Asymptotic methods of testing statistical hypotheses

by

Thuong T. M. Nguyen

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the
requirements for the degree of
Doctor of Philosophy
in Statistics.

Victoria University of Wellington
2017

Abstract

For a long time, the goodness of fit (GOF) tests have been one of the main objects of the theory of testing of statistical hypotheses. These tests possess two essential properties. Firstly, the asymptotic distribution of GOF test statistics under the null hypothesis is free from the underlying distribution within the hypothetical family. Secondly, they are of omnibus nature, which means that they are sensitive to every alternative to the null hypothesis.

GOF tests are typically based on non-linear functionals from the empirical process. The first idea to change the focus from particular functionals to the transformation of the empirical process itself into another process, which will be asymptotically distribution free, was first formulated and accomplished by **Khmaladze** [40]. Recently, the same author in consecutive papers [42] and [44] introduced another method, called here the **Khmaladze-2** transformation, which is distinct from the first Khmaladze transformation and can be used for an even wider class of hypothesis testing problems and is simpler in implementation.

This thesis shows how the approach could be used to create the asymptotically distribution free empirical process in two well-known testing problems.

The first problem is the problem of testing independence of two discrete random variables/vectors in a contingency table context. Although this problem has a long history, the use of GOF tests for it has been restricted to only one possible choice – the chi-square test and its several modifications. We start our approach by viewing the problem as one of

parametric hypothesis testing and suggest looking at the marginal distributions as parameters. The crucial difficulty is that when the dimension of the table is large, the dimension of the vector of parameters is large as well. Nevertheless, we demonstrate the efficiency of our approach and confirm by simulations the distribution free property of the new empirical process and the GOF tests based on it. The number of parameters is as big as 30. As an additional benefit, we point out some cases when the GOF tests based on the new process are more powerful than the traditional chi-square one.

The second problem is testing whether a distribution has a regularly varying tail. This problem is inspired mainly by the fact that regularly varying tail distributions play an essential role in characterization of the domain of attraction of extreme value distributions. While there are numerous studies on estimating the exponent of regular variation of the tail, using GOF tests for testing relevant distributions has appeared in few papers. We contribute to this latter aspect a construction of a class of GOF tests for testing regularly varying tail distributions.

Acknowledgments

With all gratitude, I would like to thank all people who contributed in some way to the process of completing this thesis.

First and foremost, I would like to express my sincere gratitude to my primary supervisor, Prof. Estate V. Khmaladze. I thank him for his rigorous, careful and enthusiastic supervision. I would also like to thank him for introducing me to this interesting topic, spending lots of time on training me in advanced probability and statistics and being patient with me when my progress was slow at some stage, especially after my maternity leave. Thanks are also owed to him for helping in a thorough revision of the text of this thesis. Without him, there would be no thesis. For all of his help, I am deeply grateful.

I would like to thank my secondary supervisor, Dr. Yuichi Hirose for his encouragement during my research and especially his guidance in understanding measure theory when I just started my PhD. I thank him moreover for his careful reading of this thesis and making many invaluable comments to improve the coherence of the text.

I would also like to express my great appreciation to the three examiners, Prof. Spiridon Penev, Prof. B. L. S. Prakasa Rao and Dr. Leigh Roberts for their encouragement as well as critical comments which improve the text tremendously.

I just started learning programming in R since it is necessary for doing research, and for this aspect, I am indebted to Dr. Nokuthaba Sibanda for her first introduction to R; to Joel Bancolita and Boyd Anderson for lots of useful discussions in programming.

I would also like to thank other academics and staff in the School of Mathematics and Statistics for their support and help whenever needed. Especially, I would like to thank A/Prof. Ivy Liu, A/Prof. Stefanka Chukova and Dr. Laura Dumitrescu for many pleasant and supportive conversations; and A/Prof. Richard Arnold for being so generous with his time when I was teaching assistant and tutor for the course STAT193.

I would like to acknowledge the financial support from Victoria University of Wellington through a Doctoral scholarship.

I would also like to acknowledge all my office mates in CO547, including Dr. Kemmawadee Preedalikit (Fa), Dr. Daniel Fernandze, Dr. Darcy Webber, Roy Costillia, Yuki Fujita and Doaa Aryad for their encouragement and for keeping me in good company and in a fresh working environment. I gratefully thank my friends Geoff Osborne and Roy Costillia for checking and improving English in the text. For the final proofreading, I am deeply indebted to William Roberts, who read every sentence of the text carefully.

A special thank to my dear friend Thanh Nguyen who gave me a lot of caring and support and always believed that I can overcome any difficulty. Her friendship is truly an invaluable gift.

I owe my deepest gratitude to my mother Dung and my father Thanh for their unconditional love and support, for giving me the every best opportunity in education that they could. I have received further love and support from my parents in law, Hai and Chau, which has also greatly motivated me. Additionally, thanks to my brothers and sisters Thao, Tinh and Phuong for always encouraging and believing in me.

Last but not least, I would like to thank the love of my life Hung and my little treasure An. I thank them for every single moment since I had

them. Their love is a crucial part of my life and I could not imagine my life without their love.

Contents

List of Figures

Glossary

CLT	central limit theorem
GOF	goodness of fit
MLE	maximum likelihood estimator
SLLN	strong law of large numbers
$\mathscr{A}$	a σ-algebra
A	a set in $\mathscr{A}$, i.e. an $\mathscr{A}$-measurable set
F, G, H, K	distribution functions
$F_0, F_{\boldsymbol{\theta}_0}$	hypothetical distributions
$F_a, F_{a,n}$	contiguous distributions
F_n, H_m	empirical distributions
$\mathscr{F}$	parametric family of distributions
I	indicator function
$\mathbb{I}_I, \mathbb{I}_J$	identity matrices of size $I \times I, J \times J$
$\theta, \boldsymbol{\theta}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_n$	parameters
$\Gamma, \Gamma_{\boldsymbol{\theta}}, \Gamma_H, \Gamma_G$	Fisher information/information matrix
β, β_H, β_G	normalized score functions
$\mathscr{K}[v_{nF_0}, F_0]$	transformation of v_{nF_0} and F_0
$\widehat{\mathscr{K}}[\widehat{v}_{nF_0,\hat{\boldsymbol{\theta}}_n}, \mathscr{F}]$	transformation of $\widehat{v}_{nF_0,\hat{\boldsymbol{\theta}}_n}$ and $\mathscr{F}$
v_{nF}, v_{mH}, v_{mG}	empirical process
$\widehat{v}_{nF}, \widehat{v}_{mH}$	parametric empirical process
V_F	Brownian bridge in time F
$\widehat{V}_F$	projected F-Brownian motion
W_F	Brownian motion in time F
$\xrightarrow{\mathbb{P}}$	converges in probability
$\xrightarrow{d}$	converges in distribution
$\xrightarrow{d}_F$	converges in distribution under hypothesis F
$\langle \cdot, \cdot \rangle_F$	inner product in $\mathcal{L}_2(F)$
$\Vert\cdot\Vert_F$	norm in $\mathcal{L}_2(F)$
$d(P, Q)$	distance in total variation of measures P and Q
$\mathbb{1}$	function identically equals to 1
$U_{\beta,r}, \mathbb{U}, \widehat{\mathbb{U}}$	unitary operators

Chapter 1

Introduction

1.1 Introduction

Statistical inference, a process of deducing properties of the underlying distribution F of a population by analysis of data, includes two main procedures: deriving estimates and testing hypotheses.

There are two types of hypothesis testing problems: **simple hypothesis testing** and **composite hypothesis testing**. For simple hypothesis testing, the null hypothesis is of the form $F = F_0$ for some specified distribution F_0. In addition, the alternative can be either $F = F_a$ for some particular F_a different from F_0 or just $F \neq F_0$. The simple hypothesis testing does not involve any **parameter** which labels the underlying distributions. The composite parametric hypothesis testing problem involves the parametric family

$$\mathscr{F} = \{F_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\},$$

where Θ is some set, called the **parameter space**. Then the statement

$$H_0 : F \in \mathscr{F}$$

is taken as the null hypothesis. In this case, the parameter $\boldsymbol{\theta}$ which identifies the null distribution also needs to be estimated from the given observations.

To carry out a statistical hypothesis testing problem, the decision on either to **accept** or **reject** a null hypothesis in favour of the alternative must be made based on some **test statistic**. Consider the **empirical distribution function**, which is constructed from the sample $X_1, \ldots, X_n$, as follows:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq x\}},$$

where $I_{\{X_i \leq x\}}$ are indicator functions, equal to 1 if $X_i \leq x$ and 0 otherwise. Note that if X_i are vectors in $\mathbb{R}^k$, $X_i = (X_i^{(1)}, \ldots, X_i^{(k)})$ then $X_i \leq x$ means $X_i^{(j)} \leq x^{(j)}$ for every $j = 1, \ldots, k$. Naturally, a test statistic has to involve the empirical distribution and, probably, the underlying unknown distribution F. The following:

$$\begin{aligned} &\sqrt{n} \int x[dF_n(x) - dF(x)], && \sup_{x \in \mathbb{R}^k} [F_n(x) - F(x)], \\ &\int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 dx, && \max \frac{\sqrt{n}[F_n(x) - F(x)]}{\sqrt{F(x)[1 - F(x)]}}, \end{aligned}$$

are examples of test statistics.

In 1993, **Khmaladze** [40] formulated what he called the goodness of fit (GOF) problem of testing hypotheses. In this formulation, two properties of GOF tests were considered together - its omnibus nature and its distribution free property. Obviously, not every test statistic posses these two characteristics at the same time. Those examples listed above are either not asymptotically distribution free or not of an omnibus nature.

In testing simple hypotheses for 1-dimensional continuous distributions, we have a class of distribution free GOF tests based on the **empirical process**

$$v_{nF}(x) = \sqrt{n}[F_n(x) - F(x)].$$

This class originated from the idea of the time transformation $t = F(x)$, which was first suggested by **Kolmogorov** [46] in 1933. However, this time transformation is invalid for discrete distributions as well as multidimensional distributions. For testing parametric problems, the procedure involves the parametric empirical process

$$\widehat{v}_{nF}(x) = \sqrt{n}[F_n(x) - F_{\hat{\boldsymbol{\theta}}_n}(x)]$$

where $\hat{\boldsymbol{\theta}}_n$ is an estimate of the hypothetical unknown parameter $\boldsymbol{\theta}$. The time transformation $t = F_{\hat{\boldsymbol{\theta}}_n}(x)$ does not lead to the distribution free property of the transformed process. Thus, a transformation which plays the same role as the time transformation but works for a wider class of hypothesis testing problems might have been set as a goal.

The first such transformation was considered in **Khmaladze** [38], [39], [40]. The method that the author invented works for both simple and parametric testing problems for finite-dimensional continuous distributions. The proposed method was subsequently known as the **Khmaladze transformation**. To briefly review this transformation, without losing any statistical information, it turns the empirical process v_{nF} or $\widehat{v}_{nF}$ into another process which converges in distribution to the standard Brownian motion. As a result, we are able to construct a whole class of asymptotically distribution free GOF tests on the transformed process.

Later, in a 2013 paper [42] and 2016 [44], another transformation was introduced by the same author, so let us call this later transformation **Khmaladze-2**. This transformation enables us to create a class of GOF tests for any statistical hypothesis testing problem, including problems with discrete distributions. We apply the transformation Khmaladze-2 to two problems, and that is the main content of this thesis.

1.2 Aim of the thesis

The main aim of this thesis is to show a construction of a class of GOF tests for two different hypothesis testing problems. The first problem is testing independence of two discrete random vectors/variables in the contingency table context. The other is testing regularly varying tail distributions. Ostensibly, these two problems are not connected in any statistical or mathematical sense. The former is on the class of discrete distributions and the latter is on the tails of continuous distributions. However, as mentioned, for each problem, we can build up a class of GOF tests by the same method, which is adopted from **Khmaladze** [42], [44]. Namely, the new class of GOF tests is created based on the Khmaladze-2 transformation of the empirical process. The main idea of this transformation, similar to the previous Khmaladze transformation, is not to transform test by test, from one form to another, but to transform the underlying empirical process, as a base of the test, into another version of the empirical process, which is asymptotically distribution free. Hence, any statistic based on this transformed process will be asymptotically distribution free.

In each problem, our approach is not a mere application of the Khmaladze-2 method but has its own merits and resolves its own difficulties.

For testing the independence of two random vectors in contingency tables, we view this problem as a parametric/composite hypothesis testing problem. The dimension of the parameters depends on the dimension of the table, so it can be relatively large. It was not clear in **Khmaladze** [42] whether his method would work for such large number of parameters. We will show that the method works not only reliably but also quickly. Moreover, we point out the cases where the statistical power of the new tests is better than the conventional chi-square test, which has long been the only GOF test used for this problem.

The problem of testing regularly varying tail distributions also belongs to the class of parametric hypothesis testing. However, we are not formulating any hypothesis on the underlying distribution F but only on its right tail behaviour. Consequently, not all observations are considered but only the observed values which are greater than some chosen threshold.

We construct the tail empirical process with a minor change in variable and then transform it.

We expect that this thesis can be used as a note for students studying statistics at a 400 level who are interested in empirical processes and hypothesis testing. Therefore, the structure of the thesis is organised in the way outlined below. We sometimes add some examples and material relating to the main content of the thesis in some way.

1.3 Outline of the thesis

The main content of this thesis lies in the last two chapters: Chapters 5 and 6.

In Chapter 2, we collect some basic mathematical and statistical background which helps in understanding the content of the thesis. This fundamental material includes projections, orthonormal systems, unitary operators, contiguity, regular families of distributions, maximum likelihood estimators, etc.

We present in Chapter 3 a construction of the empirical process, the tail empirical process in the case of examining the right tail of distributions and the parametric empirical process. The very rich literature of the empirical process will not be discussed; we mainly sketch the idea behind its construction and mention its limit in distribution. This presentation aims at explaining how to look at the empirical process as some kind of projection of a Brownian motion.

We devote Chapter 4 to GOF tests. To begin with, we will discuss the optimal test for the simple hypothesis testing problem where the alternative is contiguous to the null hypothesis and this contiguity is identified by a certain direction. The role of the GOF tests arises in the case when the alternative approaches the null from infinitely many possible directions. We will give some well-known GOF tests as examples and restate the formulation of the GOF testing problem.

Chapter 5 presents the method of constructing a class of GOF tests for testing independence in a contingency table, or in other words, testing independence of two random variables/vectors. Some simulation results will be given. We show in detail the limiting distribution free property of the new GOF test statistics with a large sample size. The fact that the distributions of the new tests do not depend on the parameters will also be demonstrated. Moreover, we compare the statistical power of the new tests with the conventional chi-square test for several different alternatives. We point out the case when the new tests are more powerful than the chi-square test.

Another construction of a class of GOF tests for testing distributions whose tails are regularly varying will be given in Chapter 6. In this chapter, we will also briefly review some research interests regarding regularly varying tail distributions due to its great importance in both probabilistic theory and real-life applications. Once the method of creating a class of GOF tests is established, we show some simulation results to demonstrate the distribution free property of the new tests in this specific problem.

Chapter 2

Preliminaries

In this chapter we collect most of the definitions and notations used in the rest of this thesis. The material is very basic in functional analysis and probability theory.

2.1 Limit of a sequence

2.1.1 Definitions

Definition 2.1. (Convergence in probability to 0)

A sequence of random variables $\{Z_n, n \geq 1\}$ **converges in probability to** 0 if as $n \to \infty$,

$$\mathbb{P}(|Z_n| \geq \epsilon) \to 0 \quad \text{for all } \epsilon > 0,$$

and we denote this by $Z_n \xrightarrow{\mathbb{P}} 0$.

Definition 2.2. (Convergence in probability)

A sequence of random variables $\{X_n, n \geq 1\}$ **converges in probability** to a random variable X if

$$X_n - X \xrightarrow{\mathbb{P}} 0,$$

and we denote this by $X_n \xrightarrow{\mathbb{P}} X$.

Denote by

$$F(x) = \mathbb{P}[X \leq x]$$

the **distribution function** of random variable X and denote by

$$C_F = \{x : F(x) \text{ is continous at } x\}$$

the set of continuity points of F.

Definition 2.3. (Convergence in distribution)
A sequence of random variables $\{X_n, n \geq 1\}$ **converges in distribution** to a random variable X if

$$\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x) \text{ for all } x \in C_F.$$

and we denote this by $X_n \xrightarrow{d} X$.

In this sense, if we denote by F_n the distribution function of the random variable X_n, then we say that the sequence of distributions $\{F_n, n \geq 1\}$ **converges weakly** to the distribution F and denote this by

$$F_n \xrightarrow{w} F.$$

Random variables X_n themselves do not necessarily converge to anything; only their distributions do.

2.1.2 o, O, o_P, O_P notations

Definition 2.4. Let $\{a_n\}$ and $\{b_n\}$ be two sequences of numbers in $\mathbb{R}$.

(i) $a_n = O(1)$ if $|a_n| \leq C$ for some constant C;

(ii) $a_n = o(1)$ if $|a_n| \to 0$ as $n \to \infty$;

(iii) $a_n = O(b_n)$ if $\left|\frac{a_n}{b_n}\right| \leq C$ for some constant C or $\frac{a_n}{b_n} = O(1)$;

(iv) $a_n = o(b_n)$ if $\left|\frac{a_n}{b_n}\right| \to 0$ as $n \to 0$ or $\frac{a_n}{b_n} = o(1)$.

Definition 2.5. Let $\{X_n\}$ be a sequence of random variables and $\{a_n\}$ be a sequence of non-negative real numbers.

(i) $X_n = O_P(1)$ if for any $\epsilon > 0$, there exists some $M > 0$ such that

$$\mathbb{P}\left\{|X_n| \geq M\right\} \leq \epsilon;$$

(ii) $X_n = o_P(1)$ if $X_n \overset{\mathbb{P}}{\to} 0$;

(iii) $X_n = O_P(a_n)$ if $\frac{X_n}{a_n} = O_P(1)$.

(iv) $X_n = o_P(a_n)$ if $\frac{X_n}{a_n} \overset{\mathbb{P}}{\to} 0$;

2.2 Hilbert spaces, projections and unitary operators

We briefly sketch in this section the basic theory of Hilbert spaces, their orthogonal projections and unitary operators, which are essential for understanding the method presented in this thesis. Specifically, we are mainly concerned with $\mathcal{L}_2(F)$, a specific construction of a Hilbert space, since this space is of great importance to the discussion in Chapter 6.

A thorough discussion on Hilbert spaces can be seen for example in **Debnath and Mikusinski** [12].

2.2.1 Space $\mathcal{L}_2(F)$ as a Hilbert space

Let F be a distribution function.

Definition 2.6. The space $\mathcal{L}_2(F)$ consists of all real-valued functions g such that

$$\int g^2(x)dF(x) < \infty.$$

Because of the above inequality, if we define

$$\|g\|_F = \sqrt{\int g^2(x)dF(x)},$$

then $\|\cdot\|_F$ is a function: $\mathcal{L}_2(F) \to [0, \infty)$. This function satisfies the following properties:

(i) $\|g\|_F \geq 0,\ \forall g \in \mathcal{L}_2(F)$ and $\|g\|_F = 0$ if and only if $g = 0$[1];

(ii) $\|g + h\|_F \leq \|g\|_F + \|h\|_F$ for every $g, h \in \mathcal{L}_2(F)$;

(iii) $\|cg\|_F = |c|\,\|g\|_F$ for every $c \in \mathbb{R}, g \in \mathcal{L}_2(F)$.

Conditions (i) and (iii) are easy to verify and condition (ii) follows from the well known Cauchy-Schwarz-Bunyakovsky inequality.

A function with the above properties is called a **norm** and a linear space with a norm is called a **normed space**.

The normed space $(\mathcal{L}_2(F), \|\cdot\|_F)$ is moreover **complete** in the sense that every Cauchy sequence of functions $(g_n)_{n\geq 1}$ in $\mathcal{L}_2(F)$ converges to a function g in $\mathcal{L}_2(F)$. Recall that a Cauchy sequence $(g_n)_{n\geq 1}$ is a sequence satisfying the condition that for every $\epsilon > 0$ there exists a number M such that

$$\|g_n - g_m\| < \epsilon \text{ for all } m, n > M.$$

In addition, if we define an **inner product** $\langle \cdot, \cdot \rangle_F$ as

$$\langle g, h \rangle_F = \int g(x)h(x)dF(x) \quad \text{for every } g, h \in \mathcal{L}_2(F),$$

then we have the following formula

$$\|g\|_F = \sqrt{\langle g, g \rangle_F}.$$

That is, $\|\cdot\|_F$ is the norm associated with the inner product $\langle \cdot, \cdot \rangle_F$.
Recall that $\langle \cdot, \cdot \rangle_F$ as an inner product must satisfy the following conditions:

[1] To be precise, $\|g\|_F = 0$ if and only if $g(x) = 0$ for almost every x with respect to the distribution function F. But we will not pay too much attention to this, as we will consider such a function as 0.

(i) $\langle g,h\rangle_F = \langle h,g\rangle_F$;

(ii) $\langle g+h,\ell\rangle_F = \langle g,\ell\rangle_F + \langle h,\ell\rangle_F$;

(iii) $\langle cg,h\rangle_F = c\langle g,h\rangle_F$;

(iv) $\langle g,g\rangle_F \geq 0$ and $\langle g,g\rangle_F = 0$ implies $g = 0$ for every x.

These conditions can be checked easily. Note that condition (iv) here is the same as condition (i) for the norm.

Thus, $\mathcal{L}_2(F)$ is an inner product space, whose norm is complete and so by definition, it is a **Hilbert space**. We will also need some other specific Hilbert spaces, which are not of as much use for us as $\mathcal{L}_2(F)$. They will be introduced in due course.

Note that from now on, in a general inner product space H, the inner product will be written as $\langle\cdot,\cdot\rangle$. Otherwise, we will always use the subindex F for inner products in $\mathcal{L}_2(F)$ whenever F is a distribution function.

In $\mathcal{L}_2(F)$ or more generally, in any Hilbert space H, the following concepts may be defined.

2.2.2 Orthogonality and Projections

Two vectors x and y in H are called **orthogonal**, denoted by $x \perp y$, if

$$\langle x,y\rangle = 0.$$

Let S be a non-empty subset of H. An element $x \in H$ is said to be **orthogonal** to S, denoted by $x \perp S$, if $x \perp y$ for every $y \in S$.

The set of all elements of H orthogonal to S, denoted by $S^{\perp}$, is called the **orthogonal complement** of S, i.e.,

$$S^{\perp} = \{x \in H : x \perp S\}.$$

$S^\perp$ is always a closed subspace of H. Furthermore, if S is a closed subspace of H, we have $(S^\perp)^\perp = S$ and more importantly,

$$H = S + S^\perp = \left\{y + z, y \in S, z \in S^\perp\right\}.$$

Moreover, each $x \in H$ can be expressed uniquely as

$$x = y + z, \quad y \in S, z \in S^\perp.$$

This shows that the following definition is well-defined.

Definition 2.7. (Orthogonal projection)

Let S be a closed subspace of a Hilbert space H. The operator P on H defined by

$$Px = y \text{ if } x = y + z,\, y \in S,\, z \in S^\perp, \tag{2.1}$$

is called the **orthogonal projection** onto S. The vector y is called the **projection** of x onto S.

Any orthogonal projection P is a **bounded linear operator**, where a linear mapping $L : H \to H$ is called **bounded** if there exists a number α such that

$$\|Lx\| \leq \alpha \|x\| \text{ for all } x \in H.$$

The smallest such number α is called the norm of L and we write $\|L\| = \alpha$. In the case of the orthogonal projection P, either $\|P\| = 1$ or $P = 0$. Indeed, for any $x = y + z$ as in (2.1), we have

$$\|Px\| = \|y\|$$

and

$$\|x\|^2 = \|y\|^2 + \|z\|^2$$

because $y \perp z$.

Note that every projection is an **idempotent** in the sense that $P^2 = P$. This is easy to check by the definition of the orthogonal projection: if P is a projection, then $Px \in S$ for every $x \in H$ and so $Px = Px + 0$ is the expression for Px as in (2.1). Therefore,

$$P^2(x) = P(Px) = P(x).$$

2.2.3 Orthonormal systems

A family S of non-zero vectors in a Hilbert space H is called an **orthogonal system** if $x \perp y$ for any two distinct elements of S. If, in addition, $\|x\| = 1$ for all $x \in S$, then S is called an **orthonormal system**.

A sequence of vectors which constitutes an orthonormal system is called an **orthonormal sequence**.

An orthonormal sequence S becomes an **orthonormal basis** if $x \perp S$ implies $x = 0$.

If $\{e_1, e_2, \dots\}$ is an orthonormal basis for H, then any $x \in H$ can be written as

$$x = \sum_n \alpha_n e_n$$

where

$$\alpha_n = \langle x, e_n \rangle.$$

Below are some examples of orthonormal sequences.

Example 2.1. (Haar system)

Consider the Hilbert space $\mathcal{L}^2[0,1]$, the space of all square integrable functions on the interval $[0,1]$ with respect to the Lebesgue measure. In fact, this space is nothing but $\mathcal{L}_2(F)$ where F is the uniform distribution function on $[0,1]$.

First, take $\psi_0(t) = 1$; then trivially $\|\psi_0\| = 1$. Take $\psi_1(t)$ to be a simple function that has two values 1 and -1 alternatively on the intervals $[0,1/2)$ and $[1/2,1)$; there are two choices of ψ_1.

More generally, for each $n \in \mathbb{N}$, take $\psi_n(t)$ to be a simple function that has two values 1 and -1 alternately on each pair of intervals $[2k/2^n, (2k+1)/2^n)$ and $[(2k+1)/2^n, 2(k+1)/2^n)$ with $k = 0, \dots, 2^{n-1} - 1$; there are 2^n choices of ψ_n.

If in each stage n of division, we choose only one $\psi_n(t)$ then the sequence $(\psi_n)_{n \geq 1}$ forms an orthonormal sequence on $\mathcal{L}_2([0,1])$.

Example 2.2. There is a general process to construct an orthonormal sequence in any Hilbert space H, called **Gram-Schmidt orthonormalization process**.

Assume that $(y_n)_{n\geq 1}$ is a sequence of linearly independent vectors on H. Define two sequences $(w_n)_{n\geq 1}$ and $(x_n)_{n\geq 1}$ inductively by

$$\begin{aligned} w_1 &= y_1, & x_1 &= \frac{w_1}{\|w_1\|}, \\ w_k &= y_k - \sum_{i=1}^{k-1} \langle y_k, x_i \rangle x_i, & x_k &= \frac{w_k}{\|w_k\|}, \quad \text{for } k = 2, 3, \ldots \end{aligned}$$

Intuitively, x_1 is obtained by normalizing y_1. To obtain x_2, we normalize the orthogonal component of y_2 when projecting it onto the subspace generated by x_1. Generally, x_k is obtained by normalizing the orthogonal component of y_k when projecting it onto the subspace H_k generated by $\{x_1, \ldots, x_{k-1}\}$, which is the same as the subspace generated by $\{y_1, \ldots, y_{k-1}\}$. Thus, in fact the iteration allows a quick computation of the auxiliary vector w_k.

2.2.4 Unitary operators

Definition 2.8. (Unitary operator)

A linear operator $U : H \to H$ is called a **unitary operator** on H if the following holds:

(i) U is a surjective/ onto mapping, i.e., $\forall y \in H : \exists x \in H : Ux = y$;

(ii) U preserves inner products of the Hilbert space, i.e., for every $x, y \in H$, we have
$$\langle Ux, Uy \rangle = \langle x, y \rangle.$$

It follows from (ii) that a unitary operator U is isometric, i.e.,

$$\|Ux\| = \|x\|, \ \forall x \in H.$$

In fact, these two conditions are equivalent. In particular, U is bounded with $\|U\| = 1$.

Trivial examples of unitary operators include identity operators, orthogonal matrices (i.e., such that $A^T A = AA^T = I$ where I is the identity matrix) on finite-dimensional Hilbert space.

For any unitary operator $U : H \to H$, if $\{e_1, e_2, \dots\}$ is an orthonormal sequence/ basis in H, then so too is the sequence $\{Ue_1, Ue_2, \dots\}$. Conversely, given two orthonormal bases $\{e_1, e_2, \dots\}$ and $\{f_1, f_2, \dots\}$ in H, the following mapping

$$e_n \mapsto f_n \qquad (n \in \mathbb{N})$$

extends to a unique unitary operator $U : H \to H$.

It is easy to prove the following proposition.

Proposition 2.1. *The product of a sequence of unitary operators is again a unitary operator.*

2.3 Contiguity

For the content of this section, we refer to **Oosterhoof and van Zwet** [60] and **van der Vaart** [77], Chapter 6.

Let P and Q be two probability measures on a probability space $(\Omega, \mathscr{A})$. The **distance in total variation** between P and Q is defined by

$$d(P, Q) = \sup_{A \in \mathscr{A}} |P(A) - Q(A)| .$$

We say that measure Q is **absolutely continuous** with respect to P if $P(A) = 0$ implies $Q(A) = 0$ for every measurable set A in $\mathscr{A}$. This is denoted by $Q \ll P$.

If $Q \ll P$ and $P \ll Q$, then we say that P and Q are **mutually absolutely continuous**.

If Q is absolutely continuous with respect to P then Q has a density with respect to P, denoted by $f = dQ/dP$. This density f is usually called the **Radon-Nikodym derivative**. For any measurable set A we have

$$Q(A) = \int_A f dP.$$

That means we are able to reconstruct the measure Q from the measure P and the density f.

We say that measures P and Q are **mutually singular** or **orthogonal**, denoted $P \perp Q$, if the space Ω can be partitioned as $\Omega = \Omega_P \cup \Omega_Q$ with $\Omega_P \cap \Omega_Q = \emptyset$ and $P(\Omega_Q) = Q(\Omega_P) = 0$.

Lemma 2.2. *(Lebesgue decomposition)*

Given a measure P, each measure Q has a unique decomposition of the form $Q = Q^c + Q^\perp$ where $Q^c \ll P$ and $Q^\perp \perp P$. This is called the **Lebesgue decomposition** *of Q with respect to P.*

The concept of contiguity is a natural extension of absolutely continuity of two measures in the case of a sequence of pairs of measures.

Let $(\Omega_n, \mathscr{A}_n)_{n\geq 1}$ be a sequence of measurable spaces and each space, corresponding to n, equipped with two probability measures P_n and Q_n.

The sequence $(Q_n)_{n\geq 1}$ is **contiguous** with respect to the sequence $(P_n)_{n\geq 1}$ if $P_n(A_n) \to 0$ implies $Q_n(A_n) \to 0$ as $n \to \infty$ for every sequence of measurable sets $A_n \in \mathscr{A}_n$. This is denoted by $Q_n \lhd P_n$.

The sequences P_n and Q_n are said to be **mutually contiguous** if $Q_n \lhd P_n$ and $P_n \lhd Q_n$ and we denote $P_n \lhd\rhd Q_n$.

Assume that only Q_n depends on n and P is fixed, i.e., $P_n = P$ for every n. For the sake of clarity, let us assume that all Q_n and P are given on the same sample space $(\Omega, \mathscr{A})$: for example, they are given on the real line, $\Omega = \mathbb{R}$. Let Q_n^n and P^n be direct products of measures Q_n and P

respectively, i.e., on rectangular sets $A_1 \times A_2 \times \cdots \times A_n$,

$$Q_n^n(A_1 \times A_2 \times \cdots \times A_n) = Q_n(A_1) \times Q_n(A_2) \times \cdots \times Q_n(A_n),$$
$$P^n(A_1 \times A_2 \times \cdots \times A_n) = P(A_1) \times P(A_2) \times \cdots \times P(A_n),$$

for every $A_i \in \mathscr{A}, i = 1, \dots, n$.

We would like to consider the condition when the direct product Q_n^n is contiguous with respect to P^n, $Q_n^n \lhd P^n$. Suppose also that there is a measure μ on the space $(\Omega, \mathscr{A})$ such that all Q_n and P have densities with respect to μ. Denote these densities by $q_n = dQ_n/d\mu, p = dP/d\mu$.

Note that such a μ is said to be a σ-finite measure dominating the sample space and that μ always exists. In specific cases, say, if P and Q_n are absolutely continuous distributions, then we can choose μ to be the Lebesgue measure; or if P and Q_n are discrete distributions, then we can choose μ to be the counting measure. But in general, we could always construct μ as follows

$$\mu(A) = \frac{P(A)}{2} + \frac{Q_1(A)}{4} + \frac{Q_2(A)}{8} + \dots, \text{ for all } A \in \mathscr{A}.$$

It is intuitively clear that as $n \to \infty$, a necessary condition for $Q_n^n \lhd P^n$ is that Q_n converges to P weakly. To quantify this convergence, we introduce the Hellinger distance between two probability measures Q_n and P, which is

$$\mathcal{H}_n(Q_n, P) = \Big(\int (\sqrt{q_n} - \sqrt{p})^2 d\mu \Big)^{1/2}.$$

From **Oosterhoof and van Zwet** [60], we know that in order to have $Q_n^n \lhd P^n$, the following condition must be satisfied

$$\limsup_{n\to\infty} n\mathcal{H}_n^2(Q_n, P) = \limsup_{n\to\infty} n \int (\sqrt{q_n} - \sqrt{p})^2 d\mu < \infty.$$

Let us represent q_n as

$$\sqrt{q_n(x)} = \sqrt{p(x)}[1 + \epsilon_n h_n(x)],$$

for some small ϵ_n and some function $h_n(\cdot)$. Then

$$\limsup_{n\to\infty} n\mathcal{H}_n^2(Q_n, P) = \limsup_{n\to\infty} n\epsilon_n^2 \int h_n^2(x)dP(x) < \infty.$$

For this to be true, $h_n(\cdot)$ should be bounded in the sense that

$$\limsup_{n} \int h_n^2(x)dP(x) < \infty.$$

The class of sequences of $h_n(\cdot), n = 1, 2, \ldots,$ which satisfies that condition is evidently large, so most of the time the following restriction is used in asymptotic statistics:

$$\int [h_n(x) - h(x)]^2 dP(x) \to 0 \qquad \text{as } n \to \infty,$$

for some h satisfying

$$\int h^2(x)dP(x) = c < \infty.$$

The function h can be viewed as the "direction" in which measures Q_n converge to P. At the same time, $n\epsilon_n^2$ must be bounded, which implies $\epsilon_n = O(1/\sqrt{n})$.

The contiguity of two probability measures will be discussed further in Section 4.1.1 where contiguous alternatives are defined in Definition 4.1.

2.4 Regular parametric family of distributions

Consider a parametric family of distributions

$$\mathscr{F} = \left\{F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d\right\}.$$

Denote by $f_{\boldsymbol{\theta}}(x)$ the density function corresponding to distribution function $F_{\boldsymbol{\theta}}(x)$. Even though $f_{\boldsymbol{\theta}}(x)$ or $F_{\boldsymbol{\theta}}(x)$ is a function of both x and the parameter $\boldsymbol{\theta}$, to distinguish between variables and parameters, we will write throughout the thesis $\boldsymbol{\theta}$ as a sub-index of F (or f) to indicate that the distribution (or density) function belongs to a parametric family. We also always denote by $\boldsymbol{\theta}_0$ the true unknown parameter. Suppose that the density function $f_{\boldsymbol{\theta}}(x)$ is differentiable with respect to $\boldsymbol{\theta}$ and the derivative denoted by $\dot{f}_{\boldsymbol{\theta}}(x)$. The differentiability of the density function is one of the usual conditions for the family of distributions $\mathscr{F}$ to be **regular**. A list of such conditions can be seen in **Lehmann** [51].

The **log-likelihood function** $l_{\boldsymbol{\theta}}(x)$ is defined as

$$l_{\boldsymbol{\theta}}(x) = \log f_{\boldsymbol{\theta}}(x).$$

The **score function** $\dot{l}_{\boldsymbol{\theta}}(x)$ is defined as

$$\dot{l}_{\boldsymbol{\theta}}(x) = \frac{\partial l_{\boldsymbol{\theta}}(x)}{\partial \boldsymbol{\theta}} = \frac{\dot{f}_{\boldsymbol{\theta}}(x)}{f_{\boldsymbol{\theta}}(x)}.$$

The **Fisher information matrix** Γ is defined as

$$\Gamma_{\boldsymbol{\theta}} = \mathsf{Var}(\dot{l}_{\boldsymbol{\theta}}) = \mathsf{E}(\dot{l}_{\boldsymbol{\theta}}\dot{l}_{\boldsymbol{\theta}}^T) = \int \dot{l}_{\boldsymbol{\theta}}(y)\dot{l}_{\boldsymbol{\theta}}^T(y)F_{\boldsymbol{\theta}}(dy). \tag{2.2}$$

From the definitions of the score function and the Fisher information matrix, we have the **normalized score function**, denoted by $\beta(x)$ as follows:

$$\beta(x) = \Gamma_{\boldsymbol{\theta}}^{-1/2}\dot{l}_{\boldsymbol{\theta}}(x) = \Gamma_{\boldsymbol{\theta}}^{-1/2}\frac{\dot{f}_{\boldsymbol{\theta}}(x)}{f_{\boldsymbol{\theta}}(x)}. \tag{2.3}$$

This function is of unit norm in $\mathcal{L}_2(F)$. Moreover, if we denote by $\mathbb{1}$ the function that identically equals 1, it is easy to see that $\beta \perp \mathbb{1}$ since

$$\langle \beta, \mathbb{1}\rangle_F = \Gamma_{\boldsymbol{\theta}}^{-1/2}\int \frac{\dot{f}_{\boldsymbol{\theta}}(y)}{f_{\boldsymbol{\theta}}(y)}F_{\boldsymbol{\theta}}(dy) = 0.$$

Note that this equation holds with some usual additional conditions on the density function $f_{\boldsymbol{\theta}}(x)$.

2.5 Maximum likelihood estimator

Suppose that we have a random sample $X_1, \dots, X_n$. The log-likelihood function defined on sample $X_1, \dots, X_n$ is

$$L_{\boldsymbol{\theta}}(X_1, \dots, X_n) = \sum_{i=1}^{n} l_{\boldsymbol{\theta}}(X_i) = \sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(X_i).$$

The most well-known estimator of $\boldsymbol{\theta}$ is the **maximum likelihood estimator** (MLE), defined as the estimator $\hat{\boldsymbol{\theta}}_n$ that maximizes the log-likelihood function $L_{\boldsymbol{\theta}}(\mathbf{X})$. In other words, it is a proper solution to the equation

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{l}_{\boldsymbol{\theta}}(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\dot{f}_{\boldsymbol{\theta}}(X_i)}{f_{\boldsymbol{\theta}}(X_i)} = 0.$$

The following theorem gives us the most important properties of the MLE.

Theorem 2.3. *Under usual regularity conditions, the MLE $\hat{\boldsymbol{\theta}}_n$ is consistent in the sense that it converges in probability to the true unknown parameter $\boldsymbol{\theta}_0$ and we write $\hat{\boldsymbol{\theta}}_n \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_0$. In addition, the following asymptotically linear representation of $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ is true:*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}} \Gamma_{\boldsymbol{\theta}_0}^{-1} \sum_{i=1}^{n} \frac{\dot{f}_{\boldsymbol{\theta}_0}(X_i)}{f_{\boldsymbol{\theta}_0}(X_i)} + o_P(1). \tag{2.4}$$

Moreover,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \Gamma_{\boldsymbol{\theta}_0}^{-1}),$$

where the Fisher information $\Gamma_{\boldsymbol{\theta}_0}$ is defined as in (2.2).

The term "asymptotically linear" will be explained below, in Section 3.2, where we consider the function-parametric empirical process. For the regularity conditions required in the theorem and its proof, we refer to **Lehmann** [51].

2.6 Central limit theorem and law of large numbers

Theorem 2.4. *(***Strong law of large numbers (SLLN)***) Suppose that $X_1, \dots, X_n$ is a sequence of independent, identically distributed random variables with expected value $\mathsf{E}X_i = \mu$. The strong law of large numbers states that*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu \ \ \textit{a.s.}, \qquad \textit{as } n \to \infty.$$

The finiteness of the variance of X_i is not required in this theorem. However, it is usually assumed to make the proof easier. The proof can be found in **Feller** [23], VII.8.

Theorem 2.5. *(***Central limit theorem (CLT)***) Suppose that $X_1 \dots, X_n$ is a sequence of independent, identically distributed random variables with expected value $\mathsf{E}X_i = \mu$ and finite variance $VarX_i = \sigma^2 < \infty$. Then*

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}\Big(0, \sigma^2\Big), \qquad \textit{as } n \to \infty,$$

where $\bar{X}_n$ is the sample average, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

This theorem says that the limit in distribution of the normalized sample average $\bar{X}_n$ is a normal distribution. The proof can be found, for example in **Feller** [23], VIII.4.

We will mention the normal distribution and normal random variables many times in the remaining part of the thesis, so let us agree on the following notation:

$$\begin{aligned} \mathcal{N}(\mu, \sigma^2) &: \text{normal random variable}, \\ \Phi_{\mu,\sigma^2} &: \text{normal distribution function} \end{aligned}$$

with expected value μ and variance σ^2. Without sub-indices, Φ denotes the standard normal distribution.

Chapter 3

Empirical processes

The empirical process is a key object of probability and statistics which has a very rich literature. This chapter focuses only on the construction of the empirical process and the tail empirical process. For studies of the empirical process, we refer to **Shorack and Wellner** [70], **van de Vaart and Wellner** [76] or **Pollard** [64]. For the limit in distribution of the tail empirical process, **Einmahl** [20], [21] gives a thorough investigation.

The limit in distribution of the empirical process is strictly connected with Brownian motions and Brownian bridges, so we first recall those basic concepts.

3.1 Preliminaries

The basic material in this section can be found in various books, for example in **Khmaladze** [43], Chapter 5.

3.1.1 Brownian motions

Let $t_1, t_2, \cdots, t_k$ be an arbitrary collection of k points on $[0, 1]$ such that $0 = t_0 < t_1 < \cdots < t_k < t_{k+1} = 1$. The **standard Brownian motion** on $[0, 1]$, denoted by $w(t)$, is a zero-mean Gaussian process with increments

$$\Delta w(t_j) = w(t_{j+1}) - w(t_j), \quad j = 0, \cdots, k, \text{ for any } k, \tag{3.1}$$

which are independent Gaussian random variables and each $\Delta w(t_j)$ has expected value 0 and variance

$$\Delta t_j = t_{j+1} - t_j.$$

Hence, $w(0)$ is identically 0 and $w(t)$, which can be looked at as the increment $w(t) = w(t) - w(0)$, has expected value 0 and variance $\mathsf{E}w^2(t) = t$. The covariance function of w is

$$\mathsf{E}w(t)w(t') = \min(t, t') \qquad \text{for } t, t' \in [0, 1]. \tag{3.2}$$

Let $F(x)$ be a continuous distribution function on $\mathbb{R}$. Define

$$W_F(x) = w \circ F(x) = w(F(x)), \tag{3.3}$$

then $W_F(x)$ is called a **Brownian motion in time** $F(x)$. This means, for any given k and any given collection of points $-\infty = x_0 < x_1 < \cdots < x_{k+1} = \infty$, increments

$$\Delta W_F(x_j) = W_F(x_{j+1}) - W_F(x_j), \; j = 0, \cdots, k,$$

are independent Gaussian random variables. Each increment $\Delta W_F(x_j)$ has expected value 0 and variance

$$\mathsf{E}[\Delta W_F(x_j)]^2 = F(x_{j+1}) - F(x_j) = \Delta F(x_j).$$

This also implies that $W_F(x)$ has expected value 0 and variance

$$\mathsf{E}W_F^2(x) = F(x).$$

As a consequence of (3.2), the covariance function of W_F is

$$\mathsf{E}W_F(x)W_F(x') = \min(F(x), F(x')) = F(\min(x, x')), \quad \text{for } x, x' \in \mathbb{R}. \tag{3.4}$$

In summary, Brownian motion is a Gaussian process which is completely defined by its expected value 0 and covariance (3.2) or (3.4).

There is an extended definition of the Brownian motion in terms of functions. Let $\phi(x)$ be a square integrable function with respect to F, i.e., $\phi \in \mathcal{L}_2(F)$. When it is not necessary, we omit x in the notation relating to ϕ.

The **function-parametric** $F-$**Brownian motion** is a family $W_F(\phi), \phi \in \mathcal{L}_2(F)$, of random variables, where for each $\phi \in \mathcal{L}_2(F)$, $W_F(\phi)$ is defined as

$$W_F(\phi) = \int \phi(x) dW_F(x).$$

This $W_F(\phi)$ for each ϕ is a Gaussian random variable with expected value $\mathsf{E}W_F(\phi) = 0$ and the covariance between $W_F(\phi)$ and $W_F(\phi')$ for functions ϕ and ϕ' in $\mathcal{L}_2(F)$ is

$$\mathsf{E}W_F(\phi)W_F(\phi') = \langle \phi, \phi' \rangle_F = \int \phi(x)\phi'(x)F(dx). \tag{3.5}$$

This implies that the variance of $W_F(\phi)$ is

$$\mathsf{E}W_F^2(\phi) = \|\phi\|_F^2 = \int \phi^2(x)F(dx).$$

If ϕ is an indicator function of the form $\phi_y(x) = I_{\{x \leq y\}}$ then we get back the Brownian motion in time $F(y)$, i.e.,

$$W_F(y) = W_F(\phi_y).$$

3.1.2 Brownian bridges

The **standard Brownian bridge**, denoted by $u(t)$, is the following linear transformation of the standard Brownian motion

$$u(t) = w(t) - tw(1), \quad 0 \leq t \leq 1. \tag{3.6}$$

As a linear transformation of a Gaussian process, $u(t)$ itself is a Gaussian process. It is easy to check that the expected value is $\mathsf{E}u(t) = 0$ and the covariance is

$$\begin{aligned}\mathsf{E}[u(t)u(t')] &= \mathsf{E}\big([w(t) - tw(1)][w(t') - t'w(1)]\big)\\ &= \mathsf{E}\big(w(t)w(t')\big) - t\mathsf{E}\big(w(1)w(t')\big) - t'\mathsf{E}\big(w(t)w(1)\big) + tt'\mathsf{E}w^2(1)\\ &= \min(t,t') - tt'. \end{aligned} \tag{3.7}$$

Obviously, the variance of $u(t)$ is

$$\mathsf{E}u^2(t) = t - t^2.$$

The **Brownian bridge in time** $F(x)$, denoted by $V_F(x)$, is also a linear transformation of the Brownian motion in time $F(x)$, that is,

$$V_F(x) = W(x) - F(x)W(\infty). \tag{3.8}$$

We can also write $V_F(x)$ as

$$V_F(x) = u \circ F(x) = u(F(x)). \tag{3.9}$$

As a result of a linear transformation, $V_F(x)$ is again a Gaussian process. The expected value of $V_F(x)$ is $\mathsf{E}V_F(x) = 0$ and the covariance is

$$\begin{aligned}\mathsf{E}\big(V_F(x)V_F(x')\big) &= \mathsf{E}\big([W(x) - F(x)W(\infty)][W(x') - F(x')W(\infty)]\big)\\ &= \mathsf{E}\big(W(x)W(x')\big) - F(x)\mathsf{E}\big(W(\infty)W(x')\big)\\ &\quad - F(x')\mathsf{E}\big(W(x)W(\infty)\big) + F(x)F(x')\mathsf{E}W^2(\infty)\\ &= F(\min(x,x')) - F(x)F(x'). \end{aligned} \tag{3.10}$$

This also yields the variance of the process $V_F(x)$, which is

$$\mathsf{Var}\, V_F(x) = \mathsf{E}V_F^2(x) = F(x) - F^2(x) = F(x)(1 - F(x)). \tag{3.11}$$

The transformations (3.6) and (3.8) are linear and, moreover, they are projections. To check that they are idempotent, let us denote by Π the transformation which maps $W_F(x)$ to $V_F(x)$, i.e., $V_F(x) = \Pi(W_F(x))$. Then

$$\begin{aligned}\Pi^2 W_F(x) &= \Pi(\Pi(W_F(x))) = \Pi(W_F(x) - F(x)W_F(\infty)) \\ &= \Pi(W_F(x)) - F(x)\Pi(W_F(\infty)) \\ &= \Pi(W_F(x)) - F(x)(W_F(\infty) - F(\infty)W_F(\infty)) \\ &= \Pi(W_F(x)).\end{aligned}$$

Similarly to the extension of the Brownian motion, we also have an extension of the Brownian bridge in terms of functions. That is, $V_F(\phi)$, a **function-parametric** $F-$**Brownian bridge**. This is a projection of $W_F(\phi)$ orthogonal to the function $\mathbb{1}$, i.e.,

$$V_F(\phi) = W_F(\phi) - \langle \phi, \mathbb{1} \rangle_F W_F(\mathbb{1}). \tag{3.12}$$

Again, if we set $\phi_y(x) = I_{\{x \leq y\}}$, we will get back the Brownian bridge in time $F(y)$.

3.2 Empirical processes

This section shows only the construction of empirical processes.

3.2.1 The empirical distribution

Let $X_1, \cdots, X_n$ be independent identically distributed random variables with distribution F in $\mathbb{R}$.

• *A binomial process*

For each x, denote

$$z_n(x) = \sum_{i=1}^{n} I_{\{X_i \leq x\}}. \tag{3.13}$$

Each indicator function $I_{\{X_i \leq x\}}$ at a fixed x is a Bernoulli random variable with probability

$$\mathbb{P}[I_{\{X_i \leq x\}} = 1] = \mathbb{P}[X_i \leq x] = F(x).$$

As a sum of independent identically distributed Bernoulli random variables, $z_n(x)$ is a binomial random variable. That is, for each fixed x, $z_n(x)$ follows a binomial distribution $bin(n, F(x))$, i.e., we have

$$\mathsf{E}z_n(x) = nF(x), \qquad \mathrm{Var}z_n(x) = nF(x)(1 - F(x)). \tag{3.14}$$

We call $z_n(x)$, as a function in x, a **binomial process**.

• *An empirical distribution*

An empirical distribution $F_n(x)$ is defined as

$$F_n(x) = \frac{1}{n} z_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq x\}}. \tag{3.15}$$

As a consequence of (3.14), we have

$$\mathsf{E}F_n(x) = F(x), \qquad \mathrm{Var}F_n(x) = \frac{1}{n} F(x)(1 - F(x)). \tag{3.16}$$

The main property of the empirical distribution $F_n(x)$ is given by the following theorem.

Theorem 3.1. *(Glivenko-Cantelli theorem)*

If $X_1, \cdots, X_n$ are independent and identically distributed with distribution function $F(x)$, then

$$\sup_x |F_n(x) - F(x)| \to 0, \quad \textit{as } n \to \infty. \tag{3.17}$$

This states that $F_n(x)$ converges uniformly in x to $F(x)$.

The proof of this Theorem can be found, for example in **Khmaladze** [43].

3.2.2 The empirical process

Definition 3.1. The random process

$$v_{nF}(x) = \sqrt{n}[F_n(x) - F(x)] = \frac{z_n(x) - nF(x)}{\sqrt{n}} \tag{3.18}$$

is called the **empirical process**.

It was proved that the limit in distribution of the empirical process $v_{nF}(x)$ is the Brownian bridge in time $F(x)$, that is, $V_F(x)$ in (3.8). The proof can be found in **Khmaladze** [43], Chapter 5.

The main idea for constructing the empirical process is that, centering $z_n(x)$ by its expected value $nF(x)$ and then normalizing it by $\sqrt{n}$, we will get the process $v_{nF}(x)$ having

$$\mathsf{E} v_{nF}(x) = 0, \qquad \mathsf{Var}\, v_{nF}(x) = F(x)(1 - F(x)). \tag{3.19}$$

The covariance of the process $v_{nF}(x)$ is "stable" in the sense that it does not depend on n. Moreover, this process possesses exactly the same expected value and variance as those of the Brownian bridge $V_F(x)$ (see (3.11)).

Let us also consider the **function-parametric empirical process** defined as follows:

$$v_{nF}(\phi) = \int_{\mathbb{R}} \phi(x) v_{nF}(dx) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\phi(X_i) - \mathsf{E}\phi(X_i)], \tag{3.20}$$

where $\phi(x)$ is a function in $\mathcal{L}_2(F)$.

For example, if we choose

$$\phi(x) = \Gamma_{\boldsymbol{\theta}}^{-1} \frac{\dot{f}_{\boldsymbol{\theta}}(x)}{f_{\boldsymbol{\theta}}(x)},$$

then $v_{nF}(\phi)$ will be the main part in the asymptotic representation of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ in (2.4).

The limit in distribution of the empirical process $v_{nF}(\phi)$ with proper restrictions (see, for example in **van der Vaart and Wellner** [76] or **Pollard** [64]) on the class of functions ϕ is the function-parametric F-Brownian bridge $V_F(\phi)$ defined in (3.12).

3.3 Tail empirical processes

The tail empirical process is built up based on the right tail of the distribution $F(x)$, starting from some large value x_0. Denote by

$$z_{n,x_0}(x) = \sum_{i=1}^{n} I_{\{x_0 \le X_i \le x_0 + x\}} \tag{3.21}$$

the tail binomial process in x. Let

$$z_{n,x_0}(\infty) = \sum_{i=1}^{n} I_{\{x_0 \le X_i\}} \tag{3.22}$$

denote the total number of observations which exceeds x_0. Put

$$F_{n,x_0}(x) = \frac{z_{n,x_0}(x)}{z_{n,x_0}(\infty)} = \frac{1}{z_{n,x_0}(\infty)} \sum_{i=1}^{n} I_{\{x_0 \le X_i \le x_0 + x\}}. \tag{3.23}$$

3.3.1 The unconditional tail empirical process

For some fixed x_0, we have the following properties:

$$\begin{aligned} \mathsf{E} z_{n,x_0}(x) &= n[F(x + x_0) - F(x_0)], \\ \mathsf{E} z_{n,x_0}(\infty) &= n(1 - F(x_0)), \end{aligned}$$

and

$$\operatorname{Var} z_{n,x_0}(x) = n[F(x + x_0) - F(x_0)][1 - (F(x + x_0) - F(x_0))]. \tag{3.24}$$

As in the way of constructing the usual empirical process, we center $z_{n,x_0}(x)$ by its expected value and then normalize it by $\sqrt{\mathsf{E} z_{n,x_0}(\infty)}$. The result is the **tail unconditional empirical process**

$$v_{nF,x_0}(x) = \frac{z_{n,x_0}(x) - n[F(x + x_0) - F(x_0)]}{\sqrt{n[1 - F(x_0)]}}. \tag{3.25}$$

Obviously,

$$\mathsf{E} v_{nF,x_0}(x) = 0,$$

and the variance of $v_{nF,x_0}(x)$ is

$$\operatorname{Var} v_{nF,x_0}(x) = \frac{n[F(x+x_0) - F(x_0)][1 - (F(x+x_0) - F(x_0))]}{n[1 - F(x_0)]}. \tag{3.26}$$

Assume that as $n \to \infty$, x_0 can also change and

$$n(1 - F(x_0)) \to \infty,$$

then we have

$$\lim_{n\to\infty} \frac{n(1 - F(x_0))}{z_{n,x_0}(\infty)} = \lim_{n\to\infty} \frac{\mathsf{E} z_{n,x_0}(\infty)}{z_{n,x_0}(\infty)} = 1.$$

Denote

$$G_{x_0}(x) = \frac{F(x+x_0) - F(x_0)}{1 - F(x_0)},$$

and assume that

$$\lim_{x_0\to\infty} G_{x_0}(x) = G(x), \tag{3.27}$$

for some distribution function $G(x)$. Then we can rewrite $v_{nF,x_0}(x)$ as

$$v_{nF,x_0}(x) = \sqrt{n(1 - F(x_0))}[F_{n,x_0}(x) - G(x)]. \tag{3.28}$$

As $x_0 \to \infty$, we have

$$1 - (F(x+x_0) - F(x_0)) \to 1,$$

thus from (3.26) we have

$$\lim_{x_0\to\infty} \operatorname{Var} v_{nF,x_0}(x) = G(x).$$

The unconditional tail empirical process $v_{nF,x_0}(x)$ has the same expected value and variance as the Brownian motion in time $G(x)$, which is $W_G(x)$. The proof that the limit in distribution of the unconditional tail empirical process is in fact the Brownian motion in time $G(x)$ can be found in **Einmahl** [20],[21].

Consider a slightly different form of the unconditional tail empirical process

$$\begin{aligned}\widetilde{v}_{nF,x_0}(x) &= \frac{z_{n,x_0}(x) - n[F(x+x_0) - F(x_0)]}{\sqrt{z_{n,x_0}(\infty)}} \\ &= \sqrt{z_{n,x_0}(\infty)}\Big[F_{n,x_0}(x) - \frac{n[F(x+x_0) - F(x_0)]}{z_{n,x_0}(\infty)}\Big] \\ &\approx \sqrt{z_{n,x_0}(\infty)}[F_{n,x_0}(x) - G_{x_0}(x)]. \qquad (3.29)\end{aligned}$$

It can be seen easily that

$$\lim_{n\to\infty} \frac{\widetilde{v}_{nF,x_0}(x)}{v_{nF,x_0}(x)} = 1.$$

Therefore, the normalization by $\sqrt{z_{n,x_0}(\infty)}$ does not change the asymptotic behaviour of the unconditional tail empirical process.

3.3.2 The conditional tail empirical process

The conditional tail empirical process is defined slightly differently from the unconditional tail empirical process. That is, instead of centering the binomial tail process $z_{n,x_0}(x)$ by its expected value taken under the distribution $F(x)$ of the observations, we center it by the expected value taken under the conditional distribution and as the sample size we now use $z_{n,x_0}(\infty)$.

Suppose that $z_{n,x_0}(\infty) = m$ where m behaves in such a way that

$$m \to \infty \text{ as } x_0 \to \infty \qquad \text{and} \qquad m = o_P(n) \text{ as } n \to \infty.$$

We have

$$\mathsf{E}[z_{n,x_0}(x)|z_{n,x_0}(\infty) = m] = m\frac{F(x+x_0) - F(x_0)}{1 - F(x_0)}. \qquad (3.30)$$

The **conditional tail empirical process** is defined by

$$v_{nF,x_0}(x) = \frac{z_{n,x_0}(x) - m\frac{F(x+x_0)-F(x_0)}{1-F(x_0)}}{\sqrt{m}}. \qquad (3.31)$$

Suppose that assumption (3.27) also holds then

$$v_{nF,x_0}(x) = \frac{z_{n,x_0}(x) - mG(x)}{\sqrt{m}} = \sqrt{m}[F_{n,x_0}(x) - G(x)]. \tag{3.32}$$

This process is of the same type as the usual empirical process, having a stable variance $G(x)(1 - G(x))$. The proof that the limit in distribution of the conditional tail empirical process is a Brownian bridge in time $G(x)$, again, can be found in **Einmahl** [20],[21].

3.4 The parametric empirical process

If we suppose that the distribution F is not specified, we only know that it belongs to a parametric family of distributions; that is, $F \in \mathscr{F}$ where

$$\mathscr{F} = \left\{ F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d \right\}.$$

Consider the **parametric empirical process** defined as

$$\widehat{v}_{nF,\hat{\boldsymbol{\theta}}_n}(x) = \sqrt{n}[F_n(x) - F_{\hat{\boldsymbol{\theta}}_n}(x)], \tag{3.33}$$

where $\hat{\boldsymbol{\theta}}_n = (\hat{\theta}_1, \cdots, \hat{\theta}_d)^T = \hat{\boldsymbol{\theta}}_n(X_1, \cdots, X_n)$ is an estimate of the unknown hypothetical parameter $\boldsymbol{\theta}_0$.

Suppose that $\hat{\boldsymbol{\theta}}_n$ is the MLE. Using the representation (2.4) we can rewrite the asymptotically linear expansion of $\hat{\boldsymbol{\theta}}_n$ as

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \Gamma_{\boldsymbol{\theta}_0}^{-1} \int \frac{\dot{f}_{\boldsymbol{\theta}_0}(y)}{f_{\boldsymbol{\theta}_0}(y)} v_{nF,\boldsymbol{\theta}_0}(dy) + o_P(1), \quad n \to \infty, \tag{3.34}$$

where $f_{\boldsymbol{\theta}_0}$ denotes the hypothetical density and $\dot{f}_{\boldsymbol{\theta}_0}$ the vector of its derivatives in $\boldsymbol{\theta}$.

As a consequence, the parametric empirical process has an asymptotic expansion

$$\begin{aligned}\widehat{v}_{nF,\hat{\boldsymbol{\theta}}_n}(x) &= v_{nF,\boldsymbol{\theta}_0}(x) - \sqrt{n}[F_{\hat{\boldsymbol{\theta}}_n}(x) - F_{\boldsymbol{\theta}_0}(x)]\\ &= v_{nF,\boldsymbol{\theta}_0}(x) - \int_{-\infty}^{x} \frac{\dot{f}_{\boldsymbol{\theta}_0}(y)}{f_{\boldsymbol{\theta}_0}(y)} F_{\boldsymbol{\theta}_0}(dy)\Gamma_{\boldsymbol{\theta}_0}^{-1} \int_{-\infty}^{\infty} \frac{\dot{f}_{\boldsymbol{\theta}_0}(y)}{f_{\boldsymbol{\theta}_0}(y)} v_{nF,\boldsymbol{\theta}_0}(dy) + o_P(1)\\ &= v_{nF,\boldsymbol{\theta}_0}(x) - \int_{-\infty}^{x} \beta_F^T(y) F_{\boldsymbol{\theta}_0}(dy) \int_{-\infty}^{\infty} \beta_F(y) v_{nF,\boldsymbol{\theta}_0}(dy) + o_P(1),\end{aligned} \tag{3.35}$$

where $\beta_F(x)$ is the normalized score function, defined in (2.3).

To study the limit in distribution of the parametric empirical process, it will be more convenient to use the function-parametric version of the empirical process

$$\widehat{v}_{nF,\hat{\boldsymbol{\theta}}_n}(\phi) = \int \phi(y)\widehat{v}_{nF,\hat{\boldsymbol{\theta}}_n}(dy) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [\phi(X_i) - \mathsf{E}_{\hat{\boldsymbol{\theta}}_n}\phi(X_i)].$$

The representation in (3.35) induces the following asymptotic representation

$$\widehat{v}_{nF,\hat{\boldsymbol{\theta}}_n}(\phi) = v_{nF,\boldsymbol{\theta}_0}(\phi) - \langle \beta_F, \phi\rangle_F v_{nF,\boldsymbol{\theta}_0}(\beta_F).$$

This, together with the facts that functions β_F and $\mathbb{1}$ are orthogonal and that the limit in distribution of $v_{nF}(\phi)$ is the function parametric F-Brownian bridge $V_F(\phi)$ in (3.12), yields the limit in distribution of $\widehat{v}_{nF,\hat{\boldsymbol{\theta}}_n}(\phi)$, which we denote by $\widehat{V}_F(\phi)$. That is, $\widehat{V}_F(\phi)$ is the projection of the function parametric F-Brownian motion $W_F(\phi)$ as follows

$$\begin{aligned}\widehat{V}_F(\phi) &= V_F(\phi) - \langle \beta_F, \phi\rangle_F V_F(\beta_F)\\ &= W_F(\phi) - \langle \mathbb{1}, \phi\rangle_F W_F(\mathbb{1}) - \langle \beta_F, \phi\rangle_F W_F(\beta_F).\end{aligned} \tag{3.36}$$

This was first systematically studied by **Khmaladze** [37]. In the most recent publication, the author called the process $\widehat{V}_F(\phi)$ a **β_F-projected F-Brownian motion**, see **Khmaladze** [44].

3.5 Analogue of the empirical process for discrete distributions

Let F be a discrete distribution defined by the probability $P = \{p_1, \dots, p_m\}$. Without loss of generality, we can assume that a random variable X following the distribution F can have m values $x_1, \dots, x_m$ in $\mathbb{R}$. Suppose that we have n observations classified into m groups of values with respective frequencies $\nu_1, \dots, \nu_m$. Consider a vector of normalized differences

$$Y_{in} = \Delta v_{nF}(x_i) = \frac{\Delta z_n(x_i) - n\Delta F(x_i)}{\sqrt{n}} = \frac{\nu_i - np_i}{\sqrt{n}}, i = 1, \dots, m,$$

and denote $Y_n = (Y_{in})_{i=1}^m$. This vector Y_n can be looked at as an analogue of the increments of the empirical process. The limit in distribution of Y_n has a remarkable structure, see for instance **Khmaladze** [43], page 38. To see that structure, consider the matrix

$$\mathbf{C} = \begin{pmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_m \end{pmatrix} - \begin{pmatrix} p_1 \\ \vdots \\ p_m \end{pmatrix} (p_1, \dots, p_m). \tag{3.37}$$

We have the following theorem for the limit in distribution of vector Y_n.

Theorem 3.2. *Recall that $\Phi_{\mathbf{0},\mathbf{C}}$ is the multidimensional normal distribution with mean vector* $\mathbf{0}$ *and covariance matrix* $\mathbf{C}$ *in* (3.37). *We have*

$$\mathbb{P}\{Y_{1n} \le \lambda_1, \dots, Y_{mn} \le \lambda_m\} \to \Phi_{\mathbf{0},\mathbf{C}}(\lambda), \quad \text{as } n \to \infty,$$

where $\lambda = (\lambda_1, \dots, \lambda_m)^T$.

This is an application of the Central Limit Theorem for multivariate random variables. It follows that Y_n converges in distribution to a normal random vector. Usually, we consider vector $\widetilde{Y}_n$ of the following components, as a normalized version of Y_n,

$$\widetilde{Y}_{in} = \frac{\Delta z_n(x_i) - n\Delta F(x_i)}{\sqrt{np_i}} = \frac{\nu_i - np_i}{\sqrt{np_i}}.$$

Then, as a consequence of Theorem 3.2, the limit of the distribution of vector $\widetilde{Y}_n$ is given by

$$\mathbb{P}\left\{\widetilde{Y}_{1n} \leq \lambda_1, \ldots, \widetilde{Y}_{mn} \leq \lambda_m\right\} \to \Phi_{\mathbf{0},\widetilde{\mathbf{C}}}(\lambda)$$

where

$$\widetilde{\mathbf{C}} = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} - \begin{pmatrix} \sqrt{p_1} \\ \vdots \\ \sqrt{p_m} \end{pmatrix} (\sqrt{p_1}, \ldots, \sqrt{p_m}).$$

The limit in distribution of vector $\widetilde{Y}_n$, denoted by $\widetilde{Y}$, can be represented as the projection of a standard Gaussian random vector X orthogonal to vector $\sqrt{p} = (\sqrt{p_1}, \ldots, \sqrt{p_m})^T$. That means,

$$\widetilde{Y}_n \xrightarrow{d} \widetilde{Y} = X - \langle X, \sqrt{p}\rangle \sqrt{p},$$

where $X = (X_1, \ldots, X_m)^T$ is a vector of independent standard Gaussian random variables.

3.6 Empirical processes in $\mathbb{R}^k$

Suppose that we have n independent identically distributed random vectors $X_1, \ldots, X_n$ in the space $\mathbb{R}^k$ having the same distribution F. Denote by B a Borel subset of $\mathbb{R}^k$, and consider the empirical process indexed by sets

$$v_{nF}(B) = \sqrt{n}[F_n(B) - F(B)],$$

where

$$F_n(B) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \in B\}}$$

is the empirical distribution indexed by sets. Let the Borel set be of the form $B = (0, x_1] \times (0, x_2] \times \cdots \times (0, x_k]$; then putting

$$\mathbf{x} = (x_1, x_2, \ldots, x_k),$$

we can write the empirical process in terms of points as follows:

$$v_{nF}(\mathbf{x}) = \sqrt{n}[F_n(\mathbf{x}) - F(\mathbf{x})].$$

The **function-parametric empirical process** version in $\mathbb{R}^k$ is defined in a similar way as when $k = 1$. That is,

$$v_{nF}(\phi) = \int_{\mathbb{R}^k} \phi(\mathbf{x}) v_{nF}(d\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} [\phi(X_i) - \mathsf{E}\phi(X_i)],$$

for some function $\phi(\mathbf{x}) = \phi(x_1, \ldots, x_k)$ in $\mathcal{L}_2(F)$.

Also, the limit in distribution of $v_{nF}(\phi)$ is the function-parametric F-Brownian bridge $V_F(\phi)$, defined as a linear transformation of the function-parametric F-Brownian motion $W_F(\phi)$, i.e.,

$$V_F(\phi) = W_F(\phi) - \langle \phi, \mathbb{1} \rangle_F W_F(\mathbb{1}).$$

3.7 Concluding remarks

As long ago as in 1933, **Kolmogorov** [46] realised that if $F(x)$ is a continuous distribution function on the real line, and if X follows the distribution function $F(x)$ then $U = F(X)$ is uniformly distributed on $[0, 1]$. This implies that the empirical process $v_{nF}(x)$ can be transformed into the uniform empirical process

$$u_n(t) = \sqrt{n}[F_{u,n}(t) - t] = v_{nF}(F^{-1}(t))$$

by the time transformation $t = F(x)$. At the same time, we have seen in this chapter that the limit in distribution of $u_n(t)$ is the standard Brownian bridge $u(t)$. Consequently, any statistic from $v_{nF}(x)$ that is invariant under the time transformation $t = F(x)$ will be asymptotically distribution free.

However, that fact is no longer true if F is a discrete distribution. For nearly a century, there existed only one distribution free GOF test, the chi-square test for testing statistical hypotheses on discrete distributions.

In the case of a parametric hypothesis $F \in \mathscr{F}$, the time transformation $t = F(x)$ on the parametric empirical process $\widehat{v}_{nF,\hat{\boldsymbol{\theta}}_n}(x)$ will also not be of much immediate use. The limit in distribution of this process, as can be seen from (3.36), depends not only on the hypothetical distribution $F_{\boldsymbol{\theta}_0}$ but also on the true parameter $\boldsymbol{\theta}_0$.

However, in the case of continuous distribution in multi-dimensional space and parametric hypotheses, a new methodology was introduced in **Khmaladze** [38],[39] and [40]. There, the parametric empirical process $\widehat{v}_{nF,\hat{\boldsymbol{\theta}}_n}$ was transformed into the process w_n which under the null parametric hypothesis would converge to a Brownian motion in multidimensional space.

In this thesis, we use another new transformation to construct a new class of GOF tests for two different hypothesis testing problems. The meaning of the term GOF tests will be clarified in the next chapter.

Chapter 4

Goodness of fit tests

Among the class of all test statistics for hypothesis testing problems, goodness of fit (GOF) tests are very different in nature compared to others. Generally speaking, GOF tests are of omnibus nature, which means that they are able to detect deviations in all "directions" of local alternatives from the null hypothesis. For testing simple hypotheses, if the alternative is specified, the likelihood ratio test appears as the optimal test. This will be discussed in Section 4.1. However, most of the time, the alternative is not specified, so we are in need of a test statistic which is not only asymptotically distribution free but also equally sensitive to all local deviations. Classical GOF tests will be presented in Section 4.2. We then present the formulation of the GOF testing problem in Section 4.3.

4.1 The optimal test for a particular alternative

Given a set of independent and identically distributed random variables $X_1, \ldots, X_n$, assume that they come from some unknown distribution F. Consider a simple hypothesis testing problem with the **null distribution** F_0, i.e.,

$$H_0 : F = F_0.$$

Under the alternative, assume that for each $n = 1, 2, \ldots$, the random variables $X_1, \ldots, X_n$ follow some particular distribution $F_{a,n}$. We sometimes simply write F_a when unambiguous to do so. The statement for the alternative hypothesis is

$$H_a : F = F_a.$$

4.1.1 Contiguous alternatives and distributions of a random sample

The following definition originated from contiguity theory, see Section 2.3.

Definition 4.1. The sequence of distributions $F_{a,n}$ are called **contiguous alternatives** to distribution F_0 if there exists a sequence of functions $h_n(\cdot)$ such that for each n, the Lebesgue decomposition of $F_{a,n}$ with respect to F_0, that is $F_{a,n} = F^c_{a,n} + F^{\perp}_{a,n}$, satisfies:

$$n\text{Var}(F^{\perp}_{a,n}) \to 0, \quad n \to \infty, \tag{4.1}$$

$$\left[\frac{dF^c_{a,n}}{dF_0}\right]^{1/2} = 1 + \frac{1}{2\sqrt{n}} h_n(\cdot), \tag{4.2}$$

and

$$\int [h_n(x) - h(x)]^2 F_0(dx) \to 0 \tag{4.3}$$

for some function $h(\cdot)$ where

$$\int h^2(x) F_0(dx) < \infty, \tag{4.4}$$

$$\int h(x) F_0(dx) = 0. \tag{4.5}$$

As stated in **Khmaladze** [40], the function $h(\cdot)$ involved in this definition can be viewed as a function which determines the "direction" from which the alternative distributions $(F_{a,n})$ converge to the null distribution F_0.

It is quite obvious that a function h, which satisfies conditions (4.3) and (4.4) must satisfy condition (4.5).

Since $X_1, \ldots, X_n$ are independent, the distributions of this sample under the null and alternative hypothesis are respectively the following n-fold direct products

$$\mathbb{F}_0^n = F_0 \times F_0 \times \cdots \times F_0$$

and

$$\mathbb{F}_a^n = F_{a,n} \times F_{a,n} \times \cdots \times F_{a,n}.$$

As presented in Section 2.3, as a result from **Oosterhoof** and **van Zwet** [60], for contiguous alternatives $F_{a,n}$, the sequence $\mathbb{F}_a^n$ is contiguous with respect to $\mathbb{F}_0^n$; which we write $\mathbb{F}_{a,n} \lhd \mathbb{F}_{0,n}$.

From definition 4.1 of contiguous alternatives, we see that conditions (4.3) and (4.4) require that the sequence $h_n(\cdot)$ converges in $\mathcal{L}_2(F_0)$ as an additional restriction. The choice of testing a null hypothesis against such converging contiguous alternatives is considered in a large number of studies regarding hypothesis testing problems. In some others, for example in **Khmaladze** [41], the author introduced and studied GOF tests for "chimeric alternatives". For chimeric alternatives, the convergence of the sequence $h_n(\cdot)$ was removed, and instead the sequence is required to be bounded in norm but there is no limiting point.

In the remainder of this section, suppose that $F_{a,n} \ll F_0$ for all n and the contiguous alternatives $F_{a,n}$ converge to F_0 from the direction $h(\cdot)$. Denote by f_a and f_0 the corresponding density functions of F_a and F_0 where we omit the index n. From (4.2) and (4.3) we have

$$\left[\frac{f_a(\cdot)}{f_0(\cdot)}\right]^{1/2} = 1 + \frac{1}{2\sqrt{n}} h(\cdot) + o(1/\sqrt{n}).$$

This implies

$$f_a^{1/2}(x) = \left[1 + \frac{1}{2\sqrt{n}} h(x) + o(1/\sqrt{n})\right] f_0^{1/2}(x). \tag{4.6}$$

4.1.2 The likelihood ratio test is the optimal test

Let us denote by $T_n(\mathbf{X}; F_0)$ a test statistic, as a function of $X_1, \ldots, X_n$ and probably of F_0 as well. Denote by $\mathcal{C}_T$ the **critical region** of the test statistic T_n.

Here and below, $\mathbb{P}_{F_a}$ and $\mathbb{P}_{F_0}$ respectively denote the probabilities of events under the alternative distribution F_a and the null distribution F_0. Then we have $\mathbb{P}_{F_0}\{T_n \in \mathcal{C}_T\}$ is the probability of making a **type I error**, i.e., the probability of rejecting the null hypothesis H_0 while H_0 is true. And $\mathbb{P}_{F_a}\{T_n \in \mathcal{C}_T\}$ is the **power of the test**, i.e., the probability of correctly rejecting the null hypothesis when the alternative F_a is true. If we are not considering the possibility of controlling the type I error being smaller than some level α, then the test T_n which maximizes

$$K_T = \mathbb{P}_{F_a}\{T_n \in \mathcal{C}_T\} - \mathbb{P}_{F_0}\{T_n \in \mathcal{C}_T\},$$

will be the optimal test for testing the null distribution F_0 against the particular contiguous alternative distributions F_a.

Assume that $g_0(t), g_a(t)$ are densities of T_n under the hypotheses H_0 and H_a respectively. Then we have

$$K_T = \int_{\mathcal{C}_T} [g_a(t) - g_0(t)]dt.$$

It is easy to see that the optimal critical region $\mathcal{C}_T$ should be of the form

$$\mathcal{C}_{T,max} = \Big\{t : \frac{g_a(t)}{g_0(t)} > 1\Big\}.$$

Also, denote the maximum value of K_T for each given T_n by

$$K_T^* = \int_{\mathcal{C}_{T,max}} [g_a(t) - g_0(t)]dt.$$

We can rewrite K_T^* in terms of the distributions of the sample, that is

$$\begin{aligned} K_T^* &= \mathbb{P}_{F_a}[T_n \in \mathcal{C}_{T,max}] - \mathbb{P}_{F_0}[T_n \in \mathcal{C}_{T,max}] \\ &= \mathbb{P}_{F_a}[(X_1, \ldots, X_n) \in T_n^{-1}(\mathcal{C}_{T,max})] - \mathbb{P}_{F_0}[(X_1, \ldots, X_n) \in T_n^{-1}(\mathcal{C}_{T,max})] \\ &= \int \cdots \int_{T_n^{-1}(\mathcal{C}_{T,max})} \left[d\mathbb{F}_a^n(\mathbf{x}) - d\mathbb{F}_0^n(\mathbf{x}) \right] \\ &= \int \cdots \int_{T_n^{-1}(\mathcal{C}_{T,max})} \left[\prod_{i=1}^n f_a(x_i) - \prod_{i=1}^n f_0(x_i) \right] \prod_{i=1}^n dx_i. \end{aligned} \tag{4.7}$$

Put

$$K_h = \max_{T_n} K_T^*. \tag{4.8}$$

This is the optimal value of K_T for all test statistics T_n, which depends only on $h(\cdot)$. Obviously, if the statistic $\widetilde{T}_n$ satisfies the condition

$$\widetilde{T}_n^{-1}(\mathcal{C}_{\widetilde{T},max}) = \left\{ (x_1, \ldots, x_n) : \frac{\prod_{i=1}^n f_a(x_i)}{\prod_{i=1}^n f_0(x_i)} > 1 \right\}, \tag{4.9}$$

and assigning the condition on the right hand side as a critical region L, then $\widetilde{T}_n$ is the optimal test. On the other hand, the **likelihood ratio test** (LRT) statistic of the form

$$LRT(X_1, \ldots, X_n) = \frac{\prod_{i=1}^n f_a(X_i)}{\prod_{i=1}^n f_0(X_i)}$$

has the critical region L described above, which gives it the property of being the most powerful test for testing F_0 against F_a.

4.1.3 The limit in distribution of the log-likelihood ratio test under H_0

Since the log function is increasing, instead of considering the likelihood ratio, we equivalently consider the log-likelihood ratio test statistic, that is,

$$\Lambda_h(X_1, \ldots, X_n) = \log LRT(X_1, \ldots, X_n) = \sum_{i=1}^n \log \frac{f_a(X_i)}{f_0(X_i)}.$$

We can rewrite $\Lambda_h(X_1,\ldots,X_n)$ as a function of $h(\cdot)$. That is, from (4.6) and as a consequence of the Taylor expansion, $\Lambda_h(X_1,\ldots,X_n)$ can be represented as

$$\begin{aligned}\Lambda_h(X_1,\ldots,X_n) &= 2\sum_{i=1}^{n}\log\Big(1+\frac{1}{2\sqrt{n}}h(X_i)+o_P(1/\sqrt{n})\Big)\\ &= \sum_{i=1}^{n}\Big[\frac{1}{\sqrt{n}}h(X_i)-\frac{1}{2n}h^2(X_i)\Big]+o_P(1).\end{aligned}$$

Since $h(\cdot)\in\mathcal{L}_2(F_0)$, it follows from the SLLN that

$$\frac{1}{2n}\sum_{i=1}^{n}h^2(X_i)\to\frac{1}{2}\left\|h\right\|_{F_0}^2,\ \text{as}\, n\to\infty.$$

On the other hand, by the CLT we have

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}h(X_i)\overset{d}{\to}\mathcal{N}(0,\left\|h\right\|_{F_0}^2).$$

Consequently, the limit in distribution of Λ_h under the null hypothesis is

$$\begin{aligned}\Lambda_h &= \sum_{i=1}^{n}\frac{1}{\sqrt{n}}h(X_i)-\frac{1}{2}\left\|h\right\|_{F_0}^2+o_P(1) && (4.10)\\ &\overset{d}{\to}_{F_0}\mathcal{N}(-\frac{1}{2}\left\|h\right\|_{F_0}^2,\left\|h\right\|_{F_0}^2). && (4.11)\end{aligned}$$

The rigorous proof of asymptotic normality of the log-likelihood ratio test in terms of Hellinger distance can be found in **Oosterhoof and van Zwet** [60].

4.1.4 An evaluation of the optimal test

We are going to evaluate K_h (or K_{Λ_h}), the maximum value of K_T for particular contiguous alternatives from direction $h(\cdot)$. By its definition in (4.7) and (4.8), we see that K_h itself is the distance in total variation between two probability measures $\mathbb{F}_0^n$ and $\mathbb{F}_a^n$ (see definition of the distance in total variation in Section 2.3). We have,

$$\begin{aligned} K_h &= \int \cdots \int_L d\mathbb{F}_a^n(\mathbf{x}) - d\mathbb{F}_0^n(\mathbf{x}) \\ &= \int \cdots \int_L \Big[\frac{\prod_{i=1}^n f_a(x_i)}{\prod_{i=1}^n f_0(x_i)} - 1\Big] \prod_{i=1}^n f_0(x_i) dx_i \\ &= K_L^a - K_L^0. \end{aligned} \tag{4.12}$$

Since we can rewrite L as

$$L = \{(x_1, \ldots, x_n) : \Lambda_h(x_1, \ldots, x_n) > 0\},$$

we have

$$\begin{aligned} K_L^0 = \mathbb{F}_0^n(L) &= \int \cdots \int_L \prod_{i=1}^n f_0(x_i) dx_i \\ &= \mathbb{P}_{F_0}\{\Lambda_h > 0\}. \end{aligned}$$

Moreover, from (4.11), which shows the limit in distribution of Λ_h under the null distribution F_0, we deduce

$$K_L^0 \to 1 - \Phi_{-\frac{1}{2}\|h\|_{F_0}^2, \|h\|_{F_0}^2}(0).$$

In addition,

$$\begin{aligned} K_L^a &= \int \cdots \int_L \frac{\prod_{i=1}^n f_a(x_i)}{\prod_{i=1}^n f_0(x_i)} d\mathbb{F}_0^n(\mathbf{x}) \\ &= \int \cdots \int_{\Lambda_h > 0} \exp(\Lambda_h) d\mathbb{F}_0^n(\mathbf{x}) \\ &= \int_{t>0} \exp(t) d\Upsilon(t) \end{aligned}$$

where Υ denotes the distribution of Λ_h under the measure $\mathbb{F}_0^n$, or in other words, under the hypothesis F_0. Hence, it follows from (4.11) that

$$\begin{aligned}
K_L^a \to & \frac{1}{\sqrt{2\pi \left\|h\right\|_{F_0}^2}} \int_0^\infty \exp(t) \exp\left\{ -\frac{(t + \frac{1}{2}\left\|h\right\|_{F_0}^2)^2}{2\left\|h\right\|_{F_0}^2} \right\} dt \\
&= \frac{1}{\sqrt{2\pi \left\|h\right\|_{F_0}^2}} \int_0^\infty \exp\left\{ -\frac{(t - \frac{1}{2}\left\|h\right\|_{F_0}^2)^2}{2\left\|h\right\|_{F_0}^2} \right\} dt \\
&= 1 - \Phi_{\frac{1}{2}\|h\|_{F_0}^2, \|h\|_{F_0}^2}(0).
\end{aligned}$$

Eventually, we have

$$\begin{aligned}
K_h &\to 1 - \Phi_{\frac{1}{2}\|h\|_{F_0}^2, \|h\|_{F_0}^2}(0) - (1 - \Phi_{-\frac{1}{2}\|h\|_{F_0}^2, \|h\|_{F_0}^2}(0)) \\
&= 2\Phi_{-\frac{1}{2}\|h\|_{F_0}^2, \|h\|_{F_0}^2}(0) - 1 \\
&= 2\Phi(-\frac{1}{2}\left\|h\right\|_{F_0}) - 1, \qquad (4.13)
\end{aligned}$$

where we recall that Φ, with the sub-indices omitted, denotes the standard normal distribution.

4.2 Goodness of fit tests

Frequently in testing simple hypotheses on a random variable X following a distribution F, the alternative is not specified. The null and alternative hypotheses are

$$H_0 : F = F_0$$

versus

$$H_a : F \neq F_0.$$

In a composite parametric testing problem, the null hypothesis is

$$H_0 : F \in \mathscr{F}_0 = \left\{ F_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta_0 \right\},$$

while the alternative hypothesis is usually stated as

$$H_a : F \notin \mathscr{F}_0.$$

In both problems, the optimality of the log-likelihood ratio test is lost since the alternative distributions can approach the hypothetical distribution in infinitely many different directions. We want to have a statistic which is able to detect all possible deviations of any class of local alternatives.

In the following, we are going to give examples of some GOF tests as well as non-GOF tests for testing simple hypotheses.

4.2.1 Examples of non-GOF tests

Example 4.1. (**For testing continuous distributions**)

We first see that any linear transformation of the empirical process does not form a GOF test. In fact, consider a statistic T of the form

$$T = v_{nF_0}(w) = \int w(x) dv_{nF_0}(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [w(X_i) - \mathsf{E}_{F_0}(w(X_i))].$$

Function $w(x)$ is considered as the weight function. It is clear that $\Lambda_{\widetilde{h}}$ (for some certain direction $\widetilde{h}$) belongs to this class of tests. In fact, from (4.10) we have

$$\Lambda_{\widetilde{h}} = v_{nF_0}(\widetilde{h}) - \frac{1}{2} \left\| \widetilde{h} \right\|_{F_0}^2 + o_P(1).$$

Having said that T is not a GOF test, it is of interest to see what the limit in distribution of T is under particular alternatives. Throughout the examples below, we again use $h(\cdot)$ for the direction in which a sequence of local contiguous alternative distributions F_a (with densities f_a) approach the null distribution F_0 (with density f_0). Recall that,

$$f_a^{1/2}(x) = f_0^{1/2}(x) \Big[1 + \frac{1}{2\sqrt{n}} h(x) + o(1/\sqrt{n}) \Big].$$

Recall from Chapter 3 that the limit in distribution of T, as a function parametric empirical process, is $V_{F_0}(w)$, a function-parametric F_0-Brownian bridge. Let us consider the limit in distribution of T under the alternative F_a. We can rewrite $v_{nF_0}(w)$ as

$$\begin{aligned}
v_{nF_0}(w) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[w(X_i) - \mathsf{E}_{F_0}(w(X_i))] \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{[w(X_i) - \mathsf{E}_{F_a}(w(X_i))] + [\mathsf{E}_{F_a}w(X_i) - \mathsf{E}_{F_0}(w(X_i))]\} \\
&= v_{nF_a}(w) + \sqrt{n}\Big[\int w(y)(1+\frac{1}{\sqrt{n}}h(y))F_0(dy) - \int w(y)F_0(dy)\Big] + o(1) \\
&= v_{nF_a}(w) + \int w(y)h(y)F_0(dy) + o(1).
\end{aligned}$$

It was proved, see for example in **van de Vaart and Wellner** [76], that the limit in distribution of the process $v_{nF_a}(w)$ under the alternative F_a is again $V_{F_0}(w)$. In the remainder, let us denote by $H(w)$ a functional of w, which is

$$H(w) = \langle w, h\rangle_{F_0} = \int w(y)h(y)F_0(dy). \tag{4.14}$$

This $H(w)$ is the asymptotic form of the shift of the limiting Brownian bridge since

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\mathsf{E}_{F_a}w(X_i) - \mathsf{E}_{F_0}w(X_i)] = H(w) + o(1).$$

Therefore, the limit in distribution of $v_{nF_0}(w)$ under the alternative F_a is

$$V_{F_0}(w) + H(w). \tag{4.15}$$

Consequently, for $H(w) = \langle w, h\rangle_{F_0} = 0$, or in other words, for the alternatives which approach the null distribution in a direction $h(\cdot)$ orthogonal to $w(\cdot)$, the limit in distribution of the test statistics T under the null distribution and under the alternative distribution are the same.

Example 4.2. (**For testing discrete distributions**)

Consider the case when the random variable X of interest is discrete. Denote by m the number of events and by $\{\nu_{1n}, \cdots, \nu_{mn}\}$ the observed frequencies (n is the total number of observations). The null distribution is

$$P = (p_1, \dots, p_m).$$

For each n, denote by $\widetilde{P}$ the contiguous alternative distribution,

$$\widetilde{P} = (\widetilde{p}_1, \cdots, \widetilde{p}_m).$$

By the definition of contiguous alternatives, when n is sufficiently large, we can represent $\widetilde{p}_i$ as a local deviation of p_i approximately as follows:

$$\widetilde{p}_i = p_i\left(1 + \frac{1}{\sqrt{n}} h_i\right), \tag{4.16}$$

where

(i) $\sum_{i=1}^m h_i p_i = 0$,

(ii) $\sum_{i=1}^m h_i^2 p_i < \infty$.

Note that equation (4.16) is obtained by taking the square of (4.2), approximating $h_n(\cdot)$ by its limit $h(\cdot)$ and omitting the negligible terms.

An analogous version of the statistic T_n in Example 4.1 above is the following test statistic, a linear functional of the vector of components of chi-square statistic

$$T_n = \sum_{i=1}^m w_i \frac{\nu_{in} - np_i}{\sqrt{np_i}},$$

where w_i are weights put on each component. Consider the expected value of T_n under the null distribution P and under the alternative distribution $\widetilde{P}$. It is obvious that $\mathsf{E}_P T_n = 0$. Besides,

$$\begin{aligned}\mathsf{E}_{\widetilde{P}} T_n = \mathsf{E}_{\widetilde{P}} \sum_{i=1}^m w_i \frac{\nu_{in} - np_i}{\sqrt{np_i}} &= \sum_{i=1}^m w_i \frac{n\widetilde{p}_i - np_i}{\sqrt{np_i}} \\ = \sum_{i=1}^m w_i \frac{\sqrt{n} p_i h_i}{\sqrt{np_i}} &= \sum_{i=1}^m w_i \sqrt{p_i} h_i.\end{aligned}$$

Therefore, if we choose $h = (h_1, \ldots, h_m)^T$ such that it is orthogonal to $w\sqrt{p} = (w_1\sqrt{p_1}, \ldots, w_m\sqrt{p_m})^T$ then the asymptotic behaviours of the test statistic T_n under the null and alternative hypotheses are identical.

From the above examples, we see that GOF test statistics are essentially non-linear functionals of the empirical process. We will mention some widely used examples of GOF test statistics for testing continuous and discrete distributions below.

4.2.2 Examples of GOF tests for continuous distributions

The given examples below are for testing simple hypotheses of continuous distributions on the real line $\mathbb{R}$.

Example 4.3. The **Kolmogorov-Smirnov (KS)** statistic has the following form:

$$KS_n = \sup_x |v_{nF_0}(x)| = \sup_x \sqrt{n}\,|F_n(x) - F_0(x)|\,. \tag{4.17}$$

In the function-parametric version, we may consider

$$KS_n = \sup_{\phi\in\Phi} |v_{nF_0}(\phi)|$$

for Φ a suitably chosen class of functions ϕ.

The test statistic (4.17) was introduced by **Kolmogorov** [46] in 1933. Recall that this KS statistic is invariant under the change of time $t = F(x)$, so that

$$KS_n = \sup_x |v_{nF_0}(x)| = \sup_t |u_n(t))|\,.$$

Denote by $K_n(z)$ the distribution of KS_n, Kolmogorov proved that

$$\begin{aligned}\lim_{n\to\infty} \mathbb{P}\left\{KS_n < z\right\} &= \lim_{n\to\infty} K_n(z) = K(z)\\ &= 1 - 2\sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2z^2}, \qquad 0 < z < \infty,\end{aligned} \tag{4.18}$$

and thus K is called the Kolmogorov's distribution function. Then, **Smirnov** [73] in 1939 considered the one-sided statistic

$$\sup_x v_{nF_0}(x) = \sup_x \sqrt{n}[F_n(x) - F_0(x)]$$

and derived its limit in distribution, which turned out to be simpler than Kolmogorov's distribution. In fact,

$$\mathbb{P}\left\{\sup_x \sqrt{n}[F_n(x) - F_0(x)] \leq z\right\} \to 1 - e^{-\frac{1}{\sqrt{2}}z} \quad \text{as } n \to \infty.$$

Let us consider the distribution of the KS test statistic under particular contiguous alternatives. We have

$$\begin{aligned} KS_n &= \sqrt{n}\sup_x |F_n(x) - F_0(x)| \\ &= \sqrt{n}\sup_x |F_n(x) - F_a(x) + F_a(x) - F_0(x)| \\ &= \sqrt{n}\sup_x \left|F_n(x) - F_a(x) + \frac{1}{\sqrt{n}}\int_{-\infty}^{x} h(y)f(y)dy\right| \\ &= \sup_x \left|v_{nF_a}(x) + \int_{-\infty}^{x} h(y)f(y)dy\right| \\ &= \sup_x |v_{nF_a}(x) + S(x)|, \end{aligned}$$

where

$$S(x) = \int_{-\infty}^{x} h(y)f(y)dy.$$

As mentioned above, the limit in distribution of $v_{nF_a}(x)$ under alternatives F_a is the same as that of $v_{nF_0}(x)$ under the null hypothesis. Hence the contribution of the "shift" $S(x)$ will distinguish the limit in distribution of the KS test under the null and alternative hypotheses.

Example 4.4. The **Cramér-von Mises** statistic is of the following form

$$\begin{aligned} \Omega^2[\Psi(F_0(x))] &= \int v_{nF_0}^2(x)\Psi(F_0(x))dF_0(x) \\ &= n\int [F_n(x) - F_0(x)]^2\Psi(F_0(x))dF_0(x). \end{aligned}$$

Here $\Psi(t)$ is a certain non-negative function defined on the interval $[0,1]$ such that $\Psi(t), t\Psi(t), t^2\Psi(t)$ are integrable on $[0,1]$. The original test statistic was first considered by **Cramér** [9], in 1928. He suggested the following test statistic

$$n\int [F_n(x) - F_0(x)]^2 d(x).$$

In 1931, **von Mises** [78] independently made an equivalent suggestion and developed a few properties of the test.

This current version was in fact a modification suggested by **Smirnov** [72] in 1937. He also considered the case with $\Psi(F_0(x)) \equiv 1$, and showed that in this case, the statistic

$$\Omega_n^2 = \int v_{nF_0}^2(x)dF_0(x) = n\int [F_n(x) - F_0(x)]^2 dF_0(x),$$

under the null hypothesis, has in the limit the "omega-squared" distribution, independent of the hypothetical distribution function.

To consider the limit in distribution of Ω_n^2 under contiguous alternatives, we will rewrite Ω_n^2 as follows:

$$\begin{aligned}\Omega_n^2 &= n\int [F_n(x) - F_0(x)]^2 dF_0(x)\\ &= n\int [F_n(x) - F_a(x) + F_a(x) - F_0(x)]^2 dF_0(x)\\ &= \int [v_{nF_a}(x) + S(x)]^2 dF_0(x).\end{aligned}$$

That yields the limit in distribution of Ω_n^2 under the alternative is of the quadratic form $\int [V_{F_0}(x) + S(x)]^2 dF_0(x)$. This is different from the limit under the null hypothesis, which is $\int V_{F_0}^2(x)dF_0(x)$.

Example 4.5. The **Anderson-Darling** test statistic is of the following form:

$$A_n^2 = \int \frac{v_{nF_0}^2(x)}{F_0(x)(1-F_0(x))}dF_0(x) = n\int \frac{[F_n(x) - F_0(x)]^2}{F_0(x)(1-F_0(x))}dF_0(x).$$

This A_n^2 test is a modification of the Cramér-von Mises test, giving more weight to observations in the tail of the distribution. This was introduced in 1952 by **Anderson and Darling** [3]. The asymptotically distribution free property of this test is quite clear since

$$A_n^2 = \int_0^1 \frac{u_n(t)^2}{t(1-t)} dt.$$

To see the limit in distribution of A_n^2 under contiguous alternatives, we can proceed in the same way as we did for Ω_n^2.

Example 4.6. Consider a weighted version of the empirical process

$$\widetilde{v}_{nF_0}(x) = \frac{v_{nF_0}(x)}{\lambda(x)}$$

where $\lambda(x)$ is a **Chibisov-O'Reilly** weight function. The class of Chibisov-O'Reilly functions Λ includes all *positive* functions $\lambda(t)$ on $(0, 1)$, where a positive function $\lambda(t)$ is such that

$$\inf_{\delta \leq t \leq 1-\delta} \lambda(t) > 0, \text{ for all } \delta \in (0, 1/2),$$

and where

$$I(\lambda, c) = \int_0^1 [t(1-t)]^{-1} \exp\left\{\frac{-c\lambda^2(t)}{t(1-t)}\right\} dt < \infty, \text{ for all } c > 0.$$

Members in this class can be, for example,

$$\lambda(t) = (t(1-t))^b, \ \ 0 < b < 1/2.$$

The class of Chibisov-O'Reilly functions leads to a class of GOF tests, for example, $\sup_x |\widetilde{v}_{nF_0}(x)|$. The reason to choose the weight functions is that, under the null hypothesis F_0, we have

$$\widetilde{v}_{nF_0}(x) \xrightarrow{d} \frac{V_F(x)}{\lambda(F(x))},$$

which was known as the **Chibisov-O'Reilly theorem**. This theorem was originally proved by **Chibisov** [7] in 1964 and then was re-examined by **O'Reilly** [61] in 1974.

Under contiguous alternatives, we have

$$\widetilde{v}_{nF_0}(x) \xrightarrow{d} \frac{V_F(x) + S(x)}{\lambda(F(x))}.$$

Another proof for this convergence can be seen also in **Szyszkowicz** [74].

4.2.3 Examples of GOF tests for discrete distributions

We are giving here two examples of test statistics which are sensitive to all deviations of local contiguous alternatives from the null distribution. As these examples are for discrete distributions, all assumptions are the same as in Example 4.2.

Example 4.7. Consider the chi-square statistic

$$\chi_n^2 = \sum_{i=1}^{m} \frac{(\nu_{in} - np_i)^2}{np_i}.$$

This is the most well-known test statistic, introduced by **Pearson** [62] as long ago as 1900. It is the most widely used test statistic for testing statistical hypotheses for discrete distributions or testing independence of two discrete random vectors in a contingency table context.

It is well known that χ_n^2 under the null hypothesis follows a chi-square distribution with $m-1$ degrees of freedom. Let us see what the expected values of this test are under the null and the alternative hypotheses. We have,

$$\mathsf{E}_P \chi_n^2 = \mathsf{E}_P \sum_{i=1}^{m} \frac{(\nu_{in} - np_i)^2}{np_i} = \sum_{i=1}^{m} \frac{np_i(1-p_i)}{np_i} = \sum_{i=1}^{m} (1-p_i) = m-1.$$

Under contiguous alternatives $\widetilde{P}$, we have

$$\begin{aligned}
\mathsf{E}_{\widetilde{P}} \sum_{i=1}^{m} \frac{(\nu_{in} - np_i)^2}{np_i} &= \sum_{i=1}^{m} \mathsf{E}_{\widetilde{P}} \frac{(\nu_{in} - n\widetilde{p}_i + n\widetilde{p}_i - np_i)^2}{np_i} \\
&= \sum_{i=1}^{m} \mathsf{E}_{\widetilde{P}} \frac{(\nu_{in} - n\widetilde{p}_i)^2 + n^2(\widetilde{p}_i - p_i)^2 + 2n(\nu_{in} - n\widetilde{p}_i)(\widetilde{p}_i - p_i)}{np_i} \\
&= \sum_{i=1}^{m} \frac{\widetilde{p}_i(1 - \widetilde{p}_i)}{p_i} + n \sum_{i=1}^{m} \frac{(\widetilde{p}_i - p_i)^2}{p_i} \\
&= \sum_{i=1}^{m} \left(1 + \frac{1}{\sqrt{n}} h_i\right)(1 - \widetilde{p}_i) + \sum_{i=1}^{m} p_i h_i^2 \\
&= m - 1 + \frac{1}{\sqrt{n}} \sum_{i=1}^{m} h_i (1 - \widetilde{p}_i) + \sum_{i=1}^{m} p_i h_i^2 \\
&\to m - 1 + \sum_{i=1}^{m} p_i h_i^2,
\end{aligned}$$

since $\frac{1}{\sqrt{n}} \sum_{i=1}^{m} h_i(1 - \widetilde{p}_i) \to 0$ as $n \to \infty$.

It is always true that $\sum_{i=1}^{m} p_i h_i^2 > 0$. The distribution of $\sum_{i=1}^{m} \frac{(\nu_{in} - np_i)^2}{np_i}$ under alternatives is the non-central chi-squared distribution and the "shift" $\sum_{i=1}^{m} p_i h_i^2$ is the non-centrality parameter.

Example 4.8. Consider a vector of components

$$\sum_{i=1}^{k} \frac{\nu_{in} - np_i}{\sqrt{np_i}}, \quad k = 1, \ldots, m.$$

This can be looked at as the vector of the discrete version of the Kolmogorov-Smirnov test statistic. As we have seen in Section 3.5, this vector under the null hypothesis converges in distribution to a Gaussian vector, denoted by $\{W_k\}_{k=1}^{m}$. Thus,

$$\max_k \left| \sum_{i=1}^{k} \frac{\nu_{in} - np_i}{\sqrt{np_i}} \right| \xrightarrow{d}_P \max_k |W_k| .$$

Since

$$\mathsf{E}_{\widetilde{P}} \sum_{i=1}^{k} \frac{\nu_{in} - np_i}{\sqrt{np_i}} = \sum_{i=1}^{k} \frac{n\widetilde{p}_i - np_i}{\sqrt{np_i}} = \sum_{i=1}^{k} \frac{\sqrt{n} p_i h_i}{\sqrt{np_i}} = \sum_{i=1}^{k} \sqrt{p_i} h_i = 0$$

if and only if $h_i = 0$ for every i, under contiguous alternatives we have

$$\left\{\sum_{i=1}^{k} \frac{\nu_{in} - np_i}{\sqrt{np_i}}\right\}_{k=1}^{m} \xrightarrow{d}_{\widetilde{P}} \left\{W_k + \sum_{i=1}^{k} \sqrt{p_i} h_i\right\}_{i=1}^{m}.$$

Therefore,

$$\max_k \left|\sum_{i=1}^{k} \frac{\nu_{in} - np_i}{\sqrt{np_i}}\right| \xrightarrow{d}_{\widetilde{P}} \max_k \left|W_k + \sum_{i=1}^{k} \sqrt{p_i} h_i\right|,$$

which is different from $\max\limits_k |W_k|$. However, note that the statistic

$$\max_k \left|\sum_{i=1}^{k} \frac{\nu_{in} - np_i}{\sqrt{np_i}}\right|$$

is not asymptotically distribution free as the distribution of $\{W_k\}_{k=1}^m$ depends on $\{p_i, i = 1, \dots, n\}$, see Section 3.5.

4.3 Formulation of the GOF problems

As stated in Section 3.7, the time transformation $t = F(x)$ is not of any direct use in the construction of asymptotically distribution free GOF tests for testing hypotheses on a distribution F where F is not continuous in $\mathbb{R}$. For that reason, the idea of using other transformations came up. Which properties a transformation proposed on the empirical process should possess was first formulated by **Khmaladze** [40], Section 3. The material of this section is extracted from that paper.

To begin, let us introduce in general the notation P^ν for the distribution of a random process ν or a random variable ν. Let us also use the notation $\mathcal{L}_2(F)$, ignoring the fact that parameters $\boldsymbol{\theta}_0$ or $\hat{\boldsymbol{\theta}}$ may be involved in F. Denote by $\mathscr{J}$ some subset of $\mathcal{L}_2(F)$ and by $\mathbb{L}(\mathscr{J})$ the closed linear span of $\mathscr{J}$.

For two Gaussian processes $\nu = \{\nu(f), f \in \mathscr{J}\}$ and $\eta = \{\eta(f), f \in \mathscr{J}\}$, the distance in total variation of P^{ν} and P^{η} is defined as

$$d(P^{\nu}, P^{\eta}) = \max\left\{d(P^{\nu(f)}, P^{\eta(f)}), f \in \mathbb{L}(\mathscr{J})\right\}.$$

Note that $P^{\nu(f)}$ and $P^{\eta(f)}$ for each f are Gaussian distributions of random variables $\nu(f)$ and $\eta(f)$ respectively.

The distance in total variation of $P^{V_{F_0}}$ and $P^{V_{F_0}+H}$ (see $V_{F_0} + H$ again in (4.15)) is known to be

$$d(P^{V_{F_0}}, P^{V_{F_0}+H}) = K_h,$$

see for example, **Kuo** [49].

It is essential to recall a note in Section 3.7 that the transformation

$$u_n(t) = v_{nF_0}(F_0^{-1}(t)) \tag{4.19}$$

for a continuous distribution F in $\mathbb{R}$ possesses two properties: firstly, the transformed empirical process is asymptotically distribution free; secondly, the limit in distribution of the transformed process under the null and the alternative hypotheses are different. In particular, the limit in distribution of $u_n(t)$ under contiguous alternatives is

$$u(t) + H(F_0^{-1}(t)),$$

where H is defined in (4.14). As the transformation (4.19) is a one-to-one mapping, we also have

$$d(P^{u}, P^{u+H\circ F_0^{-1}}) = K_h.$$

Recall once again that the transformation (4.19) can not be extended to multidimensional cases. Thus, one may set a goal of finding another transformation $\mathscr{K}$ which has the same properties as (4.19). The precise formulation for such $\mathscr{K}$ is stated as follows:

(i) $\mathscr{K}[v_{nF_0}, F_0] \xrightarrow{d}_{\mathbb{F}_0^n} \xi$ and the distribution P^{ξ} does not depend on F_0;

(ii) For any sequence of contiguous alternatives $F_{a,n}$, we have

$$\mathscr{K}[v_{nF_0}, F_0] \xrightarrow{d}_{\mathbb{F}_a^n} \xi'$$

such that $d(P^{\xi}, P^{\xi'}) = K_h$.

A similar formulation is needed for the problem of testing composite parametric hypotheses.

For testing parametric hypotheses of continuous distributions, recall that if the family of distributions $\mathscr{F}$ under the null hypothesis is regular then the limit in distribution, under the null hypothesis, of the function-parametric empirical process $\widehat{v}_{nF_0,\hat{\boldsymbol{\theta}}_n}$ is $\widehat{v}_{F_0}$ defined in (3.36). Define a function $\widehat{H}$ as a projection of H, that is

$$\widehat{H}(w) = H(w) - \langle w, \beta_F \rangle_{F_0} H(\beta_F).$$

Then, the limit in distribution of $\widehat{v}_{nF,\hat{\boldsymbol{\theta}}_n}$ under the contiguous alternatives $F_{a,n}$ is $\widehat{v}_{F_0} + \widehat{H}$. Moreover, if $H(\beta_F) = 0$ then

$$d(P^{\widehat{v}_{F_0}}, P^{\widehat{v}_{F_0}+\widehat{H}}) = K_h.$$

Again, the transformation (4.19) can not be extended to the parametric testing problem or the multidimensional case. The properties of the transformation $\widehat{\mathscr{K}}$, which we are looking for, of the parametric empirical process and of the hypothetical family of distribution $\mathscr{F}$ are formulated as follows:

(i) For each $\theta \in \Theta_0$, we have $\widehat{\mathscr{K}}[\widehat{v}_{nF_0,\hat{\boldsymbol{\theta}}_n}, \mathscr{F}] \xrightarrow{d}_{\mathbb{F}_0^n} \xi$ and P^{ξ} does not depend on $\mathscr{F}$ if $\mathscr{F}$ is regular;

(ii) For any sequence of contiguous alternatives $F_{a,n}$, we have

$$\widehat{\mathscr{K}}[v_{nF_0,\hat{\boldsymbol{\theta}}_n}, \mathscr{F}] \xrightarrow{d}_{\mathbb{F}_a^n} \xi'$$

such that $d(P^{\xi}, P^{\xi'}) = K_h$.

There are only two transformations known satisfying the two properties. The first one was known as the **Khmaladze transformation** (see **Khmaladze** [38], [39] [40]), which works for testing both simple and parametric hypotheses for continuous distributions in multidimensional space. The second transformation was also invented by **Khmaladze** [42],[44] recently. This new transformation is distinct from the first one and can be applied for both discrete and continuous distributions and for simple as well as parametric testing problems.

Applications of the Khmaladze-2 transformation will be presented in the two following chapters 5 and 6. The two problems are parametric hypothesis testing, for discrete distributions in one case and for the tail of continuous distributions on the other.

Chapter 5

Testing independence of two discrete random variables

The content of this chapter is extracted from **Nguyen** [58].

5.1 Introduction

The main aim of this chapter is to give a construction of a class of asymptotically distribution free GOF tests for testing independence in 2-way contingency tables. This problem is equivalent to testing the independence of two discrete random variables as the entries of the contingency tables can be either categorical frequencies or just simply numerical random variables. The problem of testing the independence of two discrete random vectors will also be treated in the same way as well. For consistency, we will use the term "contingency table" throughout this chapter.

The classical problem of testing independence in contingency tables has long been under consideration. However, there has existed essentially only one distribution free GOF test as a tool, the chi-square test. Various modifications of the chi-square tests have been employed to adapt to different circumstances. This could be found in **Haberman** [31], **Gilula** [26], **Bedrick** [4], **Holt** [35], **Koch et al.** [45], **Rao and Scott** [69] with references therein.

Suppose that we need to test the independence of two discrete random variables X and Y. Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be independent copies of (X, Y) which are the input of a table with $I + 1$ classifications of values for X and $J + 1$ classifications for Y. For each cell (i, j) of the table where $i = 1, \dots, I + 1; j = 1, \dots, J + 1$, denote by ν_{ij} the **cell counts** or **frequencies**, so that we have $\sum_{i,j} \nu_{ij} = n$. Let us denote by p_{ij} the probability that an observed value of (X, Y) is in the cell (i, j). Denote by $\{a_i\}$ and $\{b_j\}$ the marginal distributions of X and Y. Put

$$\mathbf{a} = (a_1, \dots, a_I)^T \in \mathbb{R}^I,$$
$$\mathbf{b} = (b_1, \dots, b_J)^T \in \mathbb{R}^J.$$

In some cases when the number of observations n is considered as random - most of the time a Poisson random variable - then the frequencies will also become independent Poisson random variables with some intensities γ_{ij}. We only consider n known and fixed, so the distribution of $\{\nu_{ij}\}$ is not Poisson but multinomial. The null hypothesis H_0 and the alternative H_a are stated as

$$H_0 : X \text{ and } Y \text{ are independent,}$$
$$H_a : X \text{ and } Y \text{ are dependent.}$$

The hypothesis H_0 is true when

$$p_{ij} = a_i b_j \qquad \text{for all} \quad i, j.$$

In other words, the conditional distributions under H_0 are

$$p_{i|j} = a_i,\ p_{j|i} = b_j \text{ for all } i, j.$$

For this reason, the independence of X and Y is often referred to as **homogeneity** of the conditional distributions.

Testing H_0 becomes a parametric testing problem if we view the marginal distributions $\mathbf{a}$ and $\mathbf{b}$ as parameters. However, the dimension of the parameters, $I + J$, will typically be large.

5.2 Literature review

The form of the empirical process is not exactly the same as presented in chapter 3 since it involves two marginal distributions. We will see the detail below.

5.2.1 MLEs of marginal distributions

Denote by $\boldsymbol{\theta} = (\mathbf{a}^T, \mathbf{b}^T)^T \in \mathbb{R}^d$ where $d = I + J$ the vector of parameters. Denote by $\boldsymbol{\theta}_0$ the true unknown parameter. It is well-known that the MLEs for $\boldsymbol{\theta}$ with given frequencies $\{\nu_{ij}\}$ are the sample marginal proportions or relative frequencies:

$$\hat{a}_i = \frac{\nu_{i+}}{n} = \frac{\sum_{j=1}^{(J+1)} \nu_{ij}}{n}, \quad \hat{b}_j = \frac{\nu_{+j}}{n} = \frac{\sum_{i=1}^{(I+1)} \nu_{ij}}{n}, \quad \text{for all } \ i, j. \tag{5.1}$$

Denote by $\hat{\boldsymbol{\theta}} = (\hat{a}_1, \ldots, \hat{a}_I, \hat{b}_1, \ldots, \hat{b}_J)^T$ the vector of the estimated parameters. Clearly, under the null hypothesis H_0, the estimated joint probabilities are

$$\hat{p}_{ij} = p_{ij}(\hat{\boldsymbol{\theta}}) = \hat{a}_i \hat{b}_j.$$

5.2.2 The empirical process for testing independence

In general, for testing independence of two random variables X and Y with respective marginal distributions $G(x)$ and $H(y)$, we consider the empirical process

$$v_n(x, y) = \sqrt{n}[F_n(x, y) - G_n(x)H_n(y)].$$

Here,

$$F_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \le x, Y_i \le y\}}$$

is the empirical distribution function and

$$G_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \le x\}},$$
$$H_n(y) = \frac{1}{n}\sum_{i=1}^{n} I_{\{Y_i \le y\}},$$

are marginal empirical distribution functions. If X and Y are continuous random variables, then the limit in distribution of the empirical process $v_n(x, y)$ is known to be a Brownian sheet (see for example **van de Vaart and Wellner** [76], section 3.8). And it is known that any statistic from $v_n(x, y)$ which is invariant under the time transformation $t = G(x)$ and $s = H(y)$ will be distribution free, i.e., the distribution of such a statistic will be the same for all continuous distributions G and H. This fact is no longer true when X or Y or both X and Y are discrete. It is also not valid when either X or Y is a multidimensional random variable.

5.2.3 Components of the chi-square statistic

Without loss of generality, assume that the input of the contingency tables, i.e., values of X and Y, are correspondingly enumerated by $1, \ldots, I+1$ and $1, \ldots, J+1$. Denote by

$$\widehat{T}_{n,ij} = \frac{\Delta^2 v_n(i, j)}{\sqrt{\Delta G_n(i)\Delta H_n(j)}} \tag{5.2}$$

the normalized second increments of the empirical process, where

$$\Delta^2 v_n(i, j) = v_n(i, j) - v_n(i-1, j) - v_n(i, j-1) + v_n(i-1, j-1),$$
$$\Delta G_n(i) = G_n(i) - G_n(i-1),\ \Delta H_n(j) = H_n(j) - H_n(j-1).$$

Put $\widehat{T}_n = (\widehat{T}_{n,ij})$. We can rewrite $\widehat{T}_n$ in a more conventional way. Specifically, denote by

$$T_{n,ij} = \frac{\nu_{ij} - np_{ij}(\boldsymbol{\theta}_0)}{\sqrt{np_{ij}(\boldsymbol{\theta}_0)}}$$

the components of T_n, then $\widehat{T}_{n,ij}$ is in fact the estimator of $T_{n,ij}$, i.e.,

$$\widehat{T}_{n,ij} = \frac{\nu_{ij} - np_{ij}(\hat{\boldsymbol{\theta}})}{\sqrt{np_{ij}(\hat{\boldsymbol{\theta}})}} = \frac{\nu_{ij} - n\hat{a}_i\hat{b}_j}{\sqrt{n\hat{a}_i\hat{b}_j}}. \tag{5.3}$$

Following **Khmaladze** [40] we shall call $\widehat{T}_n$ a **vector of components of the chi-square statistic**. As we know, the conventional chi-square test is of the form

$$\chi_n^2 = \sum_{i,j} \widehat{T}_{n,ij}^2 = \sum_{i,j} \frac{(\nu_{ij} - n\hat{a}_i\hat{b}_j)^2}{n\hat{a}_i\hat{b}_j} = \sum_{i,j} \frac{(\Delta^2 v_n(i,j))^2}{\Delta G_n(i)\Delta H_n(j)}.$$

Here and below, the notation $\sum_{i,j}$ means $\sum_{i=1}^{I+1}\sum_{j=1}^{J+1}$.

5.2.4 The space and limit theorem

In this section, in order to avoid any possible confusion with the operator presented later on, we will clarify the space to which $T_n, \widehat{T}_n$ belong. The limit in distribution of $\widehat{T}_n$ is essential for our method but may not be seen clearly from sections 3.4 and 3.5. Hence, we recall it here.

We consider T_n and $\widehat{T}_n$ not as matrices but as functions of two variables i and j where $i \in \mathcal{I} = \{1, \ldots, I+1\}, j \in \mathcal{J} = \{1, \ldots, J+1\}$. The space $\mathcal{C}^2_{\mathcal{I}\times\mathcal{J}}$ of such functions is equipped with the inner product and the norm defined as usual, i.e.,

$$\langle V, W \rangle = \sum_{i,j} V_{ij}W_{ij}, \qquad \|V\| = \sqrt{\sum_{i,j} V_{ij}^2} \qquad \text{for } V, W \in \mathcal{C}^2_{\mathcal{I}\times\mathcal{J}}.$$

The limit in distribution of $\widehat{T}_n$ can be deduced from Theorem 3.2. To begin, we recall that the asymptotically linear representation of the MLE $\hat{\boldsymbol{\theta}}$ holds as usual, that is,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \Gamma^{-1} \sum_{i,j} T_{ij} \frac{\dot{p}_{ij}(\boldsymbol{\theta}_0)}{\sqrt{p_{ij}(\boldsymbol{\theta}_0)}} + o_P(1), \tag{5.4}$$

where $\dot{p}_{ij}(\boldsymbol{\theta})$ is the vector of partial derivatives of p_{ij} with respect to $\boldsymbol{\theta}$ and Γ is the Fisher information matrix. Consequently we have the following representation of $\widehat{T}_n$:

$$\begin{aligned}\widehat{T}_{n,ij} &= T_{n,ij} - \frac{\dot{p}_{ij}(\boldsymbol{\theta}_0)^T}{\sqrt{p_{ij}(\boldsymbol{\theta}_0)}}\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + o_P(1)\\ &= T_{n,ij} - \frac{\dot{p}_{ij}(\boldsymbol{\theta}_0)^T}{\sqrt{p_{ij}(\boldsymbol{\theta}_0)}}\Gamma^{-1}\sum_{i',j'} T_{n,i'j'}\frac{\dot{p}_{i'j'}(\boldsymbol{\theta}_0)}{\sqrt{p_{i'j'}(\boldsymbol{\theta}_0)}} + o_P(1). \qquad (5.5)\end{aligned}$$

Denote $\beta^{(0)} = \sqrt{p} = (\sqrt{p_{ij}})$. Denote by β the vectors of normalized score functions. Components of β are

$$\beta_{ij} = \Gamma^{-1/2}\frac{\dot{p}_{ij}(\boldsymbol{\theta}_0)}{\sqrt{p_{ij}(\boldsymbol{\theta}_0)}}, \qquad \text{for all} \quad i, j, \qquad (5.6)$$

Since the dimension of the parameter is d, we have that β is a collection of d functions $\beta^{(1)}, \ldots, \beta^{(I)}, \beta^{(I+1)}, \ldots, \beta^{(d)}$ in $\mathcal{C}^2_{\mathcal{I}\times\mathcal{J}}$. The first $\beta^{(1)}$ consists of components which are based on derivatives with respect to the first element a_1 of $\boldsymbol{\theta}$, $\beta^{(I+1)}$ is the one based on derivatives with respect to b_1 and so on. From their definition in (5.6), it follows that $\beta^{(1)}, \ldots, \beta^{(d)}$ are orthonormal and each of them is orthogonal to $\beta^{(0)} = \sqrt{p}$.

Denote by $V = (V_{ij})$ an element of $\mathcal{C}^2_{\mathcal{I}\times\mathcal{J}}$ such that all V_{ij} are independent and standard normal random variables, then as stated in **Khmaladze** [42], $\widehat{T}_n$ converges in distribution to $\widehat{T}$, which is the projection of V orthogonal to the subspace generated by $\{\beta^{(0)}, \beta^{(1)}, \ldots, \beta^{(d)}\}$, i.e., we have

$$\widehat{T} = V - \sum_{\alpha=0}^{d}\langle V, \beta^{(\alpha)}\rangle\beta^{(\alpha)}. \qquad (5.7)$$

The explicit forms of $\beta^{(\alpha)}$ will be seen below.

5.3 The transformation of $\widehat{T}_n$

The limit in distribution of $\widehat{T}_n$ varies from case to case and depends on the hypothetical parameter $\boldsymbol{\theta}_0$. However, if we apply the transformation

Khmaladze-2, which is a one-to-one mapping, $\widehat{T}_n$ will be turned into $\widehat{Z}_n$ (see (5.12) below) with the specified limit in distribution to be the following:

$$\widehat{Z} = V' - \sum_{\alpha=0}^{d} \langle V', r^{(\alpha)} \rangle r^{(\alpha)}. \tag{5.8}$$

Here $V' = (V'_{ij})$, like V in (5.7), is an element of $\mathcal{C}^2_{\mathcal{I}\times\mathcal{J}}$ in which V'_{ij} are all independent and standard normal random variables. Specifically, the collection $\{r^{(0)}, \dots, r^{(d)}\}$ is an orthonormal collection of functions in $\mathcal{C}^2_{\mathcal{I}\times\mathcal{J}}$, which can be freely chosen by the users but is fixed. As demonstrated in **Khmaladze** [42], the transformation has both explicit and recursive forms, but because of its computational convenience, the latter will be used. Generally, consider a unitary operator $U_{\beta,r}$ of the form

$$U_{\beta,r} = I - \frac{1}{1 - \langle \beta, r \rangle}(r-\beta)(r-\beta)^T. \tag{5.9}$$

If β and r are two functions of unit norm then it is easy to check that $U_{\beta,r}\beta = r, U_{\beta,r}r = \beta$ and $U_{\beta,r}v = v$ for every $v \perp \beta, r$.

Consider the operator $U_{\beta^{(0)},r^{(0)}}$. Obviously,

$$U_{\beta^{(0)},r^{(0)}}\beta^{(0)} = r^{(0)}, U_{\beta^{(0)},r^{(0)}}r^{(0)} = \beta^{(0)}.$$

Denote by $\widetilde{\beta}^{(1)}$ the image of $\beta^{(1)}$ via $U_{\beta^{(0)},r^{(0)}}$. Since $U_{\beta^{(0)},r^{(0)}}$ preserves the inner product, the images of $\beta^{(0)}$ and $\beta^{(1)}$ are orthogonal, which means $r^{(0)} \perp \widetilde{\beta}^{(1)}$. Hence

$$U_{\widetilde{\beta}^{(1)},r^{(1)}}r^{(0)} = r^{(0)}, \; U_{\widetilde{\beta}^{(1)},r^{(1)}}\widetilde{\beta}^{(1)} = r^{(1)}.$$

In summary, by applying the composition $U_{\widetilde{\beta}^{(1)},r^{(1)}}U_{\beta^{(0)},r^{(0)}}$ to $\beta^{(0)}, \beta^{(1)}, \beta^{(2)}$, we get their images $r^{(0)}, r^{(1)}, \widetilde{\beta}^{(2)}$ respectively, where $\widetilde{\beta}^{(2)} \perp r^{(0)}, r^{(1)}$.

Continuing this process, generally, we can define $\widetilde{\beta}^{(\tau)}, \tau \geq 2$ recursively as

$$\widetilde{\beta}^{(\tau)} = \left(\prod_{1\leq\kappa<\tau} U_{\widetilde{\beta}^{(\kappa)},r^{(\kappa)}}U_{\beta^{(0)},r^{(0)}} \right) \beta^{(\tau)}. \tag{5.10}$$

Then define an operator $\mathbb{U}$ based on $\widetilde{\beta}^{(\tau)}$ in (5.10) by

$$\mathbb{U} = \prod_{\tau=1}^{d} U_{\widetilde{\beta}^{(\tau)},r^{(\tau)}} U_{\beta^{(0)},r^{(0)}}. \tag{5.11}$$

As a product of unitary operators, $\mathbb{U}$ is a unitary operator. The role of this operator is expressed in the following theorem.

Theorem 5.1. *The unitary operator $\mathbb{U}$ satisfies $\mathbb{U}\beta^{(\alpha)} = r^{(\alpha)}$ for all $\alpha = 0, \ldots, d$. If the operator $\mathbb{U}$ transforms $\widehat{T}_n$ into $\widehat{Z}_n$, i.e.,*

$$\widehat{Z}_n = \mathbb{U}\widehat{T}_n = \left(\prod_{\tau=1}^{d} U_{\widetilde{\beta}^{(\tau)},r^{(\tau)}} U_{\beta^{(0)},r^{(0)}}\right) \widehat{T}_n, \tag{5.12}$$

then $\widehat{Z}_n$ converges in distribution to $\widehat{Z}$ of the form given in (5.8).

This theorem is extracted from the last Remark in **Khmaladze** [42]. We will give a short proof here.

Proof. It is easy to see the first statement that $\mathbb{U}$ actually maps $\beta^{(\alpha)}$ into $r^{(\alpha)}$ for all α. We shall now show that $\widehat{Z}_n$ obtained by the equation (5.12) converges in distribution to $\widehat{Z}$.

Since $\widehat{T}_n \xrightarrow{d} \widehat{T}$, as an image of $\widehat{T}_n$ via a linear transformation, $\widehat{Z}_n = \mathbb{U}\widehat{T}_n$ will have the limit in distribution $\widehat{Z} = \mathbb{U}\widehat{T}$ whose form was given in (5.8). Indeed,

$$\begin{aligned}\widehat{Z} = \mathbb{U}\widehat{T} = \mathbb{U}[V - \sum_{\alpha=0}^{d} \langle V, \beta^{(\alpha)}\rangle \beta^{(\alpha)}] &= \mathbb{U}V - \sum_{\alpha=0}^{d} \langle V, \beta^{(\alpha)}\rangle \mathbb{U}\beta^{(\alpha)} \\ = V' - \sum_{\alpha=0}^{d} \langle V, \beta^{(\alpha)}\rangle r^{(\alpha)} &= V' - \sum_{\alpha=0}^{d} \langle V', r^{(\alpha)}\rangle r^{(\alpha)}.\end{aligned}$$

Here V' is the image of V via the unitary operator $\mathbb{U}$, which guarantees that V'_{ij} are independent and standard normal random variables, like V_{ij}. The reason why we have $\langle V, \beta^{(\alpha)}\rangle = \langle V', r^{(\alpha)}\rangle$ for all α is again because $\mathbb{U}$ is unitary, and hence preserves inner products. □

5.4 Explicit forms of matrices $\Gamma, \Gamma^{-1}, \Gamma^{-1/2}$

We present the explicit form of the Fisher information matrix, its inverse matrix and the square root of the inverse matrix. These explicit presentations help to derive the explicit form of the normalized score functions $\beta^{(i)}, i = 1, \ldots, d$.

5.4.1 The Fisher information matrix Γ

Recall that the Fisher information matrix Γ of dimension $d \times d$ is defined as

$$\Gamma = \sum_{i,j} \frac{\dot{p}_{ij}(\boldsymbol{\theta}_0)\dot{p}_{ij}(\boldsymbol{\theta}_0)^T}{p_{ij}(\boldsymbol{\theta}_0)} = \sum_{i,j} \frac{\dot{p}_{ij}(\boldsymbol{\theta}_0)}{\sqrt{p_{ij}(\boldsymbol{\theta}_0)}} \left(\frac{\dot{p}_{ij}(\boldsymbol{\theta}_0)}{\sqrt{p_{ij}(\boldsymbol{\theta}_0)}} \right)^T .$$

We shall see that under the null hypothesis, the following block matrix form is valid for Γ:

$$\Gamma = \left(\begin{array}{c|c} \Gamma_{\mathbf{a}} & \mathbf{0} \\ \hline \mathbf{0} & \Gamma_{\mathbf{b}} \end{array} \right). \tag{5.13}$$

In fact, when $p_{ij} = a_i b_j$ for every i, j, note that $a_{I+1} = 1 - \sum_{i=1}^{I} a_i, \; b_{J+1} = 1 - \sum_{j=1}^{J} b_j$, we simply have

$$\frac{\partial p_{z_1 z_2}/\partial a_i}{\sqrt{p_{z_1 z_2}}} = \sqrt{\frac{b_{z_2}}{a_{z_1}}} \left[I_{\{z_1 = i\}} - I_{\{z_1 = I+1\}} \right], \quad i = 1, \ldots, I, \tag{5.14}$$

and

$$\frac{\partial p_{z_1 z_2}/\partial b_j}{\sqrt{p_{z_1 z_2}}} = \sqrt{\frac{a_{z_1}}{b_{z_2}}} \left[I_{\{z_2 = j\}} - I_{\{z_2 = J+1\}} \right], \quad j = 1, \ldots, J. \tag{5.15}$$

Here and below, z_1, z_2 are sub-indices, $z_1 \in \mathcal{I}, z_2 \in \mathcal{J}$ and $I_{\{z=i\}}$ is the indicator function, $I_{\{z=i\}} = 1$ if $z = i$ and $I_{\{z=i\}} = 0$ otherwise. Then, if $i \neq i'$ we have

$$\sum_{z_1, z_2} \frac{\partial p_{z_1 z_2}/\partial a_i}{\sqrt{p_{z_1 z_2}}} \frac{\partial p_{z_1 z_2}/\partial a_{i'}}{\sqrt{p_{z_1 z_2}}} = \sum_{z_1, z_2} \frac{b_{z_2}}{a_{z_1}} I_{\{z_1 = I+1\}} = \frac{1}{a_{I+1}}.$$

If $i = i'$ then

$$\sum_{z_1,z_2} \frac{\partial p_{z_1z_2}/\partial a_i}{\sqrt{p_{z_1z_2}}} \frac{\partial p_{z_1z_2}/\partial a_{i'}}{\sqrt{p_{z_1z_2}}} = \sum_{z_1,z_2} \frac{b_{z_2}}{a_{z_1}} [I_{\{z_1=I+1\}} + I_{\{z_1=i\}}] = \frac{1}{a_i} + \frac{1}{a_{I+1}}.$$

For all i, j, it is easy to see that

$$\sum_{z_1,z_2} \frac{\partial p_{z_1z_2}/\partial a_i}{\sqrt{p_{z_1z_2}}} \frac{\partial p_{z_1z_2}/\partial b_j}{\sqrt{p_{z_1z_2}}} = 0,$$

which explains why we have two zero blocks in the expression of Γ. Using the notation $\mathbf{1}_I = (1, \ldots, 1)^T \in \mathbb{R}^I$ and $D(\mathbf{a}) = \text{diag}(a_1, \ldots, a_I)$, and a similar notation for $D(1/\mathbf{a}), D(\sqrt{\mathbf{a}}), \ldots$, we have

$$\Gamma_{\mathbf{a}} = D\left(\frac{1}{\mathbf{a}}\right) + \frac{1}{a_{I+1}} \mathbf{1}_I \mathbf{1}_I^T, \qquad \Gamma_{\mathbf{b}} = D\left(\frac{1}{\mathbf{b}}\right) + \frac{1}{b_{J+1}} \mathbf{1}_J \mathbf{1}_J^T.$$

5.4.2 Matrix Γ^{-1}

It is well-known that a block matrix as Γ of the form in (5.13) will have its inverse and square root of the same form. It is quite easy to verify that *the inverse matrix of the Fisher information matrix is* $\Gamma^{-1} = \left(\begin{array}{c|c} \Gamma_{\mathbf{a}}^{-1} & \mathbf{0} \\ \hline \mathbf{0} & \Gamma_{\mathbf{b}}^{-1} \end{array}\right)$ *where*

$$\Gamma_{\mathbf{a}}^{-1} = D(\mathbf{a}) - \mathbf{a}\mathbf{a}^T, \qquad \Gamma_{\mathbf{b}}^{-1} = D(\mathbf{b}) - \mathbf{b}\mathbf{b}^T. \tag{5.16}$$

Indeed,

$$\begin{aligned} \Gamma_{\mathbf{a}}\Gamma_{\mathbf{a}}^{-1} &= \left[D\left(\frac{1}{\mathbf{a}}\right) + \frac{1}{a_{I+1}} \mathbf{1}_I \mathbf{1}_I^T\right] \left[D(\mathbf{a}) - \mathbf{a}\mathbf{a}^T\right] \\ &= D\left(\frac{1}{\mathbf{a}}\right) D(\mathbf{a}) + \frac{1}{a_{I+1}} \mathbf{1}_I \mathbf{1}_I^T D(\mathbf{a}) - D\left(\frac{1}{\mathbf{a}}\right) \mathbf{a}\mathbf{a}^T - \frac{1}{a_{I+1}} \mathbf{1}_I \mathbf{1}_I^T \mathbf{a}\mathbf{a}^T \\ &= \mathbb{I}_I + \frac{1}{a_{I+1}} \mathbf{1}_I \mathbf{a}^T - \mathbf{1}_I \mathbf{a}^T - \frac{1}{a_{I+1}} (1 - a_{I+1}) \mathbf{1}_I \mathbf{a}^T = \mathbb{I}_I, \end{aligned}$$

where $\mathbb{I}_I$ is the notation for the identity matrix of size $I \times I$. A similar argument is true for $\Gamma_{\mathbf{b}}^{-1}$ which proves (5.16).

5.4.3 Matrix $\Gamma^{-1/2}$

We have the following forms for matrix $\Gamma^{-1/2}$:

$$\Gamma_{\mathbf{a}}^{-1/2} = [\mathbb{I}_I - c_{\mathbf{a}}\sqrt{\mathbf{a}}\sqrt{\mathbf{a}}^T]D(\sqrt{\mathbf{a}}), \quad \Gamma_{\mathbf{b}}^{-1/2} = [\mathbb{I}_J - c_{\mathbf{b}}\sqrt{\mathbf{b}}\sqrt{\mathbf{b}}^T]D(\sqrt{\mathbf{b}}) \quad (5.17)$$

where $c_{\mathbf{a}}, c_{\mathbf{b}}$ are constants to be determined. This is also easy to check because

$$\begin{aligned}\left(\Gamma_{\mathbf{a}}^{-1/2}\right)^T \Gamma_{\mathbf{a}}^{-1/2} &= D(\sqrt{\mathbf{a}})[\mathbb{I}_I - c_{\mathbf{a}}\sqrt{\mathbf{a}}\sqrt{\mathbf{a}}^T][\mathbb{I}_I - c_{\mathbf{a}}\sqrt{\mathbf{a}}\sqrt{\mathbf{a}}^T]D(\sqrt{\mathbf{a}}) \\ &= D(\sqrt{\mathbf{a}})\left[\mathbb{I}_I - 2c_{\mathbf{a}}\sqrt{\mathbf{a}}\sqrt{\mathbf{a}}^T + c_{\mathbf{a}}^2(1-a_{I+1})\sqrt{\mathbf{a}}\sqrt{\mathbf{a}}^T\right]D(\sqrt{\mathbf{a}}) \\ &= D(\mathbf{a}) - \left[2c_{\mathbf{a}} - c_{\mathbf{a}}^2(1-a_{I+1})\right]D(\sqrt{\mathbf{a}})\sqrt{\mathbf{a}}\sqrt{\mathbf{a}}^T D(\sqrt{\mathbf{a}}) \\ &= D(\mathbf{a}) - \left[2c_{\mathbf{a}} - c_{\mathbf{a}}^2(1-a_{I+1})\right]\mathbf{a}\mathbf{a}^T = D(\mathbf{a}) - \mathbf{a}\mathbf{a}^T = \Gamma_{\mathbf{a}}^{-1}\end{aligned}$$

if and only if $2c_{\mathbf{a}} - c_{\mathbf{a}}^2(1-a_{I+1}) = 1$ which implies that $c_{\mathbf{a}} = \frac{1}{1\pm\sqrt{a_{I+1}}}$. Similarly, $c_{\mathbf{b}} = \frac{1}{1\pm\sqrt{b_{J+1}}}$.

5.4.4 Normalized score functions

Substituting $\dot{p}_{ij}(\boldsymbol{\theta}_0)/\sqrt{p_{ij}(\boldsymbol{\theta}_0)}$ as in (5.14) and (5.15) and the explicit formula for $\Gamma^{-1/2}$ as in (5.17)) into (5.6) we get the explicit form of $\beta^{(\alpha)}$. Since matrix $\Gamma^{-1/2}$ involves $c_{\mathbf{a}}$ and $c_{\mathbf{b}}$, each of which has two possible values, there are 4 possible collections of $\beta^{(\alpha)}$ depending on the sign of $c_{\mathbf{a}}$ and $c_{\mathbf{b}}$. Specifically,

- for $1 \le \alpha \le I$, we have

$$\beta_{z_1z_2}^{(\alpha)} = \sqrt{\frac{b_{z_2}}{a_{z_1}}}\left[\sqrt{a_\alpha}I_{\{z_1=\alpha\}} - \frac{1}{1+\sqrt{a_{I+1}}}\sqrt{a_\alpha}a_{z_1}I_{\{z_1\neq I+1\}} - I_{\{z_1=I+1\}}\sqrt{a_{I+1}}\sqrt{a_\alpha}\right], \quad (5.18a)$$

or

$$\beta_{z_1z_2}^{(\alpha)} = \sqrt{\frac{b_{z_2}}{a_{z_1}}}\left[\sqrt{a_\alpha}I_{\{z_1=\alpha\}} - \frac{1}{1-\sqrt{a_{I+1}}}\sqrt{a_\alpha}a_{z_1}I_{\{z_1\neq I+1\}} + I_{\{z_1=I+1\}}\sqrt{a_{I+1}}\sqrt{a_\alpha}\right]. \quad (5.18b)$$

- for $I < \alpha \leq I + J$ we have

$$\beta^{(\alpha)}_{z_1 z_2} = \sqrt{\frac{a_{z_1}}{b_{z_2}}}\big[\sqrt{b_{\alpha-I}} I_{\{z_2=\alpha-I\}} - \frac{1}{1+\sqrt{b_{J+1}}}\sqrt{b_{\alpha-I}} b_{z_2} I_{\{z_2 \neq J+1\}} \\ - I_{\{z_2=J+1\}}\sqrt{b_{J+1}}\sqrt{b_{\alpha-I}}\big], \qquad (5.19a)$$

or

$$\beta^{(\alpha)}_{z_1 z_2} = \sqrt{\frac{a_{z_1}}{b_{z_2}}}\big[\sqrt{b_{\alpha-I}} I_{\{z_2=\alpha-I\}} - \frac{1}{1-\sqrt{b_{J+1}}}\sqrt{b_{\alpha-I}} b_{z_2} I_{\{z_2 \neq J+1\}} \\ + I_{\{z_2=J+1\}}\sqrt{b_{J+1}}\sqrt{b_{\alpha-I}}\big]. \qquad (5.19b)$$

5.5 Simulation results for distribution free property

We shall demonstrate in this section that a new GOF test based on $\widehat{Z}_n$ converges quickly to its limit in distribution and that, even for a relatively small sample size n, the asymptotic distribution of the new test statistics does not depend on the unknown hypothetical parameters $\boldsymbol{\theta}_0$.

5.5.1 Forms of GOF tests

For $i = 1, \ldots, I+1$ and $j = 1, \ldots, J+1$, let

$$V^T_{n,ij} = \sum_{(t_1,t_2) \leq (i,j)} \widehat{T}_{t_1 t_2}, \qquad V^Z_{n,ij} = \sum_{(z_1,z_2) \leq (i,j)} \widehat{Z}_{z_1 z_2}, \qquad (5.20)$$

be cumulative sums of coordinates of $\widehat{T}_n$ and $\widehat{Z}_n$ respectively, where, $(t_1, t_2) \leq (i, j)$ means $t_1 \leq i$ and $t_2 \leq j$. Put

$$V^T_n = (V^T_{n,ij}), \qquad V^Z_n = (V^Z_{n,ij});$$

and let $W = (W_{ij})$ where

$$W_{ij} = \sum_{(t,s)\le(i,j)} V_{ts}, \quad i = 1, \dots, I+1,\ j = 1, \dots, J+1,$$

in which again $V = (V_{ij}) \in \mathcal{C}^2_{\mathcal{I}\times\mathcal{J}}$, where V_{ij} are independent and standard normal random variables. From the forms of $\widehat{T}$ and $\widehat{Z}$ as in (5.7) and (5.8), we see that W is an analog of the trajectory of a Brownian motion in 2-dimensional time. Therefore, the asymptotic behaviours of V_n^T and V_n^Z are somewhat similar to Brownian bridges. More precisely, they are projected Brownian motions (see 3.36). The main difference between V_n^T and V_n^Z is that the limiting projected Brownian motion of V_n^Z has a fully prescribed distribution, while that of V_n^T depends on parameters $\{a_i\}$ and $\{b_j\}$.

Recall that the limit in distribution of $\widehat{Z}_n$ could be chosen by the users, which means that one can freely choose any collection $\left\{r^{(0)}, r^{(1)}, \dots, r^{(d)}\right\}$ provided that they are mutually orthonormal. For that reason, we chose this collection to be the specific $\left\{\beta^{(\alpha)}, \alpha = 0, \dots, d\right\}$, given by (5.18a) and (5.19a) computed in the discrete uniform case with respect to $a_i = \frac{1}{I+1}$, $b_j = \frac{1}{J+1}$ for all i, j. This choice seems natural to us since it can play the role of the one particular testing problem in the class of various parametric testing problems for given $(I+1)\times(J+1)$ tables. Let us see explicitly what $r^{(\alpha)}$ are. Simply choose

$$c_{a0} = \frac{\sqrt{I+1}}{1+\sqrt{I+1}}, \qquad c_{b0} = \frac{\sqrt{J+1}}{1+\sqrt{J+1}},$$

i.e., the " $+$ " sign is chosen (to be consistent with choosing β^α from (5.18a) and (5.19a)). Then

$$r^{(0)} = \left(\frac{1}{\sqrt{(I+1)(J+1)}}\right). \tag{5.21}$$

Other $r^{(\alpha)}, \alpha = 1, \dots, d$, have components as follows: for $\alpha \le I$,

$$r^{(\alpha)}_{z_1 z_2} = \frac{1}{\sqrt{J+1}}\left[I_{\{z_1=\alpha\}} - \frac{I_{\{z_1\neq I+1\}}}{\sqrt{I+1}(1+\sqrt{I+1})} - \frac{1}{\sqrt{I+1}} I_{\{z_1=I+1\}}\right], \tag{5.22}$$

and for $\alpha \geq I+1$,

$$r^{(\alpha)}_{z_1 z_2} = \frac{1}{\sqrt{I+1}} \left[I_{\{z_2=\alpha-I\}} - \frac{I_{\{z_2 \neq J+1\}}}{\sqrt{J+1}(1+\sqrt{J+1})} - \frac{1}{\sqrt{J+1}} I_{\{z_2=J+1\}} \right]. \tag{5.23}$$

We are choosing for demonstration two common GOF test statistics, which are the discrete versions of the Kolmogorov-Smirnov (KS) statistic and the omega-square (Ω^2) statistic, as follows:

$$KS = \max_{(1,1)\leq(i,j)\leq(I+1,J+1)} \left| V^Z_{n,ij} \right|, \tag{5.24}$$

$$\Omega^2 = \sum_{(1,1)\leq(i,j)\leq(I+1,J+1)} (V^Z_{n,ij})^2. \tag{5.25}$$

5.5.2 Distribution free property of the new GOF tests

These two test statistics, KS and Ω^2, will be shown to be asymptotically distribution free. To demonstrate this, we produced the cumulative distribution functions for various arbitrarily selected parameters $\boldsymbol{\theta}_0$ and plotted them in the same figure for each type of statistic.

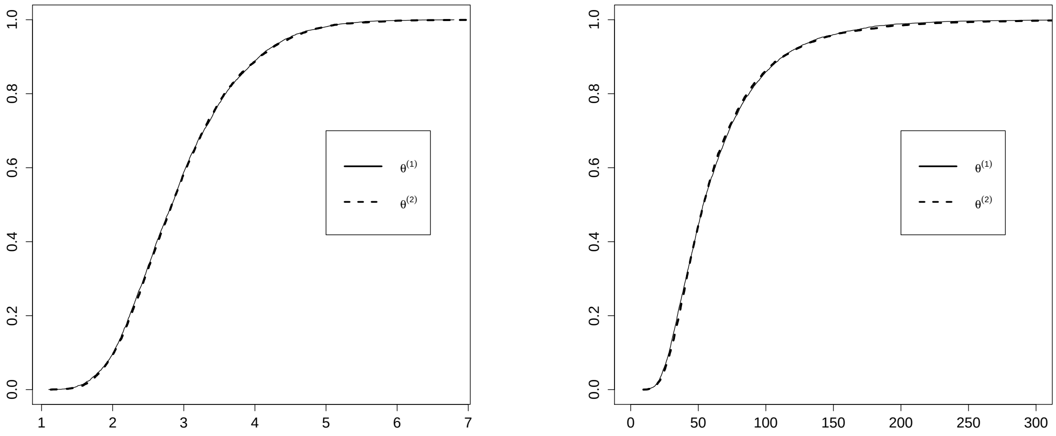
Firstly, we choose the sample size to be $n = 500$. We would expect that for tables of small dimension, this sample size is big enough for the distributions of the test statistics to reach their limits. As shown from our simulation, this suspicion is true for tables of dimension 7×7 or even more.

To create Figure 5.1, we chose the sample size $n = 500$ and table dimension 7×7, running the simulation for 5000 iterations to build up the cumulative distribution functions of KS and Ω^2 given by (5.24) and (5.25). As examples, we chose the hypothetical parameters to be

$$\boldsymbol{\theta}^{(1)} = (0.03, 0.05, 0.04, 0.17, 0.1, 0.29, 0.32, 0.04, 0.05, 0.05, 0.1, 0.32, 0.23, 0.21)^T$$

and

$$\boldsymbol{\theta}^{(2)} = (0.135, 0.25, 0.1, 0.12, 0.065, 0.23, 0.1, 0.15, 0.15, 0.1, 0.1, 0.23, 0.17, 0.1)^T.$$



(a) *Distribution functions of the KS statistics, sample size $n = 500$, with two different sets of parameters*

(b) *Distribution functions of the Ω^2 statistics, sample size $n = 500$, with two different sets of parameters*

Figure 5.1: Distributions of new test statistics based on $\widehat{Z}_n$ in the limit for 7×7 tables

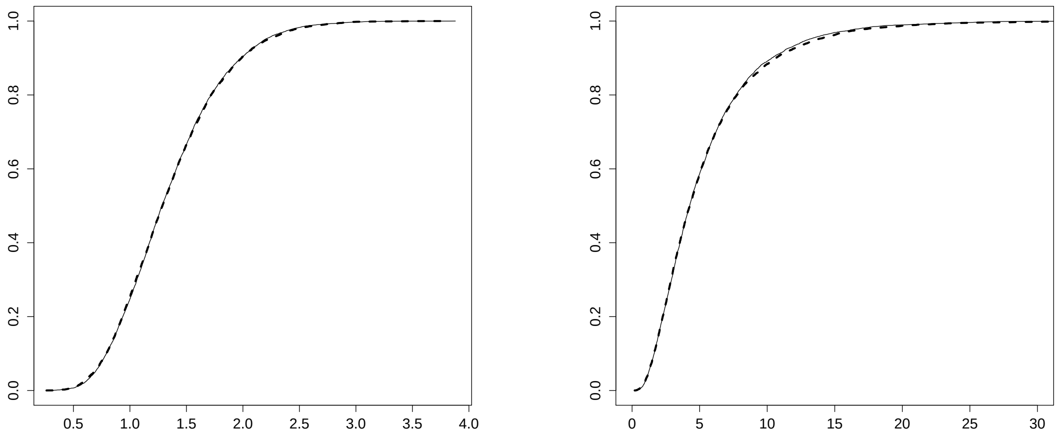
Note that here and below we write $\boldsymbol{\theta}^{(1)}$ instead of $\boldsymbol{\theta}_0^{(1)}$ for the hypothetical parameter as we are writing the full hypothetical marginal probabilities, which have 2 more elements than $\boldsymbol{\theta}_0^{(1)}$. Notice that with the first choice $\boldsymbol{\theta}^{(1)}$, we have more than 20 cells with expected value less than 5, and the biggest cell count has expected value 75.

As we can see, the two curves in each plot in Figure 5.1 are not distinguishable. In fact, any chosen parameter will give us the same result. Moreover, when we gradually reduce the sample size, the curve in each plot remains the same until n is as low as 100.

Figure 5.2 provides the same result as Figure 5.1 but for tables of dimension 5×3. Given that the result is not dependent on the choice of the parameters, we do not write them down here[1].

Secondly, we choose the sample size to be a relatively small value compares to $n = 500$. Obviously, with $n = 40$ in tables of dimension 5×8,

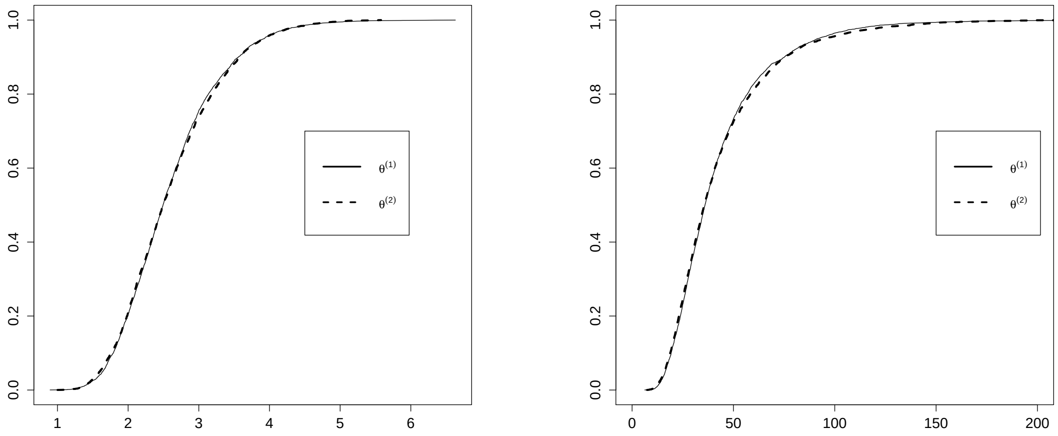
[1] Since the parameters are not given, there will be no legends in the plot.



(a) *Distributions of the KS statistics, sample size $n = 500$, with two different sets of parameters*

(b) *Distributions of the Ω^2 statistics, sample size $n = 500$, with two different sets of parameters*

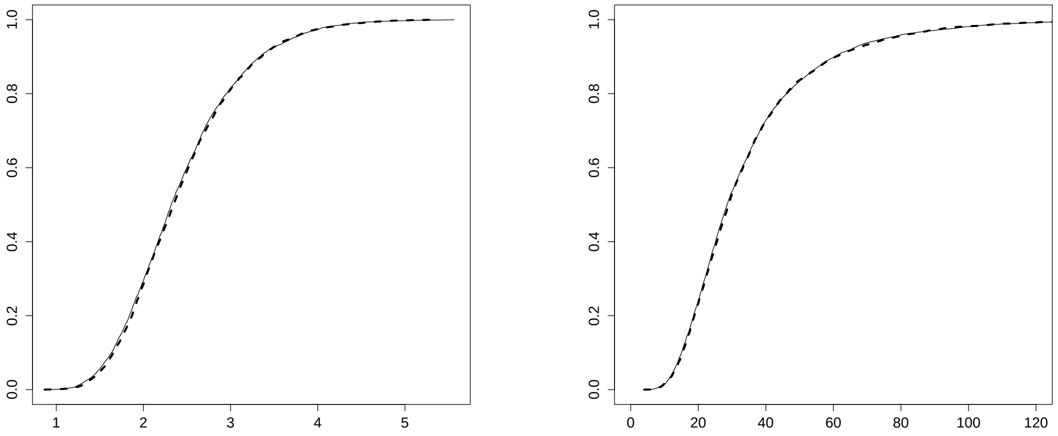
Figure 5.2: Distributions of new test statistics based on $\widehat{Z}_n$ in the limit for 5×3 tables



(a) *Distributions of the KS statistics, sample size $n = 40$, with two different sets of parameters*

(b) *Distributions of the Ω^2 statistics, sample size $n = 40$, with two different sets of parameters*

Figure 5.3: Distributions of new test statistics based on $\widehat{Z}_n$ with a small sample size for 5×8 tables



(a) *Distributions of the KS statistics, sample size $n = 60$, with two different sets of parameters*

(b) *Distributions of the Ω^2 statistics, sample size $n = 60$, with two different sets of parameters*

Figure 5.4: Distributions of new test statistics based on $\widehat{Z}_n$ with a small sample size for 6×6 tables

the distributions of test statistics have not reached their limits yet. But as can be seen in Figure 5.3, the distribution free property still holds since the two curves in each plot coincides. To create these figures, we chose the following hypothetical parameters as an example[2]:

$$\boldsymbol{\theta}^{(1)} = (0.3, 0.1, 0.21, 0.15, 0.24, 0.18, 0.05, 0.1, 0.07, 0.23, 0.1, 0.2, 0.07)^T$$

and

$$\boldsymbol{\theta}^{(2)} = (0.06, 0.35, 0.12, 0.14, 0.33, 0.1, 0.12, 0.16, 0.21, 0.08, 0.15, 0.1, 0.08)^T.$$

One more example to support the argument is given in Figure 5.4, this time for tables of dimension 6×6 with a sample size $n = 60$.

[2] It is quite redundant to write down the chosen parameters; but we hope that with these omitted, the reader can see that with whatever parameters have been chosen, the result stays the same. For this reason, we chose two sets of values which are really "far" from each other in the sense that $\boldsymbol{\theta}_i^{(1)} - \boldsymbol{\theta}_i^{(2)}$ is not always small but quite significantly big at some i.

Finally, in order to see how close the distribution of the test statistics is to its limit in distribution, we plot the cumulative distribution functions of the test statistics with two different sample sizes. This time we choose tables of size 4×6 for demonstration, with parameters

$$\boldsymbol{\theta}^{(1)} = (0.5, 0.1, 0.33, 0.07, 0.05, 0.2, 0.1, 0.15, 0.42, 0.08)^T$$

and

$$\boldsymbol{\theta}^{(2)} = (0.1, 0.2, 0.4, 0.3, 0.3, 0.1, 0.21, 0.09, 0.1, 0.2)^T.$$

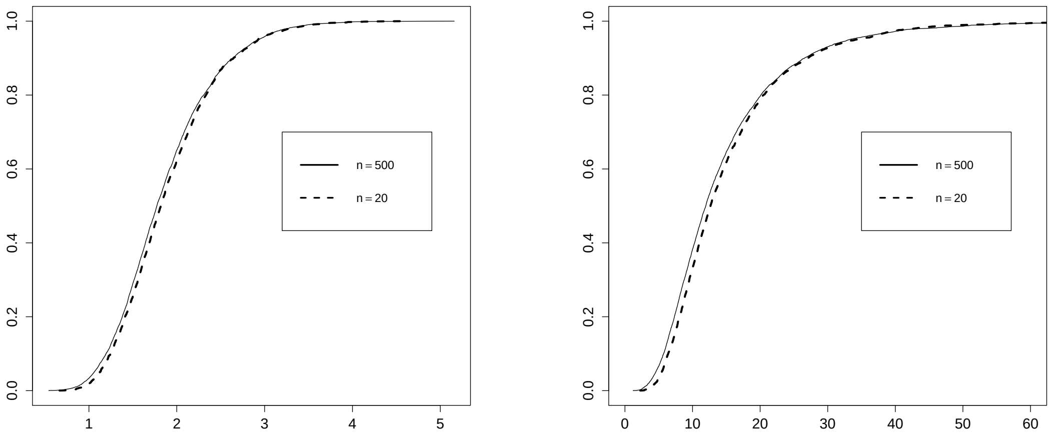
The sample sizes are chosen to be a large number $n = 500$ and two relatively small numbers $n = 60$ and $n = 20$. In fact, the curve with $n = 60$ roughly coincides with the curve with $n = 500$, so we remove it from Figure 5.5. Meanwhile that for $n = 20$ (the lower line) is not far away from them. That somewhat demonstrates how fast the distribution of the test statistics converge to their limits. Further support for this argument is provided in Figure 5.6.

Regarding the time of computational work, it usually takes 1-2 minutes to generate the cumulative distribution function of KS or Ω^2 statistic with 5000 iterations of sample size 500. Therefore, we need at most approximately 0.02 seconds to calculate the test statistics for each particular sample. This convinces us that the transformation works numerically quickly and reliably.

A further remark is that the distribution of GOF tests for tables of dimension $(I + 1) \times (J + 1)$ is the same as that for $(J + 1) \times (I + 1)$ tables. This fact is easy to see as the considered KS and Ω^2 tests are invariant with the transposition of indices i and j.

5.6 Comparison of statistical powers

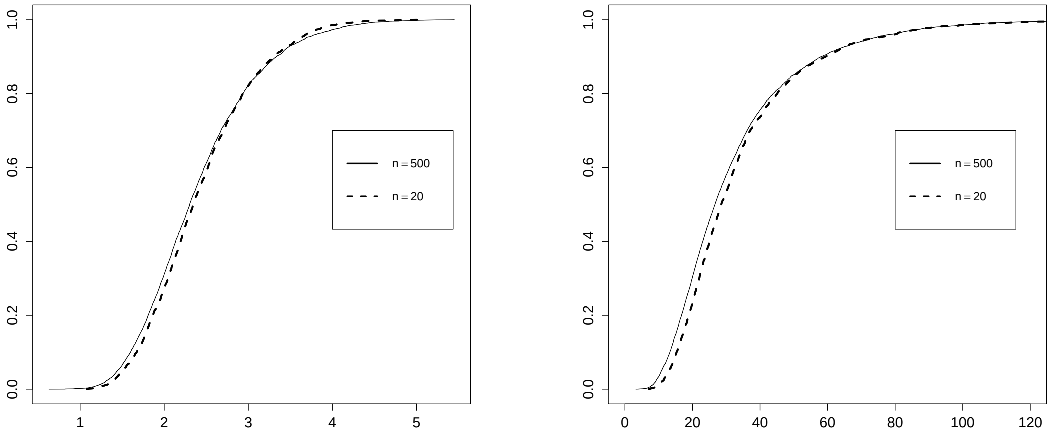
We will also take the KS and Ω^2 tests of the forms in (5.24) and (5.25) as examples for comparing the statistical powers of the new tests based on $\widehat{Z}_n$



(a) *Distributions of the KS statistics in 4×6 contingency tables: solid line ($n = 500$), dashed line (n=20)*

(b) *Distributions of the omega-square statistics in 4×6 contingency tables: solid line ($n = 500$), dashed line (n=20)*

Figure 5.5: Distributions of the new test statistics in limit and those with small sample size of 20 for 4×6 tables



(a) *Distributions of the KS statistics in 7×5 contingency tables: solid line ($n = 500$), dashed line (n=20)*

(b) *Distributions of the omega-square statistics in 7×5 contingency tables: solid line ($n = 500$), dashed line (n=20)*

Figure 5.6: Distributions of the new test statistics in limit and those with small sample size of 20 for 7×5 tables

with the conventional chi-square test. Assume that under an alternative distribution p^a the random variables X and Y are not independent, i.e., we have $p^a_{ij} \neq a_i b_j$ for some i, j. Denote by F_0 and F_a the distribution functions of test statistics under the null and the alternative respectively. The quantities

$$D = \max_{x: F_0(x) \geq 0.85} |F_0(x) - F_a(x)|$$

will be used as numerical descriptions of statistical powers of tests. We aim to compare $D(\chi^2_n)$ to $D(KS)$ and $D(\Omega^2)$.

Alternative distributions will be classified in two groups: one is the popular dependent models, the generalized RC association and correlation models; the other we created from copulas theory.

5.6.1 The alternatives are RC association and RC correlation models

Following **Goodman** [29], consider two well-known models of dependence in contingency tables, which are the generalized RC association model H_{a1} and the RC correlation model H_{a2}. These models are

$$H_{a1} : p^a_{ij} = \alpha_i \beta_j e^{\phi \mu_i \nu_j}$$

where

$$\sum_{i=1}^{I+1} \mu_i \alpha_i = \sum_{j=1}^{J+1} \nu_j \beta_j = 0, \qquad \sum_{i=1}^{I+1} \mu_i^2 \alpha_i = \sum_{j=1}^{J+1} \nu_j^2 \beta_j = 1,$$

and

$$H_{a2} : p^a_{ij} = a_i b_j (1 + \lambda \xi_i \eta_j),$$

where

$$\sum_{i=1}^{I+1} \xi_i a_i = \sum_{j=1}^{J+1} \eta_j b_j = 0, \qquad \sum_{i=1}^{I+1} \xi_i^2 a_i = \sum_{j=1}^{J+1} \eta_j^2 b_j = 1.$$

The sets $\{\mu_i\}$, $\{\xi_i\}$ and $\{\nu_j\}$, $\{\eta_j\}$ denote additional parameters pertaining to the i-th rows of X and j-th columns of Y. They are called **row and column scores**, respectively, in **Goodman** [29]. The quantities λ and ϕ are respectively called a **measure of the correlation** and a **measure of association** in the tables. Some further interpretations of λ and ϕ can be found in **Gilula et al.** [27]. It is obvious that if $\lambda = 0$ or $\phi = 0$ then X and Y are independent. The closer λ or ϕ is to 1, the bigger the deviation of the alternative is from the null distribution.

As we can see, the alternative given in model H_{a2} is a contiguous alternative (see Definition 4.1, the function $h(\cdot)$ now indicating that the deviation of the alternative from the null distribution should be a function of two variables i and j). On the other hand, model H_{a1} is mathematically very similar to model H_{a2}, as can be easily seen by taking a Taylor's expansion. However, model H_{a1} is the commonly used parametrization of the log-linear model (see, for example **Agresti** [1], Chapter 9).

In our simulations, we chose the sets $\{\mu_i\}$ (or $\{\xi_i\}$) and $\{\nu_j\}$ (or $\{\eta_j\}$) to be monotonic, i.e., the values of each are increasing or decreasing in the order of the indices. The reason is that we want to make this choice to be meaningful since the conditional distributions $\{p_{i|j}\}$ and $\{p_{j|i}\}$ will become stochastically ordered as discussed by **Goodman** [28]. We simply chose $\lambda = 0.2$ and $\phi = 0.2$ as representatives. For each model, simulations were run for at least 100 different sets of scores, then we compared $D(\chi^2_n)$ to $D(KS)$ and $D(\Omega^2)$. The result obtained was that $D(KS) > D(\chi^2_n)$ and $D(\Omega^2) > D(\chi^2_n)$ in one third of the cases. In the other third, this is reversed and in the remaining, the statistical powers are approximately the same. That means, the tests based on $\widehat{Z}_n$ have neither uniformly greater nor uniformly smaller statistical power than the conventional chi-square test.

5.6.2 Alternatives from copula theory

In the following, we make some more comparisons of the powers of test statistics for several other alternatives created from copulas theory. A full introduction and deep study regarding copula theory can be found in **Nelsen** [57] among many others.

Assume that there exists a copula $C(u,v)$ which allocates the dependence of the two random variables X and Y. Precisely, the product copula

$$\Pi(u,v) = uv$$

yields the independence of X and Y. Hence, assume that

$$C(u,v) \neq \Pi(u,v).$$

The marginal distributions are still supposed to be known as $\{a_i\}$ and $\{b_j\}$. Then with the convention that $a_0 = b_0 = 0$, the joint distribution p^a_{ij} can be defined via copula C as follows:

$$p^a_{ij} = C(\sum_{t=0}^{i} a_t, \sum_{s=0}^{j} b_s) - C(\sum_{t=0}^{i-1} a_t, \sum_{s=0}^{j} b_s) - C(\sum_{t=0}^{i} a_t, \sum_{s=0}^{j-1} b_s) + C(\sum_{t=0}^{i-1} a_t, \sum_{s=0}^{j-1} b_s),$$
$$i = 1, \dots, I+1,\ j = 1, \dots, J+1. \qquad (5.26)$$

This formula is naturally deduced from the definition of a copula and its connection with the joint distribution of the two random variables.

We simply consider three families of copulas, which appear to be good local alternatives to independence (more can also be found in **Nelsen** [57].) Those families are Cuadras-Augé $C^{(1)}_\theta$, Gumbel's bivariate exponential distribution $C^{(2)}_\theta$ and Ali-Mikhail-Haq $C^{(3)}_\theta$, which are given by

$$C^{(1)}_\theta(u,v) = [\min(u,v)]^\theta [uv]^{1-\theta}, \qquad (5.27)$$

$$C^{(2)}_\theta(u,v) = u + v - 1 + (1-u)(1-v)e^{-\theta \ln(1-u)\ln(1-v)}, \qquad (5.28)$$

and

$$C_\theta^{(3)}(u,v) = \frac{uv}{1-\theta(1-u)(1-v)} \tag{5.29}$$

with $0 \leq \theta \leq 1$. The parameter θ involved in these copulas indicates how strong the dependence of X and Y is.

Denote by $c(u,v)$ the density of copula C, i.e.,

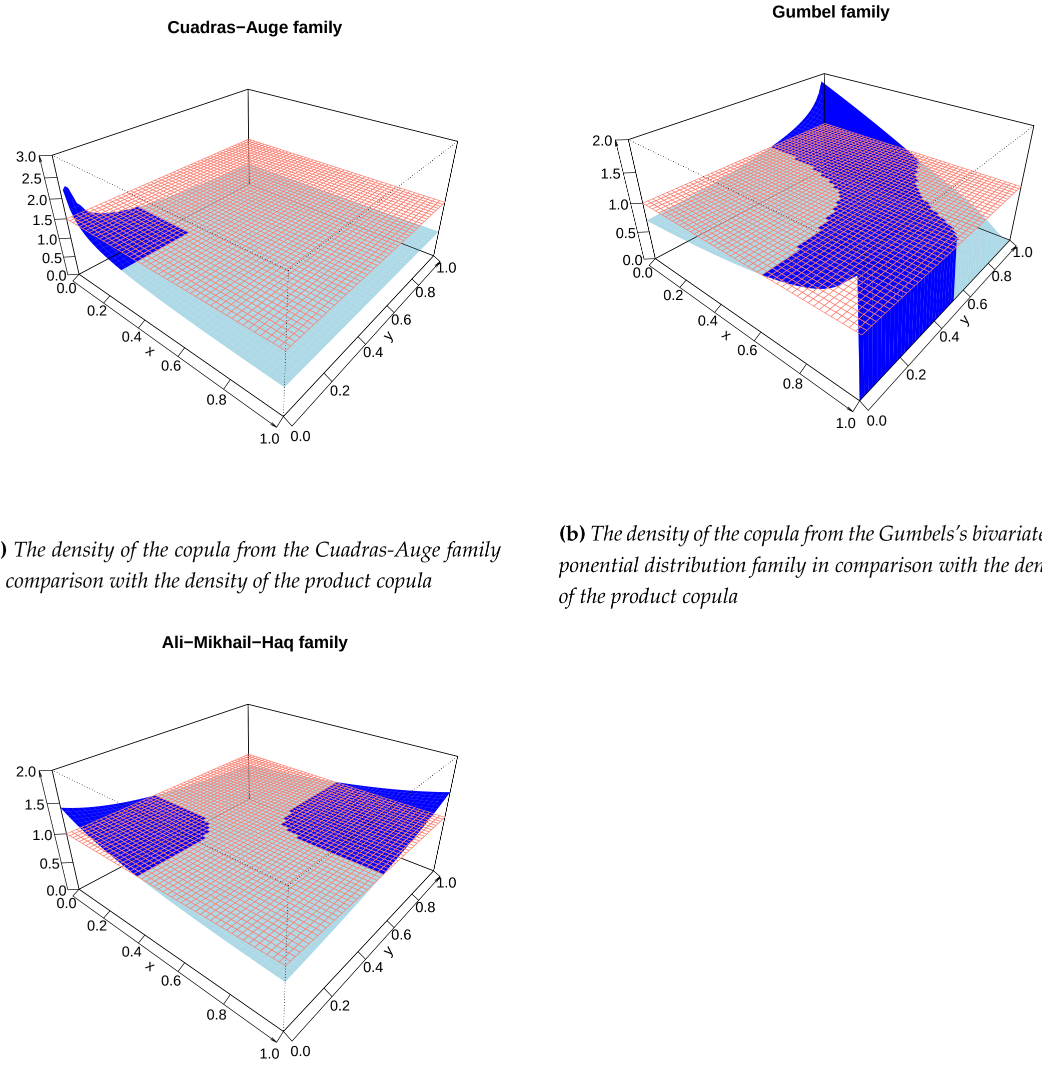
$$c(u,v) = \frac{\partial^2}{\partial u \partial v} C(u,v).$$

From (5.26), it can be seen that the density of copula $c(u,v)$ reflects the pattern of the alternative. We sketch these densities along with the densities of the product copula, which is

$$\pi(u,v) = \frac{\partial^2}{\partial u \partial v} \Pi(u,v) = 1.$$

Figure 5.7 shows the patterns of the three considered alternatives. In these plots, the density of the product copula, which is every where equal to 1, is marked in a *salmon* colour. The graphs of the other density functions are coloured in two different shades. The parts where the value of the density function is greater than 1 are made *dark blue* in the figure and the remainder is *light blue*. Note that Figure 5.7a depicts only the density function of the absolutely continuous part and the weight put on the diagonal line (which is bigger than 1) can not be displayed visibly. These figures also demonstrate how close the alternative distribution is to the null distribution. As we can see, the shift of the alternative from the null distribution is not too high at any point.

At the same time, for each family of copulas, we usually chose $\theta \in [0.3, 0.5]$ to define the joint distribution p_{ij}^a of the alternative. We then compare $D(\chi_n^2)$ with $D(KS)$ and $D(\Omega^2)$. Since the marginal distributions $\{a_i\}$ and $\{b_j\}$ can vary widely, we consider one representative which is the discrete uniform. What we observed from multiple simulations is that, with the alternative from the Cuadras-Augé family, the chi-square test performs



(a) *The density of the copula from the Cuadras-Auge family in comparison with the density of the product copula*

(b) *The density of the copula from the Gumbels's bivariate exponential distribution family in comparison with the density of the product copula*

(c) *The density of the copula from the Ali-Mikhail-Haq family in comparison with the density of the product copula*

Figure 5.7: The densities of copulas yield the patterns of the alternatives

better than the other tests based on $\widehat{Z}_n$. For example, with tables of dimension 4×4, $\theta = 0.2$ and sample size $n = 200$, we have

$$D(\chi_n^2) = 0.63,\ \ D(KS) = 0.49 \text{ and } D(\Omega^2) = 0.5.$$

With the other two, the tests based on $\widehat{Z}_n$ are more powerful. For example, with the alternative from the Ali-Mikhail-Haq family where $\theta = 0.5$, the sample size $n = 200$, table of dimension 8×7, we have

$$D(\chi_n^2) = 0.23,\ \ D(KS) = 0.38 \text{ and } D(\Omega^2) = 0.43.$$

5.7 Testing independence of two discrete random vectors

Consider now the problem of testing independence between X and Y where either or both are multidimensional random variables. Without loss of generality, assume that X is a 2-dimensional random variable of dimension $(I+1) \times (J+1)$. The random variable Y has $K+1$ values. The contingency table now should be of size $(I+1) \times (J+1) \times (K+1)$.

One may imagine that the scenario now is similar to having a three-way contingency table showing the association of 3 discrete random variables X_1, X_2 and Y. Note that in three-way contingency tables, we have several different sorts of testing independence. The so-called testing complete independence is formulated by

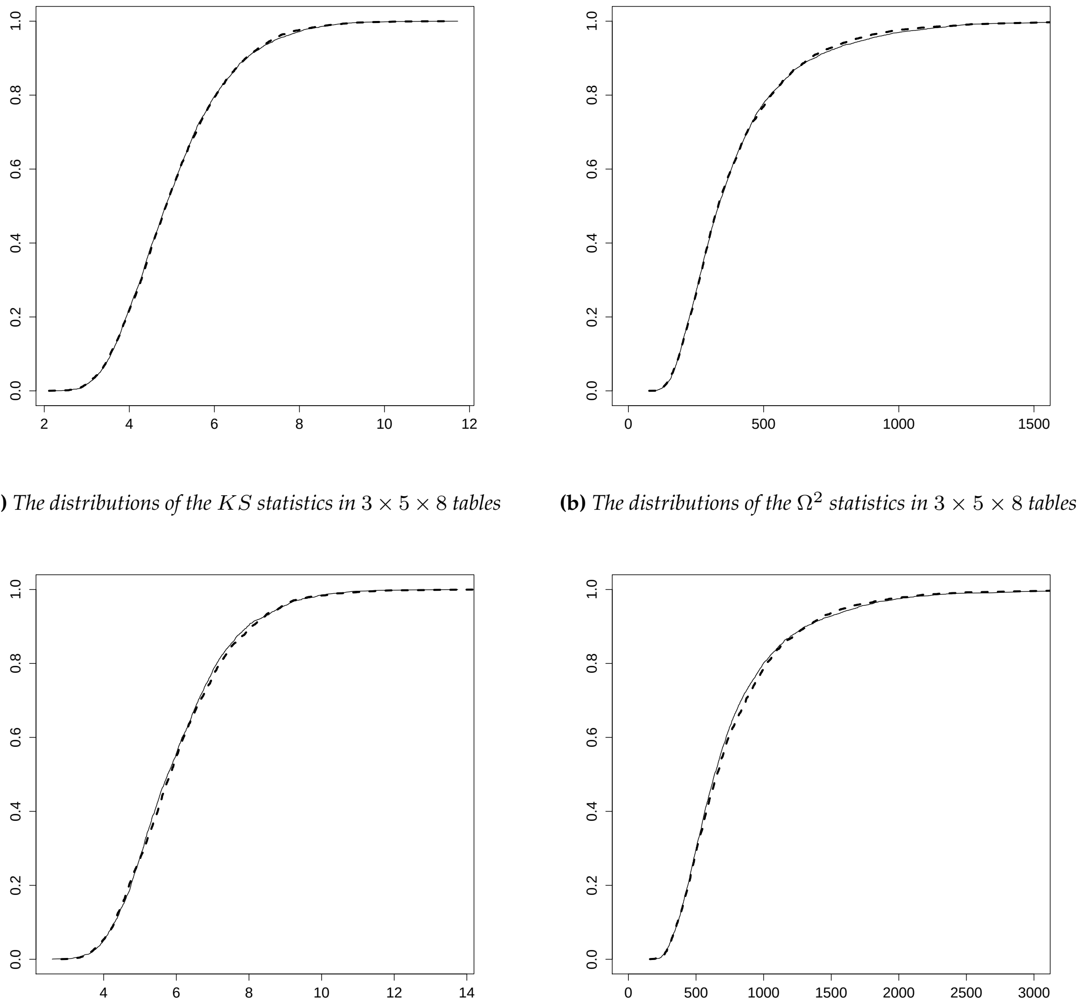
$$\mathbb{P}(X_1 = i, X_2 = j, Y = k) = \mathbb{P}(X_1 = i)\mathbb{P}(X_2 = j)\mathbb{P}(Y = k) \text{ for all } i, j, k.$$

Note that the problem of testing independence between a 2-dimensional random variable X and a random variable Y, however, is not of this scheme. The problem in fact is equivalent to the so-called testing joint independence, which is formulated by

$$\mathbb{P}(X_1 = i, X_2 = j, Y = k) = \mathbb{P}(X_1 = i, X_2 = j)\mathbb{P}(Y = k) \text{ for all } i, j, k.$$

Therefore, we can always treat this testing problem as testing independence in a two-way contingency table of size $((I+1)\times(J+1))\times(K+1)$ by a simple rearrangement. Then clearly the method could work with no modification.

Again, we use the KS and Ω^2 statistics and simulate scenarios with different dimensions of the table. Following the process in Section 5.5.2, we illustrate the distribution free property of the new test statistics, now for testing independence of two discrete random vectors. Even though the dimension of the table is increasing significantly, Figure 5.8 shows no noticeable difference between the two curves in each plot.



(a) *The distributions of the KS statistics in* $3 \times 5 \times 8$ *tables*

(b) *The distributions of the* Ω^2 *statistics in* $3 \times 5 \times 8$ *tables*

(c) *Distributions of the KS statistics in* $4 \times 6 \times 7$ *tables*

(d) *Distributions of the* Ω^2 *statistics in* $4 \times 6 \times 7$ *tables*

Figure 5.8: Distribution free property of the new test statistics for testing independence of two discrete random vectors

Chapter 6

Testing regularly varying tail distributions

In this chapter, we will give a construction of a class of asymptotically distribution free GOF tests for testing regularly varying tail distributions. The main content of this chapter lies in Section 6.3 which contains the method and Section 6.4 which presents the simulation results.

The study of testing regularly varying tail distributions is motivated by the fact that applications of these distributions can be found in various practical areas as well as in some theories of probabilities and statistics.

The practical areas in which regularly varying distributions arise as a common phenomenon include finance, insurance, physics, geology, hydrology and engineering. For instance, applications in *finance* can be seen in **Embrechts et al.** [22], **Jansen and de Vries** [36], **McCulloch** [55]. Examples in *physics* can be found in **Kotulski** [48], **Metzler and Klafter** [56], etc. For applications in *hydrology*, **Anderson and Meerschaert** [2], **Lu and Molz** [52], among others, give us some examples. In *engineering*, we refer to **Resnick** [66], **Nikias and Shao** [59], and various others.

Regular variation of the tail of a distribution also often appears as a natural condition in various probabilistic theories. The most typical example is that it is the condition for the distribution of the partial maxima to belong to the domain of attraction of extreme value distributions. This could be found in various references, the books of **de Haan and Ferreira** [15] and **Resnick** [65] among others. Another important role of regularly varying distributions is that they are involved in the characterization of the domain of attraction of an $\alpha-$stable distribution for some $\alpha \in (0, 2)$: see for example **Embrechts et al.** [22], **Geluk and de Haan**[25].

6.1 Regularly varying tail distributions and applications

We will present in this section the definitions of regular variation as well as regularly varying tail distributions.

6.1.1 Regular variation

The term **regular variation** was introduced in order to describe the deviation from pure power laws, which behaviour has been observed quite frequently in many fields of applied mathematics.

The material in this section can be found in various references, for example, **Resnick** [65].

Definition 6.1. A measurable function $f : \mathbb{R}^+ \to \mathbb{R}^+$ is called **regularly varying** at ∞ with index θ if

$$\lim_{t\to\infty} \frac{f(tx)}{f(t)} = x^\theta \quad \text{for all } \ x > 0. \tag{6.1}$$

This we denote as $f \in RV_\theta$.

We call θ the **exponent of regular variation**. If $\theta = 0$, i.e., $f \in RV_0$, then f is said to be **slowly varying** at ∞. If $f(x) \in RV_\theta$ then

$$L(x) = \frac{f(x)}{x^\theta} \in RV_0.$$

If $f(x)$ is regularly varying at ∞ then $f(x^{-1})$ is **regularly varying** at 0.

A typical example of a regularly varying function at ∞ is $f(x) = x^{\theta}$. Any polynomial function or its equivalent functions are also regularly varying at ∞. Examples of slowly varying functions at ∞ include $\log(1 + x)$, $\log\log(e + x)$. Also, any function of x with a finite limit as $x \to \infty$ is slowly varying.

6.1.2 Regularly varying tail distributions

The class of regularly varying tail distributions is classified as a sub-class of heavy tail distributions. The term **heavy-tail distributions** does not possess any universal notion or formal definition. However, regularly varying tail distributions do.

Definition 6.2. A non-negative random variable X and its distribution function $F(x)$ are said to be **regularly varying** if $1 - F(x)$, the tail, is regularly varying with index $-\theta$ with $\theta > 0$. That means,

$$\lim_{t\to\infty} \frac{1 - F(tx)}{1 - F(t)} = x^{-\theta}, \quad \text{for all } x > 0.$$

To indicate that the distribution F has regularly varying tail with exponent $-\theta$, we will simply for abbreviation write $1 - F \in RV_{\theta}$ and say that the exponent of regular variation is θ. This should not cause any contradiction or confusion since we are not talking about any regularly varying function but only about the distribution functions whose tails are regularly varying. This abbreviation will be used consistently throughout the remaining parts of the thesis.

The class of heavy tail distributions includes, for example, the log-normal distribution, log-gamma distribution, Weibull distribution, Burr, Pareto distributions, etc. Among the class of heavy tail distributions, the following distributions have been known to be regularly varying tail distributions.

- **Pareto distribution**

There is a hierarchy of the Pareto distributions, with types ranging from I to V. Type I is defined as

$$F(x) = 1 - \left(\frac{x}{x_0}\right)^{-\theta}, \qquad x \geq x_0 > 0.$$

Obviously, the Pareto distribution is regularly varying with exponent θ. Note that **Pareto-type distributions** includes distributions whose right tail is of the form

$$1 - F(x) \sim Kx^{-\theta}, \quad x \to \infty.$$

Hence, by their definitions, regularly varying tail distributions are often referred as Pareto-type distributions.

- **Cauchy distribution**

The Cauchy distribution function is

$$F(x) = \frac{1}{\pi}\arctan\left(\frac{x - x_0}{\gamma}\right) + \frac{1}{2}, \qquad \gamma > 0.$$

The Cauchy distribution is known to be regularly varying with exponent $\theta = 1$. This can be checked easily.

- **Burr distribution**

The distribution function of the Burr distribution is

$$F(x) = 1 - \left(\frac{x_0}{x_0 + x^{\tau}}\right)^{\gamma}, \qquad x_0, \gamma, \tau > 0.$$

This distribution is one of the generalizations of the Pareto distribution. When $\tau = 1$ we obtain the Pareto distribution.

- **Log-gamma distribution**

The density function of the log-gamma distribution is

$$f(x) = \frac{\alpha^{\beta}}{\Gamma(\beta)}(\log x)^{\beta-1}x^{-\alpha-1}, \qquad \alpha, \beta > 0.$$

It is easy to see that the tail of the log-gamma distribution is

$$1 - F(x) = \int_x^\infty \frac{\alpha^\beta}{\Gamma(\beta)}(\log y)^{\beta-1} y^{-\alpha-1} dy \sim \frac{\alpha^{\beta-1}}{\Gamma(\beta)}(\log x)^{\beta-1} x^{-\alpha}.$$

Hence, the log-gamma distribution is regularly varying with exponent α.

- **Stable distributions with exponent** $\alpha < 2$

Consider a sequence of independent identically distributed random variables $X_1, \cdots, X_n$ having the same distribution F as a random variable X. Consider the random walk

$$S_0 = 0,\ S_n = X_1 + \cdots + X_n,\ \ n \geq 1. \tag{6.2}$$

A random variable X is said to be **stable** or its distribution F is **stable** if for each $n \geq 1$, there exist constants $a_n > 0$ and b_n such that,

$$S_n \overset{d}{=} a_n X + b_n. \tag{6.3}$$

In other words,

$$X \overset{d}{=} \frac{S_n - b_n}{a_n}.$$

The normal distribution and the Cauchy distribution are the only known examples of stable distributions with explicit functional form for density functions.

It is known that the only possible form of the constant a_n is $a_n = n^{1/\alpha}$ with $\alpha \in (0, 2]$ (see for example **Feller**[23], VI.1), and two well-known values of α are $\alpha = 2$ and $\alpha = 1$ for normal and Cauchy distributions respectively. Stable distributions with $\alpha < 2$ are known to be regularly varying.

6.2 Research directions on regularly varying distributions

Regarding research on regularly varying tail distributions, estimating the exponent θ and using GOF tests for testing hypotheses are two main directions of interest. A brief literature review will be given in section 6.2.1 for the former and in section 6.2.2 for the latter.

6.2.1 Estimation of the exponent θ

The problem of estimating the exponent θ is the most popular research direction regarding regularly varying tail distributions. In fact, so far most studies have been focusing on finding the exponent of the regular variation in the tail and it has a very rich literature. Among various studies, the Hill estimator is the most common, see **Hill** [34].

Suppose that the random sample $X_1, \ldots, X_n$ is re-arranged in an increasing order

$$X_{1,n} \leq \cdots \leq X_{n,n},$$

which are called the **order statistics** of the sample. Then, the **Hill estimator** of the exponent θ has the following form

$$\hat{\theta}^H = \Big(\frac{1}{m}\sum_{i=1}^{m} \ln X_{n-m+i,n} - \ln X_{n-m,n}\Big)^{-1} \tag{6.4}$$

where $1 \leq m \leq n$ and m/n is called the sample fraction. This estimator $\hat{\theta}^H$ is in fact the value which maximizes the likelihood of the conditional distribution (see (6.10)).

Denote by x_0 the smallest value of the sub-sample $X_{n-m,n}, \ldots, X_{n,n}$, i.e., $x_0 = X_{n-m,n}$. Most of the time, the size m of the sub-sample, as a function of x_0 and n, is considered to satisfy the following conditions:

$$m \to \infty \text{ and } \frac{m}{n} \to 0 \quad \text{as } n \to \infty, x_0 \to \infty. \tag{6.5}$$

Under these conditions, the Hill estimator was proved to be consistent in the sense that

$$\hat{\theta}^H \xrightarrow{\mathbb{P}} \theta_0, \tag{6.6}$$

where θ_0 denotes the true unknown exponent under the assumption that the tail distribution $1 - F$ of the original sample $X_1, \ldots, X_n$ is regularly varying. The proof of this convergence can be found in **Mason** [54]. The author also proved that if m is the integer part of n^{α}, with $0 < \alpha < 1$, then $\hat{\theta}^H \to \theta_0$ almost surely as $n \to \infty$.

There have been also many studies regarding asymptotic normality of the Hill estimator, for example, **Davis and Resnick** [11], **Haeusler and Teugels** [32] , **Geluk et al.** [24], **de Haan and Resnick** [14], **Resnick and Stărică** [67], etc. Those studies indicated that together with condition (6.5) and assumption that the tail is regularly varying, the following convergence holds

$$\sqrt{m}\Big(\frac{1}{\hat{\theta}^H} - \frac{1}{\theta_0}\Big) \sim \mathcal{N}(0, \frac{1}{\theta_0^2}).$$

Besides the well-known Hill estimator, many other modifications have been proposed. Most of them are also based on the upper order statistics and are not too difficult to compute. These include the estimators introduced by **de Haan and Resnick** [13], **Hall** [33], **Pickands** [63], **Teugels** [75] among others.

Apart from estimating θ, several methods for choosing the sample fraction m/n based on survey data can be found in **Drees and Kaufmann** [18], **Danielsson et al.** [10] and **Guillou and Hall** [30].

6.2.2 GOF tests for testing regularly varying tail distributions

In contrast to the numerous approaches for estimating the exponent θ, the use of GOF tests for testing regular variation has been addressed in only a few studies.

Not long ago, in 2006, **Beirlant et al.** [5] modified the Jackson statistic - which was originally proposed as a GOF test for testing exponentiality - for testing Pareto-type data. **Koning and Peng** [47] examined the Kolmogorov-Smirnov, Berk-Jones and the score tests and their quadratic variants and compared them in terms of Bahadur efficiency. In that paper, the score test and its integrated version were shown to be the best tests.

Consider the family of the generalized Pareto distribution (GPD), which is of the form

$$F(x;k,\sigma) = \begin{cases} 1-(1-\frac{kx}{\sigma})^{1/k}, & k \neq 0, \sigma > 0 \\ 1-e^{-x/\sigma}, & k = 0, \sigma > 0 \end{cases} \tag{6.7}$$

where k and σ are shape and scale parameters and $x > 0$. Members of this family with $k < 0$ belong to class of distributions whose tails are regularly varying. To test the fit of data to a GDP, there have been several studies including **Marohn** [53], **Choulakian and Stephens** [8]. In these papers, the critical values for Cramér-von Mises statistic and Anderson-Darling statistic were given.

Recently, **Can et al.** [6] studied GOF tests for testing whether multi-dimensional distributions belong to the domain of attraction of a multi-dimensional extreme value distribution. They did not focus on any particular GOF test but on the construction of a class of asymptotically distribution free GOF tests for the hypothesis testing problem, which is the same aim as ours. In that paper, the authors used the Khmaladze innovative martingale method, or in other words, the Khmaladze transformation [40].

6.3 A construction of a class of GOF tests

We present in this section a construction of a class of GOF tests for testing regularly varying tail distributions. Again, we use the Khmaladze-2 transformation, see **Khmaladze** [44], for a parametric family of distributions.

Since the process is carried out on the tail of distributions F, we consider only the observations beyond a threshold x_0. For convenience, we use a simple change of variable and therefore change the working space into $\mathcal{L}_2(H)$ where H is the conditional distribution on the tail. In Section 6.3.1, we recall the form of the function-parametric tail empirical process $\widehat{v}_{mH}(\phi)$ (see (6.12) and (6.14)) and its limit in distribution. The method of transforming the process $\widehat{v}_{mH}$ into another process $\widehat{v}_{mG}$, which lies in another chosen space and has a specified limit in distribution, will be conducted in Section 6.3.3.

6.3.1 The tail empirical process

In this section, we will rewrite the tail empirical process and its limit in distribution in terms of tail random variables $T_i, i = 1, \ldots, m$.

From the set of observations $X_1, \ldots, X_n$, we consider only observed values on the tail, which exceed a threshold x_0. Let us denote by $\widetilde{X}_1, \cdots, \widetilde{X}_m$ the sub-sample to be considered with m satisfying (6.5). Under the hypothesis that F has regularly varying tail, the distribution of $\widetilde{X}_1, \cdots, \widetilde{X}_m$ is the conditional distribution

$$\frac{\mathbb{P}\left\{\widetilde{X} \geq x_0 t\right\}}{\mathbb{P}\left\{\widetilde{X} \geq x_0\right\}} = \frac{1 - F(x_0 t)}{1 - F(x_0)} = t^{-\theta} + o(1), \quad t \geq 1, \, \theta \geq 0. \tag{6.8}$$

Let $\widetilde{T} = \frac{\widetilde{X}}{x_0}$ and $T = \widetilde{T} - 1$. Then, the asymptotic distribution of the positive continuous random variable T under the null hypothesis is

$$H_\theta(t) = 1 - (1+t)^{-\theta}, \qquad t \geq 0. \tag{6.9}$$

Clearly, the density function is

$$h(t) = \theta(1+t)^{-(\theta+1)}.$$

Let $\hat{\theta}_m$ be the MLE, or in other words, the Hill's estimator of the true unknown exponent θ, then we can rewrite $\hat{\theta}_m$ in terms of the observed values $T_1, \ldots, T_m$ as follows:

$$\hat{\theta}_m = \frac{m}{\sum_{i=1}^m \log(T_i + 1)}. \tag{6.10}$$

The tail empirical distribution function can also be rewritten in terms of $T_1, \ldots, T_m$ as

$$H_m(t) = \frac{1}{m}\sum_{i=1}^{m} I_{\{T_i \le t\}}. \tag{6.11}$$

Then, recall from the equation (3.32) in Section 3.3 that the tail parametric empirical process $\widehat{v}_{mH}$ is

$$\widehat{v}_{mH}(t) = \sqrt{m}[H_m(t) - H_{\hat{\theta}_m}(t)] \tag{6.12}$$

and the tail empirical process is

$$v_{mH}(t) = \sqrt{m}[H_m(t) - H_{\theta_0}(t)]. \tag{6.13}$$

Similarly to what was written in Chapter 3, consider the function-parametric version of the tail empirical process

$$v_{mH}(\phi) = \int_0^\infty \phi(t) v_{mH}(dt) = \frac{1}{\sqrt{m}}\sum_{i=1}^{m}[\phi(T_i) - \mathsf{E}\phi(T_i)], \tag{6.14}$$

where ϕ is a function in $\mathcal{L}_2(H)$. With the change of variable, everything is now considered in another space, $\mathcal{L}_2(H)$.

Let us recall from (3.36) the limit in distribution of the process $\widehat{v}_{mH}(\phi)$, which is

$$\begin{aligned}\widehat{V}_H(\phi) &= V_H(\phi) - \langle \beta_H, \phi\rangle_H V_H(\beta_H)\\ &= W_H(\phi) - \langle \mathbb{1}, \phi\rangle_H W_H(\mathbb{1}) - \langle \beta_H, \phi\rangle_H W_H(\beta_H),\end{aligned} \tag{6.15}$$

where β_H is the normalized score function,

$$\beta_H(t) = \Gamma_H^{-1/2}\frac{\dot{h}_{\theta_0}(t)}{h_{\theta_0}(t)}. \tag{6.16}$$

Recall that $\Gamma_H^{-1/2}$ is the Fisher information, defined in (2.2). Recall also that $V_H(\phi)$, a function-parametric H-Brownian bridge, is the limit in distribution of $v_{mH}(\phi)$. Furthermore $\widehat{V}_H(\phi)$, as a projection of the function parametric H-Brownian motion $W_H(\phi)$ orthogonal to the subspace generated by two functions $\mathbb{1}$ and β_H, is called the β_H-projected H-Brownian motion (see Section 3.4).

6.3.2 The target distribution

One may choose any distribution G as a destination distribution, but we chose G to be the standard exponential distribution, that is,

$$G(t) = 1 - e^{-t}, \quad t \geq 0.$$

Our reason for this choice is that G is a familiar distribution and both G and H are members of the GPD family (see (6.7)). The distribution G is the limiting distribution of the GPD(k, σ) as $k \to 0$ and is scaled by $\sigma = 1$.

It is obvious that the distributions G and H are mutually absolutely continuous. It is also easy to calculate the Fisher information of G and H,

$$\begin{aligned} \Gamma_H &= \frac{1}{\theta_0^2}, \\ \Gamma_G &= 1. \end{aligned}$$

Put

$$\ell(t) = \sqrt{\frac{dG}{dH}(t)} = \theta^{-\frac{1}{2}}(1+t)^{\frac{\theta+1}{2}} e^{-\frac{t}{2}}.$$

Then this function belongs to $\mathcal{L}_2(H)$. In addition, if $\phi \in \mathcal{L}_2(G)$ then $\ell\phi \in \mathcal{L}_2(H)$ and

$$\|\phi\|_G = \|\ell\phi\|_H \,.$$

Note that when H and G are mutually absolutely continuous, it is known that there is a straightforward transformation from a H-Brownian motion into a G-Brownian motion. Namely, $W_H(\ell\phi) = W_G(\phi)$ is a G-Brownian motion in $\mathcal{L}_2(G)$. However, mapping a Brownian bridge such as V_H or $\widehat{V}_H$ into another Brownian bridge is not so straightforward any more. The fact is that the distribution of $V_H(\ell\phi)$ still depends on both H and G and so does $\widehat{V}_H(\ell\phi)$.

6.3.3 The transformation of the tail empirical process

Consider a subspace $\widehat{\mathcal{L}}$ of $\mathcal{L}_2(H)$ generated by four functions $\{\mathbb{1}, \beta_H, \ell, \ell\beta_G\}$. These functions are of unit norm in $\mathcal{L}_2(H)$. More explicitly, the score functions β_H and β_G are

$$\beta_H(t) = 1 - \theta \log(1+t), \tag{6.17}$$

$$\beta_G(t) = 1 - t. \tag{6.18}$$

For any function f and g in $\mathcal{L}_2(H)$, define the unitary operator $U_{f,g}$ on $\mathcal{L}_2(H)$ as

$$U_{f,g} = I - \frac{1}{1 - \langle f, g \rangle_H}(g - f)\langle g - f, \cdot \rangle_H$$

where I is the identity operator. This operator interchanges f and g, i.e.,

$$\begin{aligned} U_{f,g} f &= g, \\ U_{f,g} g &= f. \end{aligned}$$

Moreover, it keeps any function orthogonal to f and g unchanged, i.e.,

$$U_{f,g} v = v \text{ for all } v \perp f, g.$$

Now, as the first step, we consider the unitary operator

$$U_{\mathbb{1},\ell} = I - \frac{1}{1 - \langle \mathbb{1}, \ell \rangle_H}(\ell - \mathbb{1})\langle \ell - \mathbb{1}, \cdot \rangle_H. \tag{6.19}$$

This operator will map ℓ to $\mathbb{1}$ and $\mathbb{1}$ to ℓ. Next, consider the image of the function $\ell\beta_G$ via $U_{\mathbb{1},\ell}$, which is

$$\begin{aligned} \widetilde{\ell\beta}_G &= \ell\beta_G - \frac{1}{1 - \langle \mathbb{1}, \ell \rangle_H}(\ell - \mathbb{1})\langle \ell - \mathbb{1}, \ell\beta_G \rangle_H \\ &= \ell\beta_G - \frac{1}{1 - \int_0^\infty \ell(s)h(s)ds}(\ell - \mathbb{1}) \int_0^\infty (\ell(s) - 1)\ell(s)\beta_G(s)h(s)ds. \end{aligned}$$

Then consider the operator $U_{\beta_H, \widetilde{\ell\beta}_G}$ defined as

$$U_{\beta_H, \widetilde{\ell\beta}_G} = I - \frac{1}{1 - \langle \beta_H, \widetilde{\ell\beta}_G \rangle_H}(\widetilde{\ell\beta}_G - \beta_H)\langle \widetilde{\ell\beta}_G - \beta_H, \cdot \rangle_H.$$

Now, set

$$\widehat{\mathbb{U}} = U_{\beta_H, \widetilde{\ell\beta}_G} U_{\mathbb{1},\ell}.$$

Then this unitary operator will map ℓ to $\mathbb{1}$ and $\ell\beta_G$ to β_H. In summary,

$$\widehat{\mathbb{U}}\ell = \mathbb{1},$$
$$\widehat{\mathbb{U}}(\ell\beta_G) = \beta_H.$$

The non-uniqueness of a unitary operator like $\widehat{\mathbb{U}}$ was discussed thoroughly in Khmaladze [44], Section 3.4. Nevertheless, we believe that this operator $\widehat{\mathbb{U}}$ is simple enough for practical purposes, especially with only one parameter.

It now follows from the main result for testing parametric hypotheses, Theorem 7 in **Khmaladze** [44], that:

Theorem 6.1. *If $\widehat{V}_H$ is a β_H-projected H-Brownian motion and G is absolutely continuous with respect to H, then*

$$\widehat{V}_G(\phi) = \widehat{V}_H(\widehat{\mathbb{U}}(\ell\phi)) = \widehat{\mathbb{U}}(\widehat{V}_H(\ell\phi)) \tag{6.20}$$

is a β_G-projected G-Brownian motion.

As a consequence, transforming the function-parametric tail empirical process $\widehat{v}_{mH}(\phi)$ by $\widehat{\mathbb{U}}$, we obtain another process

$$\widehat{v}_{mG}(\phi) = \widehat{v}_{mH}(\widehat{\mathbb{U}}(\ell\phi)) = \widehat{\mathbb{U}}(\widehat{v}_{mH}(\ell\phi)), \tag{6.21}$$

which has $\widehat{V}_G(\phi)$ as a limit in distribution.
Let

$$\phi_x(t) = I_{\{t\le x\}},$$

where x runs from 0 to ∞, be a family of indicator functions depending on x defined on $\mathcal{L}_2(H)$.

Recall from (6.15) that the β_H-projected H-Brownian motion $\widehat{V}_H(\ell\phi)$ is the limit in distribution of $\widehat{v}_{mH}(\ell\phi)$; and therefore by Theorem 6.1, the image of $\widehat{v}_{mH}(\ell\phi)$ via $\widehat{\mathbb{U}}$, which is $\widehat{v}_{mH}(\widehat{\mathbb{U}}(\ell\phi_x))$ in (6.21) (or $\widehat{v}_{mH}(\widetilde{\phi}_x)$ below), has limit in distribution $\widehat{V}_H(\widehat{\mathbb{U}}(\ell\phi_x))$, a β_G-projected G-Brownian motion in

ϕ_x. Hence, any statistic as an appropriate functional based on $\widehat{v}_{mH}(\widehat{\mathbb{U}}(\ell\phi_x))$ will be asymptotically distribution free. More precisely, the limit in distributions of the test statistics does not depend either on the distribution H or on the parameters $\theta_0, \hat{\theta}_m$.

Denote by $\widetilde{\phi}_x$ the image of $\ell\phi_x$ via the operator $\widehat{\mathbb{U}}$, that is,

$$\widetilde{\phi}_x = \widehat{\mathbb{U}}(\ell\phi_x) = U_{\beta_H, \widetilde{\ell\beta}_G} U_{\mathbb{1},\ell}(\ell\phi_x).$$

For programming purposes, we need to write the explicit form of $\widetilde{\phi}_x$, that is

$$\widetilde{\phi}_x = \ell\phi_x - \frac{1}{1 - \langle \mathbb{1}, \ell\rangle_H}(\ell - 1)\langle \ell - 1, \ell\phi_x\rangle_H - \frac{1}{1 - \langle \beta_H, \widetilde{\ell\beta_G}\rangle_H}(\widetilde{\ell\beta_G} - \beta_H) \tag{6.22}$$

$$\times \Big[\langle \widetilde{\ell\beta_G} - \beta_H, \ell\phi_x\rangle_H - \frac{\langle \ell - \mathbb{1}, \ell\phi_x\rangle_H}{1 - \langle \mathbb{1}, \ell\rangle_H}\langle \widetilde{\ell\beta_G} - \beta_H, \ell - \mathbb{1}\rangle_H\Big].$$

Applying the unitary operator $\widehat{\mathbb{U}}$ on the process $\widehat{v}_{mH}(\ell\phi_x)$ we have

$$\begin{aligned}\widehat{v}_{mG}(\phi_x) = \widehat{\mathbb{U}}(\widehat{v}_{mH}(\ell\phi_x)) = \widehat{v}_{mH}(\widetilde{\phi}_x) &= \int_0^\infty \widetilde{\phi}_x(t)\widehat{v}_{mH}(dt) \\ &= \frac{1}{\sqrt{m}}\sum_{i=1}^m \big[\widetilde{\phi}_x(\widetilde{T}_i) - \mathsf{E}_{\hat{\theta}_m}\widetilde{\phi}_x(\widetilde{T}_i)\big].\end{aligned} \tag{6.23}$$

Here $\mathsf{E}_{\hat{\theta}_m}$ denotes the expected value with respect to the distribution $H_{\hat{\theta}_m}$. That means,

$$\mathsf{E}_{\hat{\theta}_m} f(\widetilde{T}) = \int_0^\infty f(s)h_{\hat{\theta}_m}(s)ds$$

for any integrable function f. We demonstrate in the next section that any functional from $\widehat{v}_{mH}(\widetilde{\phi}_x)$ is asymptotically distribution free.

6.4 Simulation results

6.4.1 Forms of GOF tests

The main purpose of this section is to show the asymptotically distribution free property of the new test statistics based on the transformed empirical process $\widehat{v}_{mH}(\widetilde{\phi}_x)$. Since the limit in distribution of $\widehat{v}_{mH}(\widetilde{\phi}_x)$ is a projected G-Brownian motion, which is completely specified and does not involve the asymptotic distribution H or any hypothetical parameter, we can take any proper functional from $\widehat{v}_{mH}(\widetilde{\phi}_x)$ to have an asymptotically distribution free GOF test. For example,

$$\int_0^\infty \widehat{v}^2_{mH}(\widetilde{\phi}_x)dx, \qquad \int_0^\infty \widehat{v}^2_{mH}(\widetilde{\phi}_x)dK(x),$$

where K is some specified measure on $(0,\infty)$, can be taken as test statistics.

For demonstration, we choose some widely used GOF tests, namely the Kolmogorov-Smirnov (KS), the Cramér-von Mises and the Anderson-Darling tests (see Chapter 4, Section 4.2.2) to illustrate their asymptotically distribution free property. These test statistics are now written as some specific functionals of the transformed process $\widehat{v}_{mH}(\widetilde{\phi}_x)$.

Namely, the form of the Kolmogorov-Smirnov test statistic is

$$KS = \max_x \left|\widehat{v}_{mH}(\widetilde{\phi}_x)\right|. \tag{6.24}$$

The Cramer-von Mises statistics is of the form

$$\Omega^2 = \int_0^\infty \widehat{v}^2_{mH}(\widetilde{\phi}_x)dG(x), \tag{6.25}$$

and its weighted version of Anderson-Darling statistic is

$$A^2 = \int_0^\infty \frac{\widehat{v}^2_{mH}(\widetilde{\phi}_x)}{G(x)(1-G(x))}dG(x). \tag{6.26}$$

In principle, we need to calculate the Ω^2 and A^2 statistics by integration from 0 to ∞. However, because

$$G(8) = 0.9997 \approx 1,$$

we can reduce the calculation time of the test value by choosing $x_{\max} = 8$ as a maximum value for x, and the range of values of x can be taken as

$$x \in \{0.1, 0.2, \cdots, 7.9, 8\}. \tag{6.27}$$

The Ω^2 and A^2 tests can be approximated as follows:

$$\Omega^2 \approx 0.1 \times \sum_x \widehat{v}_{mH}(\widetilde{\phi}_x)e^{-x},$$

$$A^2 \approx 0.1 \times \sum_x \frac{\widehat{v}_{mH}(\widetilde{\phi}_x)}{1-e^{-x}},$$

in which $\sum_x$ denotes the summation over the set in (6.27).

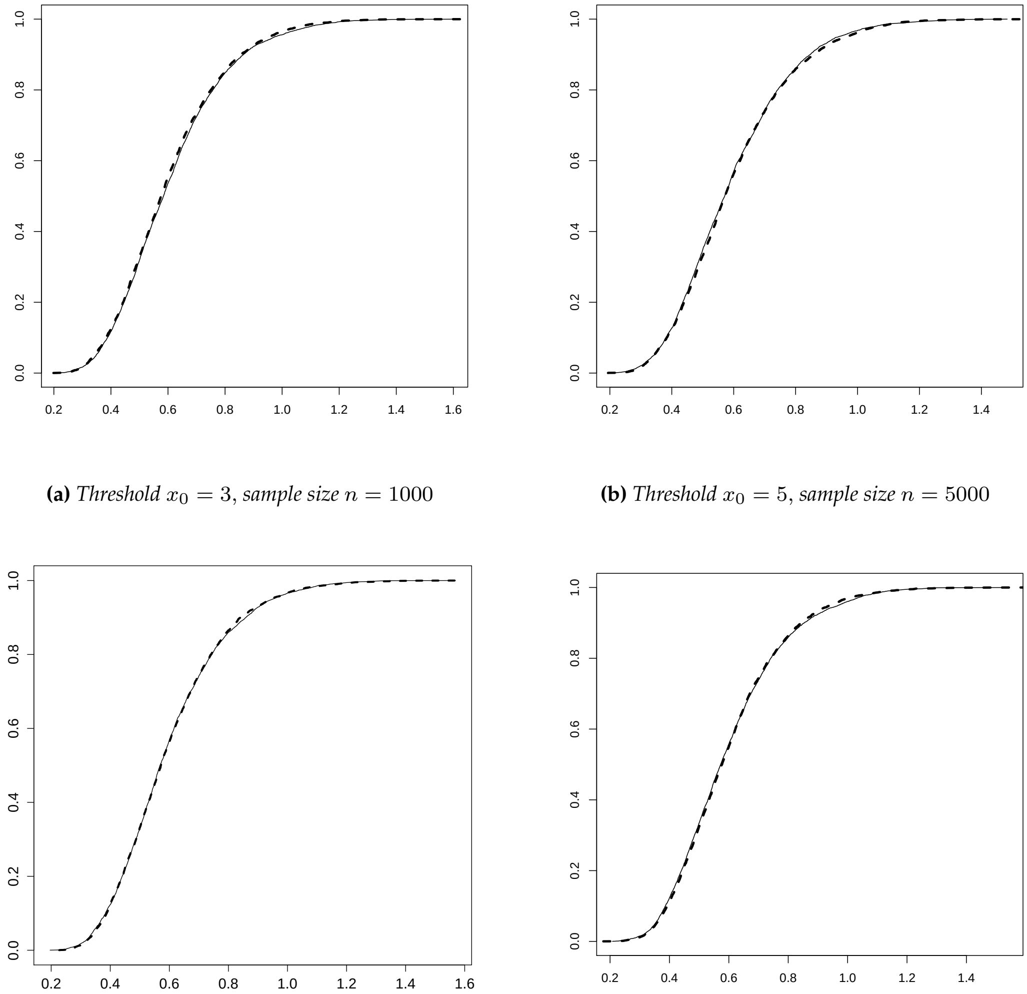
6.4.2 Distribution free property of test statistics

To create the curves of the cumulative distribution functions of the new tests, we choose two distributions whose tails are known to be regularly varying, the Pareto and Cauchy distributions, as underlying distributions.

For the Pareto distributions, the exponent is arbitrarily selected and positive. We chose some θ_0 ranging from 0.5 to 10. Note that the bigger the exponent θ_0 is, the thinner the tail of the Pareto distribution becomes. Sample size n for simulation needs to be large to guarantee that the tail is sufficiently big, namely, the number of observations m on the tail above some threshold x_0 is not less than 40. Hence, we chose a different threshold x_0, namely $x_0 = 3, 5, 8, 10$ and the sample size chosen to be increased correspondingly, namely $1000, 5000, 7000, 8000$. These choices guarantee a proper tail in the sense that the tail is thick enough. For example, for Pareto distribution with $\theta_0 = 2$, the sub-sample size is expected to be around $120, 190, 100, 80$ respectively. The Cauchy distribution appears to be thicker than the Pareto distribution where the sub-sample size is around $120, 310, 300, 250$ respectively. Usually, each curve is produced by 5000 iterations of simulation.

Figures 6.1, 6.2 and 6.3 show the plots of the cumulative distribution functions of the KS, Ω^2, A^2 test statistics respectively with different choices of the threshold x_0, different sample sizes and different parameters of the underlying Pareto distributions. As can be seen in these figures, the two curves of the cumulative distribution functions in each plot are not distinguishable, which illustrates very clearly that our approach involving the asymptotically distribution free property of the new test statistics is eminently practicable. Moreover, we notice that for different values of x_0, the difference between these curves is also very minor.

Regarding the computation time of the procedure, it took approximately 1 hour to create the cumulative distribution functions with $x_0 = 5$ and sample size $n = 5000$ by 5000 iterations for two different original distributions F. Therefore, we believe that the method is easy and efficient to implement.



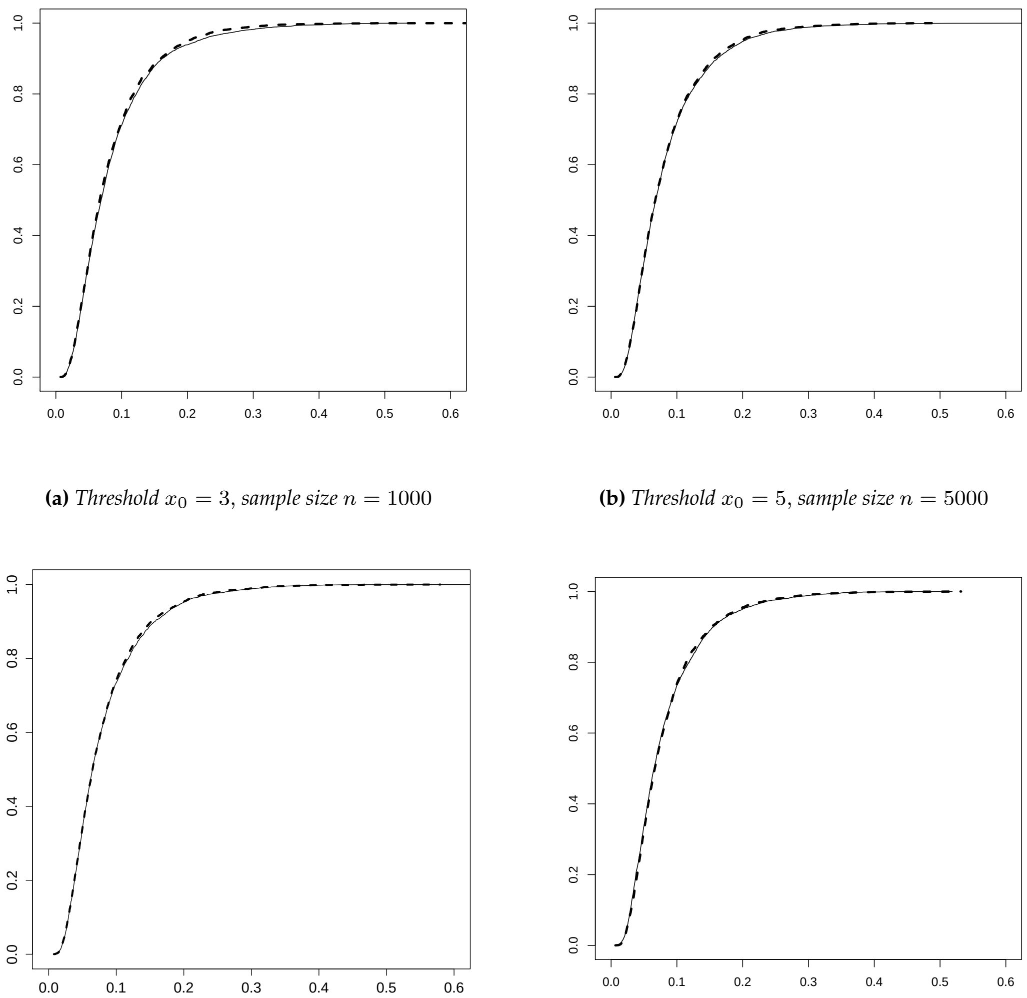
(a) *Threshold* $x_0 = 3$, *sample size* $n = 1000$

(b) *Threshold* $x_0 = 5$, *sample size* $n = 5000$

(c) *Threshold* $x_0 = 8$, *sample size* $n = 7000$

(d) *Threshold* $x_0 = 10$, *sample size* $n = 8000$

Figure 6.1: Distributions of the KS test statistics. Solid line: Pareto distribution with $\theta_0 = 3$; Dashed line: Cauchy distribution



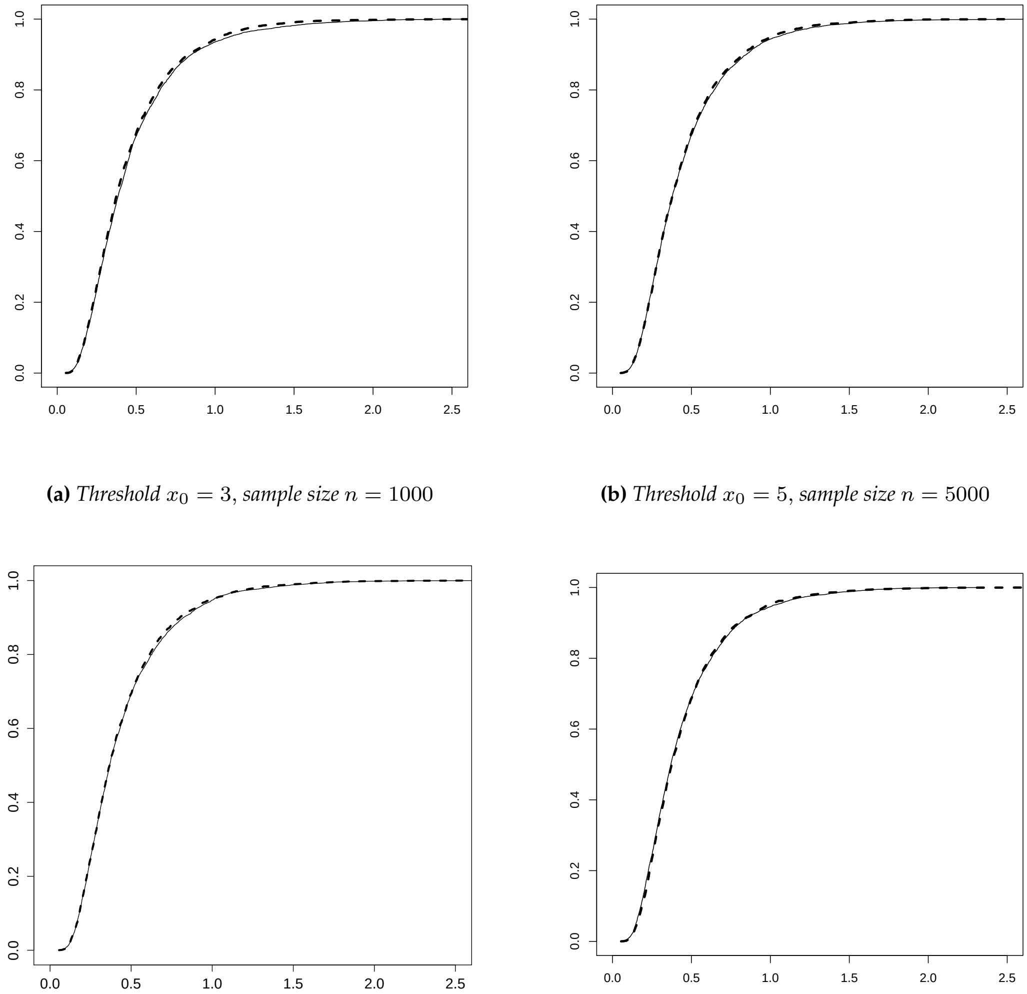
(a) *Threshold* $x_0 = 3$, *sample size* $n = 1000$

(b) *Threshold* $x_0 = 5$, *sample size* $n = 5000$

(c) *Threshold* $x_0 = 8$, *sample size* $n = 7000$

(d) *Threshold* $x_0 = 10$, *sample size* $n = 8000$

Figure 6.2: Distributions of the Ω^2 test statistics. Solid line: Pareto distribution with $\theta_0 = 2$; Dashed line: Cauchy distribution



(a) *Threshold* $x_0 = 3$, *sample size* $n = 1000$

(b) *Threshold* $x_0 = 5$, *sample size* $n = 5000$

(c) *Threshold* $x_0 = 8$, *sample size* $n = 7000$

(d) *Threshold* $x_0 = 10$, *sample size* $n = 8000$

Figure 6.3: Distributions of the A^2 test statistics. Solid line: Pareto distribution with $\theta_0 = 0.5$; Dashed line: Cauchy distribution

Chapter 7

Conclusions

In this closing part of the thesis, we will briefly review what has been done in the main content of the thesis and discuss possible extensions of the result.

Overall, this thesis did not propose any particular GOF test but gave a construction of a whole class of asymptotically distribution free GOF tests for each of two different parametric hypothesis testing problems.

Recall that the main idea of each construction is that, from the function-parametric empirical process which has a limiting distribution depending on the estimated parameter as well as the underlying distribution, one can map it to another process which has the limiting distribution specified. The specified limit is the projected Brownian motion in the time of a specified distribution. One can choose any suitable specified distribution, and we gave our choice for each problem in Chapters 5 and 6.

In Chapter 5, the chosen specified distribution for the problem of testing independence of two discrete random variables/vectors in the contingency table context is in fact a certain member of the parametric family. We created tables of critical values for the new KS and Ω^2 test given in Appendices A and B, for tables of dimension from 2×2 to 8×8. This does not mean that the maximum dimension of the table where we can apply the method presented is 8×8. We have shown that if the number of parameters $d = I + J$ (the dimension of tables is $(I+1) \times (J+1)$) is as big as 30 (see simulations in Section 5.7), the method still works well. We have not yet been able to confirm the maximum number of parameters for the method to be efficient; but it seems that the method can still work well for a larger number of parameters rather than just 30. This material has been published in [58].

The content of Chapter 6 is a part of our on-going research where the work has been done for testing regular variation in the tails of one-dimensional distributions. Noting that the choice of the specified distribution is not necessarily a member of the parametric family, we chose the specified distribution G to be another distribution which does not belong to the family of regularly varying tail distributions. In our opinion, the choice of G being a standard exponential distribution is simple enough to implement the method.

We are considering a possible extension to testing multivariate distributions whose tails are regularly varying. The motivation of this extension is that multivariate regular variation in the tail of distributions again plays an essential role in characterization of domain of attraction of multivariate extreme value distributions; see for example **Resnick** [65]. In a recent study by **Can et al.** [6], the authors showed a method of mapping the tail empirical process of extreme value distributions, which is constructed from tail copulas, to another process which converges to a standard Brownian motion. That mapping leads to a new class of GOF tests for testing multivariate extreme value distributions. We believe that we can extend the construction in Chapter 6 into testing multidimensional continuous distributions whose tails are regularly varying without changing the main spirit of the Khmaladze-2 transformation. The remaining key is to characterize the multivariate regular variation condition in a sufficiently simple way that we can find an asymptotic distribution H (see below) of the conditional distribution on the tail. In particular, let us recall the definition of multi-dimensional regularly varying distributions.

Definition 7.1. (see, for examples **Resnick** [65]) F is a distribution in $\mathbb{R}^d$. F is regularly varying tail if F satisfies the regular variation condition

$$\lim_{t\to\infty} \frac{1-F(t\mathbf{x})}{1-F(t\mathbf{1})} = H(\mathbf{x}) > 0, \quad \mathbf{x} > 0,$$

where $H(c\mathbf{x}) = c^{-\theta}H(\mathbf{x}), c > 0, \mathbf{x} > \mathbf{0}, \theta > 0$. In this case, F belongs to the domain of attraction of a multivariate extreme value distribution G where

$$G(\mathbf{x}) = \exp(-H(\mathbf{x})), \quad \mathbf{x} > \mathbf{0}.$$

Here $\mathbf{x}, \mathbf{1}, \mathbf{0}$ denote vectors in $\mathbb{R}^d$.

From the definition, the properties of $H(\mathbf{x})$ are represented only through the relation $H(c\mathbf{x}) = c^{-\theta}H(\mathbf{x})$. We need to specify the asymptotic form of H in terms of $\mathbf{x}$, and that is a part of what remains to be studied.

Appendix A

Tables of critical values for KS test

This appendix will provide the table of critical values for the KS test statistic presented in Chapter 5. Recall that

$$KS = \max_{(1,1)\leq(i,j)\leq(I+1,J+1)} \left|V_{n,ij}^{Z}\right|$$

where V_n^Z is defined in (5.20).

To get this final form of the test statistic, we map $\widehat{T}_n$ into $\widehat{Z}_n$ by the method presented in Section 5.3 and calculate $V_{n,ij}^{Z}$ as in (5.20). This given table is created with the choice of the collection of $r^{(\alpha)}$ given in (5.21), (5.22) and (5.23).

0.9	0.95	0.99	0.9	0.95	0.99	0.9	0.95	0.99	0.9	0.95	0.99	0.9	0.95	0.99	0.9	0.95	0.99	0.9	0.95	0.99
2x2 (60)			**2x3 (60)**			**2x4 (60)**			**2x5 (60)**			**2x6 (80)**			**2x7 (80)**			**2x8 (80)**		
0.83	0.98	1.31	1.12	1.29	1.62	1.34	1.55	1.92	1.54	1.75	2.19	1.74	1.98	2.45	1.9	2.19	2.73	2.06	2.32	2.86
3x3 (60)			**3x4 (60)**			**3x5 (60)**			**3x6 (60)**			**3x7 (80)**			**3x8 (80)**					
1.48	1.67	2.03	1.75	1.96	2.44	2	2.25	2.7	2.2	2.5	3.06	2.47	2.75	3.36	2.66	2.98	3.62			
4x4 (60)			**4x5 (60)**			**4x6 (60)**			**4x7 (80)**			**4x8 (80)**								
2.07	2.31	2.83	2.4	2.66	3.25	2.66	2.96	3.66	2.91	3.22	3.87	3.18	3.5	4.18						
5x5 (80)			**5x6 (80)**			**5x7 (80)**			**5x8 (80)**											
2.72	3.01	3.63	3.03	3.38	4.13	3.32	3.67	4.42	3.6	4	4.75									
6x6 (80)			**6x7 (80)**			**6x8 (80)**														
3.41	3.78	4.54	3.7	4.11	4.93	3.98	4.39	5.29												
7x7 (100)			**7x8 (100)**																	
4.05	4.51	5.39	4.41	4.88	5.83															
8x8 (100)																				
4.77	5.25	6.38																		

Table A.1: Table of critical values for KS statistics from the presented method

Appendix B

Tables of critical values for Ω^2 test

This appendix will provide the table of critical values for the Ω^2 test presented in Chapter 5. Recall that

$$\Omega^2 = \sum_{(1,1)\leq(i,j)\leq(I+1,J+1)} (V^Z_{n,ij})^2,$$

where V^Z_n is defined in (5.20).

To get this final form of the test statistic, we map $\widehat{T}_n$ into $\widehat{Z}_n$ by the method presented in Section 5.3 and calculate $V^Z_{n,ij}$ as in (5.20). This given table is created with the choice of the collection of $r^{(\alpha)}$ given in (5.21), (5.22) and (5.23).

0.9	0.95	0.99	0.9	0.95	0.99	0.9	0.95	0.99	0.9	0.95	0.99	0.9	0.95	0.99	0.9	0.95	0.99	0.9	0.95	0.99
2x2 (60)			**2x3 (60)**			**2x4 (60)**			**2x5 (60)**			**2x6 (80)**			**2x7 (80)**			**2x8 (80)**		
0.69	0.97	1.69	1.62	2.19	3.52	2.86	3.9	6.92	4.47	5.89	9.46	6.46	8.69	13.97	8.87	11.93	19.75	11.33	14.86	24.58
3x3 (60)			**3x4 (60)**			**3x5 (60)**			**3x6 (60)**			**3x7 (80)**			**3x8 (80)**					
3.85	5.12	7.76	6.75	8.73	13.74	10.77	13.65	22.4	15.32	19.6	30.46	21.35	27.03	42.18	27.6	36.12	54.38			
4x4 (60)			**4x5 (60)**			**4x6 (60)**			**4x7 (80)**			**4x8 (80)**								
11.93	15.35	24.32	19.2	24.3	37.5	27.54	34.57	53.18	37.39	47.14	70.27	48.66	61.76	95.84						
5x5 (80)			**5x6 (80)**			**5x7 (80)**			**5x8 (80)**											
29.3	37.26	57.83	42.42	54.03	83.47	57.94	72.24	110	77.8	95.8	144.4									
6x6 (80)			**6x7 (80)**			**6x8 (80)**														
62.4	78	119.6	82.7	105.5	163	109.4	136.3	207.3												
7x7 (100)			**7x8 (100)**																	
116.2	145.5	219.7	150.1	187.5	286.5															
8x8 (100)																				
196.2	248.7	367.7																		

Table B.1: Table of critical values for omega-square statistics from the presented method

Bibliography

[1] Agresti, A., *Categorical Data Analysis*. John Wiley & Sons, New York, (2003)

[2] Anderson, P. and Meerschaert, M. M., Periodic moving averages of random variables with regularly varying tails, *Annals of Statistics*, 25, pp. 771–785, (1997)

[3] Anderson, T. W. and Darlings, D. A., Asymptotic theory of certain "Goodness of fit" criteria based on stochastic processes, *Annals of Mathematical Statistics*, 23, pp. 193–212, (1952)

[4] Bedrick, E. J., Adjusted chi-squared tests for cross-classified tables of survey data, *Biometrika*, 70, pp. 591–595, (1983)

[5] Beirlant, J., de Wet, T. and Goegebeur, Y., A goodness of fit statistic for Pareto-type behaviour, *Journal of Computational and Applied Mathematics*, 186, pp. 99-116, (2006)

[6] Can, S. M, Einmahl, J. H. J, Khmaladze, E. V and Laeven, R. J. A., Asymptotically distribution-free goodness-of-fit testing for tail copulas, *Annals of Statistics*, 43 (2), pp. 878–902, (2015)

[7] Chibisov, D., Some theorems on the limiting behaviour of empirical distribution functions. *Selected Translation of Mathematical Statististics and Probabilities*, 6, pp. 147–156, (1964)

[8] Choulakian, V. and Stephens, M. A., Goodness of fit tests for the generalized Pareto distribution, *Technometrics*, 43, No. 4, pp. 478-484, (2001)

[9] Cramér, H., On the composition of elementary errors, *Skandinavisk Aktuarietidskrift*, 11, pp. 13–74 and 141–180, (1928)

[10] Danielsson, J., de Haan, L., Peng, L. and de Vries, C. G., Using a bootstrap method to choose the sample fraction in tail index estimation, *Journal of Multivariate Analysis*, 76, No. 2, pp. 226–248, (2001)

[11] Davis, R. and Resnick, S., Tail estimates motivated by extreme value theory, *Annals of Statistics*, 12, pp. 1467–1487, (1984)

[12] Debnath, L. and Mikusinski, P., *Hilbert spaces with applications*. Elsevier, 2005.

[13] De Haan, L. and Resnick, S. I., A simple asymptotic estimate for the index of a stable distribution, *Journal of Royal Statistical Society, Series B: Methodological*, 42, pp. 83–88, (1980)

[14] De Haan, L. and Resnick, S. I., On asymptotic normality of the Hill estimator, *Communications in Statistics. Stochastic Models*, 14(4), pp. 849–866, (1998)

[15] De Haan, L. and Ferreira, A., *Extreme value theory*. Springer, (2006)

[16] Dekkers, A. L. M., Einmahl, J. H. J. and de Haan, L., A moment estimator for the index of an extreme value distribution, *Annals of Statistics*, 17, pp. 1833–1855, (1989)

[17] Dietrich, D., de Haan, L. and Husler, J., Testing extreme value conditions, *Extremes*, 5, pp. 71–85, (2002)

[18] Drees, H. and Kaulfmann, E., Selecting the optimal sample fraction in univariate extreme value estimation, *Stochastic Processes and their Applications,* 75, No. 2, pp. 149–172, (1998)

[19] Drees, H., de Haan, L., Li, D., Approximations to the tail empirical distribution function with application to testing extreme value conditions, *Journal of Statistical Planning and Inference*, 136, pp. 3498–3538, (2006)

[20] Einmahl, J. H. J, The empirical distribution functions as a tail estimator, *Statistica Neerlandica*, 44, pp.79–82, (1990)

[21] Einmahl, J. H. J, Limit theorems for tail processes and application to intermediate quantile estimation, *Journal of Statistical Planning and Inference*, 32, pp. 137–145, (1992)

[22] Embrechts, P., Kluppelberg, C. and Mikosch, T., *Modelling extremal events for Insurance and Finance*. Springer-Verlag, Berlin, (1997)

[23] Feller, W., *An introduction to Probability theory and its applications*, Vol. 2. John Wiley & Sons, (1971)

[24] Geluk, J., de Haan, L., Resnick, S. and Starica, C., Second-order regular variation, convolution and the central limit theorem, *Stochastic Processes and their applications*, 69(2), pp. 139–159, (1997)

[25] Geluk, J. and de Haan, L., Stable probability distributions and their domains of attraction: a direct approach, *Probability and Mathematical Statistics*, 20, pp. 169–188, (2000)

[26] Gilula, Z. and Haberman, S. J., Canonical analysis of contingency tables by maximum likelihood, *Journal of the American Statistical Association*, 81, pp. 780–788, (1986)

[27] Gilula, Z., Krieger, A. M. and Ritov, Y., Ordinal Association in contingency tables: some interpretive aspects, *Journal of the American Statistical Association*, 83, pp. 540–545, (1988)

[28] Goodman, L. A., Association models and canonical correlation in the analysis of cross-classifications having ordered categories, *Journal of the American Statistical Association*, 76, pp. 320–334, (1981)

[29] Goodman, L. A., The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries, *Annals of Statistics*, 13, pp. 10–69, (1985)

[30] Guillou, A. and Hall, P., A diagnostic for selecting the threshold in extreme value analysis. *Journal of Royal Statistical Society, Series B: Methodology*, 63, No. 2, pp. 293–305, (2001)

[31] Haberman, S. J., Tests for independence in two-way contingency tables based on canonical correlation and on linear-by-linear interaction, *Annals of Statistics*, 9, pp. 1178–86, (1981)

[32] Haeusler, E. and Teugels, J. L., On asymptotic normality of Hill's estimator for the exponent of regular variation. *Annals of Statistics*, 13(2), pp. 743–756, (1985)

[33] Hall, P., On some simple estimates of an exponent of regular variation, *Journal of Royal Statistical Society, Series B: Methodology*, Vol. 44, No. 1, pp. 37–42, (1984)

[34] Hill, B. M., A simple general approach to inference about the tail of a distribution. *Annals of Mathematical Statistics*, 3, pp. 1163–1174, (1975)

[35] Holt, D., Scott, A. J., Ewings, P. D., Chi-squared tests with survey data, *Journal of Royal Statistical Society, Series B: Methodology*, 143, pp. 302–320, (1980)

[36] Jansen, D., and de Vries, C., On the frequency of large stock market returns: putting booms and busts into perspective, *Review of Economics and Statistics*, 73, pp. 18–24, (1991)

[37] Khmaladze, E., The use of ω^2 tests for testing parametric hypotheses, *Theory of Probability & Its Applications*, 24, pp. 283–301, (1979)

[38] Khmaladze, E., Martingale approach to the theory of goodness of fit tests, *Theory of Probability & Its Applications*, 26, pp. 240–257, (1982)

[39] Khmaladze, E., An innovation approach in goodness of fit tests in $\mathbb{R}^m$, *Annals of Statistics*, 16, pp. 1503–1516, (1988)

[40] Khmaladze, E., Goodness of fit problem and scanning innovation martingales, *Annals of Statistics*, 21, pp. 798–829, (1993)

[41] Khamaladze, E., Goodness of fit tests for "chimeric" alternatives, *Statistica Neerlandica*, 52, pp. 90–111, (1998)

[42] Khmaladze, E., Note on distribution free testing for discrete distribution, *Annals of Statistics*, 41, pp. 2979–2993, (2013)

[43] Khmaladze, E., *Statistical methods with applications to demography and life insurance*. Chapman and Hall, (2013)

[44] Khmaladze, E., Unitary transformations, empirical processes and distribution free testing, *Bernoulli*, 22, pp. 563–588, (2016)

[45] Koch, G. C., Freeman, D. H., Freeman, J. L, Strategies in the multivariate analysis of data from complex surveys, *International Statistical Review/Revue Internationale de Statistique*, 43, pp. 59–78, (1975)

[46] Kolmogorov A., Sulla determinazione empirica di una legge di distribuzione, *Giornale dell'Istituto Italiano degli Attuari*, 4, pp. 83–91, (1933)

[47] Koning, A. J. and Peng, L., Goodness of fit tests for heavy tailed distribution, *Journal of Statistical Planning and Inference*, 138, pp. 3960–3981, (2008)

[48] Kotulski, M., Asymptotic distributions of the continuous time random walks: a probabilistic approach, *Journal of Statistic Physics*, 81, pp. 777–792, (1995)

[49] Kuo, H.-H., *Gaussian measures Banach spaces*. Lecture notes in Math., 463, Springer, Berlin, (1975)

[50] Lehmann, E. L., *Testing statistical hypotheses*. Springer, (2005)

[51] Lehmann, E. L., *Theory of point estimation*. John Wiley & Sons, (1983)

[52] Lu, S. L., and Molz, F. J., How well are hydraulic conductivity variations approximated by additive stable processes?, *Advances in Environmental Research*, 5, pp. 39-45, (2001)

[53] Marohn, F., A characterization of generalized Pareto distributions by progressive censoring schemes and goodness of fit tests, *Communications in Statistics-Theory and Methods*, 31(7), pp. 1055–1065, (2002)

[54] Mason, D., Law of large numbers for sums of extreme values, *Annals of Probability*, 10, pp. 754–764, (1982)

[55] McCulloch, J., *Financial applications of stable distributions*, In: Maddala, G. S., Rao, C. R. (Eds), Statistical Methods in Finance, Handbook of Statistics, 14, North-Holland, New York, (1996)

[56] Metzler, R. and Klafter, J., The random walk's guide to anomalous diffusion: A fractional dynamics approach, *Physics report*, 339, pp. 1–77, (2000)

[57] Nelsen, R.B., *An introduction to Copulas*. Springer, (2006)

[58] Nguyen, T. M. Thuong, A new approach to distribution free tests in contingency tables, *Metrika*, 80(2), pp. 153–170, (2017)

[59] Nikias, C. and Shao, M., *Signal processing with alpha-stable distributions and applications*. Wiley, New York, (1995)

[60] Oosterhoof, J. and van Zwet, W., A note on contiguity and Hellinger distance. In *Contributions to Statistics: Jaroslav Hájek Memorial Volume*, Reidel, Dodrecht, (1979)

[61] O'Reilly, N., On the weak convergence of empirical processes in sup-norm metrics, *Annals of Probability*, 2, pp. 642–651, (1974)

[62] Pearson, K., On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, (5) 50, pp. 157–175, (1900)

[63] Pickands, J. III, Statistical inference using extreme order statistics, *Annals of Statistics*, 3, pp. 119–131, (1975)

[64] Pollard, D., *Empirical Processes: Theory and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics Vol. 2, Empirical Processes: Theory and Applications (1990), pp. i-iii+v+vii-viii+1–86, (1990)

[65] Resnick, S. I., *Extreme values, regular variation and point process*. Springer-Verlag, (1987)

[66] Resnick, S. I., Heavy tail modelling and teletraffic data, *Annals of Statistics*, 25, pp. 1805–1869, (1997)

[67] Resnick, S, I. and Stărică, C., Smoothing the Hill estimator, *Advances in Applied Probability*, 29(1), pp. 271–293, (1997)

[68] Rao, C. R, *Linear statistical inference and its applications*. John Wiley & Sons, New York, (1920)

[69] Rao, J. N. K and Scott, A. J., On chi-squared test for multiway contingency tables with cell proportions estimated from survey data, *Annals of Statistics*, 12, pp. 46-60, (1984)

[70] Shorack, G. R. and Wellner, J. A., *Empirical Processes with Applications to Statistics*. SIAM, (2009)

[71] Smith, R. L., Estimating tails of probability distributions, *Annals of Statistics*, 15, pp. 1174–1207, (1987)

[72] Smirnov, N. V., On the ω^2 distribution of von Mises, *Mat. Sb.* , 2(5), pp. 973—993 (In Russian) (French abstract), (1937)

[73] Smirnov, N. V., On the estimation of the discrepancy between empirical curves of distribution for two independent samples, *Moscow University Mathematics Bulletin*, Vol. 2, pp. 3–14, (1939)

[74] Szyszkowicz, B., The Chibisov-O'Reilly theorem for empirical processes under contiguous measures, *Statistics and Probability Letters*, 19, pp. 153–159, (1994)

[75] Teugels, J. L., Limit theorems on order statistics, *Annals of Probability*, Vol. 9, No. 5, pp. 868-880, (1981)

[76] van der Vaart, A. W. and Wellner, J. A., *Weak convergence and empirical processes: With applications to Statistics*, Springer, (1996)

[77] van der Vaart, A. W., *Asymptotic Statistics*. Cambridge University Press, Cambridge, (2000)

[78] von Mises, R., *Wahrscheinlichkeitsrechnung*. Leipzig-Wien, (1931)