boldsymbol{X}_i = \boldsymbol{\mu}.$$

Differentiating with respect to $\boldsymbol{\mu}$:

$$\begin{aligned}
\mathbf{0}_{k\times k} =& \frac{\partial}{\partial \boldsymbol{\mu}^\top}\sum_{i=1}^n p_i(\boldsymbol{X_i}-\boldsymbol{\mu}) \\
=& \frac{\partial}{\partial \boldsymbol{\mu}^\top}\sum_{i=1}^n \frac{\boldsymbol{X_i}-\boldsymbol{\mu}}{n\Big(1+\boldsymbol{\lambda}^\top(\boldsymbol{X_i}-\boldsymbol{\mu})\Big)} \\
=& \frac{1}{n}\sum_{i=1}^n \frac{-\Big(1+\boldsymbol{\lambda}^\top(\boldsymbol{X_i}-\boldsymbol{\mu})\Big)\boldsymbol{I}_k - (\boldsymbol{X_i}-\boldsymbol{\mu})\left((\boldsymbol{X_i}-\boldsymbol{\mu})^\top \dfrac{\partial \boldsymbol{\lambda}}{\partial \boldsymbol{\mu}^\top} - \boldsymbol{\lambda}^\top\right)}{\Big(1+\boldsymbol{\lambda}^\top(\boldsymbol{X_i}-\boldsymbol{\mu})\Big)^2},
\end{aligned} \tag{3.5.4}$$

where $\boldsymbol{I}_k$ denotes the $k \times k$ identity matrix. Now, we substitute (3.5.3) into (3.5.4) and then rearrange (3.5.4) to obtain

$$\begin{aligned}\mathbf{0}_{k\times k} =& \boldsymbol{I}_k + n\sum_{i=1}^{n} p_i^2 \left((\boldsymbol{X_i} - \boldsymbol{\mu})(\boldsymbol{X_i} - \boldsymbol{\mu})^\top - (\boldsymbol{X_i} - \boldsymbol{\mu}) \frac{\sum_{j=1}^n p_j^2 (\boldsymbol{X_j} - \boldsymbol{\mu})^\top}{\sum_{j=1}^n p_j^2} \right) \frac{\partial \boldsymbol{\lambda}}{\partial \boldsymbol{\mu}^\top} \\ =& \boldsymbol{I}_k + n\sum_{i=1}^{n} p_i^2 \Bigg((\boldsymbol{X_i} - \boldsymbol{\mu})(\boldsymbol{X_i} - \boldsymbol{\mu})^\top - 2(\boldsymbol{X_i} - \boldsymbol{\mu}) \frac{\sum_{j=1}^n p_j^2 (\boldsymbol{X_j} - \boldsymbol{\mu})^\top}{\sum_{j=1}^n p_j^2} \\ & + (\boldsymbol{X_i} - \boldsymbol{\mu}) \frac{\sum_{j=1}^n p_j^2 (\boldsymbol{X_j} - \boldsymbol{\mu})^\top}{\sum_{j=1}^n p_j^2} \Bigg) \frac{\partial \boldsymbol{\lambda}}{\partial \boldsymbol{\mu}^\top}.\end{aligned}$$

Since

$$\sum_{i=1}^{n} p_i^2 (\boldsymbol{X}_i - \boldsymbol{\mu}) - \sum_{i=1}^{n} p_i^2 \frac{\sum_{j=1}^n p_j^2 (\boldsymbol{X_j} - \boldsymbol{\mu})}{\sum_{j=1}^n p_j^2} = \mathbf{0}, \tag{3.5.5}$$

we have

$$\begin{aligned}\mathbf{0}_{k\times k} =& \boldsymbol{I}_k + n\sum_{i=1}^{n} p_i^2 \Bigg((\boldsymbol{X_i} - \boldsymbol{\mu})(\boldsymbol{X_i} - \boldsymbol{\mu})^\top - 2(\boldsymbol{X_i} - \boldsymbol{\mu}) \frac{\sum_{j=1}^n p_j^2 (\boldsymbol{X_j} - \boldsymbol{\mu})^\top}{\sum_{j=1}^n p_j^2} \\ & + \left(\frac{\sum_{j=1}^n p_j^2 (\boldsymbol{X_j} - \boldsymbol{\mu})}{\sum_{j=1}^n p_j^2} \right) \left(\frac{\sum_{j=1}^n p_j^2 (\boldsymbol{X_j} - \boldsymbol{\mu})}{\sum_{j=1}^n p_j^2} \right)^\top \Bigg) \frac{\partial \boldsymbol{\lambda}}{\partial \boldsymbol{\mu}^\top} \\ =& \boldsymbol{I}_k + n\sum_{i=1}^{n} p_i^2 \left((\boldsymbol{X_i} - \boldsymbol{\mu}) - \left(\frac{\sum_{j=1}^n p_j^2 (\boldsymbol{X_j} - \boldsymbol{\mu})}{\sum_{j=1}^n p_j^2} \right) \right) \\ & \cdot \left((\boldsymbol{X_i} - \boldsymbol{\mu}) - \left(\frac{\sum_{j=1}^n p_j^2 (\boldsymbol{X_j} - \boldsymbol{\mu})}{\sum_{j=1}^n p_j^2} \right) \right)^\top \frac{\partial \boldsymbol{\lambda}}{\partial \boldsymbol{\mu}^\top} \\ =& \boldsymbol{I}_k + n\boldsymbol{A} \frac{\partial \boldsymbol{\lambda}}{\partial \boldsymbol{\mu}^\top},\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{A} &= \sum_{i=1}^{n} p_i^2 \boldsymbol{\Gamma}_i \boldsymbol{\Gamma}_i^\top, \\ \boldsymbol{\Gamma}_i &= (\boldsymbol{X_i} - \boldsymbol{\mu}) - \frac{\sum_{j=1}^n p_j^2 (\boldsymbol{X_j} - \boldsymbol{\mu})}{\sum_{j=1}^n p_j^2}.\end{aligned}$$

The rank of each $\boldsymbol{\Gamma}_i$, $i = 1, 2, \cdots, n$, is no larger than 1. Then, for $i = 1, 2, \cdots, n$,

$$Rank(\boldsymbol{\Gamma}_i \boldsymbol{\Gamma}_i^\top) \leqslant 1.$$

Equality (3.5.5) implies that there exists $\mathbf{\Gamma}_l$, $l \in \{1, 2, \cdots, n\}$, linearly related with $\mathbf{\Gamma}_i$, for $i \neq l$. Hence,

$$Rank(\boldsymbol{A}) \leqslant n - 1.$$

Consider $Rank(\boldsymbol{A}) \leqslant k$. Under the assumption $n \geqslant k+1$, the k-dimensional squared matrix $\boldsymbol{A}$ has full rank k, and thus is positive definite with probability 1. It follows that $\partial\boldsymbol{\lambda}/\partial\boldsymbol{\mu}^\top$ is negative definite. That is, the second order derivative of $l_{EL}(\boldsymbol{\mu})$ is negative definite with probability 1 if $n \geqslant k+1$.

For $\boldsymbol{\mu}$ lying out of the relative interior of the convex hull of $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$, it is easy to know there does not exist $\{p_i\}_{i=1}^n$, such that $\sum_{i=1}^n p_i = 1$, and $0 < p_i < 1$. Namely, there are two cases:

1. At least one $0 < p_i \leqslant 1$, $i \in \{1, 2, \cdots, n\}$, and $p_j = 0$, for $j \neq i$.

2. There is no solution to the empirical likelihood maximisation problem with $p_i \geq 0$, $i = 1, \cdots, n$.

In case 2, by convention, the maximisation problem has solution $l_{EL}(\boldsymbol{\mu}) = -\infty$, and in the first case

$$\prod_{i=1}^{n} p_i = 0,$$

and then $l_{EL}(\boldsymbol{\mu}) = -\infty$.

Therefore, the theorem is now proved. □

For the analysis of variance, consider random samples from each of populations. Let Y_{ij} denote the random sample, $i = 1, \cdots, m$, $j = 1, \cdots, n_i$, and $\boldsymbol{\mu}_i$ denote the population mean of the ith population. Assume $N = \sum_i n_i$. Within sample i, let

$$\begin{aligned} \mathcal{R}(\boldsymbol{\mu}_i) &= \max\left\{\prod_{j=1}^{n_i} n_i p_{ij} : \sum_{j=1}^{n_i} p_{ij} = 1, \sum_{j=1}^{n_i} p_{ij}(\boldsymbol{Y}_{ij} - \boldsymbol{\mu}_i) = \mathbf{0}\right\}, \\ \mathcal{R}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_m) &= \prod_{i=1}^{m} \mathcal{R}(\boldsymbol{\mu}_i), \end{aligned} \tag{3.5.6}$$

and define $l_i(\boldsymbol{\mu}_i) = \max \sum_{j=1}^{n_i} \log p_{ij}$. The null hypothesis is $\boldsymbol{\mu}_1 = \cdots = \boldsymbol{\mu}_m$. For each random sample, we can prove that the mean is inside the convex

hull $\mathscr{X}_{n_i}^{(i)}$ of Y_{ij} for each i. Under the null hypothesis, $l_i(\boldsymbol{\mu}_i)$ is maximised in $\mathscr{X}_{n_i}^{(i)}$. Thus, $l(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_m) = \sum_{i=1}^{m} l_i(\boldsymbol{\mu}_i)$ is maximised in $\cap_{i=1}^{m} \mathscr{X}_{n_i}^{(i)}$, provided that $\cap_{i=1}^{m} \mathscr{X}_{n_i}^{(i)}$ is non-empty. Computationally, it is useful to know the object has a unique maximum, if it exists.

Remark: Alternatively, Owen (2001) used another way to approach this problem:

$$\begin{aligned}\mathcal{R}(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_m) = \max \Big\{ \prod_{i=1}^{m} \prod_{j=1}^{n_i} N p_{ij} | p_{ij} \geqslant 0, \sum_{i=1}^{m} \sum_{j=1}^{n_i} p_{ij} = 1, \\ \sum_{j=1}^{n_i} p_{ij}(\boldsymbol{Y_{ij}} - \boldsymbol{\mu}_i) = \mathbf{0}, i = 1, \cdots, k. \Big\}\end{aligned} \tag{3.5.7}$$

One difference between (3.5.6) and (3.5.7) is that (3.5.6) uses m probability vectors but (3.5.7) uses only one probability vector. However, this needs further study to identify similarities and differences.

3.6 SUMMARY

In this chapter, we first stated and proved a more general version of Wilks' theorem for empirical likelihood. Our proof involves modifying the proof of the standard version of Wilks' theorem given by Owen (2001). In Theorem 3.1, we consider the idea of local alternatives by Sugiura (1973). Local alternatives are sequences of alternative hypotheses that converge to the null hypothesis at a rate of $n^{-1/2}$, where n is the sample size. Then, Wilks' statistic for empirical likelihood evaluated at those local alternatives has an asymptotic non-central chi-squared distribution with a non-centrality parameter associated with how local alternatives are defined.

Next, we continue the work by Owen (1990) that log-empirical likelihood function of the mean parameter is concave, since Owen (1990) did not provide enough detail in the proof of the concavity.

4

EMPIRICAL LIKELIHOOD ON THE SPHERE

4.1 INTRODUCTION

The goal of this chapter is extending Wilks' theorem for empirical likelihood for a model in Euclidean space to an important and commonly used non-Euclidean space, the sphere. Spherical data will be modelled by the empirical likelihood method in this chapter. Spherical data arises naturally in many scientific fields, especially experimentation and observation in earth science and astrophysics, which involve measuring directions. As Fisher et al. (1993) stated, the spherical data gives the orientation of a straight line in 3D space. One question which arises in the analysis of spherical data is which metric to use on the sphere, for example extrinsic or intrinsic. In order to adapt Wilks' Theorem proposed by Owen (2001) to the sphere, it is helpful to find a map which transforms spherical data into a Euclidean space, such as a tangent space. As is done in Theorem 3.1, Wilks' theorem for empirical likelihood is considered to investigate the asymptotic distribution of Wilks' statistic under local alternatives. As an important application, Wilks' theorem is used to construct confidence regions on the sphere.

The outline of this chapter is as follows. As the initial work, how to define empirical likelihood on $\mathcal{S}^{k-1}$ is shown in Section 4.2. Meanwhile, in Section 4.3, Wilks' theorem for empirical likelihood under local alternatives is presented, including the theorem statement, useful lemmas and the proof of the theorem. Next, Section 4.4 describes the algorithm of calculating the

empirical likelihood against the mean direction parameter on the sphere, and presents an example. In Section 4.5, some operating characteristics of hypothesis testing are examined, including the evaluating the probability of type I error and the probability of type II error. As an important application of Wilks' theorem, confidence regions construction for empirical likelihood is presented in Section 4.6, with real data applied. The chapter summary is presented in Section 4.7.

4.2 EMPIRICAL LIKELIHOOD ON $\mathcal{S}^{k-1}$

In Euclidean space, we already know how to calculate empirical likelihood for the mean parameter based on a sample $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$. However, it is not easy to characterise the convex hull on the sphere and then evaluate the empirical likelihood. Nevertheless, the random sample on the sphere can be transformed into a sample on the Euclidean space by some means.

Recall that the tangent space $T_{\boldsymbol{\mu}}\mathcal{S}^{k-1}$ as defined in Section 2.4.1, where $\boldsymbol{\mu}$ is the extrinsic mean direction parameter, is isometric to Euclidean space. This provides a convenient method for us to evaluate empirical likelihood for the mean direction parameter $\boldsymbol{\mu} \in \mathcal{S}^{k-1}$.

4.2.1 *Empirical Likelihood Ratio for Extrinsic Mean*

Suppose $\boldsymbol{\mu}$ denotes the extrinsic mean direction parameter. There are two ways to define empirical likelihood on $\mathcal{S}^{k-1}$:

1. Map the random sample on $\mathcal{S}^{k-1}$ to a new sample on $T_{\boldsymbol{\mu}}\mathcal{S}^{k-1}$. Then apply the model in Section 2.1 to define empirical likelihood.

2. Apply the model in Section 2.1 to i.i.d. random vectors on $\mathcal{S}^{k-1}$. That is, consider the mean direction parameter, build the constraints and define empirical likelihood directly on $\mathcal{S}^{k-1}$. Then, decompose the random vectors into two components, i.e. the component parallel to

the mean direction parameter and the component perpendicular to the mean direction parameter.

Let $\boldsymbol{X}, \boldsymbol{X_1}, \boldsymbol{X_2}, \cdots, \boldsymbol{X_n}$ be *i.i.d.* random vectors on $\mathcal{S}^{k-1} \subset \mathbb{R}^k$, $\boldsymbol{\mu_0}$ be the extrinsic mean direction, and $\{\boldsymbol{\mu_n}\}_{n\geqslant 1}$ be local alternatives, which are defined as

$$\boldsymbol{\mu}_n = \boldsymbol{\mu}_0 \cos\left(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\right) + \frac{\boldsymbol{\eta}}{\|\boldsymbol{\eta}\|} \sin\left(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\right), \tag{4.2.1}$$

where $\boldsymbol{\eta} \in \mathbb{R}^k$ is a constant vector.

Remark: as is clarified in Section 2.5.3, local alternatives are defined through the idea of first order derivative. In Euclidean space, local alternatives are defined directly as

$$\boldsymbol{\mu}_n = \boldsymbol{\mu}_0 + \frac{\boldsymbol{\eta}}{\sqrt{n}}, \tag{4.2.2}$$

since the tangent space $T_{\boldsymbol{\mu}_0}\mathbb{R}^k$ is isometric to $\mathbb{R}^k$. However, variables are constrained on $\mathcal{S}^{k-1}$, instead of free variables in $\mathbb{R}^k$. From the perspective of derivatives, we prefer defining local alternatives through tangent vectors as (4.2.1). Although (4.2.2) could also be a way to find local alternatives, it does not satisfy the demand in defining local alternatives through the idea of derivatives. Thus, we do not consider (4.2.2) in $\mathcal{S}^{k-1}$.

Method 1 to Define Empirical Likelihood Ratio

Now, consider mapping the random sample on $\mathcal{S}^{k-1}$ to the tangent space $T_{\boldsymbol{\mu}}\mathcal{S}^{k-1}$. As is presented by Amaral et al. (2007) in Section 2.5.2, assume

$$\boldsymbol{\mu} = (\mu^1, \mu^2, \cdots, \mu^k)^\top, \tag{4.2.3}$$

in Euclidean coordinates and set

$$\begin{aligned} c &= \mu^k, \\ \boldsymbol{b} &= (\mu^1, \mu^2, \cdots, \mu^{k-1})^\top. \end{aligned} \tag{4.2.4}$$

Then, let

$$\boldsymbol{M} = [\boldsymbol{G}_1, \boldsymbol{G}_2]^\top, \tag{4.2.5}$$

where $\boldsymbol{G}_2 = \boldsymbol{b}$, and

$$\boldsymbol{G}_1 = \begin{cases} c|c|^{-1}\boldsymbol{I}_{k-1} - c(|c| + |c|^2)^{-1}\boldsymbol{b}\boldsymbol{b}^\top & \text{if } c \neq 0; \\ \boldsymbol{I}_{k-1} - \boldsymbol{b}\boldsymbol{b}^\top & \text{if } c = 0. \end{cases}$$

That is, there exists a matrix-valued function of $\boldsymbol{b}$ and c, and we write:

$$\boldsymbol{M} = \boldsymbol{f}(\boldsymbol{\mu}). \tag{4.2.6}$$

Not only does the matrix $\boldsymbol{M}$ map the random sample on $\mathcal{S}^{k-1}$ into a new sample on $T_{\boldsymbol{\mu}}\mathcal{S}^{k-1}$, but also decreases the dimension of i.i.d. random vectors from k to $k-1$. If the mean direction parameter is $\boldsymbol{\mu}_0$, the advantages of the coordinate matrix $\boldsymbol{M}$ are:

1. The images of the random sample on the sphere are i.i.d. as well;
2. The covariance matrix of the new i.i.d. random vectors can be assumed to have full rank since the coordinate matrix $\boldsymbol{M}$ decreases the dimension by 1.

We now define the images of i.i.d. random vectors $\boldsymbol{X}, \boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$ through the coordinate matrix $\boldsymbol{M}$. For particular choices of mean direction parameter, $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_n$, we find the corresponding matrices $\boldsymbol{M}_0$, $\boldsymbol{M}_n$, and then define

$$\begin{aligned} \boldsymbol{Y}_{\boldsymbol{i}} &= \boldsymbol{M}_0\boldsymbol{X}_{\boldsymbol{i}} \in \mathbb{R}^{k-1},\ i = 1, \cdots, n; \\ \boldsymbol{Z}_{\boldsymbol{i}} &= \boldsymbol{M}_n\boldsymbol{X}_{\boldsymbol{i}} \in \mathbb{R}^{k-1},\ i = 1, \cdots, n. \end{aligned} \tag{4.2.7}$$

Note that $\boldsymbol{Y}, \boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ are i.i.d. random vectors. Since $\mathcal{S}^{k-1} \subset \mathbb{R}^k$, we have tangent spaces $T_{\boldsymbol{\mu}_0}\mathcal{S}^{k-1} \subset \mathbb{R}^k$ and $T_{\boldsymbol{\mu}_n}\mathcal{S}^{k-1} \subset \mathbb{R}^k$. $\boldsymbol{Y}, \boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ and $\boldsymbol{Z}, \boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n$ are vectors on $T_{\boldsymbol{\mu}_0}\mathcal{S}^{k-1}$ and $T_{\boldsymbol{\mu}_n}\mathcal{S}^{k-1}$ but the coordinates are redefined through maps $\boldsymbol{M}_0$, $\boldsymbol{M}_n$.

Hence, recall the model shown in (2.1.3), the coordinate matrix $\boldsymbol{M} = \boldsymbol{f}(\boldsymbol{\mu})$ as in (4.2.6), and define the empirical likelihood ratio on $T_{\boldsymbol{\mu}}\mathcal{S}^{k-1}$:

$$\begin{aligned} \mathcal{R}(\boldsymbol{\mu}) &= \max_{p_1,\cdots,p_n} \Big\{ \prod_{i=1}^n np_i \Big| \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i\boldsymbol{M}(\boldsymbol{X}_{\boldsymbol{i}} - \boldsymbol{\mu}) = \boldsymbol{0}, p_i > 0, i = 1, \cdots, n \Big\} \\ &= \max_{p_1,\cdots,p_n} \Big\{ \prod_{i=1}^n np_i \Big| \sum_{i=1}^n p_i = 1, \sum_{i=1}^n p_i\boldsymbol{M}\boldsymbol{X}_{\boldsymbol{i}} = \boldsymbol{0}, p_i > 0, i = 1, \cdots, n \Big\}, \end{aligned} \tag{4.2.8}$$

where $\boldsymbol{M\mu} = \mathbf{0}$. The log-empirical likelihood ratio on $T_{\boldsymbol{\mu}}\mathcal{S}^{k-1}$ is defined by

$$\log \mathcal{R}(\boldsymbol{\mu}) = \max_{p_1,\cdots,p_n} \Big\{ \sum_{i=1}^{n} \log np_i \Big| \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i \boldsymbol{MX_i} = \mathbf{0}, p_i > 0, i = 1, \cdots, n \Big\}. \tag{4.2.9}$$

Applying the Lagrange Multiplier method in Section 2.1 to (4.2.9), we have the optimal solution

$$p_i(\boldsymbol{\mu}) = \frac{1}{n\Big(1 + \boldsymbol{\lambda}^\top \boldsymbol{MX_i}\Big)}, \ i = 1, 2, \cdots, n, \tag{4.2.10}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^{k-1}$ is a Lagrange Multiplier and satisfies the condition that

$$\sum_{i=1}^{n} \frac{\boldsymbol{MX_i}}{n(1 + \boldsymbol{\lambda}^\top \boldsymbol{MX_i})} = \mathbf{0}_{k-1}. \tag{4.2.11}$$

Now recall Theorem 3.1 and evaluate Wilks' statistic for empirical likelihood at $\boldsymbol{\mu_n}$:

$$\begin{aligned} -2 \log \mathcal{R}(\boldsymbol{\mu_n}) &= -2 \log \prod_{i=1}^{n} np_i(\boldsymbol{\mu_n}) \\ &= -2 \sum_{i=1}^{n} \log np_i(\boldsymbol{\mu_n}). \end{aligned}$$

Remark: Below is the explanation why the last component is chosen to construct the matrix as shown in (4.2.4). In order for Method 1 to achieve the appropriate invariances and equivariances with respect to the original choice of coordinate system, it is essential to apply a preliminary transformation at the beginning and the inverse of this transformation at the end. Write the sample as $\tilde{\boldsymbol{X}}_1, \cdots, \tilde{\boldsymbol{X}}_n$ in an arbitrary initial coordinate system. Then find an orthogonal transformation $\boldsymbol{\Gamma}$ such that the transformed sample $\boldsymbol{X}_1 = \boldsymbol{\Gamma}\tilde{\boldsymbol{X}}_1, \cdots, \boldsymbol{X}_n = \boldsymbol{\Gamma}\tilde{\boldsymbol{X}}_n$ has extrinsic sample Fréchet mean (equivalently, sample mean direction) given by $(0, \cdots, 0, 1)^\top$. The matrix $\boldsymbol{\Gamma}$ can be found explicitly in terms of the sample mean direction of the $\tilde{\boldsymbol{X}}_1, \cdots, \tilde{\boldsymbol{X}}_n$ using results from Amaral et al. (2007). Then, if we wish to express the confidence region in terms of the original coordinate system, we should apply the inverse transformation $\boldsymbol{\Gamma}^\top$ to the confidence region. The procedure leads to an equivariant confidence region with respect to the choice of original coordinate system.

Method 2 to Define Empirical Likelihood

This method was proposed by Fisher et al. (1993, Section 3). Recall the empirical likelihood model in Section 2.1. We use a similar idea to define

$$P(B) = \sum_{i=1}^{n} p_i \delta_{\boldsymbol{X}_i}(B),$$

where $\delta_{\boldsymbol{X}}(B)$ is the indicator function with respect to the set B, as is shown in Definition 2.1.1, and $B = A \cap \mathcal{S}^{k-1}$ for an arbitrary Borel set $A \in \sigma(\mathcal{B}^k)$, where $\mathcal{B}^k = \{(A_1, A_2, \cdots, A_k) : A_j \in \mathcal{B}\}$ and $\sigma(\mathcal{B}^k)$ is the σ-algebra generated on $\mathcal{B}^k$. Now define the empirical likelihood ratio on $\mathcal{S}^{k-1}$:

$$\begin{aligned} \mathcal{R}(\boldsymbol{\mu}) = \max_{p_1,\cdots,p_n} & \prod_{i=1}^{n} np_i \\ \text{subject to } & p_1 > 0, \cdots, p_n > 0; \\ & \sum_{i=1}^{n} p_i = 1; \\ & \sum_{i=1}^{n} p_i \boldsymbol{X}_i / \| \sum_{i=1}^{n} p_i \boldsymbol{X}_i \| = \boldsymbol{\mu}. \end{aligned} \tag{4.2.12}$$

Consider the projection $\boldsymbol{I}_k - \boldsymbol{\mu}\boldsymbol{\mu}^\top$ to $T_{\boldsymbol{\mu}}\mathcal{S}^{k-1}$. Let

$$\boldsymbol{W}_i = (\boldsymbol{I} - \boldsymbol{\mu}\boldsymbol{\mu}^\top)\boldsymbol{X}_i,$$

for $i = 1, 2, \cdots, n$. Note that $\boldsymbol{W}_1, \cdots, \boldsymbol{W}_n$ are i.i.d. k-dimensional random vectors.

Rewrite for $i = 1, 2, \cdots, n$,

$$\boldsymbol{X}_i = \boldsymbol{W}_i + \left(\boldsymbol{X}_i^\top \boldsymbol{\mu}\right) \boldsymbol{\mu}.$$

Then,

$$\sum_{i=1}^{n} p_i \boldsymbol{X}_i = \sum_{i=1}^{n} p_i \boldsymbol{W}_i + \sum_{i=1}^{n} p_i (\boldsymbol{X}_i^\top \boldsymbol{\mu}) \boldsymbol{\mu}.$$

Hence, rewrite (4.2.12):

$$\begin{aligned} \mathcal{R}(\boldsymbol{\mu}) = & \max_{p_1,\cdots,p_n} \prod_{i=1}^{n} np_i, \\ \text{subject to } & p_1 > 0, \cdots, p_n > 0; \\ & \sum_{i=1}^{n} p_i = 1; \\ & \sum_{i=1}^{n} p_i \boldsymbol{W}_i = \boldsymbol{0}. \end{aligned} \tag{4.2.13}$$

Remark: These two methods swap the order of

1. Mapping the random sample onto $T_{\boldsymbol{\mu}}\mathcal{S}^{k-1}$;
2. Applying empirical likelihood to the random sample.

However, they both lead to the same result in evaluating the empirical likelihood ratio against the mean direction parameter on the sphere $\mathcal{S}^{k-1}$, if the matrix $\boldsymbol{M}$ as defined in (4.2.6) is used for coordinate transformation, i.e.

$$\boldsymbol{M}(\boldsymbol{I} - \boldsymbol{\mu}\boldsymbol{\mu}^\top) = \boldsymbol{M}.$$

The only discrepancies between these two methods are:

1. Method 1 uses the coordinate matrix $\boldsymbol{M}$ for a chosen mean direction parameter $\boldsymbol{\mu}$ to decrease the dimension of images of the random sample by 1, so that in Wilks' theorem for empirical likelihood we can make the assumption that, for $\boldsymbol{\mu}_0$, both the covariance matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$ as in (4.2.7), i.e. $Cov(\boldsymbol{X})$ and $Cov(\boldsymbol{Y})$, have full rank k and $k-1$ respectively. However, in method 2, the matrix $\boldsymbol{I}_k - \boldsymbol{\mu}\boldsymbol{\mu}^\top$ is a $k \times k$ projection matrix, so that the covariance matrix of $\boldsymbol{W}$, $Cov(\boldsymbol{W})$, does not have full rank, despite the fact that the covariance matrix of $\boldsymbol{X}$ has full rank.
2. Method 1 is more straightforward, since we firstly transform the random sample into a sample in the Euclidean space, then apply the distribution model as in (2.1.3) and finally define the empirical likelihood

ratio against the mean direction parameter. The method 2 applies the distribution model as in (2.1.3) to the random sample on $\mathcal{S}^{k-1}$ first and then constructs the constraints in (4.2.12). Through decomposition in the last constraint of (4.2.12), we only consider the component that is perpendicular to $\boldsymbol{\mu}$.

4.3 WILKS' THEOREM WITH LOCAL ALTERNATIVES ON $\mathcal{S}^{k-1}$

Although the cases for $k = 2, 3$ are of the most practical interest, we shall consider a general integer $k \geqslant 2$. Now, with empirical likelihood defined on $\mathcal{S}^{k-1}$, we adapt Wilks' theorem for empirical likelihood to $\mathcal{S}^{k-1}$, considering local alternatives. It should be noted that the following theorem considers $\mathcal{S}^{k-1}$ embeded in $\mathbb{R}^k$. Thus, as (2.5.1) shows, the Fréchet mean is the population mean direction.

4.3.1 *Statement of Wilks' Theorem*

Theorem 4.1 (Wilks' Theorem on S^{k-1} under Local Alternatives) *Let $\boldsymbol{X}, \boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$ be i.i.d. random vectors on $\mathcal{S}^{k-1} \subset \mathbb{R}^k$, with common distribution P_0 and common expectation as*

$$\boldsymbol{\alpha} = \int_{\mathcal{S}^{k-1}} \boldsymbol{x} dP_0(\boldsymbol{x}). \tag{4.3.1}$$

Note that the mean direction of P_0

$$\boldsymbol{\mu}_0 = \frac{\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|}.$$

Consider again local alternatives $\{\boldsymbol{\mu}_n\}_{n\geqslant 1}$, such that

$$\boldsymbol{\mu}_n = \boldsymbol{\mu}_0 \cos\Big(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\Big) + \frac{\boldsymbol{\eta}}{\|\boldsymbol{\eta}\|} \sin\Big(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\Big),$$

where $\boldsymbol{\eta} \in \mathbb{R}^k$ *is a constant vector. Recall* $\boldsymbol{Y}, \boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ *and* $\boldsymbol{Z}, \boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_n$ *as defined in (4.2.7, and* $\boldsymbol{M}_0$ *as defined in (4.2.6) through* $\boldsymbol{\mu}_0$*. Assume that the square matrix of dimension* $k-1$

$$\boldsymbol{\Sigma} = Cov(\boldsymbol{Y})$$

has full rank $k-1$*. Then, Wilks' statistic* $-2\log\mathcal{R}(\boldsymbol{\mu_n})$ *as is defined in (4.2.9) converges in distribution to a non-central chi-squared distribution with degree of freedom* $k-1$ *and non-centrality parameter* $\|\alpha\|^2\boldsymbol{\eta}^\top\boldsymbol{M}_0^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{M}_0\boldsymbol{\eta}$*, as* $n \to \infty$*.*

Note that Wilks' theorem for empirical likelihood on $\mathcal{S}^{k-1}$ differs from Theorem 3.1. In the proof of Theorem 3.1, we defined notation associated with $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$. Similarly, notation associated with $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \cdots, \boldsymbol{Y}_n$ and $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots, \boldsymbol{Z}_n$ are defined later. However, the concern is whether the results in Section 3.3 still hold or not in this section. Recall the proof of Theorem 3.1 in Chapter 3. Some useful lemmas are presented beforehand to support the proof later. We still use the proof structure of Theorem 3.1 in this chapter, but some detail might be adapted for $\mathcal{S}^{k-1}$.

From the result in Section 3.2, the Wilks' statistic evaluated at local alternatives converges to a non-central chi-squared distribution with a non-centrality parameter, which is associated with the definition of local alternatives. As a result of Owen (2001), Wilks' theorem for empirical likelihood (Theorem 2.1) states that the Wilks' statistic, for i.i.d. Euclidean random vectors, evaluated at the population mean, asymptotically converges to a chi-squared distribution with the non-centrality parameter equal to 0, i.e. a central chi-squared distribution. That is, we can consider Theorem 2.1 as a special case of Theorem 3.1. Now, we similarly propose a corollary based on Theorem 4.1. The corollary will only focus on evaluating the Wilks' statistic for the extrinsic mean direction.

Corollary 4.3.1 (Central Wilks' Theorem on S^{k-1}**)** *Let* $\boldsymbol{\eta} = \boldsymbol{0}$ *in Theorem 4.1. Then, the Wilks' Statistic* $-2\log\mathcal{R}(\boldsymbol{\mu_0})$ *as defined in (4.2.9) converges in distribution to chi-squared distribution with degree of freedom* $k-1$ *as* $n \to \infty$*.*

4.3.2 *Rotation on $\mathcal{S}^{k-1}$*

As is defined in (4.2.6), the coordinate matrix $\boldsymbol{M}$ is introduced as a matrix-valued function of the mean direction parameter $\boldsymbol{\mu}$. We find $\boldsymbol{M}_0, \boldsymbol{M}_n$ through $\boldsymbol{\mu}_0, \boldsymbol{\mu}_n$. Amaral et al. (2007, Lemma 2) proposed a way to rotate $\boldsymbol{\mu}_n$ to $\boldsymbol{\mu}_0$. As local alternatives $\{\boldsymbol{\mu}_n\}_{n\geqslant 1}$ on the sphere is defined as

$$\boldsymbol{\mu}_n = \boldsymbol{\mu}_0 \cos\Big(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\Big) + \frac{\boldsymbol{\eta}}{\|\boldsymbol{\eta}\|}\sin\Big(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\Big),$$

where $\boldsymbol{\eta}$ is a constant vector and $\boldsymbol{\eta}^\top\boldsymbol{\mu}_0 = 0$, we define an anxillary vector

$$\begin{aligned}\boldsymbol{\mu}_A =& \boldsymbol{\mu}_n - \boldsymbol{\mu}_0\cos\left(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\right)\\ =& \frac{\boldsymbol{\eta}}{\|\boldsymbol{\eta}\|}\sin\left(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\right),\end{aligned}$$

and then

$$\begin{aligned}\boldsymbol{\mu}_B &= \frac{\boldsymbol{\mu}_A}{\|\boldsymbol{\mu}_A\|}\\ &= \frac{\boldsymbol{\eta}}{\|\boldsymbol{\eta}\|}.\end{aligned}$$

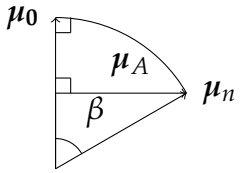


Figure 4.3.8: The anxillary vector $\boldsymbol{\mu}_A$.

We define a skew-symmetric matrix $\boldsymbol{A}$ first:

$$\begin{aligned}\boldsymbol{A} &= \boldsymbol{\mu}\boldsymbol{\mu}_B^\top - \boldsymbol{\mu}_B\boldsymbol{\mu}^\top,\\ &= \frac{\boldsymbol{\mu}_0\boldsymbol{\eta}^\top - \boldsymbol{\eta}\boldsymbol{\mu}_0^\top}{\|\boldsymbol{\eta}\|}.\end{aligned}$$

Thus, there exists a rotation matrix:

$$\begin{aligned} \boldsymbol{Q} &= \exp\left(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\boldsymbol{A}\right) \\ &= \boldsymbol{I}_k + \sum_{i=1}^{\infty}\left(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\boldsymbol{A}\right)^i / i! \\ &= \boldsymbol{I}_k + \sin\left(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\right)\boldsymbol{A} + \left\{\cos\left(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\right) - 1\right\}\left\{\boldsymbol{\mu}_0\boldsymbol{\mu}_0^\top + \frac{\boldsymbol{\eta}\boldsymbol{\eta}^\top}{\|\boldsymbol{\eta}\|^2}\right\}, \end{aligned} \tag{4.3.2}$$

such that

$$\boldsymbol{Q}\boldsymbol{\mu}_n = \boldsymbol{\mu}_0. \tag{4.3.3}$$

Pre-multiplying (4.3.3) by $\boldsymbol{M}_0$, we obtain

$$\boldsymbol{M}_0\boldsymbol{Q}\boldsymbol{\mu}_n = \boldsymbol{M}_0\boldsymbol{\mu}_0 = \boldsymbol{0}.$$

Thus,

$$\boldsymbol{M}_0\boldsymbol{Q} = \boldsymbol{M}_n. \tag{4.3.4}$$

To see this, we note that, since $\boldsymbol{M}_n$ of dimension $(k-1)\times k$ has $k-1$ mutually orthogonal row vectors and they are orthogonal to $\boldsymbol{\mu}_n$. That is, we can find a $k \times k$ orthogonal matrix

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{\mu}_n^\top \\ \boldsymbol{\mu}_{n,1}^\top \\ \vdots \\ \boldsymbol{\mu}_{n,k-1}^\top \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_n^\top \\ \boldsymbol{M}_n \end{bmatrix}.$$

Then, applying (4.3.3) to $\boldsymbol{H}$, we have

$$\begin{bmatrix} \boldsymbol{\mu}_n^\top\boldsymbol{Q}^\top \\ \boldsymbol{M}_n\boldsymbol{Q}^\top \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_0^\top \\ \boldsymbol{M}_n\boldsymbol{Q}^\top \end{bmatrix},$$

which is also an orthogonal matrix. Thus, we know $\boldsymbol{M}_0 = \boldsymbol{M}_n\boldsymbol{Q}^\top$. Now, we use (4.3.2) and (4.3.4) to rewrite $-2\log\mathcal{R}(\boldsymbol{\mu}_n)$ into a form with the leading term associated with $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ and $\boldsymbol{\eta}$. The detail is shown in the proof of Theorem 4.1 in Section 4.3.4.

4.3.3 *Some Useful Lemmas*

Recalling (4.2.10), we require a non-trivial solution $\{p_i\}_{i=1}^n$, such that for $i = 1, 2, \cdots, n$, $0 < p_i < 1$, and $\sum_{i=1}^n p_i = 1$.

Lemma 4.3.1 *Let $\boldsymbol{X}$ be a random vector on $\mathcal{S}^{k-1} \subset \mathbb{R}^k$ with probability distribution P_0, and mean direction $\boldsymbol{\mu_0}$. Suppose the covariance matrix $\boldsymbol{V}_0 = Cov(\boldsymbol{X})$ has full rank k. Consider compact sets*

$$\Theta = \mathcal{S}^{k-2} = \{\boldsymbol{\theta} \in \mathbb{R}^{k-1} : \|\boldsymbol{\theta}\| = 1\} \subset \mathbb{R}^{k-1},$$

and

$$\overline{B}_\delta(\boldsymbol{\mu_0}) = \{\boldsymbol{\mu} \in \mathcal{S}^{k-1} : \arccos \boldsymbol{\mu}^\top \boldsymbol{\mu}_0 \leqslant \delta\} \subset \mathbb{R}^k,$$

where $\delta > 0$ is given. Let $\mathcal{D}$ denote a class of subsets of $\mathcal{S}^{k-1}$, where

$$\mathcal{D} = \{D(\boldsymbol{\theta}, \boldsymbol{\mu}) : \boldsymbol{\theta} \in \Theta, \boldsymbol{\mu} \in \overline{B}_\delta(\boldsymbol{\mu_0})\},$$
$$D(\boldsymbol{\theta}, \boldsymbol{\mu}) = \{\boldsymbol{x} \in \mathbb{R}^k : \boldsymbol{\theta}^\top \boldsymbol{M} \boldsymbol{x} > 0\},$$

where $\boldsymbol{M}$ is an orthogonal matrix that depends on $\boldsymbol{\mu}$, as in (4.2.6). Then,

$$\inf_{D \in \mathcal{D}} P_0(D) > 0. \tag{4.3.5}$$

PROOF. Following the proof of Lemma 3.3.1, we argue by contradiction again. If the inequality (4.3.5) does not hold, there exists a sequence $\{\boldsymbol{\theta}_m, \boldsymbol{\mu}_m\}_{m \geqslant 1}$, such that

$$P_0\Big(D(\boldsymbol{\theta}_m, \boldsymbol{\mu}_m)\Big) < \frac{1}{m}.$$

The fact that $\boldsymbol{\Theta}$ and $\overline{B}_\delta(\boldsymbol{\mu_0})$ are compact implies that there exists a convergent subsequence $\{\boldsymbol{\theta}_{m(j)}, \boldsymbol{\mu}_{m(j)}\}_{j \geqslant 1}$ and $\{\boldsymbol{\theta}^*, \boldsymbol{\mu}^*\}$, such that

$$\boldsymbol{\theta}_{m(j)} \to \boldsymbol{\theta}^*,$$
$$\boldsymbol{\mu}_{m(j)} \to \boldsymbol{\mu}^*,$$

as $j \to \infty$.

Hence, the dominated convergence theorem (Theorem 2.4) implies that

$$P_0(D(\boldsymbol{\theta}^*, \boldsymbol{\mu}^*)) = 0,$$

and then

$$Var\Big((\boldsymbol{M}^*\boldsymbol{X})^\top \boldsymbol{\theta}^*\Big) = 0,$$

where $\boldsymbol{M}^* = f(\boldsymbol{\mu}^*)$ as defined in (4.2.6). This contradicts the assumption that $\boldsymbol{V}_0$ has full rank, and so contradiction proves the inequality (4.3.5). □

It should be noted that Lemma 4.3.1 is used for proving Lemma 4.3.2 and Lemma 4.3.3, which involves the case of local alternatives. Thus, $\overline{B}_\delta(\boldsymbol{\mu_0})$ is defined through the idea of arc-length to be consistent with $\{\boldsymbol{\mu}_n\}_{n\geqslant 1}$.

Similar to the proof of Theorem 3.1, the result in Lemma 4.3.1 needs to be extended to the measure P_n in probability.

Lemma 4.3.2 *Assume that the covariance matrix $\boldsymbol{V}_0$, as defined in Lemma 4.3.1, has full rank. Then there exists a constant $c > 0$, such that*

$$P_0\{\inf_{D\in\mathcal{D}} P_n(D) \geqslant c\} \to 1,$$

as $n \to \infty$.

Note that based on Lemma 4.3.1, this lemma is proved by modifying the proof of Lemma 3.3.2.

Lemma 4.3.3 *Let local alternatives $\{\boldsymbol{\mu}_n\}_{n\geqslant 1}$ be defined as in (4.2.1), the coordinate transforming matrix $\boldsymbol{M}_n = \boldsymbol{f}(\boldsymbol{\mu}_n)$ be defined as in (4.2.6), and $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots, \boldsymbol{Z}_n$ be defined as in (4.2.7). Then, $\boldsymbol{M}_n\boldsymbol{\mu}_n = \boldsymbol{0}_{k-1}$ lies in the interior of the convex hull of $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots, \boldsymbol{Z}_n$, with probability approaching 1, as $n \to \infty$, i.e.*

$$\lim_{n\to\infty} P_0\Big[\boldsymbol{0}_{k-1} \in int\Big(Conv(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots, \boldsymbol{Z}_n)\Big)\Big] = 1,$$

where $int(\cdot)$ denotes the interior of a set, and

$$Conv(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots, \boldsymbol{Z}_n) = \left\{ \lambda_1\boldsymbol{Z}_1 + \cdots, \lambda_n\boldsymbol{Z}_n | \lambda_1, \cdots, \lambda_n \in [0,1], \sum_{i=1}^n \lambda_i = 1 \right\}.$$

In Section 4.2, we define empirical likelihood on the subspace $\mathbb{R}^{k-1}$ through the coordinate transforming matrix $\boldsymbol{M} = \boldsymbol{f}(\boldsymbol{\mu})$, where $\boldsymbol{\mu}$ is the mean direction parameter. Local alternatives are defined via the idea of $\log_{\boldsymbol{\mu}_n}$ map. Under the assumption of antipodal random sample, the $\log_{\boldsymbol{\mu}_n}$ map is one-to-one. Thus, we only need to consider whether $\boldsymbol{M}_n\boldsymbol{\mu}_n = \boldsymbol{0}_{k-1}$ lies in the interior of the convex hull of $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots, \boldsymbol{Z}_n$, with probability converging to 1, as $n \to \infty$, in order to have the non-trivial case of $\{p_i\}_{i=1}^n$, such that $\sum_{i=1}^n p_i = 1$ and $0 < p_i < 1$, $i = 1, 2, \cdots, n$; see the analogous result of Lemma 3.3.3.

The rotation matrix $\boldsymbol{Q}$, as defined in Section 4.3.2, satisfies $\boldsymbol{M}_0\boldsymbol{Q} = \boldsymbol{M}_n$. Now, we calculate the order of the term $\boldsymbol{Q} - \boldsymbol{I}$ for the proof of Theorem 4.1.

Lemma 4.3.4 *Consider the rotation matrix $\boldsymbol{Q}$ is defined in (4.3.2). Then,*

$$\|\boldsymbol{Q} - \boldsymbol{I}\| = O_p(n^{-1/2}).$$

PROOF. From (4.3.2),

$$\begin{aligned}
\boldsymbol{Q} - \boldsymbol{I} =& \sin\Big(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\Big)\boldsymbol{A} + \Big\{\cos\Big(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\Big) - 1\Big\}\Big\{\boldsymbol{\mu}_0\boldsymbol{\mu}_0^\top + \frac{\boldsymbol{\eta}\boldsymbol{\eta}^\top}{\|\boldsymbol{\eta}\|^2}\Big\} \\
=& \Big(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}} + O_p(n^{-3/2})\Big)\Big(\frac{\boldsymbol{\mu}_0\boldsymbol{\eta}^\top}{\|\boldsymbol{\eta}\|} - \frac{\boldsymbol{\eta}\boldsymbol{\mu}_0^\top}{\|\boldsymbol{\eta}\|}\Big) + O_p(n^{-1})\Big\{\boldsymbol{\mu}_0\boldsymbol{\mu}_0^\top + \frac{\boldsymbol{\eta}\boldsymbol{\eta}^\top}{\|\boldsymbol{\eta}\|^2}\Big\} \\
=& \frac{\boldsymbol{\mu}_0\boldsymbol{\eta}^\top - \boldsymbol{\eta}\boldsymbol{\mu}_0^\top}{\sqrt{n}} + \boldsymbol{B}
\end{aligned}$$

where

$$\begin{aligned}
\sin\Big(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\Big) &= \frac{\|\boldsymbol{\eta}\|}{\sqrt{n}} + O_p(n^{-3/2}), \\
\cos\Big(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\Big) &= 1 + O_p(n^{-1}), \\
\|\boldsymbol{B}\| &= O_p(n^{-1}).
\end{aligned}$$

Thus,

$$\|\boldsymbol{Q} - \boldsymbol{I}\| = O_p(n^{-1/2}).$$

□

Similar to the proof of Theorem 3.1, in the last step to expand Wilks' statistic, we need to know $\|S_n^{-1} - \boldsymbol{\Sigma}^{-1}\|$. Here, we use Lemma 4.3.4 to calculate the bound of $S_n^{-1} - \boldsymbol{\Sigma}^{-1}$.

Lemma 4.3.5 *Let $\boldsymbol{X}, \boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$ denote k-dimensional i.i.d. random vectors on $\mathcal{S}^{k-1} \subset \mathbb{R}^k$ as in Theorem 4.1, and suppose*

$$\boldsymbol{\Sigma} = Cov(\boldsymbol{M}_0 \boldsymbol{X})$$

has full rank $k-1$. Let local alternatives be defined as in (4.2.1), and write

$$\boldsymbol{S}_n = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{M}_n \boldsymbol{X}_i \boldsymbol{X}_i^\top \boldsymbol{M}_n^\top. \tag{4.3.6}$$

Then,

$$\|S_n^{-1} - \boldsymbol{\Sigma}^{-1}\| = O_p(n^{-1/2}).$$

PROOF. Let S_n^{-1} be rewritten as in (3.3.14) and set

$$\boldsymbol{C} = \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{S}_n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1/2},$$

where $\boldsymbol{C} = \boldsymbol{C}^\top$ and $\boldsymbol{\Sigma}$ is assumed to have full rank k. We wish first to obtain a bound for $\|\boldsymbol{S}_n - \boldsymbol{\Sigma}\|$. Now

$$\begin{aligned}
\boldsymbol{S}_n =& \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{M}_n \boldsymbol{X}_i \boldsymbol{X}_i^\top \boldsymbol{M}_n^\top \\
=& \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{M}_0 \boldsymbol{Q} \boldsymbol{X}_i \boldsymbol{X}_i^\top \boldsymbol{Q}^\top \boldsymbol{M}_0^\top \\
=& \frac{1}{n} \sum_{i=1}^{n} \Big(\boldsymbol{M}_0 (\boldsymbol{I} + \boldsymbol{Q} - \boldsymbol{I}) \boldsymbol{X}_i \Big) \Big(\boldsymbol{M}_0 (\boldsymbol{I} + \boldsymbol{Q} - \boldsymbol{I}) \boldsymbol{X}_i \Big)^\top \\
=& \boldsymbol{S} + \frac{1}{n} \sum_{i=1}^{n} \Big(\boldsymbol{M}_0 \boldsymbol{X}_i \boldsymbol{X}_i^\top (\boldsymbol{Q} - \boldsymbol{I})^\top \boldsymbol{M}_0^\top + \boldsymbol{M}_0 (\boldsymbol{Q} - \boldsymbol{I}) \boldsymbol{X}_i \boldsymbol{X}_i^\top \boldsymbol{M}_0^\top \\
& + \big\{ \boldsymbol{M}_0 (\boldsymbol{Q} - \boldsymbol{I}) \boldsymbol{X}_i \big\} \big\{ \boldsymbol{M}_0 (\boldsymbol{Q} - \boldsymbol{I}) \boldsymbol{X}_i \big\}^\top \Big),
\end{aligned}$$

where

$$\boldsymbol{S} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{M}_0 \boldsymbol{X}_i \boldsymbol{X}_i^\top \boldsymbol{M}_0^\top.$$

The proof of Lemma 4.3.4 implies that

$$\|\boldsymbol{S}_n - \boldsymbol{S}\| = O_p(n^{-1/2}).$$

Since $\|\boldsymbol{M}_0\boldsymbol{X}_i\|$ is bounded for $i = 1, 2, \cdots, n$, we can apply the CLT to have

$$\|\boldsymbol{S} - \boldsymbol{\Sigma}\| = O_p(n^{-1/2}).$$

Consequently, we have

$$\|\boldsymbol{S}_n - \boldsymbol{\Sigma}\| = O_p(n^{-1/2}).$$

Hence, the bound of $\|\boldsymbol{C}\|$ satisfies

$$\|\boldsymbol{C}\| = \|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{S}_n - \boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1/2}\| = O_p(n^{-1/2}).$$

Now, consider

$$(\boldsymbol{I} + \boldsymbol{C})^{-1} = \boldsymbol{I} - \boldsymbol{C} + \boldsymbol{C}^2 - \boldsymbol{C}^3 + \cdots,$$

Therefore, we have

$$\|(\boldsymbol{I} + \boldsymbol{C})^{-1}\| = O_p(1).$$

Rewrite

$$\begin{aligned}\boldsymbol{S}_n^{-1} - \boldsymbol{\Sigma}^{-1} &= \boldsymbol{\Sigma}^{-1/2}(\boldsymbol{I} - \boldsymbol{C})^{-1}\boldsymbol{\Sigma}^{-1/2} - \boldsymbol{\Sigma}^{-1/2}\boldsymbol{I}\boldsymbol{\Sigma}^{-1/2} \\ &= -\boldsymbol{\Sigma}^{-1/2}((\boldsymbol{I} + \boldsymbol{C})^{-1}\boldsymbol{C})\boldsymbol{\Sigma}^{-1/2}.\end{aligned}$$

Therefore,

$$\|\boldsymbol{S}_n^{-1} - \boldsymbol{\Sigma}^{-1}\| = O_p(n^{-1/2}),$$

as required. □

Following the proof structure of Theorem 3.1, we also need bounds for $\|\boldsymbol{\lambda}\|$ and $\|\boldsymbol{\beta}\|$, which are given by Lemma 4.3.6 and Lemma 4.3.7.

Lemma 4.3.6 *Let $\boldsymbol{\Sigma}$ and $\boldsymbol{S}_n$ be defined as in Lemma 4.3.5, and local alternatives $\{\boldsymbol{\mu}_n\}_{n\geqslant 1}$ be defined as in (4.2.1). For $i = 1, 2, \cdots, n$, we write*

$$E_i = \boldsymbol{\lambda}^\top \boldsymbol{M}_n \boldsymbol{X}_i,$$

where $\boldsymbol{\lambda}$ is given in (4.2.10) and satisfies the condition (4.2.11) for $\boldsymbol{M}_n$.

Then,

$$\|\boldsymbol{\lambda}\| = O_p(n^{-1/2}). \tag{4.3.7}$$

PROOF. Let $\boldsymbol{Y}_1, \cdots, \boldsymbol{Y}_n$ be defined as in (4.2.7), and let

$$\overline{\boldsymbol{Y}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{Y}_i. \tag{4.3.8}$$

Note that

$$\begin{aligned} E[\overline{\boldsymbol{Y}}] =& \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{M}_0 \cdot E[\boldsymbol{X}] \\ =& \boldsymbol{M}_0 \cdot E[\boldsymbol{X}] \\ =& \boldsymbol{M}_0 \cdot \int_{\mathcal{S}^{k-1}} \boldsymbol{x} dP_0 \\ =& \boldsymbol{0}, \end{aligned}$$

where P_0 is the distribution of i.i.d. random vectors $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$, as in Theorem 4.1. Hence, CLT implies that $\|\overline{\boldsymbol{Y}}\| = O_p(n^{-1/2})$.

Recalling the proof of Lemma 3.3.6, we define

$$\begin{aligned} \overline{\boldsymbol{Z}} &= \boldsymbol{M}_n \overline{\boldsymbol{X}} \\ &= \boldsymbol{M}_0 \boldsymbol{Q} \overline{\boldsymbol{X}} \\ &= \boldsymbol{M}_0 \overline{\boldsymbol{X}} + \boldsymbol{M}_0 (\boldsymbol{Q} - \boldsymbol{I}) \overline{\boldsymbol{X}} \\ &= \overline{\boldsymbol{Y}} + \boldsymbol{M}_0 (\boldsymbol{Q} - \boldsymbol{I}) \overline{\boldsymbol{X}}. \end{aligned}$$

Then, we have

$$\|\overline{\boldsymbol{Z}}\| \leqslant \|\overline{\boldsymbol{Y}}\| + \|\boldsymbol{M}_0 (\boldsymbol{Q} - \boldsymbol{I}) \overline{\boldsymbol{X}}\|.$$

Since $E[\overline{\boldsymbol{X}}] = \boldsymbol{\alpha}$ as is shown in Theorem 4.1, CLT and Lemma 2.8.1 imply that $\|\overline{\boldsymbol{X}} - \boldsymbol{\alpha}\| = O_p(n^{-1/2})$ and then $\|\overline{\boldsymbol{X}}\| = O_p(1)$. Consequently, Lemma 4.3.4 implies that

$$\|\overline{\boldsymbol{Z}}\| = O_p(n^{-1/2}).$$

Following the proof in Lemma 3.3.6 and substituting the notation below:

$$F_n^* = \max_{1\leqslant i\leqslant n} \|\boldsymbol{M}_0\boldsymbol{Q}\boldsymbol{X}_i\|,$$
$$\boldsymbol{S}_1 = \frac{1}{n}\sum_{i=1}^{n} \frac{(\boldsymbol{M}_0\boldsymbol{Q}\boldsymbol{X}_i)(\boldsymbol{M}_0\boldsymbol{Q}\boldsymbol{X}_i)^\top}{1+E_i},$$

and $\boldsymbol{S}_n$ defined in (4.3.6) into Z_n^*, $\boldsymbol{S}_1$, $\boldsymbol{S}_n$ in Lemma 3.3.6, we prove that

$$\|\boldsymbol{\lambda}\| = O_p(n^{-1/2}).$$

□

Lemma 4.3.7 *Let E_i, $i = 1, 2, \cdots, n$, be defined as in Lemma 4.3.6. Then*

(*i*) $\max\limits_{1\leqslant i\leqslant n} |E_i| = O_p(n^{-1/2})$, *and*

(*ii*) $\max\limits_{1\leqslant i\leqslant n} \dfrac{1}{|1+E_i|} = O_p(1)$.

PROOF. To prove (i), for $i = 1, 2, \cdots, n$,

$$\begin{aligned} E_i &= \boldsymbol{\lambda}^\top \boldsymbol{M}_n \boldsymbol{X}_i \\ &= \boldsymbol{\lambda}^\top \boldsymbol{M}_0 \boldsymbol{Q} \boldsymbol{X}_i \\ &= \boldsymbol{\lambda}^\top \Big(\boldsymbol{M}_0 \boldsymbol{X}_i + \boldsymbol{M}_0 (\boldsymbol{Q} - \boldsymbol{I}) \boldsymbol{X}_i \Big) \\ &= \|\boldsymbol{\lambda}\| \Big(\boldsymbol{\theta}^\top \boldsymbol{M}_0 \boldsymbol{X}_i + \boldsymbol{\theta}^\top \boldsymbol{M}_0 (\boldsymbol{Q} - \boldsymbol{I}) \boldsymbol{X}_i \Big). \end{aligned}$$

Now,

$$\begin{aligned} \max_{1\leqslant i\leqslant n} |E_i| &= \|\boldsymbol{\lambda}\| \max_{1\leqslant i\leqslant n} |\boldsymbol{\theta}^\top \boldsymbol{M}_0 \boldsymbol{X}_i + \boldsymbol{\theta}^\top \boldsymbol{M}_0 (\boldsymbol{Q} - \boldsymbol{I}) \boldsymbol{X}_i| \\ &\leqslant \|\boldsymbol{\lambda}\| \Big(\max_{1\leqslant i\leqslant n} |\boldsymbol{\theta}^\top \boldsymbol{M}_0 \boldsymbol{X}_i| + \max_{1\leqslant i\leqslant n} |\boldsymbol{\theta}^\top \boldsymbol{M}_0 (\boldsymbol{Q} - \boldsymbol{I}) \boldsymbol{X}_i| \Big). \end{aligned}$$

Since $\boldsymbol{\theta}^\top \boldsymbol{M}_0 \boldsymbol{X}_1, \cdots, \boldsymbol{\theta}^\top \boldsymbol{M}_0 \boldsymbol{X}_n$ are bounded i.i.d. random variables,

$$\max_{1\leqslant i\leqslant n} |\boldsymbol{\theta}^\top \boldsymbol{M}_0 \boldsymbol{X}_i| = O_p(1).$$

Meanwhile,

$$\begin{aligned}\max_{1\leqslant i\leqslant n}|\boldsymbol{\theta}^\top \boldsymbol{M}_0(\boldsymbol{Q}-\boldsymbol{I})\boldsymbol{X}_i| &\leqslant \|\boldsymbol{\theta}\|\cdot\|\boldsymbol{M}_0\|\cdot\|\boldsymbol{Q}-\boldsymbol{I}\|\cdot\|\boldsymbol{X}\| \\ &= 1\cdot O_p(1)O_p(n^{-1/2})\cdot 1 \\ &= O_p(n^{-1/2}).\end{aligned}$$

Therefore, as $\|\boldsymbol{\lambda}\| = O_p(n^{-1/2})$, we have

$$\max_{1\leqslant i\leqslant n}|E_i| = O_p(n^{-1/2}).$$

To prove (ii), we follow the proof in Lemma 3.3.7, and then have

$$\max_{1\leqslant i\leqslant n}\frac{1}{|1+E_i|} = O_p(1).$$

□

Lemma 4.3.8 *Let $\boldsymbol{\Sigma}$ and $\boldsymbol{S}_n$ be defined as in Lemma 4.3.5, local alternatives $\{\boldsymbol{\mu}_n\}_{n\geqslant 1}$ be defined as in (4.2.1) and E_i, $i = 1, 2, \cdots, n$, be defined as in Lemma 4.3.6. We have*

$$\boldsymbol{\lambda} = \boldsymbol{S}_n^{-1}\boldsymbol{M}_n\overline{\boldsymbol{X}} + \boldsymbol{\beta}, \tag{4.3.9}$$

where

$$\boldsymbol{\beta} = \frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{S}_n^{-1}\boldsymbol{M}_n\boldsymbol{X}_i E_i^2}{1+E_i}, \tag{4.3.10}$$

and

$$\|\boldsymbol{\beta}\| = O_p(n^{-1}).$$

PROOF. As Lemma 4.3.7 shows,

$$\max_{1\leqslant i\leqslant n}|E_i| = O_p(n^{-1/2}).$$

We apply the formula (3.3.28) to

$$\sum_{i=1}^{n} p_i\boldsymbol{M}_n\boldsymbol{X}_i = \boldsymbol{0},$$

and then we have (4.3.9) and (4.3.10).

In order to calculate $\|\boldsymbol{\beta}\|$, we follow the proof structure of Lemma 2.8.3 to have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{M}_0\boldsymbol{Q}\boldsymbol{X}_i\|^3 \leqslant & \frac{H}{n}\sum_{i=1}^{n}\|\boldsymbol{M}_0\boldsymbol{Q}\boldsymbol{X}_i\|^2 \\ = & \frac{H}{n}\sum_{i=1}^{n}\Big\{(\boldsymbol{M}_0\boldsymbol{X}_i)^\top(\boldsymbol{M}_0\boldsymbol{X}_i) - 2\{\boldsymbol{M}_0(\boldsymbol{Q}-\boldsymbol{I})\boldsymbol{X}_i\}^\top\boldsymbol{M}_0\boldsymbol{X}_i \\ & + \{\boldsymbol{M}_0(\boldsymbol{Q}-\boldsymbol{I})\boldsymbol{X}_i\}^\top\{\boldsymbol{M}_0(\boldsymbol{Q}-\boldsymbol{I})\boldsymbol{X}_i\}\Big\},\end{aligned}$$

where

$$\begin{aligned}H = & \max_{1\leqslant i\leqslant n}\|\boldsymbol{M}_0\boldsymbol{Q}\boldsymbol{X}_i\| \\ \leqslant & \max_{1\leqslant i\leqslant n}\|\boldsymbol{M}_0\boldsymbol{X}_i\| + \max_{1\leqslant i\leqslant n}\|\boldsymbol{M}_0(\boldsymbol{Q}-\boldsymbol{I})\boldsymbol{X}_i\| \\ = & O_p(1) + O_p(n^{-1/2}) \\ = & O_p(1).\end{aligned}$$

Therefore, we have

$$\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{M}_0\boldsymbol{Q}\boldsymbol{X}_i\|^3 = O_p(1).$$

Finally, as $\|\boldsymbol{S}_n^{-1}\| = O_p(1)$, the norm bound of $\boldsymbol{\beta}$ is

$$\begin{aligned}\|\boldsymbol{\beta}\| \leqslant & \|\boldsymbol{S}_n^{-1}\|\cdot\|\boldsymbol{\lambda}\|^2\cdot\max_{1\leqslant i\leqslant n}\frac{1}{|1+E_i|}\cdot\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{M}_0\boldsymbol{Q}\boldsymbol{X}_i\|^3 \\ = & O_p(n^{-1}).\end{aligned}$$

□

4.3.4 *Proof of Theorem 4.1*

The structure of the proof is similar to that of Theorem 3.1 but with some technical diferences. In order to avoid the trivial cases that there exists an index j, such that $p_j = 0$, we require $\boldsymbol{M}_n\boldsymbol{\mu}_n = \boldsymbol{0}$ lies in the interior of the convex hull of $\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots, \boldsymbol{Z}_n$, which has been supported by Lemma 4.3.3.

The proof is organised as following steps.

1. Rewrite the formula (4.2.10) by substituting the formula (4.3.4).
2. Expand $\boldsymbol{M}_n = \boldsymbol{M}_0\boldsymbol{Q}$ by Taylor series for the next step.
3. Expand the Wilks' statistic using Taylor series, cf. proof of Theorem 3.1.

PROOF OF THEOREM 4.1 :

Step 1. Since the formula (4.3.4) shows the relation between matrices $\boldsymbol{M}_0$ and $\boldsymbol{M}_n$, we substitute $\boldsymbol{M}_n = \boldsymbol{M}_0\boldsymbol{Q}$ into $\sum_{i=1}^n p_i \boldsymbol{M}_n \boldsymbol{X}_i = \boldsymbol{0}$ to get

$$\sum_{i=1}^{n} p_i \boldsymbol{M}_0 \boldsymbol{Q} \boldsymbol{X}_i = \boldsymbol{0}_{k-1}.$$

Hence,

$$p_i = \frac{1}{n(1 + \boldsymbol{\lambda}^\top \boldsymbol{M}_0 \boldsymbol{Q} \boldsymbol{X}_i)},\ i = 1, 2, \cdots, n,$$

where $\boldsymbol{\lambda}$ satisfies

$$\sum_{i=1}^{n} \frac{\boldsymbol{M}_0 \boldsymbol{Q} \boldsymbol{X}_i}{n(1 + \boldsymbol{\lambda}^\top \boldsymbol{M}_0 \boldsymbol{Q} \boldsymbol{X}_i)} = \boldsymbol{0}_{k-1}.$$

Step 2. The matrix $\boldsymbol{Q}$ includes terms $\cos(\|\boldsymbol{\eta}\|/\sqrt{n})$ and $\sin(\|\boldsymbol{\eta}\|/\sqrt{n})$. Applying Taylor series to $\boldsymbol{Q}$ at $\boldsymbol{0}_k$, we have

$$\begin{aligned}
\boldsymbol{M}_0\boldsymbol{Q} =& \boldsymbol{M}_0 \left(\boldsymbol{I}_k + \sin\left(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\right)\boldsymbol{A} + \left\{ \cos\left(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\right) - 1 \right\} \left\{ \boldsymbol{\mu}_0\boldsymbol{\mu}_0^\top + \frac{\boldsymbol{\eta}\boldsymbol{\eta}^\top}{\|\boldsymbol{\eta}\|^2} \right\} \right) \\
=& \boldsymbol{M}_0 \left(\boldsymbol{I}_k + \frac{\boldsymbol{\mu}_0\boldsymbol{\eta}^\top - \boldsymbol{\eta}\boldsymbol{\mu}_0^\top}{\sqrt{n}} \right) + O_p(n^{-1}) \\
=& \boldsymbol{M}_0 - \frac{\boldsymbol{M}_0\boldsymbol{\eta}\boldsymbol{\mu}_0^\top}{\sqrt{n}} + O_p(n^{-1}).
\end{aligned} \tag{4.3.11}$$

where

$$\begin{aligned}
\sin\left(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\right) &= \frac{\|\boldsymbol{\eta}\|}{\sqrt{n}} + O_p(n^{-3/2}), \\
\cos\left(\frac{\|\boldsymbol{\eta}\|}{\sqrt{n}}\right) &= 1 + O_p(n^{-1}).
\end{aligned}$$

Step 3. Lemma 4.3.6 shows that

$$\|\boldsymbol{\lambda}\| = O_p(n^{-1/2}).$$

Lemma 4.3.8 shows that

$$\boldsymbol{\lambda} = \boldsymbol{S}_n^{-1}\boldsymbol{M}_n\overline{\boldsymbol{X}} + \boldsymbol{\beta},$$

where

$$\boldsymbol{\beta} = \frac{1}{n}\sum_{i=1}^{n}\frac{\boldsymbol{S}_n^{-1}\boldsymbol{M}_n\boldsymbol{X}_i E_i^2}{1+E_i},$$

and

$$\|\boldsymbol{\beta}\| = O_p(n^{-1}).$$

Now, let $\overline{\boldsymbol{Y}}$ be defined as in (4.3.8) and $\boldsymbol{\alpha}$ be defined as in (4.3.1). Note that

$$\sqrt{n}\boldsymbol{\Sigma}^{-1/2}\overline{\boldsymbol{Y}} \xrightarrow{D} N_{k-1}(\boldsymbol{0}, \boldsymbol{I}_{k-1}),$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of i.i.d. random vectors $\boldsymbol{Y}_1, \boldsymbol{Y}_2 \cdots, \boldsymbol{Y}_n$ as in the statement of Theorem 4.1. Recalling (4.3.1) and applying the CLT again, we have

$$\|\overline{\boldsymbol{X}} - \boldsymbol{\alpha}\| = O_p(n^{-1/2}).$$

We now expand Wilks' statistic (cf. proof of Theorem 3.1):

$$\begin{aligned}
-2\log\mathcal{R}(\boldsymbol{\mu}_n) =& 2\sum_{i=1}^{n}\log(1+\boldsymbol{\lambda}^\top\boldsymbol{M}_n\boldsymbol{X}_i) \\
=& 2\sum_{i=1}^{n}\boldsymbol{\lambda}^\top\boldsymbol{M}_n\boldsymbol{X}_i - \sum_{i=1}^{n}(\boldsymbol{\lambda}^\top\boldsymbol{M}_n\boldsymbol{X}_i)(\boldsymbol{\lambda}^\top\boldsymbol{M}_n\boldsymbol{X}_i)^\top + 2\sum_{i=1}^{n}\xi_i.
\end{aligned}$$

Note that

$$\begin{aligned}
\Big|\sum_{i=1}^{n}\xi_i\Big| \leqslant& \sum_{i=1}^{n}|\xi_i| \\
=& O_p(n^{-1/2}).
\end{aligned}$$

Now, we substitute (4.3.9) and (4.3.11) into $-2\log\mathcal{R}(\boldsymbol{\mu}_n)$:

$$\begin{aligned}
-2\log\mathcal{R}(\boldsymbol{\mu}_n) =& n\Big(\overline{\boldsymbol{Y}} - \frac{\boldsymbol{M}_0\boldsymbol{\eta}\boldsymbol{\mu}_0^\top\overline{\boldsymbol{X}}}{\sqrt{n}}\Big)^\top \boldsymbol{S}_n^{-1}\Big(\overline{\boldsymbol{Y}} - \frac{\boldsymbol{M}_0\boldsymbol{\eta}\boldsymbol{\mu}_0^\top\overline{\boldsymbol{X}}}{\sqrt{n}}\Big) + R_1 \\
=& n\left\{\overline{\boldsymbol{Y}} - \frac{\boldsymbol{M}_0\boldsymbol{\eta}\boldsymbol{\mu}_0^\top(\boldsymbol{\alpha}+\overline{\boldsymbol{X}}-\boldsymbol{\alpha})}{\sqrt{n}}\right\}^\top \Big(\boldsymbol{\Sigma}^{-1} + \boldsymbol{T}\Big) \\
& \cdot \left\{\overline{\boldsymbol{Y}} - \frac{\boldsymbol{M}_0\boldsymbol{\eta}\boldsymbol{\mu}_0^\top(\boldsymbol{\alpha}+\overline{\boldsymbol{X}}-\boldsymbol{\alpha})}{\sqrt{n}}\right\} + R_1 \\
=& n\left\{\overline{\boldsymbol{Y}} - \frac{\|\boldsymbol{\alpha}\|\boldsymbol{M}_0\boldsymbol{\eta}}{\sqrt{n}}\right\}^\top \boldsymbol{\Sigma}^{-1}\left\{\overline{\boldsymbol{Y}} - \frac{\|\boldsymbol{\alpha}\|\boldsymbol{M}_0\boldsymbol{\eta}}{\sqrt{n}}\right\} + R_2,
\end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{T} &= \boldsymbol{S}_n^{-1} - \boldsymbol{\Sigma}^{-1}, \\
R_1 &= -n\boldsymbol{\beta}^\top\boldsymbol{S}_n^{-1}\boldsymbol{\beta} + 2\sum_{i=1}^n \xi_i + r_1, \\
|r_1| &= O_p(n^{-1}), \\
|R_1| &= O_p(n^{-1/2}).
\end{aligned}$$

Note that

$$\begin{aligned}
\boldsymbol{\Phi} =& \overline{\boldsymbol{Y}} - \frac{\|\boldsymbol{\alpha}\|\boldsymbol{M}_0\boldsymbol{\eta}}{\sqrt{n}}, \\
\|\boldsymbol{\Phi}\| =& O_p(n^{-1/2}), \\
\boldsymbol{\Gamma} =& \frac{\boldsymbol{M}_0\boldsymbol{\eta}\boldsymbol{\mu}_0^\top(\overline{\boldsymbol{X}}-\boldsymbol{\alpha})}{\sqrt{n}} \\
\|\boldsymbol{\Gamma}\| =& O_p(n^{-1}), \\
R_2 =& R_1 + n\boldsymbol{\Phi}^\top\boldsymbol{T}\boldsymbol{\Phi} - n\boldsymbol{\Phi}^\top(\boldsymbol{\Sigma}^{-1}+\boldsymbol{T})\boldsymbol{\Gamma} \\
& - n\boldsymbol{\Gamma}^\top(\boldsymbol{\Sigma}^{-1}+\boldsymbol{T})(\boldsymbol{\Phi}-\boldsymbol{\Gamma}), \\
|R_2| =& O_p(n^{-1/2}) + n\cdot O_p(n^{-1/2})O_p(n^{-1/2})O_p(n^{-1/2}) \\
& + n\cdot O_p(n^{-1/2})O_p(1)O_p(n^{-1}) \\
& + n\cdot O_p(n^{-1})O_p(1)\Big(O_p(n^{-1/2}) + O_p(n^{-1})\Big) \\
=& O_p(n^{-1/2}) + O_p(n^{-1/2}) + O_p(n^{-1/2}) + O_p(n^{-1/2}) \\
=& O_p(n^{-1/2}).
\end{aligned}$$

Hence, we have

$$-2\log\mathcal{R}(\boldsymbol{\mu}_n) \xrightarrow{D} \chi^2_{k-1}(\|\alpha\|^2\boldsymbol{\eta}^\top \boldsymbol{M}_0^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{M}_0\boldsymbol{\eta}),$$

as $n \to \infty$.

□

4.4 EMPIRICAL LIKELIHOOD RATIO ALGORITHM

Since statistics are widely used in earth science, now we mainly focus on $\mathcal{S}^2$. The algorithm for calculating the empirical likelihood ratio against the mean direction parameter is proposed as follows.

Algorithm 1 Empirical Likelihood Ratio Evaluation at Mean Direction Parameter $\boldsymbol{\mu}$ on $\mathcal{S}^2$.

1. Given a sample on $\mathcal{S}^2$. For the mean direction parameter $\boldsymbol{\mu}$, find the 2×3 coordinate matrix $\boldsymbol{M} = \boldsymbol{f}(\boldsymbol{\mu})$, as in (4.2.5).
2. Map the sample from $\mathcal{S}^2$ onto $T_\mu\mathcal{S}^2$.
3. Use the image of the sample in $T_\mu\mathcal{S}^2$ to evaluate the empirical likelihood ratio function at $(0, 0)$.

Although Algorithm 1 focuses on the case $k = 3$, the details are similar when $k > 3$. Figure 4.4.9 shows the empirical likelihood ratio against the mean direction parameter according to the algorithm above. Note that the sample is generated from the Von-Mises Fisher distribution with parameter $\kappa = 25$, sample size 500, mean direction $(1, 0, 0)$. Within the convex hull of the sample, the empirical likelihood ratio is calculated against the mean parameter. In order to show the graph properly, we use stereographic projection, as introduced in Section 2.11, to map the points on $\mathcal{S}^2$ to the plane $\{(x, y, z) \in \mathbb{R}^3 | x = 1\}$.

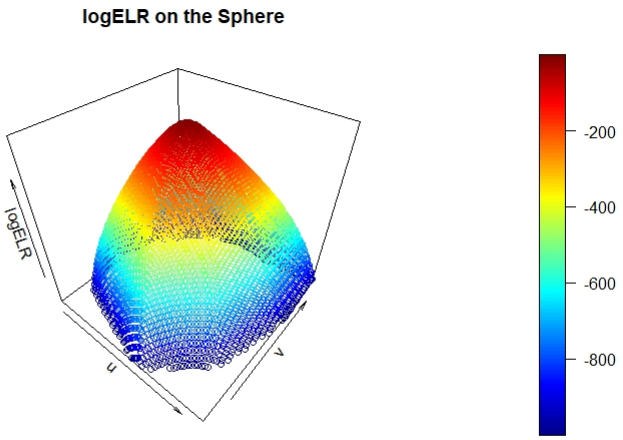


Figure 4.4.9: Empirical Likelihood Ratio at the Mean Direction Parameter.

4.5 OPERATING CHARACTERISTICS IN HYPOTHESIS TESTING

In statistical inference, since the power of the test is $1 - Pr(\text{Type II Error})$, with the probability of type I error given, the power of the test is increasing as the sample size increases. In this section, we use quantiles by checking chi-squared table and bootstrap calibration to calculate probabilities of type I error and type II error, and we name them tabular probabilities of type I and II errors and bootstrap calibration probabilities of type I and II errors. The algorithm for this numerical realization is shown in following steps.

4.5.1 *Probability of Type I Error*

In the hypothesis testing, we assume the nominal probability of type I error is α. Suppose the hypothesis test:

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0, \ v.s. \ H_1 : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0,$$

and the sample of the test is $\mathscr{X}$, as observed from i.i.d. random vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n$.

Algorithm 2 Hypothesis Testing under H_0

1. Use Algorithm 1 to calculate $-2\log\mathcal{R}(\boldsymbol{\mu}_0)$.
2. [Tabular Method] Find the tabular quantile $\chi^2_{2,1-\alpha}$. Under H_0, if $-2\log\mathcal{R}(\boldsymbol{\mu}_0) \leqslant \chi^2_{2,1-\alpha}$, accept H_0; otherwise, reject H_0.
3. Calculate the sample mean direction $\boldsymbol{\mu}_s = \sum_{i=1}^n \boldsymbol{X}_i / \| \sum_{i=1}^n \boldsymbol{X}_i \|$.
4. Generate bootstrap resamples $\mathscr{X}^*$ of size n sampled randomly with replacement, from the original sample $\mathscr{X}$. Within each resample $\mathscr{X}^*$, evaluate $W_i^* = -2\log\mathcal{R}(\boldsymbol{\mu_s}), 1 \leqslant i \leqslant B$.
5. [Bootstrap Calibration Method] Order $W_i^*, 1 \leqslant i \leqslant B$, as $W^*_{[1]} \leqslant \cdots \leqslant W^*_{[B]}$. Find $W^*_{[B(1-\alpha)]}$, the bootstrap calibration quantile. If $-2\log\mathcal{R}(\boldsymbol{\mu}_0) \leqslant W^*_{[B(1-\alpha)]}$, accept H_0; otherwise, reject H_0.

In order to calculate the probability of type I error, we set up a Monte Carlo loop to repeat the steps above M times. If the null hypothesis is accepted A times in Step 2, the tabular probability of type I error is $1 - A/M$. If the null hypothesis is accepted A' times in Step 5, the bootstrap calibration probability of type I error is $1 - A'/M$.

Von-Mises Fisher Distribution

Now, for the Monte-Carlo loop, we generate samples from a Von-Mises Fisher distribution with mean direction $(1, 0, 0)$, where $\kappa = 25, 15, 5, 1$ to test the influence of κ on the probability of type I error. In each experiment, the number of Monte-Carlo runs is 2500. For each Monte-Carlo run, bootstrap calibration is done for 200 times. Within the hypothesis test, the nominal probability of type I error is given as $\alpha = 0.05$. Note that the sample size ranges respectively at 10, 15, 25, 50, and 200. Table 2 displays the result of tabular probability of type I error and B.C. probability of type I error, where B.C. stands for bootstrap calibration.

κ	Sample Size	Tabular α	B.C. α
25	10	0.1748 (0.1201)	0.0248 (0.0467)
	15	0.1204 (0.0840)	0.0444 (0.0532)
	25	0.0856 (0.0560)	0.0560 (0.0460)
	50	0.0580 (0.0331)	0.0508 (0.0311)
	200	0.0536 (0.0159)	0.0564 (0.0163)
15	10	0.1776 (0.1209)	0.0256 (0.0256)
	15	0.1136 (0.0819)	0.0508 (0.0567)
	25	0.0728 (0.0520)	0.0428 (0.0405)
	50	0.0568 (0.0327)	0.0532 (0.0317)
	200	0.0472 (0.0150)	0.0492 (0.0153)
5	10	0.1692 (0.1186)	0.0304 (0.0543)
	15	0.1302 (0.0785)	0.0444 (0.0532)
	25	0.0688 (0.0506)	0.0504 (0.0438)
	50	0.0524 (0.0315)	0.0520 (0.0314)
	200	0.0520 (0.0157)	0.0564 (0.0163)
1	10	0.1484 (0.1124)	0.0360 (0.0589)
	15	0.0920 (0.0746)	0.0576 (0.0602)
	25	0.0616 (0.0481)	0.0500 (0.0436)
	50	0.0616 (0.0340)	0.0604 (0.0337)
	200	0.0480 (0.0151)	0.0500 (0.0154)

Table 2: Tabular probability of type I error and bootstrap calibration probability of type I error. The standard errors for both probabilities of type I error are given in brackets.

Note that the standard error of the proportion $\hat{p}$ is given by

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Here the proportion is substutited by the probability of type I and the standard error of the probability of type II error is calculated in further investigation..

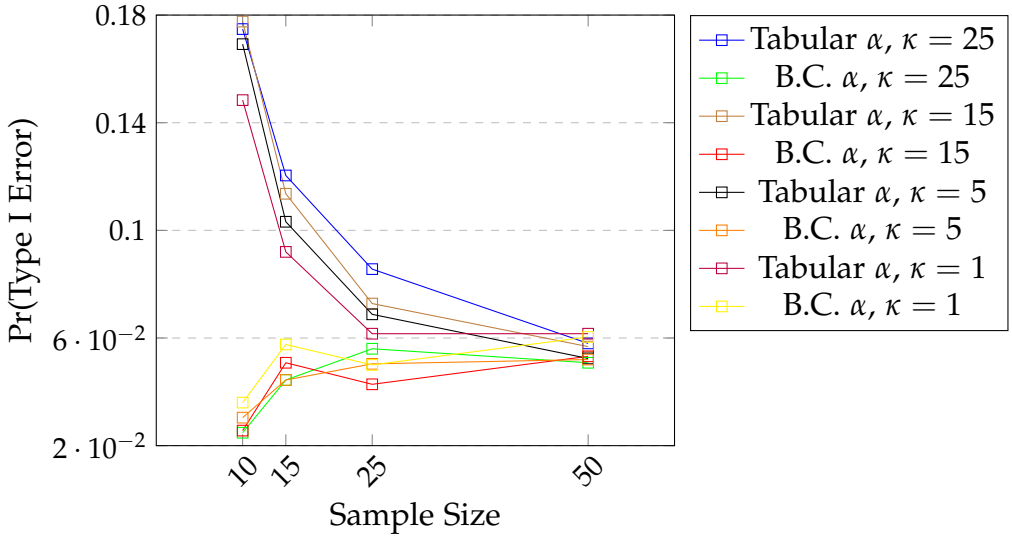


Figure 4.5.10: Tabular probability of type I error and bootstrap calibration probability of type I error. Considering the probability of type I error at sample size 200 does not differ much from that at sample size 50, they are omitted here.

Observing Table 2, Figure 4.5.10, we make the following comments:

1. For small sample sizes, say $n \leqslant 50$, the probability of type I error calculated by chi-squared tabular quantile is decreasing from a higher level to the level around 0.05, while the probability of type I error evaluated by bootstrap calibration quantile is increasing from a lower level to the level around 0.05. It suggests probabilities of type I error by tabular quantiles at small sample sizes lose the accuracy. The probabilities of type I error evaluated by bootstrap quantiles at small samples sizes except 10 still behaves well.

2. The outcome suggests that κ is a subtle factor that influences the probabilities of type I error. With κ decreasing, the probability of type I

error decreases at small sample sizes before 50. However, this outcome does not repeat in the experiment with Kent distribution; we will show it later in Figure 4.5.11. Therefore, this question needs more discussion.

3. In hypothesis test, the null hypothesis is usually carefully constructed. Then, we do not easily reject the null hypothesis without very strong evidence that the null hypothesis should be rejected. Thus, from this idea, we prefer letting the actual probability of type I error slightly smaller than the nominal probability of type I error, rather than having a larger actual probability of type I error. From this idea, the quantile calculated by the bootstrap calibration is more preferable in hypothesis testing.

Kent Distribution (5-Parameter Fisher-Bingham Distribution)

It is known in Fisher et al. (1993, Chapter 4.3.2 & 4.4.5) that von Mises Fisher distribution is symmetric about the mean direction. We would like to investigate what happens to the outcome of the experiments when the distribution is asymmetric, say Kent distribution. Now, we generate samples from a Kent distribution with mean direction $(1, 0, 0)$, where the concentration parameter κ is chosen to be 25, 15, 5, and 1 respectively. Recall that the Kent distribution is unimodal, if $\kappa/\beta \geqslant 2$. In this case, the contours of constant density have ellipse-like symmetry. Thus, we set the ellipticity parameter $\beta = \kappa/4$. In the experiment, the number of Monte-Carlo runs is 2500. For each Monte-Carlo run, bootstrap calibration is done for 200 times. Within the hypothesis test, the nominal probability of type I error is given $\alpha = 0.05$. The sample size ranges respectively at $10, 15, 25, 50$, and 200. Table 4 shows the result of tabular probability of type I error and B.C. probability of type I error, where B.C. stands for bootstrap calibration.

κ	Sample Size	Tabular α	B.C. α
25	10	0.1736 (0.1198)	0.0224 (0.0468)
	15	0.1092 (0.0805)	0.0400 (0.0506)
	25	0.0744 (0.0525)	0.0484 (0.0429)
	50	0.0548 (0.0322)	0.0440 (0.0290)
	200	0.0544 (0.0160)	0.0520 (0.0157)
15	10	0.1724 (0.1194)	0.0320 (0.0557)
	15	0.1200 (0.0839)	0.0420 (0.0518)
	25	0.0812 (0.0546)	0.0544 (0.0454)
	50	0.0532 (0.0317)	0.0488 (0.0305)
	200	0.0568 (0.0164)	0.0564 (0.0163)
5	10	0.1624 (0.1166)	0.0296 (0.0536)
	15	0.1124 (0.0816)	0.0480 (0.0552)
	25	0.0768 (0.0533)	0.0552 (0.0457)
	50	0.0584 (0.0332)	0.0592 (0.0334)
	200	0.0516 (0.0156)	0.0556 (0.0162)
1	10	0.1644 (0.1172)	0.0332 (0.0567)
	15	0.0920 (0.0746)	0.0496 (0.0561)
	25	0.0640 (0.0490)	0.0536 (0.0450)
	50	0.0520 (0.0314)	0.0556 (0.0324)
	200	0.0532 (0.0159)	0.0556 (0.0162)

Table 4: Tabular probability of type I error and bootstrap calibration probability of type I error. The standard errors for both probabilities of type I error are given in brackets.

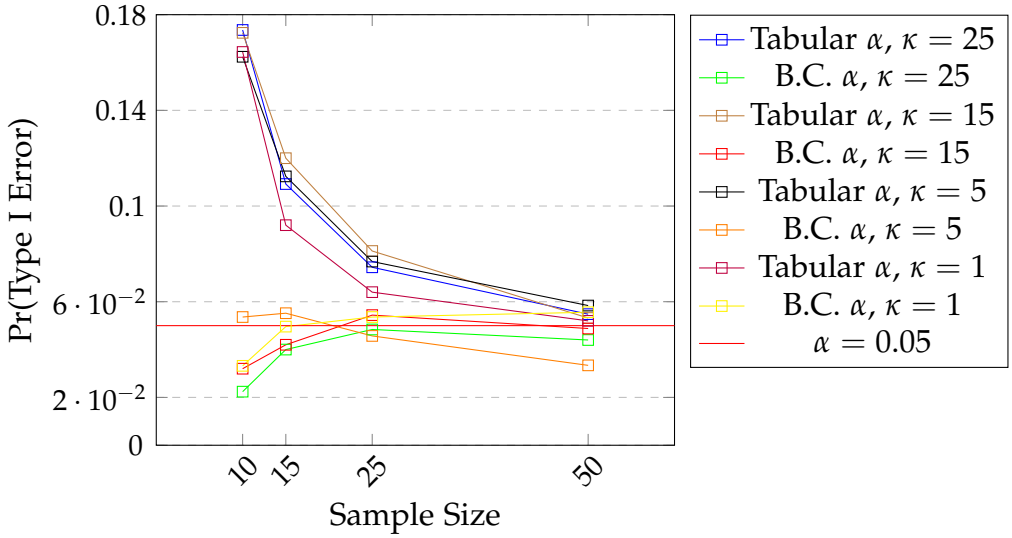


Figure 4.5.11: Tabular probability of type I error and bootstrap calibration probability of type I error of the test. Considering the probability of type I error at sample size 200 does not differ much from that at sample size 50, they are not shown here.

Observing Table 4, Figure 4.5.11, we find the same results as those in von-Mises Fisher distribution experiments except the influence made by κ.

4.5.2 *Probability of Type II Error*

Now, we give the algorithm of calculating the tabular probability of type II error and bootstrap calibration probability of type II error. We assume the nominal type I error α and a given alternative hypothesis H_1. Suppose the hypothesis test:

$$H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0, \ v.s. \ H_1 : \boldsymbol{\mu} = \boldsymbol{\mu}_1,$$

and the sample of the test $\mathscr{Y}$.

Algorithm 3 Hypothesis Testing under H_1

1. Use Algorithm 1 to calculate $-2\log\mathcal{R}(\boldsymbol{\mu}_0)$.
2. [Tabular Method] Find the tabular quantile $\chi^2_{2,1-\alpha}$. If $-2\log\mathcal{R}(\boldsymbol{\mu}_0) \leqslant \chi^2_{2,1-\alpha}$, accept H_0; otherwise, reject H_0.
3. Find the rotation matrix $\boldsymbol{Q}$ as in (4.3.2) for $\boldsymbol{\mu}_1$ to $\boldsymbol{\mu}_0$. Then, rotate $\mathscr{Y}$ to $\mathscr{Y}'$ by $\boldsymbol{Q}$. Generate bootstrap resamples $\mathscr{Y}^*$ of size n sampled randomly with replacement from the sample $\mathscr{Y}'$. Within each resample $\mathscr{Y}^*$, evaluate $-2\log\mathcal{R}(\boldsymbol{\mu_0})$, written as $W_i^*, 1 \leqslant i \leqslant B$.
4. [Bootstrap Calibration Method] Order $W_i^*, 1 \leqslant i \leqslant B$, as $W^*_{[1]} \leqslant \cdots \leqslant W^*_{[B]}$. Find $W^*_{[B(1-\alpha)]}$, the bootstrap calibration quantile. If $-2\log\mathcal{R}(\boldsymbol{\mu}_0) \leqslant W^*_{[B(1-\alpha)]}$, accept H_0; otherwise, reject H_0.

In order to calculate the probability of type II error, we set up a Monte Carlo loop to repeat the steps above M times. If the null hypothesis is accepted A times in Step 2, the tabular probability of type II error is A/M. If the null hypothesis is accepted A' times in Step 4, the bootstrap calibration probability of type II error is A'/M. Then, the power of the test is $Power = 1 - Pr(\text{Type II Error})$.

von Mises Fisher Distribution

Now, we conduct Monte-Carlo experiments for probability of type II error, following Algorithm 3. The sample is respectively generated from a von Mises Fisher distribution with mean direction $\boldsymbol{\mu}_1$ listed as follows:

1. $\boldsymbol{\mu_1} = (0.9986295, 0.052335956, 0)$, the angular distance 0.0524 to $(1, 0, 0)$;
2. $\boldsymbol{\mu_1} = (0.9992118, 0.039695808, 0)$, the angular distance 0.0397 to $(1, 0, 0)$;
3. $\boldsymbol{\mu_1} = (0.9996341, 0.027049304, 0)$, the angular distance 0.0271 to $(1, 0, 0)$;
4. $\boldsymbol{\mu_1} = (0.99988963, 0.014398469, 0)$, the angular distance 0.0144 to $(1, 0, 0)$;
5. $\boldsymbol{\mu_1} = (0.9999985, 0.001745328, 0)$, the angular distance 0.0017 to $(1, 0, 0)$.

The concentration parameter κ is chosen to be $25, 15, 5, 1$ respectively to test whether κ influences the Type II error. In each experiment, the number of Monte-Carlo runs is 500. For each Monte-Carlo run, bootstrap calibration is done for 500 times. For the hypothesis testing, the probability of nominal type I error is given as $\alpha = 0.05$. For each experiment, the sample size ranges respectively at 10, 15, 25, 50, and 200. Table 6, Table 8, Table 10 and Table 12 show the outcome of probability of type II error evaluated by chi-squared tabular quantile and bootstrap calibration quantile. Note that the standard errors of the probability of type II error are given in brackets. Figure 4.5.12, Figure 4.5.13, Figure 4.5.14 and Figure 4.5.15 present the outcome for powers of tests against the choice of these 5 alternative hypotheses.

Sample Size	Angular Distance	Tabular β	B.C. β
200	0.0524	0.076 (0.019)	0.072 (0.018)
	0.0397	0.344 (0.034)	0.338 (0.033)
	0.0271	0.624 (0.034)	0.628 (0.034)
	0.0144	0.886 (0.022)	0.874 (0.023)
	0.0017	0.936 (0.017)	0.948 (0.016)
50	0.0524	0.656 (0.067)	0.674 (0.066)
	0.0397	0.782 (0.058)	0.788 (0.058)
	0.0271	0.866 (0.045)	0.872 (0.047)
	0.0144	0.920 (0.038)	0.934 (0.035)
	0.0017	0.926 (0.037)	0.936 (0.035)
25	0.0524	0.768 (0.084)	0.828 (0.075)
	0.0397	0.840 (0.073)	0.884 (0.064)
	0.0271	0.900 (0.060)	0.926 (0.052)
	0.0144	0.916 (0.055)	0.956 (0.041)
	0.0017	0.912 (0.057)	0.946 (0.045)
15	0.0524	0.762 (0.110)	0.898 (0.078)
	0.0397	0.820 (0.099)	0.924 (0.068)
	0.0271	0.860 (0.090)	0.948 (0.057)
	0.0144	0.894 (0.079)	0.970 (0.044)
	0.0017	0.890 (0.081)	0.950 (0.056)
10	0.0524	0.716 (0.143)	0.948 (0.070)
	0.0397	0.790 (0.129)	0.968 (0.056)
	0.0271	0.808 (0.125)	0.968 (0.056)
	0.0144	0.834 (0.118)	0.984 (0.040)
	0.0017	0.832 (0.118)	0.962 (0.060)

Table 6: $Pr(\text{Type II error})$, with $\kappa = 25$.

Sample Size	Angular Distance	Tabular β	B.C. β
200	0.0524	0.292 (0.032)	0.302 (0.032)
	0.0397	0.498 (0.035)	0.508 (0.035)
	0.0271	0.800 (0.028)	0.800 (0.028)
	0.0144	0.894 (0.022)	0.894 (0.022)
	0.0017	0.950 (0.015)	0.944 (0.016)
50	0.0524	0.792 (0.057)	0.818 (0.055)
	0.0397	0.824 (0.054)	0.838 (0.052)
	0.0271	0.900 (0.042)	0.898 (0.043)
	0.0144	0.944 (0.033)	0.936 (0.035)
	0.0017	0.938 (0.034)	0.946 (0.032)
25	0.0524	0.832 (0.075)	0.872 (0.067)
	0.0397	0.864 (0.069)	0.900 (0.060)
	0.0271	0.888 (0.063)	0.918 (0.055)
	0.0144	0.922 (0.054)	0.950 (0.044)
	0.0017	0.908 (0.058)	0.934 (0.050)
15	0.0524	0.836 (0.096)	0.934 (0.064)
	0.0397	0.838 (0.095)	0.936 (0.063)
	0.0271	0.876 (0.085)	0.956 (0.053)
	0.0144	0.918 (0.071)	0.960 (0.051)
	0.0017	0.886 (0.082)	0.954 (0.054)
10	0.0524	0.746 (0.138)	0.954 (0.066)
	0.0397	0.806 (0.125)	0.962 (0.060)
	0.0271	0.828 (0.119)	0.974 (0.050)
	0.0144	0.848 (0.114)	0.976 (0.048)
	0.0017	0.808 (0.125)	0.982 (0.042)

Table 8: $Pr(\text{Type II error})$, with $\kappa = 15$.

Sample Size	Angular Distance	Tabular β	B.C. β
200	0.0524	0.742 (0.031)	0.740 (0.031)
	0.0397	0.866 (0.024)	0.850 (0.025)
	0.0271	0.892 (0.022)	0.888 (0.022)
	0.0144	0.950 (0.015)	0.946 (0.016)
	0.0017	0.934 (0.018)	0.926 (0.019)
50	0.0524	0.912 (0.040)	0.914 (0.040)
	0.0397	0.924 (0.037)	0.924 (0.037)
	0.0271	0.940 (0.034)	0.948 (0.031)
	0.0144	0.930 (0.036)	0.932 (0.036)
	0.0017	0.934 (0.035)	0.938 (0.034)
25	0.0524	0.898 (0.061)	0.924 (0.053)
	0.0397	0.914 (0.056)	0.940 (0.047)
	0.0271	0.912 (0.057)	0.946 (0.045)
	0.0144	0.952 (0.043)	0.974 (0.032)
	0.0017	0.922 (0.054)	0.950 (0.044)
15	0.0524	0.876 (0.085)	0.950 (0.056)
	0.0397	0.902 (0.077)	0.960 (0.051)
	0.0271	0.882 (0.083)	0.946 (0.058)
	0.0144	0.874 (0.086)	0.946 (0.058)
	0.0017	0.890 (0.081)	0.946 (0.058)
10	0.0524	0.820 (0.121)	0.972 (0.052)
	0.0397	0.792 (0.128)	0.966 (0.057)
	0.0271	0.832 (0.118)	0.980 (0.044)
	0.0144	0.834 (0.118)	0.974 (0.050)
	0.0017	0.864 (0.108)	0.984 (0.040)

Table 10: $Pr(\text{Type II error})$, with $\kappa = 5$.

Sample Size	Angular Distance	Tabular β	B.C. β
200	0.0524	0.946 (0.016)	0.948 (0.016)
	0.0397	0.948 (0.016)	0.944 (0.016)
	0.0271	0.942 (0.017)	0.944 (0.016)
	0.0144	0.952 (0.015)	0.946 (0.016)
	0.0017	0.938 (0.017)	0.932 (0.018)
50	0.0524	0.948 (0.031)	0.946 (0.032)
	0.0397	0.948 (0.031)	0.954 (0.030)
	0.0271	0.948 (0.031)	0.946 (0.032)
	0.0144	0.938 (0.034)	0.940 (0.034)
	0.0017	0.950 (0.031)	0.950 (0.031)
25	0.0524	0.922 (0.054)	0.940 (0.047)
	0.0397	0.948 (0.044)	0.958 (0.040)
	0.0271	0.928 (0.052)	0.940 (0.047)
	0.0144	0.922 (0.054)	0.936 (0.049)
	0.0017	0.934 (0.050)	0.948 (0.044)
15	0.0524	0.914 (0.072)	0.944 (0.059)
	0.0397	0.906 (0.075)	0.948 (0.057)
	0.0271	0.900 (0.077)	0.946 (0.058)
	0.0144	0.928 (0.067)	0.964 (0.048)
	0.0017	0.918 (0.071)	0.962 (0.049)
10	0.0524	0.828 (0.119)	0.964 (0.059)
	0.0397	0.852 (0.112)	0.982 (0.042)
	0.0271	0.866 (0.108)	0.982 (0.042)
	0.0144	0.866 (0.108)	0.968 (0.056)
	0.0017	0.870 (0.106)	0.982 (0.042)

Table 12: $Pr(\text{Type II error})$, with $\kappa = 1$.

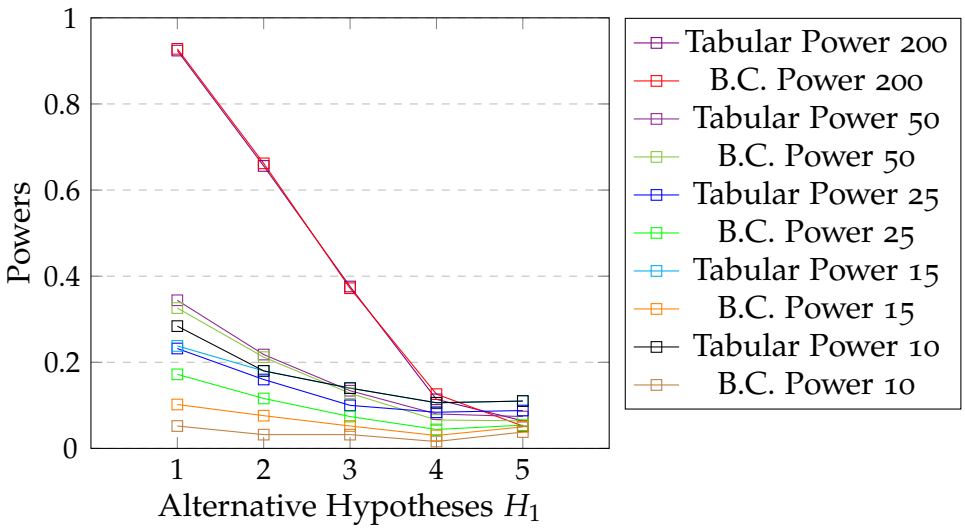


Figure 4.5.12: Tabular Powers and B.C. (Bootstrap Calibration) Powers of the test against alternative hypotheses under different sample sizes. The parameter κ of the distribution is 25 in this experiment.

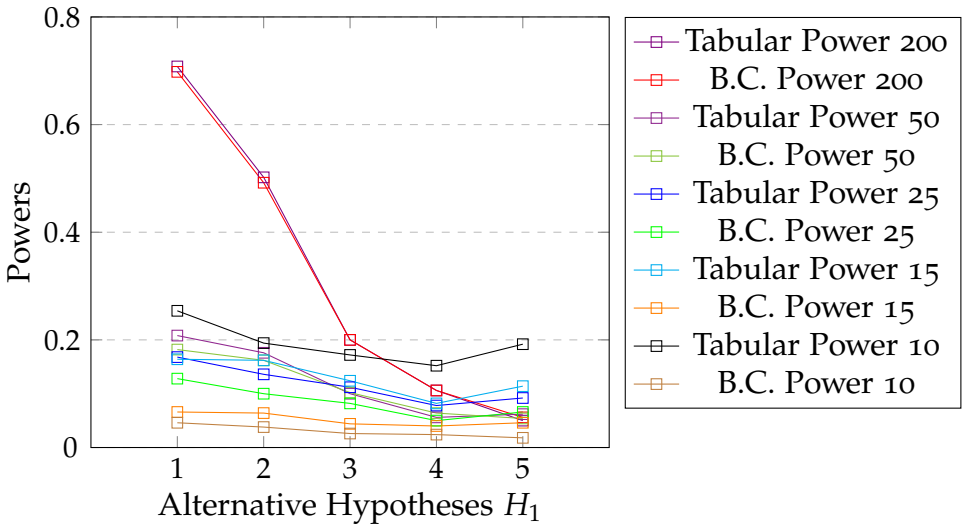


Figure 4.5.13: Tabular Powers and B.C. (Bootstrap Calibration) Powers of the test against alternative hypotheses under different sample sizes. The parameter κ of the distribution is 15 in this experiment.

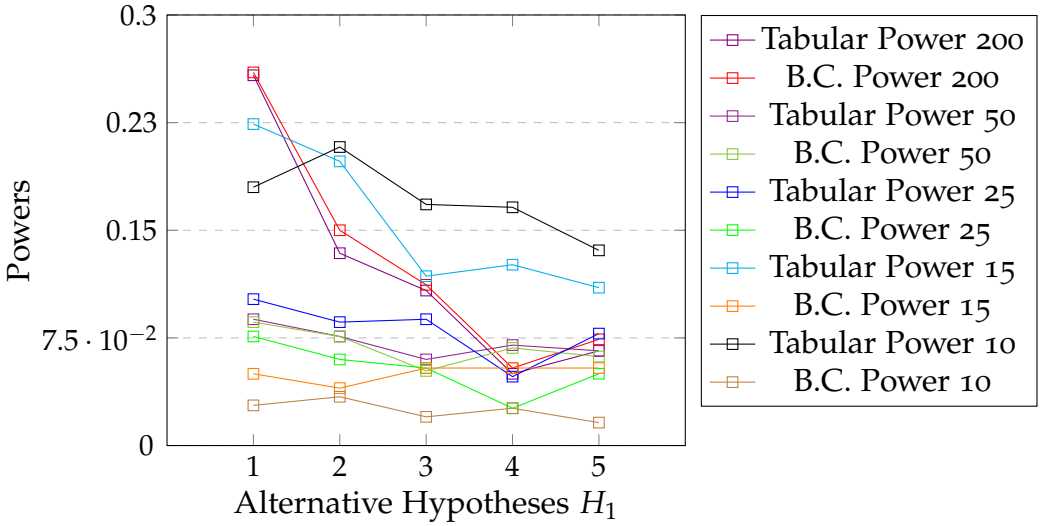


Figure 4.5.14: Tabular Powers and B.C. (Bootstrap Calibration) Powers of the test against alternative hypotheses under different sample sizes. The parameter κ of the distribution is 5 in this experiment.

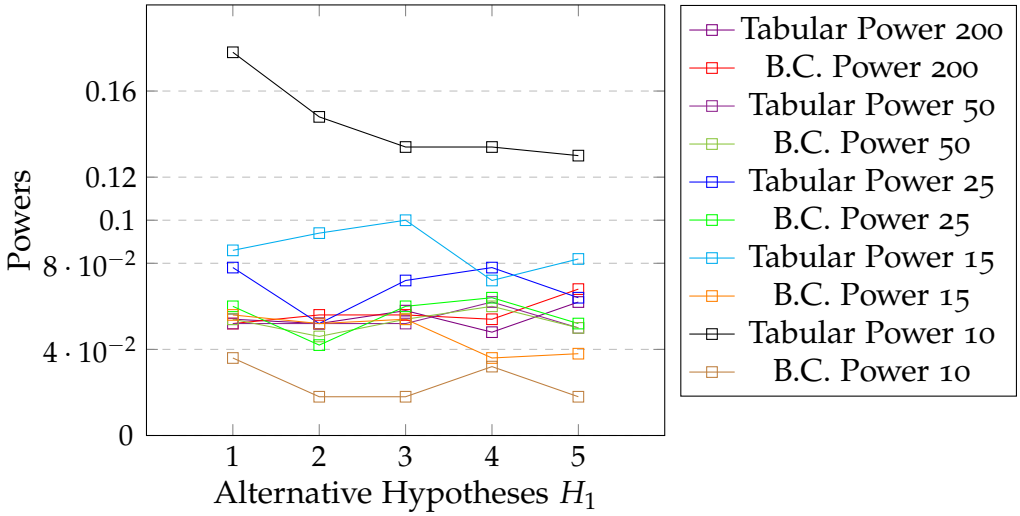


Figure 4.5.15: Tabular Powers and B.C. (Bootstrap Calibration) Powers of the test against alternative hypotheses under different sample sizes. The parameter κ of the distribution is 1 in this experiment.

Observing Tables, Figures above, we make following comments:

1. If the actual probability of type I error is too large, the power of the test loses the accuracy. For example, Figure 4.5.10 shows that probabilities of type I error calculated by the tabular quantile at small sample sizes are too large, so in Figure 4.5.12 and Table 6, the power of type II error becomes larger at sample size 10 compared with other powers at larger sample sizes.

2. Under the fixed sample size, powers (both by the tabular quantile and the bootstrap calibration quantile) have the tendency from a high level down to a lower level with the alternative hypothesis approaching the null hypothesis. Since $Power = Pr(\text{Reject } H_0|H_1)$, when the alternative hypothesis is leaving the null hypothesis, it is more likely to reject the null hypothesis.

3. For von Mises Fisher distribution on the sphere, the parameter κ determines how concentrated the sample is. The smaller κ is, the closer the distribution is to a uniform distribution, which has the mean direction everywhere on $\mathcal{S}^2$. For the sample generated from a population close to a uniform distribution, the probability of type II Error is large no matter what H_1 is and no matter what sample size it is. That explains why the powers in Figure 4.5.14 and Figure 4.5.15 are small compared with Figure 4.5.12 and Figure 4.5.13.

Kent Distribution

Now, we generate samples from Kent distribution with the concentration parameter $\kappa = 25, 15, 5, 1$ respectively, the ellipticity parameter $\beta = \kappa/4$ and the mean direction is $(1, 0, 0)$. The alternative hypotheses are the same as the choices in the experiments with von-Mises Fisher distribution. With Monte-Carlo simulation 500 times, bootstrap calibration 200 times, and nominal probability of type I error 0.05, the probability of type II error and the power of the test are shown in Table 14, Table 16, Table 18, Table 20 and Figure

4.5.16, Figure 4.5.17, Figure 4.5.18, Figure 4.5.19. Observing these outcome, we have the same conclusion as that in experiments with von-Mises Fisher distribution:

1. Probabilities of type I error calculated by the tabular quantile are too large, when the sample size is too small, say less than 50. Therefore, the test loses its accuracy, when the sample size is too small.

2. The power of the test increases, when the alternative hypothesis is away from the null hypothesis.

3. When the concentration parameter κ is approaching 0, the distribution is approaching a uniform distribution, and the probability of type II error maintains at a high level no matter of the choice of alternative hypothesis and the sample size.

Sample Size	Angular Distance	Tabular β	B.C. β
200	0.0524	0.084 (0.020)	0.084 (0.020)
	0.0397	0.326 (0.033)	0.328 (0.033)
	0.0271	0.630 (0.034)	0.638 (0.034)
	0.0144	0.874 (0.023)	0.874 (0.023)
	0.0017	0.948 (0.016)	0.942 (0.017)
50	0.0524	0.610 (0.069)	0.650 (0.067)
	0.0397	0.782 (0.058)	0.804 (0.056)
	0.0271	0.900 (0.042)	0.908 (0.041)
	0.0144	0.924 (0.037)	0.932 (0.036)
	0.0017	0.946 (0.032)	0.948 (0.031)
25	0.0524	0.792 (0.081)	0.850 (0.071)
	0.0397	0.852 (0.071)	0.908 (0.058)
	0.0271	0.874 (0.066)	0.910 (0.057)
	0.0144	0.922 (0.054)	0.948 (0.044)
	0.0017	0.912 (0.057)	0.940 (0.047)
15	0.0524	0.776 (0.107)	0.906 (0.075)
	0.0397	0.848 (0.093)	0.930 (0.066)
	0.0271	0.872 (0.086)	0.956 (0.053)
	0.0144	0.866 (0.088)	0.952 (0.055)
	0.0017	0.892 (0.080)	0.964 (0.048)
10	0.0524	0.766 (0.134)	0.954 (0.066)
	0.0397	0.760 (0.135)	0.962 (0.060)
	0.0271	0.802 (0.126)	0.950 (0.069)
	0.0144	0.812 (0.124)	0.968 (0.056)
	0.0017	0.820 (0.121)	0.976 (0.048)

Table 14: $Pr(\text{Type II error})$, with a Kent distribution of $\kappa = 25$ and $\beta = 25/4$.

Sample Size	Angular Distance	Tabular β	B.C. β
200	0.0524	0.308 (0.033)	0.308 (0.033)
	0.0397	0.576 (0.035)	0.576 (0.035)
	0.0271	0.790 (0.029)	0.778 (0.029)
	0.0144	0.906 (0.021)	0.912 (0.020)
	0.0017	0.956 (0.015)	0.956 (0.015)
50	0.0524	0.782 (0.058)	0.808 (0.056)
	0.0397	0.854 (0.050)	0.868 (0.048)
	0.0271	0.880 (0.046)	0.884 (0.045)
	0.0144	0.936 (0.035)	0.938 (0.034)
	0.0017	0.946 (0.032)	0.944 (0.033)
25	0.0524	0.814 (0.078)	0.864 (0.069)
	0.0397	0.878 (0.066)	0.906 (0.058)
	0.0271	0.892 (0.062)	0.920 (0.054)
	0.0144	0.912 (0.057)	0.934 (0.050)
	0.0017	0.924 (0.053)	0.940 (0.048)
15	0.0524	0.832 (0.097)	0.920 (0.070)
	0.0397	0.842 (0.094)	0.936 (0.063)
	0.0271	0.840 (0.095)	0.930 (0.066)
	0.0144	0.880 (0.084)	0.950 (0.056)
	0.0017	0.886 (0.082)	0.946 (0.058)
10	0.0524	0.754 (0.136)	0.950 (0.069)
	0.0397	0.778 (0.131)	0.970 (0.054)
	0.0271	0.810 (0.124)	0.972 (0.052)
	0.0144	0.798 (0.127)	0.976 (0.048)
	0.0017	0.816 (0.123)	0.966 (0.057)

Table 16: $Pr(\text{Type II error})$, with a Kent distribution of $\kappa = 15$ and $\beta = 15/4$.

Sample Size	Angular Distance	Tabular β	B.C. β
200	0.0524	0.756 (0.030)	0.744 (0.031)
	0.0397	0.836 (0.026)	0.840 (0.026)
	0.0271	0.906 (0.021)	0.902 (0.021)
	0.0144	0.940 (0.017)	0.944 (0.016)
	0.0017	0.954 (0.015)	0.948 (0.016)
50	0.0524	0.922 (0.038)	0.924 (0.037)
	0.0397	0.924 (0.037)	0.928 (0.037)
	0.0271	0.934 (0.035)	0.932 (0.036)
	0.0144	0.920 (0.038)	0.916 (0.039)
	0.0017	0.934 (0.035)	0.938 (0.034)
25	0.0524	0.888 (0.063)	0.906 (0.058)
	0.0397	0.904 (0.059)	0.924 (0.053)
	0.0271	0.938 (0.048)	0.944 (0.046)
	0.0144	0.912 (0.057)	0.928 (0.052)
	0.0017	0.926 (0.052)	0.940 (0.047)
15	0.0524	0.884 (0.083)	0.962 (0.049)
	0.0397	0.860 (0.090)	0.944 (0.059)
	0.0271	0.896 (0.086)	0.956 (0.053)
	0.0144	0.888 (0.079)	0.954 (0.054)
	0.0017	0.862 (0.089)	0.924 (0.068)
10	0.0524	0.826 (0.120)	0.980 (0.044)
	0.0397	0.826 (0.120)	0.958 (0.063)
	0.0271	0.818 (0.122)	0.958 (0.063)
	0.0144	0.838 (0.117)	0.974 (0.050)
	0.0017	0.836 (0.117)	0.968 (0.057)

Table 18: Pr(Type II error), with a Kent distribution of $\kappa = 5$ and $\beta = 5/4$.

Sample Size	Angular Distance	Tabular β	B.C. β
200	0.0524	0.920 (0.019)	0.906 (0.021)
	0.0397	0.932 (0.018)	0.926 (0.019)
	0.0271	0.952 (0.015)	0.948 (0.016)
	0.0144	0.946 (0.016)	0.934 (0.018)
	0.0017	0.958 (0.014)	0.954 (0.015)
50	0.0524	0.932 (0.036)	0.938 (0.034)
	0.0397	0.946 (0.032)	0.940 (0.034)
	0.0271	0.948 (0.031)	0.948 (0.031)
	0.0144	0.954 (0.030)	0.962 (0.027)
	0.0017	0.922 (0.038)	0.926 (0.037)
25	0.0524	0.940 (0.047)	0.948 (0.044)
	0.0397	0.950 (0.044)	0.952 (0.043)
	0.0271	0.930 (0.051)	0.952 (0.043)
	0.0144	0.930 (0.051)	0.942 (0.047)
	0.0017	0.936 (0.049)	0.940 (0.047)
15	0.0524	0.920 (0.070)	0.960 (0.051)
	0.0397	0.892 (0.080)	0.940 (0.061)
	0.0271	0.900 (0.077)	0.948 (0.057)
	0.0144	0.906 (0.075)	0.952 (0.055)
	0.0017	0.904 (0.076)	0.948 (0.057)
10	0.0524	0.854 (0.112)	0.966 (0.057)
	0.0397	0.836 (0.117)	0.960 (0.062)
	0.0271	0.860 (0.110)	0.978 (0.046)
	0.0144	0.848 (0.114)	0.962 (0.060)
	0.0017	0.842 (0.115)	0.962 (0.060)

Table 20: Pr(Type II error), with a Kent distribution of $\kappa = 1$ and $\beta = 1/4$.

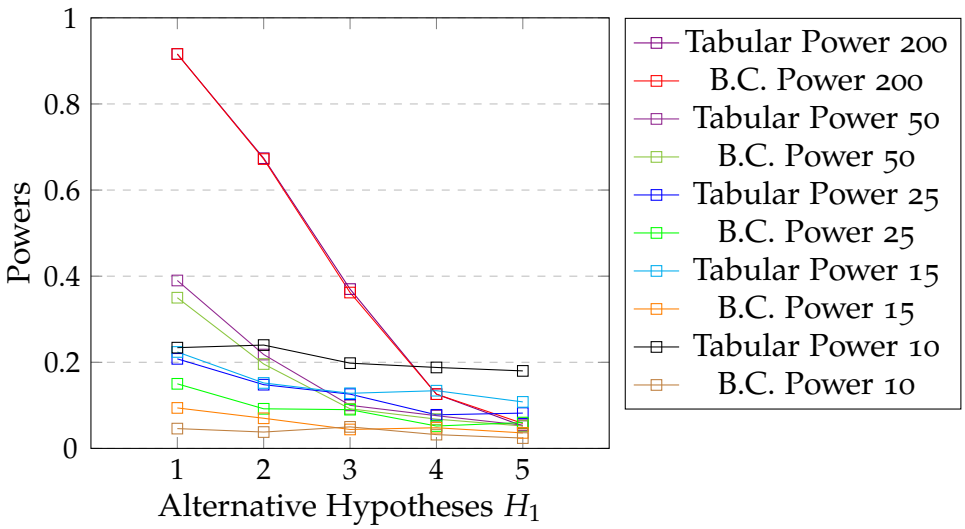


Figure 4.5.16: Tabular Powers and B.C. (Bootstrap Calibration) Powers of the test against alternative hypotheses under different sample sizes. The concentration parameter of the Kent distribution κ is 25 and the ellipticity parameter of the Kent distribution β is $25/4$.

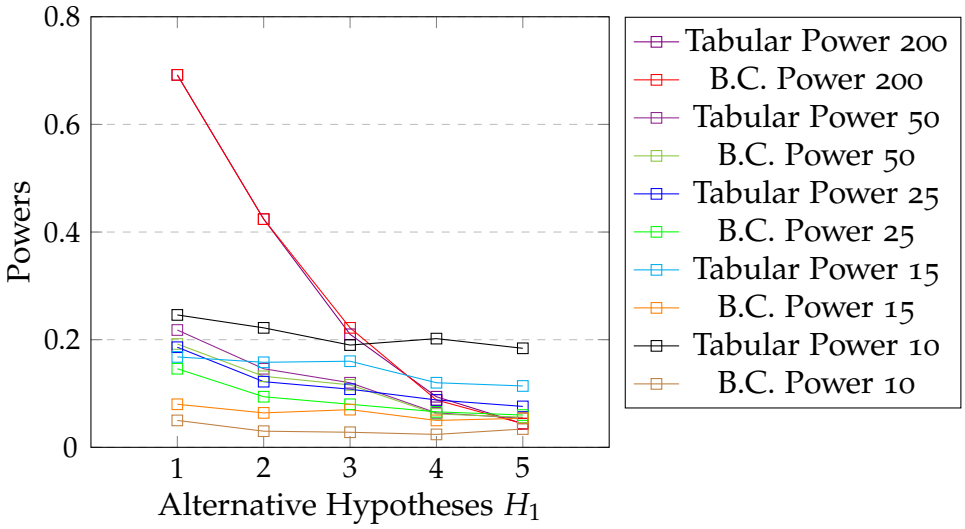


Figure 4.5.17: Tabular Powers and B.C. (Bootstrap Calibration) Powers of the test against alternative hypotheses under different sample sizes. The concentration parameter of the Kent distribution κ is 15 and the ellipticity parameter of the Kent distribution β is $15/4$.

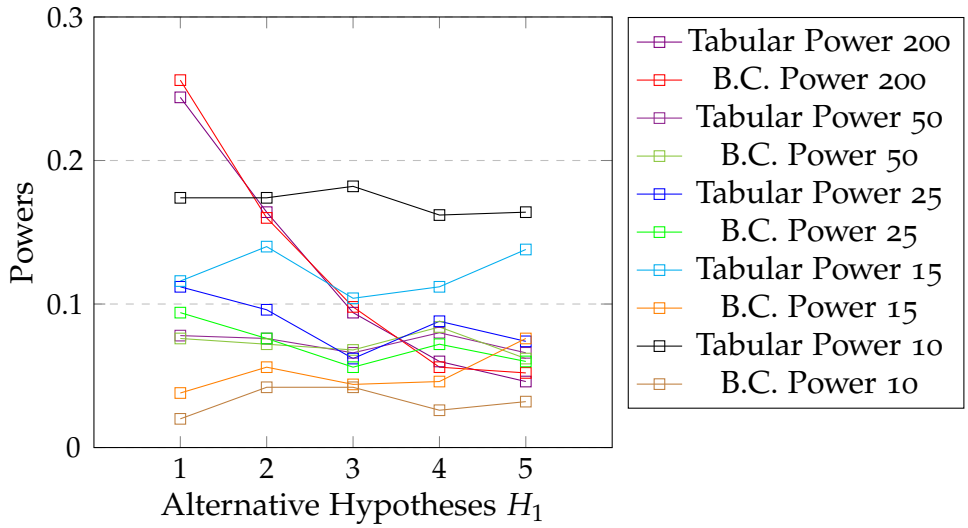


Figure 4.5.18: Tabular Powers and B.C. (Bootstrap Calibration) Powers of the test against alternative hypotheses under different sample sizes. The concentration parameter of the Kent distribution κ is 5 and the ellipticity parameter of the Kent distribution β is $5/4$.

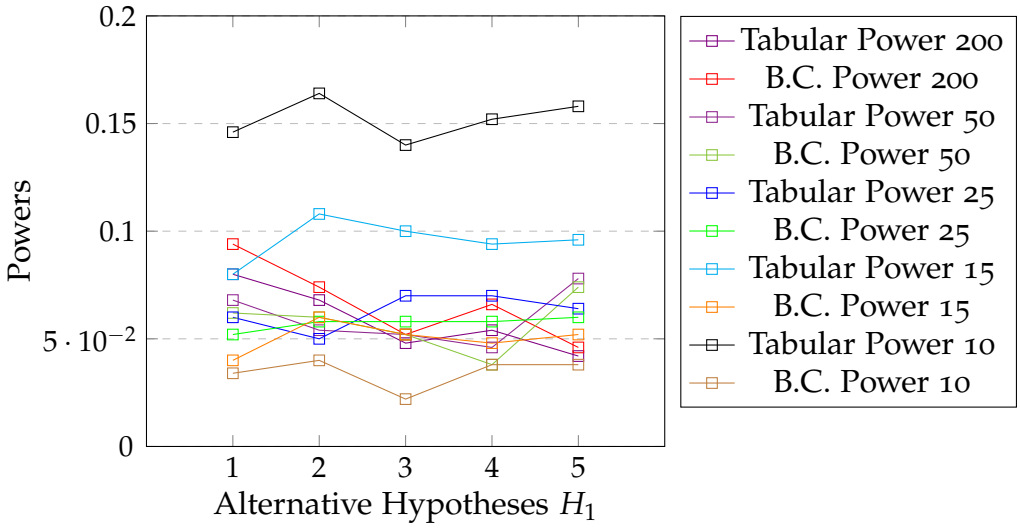


Figure 4.5.19: Tabular Powers and B.C. (Bootstrap Calibration) Powers of the test against alternative hypotheses under different sample sizes. The concentration parameter of the Kent distribution κ is 1 and the ellipticity parameter of the Kent distribution β is $1/4$.

4.6 CONFIDENCE REGION CONSTRUCTION

In hypothesis testing, the confidence region is calculated to show the area for which a null hypothesis is accepted. Now, we use Algorithm 1 to present a confidence region on the sphere $\mathcal{S}^2$, with confidence level $1-\alpha$ given. Here we use the chi-squared tabular $1-\alpha$ quantile and the bootstrap calibration $1-\alpha$ quantile; see Algorithm 2.

Example 4.6.1 *We generate a sample from the Von-Mises Fisher distribution with parameter* $\kappa = 25$*, sample size 500, mean direction* $(1,0,0)$*. Let the nominal probability of type I error* $\alpha = 0.05$*. According to Algorithm 1, we calculate the coordinates of the contour of the confidence region; see Figure 4.6.20. Note that the red boundary is the contour of the confidence region, calculated through the chi-squared tabular* $1-\alpha$ *quantile, while the blue boundary is the contour of the confidence region, calculated through the bootstrap calibration* $1-\alpha$ *quantile. The green point is the mean direction* $(1,0,0)$ *of the population. The bootstrap calibration is based on* 500 *Monte-Carlo runs.*

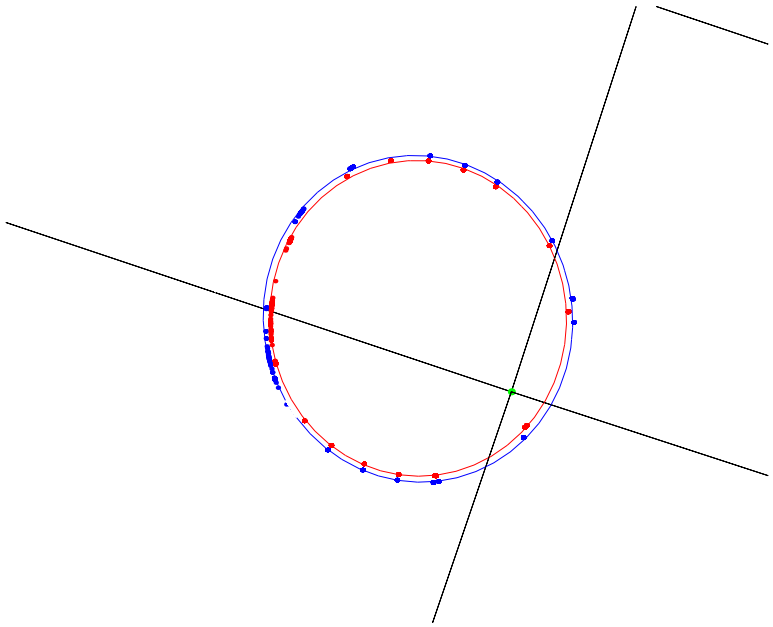


Figure 4.6.20: The contour of 95% confidence region.

Furthermore, we apply the empirical likelihood method to a real dataset and compare with the confidence region constructed using an existing method. Creer et al. (1959) measured the direction of magnetisation in 34 sites at the Great Whin Sill. Kent (1982) used the estimates of parameters in FB_5 (5-parameter Fisher-Bingham) distribution to construct the confidence region. In the example, he excluded the sites 32 and 34 in the dataset provided by Creer et al. (1959), and kept the remaining 32 sites as the sample. Now, we compare the 95%-confidence region constructed by the empirical likelihood method with the method by Kent (1982). It is noted that chi-squared 0.95 quantile and bootstrap calibration 0.95 quantile are respectively used in the empirical likelihood ratio method to construct confidence regions.

Example 4.6.2 *The magnetisation direction dataset are summarised in Fisher et al. (1993, Appendix B9). We use*

$$\theta = 90^\circ + I,$$
$$\phi = 360^\circ - D.$$

to transform the declination (D) and the inclination (I) to the polar coordinate (θ, φ)*. It is noted that the Cartesian coordinates are given by*

$$(\cos\theta, \sin\theta\cos\varphi, \sin\theta\sin\varphi).$$

The confidence regions are calculated respectively by the methods mentioned above; see Figure 4.6.21. We observe that contours (red and green) of confidence regions calculated by empirical likelihood are close to each other. The contour (blue) of confidence region shown by Kent (1982) still looks similar to other two contours, although it looks slightly away from other two contours at 50%*-level in the zoomed-in figure. In Figure 4.6.21, the black points are the sample; The blue ellipses are contours of confidence regions by Kent (1982); The red ellipses are contours of confidence regions, calculated through the chi-squared tabular quantile, while the green ellipses are contours of confidence regions, calculated through the bootstrap calibration quantile. Note that the bootstrap calibration is based on* 500 *Monte-Carlo runs.*

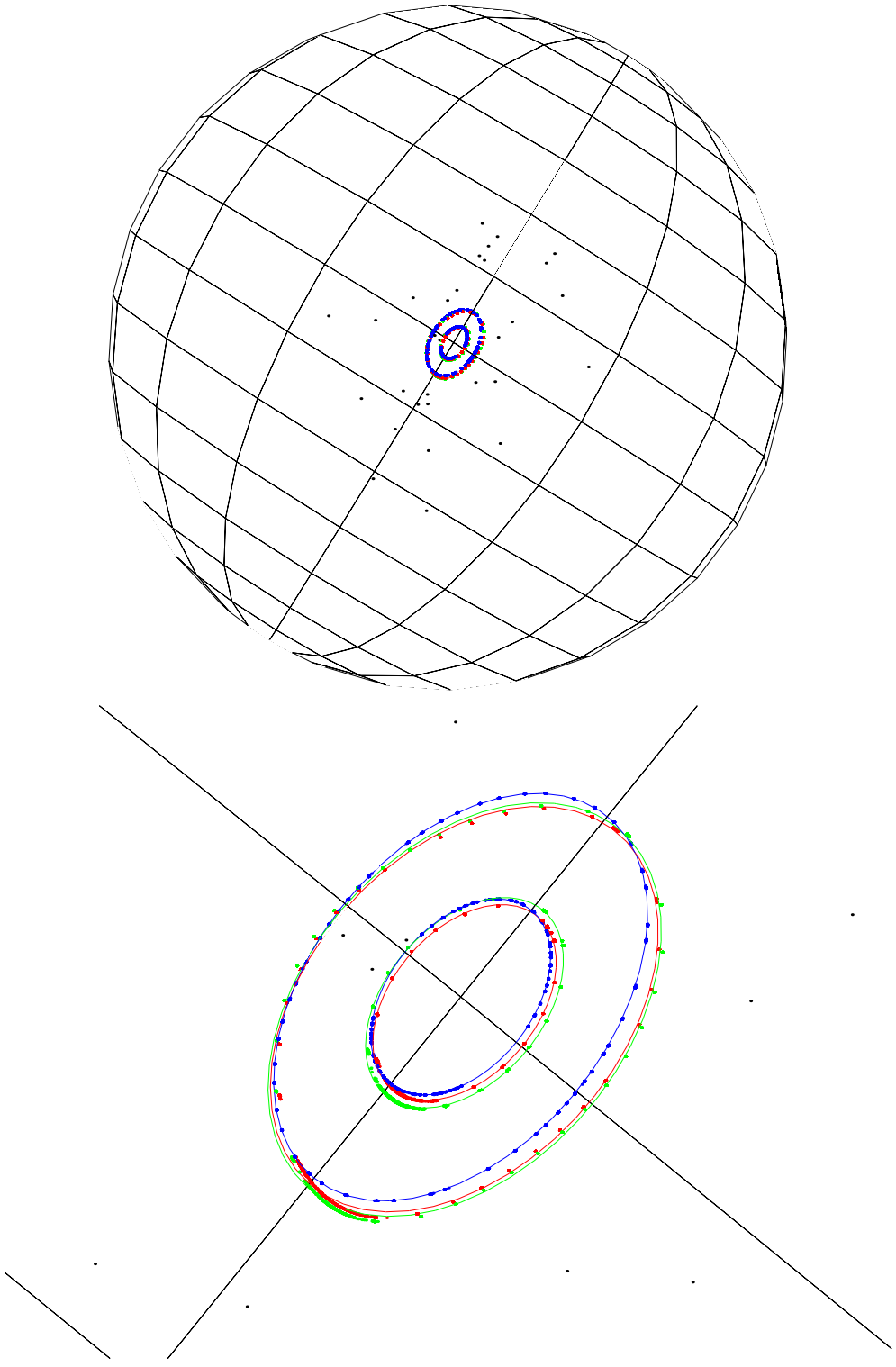


Figure 4.6.21: Contours of 95% and 50% confidence regions for the data of the direction of magnetisation at the Great Whin Sill.

Note that the method by Kent (1982) depends on parametric assumption while the empirical likelihood regions are defined non-parametrically. It is reassuring that they give such similar results in this example though we would not expect this to necessarily be the case in other examples, if the parametric assumptions do not hold.

4.7 SUMMARY

This chapter mainly focuses on extending empirical likelihood for a mean in Euclidean space $\mathbb{R}^k$ to a mean direction on the sphere $\mathcal{S}^{k-1}$.

First, we define empirical likelihood and local alternatives $\{\boldsymbol{\mu}_n\}_{n\geqslant 1}$ on $\mathcal{S}^{k-1}$. In order to prove an analogue of Theorem 3.1 to this case, we introduce a rotation matrix as in (4.2.6) which is a function of the mean direction $\boldsymbol{\mu}$. Then, we give the proof of Wilks' theorem on $\mathcal{S}^{k-1}$ considering local alternatives.

Based on Theorem 4.1, we give the algorithm to evaluate the log-empirical likelihood against the mean direction parameter on $\mathcal{S}^{k-1}$, and then construct the confidence region for the mean direction. The numerical work also includes the study of operating characteristics of the empirical likelihood procedures, such as estimating actual probabilities of type I error and type II error. Both chi-squared tabular quantiles and bootstrap calibration quantiles are considered. We find the probability of type I error evaluated under the bootstrap calibration quantile is more satisfactory than that by the tabular quantile. As for the probability of type II error, the outcome is much influenced by the probability of type I error when the sample size is small.

5

EMPIRICAL LIKELIHOOD ON 3-SPIDER

5.1 INTRODUCTION

The aim of this chapter is to explore how Wilks' theorem for empirical likelihood in a Euclidean setting can be extended to the 3-Spider. The 3-Spider, denoted by $\mathcal{SP}_3$, a non-Euclidean space, consists of three semi-infinite line segments which meet at a point; see Section 2.6. As we shall see, empirical likelihood applied to the 3-Spider exhibits some interesting and non-standard features. Thus, Wilks' theorem for empirical likelihood defined on the 3-Spider needs more consideration to understand those non-standard features. Wilks' theorem for empirical likelihood is an important implement to construct confidence regions. The 3-Spider is useful as a model for a certain type of phylogenetic tree data, as will be explained later.

The sections are organised as follows. How to define the empirical likelihood on the 3-Spider is considered first in Section 5.2, in which the folding map introduced by Hotz et al. (2013) is applied to realise the connection between empirical likelihood on the 3-Spider and empirical likelihood on Euclidean space. Wilks' theorem for empirical likelihood on the 3-Spider is given in Section 5.3, as a main part of this chapter. As we shall see, empirical likelihood behaves in a non-standard way when the population Fréchet mean lies on the cone point, the point where the three lines meet. Then, the proof of Wilks' theorem on the 3-Spider is provided in Section 5.5, making use of some technical lemmas in Section 5.4. In Section 5.6, we give the

algorithm for calculating the empirical likelihood ratio against the mean parameter. Section 5.7 presents three cases simulation of empirical likelihood ratio against the mean parameter, based on locations of the Fréchet mean. The numerical work for constructing the confidence regions is presented in Section 5.8, where the dataset used in Section 5.6 to construct the confidence region is the phylogenetic tree dataset used in the work by Nye (2011). The operating characteristics of hypothesis testing are presented in Section 5.9, including investigating the probability of type I error and type II error by tabular quantiles and bootstrap calibration quantiles.

5.2 EMPIRICAL LIKELIHOOD ON THE 3-SPIDER

As is shown in previous chapters, we first consider how to define empirical likelihood in $\mathcal{SP}_3$. Recall

$$\begin{aligned}\mathcal{SP}_3 =& L_0 \cup (\cup_{i=1}^3 L_i)\\ =& L_0 \cup (\cup_{i=1}^3 L_i^+),\end{aligned}$$

where $L_i = \mathbb{R}_{\geqslant 0}$, $L_i^+ = \mathbb{R}_{>0}$, $i = 1, 2, 3$, and $L_0 = \{0\}$, the cone point.

For a random sample $q_1, q_2, \cdots, q_n$ on $\mathcal{SP}_3$ with common distribution P, we define

$$P(A) = \sum_{i=1}^{n} p_i \delta_{q_i}(A), \tag{5.2.1}$$

where A is an arbitrary Borel-measurable subset in $\mathcal{SP}_3$, $p_i \geqslant 0$ with $\sum_{i=1}^n p_i = 1$, and

$$\delta_{q_i}(A) = \begin{cases} 1, & \text{if } q_i \in A; \\ 0, & \text{otherwise.} \end{cases}$$

Now, we define empirical likelihood on $\mathcal{SP}_3$.

Let b be the Fréchet mean parameter, which can be either off the cone point or on the cone point. The empirical likelihood ratio is a function of b. Let $d_E(\cdot,\cdot)$ denote the Euclidean distance, $d_{\mathcal{SP}_3}(\cdot,\cdot)$ denote the distance on $\mathcal{SP}_3$.

Example 5.2.1 *Suppose $q_1 \in L_1^+$ and $q_2 \in L_2^+$. For a chosen folding map F_j,*

$$d_{\mathcal{SP}_3}(q_1, q_2) \begin{cases} = d_E(F_k q_1, F_k q_2), \ k = 1, 2; \\ \neq d_E(F_k q_1, F_k q_2), \ k = 3. \end{cases}$$

That is, in order to calculate $d_{\mathcal{SP}_3}(\cdot, \cdot)$ by $d_E(\cdot, \cdot)$, the folding map needs to be well chosen. However, if b_0 is the cone point, then

$$d_{\mathcal{SP}_3}(b_0, q) = d_E(F_k b_0, F_k q) = d_E(0, F_k q),$$

for $k = 1, 2, 3$.

5.2.1 *Empirical Likelihood Ratio when b is off the Cone Point*

Let $q = (q_0, k) \in \mathcal{SP}_3$ with index $k \in \{1, 2, 3\}$. Then, for $j \in \{1, 2, 3\}$, we recall the folding map $F_j : \mathcal{SP}_3 \to \mathbb{R}$ is defined as

$$F_j q = \begin{cases} q_0 & \text{if } k = j, \\ -q_0 & \text{if } k \neq j. \end{cases} \tag{5.2.2}$$

In particular, let b_0 denote the cone point. Then, for folding maps F_k, $k = 1, 2, 3$, we have $F_k b_0 = 0$.

Let $q_{\max}^{(j)} = \max_{1 \leqslant i \leqslant n} F_j q_i$, $j = 1, 2, 3$. If b is off the cone point, then

$$b \in \cup_{j=1}^{3} \{L_j^+ \cap (0, q_{\max}^{(j)})\},$$

determines a set $\{p_i\}_{i=1}^n$, such that $\sum_{i=1}^n p_i = 1$ and $0 < p_i < 1$, $i = 1, 2, \cdots, n$.

Assume $b \in L_j^+$, $j \in \{1, 2, 3\}$. There correspondingly exists a set of $\{p_i\}_{i=1}^n$ for the images of $q_1, q_2, \cdots, q_n$ through the folding map F_j such that

$$\begin{aligned} &\sum_{i=1}^{n} p_i F_j q_i = F_j b, \\ &\sum_{i=1}^{n} p_i = 1, \\ &0 \leqslant p_i \leqslant 1, \ i = 1, 2, \cdots, n. \end{aligned} \tag{5.2.3}$$

Example 5.2.2 *Assume $b \in L_1^+$. We suppose there exists a set of $\{p_i\}_{i=1}^n$ such that*

$$\sum_{i=1}^{n} p_i F_1 q_i = F_1 b,$$
$$\sum_{i=1}^{n} p_i = 1,$$
$$0 \leqslant p_i \leqslant 1,\ i = 1, 2, \cdots, n.$$

If we use F_2, all the sample on L_1 and L_3 are mapped to $\mathbb{R}_{\leqslant 0}$, while the sample on L_2 are mapped to $\mathbb{R}_{\geqslant 0}$. Then, for the same set $\{p_i\}_{i=1}^n$,

$$\sum_{i=1}^{n} p_i F_2 q_i \neq F_2 b.$$

except possibly in exceptional circumstances.

Hence, the empirical likelihood ratio is defined as:

$$\mathcal{R}(b) = \max_{p_1, \cdots, p_n} \Big\{ \prod_{i=1}^{n} n p_i | p_i \geqslant 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i F_j q_i = F_j b, \text{for } b \in L_j^+ \Big\}. \tag{5.2.4}$$

Then, defining $\mathcal{R}(b)$ in $\mathcal{SP}_3$ is equivalent to defining $\mathcal{R}(F_j b)$ in $\mathbb{R}$.

5.2.2 *Empirical Likelihood Ratio when b is on the Cone Point*

We now introduce the following lemma containing a necessary and sufficient condition for the Fréchet mean to be on the cone point, which is useful for calculating empirical likelihood at the cone point.

Lemma 5.2.1 (Fréchet Mean on the Cone Point) *Let let q be a random variable on $\mathcal{SP}_3$ with distribution P. The Fréchet mean is on the cone point if and only if*

$$\int_{\mathcal{SP}_3} F_k q dP(q) \leqslant 0, \tag{5.2.5}$$

for all $k = 1, 2, 3$, where the folding map F_k is defined as in Definition 2.6.2.

PROOF. Let b_0 denote the cone point of $\mathcal{SP}_3$. Suppose that the Fréchet mean of P is at the cone point, i.e.

$$b_0 = \underset{b\in\mathcal{SP}_3}{\operatorname{argmin}} \int_{\mathcal{SP}_3} d^2_{\mathcal{SP}_3}(b,q)dP(q).$$

Then, for any arbitrary point $b_1 \in \mathcal{SP}_3$ and $b_1 \neq b_0$,

$$\int_{\mathcal{SP}_3} d^2_{\mathcal{SP}_3}(b_0,q)dP(q) \leqslant \int_{\mathcal{SP}_3} d^2_{\mathcal{SP}_3}(b_1,q)dP(q). \tag{5.2.6}$$

For folding maps F_k, $k = 1,2,3$, $F_k b_0 = 0$. We use folding maps F_k, $k = 1,2,3$, to rewrite (5.2.6).

1. If $b_1 \in L_j^+$, $j \in \{1,2,3\}$, then we recall Example 5.2.1 and rewrite the inequality (5.2.6) as

$$\begin{aligned}\int_{\mathcal{SP}_3} d^2_E(F_j b_0, F_j q)dP(q) &= \int_{\mathcal{SP}_3} d^2_E(0, F_j q)dP(q)\\ &\leqslant \int_{\mathcal{SP}_3} d^2_E(F_j b_1, F_j q)dP(q).\end{aligned}$$

2. If $b_1 \notin L_j^+$, say $b_1 \in L_l^+$, $l \in \{1,2,3\}/\{j\}$, then we recall Example 5.2.1 again and rewrite the inequality (5.2.6) as

$$\begin{aligned}\int_{\mathcal{SP}_3} d^2_E(F_j b_0, F_j q)dP(q) &= \int_{\mathcal{SP}_3} d^2_E(F_l b_0, F_l q)dP(q)\\ &= \int_{\mathcal{SP}_3} d^2_E(0, F_l q)dP(q)\\ &\leqslant \int_{\mathcal{SP}_3} d^2_E(F_l b_1, F_l q)dP(q).\end{aligned}$$

That is, for $k = 1,2,3$ and $b_1 \in L_k^+$,

$$\int_{\mathcal{SP}_3} d^2_E(0, F_k q)dP(q) \leqslant \int_{\mathcal{SP}_3} d^2_E(F_k b_1, F_k q)dP(q). \tag{5.2.7}$$

Then, we calculate the square $d^2_E(\cdot,\cdot)$, rearrange the inequality (5.2.7) and have

$$2F_k b_1 \int_{\mathcal{SP}_3} F_k q dP(q) - (F_k b_1)^2 \leqslant 0. \tag{5.2.8}$$

Now, since $F_k b_1 > 0$, we rewrite (5.2.8) as

$$F_k b_1 \Big(\int_{\mathcal{SP}_2} F_k q - \frac{F_k b_1}{2} dP(q) \Big) \leqslant 0. \tag{5.2.9}$$

Let b_1 approach the cone point. Inequality (5.2.9) gives

$$\int_{\mathcal{SP}_3} F_k q dP(q) \leqslant 0.$$

□

Now, consider the Stieltjes integral under step functions; see Section 2.8.2. Since the empirical distribution function is a step function, we apply the distribution in (5.2.1) to (5.2.6) and define the empirical likelihood ratio for b on the cone point:

$$\mathcal{R}(b) = \max_{p_1,\cdots,p_n} \Big\{ \prod_{i=1}^{n} np_i | p_i \geqslant 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i F_j q_i \leqslant 0, j = 1,2,3 \Big\}. \tag{5.2.10}$$

Now, we consider how to evaluate $\mathcal{R}(b)$ when b is on the cone point. There are two possible ways to evaluate $\mathcal{R}(b)$ at the cone point:

1. There exists $j \in \{1,2,3\}$, such that $\sum_{i=1}^{n} p_i F_j q_i = 0$, $\sum_{i=1}^{n} p_i F_k q_i < 0$, $k \neq j$.

2. For $k = 1,2,3$, $\sum_{i=1}^{n} p_i F_k q_i < 0$.

Now, we recall two definitions: the first population moment and the first sample moment after applying the folding maps F_k, $k = 1,2,3,$.

$$\begin{aligned} m_k &= \int_{\mathcal{SP}_3} (F_k q) dP(q), \\ \hat{m}_k &= \frac{1}{N} \sum_{i=1}^{N} F_k q_i. \end{aligned} \tag{5.2.11}$$

Then, we introduce an important lemma as follows, considering possibilities 1 and 2 above when $\mathcal{R}(b)$ is evaluated at the cone point.

Lemma 5.2.2 *Rewrite (5.2.10) as:*

$$\begin{aligned} & l = \max \sum_{i=1}^{n} \log np_i \\ \textit{subject to } & \sum_{i=1}^{n} p_i F_k q_i \leqslant 0, \ k = 1,2,3, \\ & \sum_{i=1}^{n} p_i = 1, \\ & p_i \geqslant 0, \ i = 1,\cdots,n. \end{aligned} \tag{5.2.12}$$

Rewrite

$$\hat{m}_j = \frac{1}{n}(\hat{S}_j - \sum_{i \neq j} \hat{S}_i), \tag{5.2.13}$$

where $\hat{S}_j = \sum_{i:q_i \in L_j} F_j q_i$.

Case 1. Suppose there exists $j \in \{1,2,3\}$, *such that* $\hat{m}_j \geqslant 0$ *and* $\hat{m}_k < 0$, *for* $k \neq j$. *Solving (5.2.12) under the folding map* F_j, *i.e.*

$$l_j(s) = \max \sum_{i=1}^{n} \log N p_i,$$

where j *and* s *are specified by the constraints:*

$$\begin{aligned} &\sum_{i=1}^{n} p_i F_j q_i = s, \\ &\sum_{i=1}^{n} p_i = 1, \\ &p_i \geqslant 0,\ i = 1, \cdots, n, \end{aligned} \tag{5.2.14}$$

for $s = 0$ *gives the solution to (5.2.10), i.e.* $l = l_j(0)$.

Case 2. If $\hat{m}_k < 0$ *for* $k = 1,2,3$, *the solution to (5.2.10) is*

$$p_1 = \cdots = p_n = \frac{1}{n},$$

i.e. $l = 0$.

PROOF.

Case 1. Here we assume that there exists $j \in \{1,2,3\}$, such that $\hat{m}_j \geqslant 0$ and $\hat{m}_k < 0$, for $k \neq j$. First, we know the sample mean of $F_j q_1, \cdots, F_j q_n$ is non-negative. Now consider two intermediate optimization problems:

$$\begin{aligned} &l_j = \max \sum_{i=1}^{n} \log n p_i, \\ \text{subject to } &\sum_{i=1}^{n} p_i F_j q_i \leqslant 0, \\ &\sum_{i=1}^{n} p_i = 1, \\ &p_i \geqslant 0,\ i = 1, \cdots, n, \end{aligned} \tag{5.2.15}$$

and

$$\begin{aligned} & l_j^* = \max \sum_{i=1}^{n} \log np_i, \\ \text{subject to } & \sum_{i=1}^{n} p_i F_j q_i = 0 \\ & \sum_{i=1}^{n} p_i F_k q_i \leqslant 0,\ k \neq j \\ & \sum_{i=1}^{n} p_i = 1, \\ & p_i \geqslant 0,\ i = 1, \cdots, n. \end{aligned} \tag{5.2.16}$$

Now, we use these new intermediate problems (5.2.15) and (5.2.16) to prove Case 1 of the lemma, following Step 1 and Step 2 below.

Step 1. Use l_j to prove $l \leqslant l_j(0)$. Firstly, the fact that the constraint of (5.2.12) is more restrictive than that of (5.2.15) implies

$$l \leqslant l_j.$$

Secondly, Theorem 3.2 implies that $l_j(s)$ is concave with respect to s. Then, $\hat{m}_j \geqslant 0$ implies that

$$\operatorname*{argmax}_{s \in \mathbb{R}} l_j(s) \geqslant 0.$$

Therefore,

$$\max_{s \leqslant 0} l_j(s) = l_j(0).$$

It follows that

$$l_j = l_j(0),$$

and then

$$l \leqslant l_j(0).$$

Step 2. Use l_j^* in (5.2.15) to prove $l \geqslant l_j(0)$. Firstly, since the constraint of (5.2.16) is more restrictive than that of (5.2.12), we have

$$l_j^* \leqslant l.$$

Although the constraint of (5.2.16) looks more restrictive than that of (5.2.14) at $s = 0$, they are essentially the same under any non-negative set of $\{p_i\}_{i=1}^n$, such that $\sum_{i=1}^n p_i = 1$. To see this, we define

$$S_k(p) = \sum_{i:q_i \in L_k} p_i F_k q_i \geqslant 0,\ k = 1, 2, .3$$

Without losing generality, we let $j = 1$. Then, as the constraint of (5.2.16) shows,

$$\sum_{i=1}^{n} p_i F_1 q_i = S_1(p) - S_2(p) - S_3(p) = 0.$$

Under the same $\{p_i\}_{i=1}^n$, changing the folding map to, say F_2, will give us

$$\sum_{i=1}^{n} p_i F_2 q_i = S_2(p) - S_1(p) - S_3(p) = -2S_3(p) \leqslant 0.$$

That is to say, for $k = 1, 2, 3$,

$$\sum_{i=1}^{n} p_i F_k q_i \leqslant 0.$$

Thus, the inequalities in the constraints of (5.2.16) are redundant. It follows that the optimization problem (5.2.14) at $s = 0$ and (5.2.16) are equivalent, i.e. $l_j(0) = l_j^*$. Then, we have

$$l \geqslant l_j(0).$$

Finally, we have

$$l = l_j(0).$$

Case 2. $\hat{m}_k < 0$ for $k = 1, 2, 3$. Thus, we know the sample means of $F_k q_1, \cdots, F_k q_n$ are all negative, for $k = 1, 2, 3$. Now, consider

$$\begin{aligned} & l_0 = \max \sum_{i=1}^{n} \log n p_i \\ \text{subject to } & \sum_{i=1}^{n} p_i = 1, \\ & p_i \geqslant 0,\ i = 1, \cdots, n. \end{aligned} \tag{5.2.17}$$

The constraints of (5.2.12) are more strict than the constraint of (5.2.17). Hence, we have $l \leqslant l_0$.

The Lagrange Multiplier method gives the solution, for $i = 1, 2, \cdots, n$,

$$p_i = \frac{1}{n}. \tag{5.2.18}$$

The solution (5.2.18) is also a solution to (5.2.12), since the solution fits the constraint

$$\sum_{i=1}^{n} p_i F_k q_i \leqslant 0.$$

Under this solution,

$$\sum_{i=1}^{n} \log n p_i = 0,$$

which is the maximum that l_0 can reach. Now, the solution (5.2.18) also makes l reach the maximum 0. Hence, for $i = 1, 2, \cdots, n$,

$$p_i = \frac{1}{n},$$

gives the solution to (5.2.12), i.e. $l = 0$.

□

5.3 WILKS' THEOREM ON 3-SPIDER

Consider a random sample $q_1, q_2, \cdots, q_n$ on $\mathcal{SP}_3$ with common population distribution P_0 and finite variance. Let $\bar{b}$, b_n be the Fréchet mean and the sample Fréchet mean on $\mathcal{SP}_3$ respectively, where

$$\bar{b} = \underset{b \in \mathcal{SP}_3}{\operatorname{argmin}} \int_{\mathcal{SP}_3} d_{\mathcal{SP}_3}(b, q)^2 dP_0(q).$$

Correspondingly, the sample Fréchet mean is given by

$$b_N = \underset{b \in \mathcal{SP}_3}{\operatorname{argmin}} \sum_{i=1}^{n} d_{\mathcal{SP}_3}(b, q_i)^2.$$

Let m_k, $\hat{m}_k$, $j = 1, 2, 3$, be defined as in Definition 5.2.11. Now, define Wilks' statistic in $\mathcal{SP}_3$ as $-2\log\mathcal{R}(\bar{b})$, where $\mathcal{R}(\bar{b})$ is defined as in Section 5.2.1 and Section 5.2.2 according to the location of $\bar{b}$ in $\mathcal{SP}_3$. Now, we give Wilks' Theorem on $\mathcal{SP}_3$.

Theorem 5.1 (Wilks' Theorem on 3-Spider) *Assume nondegeneracy*

$$P_0(L_k^+) \in (0,1), \quad k = 1, 2, 3. \tag{5.3.1}$$

1. ***Sticky Case**: Suppose $m_k < 0$ for all $k \in \{1, 2, 3\}$. There exists a random integer $N_0(\omega)$ with probability 1, such that for all $n \geqslant N_0(\omega)$, $-2\log\mathcal{R}(\bar{b}) = 0$.*

2. ***Non-Sticky Case**: Suppose $m_j > 0$, for some $j \in \{1, 2, 3\}$. Then,*

$$-2\log\mathcal{R}(\bar{b}) \xrightarrow{D} \chi_1^2,$$

   as $n \to \infty$.

3. ***Partially-Sticky Case**: Suppose $m_k = 0$, for some $j \in \{1, 2, 3\}$ and write $W(\bar{b}) = -2\log\mathcal{R}(\bar{b})$. Then,*

$$Pr\left[W(\bar{b}) \leqslant w\right] \to F(w) = \begin{cases} 0 & w < 0, \\ \frac{1}{2} & w = 0, \\ \frac{1}{2} + \frac{1}{2}F_V(w) & w > 0, \end{cases} \tag{5.3.2}$$

   as $n \to \infty$, and where $V \sim \chi_1^2$, and $F_V(\cdot)$ denotes the cumulative distribution function of V. In (5.3.2), pointwise convergence occurs at all points of continuity of F, i.e. at all points except $w = 0$.

In the non-sticky case, Wilks' theorem for empirical likelihood takes a familiar form to that in Euclidean space. However, since the cone point is the singularity point in $\mathcal{SP}_3$, Lemma 5.2.2 explains the behaviour of Wilks' statistic under the sticky case and the partially-sticky case. It is noted that Theorem 5.1 can be used to construct the confidence regions for the Fréchet mean by the empirical likelihood method.

5.4 TWO USEFUL LEMMAS

The lemmas shown below will be used in the proof of Theorem 5.1.

Lemma 5.4.1 *Assume the non-degeneracy (5.3.1). If there exists $j \in \{1,2,3\}$, such that the first moment $m_j = 0$, then $m_k < 0$, $k \neq j$.*

PROOF. Define $S_k = \int_{L_k} F_k q P_0(q)$, $k = 1,2,3$. Non-degeneracy implies that $S_k > 0$, $k = 1,2,3$. Then,

$$\begin{aligned} 0 = m_j =& \int_{\mathcal{SP}_3} F_j q dP_0(q) \\ =& \int_{L_j} F_j q dP_0(q) - \sum_{k\neq j} \int_{L_k} F_k q dP_0(q) \\ =& S_j - \sum_{k\neq j} S_k. \end{aligned}$$

Hence,

$$\begin{aligned} S_k =& S_j - \sum_{h\neq j,k} S_h \\ <& S_j + \sum_{h\neq j,k} S_h. \end{aligned}$$

It follows that

$$m_k = S_k - \sum_{m\neq k} S_m < 0,$$

for any $k \neq j$. □

Note that Lemma 5.4.1 is the population version of the proof in Step 2 of Case 1 in Lemma 5.2.2.

Lemma 5.4.2 *Assume the square integrability*

$$\int_{\mathcal{SP}_3} (F_k q)^2 dP_0(q) < \infty.$$

where $k = 1,2,3$. Let m_k, $\hat{m}_k$ be defined as in (5.2.11). If $m_j < 0$, for $j \in \{1,2,3\}$, then

$$Pr(\hat{m}_j \geqslant 0 \text{ happens finitely often}) = 1.$$

PROOF. By the strong LLN:

$$\lim_{N\to\infty} \hat{m}_j = m_j,$$

for $j \in \{1,2,3\}$, holds almost surely.

That is, there exists a null set

$$A = \{\omega \in \Omega | \hat{m}_j(\omega) \geqslant 0 \text{ happens infinitely often}\},$$

such that $P(A) = 0$.

For any $\omega \notin A$,

$$\lim_{N\to\infty} \hat{m}_j(\omega) = m_j(\omega) < 0.$$

i.e.

$$Pr(\hat{m}_j \geqslant 0 \text{ happens finitely often}) = 1.$$

□

5.5 PROOF OF THEOREM 5.1

Now, with lemmas provided in Section 5.4, we give the proof of Theorem 5.1 as follow:

PROOF OF THEOREM 5.1 : According to the statement in Theorem 5.1, we prove the theorem under three different cases: sticky case, non-sticky case and partially sticky case. Let $W(b) = -2\log\mathcal{R}(b)$. Thus, $W(\bar{b})$ is Wilks' statistic at the Fréchet mean.

Case 1. **Sticky Case**. Definition 2.6.6 shows that $m_k < 0$, $k = 1,2,3$. Theorem 2.3 implies that the population Fréchet mean is on the cone point.

Lemma 5.4.2 points out that $\hat{m}_j \xrightarrow{a.s.} m_j$. The almost surely convergence implies that there exists a null set A, such that $P(A) = 0$, and for any $\omega \notin A$, there are only finite times when $\hat{m}_k \geqslant 0$, $k = 1,2,3$. Now, set

$$N_k(\omega) = \sup\{n \in \mathbb{N} | \hat{m}_k \geqslant 0\},$$

i.e. the last time when the event $\{\hat{m}_k \geqslant 0\}$ happens, $k = 1, 2, 3$.

Then, let

$$N_0(\omega) = \max_{k\in\{1,2,3\}} N_k(\omega).$$

For $n \geqslant N_0(\omega)$, $\hat{m}_k < 0$, $k = 1, 2, 3$, and Lemma 5.2.2 implies that $W(\bar{b})(\omega) = W(b_n)(\omega) = 0$, where $\bar{b}$ and b_n are both on the cone point.

Case 2. **Non-Sticky Case**. From Definition 2.6.6, we know $m_j > 0$ for some $j \in \{1, 2, 3\}$. Theorem 2.3 implies that the population Fréchet mean $\bar{b} \in L_j^+$. Thus, we use (5.2.4) to evaluate $W(\bar{b})$.

Under the folding map F_j, $F_j q_1, F_j q_2, \cdots, F_j q_n$ are i.i.d. random variables with the common finite mean and finite variance.

Since $m_j = F_j \bar{b} > 0$,

$$W(\bar{b}) = -2 \log \mathcal{R}(m_j),$$

where $-2 \log \mathcal{R}(m_j)$ is Wilks' statistic in Theorem 2.1. Hence,

$$W(\bar{b}) \xrightarrow{D} \chi_1^2,$$

as $n \to \infty$.

Case 3. **Partially-Sticky Case**. If $m_j = 0$ for some $j \in \{1, 2, 3\}$, then

$$m_j = \int_{\mathcal{SP}_3} F_j q dP_0(q) = 0,$$

and Lemma 5.4.1 implies that $m_k < 0$ for $k \neq j$. Theorem 2.3 implies that $\bar{b}$ is at the cone point.

Let S_j be defined as in Lemma 5.4.1. Without lossing generality, we assume $j = 1$. Thus, $S_1 = S_2 + S_3$, which means the weight on L_1 is equal to the weight on other two legs together. Hence, according to the law of large number, $m_k < 0$, for $k \neq j$, implies that

$$Pr(\hat{m}_k \geqslant 0, k \neq j, \text{ happens finitely often}) = 1.$$

For $m_j = 0$, we consider two cases:

1. $Pr(\hat{m}_j < 0) \to \frac{1}{2}$, as $n \to \infty$.

   Under the condition that $\hat{m}_k < 0$, $k = 1, 2, 3$, Both the sample Fréchet mean and the population Fréchet mean are on the cone point. Then, Case 2 of Lemma 5.2.2 implies that

   $$Pr(W(\bar{b}) = 0) = Pr(\{\hat{m}_k < 0, k = 1, 2, 3\}) \to \frac{1}{2},$$

   as $n \to \infty$.

2. $Pr(\hat{m}_j \geqslant 0) \to \frac{1}{2}$, as $n \to \infty$.

   Recalling $\hat{S}_k$, $k = 1, 2, 3$, as in (5.2.13), we know, under the condition that $\hat{m}_j \geqslant 0$, $\hat{m}_k < 0$, $k \neq j$, Case 1 of Lemma 5.2.2 implies that $W(\bar{b})$ is evaluated under the folding map F_j.

   The partially-sticky case $m_j = 0$ implies that $F_j q_1, \cdots, F_j q_n$ has common mean 0. With common finite variance σ^2, CLT implies that

   $$\frac{\sqrt{n}\hat{m}_j}{\sigma} \Big| \{\hat{m}_j \geqslant 0\} \xrightarrow{D} \widetilde{N},$$

   where $\widetilde{N}$ is the half-normal distribution derived from $N(0, \sigma^2)$.

   It follows that

   $$\frac{n\hat{m}_j^2}{\sigma^2} \xrightarrow{D} \chi_1^2,$$

   as $n \to \infty$.

   Now, write

   $$W(\bar{b}) = \begin{cases} 0, & \hat{m}_j < 0, \\ \widetilde{W}(\bar{b}), & \hat{m}_j \geqslant 0. \end{cases}$$

   The proof of Theorem 2.1 implies that $n\hat{m}_j^2/\sigma^2$ is the leading term of $\widetilde{W}$, and thus

   $$\widetilde{W}(\bar{b}) \xrightarrow{D} \chi_1^2,$$

   as $n \to \infty$.

Let V be a Chi-square random variable with one degree of freedom. Let $F_V(\cdot)$ be its cumulative distribution function. Then,

$$Pr\left[W(\bar{b}) \leqslant w\right] \to F(w) = \begin{cases} 0 & w < 0, \\ \frac{1}{2} & w = 0, \\ \frac{1}{2} + \frac{1}{2}F_V(w) & w > 0, \end{cases}$$

as $n \to \infty$, where pointwise convergence occurs at all points of continuity of F, i.e. at all points except $w = 0$.

□

5.6 ALGORITHM: EMPIRICAL LIKELIHOOD RATIO ON 3-SPIDER

Following Section 5.2.1 and Section 5.2.2, we propose an algorithm for evaluating the empirical likelihood ratio on $\mathcal{SP}_3$ as follows. Suppose we are given a sample in $\mathcal{SP}_3$, $\mathscr{X} = \{q_1, q_2, \cdots, q_n\}$.

Algorithm 4 Empirical Likelihood Ratio Calculation on 3-Spider

Step 1. For the sample Fréchet mean parameter $b \in \mathcal{SP}_3$, if $b \notin L_0$, go to Step 2; if $b \in L_0$, go to Step 3.

Step 2. Suppose $b \in L_j^+$, $j \in \{1, 2, 3\}$. Solving (5.2.4) with this choice of j gives the empirical likelihood ratio.

Step 3. Suppose $b \in L_0$, i.e. the cone point. Consider the sample first moments $\hat{m}_k$, for $k = 1, 2, 3$:

1. Suppose $\hat{m}_k < 0$, for $k = 1, 2, 3$. Then, $\mathcal{R}(b) = 1$.
2. Suppose there exists j such that $\hat{m}_j \geqslant 0$ and $\hat{m}_k < 0$, for $k \neq j$. $\mathcal{R}(b)$ is evaluated by $\mathcal{R}(0)$ under the folding map F_j.

Following **Algorithm** 4, we shall in the next two sections conduct experiments:

1. Evaluating the empirical likelihood ratios $\mathcal{R}(b)$ for the Fréchet mean parameter b, to study how to construct confidence regions.

2. Using bootstrap calibration to evaluate the contour of the confidence region for the Fréchet mean, and to investigate operating characteristics in hypothesis testing, including the probability of type I error and the probability of type II error. Comparing these probabilities with the ones calculated by checking chi-squared table, as was done in Section 4.5.

5.7 SIMULATION

Now, we use Algorithm 4.5 in three simulation cases, fully-sticky, partially-sticky and non-sticky, to calculate empirical likelihood ratio against the mean parameter on each leg.

5.7.1 *Non-Sticky Case*

In this simulation, the distribution has such settings:

1. Leg 1, 2 and 3 respectively have proportions 1/2, 1/3, and 1/6.

2. From Leg 1 to 3, the distribution is set to be exponential with parameter 1, 2, and 1 respectively.

3. The population Fréchet mean is 1/6 on Leg 1.

The graph of empirical likelihood ratio against the mean parameter on each leg is shown in Figure 5.7.22. The sample Fréchet mean is 0.142.

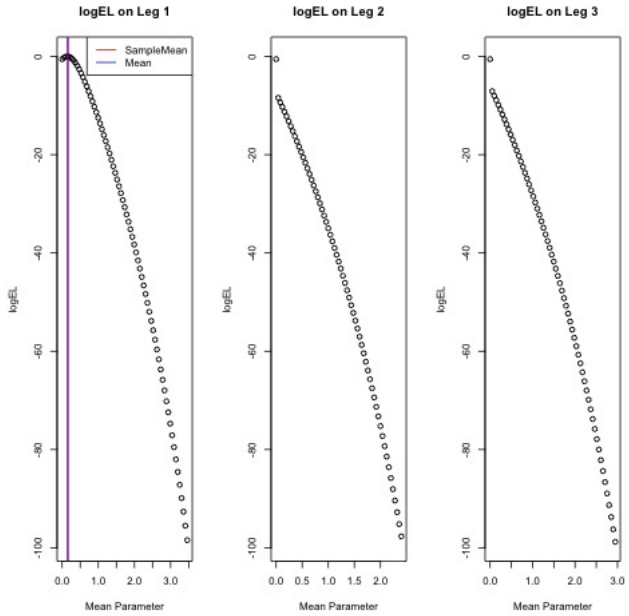


Figure 5.7.22: Empirical Likelihood Ratio against Fréchet Mean Parameter on $\mathcal{SP}_3$ (Non-Sticky).

5.7.2 *Partially-Sticky Case*

In this simulation, the distribution has such settings:

1. Leg 1, 2 and 3 respectively have proportions 1/2, 1/4, and 1/4.
2. From Leg 1 to 3, the distribution is set to be exponential with parameter 1, 1, and 1 respectively.
3. The population Fréchet mean is on the cone point.

The graph of empirical likelihood ratio against the mean parameter on each leg is shown in Figure 5.7.23 and Figure 5.7.24. As is expected from Theorem 5.1, the sample Fréchet mean is either on the cone point or strictly on a certain leg.

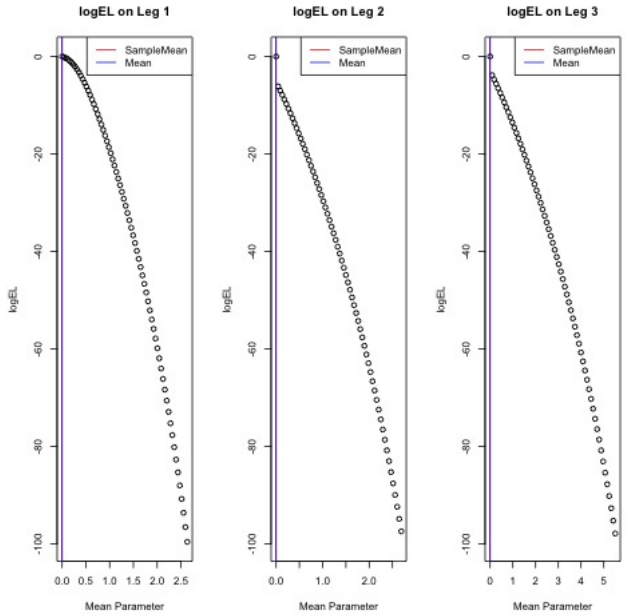


Figure 5.7.23: Empirical Likelihood Ratio against Fréchet Mean Parameter on $\mathcal{SP}_3$ (Partially-Sticky). The sample Fréchet mean is on the cone point.

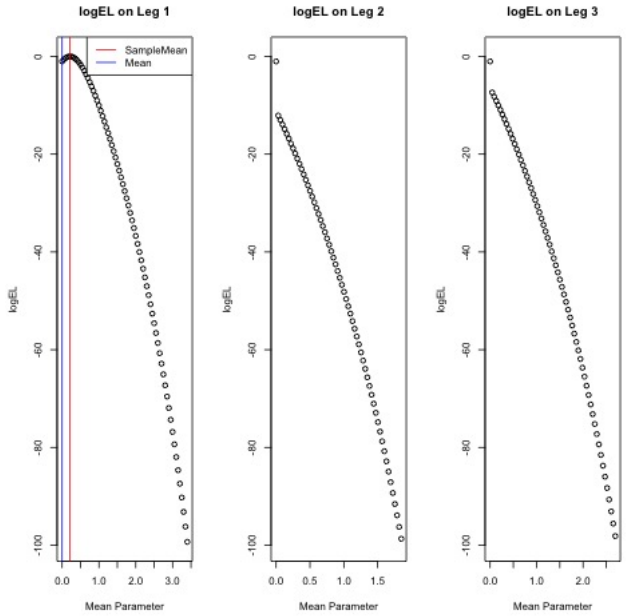


Figure 5.7.24: Empirical Likelihood Ratio against Fréchet Mean Parameter on $\mathcal{SP}_3$ (Partially-Sticky). The sample Fréchet mean is away the cone point.

5.7.3 *Fully-Sticky Case*

In this simulation, the distribution has such settings:

1. Leg 1, 2 and 3 respectively have proportions 1/3, 1/3, and 1/3.
2. From Leg 1 to 3, the distribution is set to be exponential with parameter 1, 1, and 1 respectively.
3. The population Fréchet mean is on the cone point.

The graph of empirical likelihood ratio against the mean parameter on each leg is shown in Figure 5.7.25. The sample Fréchet mean is on the cone point.

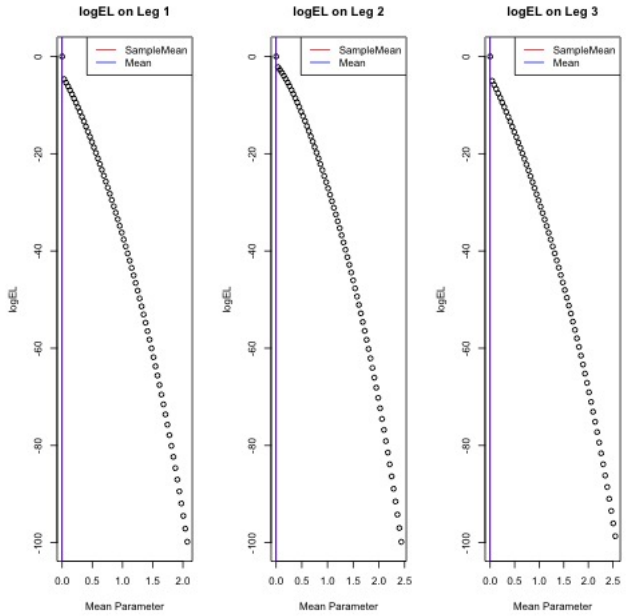


Figure 5.7.25: Empirical Likelihood Ratio against Fréchet Mean Parameter on $\mathcal{SP}_3$ (Fully-Sticky).

5.8 PHYLOGENETIC TREE DATA IN 3-SPIDER

First, we evaluate the empirical likelihood ratio for the Fréchet mean parameter on $\mathcal{SP}_3$. Instead of simulated data, we use a real dataset, the metazoan

data in Nye (2011), which includes 106 gene trees from 21 metazoan species. We first connect the phylogenetic tree data with the points on $\mathcal{SP}_3$.

5.8.1 *Phylogenetic Tree Data on 3-Spider*

As is shown in Subfigure (a) of Figure 2.7.5, the phylogenetic tree dataset is composed of the lengths of internal edges and external edges. Consider those three types of phylogenetic tree datasets in Figure 2.7.5. If the internal edge a shrinks to 0, then we have a new phylogenetic tree data, as is shown in Subfigure (b) of Figure 2.7.5. Then, extending the internal edge a' from 0, we have another new phylogenetic tree data, as is in Subfigure (c) of Figure 2.7.5.

Only considering the internal edges, those three types of phylogenetic tree data can be allocated to the corresponding locations on $\mathcal{SP}_3$: moving a point from L_i^+ to L_j^+, $i \neq j$. It involves the same idea of transforming Subfigure (a) to Subfigure (c) in Figure 2.7.5. Shrinking an internal edge to zero can be regarded as moving a point to the cone point. Extending the interior edge gradually from 0 in another internal edge involves the same idea of moving the point from the cone point to another leg.

Nye (2011) used metazoan data for a principal component analysis. We observe the structure of each tree dataset, and finally find most of the tree datasets (99 out of 106) have a common feature that “Calb", “Scas" and “Sklu" stick together. Thus, ignoring tree datasets without that feature, and treating the tree data with such a feature as the source in Figure 5.8.26, we allocate these 99 tree data onto $\mathcal{SP}_3$. Now we apply Algorithm 4, and plot the log-empirical likelihood against the Fréchet mean parameter as follows.

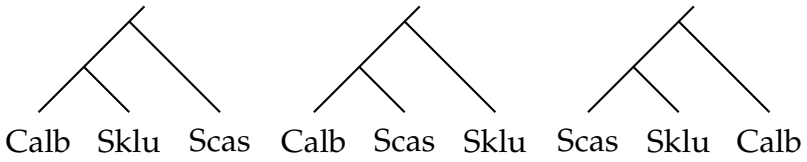


Figure 5.8.26: Only considering tree data with sub-structure above, 99 out of 106 metazoan dataset are divided into 3 clusters and allocated onto $\mathcal{SP}_3$ according to their internal edges.

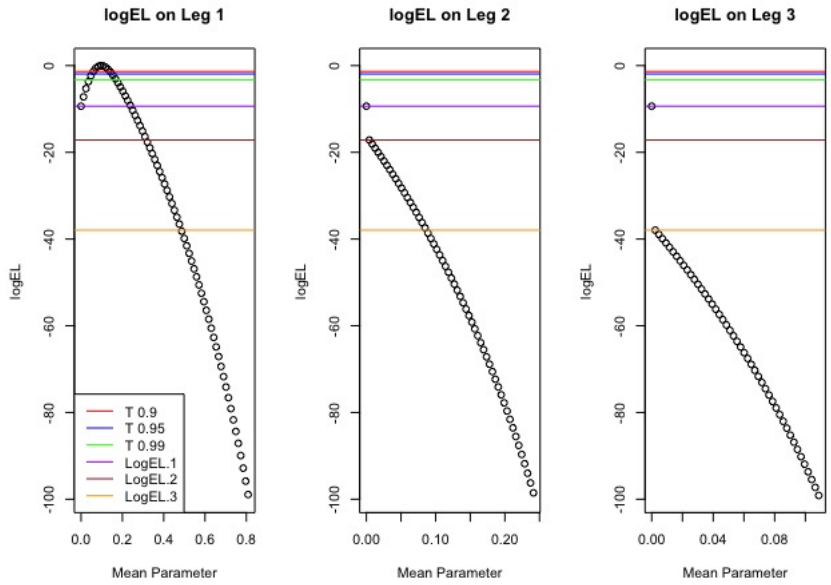


Figure 5.8.27: Empirical Likelihood Ratio against Fréchet Mean Parameter on $\mathcal{SP}_3$ (Metazoan). Quantiles are from chi-squared Table.

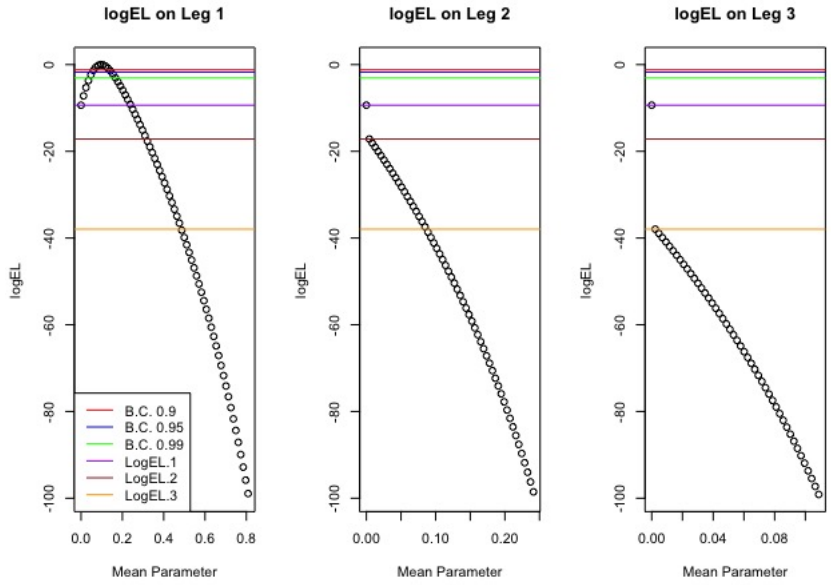


Figure 5.8.28: Empirical Likelihood Ratio against Fréchet Mean Parameter on $\mathcal{SP}_3$ (Metazoan). Quantiles are from bootstrap calibration.

It should be noted that in Figure 5.8.27 and Figure 5.8.28, the phylogenetic tree data with structures shown in Figure 5.8.26 (from left to right in order) are assigned to Leg $1, 2, 3$ respectively. In Figure 5.8.27 and 5.8.28, we either choose quantiles from chi-squared table or calculate bootstrap calibration quantiles, and then calculate the corresponding log-empirical likelihood ratios $\log \mathcal{R}(b)$, shown by horizontal lines with different colours. We will introduce algorithms for calculating bootstrap calibration quantiles in Section 5.9, and then show the quantiles and their corresponding confidence regions, as a comparison with tabular quantiles and corresponding regions.

5.8.2 *Confidence Region Construction*

An important application of Wilks' theorem for empirical likelihood is calculating the confidence region for a chosen confidence level. In the 3-Spider, Theorem 5.1 is used for confidence region construction. Now, we take the case in Figure 5.8.27 and 5.8.28 as examples, and first explain the following lines with colours:

1. Red Line: the corresponding $\log \mathcal{R}(b)$ of 0.9 quantile;

2. Blue Line: the corresponding $\log \mathcal{R}(b)$ of 0.95 quantile;

3. Green Line: the corresponding $\log \mathcal{R}(b)$ of 0.99 quantile;

4. Purple Line: $\log \mathcal{R}(b)$ at the cone point under the folding map F_1, written as $l(\alpha_1)$, with the corresponding confidence level α_1,, where $\alpha_1 = 0.99998$;

5. Brown Line: $\log \mathcal{R}(b)$ at the cone point under the folding map F_2, written as $l(\alpha_2)$, with the corresponding confidence level $\alpha_1 > \alpha_2$, where $\alpha_2 \approx 1$;

6. Orange Line: $\log \mathcal{R}(b)$ at the cone point under the folding map F_3, written as $l(\alpha_3)$, with the corresponding confidence level $\alpha_2 > \alpha_3$, where $\alpha_3 \approx 1$.

The intersect points outline the confidence regions. However, the cone point in $\mathcal{SP}_3$ is a singularity point. Confidence regions may be constructed in a non-standard way. Now, we construct confidences regions as follows.

In this thesis, we are guided by empirical likelihood. The definition of empirical likelihood at the cone point makes the discontinuity on some legs at the cone point. Therefore, some confidence regions are non-standard. For example, in Figure 5.8.27 and 5.8.28 from the purple line and the brown line, one confidence region point will always stay at the cone point, but the confidence level increases from α_1 to α_2. Then, from the brown line to the orange line, the confidence region comprises of two intervals, one on L_1 and the other one on L_2, as is shown in Figure 5.8.27 and 5.8.28. Note that α_1, α_2 and α_3 are too close to each of other two, so another case for the confidence region construction needs investigating to observe these three confidence levels.

Now, given α, we construct confidence regions (labelled with green colour) in a non-sticky case as Figure 5.8.29.

1. $0 \leqslant \alpha \leqslant \alpha_1$. The confidence regions are constructed in the standard way, as in Subfigure (a).

2. $\alpha_1 < \alpha \leqslant \alpha_2$. One confidence region point stays at the cone point, but the other point is increasing as α increases, as is shown in Subfigure (b).

3. $\alpha_2 < \alpha \leqslant \alpha_3$. The confidence region now consists of two parts: one segment in L_1 and the other segment in L_2, as Subfigure (c) presents.

4. $\alpha_3 < \alpha \leqslant 1$.The confidence region is made up of three parts, with respectively one interval on each leg; see Subfigure (d).

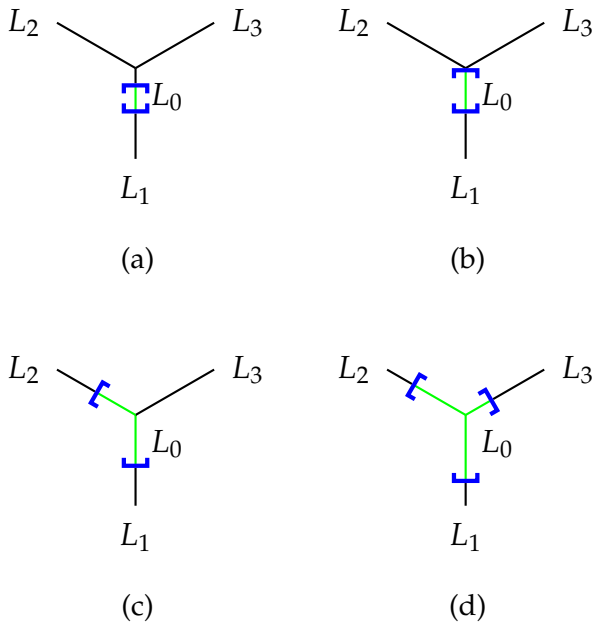


Figure 5.8.29: Confidence Regions Construction for the Fréchet mean on $\mathcal{SP}_3$ in a non-sticky case.

Remark: as is shown in Figure 5.8.29, at different levels of confidence, the confidence region for the Fréchet mean has 4 different forms. For example, as is shown in Figure 5.8.26, let the tree data with these three structures (from left to right) respectively be assigned to L_1, L_2 and L_3. If we regard these three structures as three categorization for species shown in Figure

5.8.26, the confidence region will provide a straightforward outcome about which categorization for species is acceptable at a certain confidence level. It would be interesting to discuss these results with a biologist.

As for the sticky case, the definition of empirical likelihood at the cone point makes the discontinuity for the cone point to all legs. Now in order to investigate the confidence region construction in sticky case, the Metazoan sample is switched to a fully-sticky case. Here for Metazoan phylogenetic tree data in $\mathcal{SP}_3$, we shrink the sample on L_1 so that the sample Fréchet mean is strictly on the cone point. Note that shrinking means the sample on Leg 1 are divided by 4, so that we get a fully-sticky case. As is shown in Figure 5.8.30, the log-empirical likelihood is not continuous at the cone point on each leg.

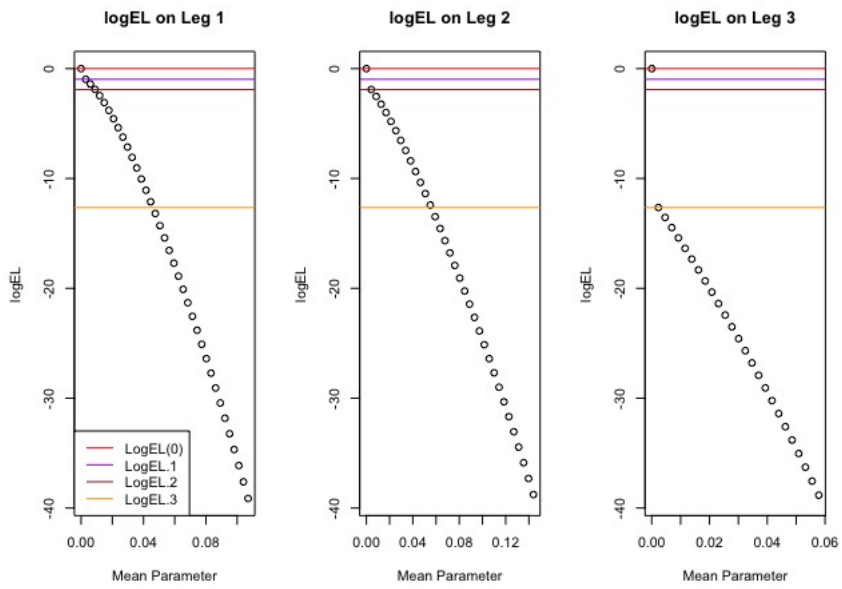


Figure 5.8.30: Empirical Likelihood Ratio against Fréchet Mean Parameter on $\mathcal{SP}_3$, with Metazoan Data shrunk in L_1.

In order to calculate the confidence regions, now we explain the following lines with colours in Figure 5.8.30:

1. Red Line: $\log \mathcal{R}(b)$ at the cone point, written as $l(0)$.

2. Purple Line: $\log \mathcal{R}(b)$ at the cone point under the folding map F_1, written as $l(\alpha_1)$, with the corresponding confidence level α_1, where $\alpha_1 = 0.514$;

3. Brown Line: $\log \mathcal{R}(b)$ at the cone point under the folding map F_2, written as $l(\alpha_2)$, with the corresponding confidence level $\alpha_1 > \alpha_2$, where $\alpha_2 = 0.671$;

4. Orange Line: $\log \mathcal{R}(b)$ at the cone point under the folding map F_3, written as $l(\alpha_3)$, with the corresponding confidence level $\alpha_2 > \alpha_3$, where $\alpha_3 = 0.988$.

Now, given α, we construct confidence regions (labelled with green colour) in a sticky case as Figure 5.8.31.

1. $0 \leqslant \alpha \leqslant \alpha_1$. The corresponding confidence regions are constructed within the cone point, i.e. L_0, as Subfigure (a).

2. $\alpha_1 < \alpha \leqslant \alpha_2$. One confidence region point stays at the cone point, but the other point is increasing as α increases, as is shown in Subfigure (b).

3. $\alpha_2 < \alpha \leqslant \alpha_3$. The confidence region now consists of two parts: one segment in L_1 and the other segment in L_2, as Subfigure (c) presents.

4. $\alpha_3 < \alpha \leqslant 1$.The confidence region is made up of three parts, with respectively one interval on each leg; see Subfigure (d).

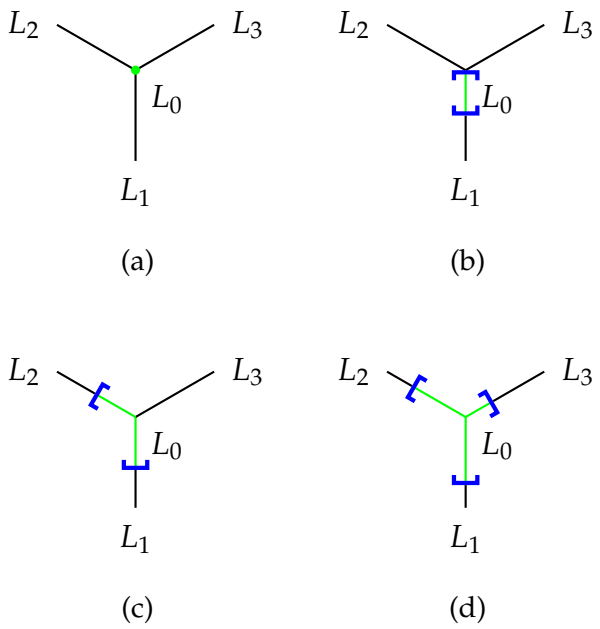


Figure 5.8.31: Confidence Regions Construction for the Fréchet mean on $\mathcal{SP}_3$ in a sticky case.

Now, we list the steps to build the confidence region C_α, for a given α:

Step 1. Use the confidence level α to find the corresponding quantile, either bootstrap calibration or chi-squared table.

Step 2. Calculate the corresponding log-empirical likelihood ratio.

Step 3. As in the horizontal lines in Figure 5.8.27 and 5.8.28, calculate the confidence region C_α in $\mathcal{SP}_3$.

We have investigated the confidence region construction with the confidence level given. On the other hand, if the confidence region C_α is given, we can also calculate the corresponding confidence level α by reversing the steps above:

Step 1. For a specific confidence region C_α, calculate the log-empirical likelihood ratio of confidence region contour points.

Step 2. Evaluate the Wilks' statistic for empirical likelihood.

Step 3. Find the confidence level α either by chi-squared table or bootstrap calibration.

5.9 NUMERICAL REALIZATION

As is discussed in Section 4.5, probabilities of type I error and type II error are calculated by Monte-Carlo and bootstrap calibration, given the nominal Type I error α. Now, we conduct analogous experiments on $\mathcal{SP}_3$. Here, we consider the example that

$$H_0 : \bar{b} = \bar{b}_0 \in L_1^+.$$

5.9.1 *Probability of Type I Error*

First, we give the algorithm to calculate the probability of type I error by checking chi-squared table and bootstrap calibration. Suppose the hypothesis test is:

$$H_0 : \bar{b} = \bar{b}_0 \in L_1^+, \ v.s. \ H_1 : \bar{b} \neq \bar{b}_0.$$

Suppose the sample in this hypothesis testing is $\mathscr{X}$, sampled under hypothesis H_0.

Algorithm 5 Hypothesis Testing under H_0 (3-Spider)

1. Use Algorithm 4 to calculate $-2\log\mathcal{R}(\bar{b}_0)$.
2. [Tabular Method] For a confidence level $1-\alpha$, find the tabular quantile $\chi^2_{1,1-\alpha}$. If $-2\log\mathcal{R}(\bar{b}_0) \leqslant \chi^2_{1,1-\alpha}$, accept H_0; otherwise, reject H_0.
3. Generate bootstrap resamples $\mathscr{X}^*$ of size N sampled randomly with replacement from the sample $\mathscr{X}$. Within each resample $\mathscr{X}^*$, evaluate $-2\log\mathcal{R}(b_n)$, where b_N denotes the sample Fréchet mean of $\mathscr{X}$, written as $W_t^*, 1 \leqslant t \leqslant B$.
4. [Bootstrap Calibration Method] Order $W_t^*, 1 \leqslant t \leqslant B$, as $W^*_{[1]} \leqslant \cdots \leqslant W^*_{[B]}$. Find $W^*_{[B(1-\alpha)]}$, the bootstrap quantile. Then, if $-2\log\mathcal{R}(\bar{b}) \leqslant W^*_{[B(1-\alpha)]}$, accept H_0; otherwise, reject H_0.

Now, we use Algorithm 5 to calculate the probability of type I error. First, we set up a Monte Carlo loop to repeat the steps above for M times. If the null hypothesis is accepted A times in Step 2, the tabular probability of type I error is $1 - A/M$. If the null hypothesis is accepted A′ times in Step 4, the bootstrap calibration probability of type I error is $1 - A'/M$.

Now, we conduct Monte-Carlo experiments for probability of type I error, following Algorithm 2. The sample is generated with Fréchet mean $9/4$ on L_1^+. In each experiment, the number of Monte-Carlo runs is 500. For each Monte-Carlo run, bootstrap calibration is done for 500 times. Within each hypothesis testing, the nominal probability of type I error is given $\alpha = 0.05$. For each experiment, the sample size ranges respectively at 10, 20, 50, and 200. Table 22 and Figure 5.9.32 presents the result of tabular probability of type I error and bootstrap calibration probability of type I error.

Sample Size	Tabular α	B.C. α
10	0.1530 (0.1138)	0.0695 (0.0804)
20	0.1005 (0.0672)	0.0690 (0.0567)
50	0.0595 (0.0335)	0.0545 (0.0321)
200	0.0670 (0.0177)	0.0630 (0.0172)

Table 22: Probability of type I error under the hypothesis test $H_0 : \bar{b} = 9/4 \quad v.s. \quad H_1 : \bar{b} \neq 9/4$ under different sample sizes. The standard errors of probabilities of type I error are given in brackets.

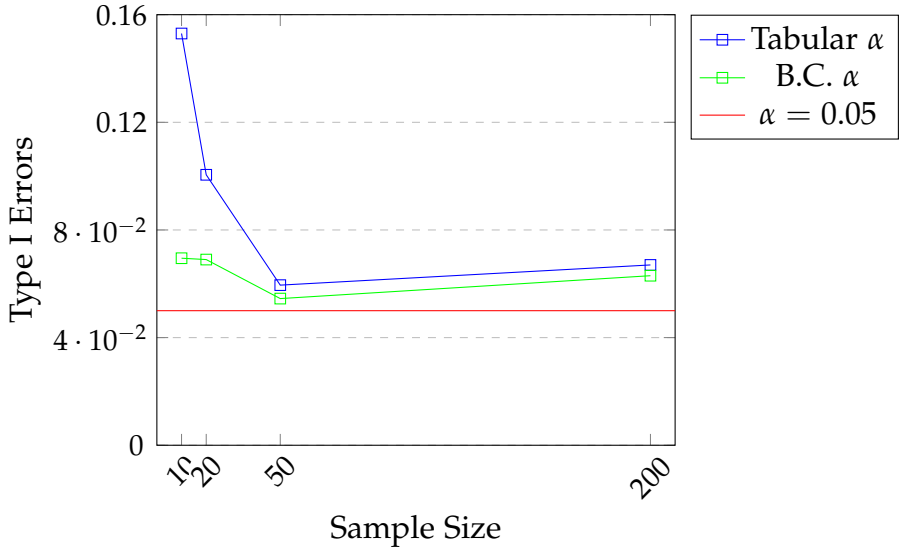


Figure 5.9.32: Tabular probability of type I error and bootstrap calibration probability of type I error of the test under different sample sizes.

Observing Table 22, and Figure 5.9.32, we make the following comments:

1. The outcome suggests that at small sample sizes, i.e. less than 50, the probability of type I error evaluated by the tabular quantile is much larger than the nominal probability of type I error 0.05, while the probability of type I error calculated by the bootstrap calibration quantile is slightly larger than 0.05. As is the comment in Section 4.5.1, bootstrap calibration probability of type I error is preferable.

2. When the sample size is larger than 50, the probabilities of type I error calculated both by the tabular quantile and the bootstrap calibration quantile are close to each other, although from Figure 5.9.32 the probability of type I error evaluated by the tabular quantile is still slightly larger than that by the bootstrap calibration quantile.

5.9.2 *Probability of Type II Error*

In order to calculate the power of the test, we set the test is:

$$H_0 : \bar{b} = \bar{b}_0. \; v.s. \; H_1 : \bar{b} = \bar{b}_1.$$

Note that the nominal probability of type I error is α, and we assume $\bar{b}_0 \in L_j^+$, $j \in \{1, 2, 3\}$. Now, we list the algorithm to calculate the probability of type II error on $\mathcal{SP}_3$. Suppose the sample in hypothesis testing is $\mathscr{Y}$ from H_1.

Algorithm 6 Hypothesis Testing under H_1 (3-Spider)

1. Use Algorithm 4 to calculate $-2\log\mathcal{R}(\bar{b}_0)$.
2. [Tabular Method] For a confidence level $1-\alpha$, find the tabular quantile $\chi^2_{2,1-\alpha}$. Then, if $-2\log\mathcal{R}(\bar{b}_0) \leqslant \chi^2_{2,1-\alpha}$, accept H_0; otherwise, reject H_0.
3. Use the folding map F_j to map the sample $\mathscr{Y}$ from $\mathcal{SP}_3$ to $\mathbb{R}$, labelled as $\mathscr{Y}_j$. Let each observation in $\mathscr{Y}_j$ be deducted by $\Delta = (F_j\bar{b}_1 - F_j\bar{b}_0)$, i.e. $\mathscr{Y}_j' = \mathscr{Y}_j - \Delta$.
4. Generate bootstrap resamples $\mathscr{Y}_j^*$ of size N sampled randomly with replacement from the sample $\mathscr{Y}_j'$. Within each resample $\mathscr{Y}_j^*$, evaluate $-2\log\mathcal{R}(\bar{b}_0)$, written as $W_t^*, 1 \leqslant t \leqslant B$.
5. [Bootstrap Calibration Method] Order $W_t^*, 1 \leqslant t \leqslant B$, as $W_{[1]}^* \leqslant \cdots \leqslant W_{[B]}^*$. Find $W_{[B(1-\alpha)]}^*$, the bootstrap quantile. Then, if $-2\log\mathcal{R}(\boldsymbol{\mu}_0) \leqslant W_{[B(1-\alpha)]}^*$, accept H_0; otherwise, reject H_0.

In order to calculate the power of the test, we set up a Monte Carlo loop to repeat the steps above for M times. If the null hypothesis is accepted A times in Step 2, the tabular probability of type II error is A/M. If the

null hypothesis is accepted A' times in Step 4, the bootstrap calibration probability of type II error is A'/M. Then,

$$Power = 1 - Pr(\text{Type II Error}).$$

Within each experiment, the sample size ranges respectively at 10, 20, 50 and 200. Without loss of generality, we assume $\bar{b}_0$ is 9/4 on L_1^+. In the experiments, the alternative hypotheses are selected as follows:

1. $\bar{b}_1 = 13/4 \in L_1^+$;
2. $\bar{b}_1 = 12/4 \in L_1^+$;
3. $\bar{b}_1 = 11/4 \in L_1^+$;
4. $\bar{b}_1 = 10/4 \in L_1^+$.

Table 24 and Figure 5.9.33 show the probabilities of type II error by the tabular quantile and bootstrap calibration quantile against each alternative hypothesis under different sample sizes.

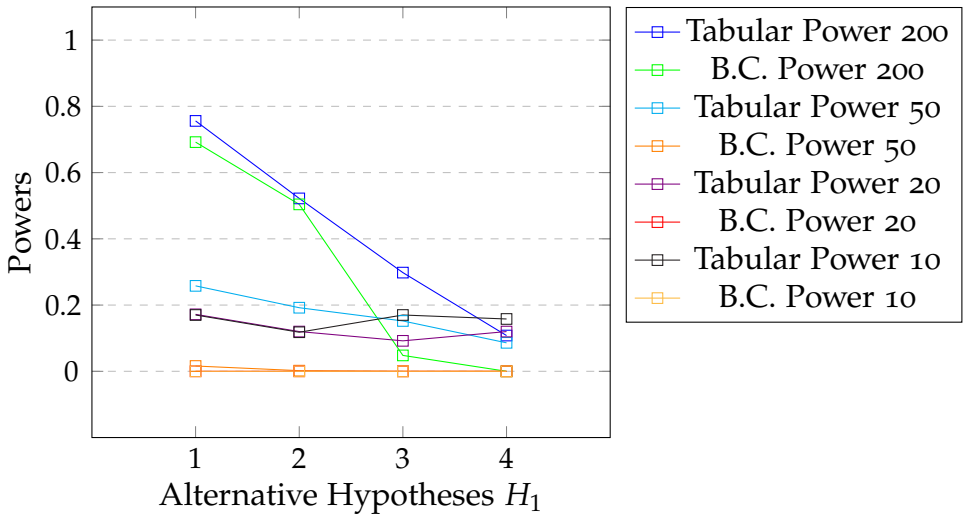


Figure 5.9.33: Tabular Powers and Bootstrap Calibration (B.C.) Powers of the test against the alternative hypotheses under different sample sizes.

Sample Size	$\bar{b}_1$	Tabular β	B.C. β
200	$\frac{13}{4}$	0.244 (0.030)	0.308 (0.033)
	$\frac{12}{4}$	0.478 (0.035)	0.496 (0.035)
	$\frac{11}{4}$	0.702 (0.032)	0.952 (0.015)
	$\frac{10}{4}$	0.892 (0.022)	1.000 (0.000)
50	$\frac{13}{4}$	0.742 (0.062)	0.984 (0.018)
	$\frac{12}{4}$	0.808 (0.056)	0.998 (0.006)
	$\frac{11}{4}$	0.848 (0.051)	1.000 (0.000)
	$\frac{10}{4}$	0.914 (0.040)	1.000 (0.000)
20	$\frac{13}{4}$	0.828 (0.084)	1.000 (0.000)
	$\frac{12}{4}$	0.880 (0.073)	1.000 (0.000)
	$\frac{11}{4}$	0.908 (0.065)	1.000 (0.000)
	$\frac{10}{4}$	0.880 (0.073)	1.000 (0.000)
10	$\frac{13}{4}$	0.830 (0.119)	1.000 (0.000)
	$\frac{12}{4}$	0.882 (0.102)	1.000 (0.000)
	$\frac{11}{4}$	0.830 (0.119)	1.000 (0.000)
	$\frac{10}{4}$	0.842 (0.115)	1.000 (0.000)

Table 24: Probability of type II error for the hypothesis test $H_0 : \bar{b} = \bar{b}_0$ *v.s.* $H_0 : \bar{b} = \bar{b}_1 \neq \bar{b}_0$ under different sample sizes. The standard errors of probabilities of type II error are given in brackets.

Observing the Figure 5.9.33 and Table 24, we make the following comment:

1. Table 5.9 and Figure 5.9.32 already suggest for small sample sizes, such as 10 and 20, the probability of type I error calculated by the tabular quantile is too large. Then we know the probability of type II error evaluated by the tabular quantile loses the accuracy.

2. For large sample sizes, the outcome indicates that the power of the test decreases when H_1 is close to H_0, since the closer H_1 is to H_0, the more likely it is for the type II error to happen.

3. The power line at sample size 200 and 50 shows that the power of the test increases when the sample size increases.

5.9.3 *Application in Phylogenetic Tree Data*

Observing Figure 5.8.27 and 5.8.28, we find the corresponding log-empirical likelihood values of the lines for the same colour are close to each other. Thus, confidence regions are also expected to be similar; see Table 25. The outcome in Section 5.9.1 has shown the similar result for the sample size over 50.

C.L. \ C.R.	B.C. Confidence Region	Tabular Confidence Region
0.9	(0.0613,0.1376)	(0.0590,0.1404)
0.95	(0.0544,0.1465)	(0.0522,0.1498)
0.99	(0.0407,0.1668)	(0.0387,0.1698)

Table 25: The confidence regions for the Fréchet mean, calculated by bootstrap calibration quantiles and chi-squared tabular quantiles. C.L.: Confidence Level. C.R.: Confidence Region. B.C.:Bootstrap Calibration.

5.10 SUMMARY

In this chapter, Wilks' theorem is mainly adapted to the non-Euclidean space, the 3-Spider. First, the log-empirical likelihood ratio function for the Fréchet mean parameter on the 3-Spider is defined. In this step, the object of the study is how to evaluate empirical likelihood for the Fréchet mean

parameter on and off the cone point. The folding map that sends the random variable on 3-Spider to Euclidean space is introduced simultaneously to evaluate empirical likelihood.

Since the cone point in the 3-Spider is a singularity point, there are three different cases, sticky case, non-sticky case and partially sticky case, for the locations of Fréchet mean in the 3-Spider, according to Hotz et al. (2013). Thus, the asymptotic distribution of Wilks' statistic is investigated following these three cases, i.e. Wilks' theorem for empirical likelihood on the 3-Spider.

Based on Wilks' theorem on the 3-Spider, the algorithm of calculating the log-empirical likelihood ratio is given. Then, the confidence region is constructed. We also use a real dataset, the metazoan phylogenetic tree datasets, to study how to construct the confidence region. Due to the fact that the cone point is a singularity point, the confidence region on $\mathcal{SP}_3$ is built in a non-standard way. Operating characteristics in hypothesis testing are also studied, including the probability of type I error and the probability of type II error. We find that with sample size over 50, there is no much difference for the outcome evaluated by bootstrap calibration quantiles and chi-squared tabular quantiles; for sample sizes smaller than 50, bootstrap calibration method is preferable. Other conclusions are similar to conclusions in the study of operating characteristics in hypothesis testing in the last chapter.

6

CONCLUSION AND FURTHER WORK

The aim of this thesis is to explore theoretical and computational aspects of empirical likelihood when applied to data from different types of spaces. A key theoretical goal is to provide a version of Wilks' theorem which opens the way to the application of methods of inference including the construction of confidence regions and hypothesis testing.

Wilks' theorems in different spaces and the concavity of the log-empirical likelihood in $\mathbb{R}^k$ are presented in Section 6.1. Conclusions about using empirical likelihood method for constructing confidence regions and operating characteristics are given in Section 6.2. In Section 6.3, further research is discussed.

6.1 WILKS' THEOREMS AND CONCAVITY OF log-EMPIRICAL LIKELIHOOD IN $\mathbb{R}^k$

First, based on Wilks' theorem by Owen (2001), we consider the idea of local alternatives by Sugiura (1973) to obtain a more general version of Wilks' theorem in $\mathbb{R}^k$. Theorem 3.1 states that Wilks' statistic evaluated at local alternatives has an asymptotic non-central chi-squared distribution with a non-centrality parameter associated with local alternatives.

In order to adapt Wilks' theorem to $\mathcal{S}^2$, we apply tangent spaces at chosen points on $\mathcal{S}^{k-1}$ to build the connection between empirical likelihood on $\mathcal{S}^{k-1}$ and empirical likelihood on $\mathbb{R}^k$. With a chosen rotation matrix (4.2.6),

the random vector on $\mathcal{S}^{k-1}$ is transformed to a random vector in $\mathbb{R}^{k-1}$, under the assumption that the covariance matrix of the random vector has full rank. Theorem 4.1 also states that Wilks' statistic for empirical likelihood evaluated at local alternatives has an asymptotic non-central chi-squared distribution with a non-centrality parameter associated with local alternatives.

As for the 3-Spider, $\mathcal{SP}_3$, Hotz et al. (2013) showed that there are three different cases, sticky case, non-sticky case and partially-sticky case, for the central limit theorem for the Fréchet mean of the 3-Spider. According to this, Theorem 5.1 states that

1. In the sticky case, Wilks' statistic for empirical likelihood evaluated at population Fréchet mean converges almost surely to 0.

2. In the non-sticky case, Wilks' statistic for empirical likelihood evaluated at population Fréchet mean has an asymptotic chi-squared distribution.

3. In the partially-sticky case, Wilks' statistic for empirical likelihood evaluated at population Fréchet mean has an asymptotic distribution which is a mixture of the sticky case and the non-sticky case.

Owen (1990) mentioned the concavity property of the log-empirical likelihood for the mean parameter. Theorem 3.2 offers full detail of the proof, and states that for random vectors $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n \in \mathbb{R}^k$ with covariance matrix of full rank, if the mean parameter lies in the relative interior of the convex hull of $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$, the log-empirical likelihood is a concave function of the mean parameter.

6.2 SIMULATION STUDY

6.2.1 *Confidence Region Construction*

One important application of Wilks' theorem is to construct the confidence regions for the mean. In Section 4.6 and Section 5.8, confidence regions are constructed respectively in $\mathcal{S}^2$ and $\mathcal{SP}_3$. In Section 4.6, we use the data of the direction of magnetisation in 34 sites at the Great Whin Sill, collected by Creer et al. (1959). In Figure 4.6.21, we outline the construction of confidence regions by the empirical likelihood method. Both chi-squared tabular quantiles and bootstrap calibration quantiles are used in evaluating the contours of the confidence regions. Compared with the confidence region estimated by Kent (1982), the outcome suggests confidence regions by empirical likelihood method are close to the result by Kent (1982). In Section 5.8, we use the metazoan phylogenetic tree data. As Figure 5.8.29 and Figure 5.8.31 show, we calculate the confidence regions by the empirical likelihood method in a non-standard way, since the cone point is the singularity point in $\mathcal{SP}_3$. In the construction of the confidence regions, both chi-squared tabular quantiles and bootstrap calibration quantiles are used for comparison.

6.2.2 *Operating Characteristics*

Another application is to investigate operating characteristics in hypothesis testing, such as probabilities of type I and II errors by empirical likelihood method, using chi-squared tabular quantiles and bootstrap calibration quantiles. Experiments in calculating the probability of type I error are conducted first both in Section 4.5 and Section 5.9, and then experiments in calculating the probability of type II error with several chosen alternative hypotheses.

As for the probability of type I error experiments in Section 4.5, we find that when sample size is smaller than 50, the probability of type I error evaluated by tabular quantile is too large. However, we normally build the hypothesis test carefully and wish the null hypothesis not to be rejected easily. That means, for the nominal probability of type I error α, it is better if the actual probability of type I error is not larger than α by too much. Even though the probability of type I error by bootstrap calibration quantile at an extreme case, sample size 10, is slightly above 0.02, it still looks much more acceptable than that by chi-squared tabular quantiles. For large sample sizes ($\geqslant 50$), there is no much difference between results by bootstrap calibration quantiles and chi-squared tabular quantiles.

For the probability of type II error at small sample sizes, since the probability of type I error by tabular quantile is too large, that makes the power of the hypothesis test less accurate. As for different choices of alternative hypotheses, the powers of the tests decrease, when the alternative hypothesis comes closer to the null hypothesis. The concentration parameter κ also influences the power. When κ gets closer to 0, the distribution looks more like an uniform distribution, which has the mean everywhere on the sphere, the type II error is more likely to happen.

In Section 5.9, we only consider conducting experiments in the non-sticky case, since we already know the behaviour of Wilks' statistic in the sticky case and the partially-sticky case is a boundary case. Observing the outcome of those experiments, we draw similar conclusions to results in Section 4.5.

6.3 FURTHER RESEARCH

We consider several possible directions for the further study of the empirical likelihood method. First, one direction is to use the intrinsic distance on $\mathcal{S}^{k-1}$ in Chapter 4. Second, we can consider local alternatives in $\mathcal{SP}_3$. The third direction is adapting Wilks' theorem on $\mathcal{SP}_3$ to a more complicated

space, say multi-dimensional open book and $\mathcal{Q}_5$; we refer to these spaces later in 6.3.3.

6.3.1 *Intrinsic Distance on $\mathcal{S}^{k-1}$*

In Chapter 4, we embed $\mathcal{S}^{k-1}$ into $\mathbb{R}^k$ first, and then use Euclidean coordinates for all points on $\mathcal{S}^{k-1}$, including local alternatives on $\mathcal{S}^{k-1}$.

Now, we can consider points on $\mathcal{S}^{k-1}$ without Euclidean coordinates. Let $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n \in \mathcal{S}^{k-1}$, and define local alternatives $\{\boldsymbol{\mu}_n\}_{n\geqslant 1}$ as

$$\boldsymbol{\mu}_n = \exp_{\boldsymbol{\mu}_0}\Big(\frac{\boldsymbol{\eta}}{\sqrt{n}}\Big),$$

where $\boldsymbol{\eta}$ is a constant vector, and let

$$\boldsymbol{Y}_i(\boldsymbol{\mu}) = \log_{\boldsymbol{\mu}} \boldsymbol{X}_i,$$

be the image of $\log_{\boldsymbol{\mu}}$-map, for $i = 1, 2, \cdots, n$, where $\boldsymbol{\mu}$ is the intrinsic mean parameter.

The proof of Theorem 4.1 shows the key problem is how to build a connection between $\boldsymbol{Y}_i(\boldsymbol{\mu}_0)$ and $\boldsymbol{Y}_i(\boldsymbol{\mu}_n)$, for $i = 1, \cdots, n$. However, in the proof of Theorem 4.1, the rotation matrix in Section 4.3.2 solves the above problem. Now, the question is how to find an approach to solve the above problem, using the intrinsic distance.

6.3.2 *Local Alternatives in $\mathcal{SP}_3$*

Theorem 3.1 and Theorem 4.1 consider the asymptotic distribution of Wilks' statistic evaluated at local alternatives. However, in Theorem 5.1, local alternatives are not considered. For example, for the population Fréchet mean $\bar{b}$ on the cone point, according to the idea of local alternatives in Section 2.3, it is possible that local alternatives $\{\bar{b}_n\}_{n\geqslant 1}$ are located off the cone point. Thus, it needs more consideration about the asymptotic behaviour of Wilks' statistic evaluated at local alternatives.

6.3.3 *Open Books and $\mathcal{Q}_5$*

In this thesis, we only consider two non-Euclidean spaces, the sphere $\mathcal{S}^{k-1}$ and the 3-Spider $\mathcal{SP}_3$.

Open Books

As is shown by Hotz et al. (2013), the 3-Spider is the simplest open book. Figure 6.3.34 explains the concept of open books.

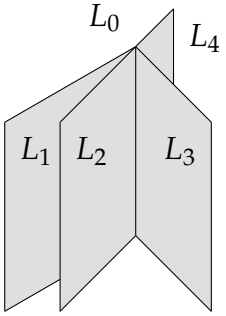


Figure 6.3.34: Open book of dimension 2 with 4 "pages", which intuitively looks like an ordinary open book. Generally, the "pages" would extend to infinity vertically, both up and down, also away from the spine on each "page". Anologous to 3-Spider, the spine refers to the cone point, and pages refer to legs in $\mathcal{SP}_3$.

Now, for a general case, let $S = \mathbb{R}^{k-1}$. The closed half-space $\overline{H}_+ = \mathbb{R}_{\geq 0} \times S$ denotes the subspace of $\mathbb{R}^k = \mathbb{R} \times S$ with boundary S.

Definition 6.3.1 (Open Books) *The open book $\mathcal{O}$ consists of at least 3 pages L_j, for $j = 1, 2, 3, \cdots$, of dimension k and defined by*

$$L_j = \overline{H}_+ \times \{j\}.$$

Let the spine L_0 be identified with the hyperplane $H = \{0\} \times S$. Thus, the open book $\mathcal{O}$ is the disjoint union

$$\mathcal{O} = L_0 \cup L_1^+ \cup \cdots \cup L_K^+,$$

where L_0 is the spine and $L_k^+ = L_k \backslash L_0$ is the interior of the page, $j = 1, 2, 3, \cdots$.

Three cases about the location of Fréchet mean also work for open books, if we only consider the first coordinate of random vectors to define the first moment as Definition 2.6.5. As for non-sticky case, we find Wilks' statistic for empirical likelihood evaluated at the population Fréchet mean has an asymptotic chi-squared distribution. However, for sticky and partially-sticky cases, more detail needs to be considered, since we not only consider the behaviour of the first coordinate but also need to study the behaviour of the rest coordinates.

$\mathcal{Q}_5$

In Chapter 5, $\mathcal{SP}_3$ is a model for phylogenetic tree data with 3 external taxa. As for the model fitting phylogenetic tree data with 4 external taxa, $\mathcal{Q}_5$ is an option in Barden et al. (2013). $\mathcal{Q}_5$ is another non-Euclidean space as a union of 5 quadrants sticking to the origin. Each quadrant is a subplane of $\mathbb{R}^2$. Figure 6.3.35 gives a concept of this space.

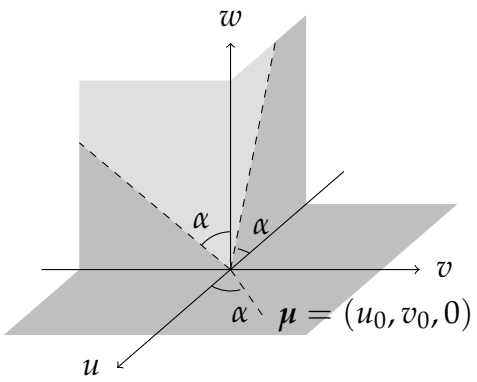


Figure 6.3.35: The space $\mathcal{Q}_5$.

According to the structure of $\mathcal{Q}_5$, we define the coordinates of points in $\mathcal{Q}_5$ in the format of $(u, v, w)^\top$. Now, we define $d_{\mathcal{Q}_5}(\cdot, \cdot)$ the distance function in $\mathcal{Q}_5$. As Figure 6.3.35 shows, we consider a point off the origin, say $\boldsymbol{\mu} = (u, v, 0)^\top$. In order to calculate the distance from $\boldsymbol{\mu}$ to any other point in $\mathcal{Q}_5$, we look for the geodesic between those two points. The dark shaded

area $\mathcal{I}_{\boldsymbol{\mu}}$ refers to the set of points whose geodesics to $\boldsymbol{\mu}$ can be directly expressed as the straight linear segments when the relevant quadrant is properly folded down into the (u, v)-plane. The light area $\mathcal{U}_{\boldsymbol{\mu}}$ refers to the set of points whose geodesics to $\boldsymbol{\mu}$ can only be expressed as a bent line through the origin, which intuitively has the same idea as the geodesics in $\mathcal{SP}_3$. In mathematics,

$$\mathcal{U}_{\boldsymbol{\mu}} = \{\boldsymbol{x} \in \mathcal{Q}_5 | d^2_{\mathcal{Q}_5}(\boldsymbol{x}, \boldsymbol{\mu}) = \{\|\boldsymbol{x}\| + \|\boldsymbol{\mu}\|\}^2\},$$

and $\mathcal{I}_{\boldsymbol{\mu}}$ is the complementary set of $\mathcal{U}_{\boldsymbol{\mu}}$ in $\mathcal{Q}_5$.

Let $\boldsymbol{X_1}, \boldsymbol{X_2}, \cdots, \boldsymbol{X_n}$ be i.i.d. random vectors in $\mathcal{Q}_5$. For $i = 1, 2, \cdots, n$, let $\boldsymbol{X}_i = (u_i, v_i, w_i)$. Now, we define

$$d^2_{\mathcal{Q}_5}(\boldsymbol{X_i}, \boldsymbol{\mu})|_{\boldsymbol{\mu}=(u,v,0)^\top \neq (0,0,0)^\top}$$
$$= \begin{cases} (u-u_i)^2 + (v-v_i)^2 & \text{if } w_i = 0. \\ (u-u_i)^2 + (v+w_i)^2 & \text{if } w_i > 0, v = 0 \text{ and } -w_i/u_i < v/u. \\ (u+w_i)^2 + (v-v_i)^2 & \text{if } w_i > 0, u = 0 \text{ and } -v_i/w_i > v/u. \\ \{\|\boldsymbol{\mu}\| + \|\boldsymbol{X_i}\|\}^2 & \text{if } \boldsymbol{X_i} \in \mathcal{U}_{\boldsymbol{\mu}}. \end{cases}$$

Then, we define the geodesic from $\boldsymbol{X_1}, \boldsymbol{X_2}, \cdots, \boldsymbol{X_n}$ to $\boldsymbol{\mu}$ as follows.

$$\boldsymbol{Y_i}(\boldsymbol{\mu}) = -\frac{1}{2}\boldsymbol{grad}(d^2_{\mathcal{Q}_5}(\boldsymbol{X_i}, \boldsymbol{\mu}))|_{\boldsymbol{\mu}=(u,v,0)\top}$$
$$= \begin{cases} (u_i - u, v_i - v) & \text{if } w_i = 0. \\ (u_i - u, -w_i - v) & \text{if } w_i > 0, v = 0 \text{ and } -w_i/u_i < v/u. \\ (-w_i - u, v_i - v) & \text{if } w_i > 0, u = 0 \text{ and } -v_i/w_i > v/u. \\ -(1 + \frac{\|\boldsymbol{X}_i\|}{\|\boldsymbol{\mu}\|})\boldsymbol{\mu} & \text{if } \boldsymbol{X_i} \in \mathcal{U}_{\boldsymbol{\mu}}. \end{cases}$$

In order to adapt Wilks' theorem to $\mathcal{Q}_5$, we need to define empirical likelihood first in $\mathcal{Q}_5$. It is expected there are two cases for empirical likelihood, the Fréchet mean parameter at and off the origin. Then, Wilks' theorem for empirical likelihood will mainly discuss the asymptotic behaviour of Wilks' statistic, when the population Fréchet mean is at and off the origin.

BIBLIOGRAPHY

G. A. Amaral, I. Dryden, and A. T. A. Wood. Pivotal Bootstrap Methods for k-Sample Problems in Directional Statistics and Shape Analysis. *Journal of the American Statistical Association*, 102(478):695–707, 2007.

D. Barden, H. Le, and M. Owen. Central Limit Theorems for Fréchet Means in the Space of Phylogenetic Trees. *Electronic Journal of Probability*, 18, 2013.

D. A. Baum and S. D. Smith. *Tree Thinking: An Introduction to Phylogenetic Biology*. Roberts Greenwood Village, CO, 2013.

R. Bhattacharya and V. Patrangenaru. Large Sample Theory of Intrinsic and Extrinsic Sample Means on Manifolds. *The Annals of Statistics*, 31(1):1–29, 2003.

R. Bhattacharya and V. Patrangenaru. Large Sample Theory of Intrinsic and Extrinsic Sample Means on Manifolds: II. *Annals of Statistics*, pages 1225–1259, 2005.

Y. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. Springer Science & Business Media, 2007.

E. Bolthausen, E. Perkins, and A. W. van der Vaart. *Lectures on Probability Theory and Statistics: Ecole D'Ete de Probabilites de Saint-Flour XXIX-1999*. Springer Science & Business Media, 2002.

W. M. Boothby. *An Introduction to Differentiable Manifolds and Riemannian Geometry*, volume 120. Academic Press, 1986.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

K. L. Chung. *A Course in Probability Theory*. Academic Press, 2001.

J. B. Conway. *A Course in Functional Analysis*, volume 96. Springer Science & Business Media, 2013.

K. Creer, E. Irving, and A. Nairn. Palaeomagnetism of the Great Whin Sill. *Geophysical Journal International*, 2(4):306–323, 1959.

M. Dresher. Games of Strategy: Theory and Applications. Technical report, RAND CORP SANTA MONICA CA, 1961.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. CRC Press, 1994.

N. I. Fisher, T. Lewis, and B. J. Embleton. *Statistical Analysis of Spherical Data*. Cambridge University Press, 1993.

N. I. Fisher, P. Hall, B.-Y. Jing, and A. T. Wood. Improved Pivotal Methods for Constructing Confidence Regions with Directional Data. *Journal of the American Statistical Association*, 91(435):1062–1070, 1996.

M. Fréchet. Les Éléments Aléatoires de Nature Quelconque dans un Espace Distancié. In *Annales de l'institut Henri Poincaré*, volume 10, pages 215–310, 1948.

B. G. Hall. *Phylogenetic Trees Made Easy: A How-to Manual*. Number 576.88 H174p. Sinauer, Associates, Inc, 2011.

P. Hall and B. La Scala. Methodology and Algorithms of Empirical Likelihood. *International Statistical Review/Revue Internationale de Statistique*, pages 109–127, 1990.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

T. Hotz, S. Huckemann, H. Le, J. S. Marron, J. C. Mattingly, E. Miller, J. Nolen, M. Owen, V. Patrangenaru, and S. Skwerer. Sticky Central Limit

Theorems on Open Books. *The Annals of Applied Probability*, 23(6):2238–2258, 2013.

H. Jeffreys, B. Jeffreys, and B. Swirles. *Methods of Mathematical Physics*. Cambridge University Press, 1999.

N. L. Johnson, S. Kotz, and N. Balakrishnan. Continuous Univariate Distributions, Vol. 1. 1994.

J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, 1980.

J. T. Kent. The Fisher-Bingham Distribution on the Sphere. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(1):71–80, 1982.

K. Mardia, J. Kent, and J. Bibby. *Multivariale Analysis*. Academic Press, 1979.

K. V. Mardia and P. E. Jupp. *Directional Statistics*, volume 494. John Wiley & Sons, 2009.

C. D. Meyer. *Matrix Analysis and Applied Linear Algebra*, volume 71. Siam, 2000.

J. Neyman. "Smooth" Test for Goodness of fit. *Skandinavisk Aktuarietiskrift*, 20:149–199, 1937.

T. M. Nye. Principal Components Analysis in the Space of Phylogenetic Trees. *The Annals of Statistics*, 39(5):2716–2739, 2011.

A. B. Owen. Empirical likelihood Ratio Confidence Intervals for a Single Functional. *Biometrika*, 75(2):237–249, 1988.

A. B. Owen. Empirical Likelihood Ratio Confidence Regions. *The Annals of Statistics*, 18(1):90–120, 1990.

A. B. Owen. *Empirical Likelihood*. Wiley Online Library, 2001.

D. Pollard. *Convergence of Stochastic Processes*. Springer Science & Business Media, 2012.

G. Stengle and J. Yukich. Some New Vapnik-Chervonenkis Classes. *The Annals of Statistics*, 17(4):1441–1446, 1989.

G. Strang. *Introduction to Linear Algebra*, volume 3. Wellesley-Cambridge Press Wellesley, MA, 1993.

N. Sugiura. Asymptotic Non-Null Distributions of the Likelihood Ratio Criteria for Covariance Matrix under Local Alternatives. *The Annals of Statistics*, 1(4):718–728, 1973.

D. R. Thomas and G. L. Grunkemeier. Confidence Interval Estimation of Survival Probabilities for Censored Data. *Journal of the American Statistical Association*, 70(352):865–871, 1975.

A. W. van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 1998.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence*. Springer, 1996.

V. N. Vapnik and A. Y. Chervonenkis. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability and Its Applications*, 16:264–280, 1971.

S. S. Wilks. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62, 1938.