Technische Universität München

Department of Mathematics

Clarke's Test For Non-Nested Model Comparison

Master Thesis

by

Florian Brück

Supervisor:	PD Dr. Aleksey Min
	Prof. Jean-David Fermanian
Advisor:	PD Dr. Aleksey Min
	Prof. Jean-David Fermanian
Submission Date:	June 28, 2019

I hereby declare that this hesis is my own work and that no other sources have been used except those clearly indicated and referenced.

Munich, 28.06.2019

Abstract

We study the theoretical properties of the model comparison test, introduced by Clarke in [6]. First, we show that the proposed test statistic is not Binomial distributed, contrary to what has originally been stated. Second, we propose a corrected test statistic and derive its asymptotic Normal distribution under some regularity assumptions. Further, we propose two estimators for the asymptotic variance of the new test statistic. In a Monte Carlo simulation study, we verify the above distributional approximation and investigate the empirical level and power of old and new Clarke's test statistics.

Zusammenfassung

Wir untersuchen die mathematischen Eigenschaften eines statistischen Tests zum Modellvergleich, der 2007 von Clarke in [6] vorgestellt wurde. Zuerst zeigen wir, dass die von Clarke vorgeschlagene Teststatistik nicht binomialverteilt ist, im Gegensatz zur Vermutung in [6]. Alternativ schlagen wir eine korrigierte Teststatistik vor und leiten ihre asymptotische Normalverteilung unter einigen Regularitätsannahmen her. Des Weiteren werden zwei Schätzer der asymptotischen Varianz dieser Teststatistik vorgestellt. In einer Monte-Carlo Simulationsstudie verifizieren wir die Verteilungsapproximation und untersuchen das Signifikanzniveau sowie die Trennschärfe des vorgeschlagenen Tests.

Contents

Chapter 1

Introduction

In statistics, independent realizations of a random vector X are often given. For many applications, it is crucial to deduce a suitable distributional approximation of the random vector X, based on the given observations. The maximum likelihood approach requires the choice of an appropriate density function for X. Usually, several candidate densities are available. Among these candidates, the density function for X with the best approximation should be selected. For this purpose, several approaches for density comparisons have been introduced in the past.

The most famous density comparison criteria are the AIC and the BIC proposed by Akaike [1] and Schwarz [16], respectively. They are based on the Kullback-Leibler Information Criterion (KLIC) [11], which measures the pseudo-distance between a proposed density and the true density of X. Both criteria penalize the negative log-likelihood of the proposed density proportional to the number of estimated parameters. The model with the lowest AIC or BIC is selected. However, neither approach does provide any information about the statistical significance of this choice.

This drawback of the AIC and BIC can be mitigated by the test for model comparison introduced by Vuong in [20], which is also based on the KLIC. Vuong derived the asymptotic distribution of the log-likelihood ratio of two competing densities, whose parameters are estimated. Based on this result, it is possible to test whether two competing densities are equally approximating the unknown density of X. Vuong's test is applicable for nested and non-nested model comparisons, meaning that one competing family is a subset of the other or that the competing densities do not coincide, respectively. Vuong's test has been generalized by Chen and Fan [5] for parametric copula density comparisons based on pseudo observations.

In [6], Clarke combines the ideas of Vuong's test and the paired sign test. He proposed

a test for non-nested model selection based on the median of the log-likelihood ratio of two competing density families. The test statistic simply counts the number of positive likelihood ratios greater than 1 and is claimed to be Binomial distributed. Clarke's test is widely applicable, since the median of a continuous distribution always exists. However, his paper lacks mathematical formalism.

In this master thesis, we provide a mathematical framework for Clarke's test. We introduce a rigorous mathematical formalism and present two examples of density comparisons satisfying Clarke's null hypothesis. The examples illustrate that Clarke's test statistic is not (asymptotically) Binomial distributed. Moreover, we prove that the Binomial distribution is not even a viable asymptotic approximation of the distribution of Clarke's test statistic.

Our main contribution is the derivation of the correct asymptotic variance of Clarke's test statistic. Using techniques from empirical process theory, we show that the properly normalized Clarke's test statistic is asymptotically Normal distributed. Since the derived asymptotic variance is not available in closed form, we propose a weakly consistent estimator. Additionally, we show the bootstrap consistency of an adjusted statistic. Based on both variance estimates, we introduce two test statistics for non-nested model selection. Furthermore, we conduct a simulation study to asses the finite sample behavior of the proposed test statistics and compare our results with those obtained by Vuong.

The master thesis is organized as follows. In Chapter 2, we recall important definitions and results from functional analysis, measure theory and probability theory. On the basis of these results, we introduce weak convergence and empirical process theory in Chapter 3. Chapter 4 is split into two subsections. In Subsection 4.1, we shortly motivate Clarke's test statistic and introduce several important notions. To specify the mathematical framework, we formally introduce Clarke's test in Subsection 4.2. In Subsection 4.3, we prove that Clarke's test statistic $\hat{B}_n$ is in general not Binomial distributed. In Chapter 5, we prove the asymptotic Normality of a suitably normalized Clarke's test statistic $n^{-1/2}(\hat{B}_n - n/2)$ under various assumptions on the involved families of densities and estimators. Moreover, we propose two approaches to estimate the asymptotic variance of the statistic $n^{-1/2}(\hat{B}_n - n/2)$. One estimator is based on numerical differentiation in combination with the usual sample variance estimation of a particular random variable, whereas the other estimator is based on a bootstrap approach. Combining these results, we are able to define two asymptotically standard Normal distributed test statistics for non-nested model selection. In Chapter 6, we compare the modified Clarke's test statistic introduced in this master thesis with the test proposed by Vuong in [20]. In Chapter 7, we calculate all relevant estimators and theoretical quantities introduced in Chapter 5

for two examples of competing strictly non-nested density families. On the basis of these calculations we conduct a simulation study to investigate the finite sample behavior of the estimators proposed in Chapter 5 and the test statistics given in (5.5) and (7.3). In Chapter 8, we give a summary of our results and discuss topics for future research. Most of the proofs can be found in Appendix A.

Chapter 2

Mathematical Preliminaries

In this chapter, we recall some important definitions and results from functional analysis, measure theory and probability theory, since they are crucial to understand the theoretical framework of empirical process theory. Most of the results in the probability theory part can be generalized to statements with outer probability, which will be presented in Chapter 3. Some theorems are reformulated to suit the context of this master thesis.

2.1 Preliminaries From Functional Analysis

The results and definitions from functional analysis are the basis to understand empirical process theory and are stated without comment. This section is based on [15].

Definition 2.1.1
A collection of subset τ of some set Ω is a topology if the following is true:

1. *$\emptyset \in \tau$ and $\Omega \in \tau$;*
2. *If for all $(A_i)_{1 \leq i \leq n} \in \tau$, then $\cap_{1 \leq i \leq n} A_i \in \tau$ for all $n \in \mathbb{N}$;*
3. *If $(A_i)_{i \in I} \in \tau$ for some index set I, then $\cup_{i \in I} A_i \in \tau$.*

The sets in τ are called the open sets and the tuple (Ω, τ) is called a topological space.

Definition 2.1.2
A function $f : (\Omega_1, \tau_1) \to (\Omega_2, \tau_2)$ is continuous if for every $A \in \tau_2$ we have $f^{-1}(A) \in \tau_1$.

Definition 2.1.3
The complement of a set A is defined as $A^{\complement} := \Omega \setminus A$. The closure of a set A is defined as $\overline{A} := \cap_{A \subseteq B; B^{\complement} \in \tau} B$ and the interior of a set is defined as $\mathring{A} := \cup_{B \subseteq A; B \in \tau} B$.

Definition 2.1.4
A map $\rho : \Omega \times \Omega \to [0, \infty)$ is called a semi-metric if

1. $\rho(x, y) = \rho(y, x)$,
2. $\rho(x, y) \leq \rho(x, z) + \rho(z, y)$.

Additionally, if $\rho(x, y) = 0$ is equivalent to $x = y$, ρ is called a metric. The tuple (Ω, ρ) is called a (semi-)metric space.

Lemma 2.1.5
Every semimetric on Ω induces a topology τ. τ is defined as the smallest topology containing the collection of sets $\{B_r(x) \mid r \in \mathbb{Q}; x \in \Omega\}$, where $B_r(x) := \{y \in \Omega \mid \rho(x, y) < r\}$.

Remark 1
Every semimetric is a metric on the space of equivalence classes of Ω. A point x is equivalent to a point y if $\rho(x, y) = 0$. Therefore, we can treat every semimetric space as metric space defined on the equivalence classes of Ω.

Definition 2.1.6
A topological space (Ω, τ) is separable if there exists a countable set $A \subseteq \Omega$ with $\overline{A} = \Omega$.

Lemma 2.1.7
Let T be an arbitrary uncountable set. The space $l^\infty(T) := \{f \mid f : T \to \mathbb{R}; sup_{t \in T}|f(t)| < \infty\}$ equipped with the norm $\|f\|_\infty := sup_{t \in T}|f(t)|$ is a non-separable Banach space.

Definition 2.1.8
A set $K \subseteq \Omega$ is compact if every open cover of K has a finite subcover. A set $K \subset \Omega$ is σ-compact if K is the countable union of compact sets.

Lemma 2.1.9
A σ-compact set in a metric space is separable.

Definition 2.1.10
A set $K \subset (\Omega, \rho)$ is totally bounded if there exist finitely many ρ-balls covering K.

Definition 2.1.11
Consider the space $(l^\infty(T), \|\cdot\|_\infty)$, where (T, ρ) is a semimetric space. The subspace of $l^\infty(T)$ containing all functions $f : T \to \mathbb{R}$ satisfying

$$\lim_{\delta \to 0} \sup_{\rho(s,t) < \delta} |f(s) - f(t)| = 0$$

is called the space of uniformly continuous functions w.r.t. ρ and is denoted as $UC(T, \rho)$.

Theorem 2.1.12
The closure of a set $K \in (l^\infty(T), \|\cdot\|_\infty)$ is σ-compact iff $K \subset UC(T, \rho)$ for some semi-metric ρ making T totally bounded.

2.2 Preliminaries From Measure Theory

In this section, we recall some important notions from measure theory, which will be used frequently in Chapter 3. The results are based on [7] as well as on [15] and are stated without comment.

Definition 2.2.1
A collection of sets $\mathcal{A} \subset \Omega$ is called a sigma algebra if it satisfies the following properties:

1. *$\Omega \in A$;*
2. *If $A \in \mathcal{A}$ then $A^\complement \in \mathcal{A}$;*
3. *If $(A_i)_{i\in\mathbb{N}} \in \mathcal{A}$ then $\cup_{i\in\mathbb{N}} A_i \in \mathcal{A}$.*

The tuple $(\Omega, \mathcal{A})$ is called a measurable space.

Remark 2
Every topological space induces a measurable space by choosing $\mathcal{A}$ to be equal to the smallest sigma-algebra containing τ.

Definition 2.2.2
Let (Ω, τ) denote a topological space. The Borel sigma-algebra $\mathcal{B}(\Omega)$ on Ω is defined as the smallest sigma-algebra on Ω containing τ.

Definition 2.2.3
Let $(\Omega_1, \mathcal{A}_1)$ denote a measurable space and (Ω_2, τ) denote a topological space. A function $f : \Omega_1 \to \Omega_2$ is called Borel-measurable if $f^{-1}(A_2) \in \mathcal{A}_1$ for every $A_2 \in \tau$.

Remark 3
By the definition of the Borel sigma-algebra on a topological space (Ω, τ), it is obvious that the Borel sigma-algebra is the smallest sigma-algebra making all functions in the set $C_b(\Omega) := \{f \mid f : \Omega \to \mathbb{R}; f$ is bounded and continuous$\}$ measurable.

Definition 2.2.4
Let $(\Omega, \mathcal{A})$ be a measurable space. A map $\mu : \mathcal{A} \to [0, \infty]$ is called a measure if it satisfies the following properties

1. $\mu(\emptyset) = 0$;
2. *If* $(A_i)_{i\in\mathbb{N}}$ *are disjoint sets* $\mu(\cup_{i\in\mathbb{N}} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$.

Definition 2.2.5
Let μ *denote a measure on the measurable space* $(\Omega, \mathcal{A})$. *The triplet* $(\Omega, \mathcal{A}, \mu)$ *is called a measure space. If* $\mu(\Omega) = 1$, $(\Omega, \mathcal{A}, \mu)$ *is called a probability space.*

Definition 2.2.6
For any Borel-measurable function $f : (\Omega, \mathcal{A}, \mu) \to \mathbb{R} \cup \{\pm\infty\}$, *we define*

$$\int f d\mu = \int \max\{f, 0\} d\mu - \int \max\{-f, 0\} d\mu$$

if $\int \max\{f, 0\} d\mu < \infty$ *or* $\int \max\{-f, 0\} d\mu < \infty$.

Theorem 2.2.7
There exists a set $A \subset [0, 1]$ *which is not Borel-measurable.*

2.3 Preliminaries From Probability Theory

In this section, we recall some important notions from probability theory. This section is based on [7].

Definition 2.3.1
Let $(\Omega, \mathcal{A}, P)$ *denote a probability space and* $(\tilde{\Omega}, \rho)$ *denote a semi-metric space. A map* $X : (\Omega, \mathcal{A}, P) \to (\tilde{\Omega}, \rho)$ *a called a random variable if it is measurable w.r.t. the Borel sigma-algebra on* $\tilde{\Omega}$ *generated by* ρ.

Remark 4
Note that a random variable X *induces a probability measure* P_X *on* $\tilde{\Omega}$ *by defining* $P_X(A) = P(X \in A)$ *for any set* A *in the Borel sigma-algebra on* $\tilde{\Omega}$. *In abuse of the notation, we will sometimes use* P *and* P_X *interchangeably, since we are usually interested in* P_X *only and do not care about the underlying probability space.*

Now, we present three modes of convergence of random variables, namely almost sure convergence, convergence in probability and convergence in distribution. We start with almost sure convergence, which is defined in the following definition.

Definition 2.3.2
A sequence of random variables $(X_i)_{i\in\mathbb{N}}$ *converges* P*-almost surely to a random variable* X, *denoted as* $X_i \overset{a.s.}{\to} X$, *if*

$$P\left(\lim_{i\to\infty} \rho(X_i, X) = 0\right) = 1.$$

The P in P-almost surely is usually suppressed if the underlying probability measure is known from the context. Almost sure convergence is the strongest of the three modes of convergence introduced in this chapter. It implies convergence in probability, defined in the next definition.

Definition 2.3.3
A sequence of random variables $(X_i)_{i\in\mathbb{N}}$ converges in P-probability to a random variable X, denoted as $X_i \xrightarrow{P} X$, if for all $\epsilon > 0$:

$$\lim_{i\to\infty} P\left(\rho(X_i, X) > \epsilon\right) = 0.$$

Again, the P in P-probability is usually suppressed if the underlying probability measure is known from the context. Convergence in probability implies convergence in distribution, which is the weakest mode of convergence introduced in this thesis.

Definition 2.3.4
A sequence of random variables $(X_i)_{i\in\mathbb{N}}$ converges in distribution to a random variable X, denoted as $X_i \xrightarrow{d} X$, if for all $f \in C_b(\tilde{\Omega}) := \{f : \tilde{\Omega} \to \mathbb{R};\ f \text{ is bounded and continuous}\}$:

$$\lim_{i\to\infty} \mathbb{E}\left[f(X_i)\right] = \mathbb{E}\left[f(X)\right].$$

As it is not easy to check the condition $\mathbb{E}\left[f(X_i)\right] \to \mathbb{E}\left[f(X)\right]$ for arbitrary random variables $(X_i)_{i\in\mathbb{N}}$ and X, it is useful to have equivalent conditions to verify convergence in distribution. The following theorem is known as the Portmanteau-Theorem and states several equivalent conditions for convergence in distribution.

Theorem 2.3.5
The following statements are equivalent:

1. $X_i \xrightarrow{d} X$
2. *For every open set $G \subseteq \tilde{\Omega}$: $\liminf_{i\to\infty} P\left(X_i \in G\right) \geq P\left(X \in G\right)$*
3. *For every closed set $F \subseteq \tilde{\Omega}$: $\limsup_{i\to\infty} P\left(X_i \in F\right) \leq P\left(X \in F\right)$*
4. *For every Borel-measurable set $B \subseteq \tilde{\Omega}$ with $P\left(X \in \overline{B} \setminus \mathring{B}\right) = 0$: $\lim_{i\to\infty} P\left(X_i \in B\right) = P\left(X \in B\right)$.*

If $\tilde{\Omega} \subseteq \mathbb{R}^d$, the following condition is equivalent to convergence in distribution.

5. *For any continuity point c of the distribution function of X : $\lim_{i\to\infty} P(X_i \leq c) = P(X \leq c)$.*

We previously mentioned that the three introduced modes of convergence are related. The following theorem summarizes these relations and provides other important properties of the three modes of convergence.

Theorem 2.3.6
The following statements are true:

1. *If* $X_i \xrightarrow{a.s.} X$, *then* $X_i \xrightarrow{P} X$.
2. *If* $X_i \xrightarrow{P} X$, *then* $X_i \xrightarrow{d} X$.
3. *If* $X_i \xrightarrow{d} a$, *where* a *is a constant real number, then* $X_i \xrightarrow{P} a$.
4. *If* $X_i \xrightarrow{P} X, Y_i \xrightarrow{P} Y$ *and* $Z_i \xrightarrow{P} Z$, *then* $X_i Y_i + Z_i \xrightarrow{P} XY + Z$.

Remark 5
Note that, in general, we do not have a similar statements as 4. *of Theorem 2.3.6 for convergence in distribution. To see this, assume that* X *is a non-constant and* 0*-symmetric random variable taking values in* $\mathbb{R}$. *Then the distribution of* X *and* $-X$ *are identical, i.e.* $\mathbb{E}[f(X)] = \mathbb{E}[f(-X)]$ *for all* $f \in C_b(\mathbb{R})$. *Define* $X_i := X$ *and* $Y_i := -X$ *and observe that* $X_i \xrightarrow{d} X$ *as well as* $Y_i \xrightarrow{d} X$, *but* $X_i + Y_i = 0 \not\xrightarrow{d} 2X$. *However, one can show that in the case of independent sequences, statement* 4. *in Theorem 2.3.6 is valid with* $\xrightarrow{P}$ *replaced by* $\xrightarrow{d}$.

The following theorem is known as Slutsky's Lemma and will be important in later applications.

Theorem 2.3.7
Let $(X_i)_{i\in\mathbb{N}}, (Y_i)_{i\in\mathbb{N}}$ *and* $(Z_i)_{i\in\mathbb{N}}$ *denote sequences of random variables with* $X_i \xrightarrow{d} X$, $Y_i \xrightarrow{P} a$ *and* $Z_i \xrightarrow{P} b$, *where* a *and* b *are constant real numbers. Then*

$$Y_i X_i + Z_i \xrightarrow{d} aX + b.$$

The next result, known as the Glivenko-Cantelli Theorem, is presented because one of the main motivations behind empirical process theory is to find generalizations of this statement.

Theorem 2.3.8
Let $(X_i)_{i\in\mathbb{N}}$ *be independent and identically distributed (i.i.d.) random variables in* $\mathbb{R}^d$ *with distribution function* F. *Then*

$$\lim_{n\to\infty} \sup_{t\in\mathbb{R}} \left| F(t) - \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \le t\}} \right| = 0 \text{ almost surely.}$$

Having the Glivenko-Cantelli Theorem at hand, one can ask the question whether the rate of convergence of $n^{-1}\sum_{i=1}^{n}\mathbf{1}_{\{X_i\leq\cdot\}}$ to $F(\cdot)$ is uniform in $l^\infty(\mathbb{R}^d)$. In Chapter 3, we present the answer to this question, known as Donsker's Theorem. It tells us that the uniform rate of convergence is $\sqrt{n}$. Another main motivation of empirical process theory is to provide general conditions implying the rate of convergence in a generalized Glivenko-Cantelli Theorem to be uniform. To understand the arising limit objects, we need to introduce a Brownian motion and Gaussian processes, which are stochastic processes defined on some probability space.

Definition 2.3.9
Let T be an arbitrary set and $X_t : (\Omega, \mathcal{A}, P) \to (\tilde{\Omega}, \rho)$ be Borel-measurable for all $t \in T$. The collection of random variables $(X_t)_{t\in T}$ is called a stochastic process indexed by T.

Definition 2.3.10
A Brownian motion B is a stochastic process $B(t)_{t\in[0,\infty)}$ defined on a probability space $(\Omega, \mathcal{A}, P)$ such that the following is true:

1. *$B(0) = 0$ almost surely,*
2. *$B : [0,\infty) \to \mathbb{R}$ is continuous almost surely,*
3. *For $0 \leq t_0 < ... < t_n$: The increments $(B(t_n) - B(t_{n-1})), ..., (B(t_1) - B(t_0))$ are independent and $B(t_i) - B(t_{i-1})$ is Normal distributed with mean zero and variance $(t_i - t_{i-1})$.*

A Brownian motion belongs to the class of Gaussian processes, defined in the next definition.

Definition 2.3.11
A Gaussian process $(G(t))_{t\in[0,\infty)}$ is a stochastic process such that for any $(t_i)_{1\leq 1\leq n} \in [0,\infty)$, the vector $(G(t_1), ..., G(t_n))$ has a multivariate Normal distribution.

The next theorem collects some useful properties of Brownian motion, which are important to understand the limit objects of empirical process theory.

Theorem 2.3.12
Let $\mathcal{A}_t$ denote the smallest sigma-algebra such that $B(s)_{s\leq t}$ is measurable on $(\Omega, \mathcal{A}_t)$. Then $B(t)$ is adapted to the filtration $(\mathcal{A}_t)_{t\in[0,\infty)}$ and the following is true

1. *$B(t)$ is a martingale w.r.t. $\left(\Omega, (\mathcal{A}_t)_{t\in[0,\infty)}, P\right)$.*
2. *$B(t)$ is a Gaussian process with covariance function $\mathrm{Cov}(B(s), B(t)) = \min\{s,t\}$.*

3. *For* $t \in [0,1] : \mathbb{G}(t) := B(t) - tB(1)$ *is a continuous, mean zero Gaussian process with covariance function* $\mathrm{Cov}(\mathbb{G}(s), \mathbb{G}(t)) = \min\{s,t\} - st$. $\mathbb{G}$ *is called a Brownian bridge on* $[0,1]$

The last important notion needed to introduce empirical process theory is given in the following definition.

Definition 2.3.13
A probability measure P *is tight if for every* $\epsilon > 0$ *there exists a compact set* $K \subseteq \Omega$ *such that* $P(K) \geq 1 - \epsilon$.

Remark 6
From the definition of tight laws, we deduce that every tight P *resides in a sigma compact set* $\tilde{K}$ *with probability* 1.

In Chapter 3, we will see that many limit laws on $l^\infty(T)$ are tight. A tight law on $l^\infty(T)$ resides in the set $UC(T,\rho)$ with probability 1 for some semimetric ρ making T totally bounded, by Theorem 2.1.12. Therefore, many of the limit laws in Chapter 3 will possess certain continuity properties.

Chapter 3

Empirical Process Theory

3.1 Motivation

Consider i.i.d. random variables $(X_i)_{i\in\mathbb{N}}$ distributed according to a distribution function F. The empirical distribution function of $(X_i)_{1\leq i\leq n}$ is defined as $F_n(t) := n^{-1}\sum_{i=1}^n \mathbf{1}_{\{X_i\leq t\}}$. The functions F_n and F can be viewed as functions in $l^\infty(\mathbb{R})$. By the Glivenko-Cantelli Theorem, Theorem 2.3.8, the function F_n almost surely converges to the function F in the space $(l^\infty(\mathbb{R}), \|\cdot\|_\infty)$. This is a uniform strong law of large numbers for the random variables $F_n(t)_{t\in\mathbb{R}}$. A natural question to ask is whether there also a uniform central limit theorem for the random variables $F_n(t)_{t\in\mathbb{R}}$? Mathematically, we can phrase the question as follows: Does the process $\mathbb{G}_n(\cdot) := \sqrt{n}\,(F_n(\cdot) - F(\cdot))$ converge in distribution to some Gaussian process $\mathbb{G}$ in $l^\infty(\mathbb{R})$?
The answer to this question is "yes" and is given by empirical process theory. However, we have to pay a price for considering F_n as a function in the space $l^\infty(\mathbb{R})$ equipped with the norm $\|\cdot\|_\infty$. The problem arises from the following example:

Consider $(X_i)_{i\in\mathbb{N}}$ as i.i.d. random variables on the product space $\big([0,1]^{\mathbb{N}}, \mathcal{B}[0,1]^{\mathbb{N}}, \lambda\!\!\!^{-}[0,1]^{\mathbb{N}}\big)$, where $\lambda\!\!\!^{-}[0,1]$ denotes the Lebesgue measure on $[0,1]$. In this case, we can view the resulting process F_n as a map in $(l^\infty([0,1]), \|\cdot\|_\infty)$. Now, choose a set $A \subset [0,1]$ which is not Borel-measurable (such a set exists by Theorem 2.2.7) and consider the set $C := \{\mathbf{1}_{\{a\leq\cdot\}} \mid a \in A\} \subset l^\infty([0,1])$. Note that C is a closed set in $l^\infty([0,1])$. To see this, choose an arbitrary $f \in C^{\complement}$ and observe that every function in C only takes values in $\{0,1\}$. If f does not exclusively take values in $\{0,1\}$, there exists $t \in [0,1]$ with $f(t) = y \notin \{0,1\}$. Choosing $r = \min\{|y|/2, (|1-y|)/2\}$, we get that $B_r(f) \subset C^{\complement}$, because for all $g \in B_r(f)$, we have $g(t) \notin \{0,1\}$. If f only takes values in $\{0,1\}$, we can deduce that $\|f - \mathbf{1}_{\{a\leq\cdot\}}\|_\infty = 1$, since $|f(t) - \mathbf{1}_{\{a\leq t\}}| \in \{0,1\}$ for any $t \in [0,1]$. Therefore, $B_{\frac{1}{2}}(f) \subset C^{\complement}$. Combining the

arguments, we get that $C^{\complement}$ is open in $l^\infty([0,1])$. Since every closed set in $l^\infty([0,1])$ is Borel-measurable, C is a Borel-measurable set. Furthermore, the set $\tilde{C} := \{f - F \mid f \in C\}$ is a Borel-measurable set since the translation of a closed set is closed.
To prove that the process $\mathbb{G}_n(\cdot) := \sqrt{n}\,(F_n(\cdot) - F(\cdot))$ converges in distribution to some Gaussian process $\mathbb{G}$ in $l^\infty([0,1])$, we need to verify one of the four equivalent conditions in Theorem 2.3.5. However, to be able to formulate any of these conditions, it is required that the map

$$\mathbb{G}_1 : \ ([0,1], \mathcal{B}[0,1], \lambda[0,1]) \to (l^\infty[0,1], \|\cdot\|_\infty) \ ; \ \omega \mapsto \mathbf{1}_{\{\omega \leq \cdot\}} - F(\cdot)$$

is Borel-measurable. Unfortunately,

$$\mathbb{G}_1^{-1}(\tilde{C}) = \{\omega \mid \mathbf{1}_{\{\omega \leq \cdot\}} \in C\} = A \notin \mathcal{B}[0,1],$$

thus $\mathbb{G}_1$ is not Borel-measurable. This lack of measurability persists for any fixed $n \in \mathbb{N}$. Therefore, we are not able to formulate any of the conditions of Theorem 2.3.5.

There are two possible solutions to this problem. The first solution is to abandon the metric $\|\cdot\|_\infty$ and to reduce the function space $l^\infty(\mathbb{R})$ to the space of Càdlàg functions on $\mathbb{R}$, equipped with the Skorokhod metric. The Skorokhod metric is a weaker metric than the metric induced by $\|\cdot\|_\infty$. This implies that the Borel sigma-algebra on $l^\infty(\mathbb{R})$ contains fewer sets and that $\mathbb{G}_n$ is measurable. In this particular setting, the definition of convergence in distribution is well defined and one can prove a functional central limit theorem, which is known as Donsker's Theorem.

Theorem 3.1.1
Let F_n denote the empirical distribution function of a sequence of i.i.d. random variables $(X_i)_{i\in\mathbb{N}}$ with distribution function F. Then,

$$\sqrt{n}\,(F_n(\cdot) - F(\cdot)) \xrightarrow{d} \mathbb{G}$$

in the space of Càdlàg functions on $\mathbb{R}$, equipped with the Skorokhod metric. $\mathbb{G}$ is a Gaussian process with covariance function $\operatorname{Cov}(\mathbb{G}(s), \mathbb{G}(t)) = \min\{F(s), F(t)\} - F(s)F(t)$ and can be written as $\mathbb{G}(t) = G(F(t))$, where G is a Brownian Bridge on $[0,1]$.

It turns out that this approach is useful in the case of empirical distribution functions, but does not allow for much greater generality.

The second solution to the problem described above allows us to keep the metric $\|\cdot\|_\infty$ at the cost of abandoning the measurability of the process $\mathbb{G}_n$. As long as the limit process $\mathbb{G}$

is Borel-measurable, one can replace the expectations in Definition 2.3.4 by a more general construct, which handles non-measurable quantities. It turns out that this approach allows us to define a useful notion of convergence in distribution in $l^\infty(T)$, from now on referred to as weak convergence. This approach is known as empirical process theory and will be introduced in the following. We start by introducing outer probability, which is the basis of the new definition of convergence in distribution, introduced in Chapter 3.3. Chapter 3.4 is a collection of some important theorems from empirical process theory, focusing on results used in Chapters 4-7. Most of the following is based on [19] Chapters 1-2 and [10] Chapters 6-8.

3.2 Outer Probability

Consider an arbitrary, not necessarily Borel-measurable subset A of $[0,1]$. What could be an appropriate "volume" of this set? One idea is to take the smallest Borel-measurable set A^* such that $A \subseteq A^*$ and assign the measure of A^* to A. Note that this is a well defined procedure, since such a set A^* exists by Corollary 3.2.2 below. An equally valid approach is to assign the measure of largest Borel-measurable set $A_* \subseteq A$ to A, which is also well defined by Corollary 3.2.2. If A itself is Borel-measurable, both approaches yield the same result. In empirical process theory, we usually deal with sequences of non-Borel-measurable quantities and we can exploit the just presented ideas. Luckily, the limit process is usually measurable and it turns out that for large n the empirical process is "almost" measurable. A rigorous formulation of both approaches is given in the following.

Lemma 3.2.1
Consider an arbitrary map $T : (\Omega, \mathcal{A}, P) \to \mathbb{R} \cup \{\pm\infty\} =: \overline{\mathbb{R}}$. There exist measurable maps T^* and T_* with the following properties:

1. *$T^* \geq T$ and $T_* \leq T$;*
2. *For every measurable $U \geq T$ almost surely, we have $T^* \leq U$ almost surely;*
3. *For every measurable $U \leq T$ almost surely, we have $T_* \geq U$ almost surely.*

T^* is called minimal measurable majorant and T_* is called maximal measurable minorant of T.

Note that a similar result also applies to arbitrary subsets of Ω.

Corollary 3.2.2
Let A be an arbitrary subset of Ω. Then there exist sets A_* and $A^* \in \mathcal{A}$ such that

1. *$A \subseteq A^*$ and $A_* \subseteq A$;*
2. *For all measurable B with $A \subseteq B$, we have $P(A^*) \leq P(B)$;*
3. *For all measurable B with $B \subseteq A$, we have $P(A_*) \geq P(B)$.*

$P(A^*) =: P^*(A)$ is called outer probability of A and $P(A_*) =: P_*(A)$ is called inner probability of A.

Having these tools at hand, we can define "expectation" for non-measurable quantities in a meaningful way.

Definition 3.2.3
Consider an arbitrary map $T : (\Omega, \mathcal{A}, P) \to \overline{\mathbb{R}}$. We define the outer expectation of T as

$$\mathbb{E}^*[T] := \inf\left\{\mathbb{E}[U] \mid U \geq T;\ \mathbb{E}[U] \ \text{exists}\right\} \tag{3.1}$$

and the inner expectation of T as

$$\mathbb{E}_*[T] := \sup\left\{\mathbb{E}[U] \mid U \leq T;\ \mathbb{E}[U] \ \text{exists}\right\}, \tag{3.2}$$

where $\inf \emptyset := \infty$ and $\sup \emptyset := -\infty$.

If T is measurable the inner and outer expectation of T are both equal to $\mathbb{E}[T]$, provided $\mathbb{E}[T]$ exists. The following relationships hold for the minimal measurable majorant and the maximal measurable minorant of an arbitrary map T.

Lemma 3.2.4
For any subset $A \subseteq \Omega$ and an arbitrary map $T : (\Omega, \mathcal{A}, P) \to \overline{\mathbb{R}}$, we have

1. *If $\mathbb{E}[T^*]$ exists, $\mathbb{E}^*[T] = \mathbb{E}[T^*]$.*
2. *If $\mathbb{E}[T_*]$ exists, $\mathbb{E}_*[T] = \mathbb{E}[T_*]$.*
3. *$P(A^*) = \mathbb{E}^*[\mathbf{1}_A] = P^*(A)$ and $P(A_*) = \mathbb{E}_*[\mathbf{1}_A] = P_*(A)$.*

Now, we specify the notion of asymptotic measurability. In Chapter 3.3, we will see that many weakly converging sequences are "almost" measurable in this sense.

Definition 3.2.5
Let $(\mathbb{D}, d)$ be a metric space. A sequence of maps $(T_n)_{n\in\mathbb{N}} : (\Omega, \mathcal{A}, P) \to (\mathbb{D}, d)$ is asymptotically measurable if for all $f \in C_b(\mathbb{D})$:

$$\lim_{n\to\infty} \mathbb{E}^*[f(T_n)] - \mathbb{E}_*[f(T_n)] = 0.$$

As already mentioned in Chapter 2, tightness of the limit process is often related to certain continuity properties. In fact, the tightness of the limit process can be deduced from the asymptotic tightness of the converging sequence, which is defined below.

Definition 3.2.6
Let $(\mathbb{D}, d)$ be a metric space. A sequence of maps $(T_n)_{n\in\mathbb{N}} : (\Omega, \mathcal{A}, P) \to (\mathbb{D}, d)$ is asymptotically tight if for every $\epsilon > 0$ there exists a compact set K such that $\liminf_{n\to\infty} P_\star (T_n \in G) \geq 1 - \epsilon$ for every open $G \supset K$.

Similar to measurable maps, we have a version of Chebyshev's inequality for outer probability.

Lemma 3.2.7
Let $\phi : [0, \infty) \to [0, \infty)$ be convex, non-decreasing and strictly positive on $(0, \infty)$. For an arbitrary map $T : (\Omega, \mathcal{A}, P) \to \overline{\mathbb{R}}$ we have the following outer probability version of Chebyshev's inequality

$$P^* (|T| > a) \leq \frac{\mathbb{E}^* [\phi(T)]}{\phi(a)}.$$

Moreover, there is a version of the Dominated Convergence Theorem.

Lemma 3.2.8
Let $(T_n)_{n\in\mathbb{N}}, T, S : (\Omega, \mathcal{A}, P) \to \overline{\mathbb{R}}$ be a arbitrary maps. If $|T_n - T|^* \overset{n\to\infty}{\to} 0$ almost surely and $|T_n| \leq S$ for all $n \in \mathbb{N}$ with $\mathbb{E}^* [S] < \infty$, the following version of the Dominated Convergence Theorem holds

$$\lim_{n\to\infty} \mathbb{E}^* [T_n] = \mathbb{E}^* [T] .$$

Since we frequently encounter random variables defined on product spaces, it is useful to have a version of Fubinis theorem for outer expectation. First, we have to clarify how repeated inner and outer expectations are defined. Consider $T : (\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2, P_1 \otimes P_2) \to \mathbb{R}$. For fixed ω_1, define $\mathbb{E}_2^*[T]$ as the outer expectation of the random variable $T(\omega_1, \cdot)$ on Ω_2. Now, $\mathbb{E}_2^*[T]$ is a map on Ω_1 and we define $\mathbb{E}_1^* [\mathbb{E}_2^* [T]] =: \mathbb{E}_1^*\mathbb{E}_2^* [T]$ as the outer expectation of the map $\mathbb{E}_2^*[T]$ on Ω_1. Repeated inner expectation is defined similarly.

Lemma 3.2.9
Let $T : (\Omega_1 \times \Omega_2, \mathcal{A}_1 \otimes \mathcal{A}_2, P_1 \otimes P_2) \to \mathbb{R}$ be an arbitrary map on a product probability space. Then we have the following outer version of Fubinis theorem

$$\mathbb{E}_* [T] \leq \mathbb{E}_{*,1}\mathbb{E}_{*,2} [T] \leq \mathbb{E}_1^*\mathbb{E}_2^* [T] \leq \mathbb{E}^* [T] .$$

This outer version of Fubinis theorem is the reason for the measurability assumptions appearing in Theorem 3.4.5, since we need to avoid disturbing inequalities when changing the order of expectations. Before introducing weak convergence, we present the definitions of two other modes of convergence of non-measurable maps.

Definition 3.2.10
Let $(\mathbb{D}, d)$ be a metric space and let $(T_n)_{n\in\mathbb{N}} : (\Omega, \mathcal{A}, P) \to (\mathbb{D}, d)$ arbitrary random maps with a measurable random map $T : (\Omega, \mathcal{A}, P) \to (\mathbb{D}, d)$

1. *We say that T_n converges to T in outer probability if for all $\epsilon > 0$ we have that $\lim_{n\to\infty} P^* (d(T_n, T) > \epsilon) = 0$.*
2. *We say that T_n converges to T outer almost surely if there exists a measurable sequence of random variables Δ_n such that $d(T_n, T) \leq \Delta_n$ and $P (\limsup_{n\to\infty} \Delta_n = 0) = 1$.*

If T_n is measurable, outer almost sure convergence and convergence in outer probability are equivalent to the usual definitions of almost sure convergence and convergence in probability for random variables.

Using the concept of outer probability, we are able to introduce a new notion of convergence in distribution of non-measurable quantities.

3.3 Weak Convergence

In this subsection, we introduce weak convergence for sequences of random maps Y_n, which are not necessarily measurable, but converge to a (Borel-)measurable limit Y.

Similarly to Definition 2.3.4, we give the following definition of weak convergence.

Definition 3.3.1
Let Y_n denote a sequence of maps from a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ to some metric space $(\mathbb{D}, d)$. We say that Y_n converges weakly to a Borel-measurable limit Y if for every $f \in C_b(\mathbb{D})$:

$$\lim_{n\to\infty} \mathbb{E}^* [f(Y_n)] = \mathbb{E} [f(Y)] . \tag{3.3}$$

Weak convergence of Y_n to Y is denoted as $Y_n \rightsquigarrow Y$.

If the sequence of random variables Y_n is measurable, Definition 3.3.1 is equivalent to convergence in distribution, defined in Defintion 2.3.4. Similarly to the Portmanteau-

Theorem for convergence in distribution, we have a version of the Portmanteau-Theorem for weak convergence.

Theorem 3.3.2
The following statements are equivalent:

1. $Y_n \rightsquigarrow Y$;
2. $\liminf_{n\to\infty} \mathbb{P}_*(Y_n \in G) \geq \mathbb{P}(Y \in G)$ *for every open* G;
3. $\limsup_{n\to\infty} \mathbb{P}^*(Y_n \in F) \leq \mathbb{P}(Y \in G)$ *for every closed* F;
4. $\lim_{n\to\infty} \mathbb{P}^*(Y_n \in B) = \lim_{n\to\infty} \mathbb{P}_*(Y_n \in B) = \mathbb{P}(Y \in B)$ *for any Borel measurable set* B *with* $\mathbb{P}(\overline{B} \setminus \mathring{B}) = 0$.

If there exists a measurable and separable set S *with* $\mathbb{P}(Y \in S) = 1$, *then* $1. - 4.$ *are also equivalent to*

5. $\lim_{n\to\infty} \sup_{f\in BL_1(\mathbb{D})} \left|\mathbb{E}^*\left[f(Y_n)\right] - \mathbb{E}\left[f(Y)\right]\right| = 0$, *where* $BL_1(\mathbb{D}) := \left\{f : \mathbb{D} \to \mathbb{R} \;\middle|\; |f(x) - f(y)| \leq d(x,y)\right\}$.

Note that the Portmanteau-Theorem can also be formulated in terms of the laws of Y_n and Y. Simply replace the expectations by integrals with respect to the laws of Y_n and Y and one obtains the weak convergence of the laws.

Remark 7
Later, we will see that the limit processes in $l^\infty(T)$ *often reside in a measurable and separable subset of* $l^\infty(T)$, *which implies that weak convergence is metrizable in the following sense:*
$Y_n \rightsquigarrow Y$ *is equivalent to* $\rho(Y_n, Y) := \sup_{f\in BL_1(\mathbb{D})} \left|\mathbb{E}^*\left[f(Y_n) - f(Y)\right]\right| \to 0$. *This "semi-metric" will be very useful to define the weak convergence of the bootstrapped empirical process.*

Moreover, there is an analogue of Slutsky's Lemma for weak convergence.

Theorem 3.3.3
Assume $Y_n \rightsquigarrow Y$ *and* $Z_n \rightsquigarrow c$, *where* Y *is separable and* c *is a constant. Then*

1. $(Y_n, Z_n) \rightsquigarrow (Y, c)$
2. $Y_n + Z_n \rightsquigarrow Y + c$

Before studying further properties of weak convergence, we investigate the properties of a Borel measurable Gaussian process Y in $l^\infty(T)$, which is the most frequently occuring limit process. First, we clarify the definition of a Gaussian Process indexed by some arbitrary set T.

Definition 3.3.4
A stochastic process $\mathbb{G}$ in $l^\infty(T)$ is Gaussian if for every $t_1, ..., t_k \in T$ the vector $(\mathbb{G}(t_1), ..., \mathbb{G}(t_k))$ has a multivariate Normal distribution.

The following lemma connects the tightness and the continuity of Borel measurable random elements in $l^\infty(T)$ and treats the special case of tight Gaussian processes (ref. [10] p. 106, [19] p. 39-41 and Theorem 2.1.12).

Lemma 3.3.5
Let Y be a Borel measurable random element in $l^\infty(T)$. Then the following statements are equivalent:

1. *Y is tight;*
2. *There exists a semimetric ρ making T totally bounded with $\mathbb{P}(Y \in UC(T, \rho)) = 1$.*

If Y is Gaussian, the following is also equivalent to 1. *and* 2.*:*

3. *For all $p \geq 1$ $\rho_p(s,t) := (\mathbb{E}[|Y(s) - Y(t)|^p])^{1/p}$ defines a semimetric on T, making T totally bounded with $\mathbb{P}(Y \in UC(T, \rho_p)) = 1$.*

Lemma 3.3.5 tells us that a tight Gaussian process in $l^\infty(T)$ has to be uniformly continuous w.r.t. ρ_p with probability 1. The next remark gives an alternative characterization of the metric ρ_2 in the empirical processes setting.

Remark 8
Assume that we are in the setting of an empirical process stemming from i.i.d. observations $(X_i)_{i\in\mathbb{N}}$ of some random variable X. Furthermore, assume that the limit process $\mathbb{G}$ is a mean zero and tight Gaussian process indexed by a set of functions $\mathcal{F}$. Then $\mathbb{G}$ has covariance function $\mathrm{Cov}(\mathbb{G}f, \mathbb{G}g) = \mathbb{E}[\mathbb{G}f\mathbb{G}g] = \mathbb{E}[(f(X) - \mathbb{E}[f(X)])(g(X) - \mathbb{E}[g(X)])]$. *This implies that $\rho_2(f,g) = (\mathbb{E}[|\mathbb{G}f - \mathbb{G}g|^2])^{\frac{1}{2}}$ is the square root of the second moment of $X_1 - X_2$, where (X_1, X_2) is a multivariate Normal random vector with mean zero and covariance matrix*

$$\Sigma = \begin{pmatrix} \mathrm{Var}(f(X)) & \mathbb{E}[\mathbb{G}f\mathbb{G}g] \\ \mathbb{E}[\mathbb{G}f\mathbb{G}g] & \mathrm{Var}(g(X)) \end{pmatrix}.$$

Thus, $X_1 - X_2 \sim \mathcal{N}\left(0, \mathbb{E}\left[\left(f(X) - \mathbb{E}\left[f(X)\right] - g(X) + \mathbb{E}\left[g(X)\right]\right)^2\right]\right)$*, since*

$$\begin{aligned}\mathrm{Var}(X_1 - X_2) &= \mathrm{Var}\left(f(X)\right) - 2\mathbb{E}\left[\left(f(X) - \mathbb{E}\left[f(X)\right]\right)\left(g(X) - \mathbb{E}\left[g(X)\right]\right)\right] + \mathrm{Var}\left(g(X)\right) \\ &= \mathbb{E}\left[\left(f(X) - \mathbb{E}\left[f(X)\right] - g(X) + \mathbb{E}\left[g(X)\right]\right)^2\right].\end{aligned}$$

Therefore, $\mathbb{G}$ *is uniformly continuous w.r.t.* $\rho_2(f,g) = \left(\mathbb{E}\left[\left(f(X) - \mathbb{E}\left[f(X)\right] - g(X) + \mathbb{E}\left[g(X)\right]\right)^2\right]\right)^{1/2}$ *and* $\mathcal{F}$ *is totally bounded w.r.t.* ρ_2*.*

This shows that we can always work with the semimetric ρ_2 *and the distribution of the random variable* X *if* $\mathcal{F}$ *is a Donsker-class (see Theorem 3.4.1).*

The properties of the weakly converging sequence $(Y_n)_{n\in\mathbb{N}}$ and the limit process Y are closely related, which is shown in the following lemma.

Lemma 3.3.6
Assume $Y_n \rightsquigarrow Y$*, then*

1. Y_n *is asymptotically measurable;*
2. Y_n *is asymptotically tight iff* Y *is tight.*

Lemma 3.3.6 shows that we automatically get the asymptotic measurability of $(Y_n)_{n\in\mathbb{N}}$ by the weak convergence of Y_n to Y. Further, the tightness of Y is equivalent to the asymptotic tightness of $(Y_n)_{n\in\mathbb{N}}$.

Recall that the tightness of $Y \in l^\infty(T)$ is related to certain continuity properties of the process Y by Lemma 3.3.5. We will see that also the sequence $(Y_n)_{n\in\mathbb{N}}$ fulfills certain asymptotic continuity properties known as asymptotic equicontinuity, defined in the following definition.

Definition 3.3.7
A sequence of processes $(Y_n)_{n\in\mathbb{N}}$ *in* $l^\infty(T)$ *is asymptotically uniformly* ρ*-equicontinuous in probability if*

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\mathbb{P}^*\left(\sup_{s,t\in T;\rho(s,t)<\delta}|Y_n(s) - Y_n(t)| > \epsilon\right) = 0.$$

Combining the results presented above, we are able to state some equivalent conditions to weak convergence in $l^\infty(T)$. These conditions are easier to verify than the conditions given in the Portmanteau-Theorem and relate the weak convergence of Y_n to a tight Y to the asymptotic equicontinuity of $(Y_n)_{n\in\mathbb{N}}$ (ref. [10] Thm 2.1/7.17).

Theorem 3.3.8
The following statements are equivalent:

1. *The sequence Y_n converges weakly to a tight limit $Y \in UC(T, \rho) \subseteq l^\infty(T)$, where ρ is a semimetric making T totally bounded.*
2. *Y_n is asymptotically tight and all finite dimensional marginals $(Y_n(t_1), ..., Y_n(t_k))$ converge weakly to the finite dimensional marginals of the process Y.*
3. *(i) All finite dimensional marginals $(Y_n(t_1), ..., Y_n(t_k))$ converge weakly to the finite dimensional marginals of the process Y.*
   (ii) Y_n is asymptotically uniformly ρ-equicontinuous w.r.t. some semimetric ρ making T totally bounded.

Remark 9
Let $\mathbb{G}_n \rightsquigarrow \mathbb{G}$, where $\mathbb{G}$ is a tight Gaussian process in $l^\infty(T)$. Then $\mathbb{G}_n$ is uniformly ρ_2-equicontinuous by an extension of Theorem 3.3.8. It essentially states that the existence of an arbitrary semimetric ρ, which makes T totally bounded with $\mathbb{P}(\mathbb{G} \in UC(T, \rho)) = 1$ is equivalent to asymptotic ρ_2-equicontinuity of Y_n.

3.4 Donsker Theorems

In this section, we formally introduce the empirical process and *Donsker*-classes. Moreover, we state two sufficient conditions for a class of functions $\mathcal{F}$ to be a *Donsker*-class. This section is based on [10] Section 8.4 and [19] Section 2.5.

First, recall the mathematical setting of empirical processes. $(X_i)_{i\in\mathbb{N}}$ are i.i.d. observations of some random variable $X \in \mathbb{R}^d$ defined on a probability space $(\Omega, \mathcal{A}, P)$. We are interested in the uniform limit behavior of $\sqrt{n}\,(n^{-1}\sum_{i=1}^n f(X_i) - \mathbb{E}\,[f(X)])$ over a class of functions $\mathcal{F} \subset \{f \mid f : \mathbb{R}^d \to \mathbb{R}\}$. In the setting of Theorem 3.1.1, $\mathcal{F}$ would correspond to the class of functions $\{\mathbf{1}_{\{\cdot\leq t\}} \mid t \in \mathbb{R}\}$. Let $\mathbb{P}_n := n^{-1}\sum_{i=1}^n \delta_{X_i}$ denote the empirical measure, where δ_{X_i} denotes the Dirac measure at X_i. We define the random operator $\mathbb{P}_n : \mathcal{F} \to \mathbb{R}$ as $\mathbb{P}_n f := \int f d\mathbb{P}_n = n^{-1}\sum_{i=1}^n f(X_i)$. Further, we define the operator $P : \mathcal{F} \to \mathbb{R}$ as $Pf := \int f dP$. Now, $\sqrt{n}\,(n^{-1}\sum_{i=1}^n f(X_i) - \mathbb{E}\,[f(X)])$ can be rewritten as $\sqrt{n}\,(\mathbb{P}_n f - Pf) = \sqrt{n}\,(\mathbb{P}_n - P)\,f =: \mathbb{G}_n f$. $\mathbb{G}_n$ is called the empirical process and can be viewed as a bounded stochastic process indexed by the functions $f \in \mathcal{F}$, which means that it is a stochastic process in $l^\infty(\mathcal{F})$. For fixed f and under some moment conditions on $f(X)$, we know that $\mathbb{G}_n f$ converges in distribution to a Normal distributed random variable with mean zero and variance $\mathrm{Var}\,(f(X))$. This property almost ensures the weak convergence of the sequence $\mathbb{G}_n$ and motivates the definition of a $P - Donsker$ class.

Definition 3.4.1
A class of measurable functions $\mathcal{F}$ with $\sup_{f\in\mathcal{F}} f(x) < \infty$ for all $x \in \mathbb{R}^d$ is called $P-$ Donsker (or a Donsker class) if $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ in $l^\infty(\mathcal{F})$, where $\mathbb{G}$ is a tight Gaussian process with mean zero and covariance function $\operatorname{Cov}(\mathbb{G}f, \mathbb{G}g) = \mathbb{E}[f(X)g(X)] - \mathbb{E}[f(X)]\,\mathbb{E}[g(X)]$.

Informally, a *Donsker*-class is a class of functions, where every projection $\mathbb{G}_n f$ converges to a mean zero Normal random variable, while the limit process $\mathbb{G}$ is continuous in $\mathcal{F}$.

Remark 10
Since the limiting process $\mathbb{G}$ of a Donsker class is tight, we know that $\mathbb{G}$ is in $UC(\mathcal{F}, \rho_2)$, by Remark 8.

Clearly, the condition that $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ over the class of functions $\mathcal{F}$ requires some bounds on the complexity of $\mathcal{F}$. We will see that this complexity of the class $\mathcal{F}$ can be measured in terms of entropy.

First, we define the concept of bracketing entropy. For any probability measure Q define $L_2(Q) := \{g \mid \int g^2 dQ < \infty\}$ and define the ball of radius ϵ around a function g in $L_2(Q)$ as $B_\epsilon(g) := \{f \mid \left(\int (f-g)^2 dQ\right)^{1/2} < \epsilon\}$.

Definition 3.4.2
Let $\epsilon > 0$. Assume that there exists $K \in \mathbb{N}$ such that for $1 \le i \le K$ we have $g_{i,1}, g_{i,2} \in L_2(P)$, $g_{i,1} \le g_{i,2}$ and $\left(\int (g_{i,1} - g_{i,2})^2 dP\right)^{1/2} < \epsilon$. Additionally, for any $f \in \mathcal{F}$ there exists $i \in \{1, ..., K\}$ such that $g_{i,1} \le f \le g_{i,2}$. The minimal K satisfying the requirements above is defined as the $L_2(P)$ bracketing number of $\mathcal{F}$ and is denoted as $N_{[]}(\epsilon, \mathcal{F}, L_2(P))$. Furthermore, $\log\left(N_{[]}(\epsilon, \mathcal{F}, L_2(P))\right)$ *is defined as the bracketing entropy of $\mathcal{F}$.*

Intuitively, the bracketing number is the minimal number of functions needed to put every $f \in \mathcal{F}$ between two functions from the bracketing cover. Note that the functions in the bracketing cover do not need to be elements of $\mathcal{F}$.

Another concept of entropy is that of uniform entropy. Let Q be a discrete probability measure, i.e. Q can be written as $Q = \sum_{i=1}^n \delta_{x_i}$ for some $(x_i)_{1\le i\le n} \in \mathbb{R}^d$.

Definition 3.4.3
Let $\epsilon > 0$ arbitrary.

1. *Consider a discrete probability measure Q and assume that there exists $K \in \mathbb{N}$ and $(g_i)_{1\le i\le K} \in L_2(Q)$ such that for every f in $\mathcal{F}$ there is some $i \in \{1, ..., K\}$ with $f \in B_\epsilon(g_i)$. The covering number of $\mathcal{F}$ in $L_2(Q)$ is the minimal number K satisfying the requirements above and is defined as $N(\epsilon, \mathcal{F}, L_2(Q))$.*

2. *Let Q be a discrete probability measure and define $\|F\|_Q := \left(\int sup_{f\in\mathcal{F}} f^2 dQ\right)^{1/2}$. The uniform covering number of $\mathcal{F}$ is defined as the supremum of the covering numbers over all discrete probability measures Q with $\|F\|_Q > 0$ weighted by $\|F\|_Q$. Mathematically, this translates to*

$$\sup_{Q \text{ is discrete; } \|F\|_Q>0} N(\epsilon\|F\|_Q, \mathcal{F}, L_2(Q)).$$

The uniform entropy of $\mathcal{F}$ is defined as

$$\log\left(\sup_{Q \text{ is discrete;}\|F\|_Q>0} N(\epsilon\|F\|_Q, \mathcal{F}, L_2(Q))\right).$$

The covering number of $\mathcal{F}$ is the minimal number of $L_2(Q)$ balls needed to cover $\mathcal{F}$. The uniform covering number is the maximum of all covering numbers of $\mathcal{F}$ over all discrete probability measures, with the radius weighted by the Q-expectation of a majorant of the class $\mathcal{F}$. Note that the uniform covering number is independent of the underlying probability measure P.

The last definition needed in order to state sufficient conditions for a class $\mathcal{F}$ to be a *Donsker*-class is that of $P-$ measurability. It is essentially required to avoid the disturbing inequalities in the outer version of Fubinis Theorem (Theorem 3.2.9).

Definition 3.4.4
A class of functions $\tilde{\mathcal{F}}$ is called a P-measurable class if for every $n \in \mathbb{N}$ and $(e_1, ..., e_n) \in \mathbb{R}^n$ the function

$$(X_1, ..., X_n) \mapsto \sup_{f\in\tilde{\mathcal{F}}} \sum_{i=1}^{n} e_i f(X_i)$$

is measurable on the completion of $\left(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), P_X^n\right)$.

The next theorem combines the results and definitions introduced in the previous chapters to state two sufficient conditions for the class $\mathcal{F}$ to be a *Donsker*-class.

Theorem 3.4.5
If $sup_{f\in\mathcal{F}} f(x) < \infty$ for $x \in \mathbb{R}^d$ and one of the following conditions is satisfied

1. $$\int_0^\infty \sqrt{log\left(N_{[]}(\epsilon, \mathcal{F}, L_2(P))\right)} d\epsilon < \infty;$$

2. *The classes of functions $(\mathcal{F})_\delta := \{f - g \mid \int (f-g)^2 dP < \delta\}$ and $\{h^2 \mid h \in (\mathcal{F})_\infty\}$*

are P-measurable for every $\delta \in (0, \infty)$ and $\mathbb{E}^* \left[\sup_{f \in \mathcal{F}} f(X)^2\right] < \infty$. Additionally,

$$\int_0^1 \sqrt{\log \left(\sup_{Q \text{ is discrete}; \|F\|_Q > 0} N(\epsilon \|F\|_Q, \mathcal{F}, L_2(Q)) \right)} d\epsilon < \infty;$$

then $\mathcal{F}$ is P-Donsker.

3.5 Bootstrapping The Empirical Process

In this subsection, we introduce the bootstrapped empirical process, which can be used to construct confidence intervals or to estimate the asymptotic variance of a test statistic. Luckily, if we already know that a class of functions is *P-Donsker*, most of the bootstrap results require only mild additional assumptions. This section is based on Section 3.6 in [19].

The main idea of bootstrapping is that a sample of the original sample should behave similarly to the original sample. If this is the case, one can create arbitrarily many samples of the original sample, each of them similar to the original sample. For example, one could estimate the variance of an observed statistic by computing the empirical variance of this statistic for a large number of bootstrap samples. However, the bootstrap samples are obviously dependent, because each sample is drawn from the original sample. Therefore, mathematically precise results can only be obtained asymptotically.
One possibility to create bootstrap samples is to draw with replacement from the original sample. This approach is known as Efron's Bootstrap and is incorporated in the bootstrap scheme introduced below. Note that this procedure is equivalent to re-weighting the initial sample with multinomial weights. When drawing weights for each observation in the original sample, there is no need to require these weights to be natural numbers. This idea leads to exchangeable bootstraps. The general approach is to assign identically distributed random weights to each observation in the original sample, which is formalized below.

Before introducing the exchangeable bootstrap, we have to modify the probability space to account for the additional randomness induced by the weights. Define $\mathbb{P}_X := \otimes_{i \in \mathbb{N}} P$ and assume that the observations $(X_i)_{i \in \mathbb{N}}$ only depend on the first and the weights $(\xi_{i,n})_{1 \leq i \leq n, n \in \mathbb{N}}$ only depend on the second coordinate of a product probability space $(\mathcal{X}^\infty, \mathcal{A}^\infty, \mathbb{P}_X) \otimes (\mathcal{Z}, \mathcal{W}, \mathbb{P}_W)$. Note that this structure implies that the weights are independent of the observations and that their joint distribution is described by the product measure $\mathbb{P}_{XW} := \mathbb{P}_X \otimes \mathbb{P}_W$. Furthermore, assume that the weights $(\xi_{i,n})_{1 \leq i \leq n, n \in \mathbb{N}}$ are non-negative and exchangeable, which is defined in the following definition.

Definition 3.5.1

A random vector $(\xi_{1,n}, ..., \xi_{n,n})$ *is exchangeable if for every* $n \in \mathbb{N}$ *and every permutation* π *of* $(1, ..., n)$ *the vector* $(\xi_{\pi(1),n}, ..., \xi_{\pi(n),n})$ *has the same distribution as* $(\xi_{1,n}, ..., \xi_{n,n})$.

We have to impose some conditions on the vector $(\xi_{1,n}, ..., \xi_{n,n})$ such that it can be considered as meaningful weights for a bootstrap scheme. In particular, define $\bar{\xi} = n^{-1} \sum_{i=1}^{n} \xi_i$ and assume

1. $$\sup_n \int_0^\infty \sqrt{P\left(|\xi_{1,n} - \bar{\xi}| > x\right)} dx < \infty$$

2. $$\frac{1}{\sqrt{n}} \mathbb{E}\left[\max_{1 \le i \le n} \left|\xi_{i,n} - \bar{\xi}\right|\right] \to 0$$

3. $$\sum_{i=1}^{n} \left(\xi_{i,n} - \bar{\xi}\right)^2 \to c^2 \text{ in } \mathbb{P}_W\text{-probability.}$$

For non-negative exchangeable weights satisfying conditions $1. - 3.$, we define the exchangeable bootstrap empirical measure as $\tilde{\mathbb{P}}_n := n^{-1} \sum_{i=1}^{n} \xi_{i,n} \delta_{X_i}$ and define the bootstrap empirical process as $\tilde{\mathbb{G}}_n := \sqrt{n} c^{-1} (\tilde{\mathbb{P}}_n - \bar{\xi} \mathbb{P}_n)$.

We want to define convergence of the bootstrapped empirical process conditioned on the original sample $(X_i)_{i \in \mathbb{N}}$. Therefore, recall the bounded Lipschitz metric defined in Remark 7, which tells us that weak convergence to a tight limit is metrizable in $l^\infty(\mathcal{F})$. To see this, note that tight limits reside in σ-compact subsets of $l^\infty(\mathcal{F})$ with probability 1. In a metric space σ-compact sets are separable, which allows us to use the bounded Lipschitz metric. Assume that we have a realization of the original sample $(X_i)_{i \in \mathbb{N}}$. If the bootstrapped sample behaves similarly as the original sample then $\tilde{\mathbb{P}}_n$ behaves similarly as $\mathbb{P}_n$. Since we know that $\mathbb{G}_n$ converges weakly to some tight Gaussian process $\mathbb{G}$, the same should be true for $\tilde{\mathbb{G}}_n$, given the original sample $(X_i)_{i \in \mathbb{N}}$. Intuitively, the law of the bootstrapped empirical process given $(X_i)_{i \in \mathbb{N}}$ converges to the "law" of the process $\mathbb{G}$ if $\mathbb{E}_{\mathbb{P}_W}\left[h(\tilde{\mathbb{G}}_n)\right] \to \mathbb{E}\left[h(\mathbb{G})\right]$ for all $h \in BL_1(l^\infty(\mathcal{F}))$, where $\mathbb{E}_{\mathbb{P}_W}$ denotes expectation w.r.t. the weights only. The following definitions formalize this reasoning.

Let $Y(\omega_1, \omega_2)$ be measurable w.r.t. $(\mathcal{Z}, \mathcal{W}, \mathbb{P}_W)$ for every fixed ω_1. We define $\mathbb{E}_{\mathbb{P}_W}\left[Y(\omega_1, \omega_2)\right]$ as the expectation of $Y(\omega_1, \cdot)$ w.r.t. $(\mathcal{Z}, \mathcal{W}, \mathbb{P}_W)$, treating the random quantity ω_1 as a constant. Using this formalism, we are able to define conditional weak convergence.

Definition 3.5.2

Let $\mathbb{G} \in l^\infty(\mathcal{F})$ be tight. We say that $\tilde{\mathbb{G}}_n$ converges to $\mathbb{G}$ conditionally on $(X_i)_{i\in\mathbb{N}}$ if

$$\sup_{h\in BL_1(l^\infty(\mathcal{F}))} \left| \mathbb{E}_{\mathbb{P}_W}\left[h(\tilde{\mathbb{G}}_n)\right] - \mathbb{E}[h(\mathbb{G})] \right| \to 0$$

in outer $\mathbb{P}_X$-probability.

Note that $h(\tilde{\mathbb{G}}_n)$ is measurable w.r.t. $(\mathcal{Z}, \mathcal{W}, \mathbb{P}_W)$ if we assume $(X_i)_{i\in\mathbb{N}}$ to be known, since it is just a composition of a continuous function h with the weighted sum of random variables $(\xi_{i,n})_{1\leq i\leq n, n\in\mathbb{N}}$.
Having this definition at hand, we are able to state sufficient conditions for $\tilde{\mathbb{G}}_n$ to converge to the same limit as $\mathbb{G}_n$, conditionally on the initial sample.

Theorem 3.5.3
Let $\mathcal{F}$ be a Donsker class and assume $\{f-g \mid \int(f-g)^2 dP < \delta\}$ is P-measurable for every $\delta > 0$. Additionally, assume that the weights $(\xi_{i,n})_{1\leq i\leq n, n\in\mathbb{N}}$ are non-negative, exchangeable and satisfy conditions $1.-3.$ Then

$$\sup_{h\in BL_1(l^\infty(\mathcal{F}))} \left| \mathbb{E}_{\mathbb{P}_W}\left[h(\tilde{\mathbb{G}}_n)\right] - \mathbb{E}[h(\mathbb{G})] \right| \to 0$$

in outer $\mathbb{P}_X$-probability. Furthermore, $\mathbb{E}_{\mathbb{P}_W}\left[h(\tilde{\mathbb{G}}_n)\right]_* - \mathbb{E}_{\mathbb{P}_W}\left[h(\tilde{\mathbb{G}}_n)\right]^* \to 0$ in outer $\mathbb{P}_X$-probability, i.e. $\mathbb{E}_{\mathbb{P}_W}\left[h(\tilde{\mathbb{G}}_n)\right]$ is asymptotically measurable w.r.t. $(\mathcal{X}^\infty, \mathcal{A}^\infty, \mathbb{P}_X)$.

Theorem 3.5.3 tells us that we obtain the convergence of the bootstrap whenever $\mathcal{F}$ is *Donsker* and suitably measurable.

Chapter 4

Clarke's Test

4.1 Clarke's And Vuong's Test

We shortly motivate Clarke's test statistic and define several important concepts.

Let $(\Omega, \mathcal{A}, P)$ be a probability space and $X \in \mathbb{R}^d$ be a random vector on this probability space. Assume that the vector X can be split into two subvectors Y and Z, where Z can be interpreted as a vector of covariates. Now, the conditional distribution of Y given Z induces a law $P_{Y|Z}$ on $\mathbb{R}^{d'}$, $d' \leq d$, whose density w.r.t. the Lebesgue measure is denoted by $p_{Y|Z}$.

Consider an i.i.d. sample $(X_i)_{i\in\mathbb{N}}$ from X. We want to find an appropriate probability density to describe the distribution of Y given Z. Assume we have a candidate probability density f. To determine the "closeness" of the density f to the true conditional density $p_{Y|Z}$, the Kullback-Leibler Information Criterion from [11] can be used. It measures the pseudo-distance of two densities of random vectors and is defined as follows:

Definition 4.1.1
The Kullback-Leibler Information Criterion (KLIC) of f w.r.t. to the true conditional density $p_{Y|Z}$ is defined as

$$K(f) := \mathbb{E}[\log(p_{Y|Z}(Y \mid Z))] - \mathbb{E}[\log(f(Y \mid Z))].$$

Note that $K(f) \geq 0$ and $K(f) = 0$ if and only if $p_{Y|Z} = f$ almost surely. Unfortunately, the $KLIC$ is not a true metric, since it is not symmetric and it does not satisfy the triangle inequality.

As the true density $p_{Y|Z}$ is unknown, $K(f)$ is unknown. However, it is possible to minimize

the KLIC by maximizing $\mathbb{E}[\log(f(Y \mid Z))]$. Suppose we have two candidates of parametric density families and we want to decide which family maximizes the KLIC. Let the first and the second family be denoted by $f_\alpha(\cdot \mid Z)_{\alpha\in\Theta_\alpha}$ and $g_\beta(\cdot \mid Z)_{\beta\in\Theta_\beta}$, respectively, where α takes values in a compact set $\Theta_\alpha \subset \mathbb{R}^{d_\alpha}$ and β takes values in a compact set $\Theta_\beta \subset \mathbb{R}^{d_\beta}$. Further, assume that there exist some unique pseudo-true values $\alpha^\star$ and $\beta^\star$, such that

$$\alpha^\star = \arg\max_{\alpha\in\Theta_\alpha} \mathbb{E}[\log(f_\alpha(Y \mid Z))] \quad \text{and} \quad \beta^\star = \arg\max_{\beta\in\Theta_\beta} \mathbb{E}[\log(g_\beta(Y \mid Z))].$$

Comparing these two families in terms of their KLIC, we say that the family of densities $f_\alpha(\cdot \mid Z)_{\alpha\in\Theta_\alpha}$ is closer to the true model than the family of densities $g_\beta(\cdot \mid Z)_{\beta\in\Theta_\beta}$ if $\mathbb{E}[\log(f_{\alpha^\star}(Y \mid Z))] > \mathbb{E}[\log(g_{\beta^\star}(Y \mid Z))]$ and vice versa, if both expectations exist. Two families $f_\alpha(\cdot \mid Z)_{\alpha\in\Theta_\alpha}$ and $g_\beta(\cdot \mid Z)_{\beta\in\Theta_\beta}$ are equally close to the true distribution if $\mathbb{E}[\log(f_{\alpha^\star}(Y \mid Z))] = \mathbb{E}[\log(g_{\beta^\star}(Y \mid Z))]$.

Since the quantities $\mathbb{E}[\log(f_{\alpha^\star}(Y \mid Z))]$ and $\mathbb{E}[\log(g_{\beta^\star}(Y \mid Z))]$ are unknown, they should be estimated. Therefore, let $\hat{\alpha}_n$ and $\hat{\beta}_n$ denote the pseudo-maximum likelihood estimators of $\alpha^\star$ and $\beta^\star$, which are defined as measurable functions of $(X_1, .., X_n)$ satisfying

$$\hat{\alpha}_n = \arg\max_{\alpha\in\Theta_\alpha} \sum_{i=1}^{n} \log(f_\alpha(Y_i \mid Z_i)) \quad \text{and} \quad \hat{\beta}_n = \arg\max_{\beta\in\Theta_\beta} \sum_{i=1}^{n} \log(g_\beta(Y_i \mid Z_i)).$$

Under some general conditions, see e.g. [21], one can show that these estimators are consistent and asymptotically Normal distributed.

Using the estimators $\hat{\alpha}_n$ and $\hat{\beta}_n$ as the estimated parameters of the competing families of densities, we are able to evaluate the random quantities $\left(\log(f_{\hat{\alpha}_n}(Y_i \mid Z_i))\right)_{1\le i\le n}$ and $\left(\log(g_{\hat{\beta}_n}(Y_i \mid Z_i))\right)_{1\le i\le n}$. In a more restrictive setting than [21] and under the assumption that $f_\alpha(\cdot \mid Z = z) \neq g_\beta(\cdot \mid Z = z)$ for all z and $(\alpha, \beta) \in \Theta_\alpha \times \Theta_\beta$ it was shown in [20] that

$$\frac{1}{n}\sum_{i=1}^{n} \log\left(f_{\hat{\alpha}_n}(Y_i|Z_i)\right) \xrightarrow{a.s.} \mathbb{E}\left[\log\left(f_{\alpha^\star}(Y \mid Z)\right)\right],$$

and

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{f_{\hat{\alpha}_n}(Y_i|Z_i)}{g_{\hat{\beta}_n}(Y_i|Z_i)}\right) - \mathbb{E}_P\left[\log\left(\frac{f_{\alpha^\star}(Y|Z)}{g_{\beta^\star}(Y|Z)}\right)\right]\right) \xrightarrow{d} \tilde{\mathcal{N}},$$

where $\tilde{\mathcal{N}}$ denotes a mean zero normally distributed random variable. Exploiting this result, one can deduce a hypothesis tests for model selection, which is known as Vuong's test. Its null hypothesis is given by

$$\mathcal{H}_0^V : \ \mathbb{E}\left[\log\left(f_{\alpha^\star}(Y \mid Z)\right) - \log\left(g_{\beta^\star}(Y \mid Z)\right)\right] = 0.$$

If two competing density families satisfy the condition $f_\alpha(\cdot \mid Z = z) \neq g_\beta(\cdot \mid Z = z)$ for all z and $(\alpha, \beta) \in \Theta_\alpha \times \Theta_\beta$ they are called non-nested density families.

Clarke introduced an alternative to Vuong's test in [6], assuming that the competing models are non-nested. Instead of comparing two models in terms of their $KLIC$, Clarke proposed a "distribution-free" alternative by considering the median of $\log\left(f_\alpha(Y \mid Z)\right) - \log\left(g_\beta(Y \mid Z)\right)$. Then, under Clarke's approach, a model $f_\alpha(\cdot \mid Z)_{\alpha\in\Theta_\alpha}$ is equally suited to describe the distribution of $Y \mid Z$ as a model $g_\beta(\cdot \mid Z)_{\beta\in\Theta_\beta}$ if the median of $\log\left(f_{\alpha^\star}(Y \mid Z)\right) - \log\left(g_{\beta^\star}(Y \mid Z)\right)$ is equal to $1/2$. This translates to the null hypothesis

$$\mathcal{H}_0 : \; P\left(\log\left(\frac{f(Y \mid Z, \alpha^\star)}{g(Y \mid Z, \beta^\star)}\right) > 0\right) = \frac{1}{2}.$$

For $\mathcal{H}_0$, Clarke proposed the test statistic

$$\hat{B}_n := \sum_{i=1}^{n} \mathbf{1}\left\{\log\left(\frac{f(Y_i \mid Z_i, \hat{\alpha}_n)}{g(Y_i \mid Z_i, \hat{\beta}_n)}\right) > 0\right\}$$

and conjectured that it is Binomial distributed. The theoretical counterpart

$$B_n := \sum_{i=1}^{n} \mathbf{1}\left\{\log\left(\frac{f(Y_i \mid Z_i, \alpha^\star)}{g(Y_i \mid Z_i, \beta^\star)}\right) > 0\right\}$$

of $\hat{B}_n$ is indeed Binomial distributed. However, as we will see in Section 4.3, $\hat{B}_n$ is generally not Binomial distributed due to the additional randomness induced by the estimators $\hat{\alpha}_n$ and $\hat{\beta}_n$. Unfortunately, this method has been extensively applied in many applied works and in many different fields: finance, economics, accounting, political science, etc: see [2], [8], [13], [12], among others.

It is worth noting that, in general, Vuong's null hypothesis $\mathcal{H}_0^V$ does not imply Clarke's null hypothesis $\mathcal{H}_0$ and vice versa. A more detailed discussion of the similarities and differences between Vuong's test and Clarke's test is available in Chapter 6. The main purpose of this master thesis is to state the correct asymptotic distribution of $\hat{B}_n$ in a slightly modified setting compared to [6] and [20].

4.2 Theoretical Framework

In this section, we introduce the mathematical setup of Clarke's test statistic $\hat{B}_n$ and define several abbreviations, which are frequently used throughout the master thesis.

Recall that the observations $(X_i)_{i\in\mathbb{N}}$ are independent copies of a random vector $X \in \mathbb{R}^d$,

which is distributed according to some probability measure P. The measure $\mathbb{P}$ will refer to the joint law of $(X_i)_{i\in\mathbb{N}}$. In contrast to the papers [20] and [6], we do not assume that our observations are conditioned on some random vector Z, to simplify our theoretical developments. In the framework of [20], this would correspond to the case of no covariates. An extension of our results to the case of a covariate vector Z, whose law is identical under the potential models $f(\cdot,\alpha)$ and $g(\cdot,\beta)$, is straightforward. A proof of this statement can be found in Appendix A.1. Moreover, in this master thesis, we only consider strictly non-nested models, which is needed to exclude the possibility of $f_{\alpha^\star}(X)/g_{\beta^\star}(X)$ having an atom at 1. The definition is as follows:

Definition 4.2.1
A couple of parametric density families $(f_\alpha)_{\alpha\in\Theta_\alpha}$ and $(g_\beta)_{\beta\in\Theta_\beta}$ is strictly non-nested on $\overline{\Theta}_\alpha\times\overline{\Theta}_\beta\subseteq\Theta_\alpha\times\Theta_\beta$ if $f_\alpha(X)\neq g_\beta(X)$ a.s. for all $\alpha\in\overline{\Theta}_\alpha$ and $\beta\in\overline{\Theta}_\beta$.

Note that this definition is stronger than the definition of non-nested models in [20], since we do not allow that two rival models f_α and g_β coincide on a set with positive probability. This case is not excluded in [20], who required that the functions $f_\alpha(\cdot\mid Z=z)$ and $g_\beta(\cdot\mid Z=z)$ are not equal for all z and $(\alpha,\beta)\in\Theta_\alpha\times\Theta_\beta$ ("strictly non-nested" models, Definition 2). However, in most applications, it is rarely the case that two competing models do not satisfy Definition 4.2.1 but would satisfy the strict non-nestedness Definiton 2 in [20].

Our competing families of densities are $(f(\cdot,\alpha))_{\alpha\in\Theta_\alpha}$ and $(g(\cdot,\beta))_{\beta\in\Theta_\beta}$, where $\Theta_\alpha\subset\mathbb{R}^{d_\alpha}$ and $\Theta_\beta\subset\mathbb{R}^{d_\beta}$ are compact sets with possibly different dimensions. Furthermore, for each $(\alpha,\beta)\in\Theta_\alpha\times\Theta_\beta$ the functions $f(\cdot,\alpha)$ and $g(\cdot,\beta)$ are probability densities on $\mathbb{R}^d$, which are almost surely positive on the support of X. This is not really a lack of generality in practice. Indeed, when some realizations of X do not belong to the support of $f(\cdot,\alpha)$ and/or $g(\cdot,\beta)$, it would not be realistic to select one or both of the latter candidate models. Further, it is important to note that the latter requirement does not imply $f(\cdot,\alpha)$ and $g(\cdot,\beta)$ to have the same support. It is still possible that $f(x,\alpha)=0$ and $g(x,\beta)>0$ for all x in an open subset of $\mathbb{R}^d$. This event just occurs with probability 0 under P. Due to the reasoning above, it is safer and simpler to assume that both families are almost surely positive on the support of X. We always assume that the pseudo-true values $\alpha^\star$ and $\beta^\star$, as defined in [21], belong to the interior of Θ_α and Θ_β, respectively.

For any $\delta>0$, denote $E_\delta:=[\alpha_1^\star-\delta,\alpha_1^\star+\delta]\times\ldots\times[\alpha_{d_\alpha}^\star-\delta,\alpha_{d_\alpha}^\star+\delta]\times[\beta_1^\star-\delta,\beta_1^\star+\delta]\times\ldots\times[\beta_{d_\beta}^\star-\delta,\beta_{d_\beta}^\star+\delta]$ and

$$\psi(x,\alpha,\beta):=\log\left(\frac{f(x,\alpha)}{g(x,\beta)}\right).$$

Moreover, for $\gamma > 0$ and $E_\gamma \subseteq \Theta_\alpha \times \Theta_\beta$ we denote $\mathcal{F}_\gamma := \{\mathbf{1}_{\{\psi(\cdot,\alpha,\beta)>0\}} \mid (\alpha,\beta) \in E_\gamma\}$. From the assumption that $(\alpha^\star, \beta^\star)$ is an interior point of $\Theta_\alpha \times \Theta_\beta$ we deduce that such a $\gamma > 0$ always exists, which implies that $\mathcal{F}_\gamma$ is well defined.

We recall the Clarke's statistic as

$$\hat{B}_n := \sum_{i=1}^{n} \mathbf{1}\left\{ \log\left(\frac{f(X_i, \hat{\alpha}_n)}{g(X_i, \hat{\beta}_n)}\right) > 0 \right\}, \tag{4.1}$$

where $\hat{\alpha}_n$ and $\hat{\beta}_n$ are estimators of the pseudo-true values $\alpha^\star$ and $\beta^\star$. Additionally, the "unfeasible" Binomial distributed statistic is

$$B_n := \sum_{i=1}^{n} \mathbf{1}\left\{ \log\left(\frac{f(X_i, \alpha^\star)}{g(X_i, \beta^\star)}\right) > 0 \right\}.$$

From now on, if not mentioned otherwise, we always work under the null hypothesis of Clarke's test

$$\mathcal{H}_0 : \; P\left(\log\left(\frac{f(X, \alpha^\star)}{g(X, \beta^\star)}\right) > 0\right) = \frac{1}{2}. \tag{4.2}$$

Since f_α and g_β do not need to have the same support, it can occur that $\log(f(x,\alpha)/g(x,\beta))$ is not well defined when one or both densities are zero. In such cases (these events occur with probability zero), formally set $\log(f(x,\alpha)/0) := +\infty$, $\log(0/g(x,\beta)) := -\infty$ and $\log(0/0) := 0$.

4.3 Non Binomial Distribution Of Clarke's Test Statistic

Now, we present a short counterexample to prove that $\hat{B}_n$ is generally not Binomial distributed. In other words, the conclusions drawn in Chapter 2.2 of [6] are incorrect. To overcome this problem, we will propose two alternative asymptotically standard Normal distributed test statistics in (5.5) and (7.3) hereafter.

Let the true distribution P follow a Normal distribution with mean μ_0 and fixed variance 1. We compare the density of a Normal distribution with fixed variance σ_f^2 and the density of a Normal distribution with fixed variance σ_g^2, while estimating the mean for both of these families. This translates to the families of densities $f(\cdot,\alpha) = (2\pi\sigma_f^2)^{-1/2}\exp\left(-(\cdot-\alpha)^2/2\sigma_f^2\right)$ and $g(\cdot,\beta) = (2\pi\sigma_g^2)^{-1/2}\exp\left(-(\cdot-\beta)^2/2\sigma_g^2\right)$. Note that the two models are strictly non-nested according to Definition 4.2.1.

In the following, we will choose $\sigma_f \neq \sigma_g$ not equal to one, such that the null hypothesis of Clarke's test is satisfied. Before fixing σ_f and σ_g, we need to find the pseudo-true values $\alpha^\star$ and $\beta^\star$. First, we show that $\alpha^\star$ and $\beta^\star$ are both equal to μ_0.

$$\begin{aligned}
\mathbb{E}\left[\log\big(f(X,\alpha)\big)\right] &= \mathbb{E}\left[-\log(\sqrt{2\pi\sigma_f^2}) - \frac{(X-\alpha)^2}{2\sigma_f^2}\right] \\
&= -\log\left(\sqrt{2\pi\sigma_f^2}\right) - \frac{1}{2\sigma_f^2}\Big(\mathrm{Var}(X) + 2\mathbb{E}\left[(X-\mathbb{E}\left[X\right])(\mathbb{E}\left[X\right]-\alpha)\right] \\
&\quad + (\mathbb{E}\left[X\right]-\alpha)^2\Big) \\
&= -\log\left(\sqrt{2\pi\sigma_f^2}\right) - \frac{1}{2\sigma_f^2}\,\mathrm{Var}(X) - \frac{1}{2\sigma_f^2}(\mu_0-\alpha)^2 \\
&\leq -\log\left(\sqrt{2\pi\sigma_f^2}\right) - \frac{1}{2\sigma_f^2}\,\mathrm{Var}(X),
\end{aligned}$$

where the last inequality is an equality if we choose $\alpha = \mu_0$. Thus, μ_0 maximizes $\alpha \mapsto \mathbb{E}\left[\log\big(f(X,\alpha)\big)\right]$, i.e the pseudo-true value $\alpha^\star$ is equal to μ_0. Because of the symmetry w.r.t. σ_g, $\beta^\star$ is also equal to μ_0.

In the next step, we need to compute the estimators of the pseudo-true values $\alpha^\star$ and $\beta^\star$ in the case of known variances. We find the pseudo maximum likelihood estimators for α and β by solving

$$\hat{\beta}_n = \hat{\alpha}_n = \arg\max_{\mu\in\mathbb{R}} \log\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_f^2}} \exp\Big(\frac{-(X_i-\mu)^2}{2\sigma_f^2}\Big)\right).$$

This is obviously the usual ML estimator for the mean of a Normal distribution, i.e. $\hat{\beta}_n = \hat{\alpha}_n = \bar{X}$.

Now, we choose σ_f and σ_g such that the null hypothesis of Clarke's test is satisfied. Thus, let us calculate the probability of the set

$$\begin{aligned}
\left\{\log\Big(\frac{f(X,\alpha)}{g(X,\beta)}\Big) > 0\right\} &= \left\{\log\Big(\frac{\sigma_g}{\sigma_f}\Big) - \frac{(X-\alpha)^2}{2\sigma_f^2} + \frac{(X-\beta)^2}{2\sigma_g^2} > 0\right\} \\
&= \left\{\frac{(X-\beta)^2\sigma_f^2 - (X-\alpha)^2\sigma_g^2}{2\sigma_f^2\sigma_g^2} > \log\left(\frac{\sigma_f}{\sigma_g}\right)\right\}.
\end{aligned}$$

Replacing α and β with their pseudo-true value μ_0, and assuming w.l.o.g. that $\sigma_g < \sigma_f$, we get

$$\left\{\frac{(X-\beta)^2\sigma_f^2 - (X-\alpha)^2\sigma_g^2}{2\sigma_f^2\sigma_g^2} > \log\left(\frac{\sigma_f}{\sigma_g}\right)\right\}$$

$$= \left\{ (X-\mu_0)^2 > \log\left(\frac{\sigma_f}{\sigma_g}\right) \frac{2\sigma_f^2\sigma_g^2}{\sigma_f^2-\sigma_g^2} \right\}.$$

Since $(X-\mu_0)$ is standard Normal distributed, we calculate

$$\begin{aligned} P\Big(\Big\{\log\Big(\frac{f(X,\alpha)}{g(X,\beta)}\Big) > 0\Big\}\Big) &= \Phi\left(-\sqrt{\log\left(\frac{\sigma_f}{\sigma_g}\right)\frac{2\sigma_f^2\sigma_g^2}{\sigma_f^2-\sigma_g^2}}\right) \\ &\quad + (1-\Phi)\left(\sqrt{\log\left(\frac{\sigma_f}{\sigma_g}\right)\frac{2\sigma_f^2\sigma_g^2}{\sigma_f^2-\sigma_g^2}}\right) \\ &= 2\Phi\left(-\sqrt{\log\left(\frac{\sigma_f}{\sigma_g}\right)\frac{2\sigma_f^2\sigma_g^2}{\sigma_f^2-\sigma_g^2}}\right). \end{aligned}$$

E.g. setting $\sigma_g = 1/2$, it is easy to see that the map $\sigma_f \mapsto \left\{ \log(2\sigma_f)\sigma_f^2/\ (2\sigma_f^2 - 0.5)\right\}^{1/2}$ attains all values in $\mathbb{R}_+$. Therefore, we can find $\overline{\sigma}_f$ such that $-\big(\log(2\overline{\sigma}_f)\overline{\sigma}_f^2/\ (2\overline{\sigma}_f^2 - 0.5)\big)^{1/2} = u_{0.25}$, where $u_{0.25}$ is the 0.25 quantile of the standard Normal distribution and $\mathcal{H}_0$ from (4.2) is satisfied.

In the case $\sigma_g = 1/2$, we get an approximated value of $\overline{\sigma}_f \approx 0.98$. Figure 4.1 shows the curve of pairs (σ_g, σ_f), such that the null-hypothesis is satisfied.

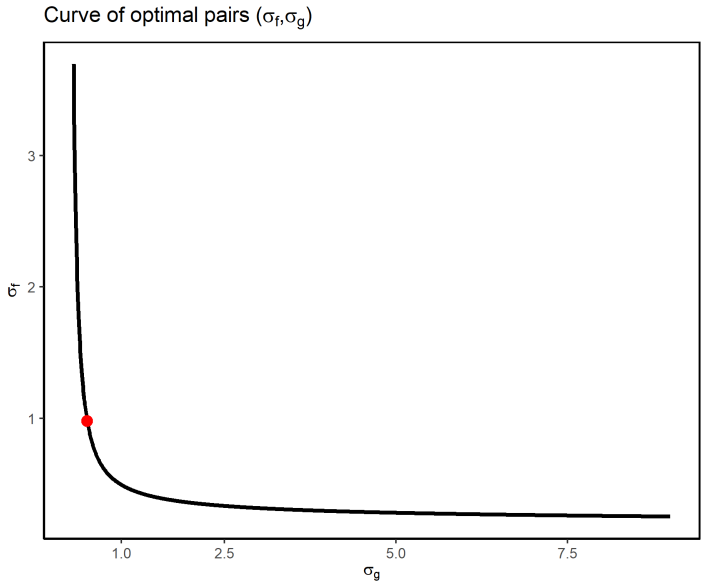

Figure 4.1: Curve of pairs (σ_g, σ_f) satisfying $\mathcal{H}_0$ (Normal laws). The red dot displays the point $(1/2, 0.98)$.

From now on, consider that we have chosen $\sigma_f \neq 1$, $\sigma_g \neq 1$ and $\sigma_f > \sigma_g$ such that

$$P\left(\left\{(X-\mu_0)^2 > \log\left(\frac{\sigma_f}{\sigma_g}\right)\frac{2\sigma_f^2\sigma_g^2}{\sigma_f^2-\sigma_g^2}\right\}\right) = \frac{1}{2}.$$

Therefore, the null hypothesis of Clarke's test is satisfied for the chosen values of σ_f and σ_g. Note that one would clearly prefer the family $f(\cdot,\alpha)$ with variance $\overline{\sigma}_f \approx 0.98$ over the family $g(\cdot,\beta)$ with variance $\overline{\sigma}_g = 1/2$, but Clarke's test considers them as equivalently suited to approximate a standard Normal distribution. A thorough discussion of this issue can be found in Chapter 6 and in the discussion of model comparison (7.2) in Section 7.3.

Clarke states in [6] that the statistic $\hat{B}_n$ is Binomial distributed with parameter $p = 0.5$. We show that $\hat{B}_n$ is generally not even Binomial distributed for any $p \in [0,1]$. For $n = 2$ the test statistic $\hat{B}_2$ takes the form:

$$\mathbf{1}\left\{\log\left(\frac{f(X_1,\hat{\alpha}_2)}{g(X_1,\hat{\beta}_2)}\right)) > 0\right\} + \mathbf{1}\left\{\log\left(\frac{f(X_2,\hat{\alpha}_2)}{g(X_2,\hat{\beta}_2)}\right) > 0\right\},$$

which is equal to

$$\mathbf{1}\left\{(X_1-\bar{X})^2 > \log\left(\frac{\sigma_f}{\sigma_g}\right)\frac{2\sigma_f^2\sigma_g^2}{\sigma_f^2-\sigma_g^2}\right\} + \mathbf{1}\left\{(X_2-\bar{X})^2 > \log\left(\frac{\sigma_f}{\sigma_g}\right)\frac{2\sigma_f^2\sigma_g^2}{\sigma_f^2-\sigma_g^2}\right\}$$

by the previous calculations and the fact that $\hat{\alpha}_2 = \hat{\beta}_2 = \bar{X}$. Noting that $\bar{X} = X_1/2 + X_2/2$, we get

$$(X_1-\bar{X})^2 = \left(\frac{X_1}{2}-\frac{X_2}{2}\right)^2 = \left(-\frac{X_1}{2}+\frac{X_2}{2}\right)^2 = (X_2-\bar{X})^2.$$

Therefore, this yields

$$B_2 = 2 \times \mathbf{1}\left\{\left(\frac{X_1}{2}-\frac{X_2}{2}\right)^2 > \log\left(\frac{\sigma_f}{\sigma_g}\right)\frac{2\sigma_f^2\sigma_g^2}{\sigma_f^2-\sigma_g^2}\right\},$$

which takes values in $\{0,2\}$ and is clearly not Binomial distributed.

A remaining question is whether $\hat{B}_n$ may be asymptotically Binomial distributed. The answer to this question is given in Corollary 5.0.2, which describes the asymptotic distance between B_n and $\hat{B}_n$.

Chapter 5

Asymptotic Normality of $\hat{\mathbf{B}}_\mathbf{n}$

In this section, we derive the asymptotic distribution of $n^{-1/2}(\hat{B}_n - n/2)$ under various assumptions and we propose two estimators for its asymptotic variance. Additionally, we will answer the question whether $\hat{B}_n$ is asymptotically Binomial distributed in Corollary 5.0.2. Note that, if not explicitly stated otherwise (e.g. as in Theorem 5.0.1), we will always assume that the null hypothesis of Clarke's test is satisfied.

Following the notation in [19], we denote $Pf = \int_\Omega f(X(\omega))dP(\omega) = \mathbb{E}[f(X)]$ for some measurable function $f : \mathbb{R}^d \mapsto \mathbb{R}$. The empirical measure associated with the sequence of random vectors $(X_i)_{i\in\mathbb{N}}$ is defined as $\mathbb{P}_n := n^{-1}\sum_{i=1}^n \delta_{X_i}$, where δ_{X_i} is the Dirac measure at X_i. Similarly to Pf, define $\mathbb{P}_n f := n^{-1}\sum_{i=1}^n f(X_i)$. To prove the weak convergence of Clarke's test statistic, we frequently use the expression $\mathbb{G}_n f := \sqrt{n}(\mathbb{P}_n - P)f$. We use $\rightsquigarrow$ to denote weak convergence in $l^\infty(\mathcal{F}_\gamma) := \{h \mid \ h : \mathcal{F}_\gamma \mapsto \mathbb{R}; \sup_{f\in\mathcal{F}_\gamma} |h(f)| < \infty\}$ equipped with the supremum norm $d(h,k) := \sup_{f\in\mathcal{F}_\gamma} |h(f) - k(f)|$ and refer to outer probabilities and outer expectations, as defined in Chapter 1 of [19]. Furthermore, denote by $u \cdot v$ the Euclidean scalar product of two vectors u and v and denote $\phi_n := \mathbf{1}\{\psi(X, \hat{\alpha}_n, \hat{\beta}_n) > 0\}$ as well as $\phi_\star := \mathbf{1}\{\psi(X, \alpha^\star, \beta^\star) > 0\}$. The Normal distribution with mean μ and variance σ^2 is denoted as $\mathcal{N}(\mu, \sigma^2)$.

We need the following assumptions to prove the asymptotic Normality of a modified Clarke's test statistic.

Assumptions

B1 *The pseudo-true values $\alpha^\star$ and $\beta^\star$ exist and are unique.*

B2 *There exists $\gamma > 0$ such that, for any $x \in Range(X)$, the function $\psi(x, \cdot, \cdot)$ is continuous on E_γ.*

B3 *There exists $\gamma > 0$ such that the models (f_α) and (g_β) are strictly non-nested on E_γ.*

B4 *There exists $\gamma > 0$ such that the function*

$$h: \ E_\gamma \to [0,1] \ ; \ (\alpha,\beta) \mapsto \int \mathbf{1}_{\{\psi(x,\alpha,\beta)>0\}} dP(x) = P(\psi(X,\alpha,\beta) > 0)$$

is continuously differentiable at $(\alpha^\star, \beta^\star)$. Denote $h_1(\alpha,\beta) := \left(\frac{\partial}{\partial\alpha_1}h(\alpha,\beta), \ldots, \frac{\partial}{\partial\alpha_{d_\alpha}}h(\alpha,\beta)\right)^\intercal$ and $h_2(\alpha,\beta) := \left(\frac{\partial}{\partial\beta_1}h(\alpha,\beta), \ldots, \frac{\partial}{\partial\beta_{d_\beta}}h(\alpha,\beta)\right)^\intercal$ the column vectors of the partial derivatives of h w.r.t. α and β.

B5 *The (measurable) estimators $\hat{\alpha}_n$ and $\hat{\beta}_n$ can be written as $\hat{\alpha}_n = \mathbb{P}_n s_1 + o_{\mathbb{P}}(1)$ and $\hat{\beta}_n = \mathbb{P}_n s_2 + o_{\mathbb{P}}(1)$ for some measurable functions s_1 and s_2 with $\mathbb{E}[s_1(X)^2] < \infty$ and $\mathbb{E}[s_2(X)^2] < \infty$. Additionally, the estimators are strongly consistent and it is possible to write $\sqrt{n}(\hat{\alpha}_n - \alpha^\star) = \mathbb{G}_n(s_1) + o_{\mathbb{P}}(1)$ and $\sqrt{n}(\hat{\beta}_n - \beta^\star) = \mathbb{G}_n(s_2) + o_{\mathbb{P}}(1)$.*

Remark 11

1. *B5 is true for many M- and Z-estimators by Theorem 2.10/2.12 in [10].*
2. *Assume that the true density p_X is continuous, X is real valued and that for all $(\alpha,\beta) \in E_\gamma$ the number of zeros of the function $\psi(\cdot,\alpha,\beta)$ is bounded by a universal constant. Additionally, assume that every zero of ψ can be represented as a continuously differentiable function (w.r.t. α and β). Then Assumption B4 is satisfied, since $h(\alpha^\star,\beta^\star)$ can be written as a finite sum of terms of the form $P(X \le x_i)$ or $1 - P(X \le x_i)$, where $x_i = x_i(\alpha^\star,\beta^\star)$ are the continuously differentiable zeros of $\psi(\cdot,\alpha^\star,\beta^\star)$.*

Using Assumption $B1 - B5$, we are ready to state the main theorem of this master thesis. The first part of Theorem 5.0.1 shows that $n^{-1/2}(\hat{B}_n - n/2)$ is asymptotically Normal distributed under $\mathcal{H}_0$. Parts *(ii)* and *(iii)* handle the two possible cases under the alternative.

Theorem 5.0.1
Assume that there exists $\gamma > 0$ such that $\mathcal{F}_\gamma$ is $P-$Donsker and Assumptions $B1 - B5$ are satisfied. Then the following statements are valid.

(i) *Under $\mathcal{H}_0$, i.e. if $P(\psi(X,\alpha^\star,\beta^\star) > 0) = \frac{1}{2}$, we have*

$$\frac{1}{\sqrt{n}}\left(\hat{B}_n - \frac{n}{2}\right) \rightsquigarrow \mathcal{N}(0,\sigma_\psi^2), \tag{5.1}$$

where $\sigma_\psi^2 = \mathrm{Var}\left(\mathbf{1}\{\psi(X,\alpha^\star,\beta^\star) > 0\} + h_1(\alpha^\star,\beta^\star)\cdot s_1(X) + h_2(\alpha^\star,\beta^\star)\cdot s_2(X)\right)$.

(ii) *If $P(\psi(X,\alpha^\star,\beta^\star) > 0) < 1/2$, then $n^{-1/2}(\hat{B}_n - n/2) \to -\infty$ $\mathbb{P}$-almost surely.*

(iii) If $P(\psi(X, \alpha^\star, \beta^\star) > 0) > 1/2$, *then* $n^{-1/2}(\hat{B}_n - n/2) \to +\infty$ $\mathbb{P}$*-almost surely.*

The proof of the Theorem has been postponed to Appendix A.2.

From the proof of Theorem 5.0.1, the following corollary can be deduced. It provides information about the asymptotic distance between $\hat{B}_n$ and the unknown statistic B_n.

Corollary 5.0.2
Under the assumptions of Theorem 5.0.1

$$\frac{1}{\sqrt{n}}\left(\hat{B}_n - B_n\right) \rightsquigarrow \mathcal{N}(0, \sigma_h^2),$$

where $\sigma_h^2 = \mathrm{Var}\left(h_1(\alpha^\star, \beta^\star) \cdot s_1(X) + h_2(\alpha^\star, \beta^\star) \cdot s_2(X)\right)$.

Corollary 5.0.2 tells us that the difference between $\hat{B}_n$ and B_n goes to infinity if $h_1(\alpha^\star, \beta^\star) \cdot s_1(X) + h_2(\alpha^\star, \beta^\star) \cdot s_2(X)$ is not equal to 0. This shows that $\hat{B}_n$ is not asymptotically Binomial distributed in general. Furthermore, it is interesting to note that Corollary 5.0.2 is valid independent of the value of $P(\psi(X, \alpha^\star, \beta^\star) > 0)$. This means that $\mathcal{H}_0$ does not have to be satisfied for Corollary 5.0.2 to be true.

Hereafter, we provide a more explicit expression of the variance σ_h^2. Define the matrices $A_f(\alpha)$ and $B_f(\alpha)$ with entries

$$A_f(\alpha)_{i,j} := \mathbb{E}\left[\frac{\partial^2 \log(f(X, \alpha))}{\partial \alpha_i \partial \alpha_j}\right]$$

and

$$B_f(\alpha)_{i,j} := \mathbb{E}\left[\frac{\partial \log(f(X, \alpha))}{\partial \alpha_i}\frac{\partial \log(f(X, \alpha))}{\partial \alpha_j}\right], \quad 1 \le i, j \le d_\alpha.$$

The matrices $A_g(\beta), B_g(\beta)$ are defined similarly. Furthermore, for $1 \le i \le d_\alpha$ and $1 \le j \le d_\beta$, define the matrix $B_{f,g}(\alpha, \beta)$ with entries

$$B_{f,g}(\alpha, \beta)_{i,j} := \mathbb{E}\left[\frac{\partial \log(f(X, \alpha))}{\partial \alpha_i}\frac{\partial \log(g(X, \beta))}{\partial \beta_j}\right].$$

In [20], it is shown that

$$\sqrt{n}\begin{pmatrix}\hat{\alpha}_n - \alpha^\star \\ \hat{\beta}_n - \beta^\star\end{pmatrix} \rightsquigarrow N\big(0, \Sigma(\alpha^\star, \beta^\star)\big),$$

where

$$\Sigma(\alpha^\star, \beta^\star) = \begin{pmatrix} A_f^{-1}(\alpha^\star)^{-1} B_f(\alpha^\star) A_f^{-1}(\alpha^\star); & A_f^{-1}(\alpha^\star)^{-1} B_{f,g}(\alpha^\star, \beta^\star) A_g^{-1}(\beta^\star) \\ A_g^{-1}(\beta^\star)^{-1} B_{g,f}(\alpha^\star, \beta^\star) A_f^{-1}(\alpha^\star); & A_g^{-1}(\beta^\star)^{-1} B_g(\beta^\star) A_g^{-1}(\beta^\star) \end{pmatrix}$$
$$:= \begin{pmatrix} \Sigma_f & \Sigma_{f,g} \\ \Sigma_{g,f} & \Sigma_g \end{pmatrix}.$$

We use the just stated results to express the asymptotic variance of Corollary 5.0.2.

Proposition 5.0.3
If we assume all assumptions of Theorem 5.0.1, Assumptions $A1 - A5$ from [20] and that $n\left(\hat{\alpha}_n - \alpha^\star\right)^2$ and $n(\hat{\beta}_n - \beta^\star)^2$ are uniformly integrable, the variance σ_h^2 simplifies to

$$\begin{aligned} \sigma_h^2 = {} & h_1(\alpha^\star, \beta^\star)^\intercal \Sigma_f(\alpha^\star, \beta^\star) h_1(\alpha^\star, \beta^\star) + h_2(\alpha^\star, \beta^\star)^\intercal \Sigma_g(\alpha^\star, \beta^\star) h_2(\alpha^\star, \beta^\star) \\ & + h_1(\alpha^\star, \beta^\star)^\intercal \Sigma_{f,g}(\alpha^\star, \beta^\star) h_2(\alpha^\star, \beta^\star) + h_2(\alpha^\star, \beta^\star)^\intercal \Sigma_{g,f}(\alpha^\star, \beta^\star) h_1(\alpha^\star, \beta^\star). \end{aligned}$$

Furthermore, Σ can be consistently estimated by its sample equivalent $\hat{\Sigma}_n(\hat{\alpha}_n, \hat{\beta}_n)$.

The proof is detailed Appendix A.3.

Remark 12
Note that the asymptotic variance of the test statistic proposed in [20] can be estimated by its usual sample counterparts. Unfortunately, this is generally not the case for σ_ψ^2. To verify this, note that the natural sample estimator of the variance σ_ψ^2 is given by $\mathbb{P}_n \phi_n^2 - (\mathbb{P}_n \phi_n)^2$, which is not consistent for σ_ψ^2. To see this, note that $\phi_n = \phi_n^2$ and therefore

$$\mathbb{P}_n \phi_n^2 = \mathbb{P}_n \phi_n = \mathbb{P}_n(\phi_n - \phi_\star) + \mathbb{P}_n(\phi_\star) \to \frac{1}{2}$$

in outer probability, when n tends to infinity. Therefore, we always have $\mathbb{P}_n \phi_n^2 - (\mathbb{P}_n \phi_n)^2 \to 1/4$ in outer probability, but in Example 2, which is presented in Section 7.1.2, we will see that σ_ψ^2 is not always equal to $1/4$. Thus, σ_ψ^2 is not consistently estimated by its sample equivalent in general. This may be explained by the additional noise introduced by taking $n/2$ to center $\hat{B}_n$. If we used the unknown quantity $nP(\psi(X, \hat{\alpha}_n, \hat{\beta}_n) > 0)$ instead of $n/2$ in (5.1), the asymptotic variance of the test statistic would always be equal to $1/4$. This phenomenon does not occur for Vuong's test, since the error from considering $\mathbb{E}\left[\psi(X, \alpha^\star, \beta^\star)\right]$ instead of $\mathbb{E}\left[\psi(X, \hat{\alpha}_n, \hat{\beta}_n)\right]$ can be shown to be of order $o_{\mathbb{P}}(1)$.

To construct a consistent estimator, we directly estimate the partial derivatives of the function h. To this goal, let $u^\alpha = (0, ..., 0, 1, ..., 0) \in \mathbb{R}^{d_\alpha}$ denote the i-th unit vector in $\mathbb{R}^{d_\alpha}$, and $u_j^\beta = (0, ..., 0, 1, ..., 0) \in \mathbb{R}^{d_\beta}$ denote the j-th unit vector in $\mathbb{R}^{d_\beta}$.

Lemma 5.0.4
Consider an arbitrary positive function $e(n)$ with $\lim_{n\to\infty} e(n) = 0$ and $\lim_{n\to\infty} \sqrt{n}e(n) > 0$. Define $h_n : E_\gamma \mapsto [0,1]$, $h_n(\alpha,\beta) = \int \mathbf{1}\{\psi(x,\alpha,\beta) > 0\} d\mathbb{P}_n(x)$. Under the same assumptions as in Theorem 5.0.1 and for every $i \in \{1,\dots,d_\alpha\}$, we have,

$$\hat{h}_{1,n,i} := \frac{h_n(\hat{\alpha}_n + e(n)u_i^\alpha, \hat{\beta}_n) - h_n(\hat{\alpha}_n - e(n)u_i^\alpha, \hat{\beta}_n)}{2e(n)} \to \frac{\partial}{\partial \alpha_i} h(\alpha^\star, \beta^\star) \tag{5.2}$$

and, for every $j \in \{1,\dots,d_\beta\}$,

$$\hat{h}_{2,n,j} := \frac{h_n(\hat{\alpha}_n, \hat{\beta}_n + e(n)u_j^\beta) - h_n(\hat{\alpha}_n, \hat{\beta}_n - e(n)u_j^\beta)}{2e(n)} \to \frac{\partial}{\partial \beta_j} h(\alpha^\star, \beta^\star) \tag{5.3}$$

in outer probability.

A proof is provided in Appendix A.4.

Usually, the functions s_1 and s_2 are known. Therefore, we can propose the following estimator of σ_ψ^2.

Theorem 5.0.5
Under the same assumptions as in Theorem 5.0.1, we have

$$\hat{\sigma}_\psi^2 := \mathbb{P}_n\left(\phi_n + \hat{h}_{1,n} \cdot s_1 + \hat{h}_{2,n} \cdot s_2 - \mathbb{P}_n\big(\phi_n + \hat{h}_{1,n} \cdot s_1 + \hat{h}_{2,n} \cdot s_2\big)\right)^2 \to \sigma_\psi^2 \tag{5.4}$$

in outer probability.

Again, the proof is postponed to Appendix A.5.

By inspecting the proof of Theorem 5.0.1, we deduce that Theorem 5.0.5 is also valid if the null hypothesis in not satisfied. In this case, σ_ψ^2 is the asymptotic variance of the expression $n^{-1/2}\big(\hat{B}_n - nP(\psi(X,\alpha^\star,\beta^\star) > 0)\big)$. This remark is important if we want to apply the test under the alternative. It shows that the variance estimator $\hat{\sigma}_\psi^2$ converges to some finite real number and that the test statistic

$$\frac{\hat{B}_n - n/2}{\sqrt{n\hat{\sigma}_\psi^2}} =: T_{1,n} \tag{5.5}$$

converges to $+\infty$ or $-\infty$, depending on the sign of $P(\psi(X,\alpha^\star,\beta^\star) > 0) - 1/2$. Furthermore, by an application of Lemma 7.15 in [10], $T_{1,n}$ is asymptotically standard Normal distributed under $\mathcal{H}_0$, i.e. if $P(\psi(X,\alpha^\star,\beta^\star) > 0) = 1/2$

$$T_{1,n} \rightsquigarrow \mathcal{N}(0,1).$$

Remark 13
Note that σ_h^2 can also be estimated due to the results of Lemma 5.0.4. Mimicking the proof of Theorem 5.0.5, we deduce the estimator

$$\mathbb{P}_n\left(\hat{h}_{1,n}\cdot s_1+\hat{h}_{2,n}\cdot s_2-\mathbb{P}_n\big(\hat{h}_{1,n}\cdot s_1+\hat{h}_{2,n}\cdot s_2\big)\right)^2$$

of σ_h^2. Additionally, if all assumptions of Proposition 5.0.3 are satisfied, one can also estimate σ_h^2 via

$$\begin{aligned}&\hat{h}_1(\alpha^\star,\beta^\star)^\intercal\hat{\Sigma}_f(\alpha^\star,\beta^\star)\hat{h}_1(\alpha^\star,\beta^\star)+\hat{h}_2(\alpha^\star,\beta^\star)^\intercal\hat{\Sigma}_g(\alpha^\star,\beta^\star)\hat{h}_2(\alpha^\star,\beta^\star)\\&+\hat{h}_1(\alpha^\star,\beta^\star)^\intercal\hat{\Sigma}_{f,g}(\alpha^\star,\beta^\star)\hat{h}_2(\alpha^\star,\beta^\star)+\hat{h}_2(\alpha^\star,\beta^\star)^\intercal\hat{\Sigma}_{g,f}(\alpha^\star,\beta^\star)\hat{h}_1(\alpha^\star,\beta^\star).\end{aligned}$$

An alternative to the estimator $\hat{\sigma}_\psi^2$ is the bootstrap estimator proposed in Section 5.3, which can be applied even if the functions s_1 and s_2 are unknown.

So far, we have assumed that $\mathcal{F}_\gamma$ is $P-Donsker$. Since this requirement is far from being trivial, it would be of interest to provide some sufficient conditions so that F_γ is $P-Donsker$. In the following sections, we separately tackle the case of univariate and multivariate random vectors.

5.1 Real Valued Random Variables

Here, we solely consider univariate random variables. If X is a random variable in $\mathbb{R}$, the indicator function $\mathbf{1}_{\{\psi(\cdot,\alpha,\beta)>0\}}$ translates into sums of indicator functions over intervals, in most of the cases. The following assumption exploits this behavior to state a sufficient condition for $\mathcal{F}_\gamma$ being $P-Donsker$.

Assumption *B6 For some $\gamma>0$ and every $0<\epsilon\leq 1$, there exist some constants $M_1\leq M_2$ such that $P\left(X\in[M_1,M_2]\right)\geq 1-\epsilon^2$ and, for all $(\alpha,\beta)\in E_\gamma$, the indicator function $\mathbf{1}_{\{\psi(\cdot,\alpha,\beta)\}}$ can be written as the sum of at most $\lfloor K(\epsilon)\rfloor$ indicator functions of the form $\mathbf{1}_{(a,b]},\mathbf{1}_{(a,b)},\mathbf{1}_{[a,b]},\mathbf{1}_{[a,b)}$ on $[M_1,M_2]$. Moreover, the latter function $K:(0,1]\to[0,\infty)$ is differentiable, $\lim_{\epsilon\to 0}\epsilon K(\epsilon)=0$ and $\epsilon K(\epsilon)+\epsilon$ is strictly increasing on $[0,a]$, for $a:=(\epsilon K(\epsilon)+\epsilon)^{-1}(1)$. Furthermore,*

$$\int_0^a\sqrt{\lceil K(\epsilon)+1\rceil\log(64)-4\lceil K(\epsilon)+1\rceil\log(\epsilon)}(K'(\epsilon)\epsilon+K(\epsilon)+1)d\epsilon<\infty.$$

Lemma 5.1.1
Under Assumptions $B1,B2$ and $B6$, there exists some $\gamma>0$ such that the class F_γ is

$P-Donsker$.

Again, the proof is postponed into Appendix A.6.

An easily verifiable condition ensuring that Assumption $B6$ is satisfied is given in the following proposition.

Proposition 5.1.2
If there exist $\gamma > 0$ such that $\psi(\cdot, \alpha, \beta)$ is continuous for all $(\alpha, \beta) \in E_\gamma$ and there exits an integer K such that $\psi(\cdot, \alpha, \beta)$ has at most K zeros for any $(\alpha, \beta) \in E_\gamma$, then Assumption $B6$ is satisfied.

Proof. For the first assertion choose $M_1 = -\infty, M_2 = \infty$ and $K(\epsilon) = K$. Then, choosing the zeros of ψ as the left and right endpoints of intervals, one can write ψ as a finite sum of at most K indicator functions of the form $\mathbf{1}_{(-\infty,a)}, \mathbf{1}_{(a,b)}, \mathbf{1}_{(a,b]}, \mathbf{1}_{[a,b)}, \mathbf{1}_{[a,b]}, \mathbf{1}_{(a,\infty)}$, which is a *Donsker* class by Corollary 9.32 in [10]. □

5.2 Random Vectors In $\mathbb{R}^d$

In this subsection, we consider the more complex case of a random vector $X \in \mathbb{R}^d$. The complication arises from the shape of the boundary of the set $\{x \in \mathbb{R}^d \mid \psi(x, \alpha, \beta) > 0\}$. Since the shape of these sets strongly depends on the function ψ, we cannot state general conditions so that $\mathcal{F}_\gamma$ is $P-Donsker$. However, we state three relatively high-level conditions, which are sufficient for $\mathcal{F}_\gamma$ to be a *Donsker* class. For the sake of readability we write $x \in \mathbb{R}^d$, but all of the following assertions only need to be satisfied for $x \in Range(X)$.

Lemma 5.2.1
If, additionally to Assumptions $B1$ and $B2$, one of the following conditions is satisfied, then $\mathcal{F}_\gamma$ is $P-Donsker$.

(i) There exists $\gamma > 0$ such that, for any $(\alpha, \beta) \in E_\gamma$,

$$\{\psi(x, \alpha, \beta) > 0\} = \{\xi_1(x) > \xi_2(\alpha, \beta)\}$$

for some measurable functions ξ_1 and ξ_2 from $\mathbb{R}^d$ and E_γ, respectively, to $\mathbb{R}$.

(ii) There exists $\gamma > 0$ such that

$$\left\{ \{x \in \mathbb{R}^d \mid \psi(x, \alpha, \beta) > 0\} \mid (\alpha, \beta) \in E_\gamma \right\}$$

is a VC class of sets and the classes of functions $(\mathcal{F}_\gamma)_\delta := \{f - g \mid \int (f-g)^2 dP < \delta; f, g \in \mathcal{F}_\gamma\}$ *and* $\{h^2 \mid h \in (\mathcal{F}_\gamma)_\infty\}$ *are* P*-measurable for every* $\delta > 0$*, according to Definition 2.3.3. in [19].*

(iii) ψ *satisfies* $|\psi(x, \alpha_1, \beta_1) - \psi(x, \alpha_2, \beta_2)| \leq L(x)\|(\alpha_1, \beta_1) - (\alpha_2, \beta_2)\|_r$ *for some* $r \in \mathbb{N}$ *and* $L := \sup_x L(x) < \infty$*. Additionally, there exists* $\gamma > 0$ *and* $A \in \mathbb{R}$ *such that*

$$\lim_{\epsilon \to 0} \sup_{(\alpha,\beta) \in E_\gamma} \frac{P\left(\psi(X, \alpha, \beta) \in [-\epsilon, \epsilon]\right)}{\epsilon} < A.$$

The proof appears in Appendix A.7.

Remark 14

1. *Condition (ii) is satisfied by multivariate polynomials of bounded degree by Exercise 6.12 in [17].*
2. *Note that an equivalent condition to* $\mathcal{F}_\gamma$ *being a VC subgraph class is* $\tilde{\mathcal{F}}_\gamma := \{\{(\alpha, \beta) \mid (\alpha, \beta) \in E_\gamma,\ \psi(x, \alpha, \beta) > 0\} \mid x \in \mathbb{R}^d\}$ *being a VC-class of sets, which is proven in [9]. This condition may be easier to verify in some cases.*

5.3 The Bootstrap

Now, we present an alternative estimation procedure of the asymptotic variance σ_ψ^2, given in Equation (5.1). We need to introduce a slightly modified mathematical framework to account for the additional randomness induced by the bootstrap weights.

Let $(\xi_{i,n})_{1 \leq i \leq n; n \in \mathbb{N}}$ be an exchangeable triangular array of non-negative random variables on a probability space Ω_2 with probability measure $\mathbb{P}_W$. Assume that the $\xi_{i,n}$ satisfy the following conditions, which are given in [19] Chapter 3.6.2

Assumptions

W1
$$\sum_{i=1}^{n} \xi_{i,n} = n;$$

W2
$$\sup_n \int_0^\infty \sqrt{\mathbb{P}_W(|\xi_{1,n} - 1| > x)}\, dx < \infty;$$

W3
$$\frac{1}{\sqrt{n}} \mathbb{E}_{\mathbb{P}_W}\left[\max_{1 \leq i \leq n} |\xi_{i,n} - 1|\right] \to 0;$$

W4 For some constant $c > 0$, $n^{-1}\sum_{i=1}^{n}(\xi_{i,n}-1)^2 \to c^2$ in $\mathbb{P}_W$-probability;

Remark 15

Consider i.i.d. non-negative random variables $(\tilde{\xi}_i)_{i\in\mathbb{N}}$ with mean $0<\mu<\infty$ and variance $0<\tau^2<\infty$, such that $\int_0^\infty \left(\mathbb{P}_W(|\tilde{\xi}_1|>x)\right)^{1/2}dx<\infty$. Then, the conditions $W1-W4$ are satisfied for $\xi_{i,n} := n\tilde{\xi}_i\left(\sum_{1\le i\le n}\tilde{\xi}_i\right)^{-1}$, with $c=\tau/\mu$.

Define the exchangeable bootstrap empirical measure as $\tilde{\mathbb{P}}_n := n^{-1}\sum_{i=1}^{n}\xi_{i,n}\delta_{X_i}$. Furthermore, define the bootstrap empirical process as $\tilde{\mathbb{G}}_n := \sqrt{n}c^{-1}(\tilde{\mathbb{P}}_n-\mathbb{P}_n)$ and assume that the sequences $(X_i)_{i\in\mathbb{N}}$ and $(\xi_{i,n})_{1\le i\le n;n\in\mathbb{N}}$ originate from a probability space with product structure as defined in [4], Section 3. This implies that the sequence $(X_i)_{i\in\mathbb{N}}$ only depends on the first coordinate of some probability space $(\Omega := \Omega_1\times\Omega_2, \mathcal{A} := \mathcal{A}_1\times\mathcal{A}_2, \mathbb{P}_{XW})$ equipped with a probability measure $\mathbb{P}_{XW} := \mathbb{P}_X\otimes\mathbb{P}_W$. On the other side, the triangular array $(\xi_i)_{1\le i\le n;n\in\mathbb{N}}$ only depends on the second component of the latter space. This specific structure implies that $(X_i)_{i\in\mathbb{N}}$ and $(\xi_i)_{1\le i\le n;n\in\mathbb{N}}$ are independent. We write $\tilde{\mathbb{G}}_n \underset{\xi}{\rightsquigarrow} \mathbb{G}$ to denote convergence in the sense of the bounded Lipschitz metric, i.e.

$$\mathbb{E}_{\mathbb{P}_W}\left[h(\tilde{\mathbb{G}}_n)\right]^* - \mathbb{E}_{\mathbb{P}_W}\left[h(\tilde{\mathbb{G}}_n)\right]_* \to 0 \text{ in } \mathbb{P}_X\text{-probability and}$$

$$\sup_{h\in BL_1}\left|\mathbb{E}_{\mathbb{P}_W}\left[h(\tilde{\mathbb{G}}_n)\right] - \mathbb{E}\left[h(\mathbb{G})\right]\right| \to 0 \text{ in outer } \mathbb{P}_X\text{-probability,}$$

where

- $\mathbb{E}_{\mathbb{P}_W}$ denotes expectation w.r.t. the weights $(\xi_{i,n})_{1\le i\le n,n\in\mathbb{N}}$ treating the sample $(X_i)_{1\le i\le n}$ as constants,
- $BL_1 := \{h \mid h : l^\infty(\mathcal{F}_\gamma)\mapsto\mathbb{R};\ \sup_{x\in l^\infty(\mathcal{F}_\gamma)}|h(x)|\le 1;\ |h(x)-h(y)|\le \sup_{f\in\mathcal{F}_\gamma}|x(f)-y(f)|\}$
- Y^* and Y_* denote the measurable majorant and minorant of a random map Y as defined in [19].

In order to prove the convergence of the bootstrapped process, we need additional assumptions on the estimators $\hat{\alpha}_n$ and $\hat{\beta}_n$ and on the class of functions $\mathcal{F}_\gamma$. Let $\tilde{\alpha}_n$ and $\tilde{\beta}_n$ denote the bootstrapped estimators for $\alpha^\star$ and $\beta^\star$, i.e. the estimators of the pseudo-true values calculated from the bootstrap sample.

Assumptions

B7 The estimators $\hat{\alpha}_n, \tilde{\alpha}_n$ and $\hat{\beta}_n, \tilde{\beta}_n$ satisfy $\sqrt{n}(\tilde{\alpha}_n-\hat{\alpha}_n) = \sqrt{n}\left(\tilde{\mathbb{P}}_n-\mathbb{P}_n\right)s_1 + o_{\mathbb{P}^*_{XW}}(1)$ and $\sqrt{n}\left(\tilde{\beta}_n-\hat{\beta}_n\right) = \sqrt{n}\left(\tilde{\mathbb{P}}_n-\mathbb{P}_n\right)s_2 + o_{\mathbb{P}^*_{XW}}(1)$.

B8 *There exists $\gamma > 0$ such that the class of functions $\{f - g \mid \int (f-g)^2 dP < \delta, (f,g) \in \mathcal{F}_\gamma\}$ is P-measurable for every $\delta > 0$, according to Definition 2.3.3. in [19].*

B9 *For any compact set $K \subset Range(X)$, ψ satisfies the following uniform continuity condition: for all $\epsilon > 0$, there exists $\delta > 0$ such that*

$$\|(\alpha_1, \beta_1) - (\alpha_2, \beta_2)\|_1 \leq \delta \Rightarrow \sup_{x \in K} |\psi(x, \alpha_1, \beta_1) - \psi(x, \alpha_2, \beta_2)| \leq \epsilon.$$

Remark 16

1. *Let $\mathbb{Q}$ denote the set of rational numbers. We require Assumption B8 solely to satisfy the measurability condition of Theorem 3.6.16 in [19], which allows us to freely use Fubini's theorem. If we restrict the discussion to the bootstrap schemes described in [10], we do not need Assumption B8. Furthermore, if we restrict ourselves to rational parameters, i.e. to $E_\gamma \cap \mathbb{Q}^{d_\alpha + d_\beta}$, Assumption B8 would be satisfied.*

2. *If there exists $\gamma > 0$ such that $\psi(x, \alpha, \beta)$ is continuous for all $(x, \alpha, \beta) \in \mathbb{R}^d \times E_\gamma$, then Assumption B9 is satisfied (uniform continuity on a compact subset).*

In the following lemma, we state a mild condition such that Assumption $B8$ is satisfied.

Lemma 5.3.1
Assume there exists $\gamma > 0$ such that Assumption B2 is satisfied. Moreover, for all $(\alpha, \beta) \in E_\gamma$, there exists $(\bar{\alpha}_n, \bar{\beta}_n) \in E_\gamma \cap \mathbb{Q}^{d_\alpha + d_\beta} \setminus (\alpha, \beta)$ with $\lim_{n\to\infty}(\bar{\alpha}_n, \bar{\beta}_n) = (\alpha, \beta)$ and $\psi(x, \bar{\alpha}_n, \bar{\beta}_n)) \leq 0$ for all $x \in \mathbb{R}^d$ with $\psi(x, \alpha, \beta) = 0$. Then, Assumption B8 is satisfied.

Proof. It is sufficient to show that $\mathcal{F}_\gamma$ is a pointwise measurable class, according to Proposition 8.11 in [10]. To show that $\mathcal{F}_\gamma$ is pointwise measurable choose $\mathcal{G} := \{\mathbf{1}_{\{\psi(\cdot,\alpha,\beta)>0\}} \big| \ (\alpha, \beta) \in E_\gamma \cap \mathbb{Q}^{d_\alpha + d_\beta}\}$. Obviously, $\mathcal{G}$ is countable. Now, choose an arbitrary $\mathbf{1}\{\psi(\cdot, \alpha, \beta) > 0\} \in \mathcal{F}_\gamma$. By Assumption $B2$ and the stated condition, we can choose $(\bar{\alpha}_n, \bar{\beta}_n) \in E_\gamma \cap \mathbb{Q}^{d_\alpha + d_\beta}$ with $\lim_{n\to\infty}(\bar{\alpha}_n, \bar{\beta}_n) = (\alpha, \beta)$, such that for all $x \in \mathbb{R}^d$ with $\psi(x, \alpha, \beta) = 0$:

$$\mathbf{1}\{\psi(x, \alpha, \beta)\} = 0 = \lim_{n\to\infty} \mathbf{1}\{\psi(x, \bar{\alpha}_n, \bar{\beta}_n) > 0\}.$$

If $\psi(x, \alpha, \beta) \neq 0$, Assumption $B2$ implies that there exists $N \in \mathbb{N}$ such that for all $n \geq N$: $sign(\psi(x, \alpha, \beta)) = sign(\psi(x, \bar{\alpha}_n, \bar{\beta}_n))$. This shows that for all $x \in \mathbb{R}^d$

$$\lim_{n\to\infty} \mathbf{1}\{\psi(x, \bar{\alpha}_n, \bar{\beta}_n) > 0\} = \mathbf{1}\{\psi(x, \alpha, \beta) > 0\}.$$

Therefore, $\mathcal{F}_\gamma$ is pointwise measurable and the claim follows. □

We define the bootstrapped version of $\hat{B}_n$ as

$$\tilde{B}_n := n\tilde{\mathbb{P}}_n \mathbf{1}\{\psi(\cdot, \tilde{\alpha}_n, \tilde{\beta}_n) > 0\} = \sum_{i=1}^{n} \xi_{i,n} \mathbf{1}\{\psi(X_i, \tilde{\alpha}_n, \tilde{\beta}_n) > 0\}. \tag{5.6}$$

The next theorem shows that the bootstrapped statistic $\tilde{B}_n$ behaves similarly to $\hat{B}_n$, knowing the initial sample $(X_i)_{i=1,\dots,n}$.

Theorem 5.3.2
Assume Assumptions $B1 - B5$ and $B7 - B9$ are satisfied. Moreover, assume there exists $\gamma > 0$ such that $\mathcal{F}_\gamma$ is $P -$ Donsker. Then, the following statements are true for the bootstrapped version $\tilde{B}_n$ of $\hat{B}_n$:

(i)

$$\frac{1}{c\sqrt{n}} \left(\tilde{B}_n - \hat{B}_n \right) = \tilde{\mathbb{G}}_n \left(\mathbf{1}\{\psi(X, \alpha^\star, \beta^\star) > 0\} + h_1(\alpha^\star, \beta^\star) \cdot s_1 + h_2(\alpha^\star, \beta^\star) \cdot s_2 \right) + o_{\mathbb{P}^*_{XW}}(1).$$

(ii)

$$\frac{1}{c\sqrt{n}} \left(\tilde{B}_n - \hat{B}_n \right) \underset{\xi}{\rightsquigarrow} \mathcal{N}(0, \sigma_\psi^2).$$

(iii) Statement (ii) is also satisfied unconditionally, i.e. with $\underset{\xi}{\rightsquigarrow}$ replaced by $\rightsquigarrow$ w.r.t. $\mathbb{P}_{XW}$.

See the proof in Appendix A.8, which utilizes Lemma B.0.1 from Appendix B.

Chapter 6

Comparison With Vuong's Test

In this chapter, we compare Clarke's null hypothesis $\mathcal{H}_0$ with Vuong's null hypothesis $\mathcal{H}_0^V$. Moreover, we compare our Assumptions $B1 - B9$ with the assumptions by Vuong in [20].

Recall that our definition of non-nested models is slightly stronger than the definition of Vuong, since we do not allow the families of competing densities $(f_\alpha)_{\alpha\in\Theta_\alpha}$ and $(g_\beta)_{\beta\in\Theta_\beta}$ to intersect on a set with positive probability. Further, recall that the null hypothesis of Clarke's test is

$$\mathcal{H}_0 : \; P\left(\log\left(\frac{f(X,\alpha^\star)}{g(X,\beta^\star)}\right) > 0\right) = \frac{1}{2},$$

and the null hypothesis in [20] is

$$\mathcal{H}_0^V : \; \mathbb{E}\left[\log\left(\frac{f(X,\alpha^\star)}{g(X,\beta^\star)}\right)\right] = 0.$$

Obviously, if the distribution of $\log(f(X,\alpha^\star)/g(X,\beta^\star))$ is symmetric around 0, both null hypotheses coincide, in case the expectation in $\mathcal{H}_0^V$ is finite. Apart from symmetry around 0, it is hard to state any general condition which would imply the equivalence of both null hypotheses. Presumably, this is hardly ever the case if the distribution is not symmetric around 0.

Comparing $\mathcal{H}_0$ and $\mathcal{H}_0^V$, we observe that $\mathcal{H}_0^V$ cannot be defined in case $\mathbb{E}[\log(f(X,\alpha^\star))]$ or $\mathbb{E}[\log(g(X,\beta^\star))]$ does not exist or is infinite. However, in many of these cases, $\mathcal{H}_0$ is still well defined. For example, consider a Normal distribution compared to a t-distribution and assume that the true distribution is Cauchy. In this case $\mathcal{H}_0^V$ is undefined, whereas $\mathcal{H}_0$ is well defined. On the other hand, if $\mathcal{H}_0^V$ is well defined, $\mathcal{H}_0$ is also well defined. This shows that Clarke's test can be applied in a more general setting than Vuong's test.

In contrast to Vuong's test, we lack the relation of test statistic $T_{1,n}$ to the KLIC of the

competing models. In fact, for skewed distributions in particular, it may be the case that the two tests prefer different models. An example of such a discrepancy is yielded by the model comparison in (7.2) in our empirical section. A consequence from comparing the $KLIC$ of the competing models in Vuong's approach is that the null assumption $\mathcal{H}_0^V$ excludes that one or both of the models are correctly specified. To see the claim, recall Definition 4.1.1 and note that $\mathcal{H}_0^V$ implies $K(f_{\alpha^\star}) = K(g_{\beta^\star})$. Now, if at least one of the models is correctly specified, it is true that $0 = K(f_{\alpha^\star}) = K(g_{\beta^\star})$, which is equivalent to $p_X = f_{\alpha^\star} = g_{\beta^\star}$. Therefore, the models are nested, contradicting the assumption of non-nestedness. Unsurprisingly, $\mathcal{H}_0$ does not exclude the possibility that either of the models is correctly specified, which shows that Clarke's test may not be able to find the true model, even though $f_{\alpha^\star}$ or $g_{\beta^\star}$ is equal to the true model.

Next, we compare our Assumptions $B1 - B9$ with the assumptions in [20].

- As stated above, Vuong's test requires that both $\mathbb{E}[\log(f(X, \alpha^\star))]$ and $\mathbb{E}[\log(g(X, \beta^\star))]$ exist and at least one of them is finite. Clarke's test does not require such moment conditions on the log-likelihood. Furthermore, we do not require any moment condition on (the derivatives of) $\psi(X, \alpha, \beta)$, but we implicitly impose some moment conditions on the true distribution via Assumption $B5$. However, under the conditions $A1 - A5$ in [20], the implicit moment conditions of Assumption $B5$ are satisfied.

- Comparing the regularity Assumption $A4$ in Vuong with Assumption $B2$, we observe that we only require $\psi(x, \alpha, \beta)$ to be continuous in (α, β), instead of being twice continuously differentiable. For example, a Laplace density is not excluded by our conditions, whereas it is excluded in Vuong's framework. In contrast to Vuong, we require a mild differentiability condition on $P(\psi(X, \alpha, \beta) > 0)$.

- We require the estimators and the class of functions $\mathcal{F}_\gamma$ to be *Donsker*. The *Donsker* property of $\mathcal{F}_\gamma$ is likely to be satisfied if the function ψ has sufficiently "regular" behavior on the boundary of the set $\{x \in \mathbb{R}^d \mid \psi(x, \alpha, \beta) > 0\}$, which is probably the case for many competing densities. Vuong requires the asymptotic Normality of $\sqrt{n}\,(n^{-1}\sum_{i=1}^n \log(f(X_i, \alpha)) - \mathbb{E}[\log(f(X, \alpha))])$ and $\sqrt{n}\,(n^{-1}\sum_{i=1}^n \log(g(X_i, \beta)) - \mathbb{E}[\log(f(X, \beta))])$ for all $(\alpha, \beta) \in \Theta_\alpha \times \Theta_\beta$. If an additional uniform convergence condition would be satisfied, the class of functions $\mathcal{F}_V := \{\log(f(\cdot, \alpha)/g(\cdot, \beta)) \mid (\alpha, \beta) \in \Theta_\alpha \times \Theta_\beta\}$ would be a *Donsker* class.

Finally, one can conclude that the assumptions of Clarke's test and Vuong's test are similar, but differ slightly in some moment and "smoothness" conditions. Even if Clarke's

test can be applied more often, all of our Assumptions $B1 - B9$ and all assumptions of Vuong will be satisfied in many cases. A power analysis would help to discriminate between both approaches.

Chapter 7

Examples And Simulations

In this chapter, we present two examples of competing models, which satisfy the null hypothesis $\mathcal{H}_0$ and we conduct a small simulation study to assess the finite sample behavior of the estimators introduced in Chapter 5. Additionally, we simulate Clarke's test for competing models that do not satisfy $\mathcal{H}_0$ in order to investigate the empirical power of Clarke‘s test.

7.1 Examples

Let us present two examples of strictly non-nested models satisfying $\mathcal{H}_0$ and Assumptions $B1 - B9$. Our goal is to illustrate the theoretical quantities involved in the formulation of the test statistic $T_{1,n}$ and to use them as a benchmark to assess the accuracy of the estimators in finite samples in Section 7.2.

7.1.1 Example 1

First, we come back to the example from Section 4.3. We need to verify Assumptions $B1 - B9$. It is obvious that Assumptions $B1 - B3$ and $B6$ are satisfied for this model comparison. Since $\hat{\beta}_n = \hat{\alpha}_n = n^{-1}\sum_{i=1}^{n} X_i$, we get that Assumptions $B5$ and $B7$ are satisfied with $s_1(x) = s_2(x) = x$. Due to the smoothness of $\psi(x, \alpha, \beta)$ Assumptions $B8$ and $B9$ are satisfied. It remains to verify Assumption $B4$.
Again, assuming w.l.o.g that $\sigma_f > \sigma_g$, we calculate

$$P\Big(\log\Big(\frac{f(X,\alpha)}{g(X,\beta)}\Big) > 0\Big) = P\Big(-\frac{(X-\alpha)^2}{2\sigma_f^2} + \frac{(X-\beta)^2}{2\sigma_g^2} + \log\left(\frac{\sigma_g}{\sigma_f}\right) > 0\Big)$$
$$= \Phi\left(x_2(\alpha,\beta) - \mu_0\right) - \Phi\left(x_1(\alpha,\beta) - \mu_0\right),$$

where $x_2(\alpha,\beta) > x_1(\alpha,\beta)$ are the zeros (in x) of the function $-\frac{(x-\alpha)^2}{2\sigma_f^2} + \frac{(x-\beta)^2}{2\sigma_g^2} + \log\left(\frac{\sigma_g}{\sigma_f}\right)$. $x_1(\alpha,\beta)$ and $x_2(\alpha,\beta)$ can be calculated by the usual formula for zeros of polynomials of degree 2 and are given by

$$x_{1,2}(\alpha,\beta) = \frac{-\left(\frac{\alpha}{\sigma_f^2} - \frac{\beta}{\sigma_g^2}\right)}{\frac{1}{\sigma_g^2} - \frac{1}{\sigma_f^2}} \pm \sqrt{\frac{\left(\frac{\alpha}{\sigma_f^2} - \frac{\beta}{\sigma_g^2}\right)^2}{\left(\frac{1}{\sigma_g^2} - \frac{1}{\sigma_f^2}\right)^2} - 2\frac{\left(-\frac{\alpha^2}{2\sigma_f^2} + \frac{\beta^2}{2\sigma_g^2} + \log\left(\frac{\sigma_g}{\sigma_f}\right)\right)}{\frac{1}{\sigma_g^2} - \frac{1}{\sigma_f^2}}}.$$

One can check that x_1 and x_2 are real numbers for sufficiently small $\gamma > 0$. To see this, recall that $\alpha^\star = \beta^\star = \mu_0$ and note that x_1 and x_2 are continuous in $(\alpha^\star, \beta^\star)$. Further, we have

$$\frac{\partial}{\partial\alpha} x_{1,2}(\alpha,\beta) = \frac{-1}{\sigma_f^2\left(\frac{1}{\sigma_g^2} - \frac{1}{\sigma_f^2}\right)} \pm \left(\frac{\left(\frac{\alpha}{\sigma_f^2} - \frac{\beta}{\sigma_g^2}\right)^2}{\left(\frac{1}{\sigma_g^2} - \frac{1}{\sigma_f^2}\right)^2} - 2\frac{\left(-\frac{\alpha^2}{2\sigma_f^2} + \frac{\beta^2}{2\sigma_g^2} + \log\left(\frac{\sigma_g}{\sigma_f}\right)\right)}{\frac{1}{\sigma_g^2} - \frac{1}{\sigma_f^2}}\right)^{-\frac{1}{2}} \left(\frac{\left(\frac{\alpha}{\sigma_f^2} - \frac{\beta}{\sigma_g^2}\right)}{\sigma_f^2\left(\frac{1}{\sigma_g^2} - \frac{1}{\sigma_f^2}\right)^2} + \frac{\alpha}{\sigma_f^2\left(\frac{1}{\sigma_g^2} - \frac{1}{\sigma_f^2}\right)}\right)$$

and

$$\frac{\partial}{\partial\beta} x_{1,2}(\alpha,\beta) = \frac{1}{\sigma_g^2\left(\frac{1}{\sigma_g^2} - \frac{1}{\sigma_f^2}\right)} \pm \left(\frac{\left(\frac{\alpha}{\sigma_f^2} - \frac{\beta}{\sigma_g^2}\right)^2}{\left(\frac{1}{\sigma_g^2} - \frac{1}{\sigma_f^2}\right)^2} - 2\frac{\left(-\frac{\alpha^2}{2\sigma_f^2} + \frac{\beta^2}{2\sigma_g^2} + \log\left(\frac{\sigma_g}{\sigma_f}\right)\right)}{\frac{1}{\sigma_g^2} - \frac{1}{\sigma_f^2}}\right)^{-\frac{1}{2}} \left(-\frac{\left(\frac{\alpha}{\sigma_f^2} - \frac{\beta}{\sigma_g^2}\right)}{\sigma_g^2\left(\frac{1}{\sigma_g^2} - \frac{1}{\sigma_f^2}\right)^2} - \frac{\beta}{\sigma_g^2\left(\frac{1}{\sigma_g^2} - \frac{1}{\sigma_f^2}\right)}\right).$$

Using that $\mu_0 = \alpha^\star = \beta^\star$ and after some calculations, we get

$$\begin{aligned} h_1(\alpha^\star,\beta^\star) &= \varphi\left(x_2(\alpha^\star,\beta^\star) - \mu_0\right)\frac{\partial}{\partial\alpha}x_2(\alpha^\star,\beta^\star) - \varphi\left(x_1(\alpha^\star,\beta^\star) - \mu_0\right)\frac{\partial}{\partial\alpha}x_1(\alpha^\star,\beta^\star) \\ &= \varphi\left(\sqrt{2\frac{\log\left(\frac{\sigma_f}{\sigma_g}\right)}{\frac{1}{\sigma_g^2} - \frac{1}{\sigma_f^2}}}\right)\frac{-1}{\sigma_f^2\left(\frac{1}{\sigma_g^2} - \frac{1}{\sigma_f^2}\right)} - \varphi\left(-\sqrt{2\frac{\log\left(\frac{\sigma_f}{\sigma_g}\right)}{\frac{1}{\sigma_g^2} - \frac{1}{\sigma_f^2}}}\right)\frac{-1}{\sigma_f^2\left(\frac{1}{\sigma_g^2} - \frac{1}{\sigma_f^2}\right)} \\ &= 0. \end{aligned}$$

Similarly,

$$h_2(\alpha^\star,\beta^\star) = \varphi\left(x_2(\alpha^\star,\beta^\star) - \mu_0\right)\frac{\partial}{\partial\beta}x_2(\alpha^\star,\beta^\star) - \varphi\left(x_1(\alpha^\star,\beta^\star) - \mu_0\right)\frac{\partial}{\partial\beta}x_1(\alpha^\star,\beta^\star)$$

$$= \varphi\left(\sqrt{2\frac{\log\left(\frac{\sigma_f}{\sigma_g}\right)}{\frac{1}{\sigma_g^2}-\frac{1}{\sigma_f^2}}}\right)\frac{1}{\sigma_g^2\left(\frac{1}{\sigma_g^2}-\frac{1}{\sigma_f^2}\right)} - \varphi\left(-\sqrt{2\frac{\log\left(\frac{\sigma_f}{\sigma_g}\right)}{\frac{1}{\sigma_g^2}-\frac{1}{\sigma_f^2}}}\right)\frac{1}{\sigma_g^2\left(\frac{1}{\sigma_g^2}-\frac{1}{\sigma_f^2}\right)}$$
$$= 0.$$

Obviously, h_1 and h_2 are continuous at $(\alpha^\star, \beta^\star)$, which verifies $B4$. Thus, all Assumptions $B1 - B9$ are verified and Theorem 5.0.1 is applicable.

By Theorem 5.0.1, the asymptotic variance of Clarke's test is $1/4$, independent of the choice of σ_f, σ_g and μ_0, since $h_1 = h_2 = 0$. This result is surprising, since we would expect the asymptotic variance to exceed the asymptotic variance of B_n, due to the randomness induced by $\hat{\alpha}_n$ and $\hat{\beta}_n$. However and as expected, as we will see in Example 2 below, there exist cases in which the asymptotic variance exceeds $1/4$.

Remark 17
When carefully looking at Example 1, we can observe that the previous results can be generalized as described in the following: Assume that the true distribution has a continuous and symmetric density around its expected value. Then $\alpha^\star = \beta^\star = \mathbb{E}[X]$ as $\mathbb{E}[X]$ minimizes the function $a \mapsto \mathbb{E}[(X-a)^2]$. Repeating the above calculations, one can observe that for arbitrary σ_f and σ_g, i.e. we do not have to satisfy the null hypothesis, the asymptotic variance of Clarke's test is always $P(\psi(X, \alpha^\star, \beta^\star > 0)\,(1 - P(\psi(X, \alpha^\star, \beta^\star > 0))$, since $h_1 = h_2 = 0$ in every of these cases. Comparing this with Vuong's test, we calculate

$$\begin{aligned}
\operatorname{Var}\left(\log\Big(\frac{f(X,\alpha^\star)}{g(X,\beta^\star)}\Big)\right) &= \operatorname{Var}\left(\frac{\sigma_f^2-\sigma_g^2}{2\sigma_f^2\sigma_g^2}(X-\mathbb{E}[X])^2\right)\\
&= \left(\frac{\sigma_f^2-\sigma_g^2}{2\sigma_f^2\sigma_g^2}\right)^2 \operatorname{Var}\left((X-\mathbb{E}[X])^2\right)\\
&= \left(\frac{\sigma_f^2-\sigma_g^2}{2\sigma_f^2\sigma_g^2}\right)^2 \left(\mathbb{E}\left[(X-\mathbb{E}[X])^4\right] - \operatorname{Var}(X)^2\right),
\end{aligned}$$

which clearly varies with the chosen values of σ_f and σ_g. Therefore, when applying Clarke's test to compare two Normal distributions with fixed variance, one does not need to estimate the asymptotic variance, whereas one needs to do so when applying Vuong's test.

7.1.2 Example 2

As as second example, we consider random variables supported on the positive real line. The main purpose of this example is to find competing models that satisfy $\mathcal{H}_0$ and have

asymptotic variance larger than 1/4.

Choose the true distribution P as a member of the family of generalized Gamma distributions as defined in [14], which is defined by the densities

$$q(x,a,d,p) = \frac{p}{a^d\Gamma(d/p))}x^{d-1}e^{-\left(\frac{x}{a}\right)^p}\mathbf{1}_{\{x>0\}}, \quad a,d,p>0.$$

Assume that the competing models follow a Weibull distribution, given by the family of densities

$$w(x,\alpha_1,\alpha_2) = \frac{\alpha_2}{\alpha_1}x^{\alpha_2-1}\exp\left(-\frac{x^{\alpha_2}}{\alpha_1}\right)\mathbf{1}_{\{x>0\}}, \quad \alpha_1,\alpha_2>0$$

and a Gamma distribution, whose densities are

$$g(x,\beta_1,\beta_2) = \frac{1}{\beta_1^{\beta_2}\Gamma(\beta_2)}x^{\beta_2-1}\exp\left(-\frac{1}{\beta_1}x\right)\mathbf{1}_{\{x>0\}}, \quad \beta_1,\beta_2>0.$$

Note that for $p=d$ the generalized Gamma distribution becomes a Weibull distribution and for $p=1$ the generalized Gamma distribution becomes a Gamma distribution. In order to satisfy Assumption $B5$, this particular parametrization of the competing models is convenient, because we need to represent the estimators $\hat{\alpha}_n$ and $\hat{\beta}_n$ as $\hat{\alpha}_n = \mathbb{P}_n s_1 + o_P(1)$ and $\hat{\beta}_n = \mathbb{P}_n s_2 + o_P(1)$ for some measurable functions s_1 and s_2.

Obviously, if $\alpha_2=\beta_2=1$, the models are nested. To satisfy Assumption $B3$, we choose $\alpha_2=\beta_2=2$. As both families only depend on one remaining parameter, we denote $\alpha_1 =: \alpha$ and $\beta_1 =: \beta$. First, we calculate the pseudo maximum likelihood estimators of $\alpha^\star$ and $\beta^\star$. The partial derivative of $\sum_{i=1}^n \log(w(X_i,\alpha,2))$ w.r.t. α is given by

$$\frac{\partial}{\partial\alpha}\sum_{i=1}^n \log\left(w(X_i,\alpha,2)\right) = -\frac{n}{\alpha}+\sum_{i=1}^n\frac{X_i^2}{\alpha^2}.$$

Thus, the pseudo maximum likelihood estimator of $\alpha^\star$ is equal to $\hat{\alpha}_n = n^{-1}\sum_{i=1}^n X_i^2$. Next, we calculate the pseudo maximum likelihood estimator of $\beta^\star$. The partial derivative of $\sum_{i=1}^n \log(g(X_i,\beta,2))$ w.r.t. β is given by

$$\frac{\partial}{\partial\beta}\sum_{i=1}^n \log(g(X_i,\beta,2)) = -\frac{2n}{\beta}+\sum_{i=1}^n\frac{X_i}{\beta^2}.$$

Therefore, the pseudo maximum likelihood estimator of $\beta^\star$ is equal to $\hat{\beta}_n = (n)^{-1}\sum_{i=1}^n X_i/2$.

In the next step, we calculate the pseudo-true values $\alpha^\star$ and $\beta^\star$ explicitly, in case the true distribution belongs to the family of generalized Gamma distributions. The calculations

of $\alpha^\star$ and $\beta^\star$ involve the calculations of $\mathbb{E}\left[\log(w(X,\alpha,2))\right]$ and $\mathbb{E}\left[\log(g(X,\beta,2))\right]$, which is a non-trivial task. We will use the formulas given in [3] and [14] and the properties of the $KLIC$ to split the expressions in several terms, which can be calculated separately. The details of the calculations are available in Appendix E and yield the following expressions:

$$\mathbb{E}\left[\log\left(w(X,\alpha,2)\right)\right] = \log\left(\frac{2a}{\alpha}\right) + \frac{1}{p}\tau\left(\frac{d}{p}\right) - \frac{\Gamma\left(\frac{d+2}{p}\right)}{\Gamma\left(\frac{d}{p}\right)}\frac{a^2}{\alpha}$$

and

$$\mathbb{E}\left[\log\left(g(X,\beta,2)\right)\right] = \log\left(\frac{a}{\beta^2}\right) + \frac{1}{p}\tau\left(\frac{d}{p}\right) - \frac{\Gamma\left(\frac{d+1}{p}\right)}{\Gamma\left(\frac{d}{p}\right)}\frac{a}{\beta},$$

where $\tau(y) = \Gamma'(y)/\Gamma(y)$ denotes the digamma function.
Therefore, we get

$$\frac{\partial}{\partial\alpha}\mathbb{E}\left[\log\left(g(X,\beta,2)\right)\right] = -\frac{1}{\alpha} + \frac{\Gamma\left(\frac{d+2}{p}\right)}{\Gamma\left(\frac{d}{p}\right)}\frac{a^2}{\alpha^2},$$

which implies $\alpha^\star = a^2\Gamma\left(\frac{d+2}{p}\right)/\Gamma\left(\frac{d}{p}\right)$. Similarly,

$$\frac{\partial}{\partial\beta}\mathbb{E}\left[\log\left(g(X,\beta,2)\right)\right] = -\frac{2}{\beta} + \frac{\Gamma\left(\frac{d+1}{p}\right)}{\Gamma\left(\frac{d}{p}\right)}\frac{a}{\beta^2},$$

which implies $\beta^\star = a\Gamma\left(\frac{d+1}{p}\right)/2\Gamma\left(\frac{d}{p}\right)$.

Next, we fix the parameters of the underlying generalized Gamma distribution, such that $\mathcal{H}_0$ is satisfied. To calculate the probability of $\psi(X,\alpha^\star,\beta^\star) > 0$, we need to find the zeros (in x) of the continuous function

$$\log\left(\frac{g(x,\beta^\star,2)}{w(x,\alpha^\star,2)}\right) = \log\left(\frac{x\exp\left(-\frac{x}{\beta^\star}\right)\alpha^\star}{2(\beta^\star)^2 x\exp\left(-\frac{x^2}{\alpha^\star}\right)}\right) = \log\left(\frac{\alpha^\star}{2(\beta^\star)^2}\right) - \frac{x}{\beta^\star} + \frac{x^2}{\alpha^\star}.$$

By the usual formula for zeros of a polynomial of degree 2 we get,

$$x_{1,2} = \frac{\alpha^\star}{2\beta^\star} \pm \sqrt{\left(\frac{\alpha^\star}{2\beta^\star}\right)^2 - \alpha^\star\log\left(\frac{\alpha^\star}{2(\beta^\star)^2}\right)}$$

$$= \frac{a\Gamma\left(\frac{d+2}{p}\right)}{\Gamma\left(\frac{d+1}{p}\right)} \pm \sqrt{\left(\frac{a\Gamma\left(\frac{d+2}{p}\right)}{\Gamma\left(\frac{d+1}{p}\right)}\right)^2 - \frac{a^2\Gamma\left(\frac{d+2}{p}\right)}{\Gamma\left(\frac{d}{p}\right)} \log\left(\frac{\Gamma\left(\frac{d+2}{p}\right) 2\Gamma\left(\frac{d}{p}\right)}{\Gamma\left(\frac{d+1}{p}\right)^2}\right)}.$$

One can show that x_1 and x_2 are real number for sufficiently small $\gamma > 0$ and we can w.l.o.g. assume that $x_1 < x_2$. Denoting the distribution function of the generalized Gamma distribution by $Q_{a,d,p}$, we need to find a triplet of parameters (a, d, p) such that

$$P\left(\psi(X, \alpha^\star, \beta^\star) > 0\right) = Q_{a,d,p}(x_1(\alpha^\star, \beta^\star)) + 1 - Q_{a,d,p}(x_2(\alpha^\star, \beta^\star)) = \frac{1}{2}$$

in order to satisfy the null hypothesis $\mathcal{H}_0$.

By fixing $a = 1/2$ and $d = 3$, we can solve for p numerically and obtain an approximated value $p \approx 0.6457$. Fixing a, d and p as above ensures that $\mathcal{H}_0$ is satisfied with $\alpha^\star \approx 52.85873$ and $\beta^\star \approx 2.938702$.

It remains to verify Assumptions $B1 - B9$. Obviously, Assumptions $B1 - B3$ and $B9$ are satisfied. Additionally $B6$ is satisfied with $K(\epsilon) = 2$. Due to the special structure of x_1 and x_2, Assumption $B8$ is also satisfied. To verify Assumptions $B5$ and $B7$, we observe that $s_1(x) = x^2$ and $s_2(x) = x/2$. Since $\mathbb{E}[X^4] < \infty$, Assumptions $B5$ and $B7$ are satisfied. To verify $B4$, it is enough to observe that the zeros $x_1(\alpha, \beta)$ and $x_2(\alpha, \beta)$ are continuously differentiable at $(\alpha^\star, \beta^\star)$ by Remark 1. However, since we want to calculate the explicit value of the asymptotic variance σ^2_ψ, we will verify $B4$ by explicitly calculating $h_1(\alpha, \beta)$ and $h_2(\alpha, \beta)$.

Since $\psi(\frac{x_1(\alpha^\star,\beta^\star)+x_2(\alpha^\star,\beta^\star)}{2}, \alpha^\star, \beta^\star) < 0$, we get

$$h(\alpha^\star, \beta^\star) = Q_{a,d,p}(x_1(\alpha^\star, \beta^\star)) + 1 - Q_{a,d,p}(x_2(\alpha^\star, \beta^\star)).$$

Therefore, we obtain

$$\begin{aligned} h_1(\alpha^\star, \beta^\star) = q\left(x_1(\alpha^\star, \beta^\star), a, d, p\right) & \left(\frac{1}{2\beta^\star} - \frac{1}{2}\left(\left(\frac{\alpha^\star}{2\beta^\star}\right)^2 - \alpha^\star \log\left(\frac{\alpha^\star}{2(\beta^\star)^2}\right)\right)^{-1/2}\right. \\ & \left.\left(\frac{\alpha^\star}{2(\beta^\star)^2} - \log\left(\frac{\alpha^\star}{2(\beta^\star)^2}\right) - 1\right)\right) - q\left(x_2(\alpha^\star, \beta^\star), a, d, p\right)\left(\frac{1}{2\beta^\star} + \right. \\ & \left.\frac{1}{2}\left(\left(\frac{\alpha^\star}{2\beta^\star}\right)^2 - \alpha^\star \log\left(\frac{\alpha^\star}{2(\beta^\star)^2}\right)\right)^{-1/2}\left(\frac{\alpha^\star}{2(\beta^\star)^2} - \log\left(\frac{\alpha^\star}{2(\beta^\star)^2}\right) - 1\right)\right) \\ & \approx 0.004467893 \end{aligned}$$

as well as

$$h_2(\alpha^\star,\beta^\star) = q\left(x_1(\alpha^\star,\beta^\star),a,d,p\right)\left(-\frac{\alpha^\star}{2(\beta^\star)^2}-\frac{1}{2}\left(\left(\frac{\alpha^\star}{2\beta^\star}\right)^2-\alpha^\star\log\left(\frac{\alpha^\star}{2(\beta^\star)^2}\right)\right)^{-1/2}\right.$$
$$\left.\left(-\frac{(\alpha^\star)^2}{2(\beta^\star)^3}+\alpha^\star\frac{2}{\beta^\star}\right)\right) - q\left(x_2(\alpha^\star,\beta^\star),a,d,p\right)$$
$$\left(-\frac{\alpha^\star}{2(\beta^\star)^2}+\frac{1}{2}\left(\left(\frac{\alpha^\star}{2\beta^\star}\right)^2-\alpha^\star\log\left(\frac{\alpha^\star}{2(\beta^\star)^2}\right)\right)^{-1/2}\left(-\frac{(\alpha^\star)^2}{2(\beta^\star)^3}+\alpha^\star\frac{2}{\beta^\star}\right)\right)$$
$$\approx -0.04957933.$$

Observing that h_1 and h_2 are continuous at $(\alpha^\star,\beta^\star)$, we have verified Assumption $B4$ explicitly.

The asymptotic variance of Clarke's test is given by

$$\sigma_\psi^2 = \mathrm{Var}\left(\mathbf{1}\{\psi(X,\alpha^\star,\beta^\star)>0\}+h_1(\alpha^\star,\beta^\star)X^2+\frac{h_2(\alpha^\star,\beta^\star)}{2}X\right)\approx 0.3475695.$$

Again, details of the tedious calculations can be found in Appendix E.

Remark 18
Since the asymptotic variance σ_ψ^2 is larger than $1/4$, we have shown that it is not a conservative approach to use $1/4$ as an estimator of σ_ψ^2.

7.2 Simulations

In this subsection, we investigate the finite sample behavior of the estimators introduced in Chapter 5. To this goal, we conduct a small study of the empirical levels and powers of Clarke's test for the examples from Section 7.1 and several other model comparisons introduced below. For each simulation study, two competing, non-nested density-families are chosen. Moreover, the true unknown distribution P is fixed in such a way that Assumptions $B1-B6$ are satisfied. We draw 1000 samples of size n from a distribution P with absolutely continuous density p w.r.t. the Lebesgue measure and we compute $\hat{B}_n$, $(\hat{h}_{1,n,i})_{1\le i\le d_\alpha}$, $(\hat{h}_{2,n,i})_{1\le i\le d_\beta}$ and $\hat{\sigma}_\psi^2$ defined in Equations (4.1),(5.2),(5.3) and (5.4), respectively. For Examples 1 and 2, the theoretical values of $(h_{1,i})_{1\le i\le d_\alpha}$, $(h_{2,i})_{1\le i\le d_\beta}$ and σ_ψ^2 are known and we can assess the precision of their estimation.

Therefore, we proceed as follows:

1. Draw samples of size n from the distribution P, where $n\in\{50,100,250,500,$

$1000, 10000\}$.

2. Calculate $\hat{B}_n, (\hat{h}_{1,n,i})_{1\leq i\leq d_\alpha}, (\hat{h}_{2,n,i})_{1\leq i\leq d_\beta}$ and $\hat{\sigma}^2_\psi$.

3. Calculate $T_{1,n}$ (and other test statistics, if they are of interest).

In each simulation study, we obtain 1000 realizations of the estimators $\hat{B}_n$, $(\hat{h}_{1,n,i})_{1\leq i\leq d_\alpha}$, $(\hat{h}_{2,n,i})_{1\leq i\leq d_\beta}$ and $\hat{\sigma}^2_\psi$ for a sample size n. Under the null hypothesis, B_n is Binomial distributed with parameter $1/2$ and size n. Therefore, the empirical mean of the $\hat{B}_n$'s over 1000 samples estimates the mean of this Binomial distribution. Further, the empirical means of $(\hat{h}_{1,n,i})_{1\leq i\leq d_\alpha}, (\hat{h}_{2,n,i})_{1\leq i\leq d_\beta}$ and $\hat{\sigma}^2_\psi$ approximate their theoretical counterparts.

After repeating the procedure for various sample sizes n, we summarize the results in a corresponding table. For each simulated sample size, we present the mean of the 1000 realizations of the estimators. Additionally, we present the empirical level or power, i.e. the percentage of rejections, of each of the calculated test statistics. The significance level of the considered test is fixed at 5%. To investigate the stability of the considered estimators, we report their empirical variances in Appendix D. The following abbreviations are used in the tables below:

$T_{1,n}$ - the reference test statistic given in (5.5), i.e. $\big(\hat{B}_n - n/2\big)/\sqrt{n\hat{\sigma}^2_\psi}$.

$T_{2,n}$ - the test statistic $\big(\hat{B}_n - n/2\big)/\sqrt{n\sigma^2_\psi}$.

$T_{3,n}$ - the test statistic $\big(\hat{B}_n - n/2\big)/\sqrt{n/4}$.

Due to Theorem 5.0.5, the parameter $e(n)$ has to be chosen in order to estimate the asymptotic variance σ^2_ψ from (5.1). The bandwidth $e(n)$ essentially determines the rate of convergence of the partial derivative estimators $\hat{h}_{1,n}$ and $\hat{h}_{2,n}$. However, there is a trade off between accuracy and shrinkage of the error terms. On the one hand, the closer we choose $e(n)$ to $1/\sqrt{n}$, the faster the estimators of the partial derivatives converge. On the other hand, the shrinkage factor of the (asymptotically) Normal distributed error terms is $1/(\sqrt{n}e(n))$, due to Equation (A.1). Therefore, the shrinkage factor is close to 1 if $e(n)$ is close to $1/\sqrt{n}$. To find an optimal $e(n)$, we computed the estimators $\hat{h}_{1,n}, \hat{h}_{2,n}$ and $\hat{\sigma}^2_\psi$ for the grid of values $n^{-1/2.5}, n^{-1/3}, n^{-1/3.5}, n^{-1/4}, n^{-1/5}$. After investigating the results, we chose $e(n) = n^{-1/3}$, since it resulted in the smallest average empirical variance of the estimators.

7.2.1 Simulations For Example 1

We present our numerical results for Example 1, choosing the standard deviations of the competing models as $\sigma_g = 1/2$ and $\sigma_f \approx 0.98$. Figure 7.1 displays 1000 realizations of $T_{1,n}$ (left plot) and of the test statistic $T_{2,n}$ (right plot) together with their respective boxplots.

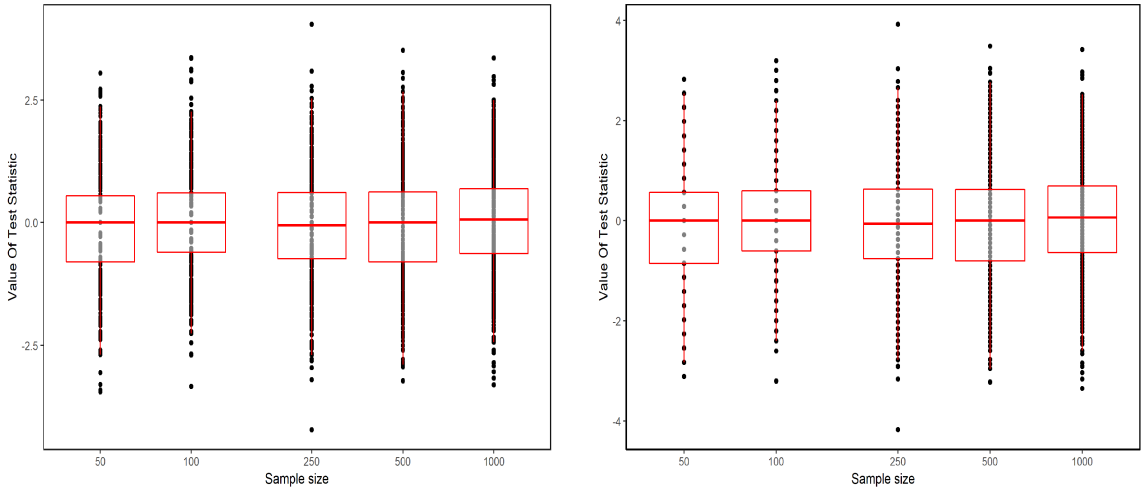

Figure 7.1: Plot of the simulated test statistics for Example 1. On the left is the plot of $T_{1,n}$ and on the right is the plot of the test statistic $T_{2,n}$

Since $\mathcal{H}_0$ is satisfied for Example 1, both statistics are asymptotically standard Normal distributed, which is partially confirmed by the boxplots. All observed values of the test statistics lie in $[-4, 4]$, showing that no outliers are present. The empirical 25% and 75% quantiles of the data lie in $[-1, 1]$, which indicates that most of the probability mass in located around 0.

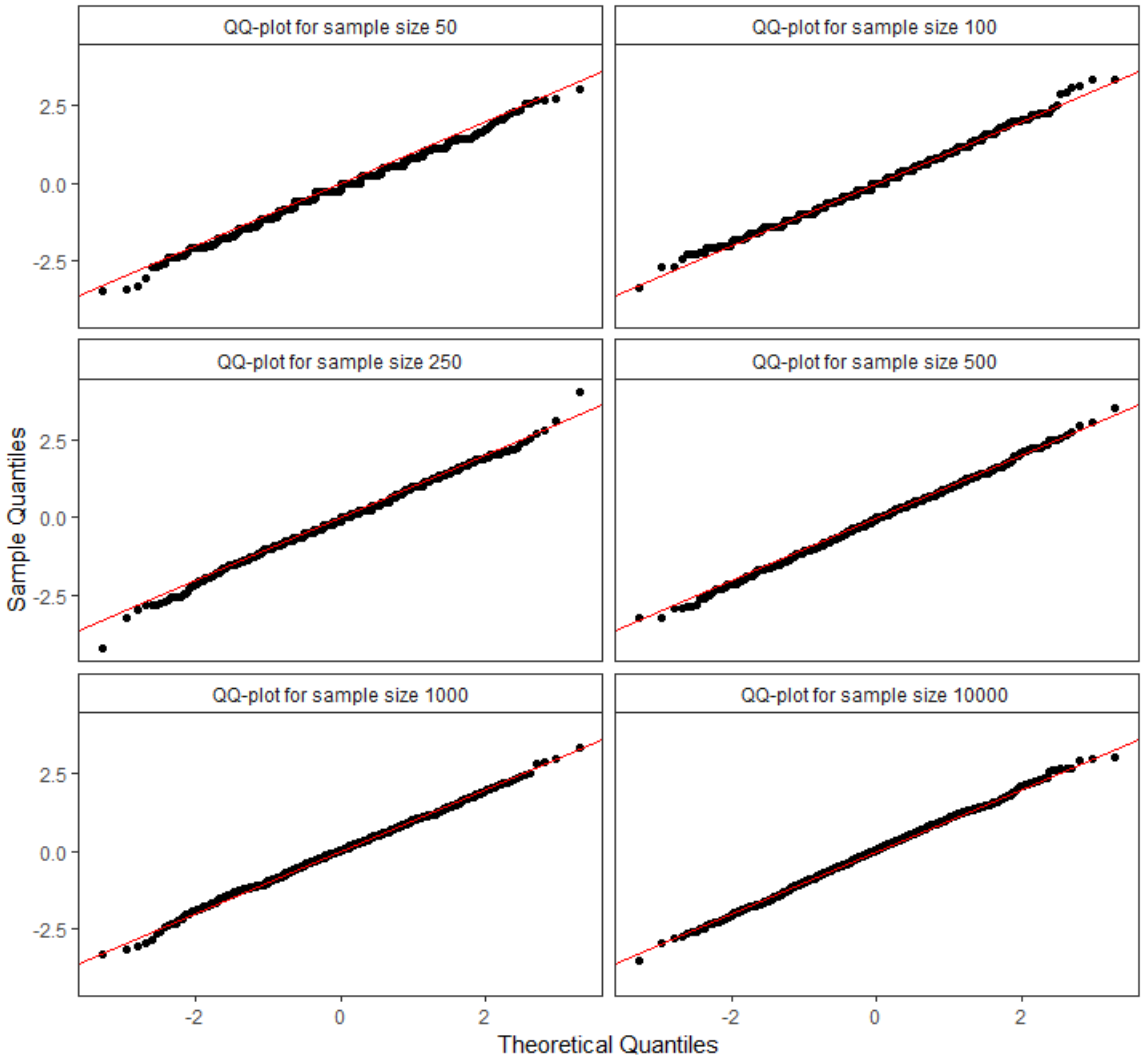

Figure 7.2: QQ-plot of the observed $T_{1,n}$.

To further verify the assumption of Normality, we present the QQ-plots of $T_{1,n}$ in Figure 7.2 for sample sizes $50, 100, 250, 500, 1000$ and 10000. The red line is the 45 degree line. Surprisingly, the QQ-plots display an accurate distributional approximation even for small sample sizes. The slight step patterns in the QQ-plots for small sample sizes are due to the discrete values of $\hat{B}_n$. This results in finitely many possible values of the estimators $\hat{h}_{1,i}, \hat{h}_{2,i}$ and $\hat{\sigma}^2_\psi$, explaining discrete values of $T_{1,n}$. Altogether, the QQ-plots confirm the asymptotic normality of $T_{1,n}$.

Table 7.1 summarizes further simulation results. We observe that the empirical mean of $\hat{B}_n$ is very close to the true value $n/2$ and that the mean of the estimated variance $\hat{\sigma}^2_\psi$ is close to the true value $1/4$, for every sample size n. The mean of the estimated variance $\hat{\sigma}^2_\psi$ converges with increasing number of observations. The empirical mean of the partial derivative estimates are also close to their true value 0. The empirical levels of

test statistics $T_{1,n}$ and $T_{2,n}$ are close to the (asymptotic) significance level 5%. Table D.1 in Appendix D shows the low variability of the estimators, which are spread around their true values. It should be noted that the test statistic $T_{3,n}$ is identical to $T_{2,n}$ for Example 1.

	Empirical Mean Of				Empirical Level Of	
n	$\hat{B}_n$	$\hat{h}_1$	$\hat{h}_2$	$\hat{\sigma}^2_\psi$	$T_{1,n}$	$T_{2,n}$
50	24.62	-0.003058	0.000589	0.271854	0.0610	0.0760
100	49.80	-0.000186	0.004850	0.263450	0.0530	0.0590
250	124.31	-0.000491	0.002835	0.257653	0.0600	0.0620
500	249.68	-0.000302	-0.000706	0.254543	0.0670	0.0720
1000	499.23	0.000065	0.003345	0.253132	0.0420	0.0450
10000	5000.65	0.000386	-0.001303	0.250676	0.0500	0.0500

Table 7.1: Empirical mean and empirical level of the estimators for Example 1 under H_0.

7.2.2 Simulations For Example 2

For Example 2, we simulate 1000 samples with the parameters from Section 7.1, namely $\alpha_2 = \beta_2 = 2$, $a = 1/2$ and $d = 3$. Figure 7.3 shows the corresponding boxplots of $T_{1,n}$ (left plot) and $(\hat{B}_n - n/2)/\sqrt{n\sigma^2_\psi}$ (right plot).

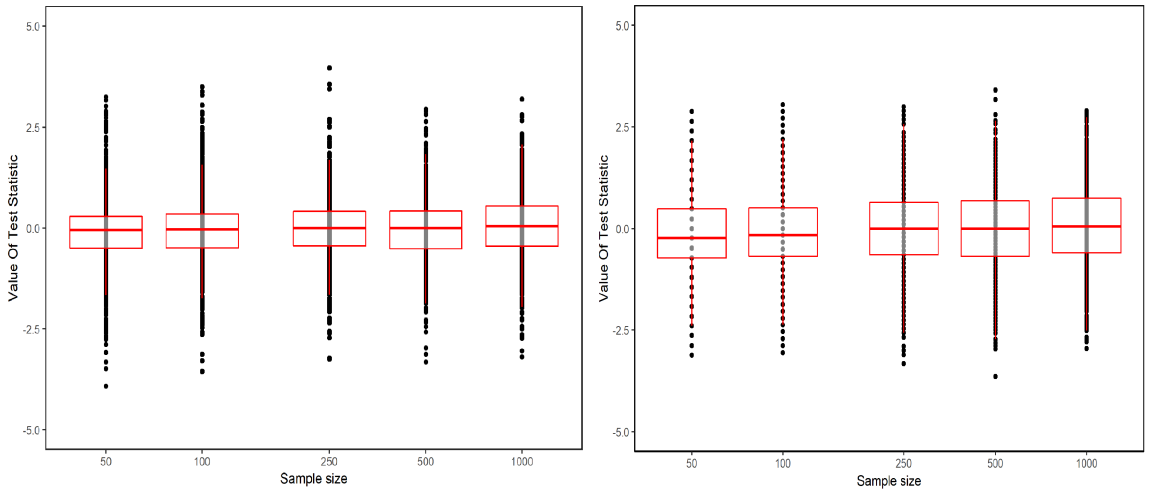

Figure 7.3: Plot of the simulated test statistics for Example 2. On the left is the plot of $T_{1,n}$ and on the right is the plot of the test statistic $T_{2,n}$.

The null hypothesis is satisfied for this example and therefore both test statistics are asymptotically standard Normal distributed. Most of the observed values of both test statistics lie in $[-4, 4]$, but some outliers are present in the left plot. The empirical 25% and 75% quantiles of the observations lie in $[-1, 1]$, which indicates that most of the

probability mass is located around 0. The quantiles in the left plot seem to be smaller than the quantiles in right plot, even though outliers are present in the left plot.

In the case of $T_{1,n}$, it is interesting to present the QQ-plot to check the assumption of Normality.

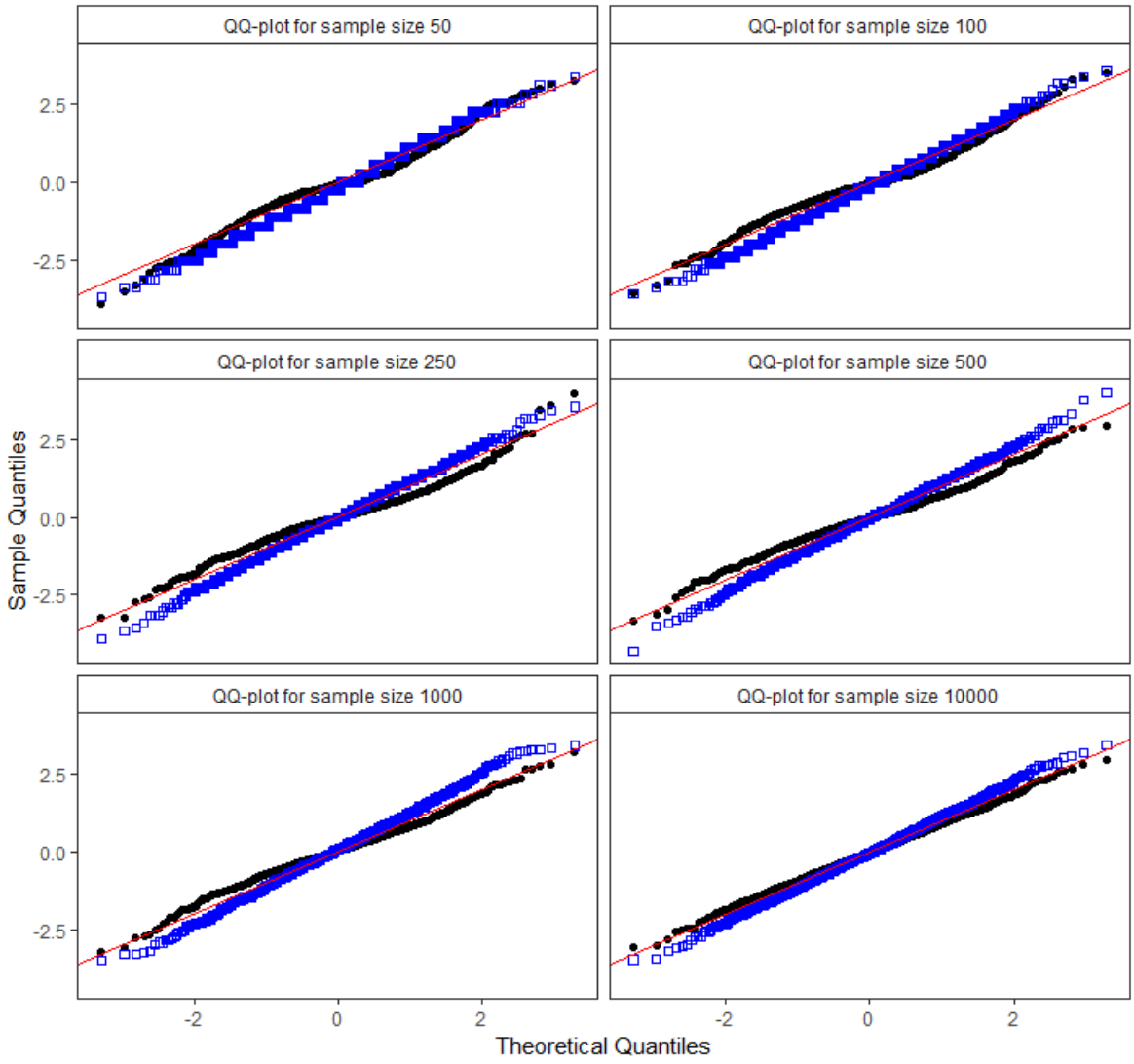

Figure 7.4: QQ-plot of $T_{1,n}$ (black dots) and the QQ-plot of $T_{3,n}$ (blue dots).

Figure 7.4 shows the QQ-plot of $T_{1,n}$ with black dots and test statistic $T_{3,n}$ with blue dots for each sample size. The red line is the 45 degree line. We observe that the black dots exhibit a S-shaped form around the red line. Further, we see more observations around the origin as we theoretically expect resulting in the S-shape of the black dots. The accumulation of observations close to 0 is due the fact that $T_{1,n}$ suffers from some unusual high variance estimates. These high variance estimates shrink the term $(\hat{B}_n - n/2)/\hat{\sigma}_\psi^2$ towards 0. The distributional approximation of $T_{1,n}$ is improved if we increase the sample

size.

The QQ-plot of the test statistic $T_{3,n}$ visualizes that the asymptotic variance proposed by Clarke is too small for Example 2. Indeed, the blue dots are consistently below the red line for quantiles less than 0 and above the red line for quantiles larger than 0. This resembles the fact that the asymptotic variance of test statistic $T_{3,n}$ is larger than 1. The effect is not strongly visible, since the slope of the QQ-line for the blue dots is approximately 1.18, which is close to the slope of the red line.

Further simulation results are summarized in Table 7.2. It can be observed that the empirical mean of the $\hat{B}_n$'s is close to the true value $n/2$ for all sample sizes. The empirical means of the partial derivative estimates $\hat{h}_1$ and $\hat{h}_2$ are close to their true values from sample sizes $n \geq 250$ on. However, the empirical mean of the estimated variances is much larger than the true value $\sigma_\psi^2 \approx 0.348$. This behavior explains the S-shape of the black dots in Figure 7.4. Table D.2 in Appendix D additionally shows that the variability of the variance estimator is huge for small sample sizes and still significantly large for sample size 10 000.

We can give two possible explanations for this phenomenon. First, the partial derivatives $h_1 \approx 0.004$ and $h_2 \approx -0.05$ are very small, but have a large impact on the variance. Small deviations from the true values of h_1 and h_2 can lead to a large change in the asymptotic variance. Therefore, the estimated variance can vary largely due to small inaccuracies in the partial derivative estimates. The second explanation is that $e(n)$ was chosen to be equal to $n^{-1/3}$ for both partial derivative estimates $\hat{h}_1$ and $\hat{h}_2$. For $n = 100$, $n^{-1/3} \approx 0.22$ is relatively small in comparison to $\alpha^\star$, but still relatively large in comparison to $\beta^\star$. To validate this explanation, we conducted another simulation study and used the theoretical values h_1 and h_2 instead of $\hat{h}_1$ and $\hat{h}_2$ in the calculation of $\hat{\sigma}_\psi^2$. The empirical mean of the resulting variance estimates is very close to the true variance across all sample sizes and confirms our statements.

Furthermore, we observe that the empirical level of $T_{1,n}$ deviates from 5% a bit more in comparison to Table 7.1. The empirical level of $T_{2,n}$ is close to 5% for every sample size and this indicates that $T_{1,n}$ suffers from outliers of the variance estimate. The last column of Table 7.2 shows the empirical level of $T_{3,n}$, which is based on the original Clarke's test statistic $\hat{B}_n$. We observe that the original approach from Clarke is too conservative, as we reject the null hypothesis in about 10% of the cases. This result confirms the theoretical results of Chapter 5, since it shows that the proposed variance in [6] is false. Therefore, the approach by Clarke can lead to misleading results and it is advisable to use the approach proposed in this master thesis. Moreover, the deviation of the level of Clarke's test explains the observation on the blue dots in Figure 7.3 and the statement in Remark

18. Summarizing, we can say that the test based on $T_{1,n}$ keeps its level better than Clarke's test based on $T_{3,n}$.

	Empirical Mean Of				Empirical Level Of		
n	$\hat{B}_n$	$\hat{h}_1$	$\hat{h}_2$	$\hat{\sigma}^2_\psi$	$T_{1,n}$	$T_{2,n}$	$T_{3,n}$
50	24.42	0.003389	-0.051429	3.393517	0.0700	0.0470	0.1220
100	49.58	0.003365	-0.052404	2.244614	0.0400	0.0440	0.0910
250	124.61	0.004700	-0.051065	1.501887	0.0380	0.0550	0.1000
500	249.49	0.004516	-0.050654	1.044610	0.0290	0.0430	0.0940
1000	499.55	0.004525	-0.051370	0.860713	0.0390	0.0600	0.1010
10000	5000.91	0.004405	-0.049034	0.446276	0.0480	0.0540	0.1140

Table 7.2: Empirical mean and empirical level of the estimators for Example 2 under H_0.

7.3 Empirical Power Study

We present our results of the empirical power study for several model comparisons that are introduced hereafter. For each example, we introduce the competing density families together with the true distribution P and estimate $(\hat{h}_{1,n,i})_{1\leq i\leq d_\alpha}, (\hat{h}_{2,n,i})_{1\leq i\leq d_\beta}$ and $\hat{\sigma}^2_\psi$ to compute $T_{1,n}$. Note that we cannot investigate the behavior of $T_{2,n}$, because the computation of the theoretical variance σ^2_ψ is very difficult. Moreover, we do not consider $T_{3,n}$, since its asymptotic distribution is not standard Normal under $\mathcal{H}_0$. This implies that $T_{3,n}$ does not keep its level under $\mathcal{H}_0$, which would be a necessary requirement to investigate its power.

We summarize the empirical results with the empirical mean of the estimators $(\hat{h}_{1,n,i})_{1\leq i\leq d_\alpha}, (\hat{h}_{2,n,i})_{1\leq i\leq d_\beta}$ and $\hat{\sigma}^2_\psi$ and the empirical power of $T_{1,n}$. Note that the empirical power of $T_{1,n}$ is displayed as the percentage of "correct" rejections. This means that the critical region is defined as $[1.96, \infty]$ if $P\left(\log\left(f(X,\alpha^\star)/g(X,\beta^\star)\right) > 0\right) > 1/2$ and as $[-\infty, -1.96]$ if $P\left(\log\left(f(X,\alpha^\star)/g(X,\beta^\star)\right) > 0\right) < 1/2$. If we do not classify rejections as "correct" or "incorrect" then the reported empirical powers would be even slightly higher. Since this gap is negligible, we have preferred to consider unilateral critical regions.

We start with a simple model comparison. We compare a Normal distribution with variance 2 against a Normal distribution with variance 3. The underlying distribution P is a standard Normal distribution. It is intuitively clear that the model with variance 2 should be considered as the "better" model. Table 7.3 shows the simulation results. The test consistently prefers the Normal distribution with variance 2, since the empirical mean of $\hat{B}_n$ is larger than $n/2$. In fact, the null hypothesis was rejected in 100% of the cases in favor of the model with variance 2, for all sample sizes. Note, that in this example $h_1 = h_2 = 0$

by Remark 17 and we can asses the precision of the estimators $\hat{h}_1$ and $\hat{h}_2$. The empirical means of $\hat{h}_1$ and $\hat{h}_2$ are close to 0 across all sample sizes indicating their unbiasedness. Table D.3 shows the low variability of all estimators, illustrating stable estimation results for all sample sizes.

	Empirical Mean Of				Emp. Pow.
n	$\hat{B}_n$	$\hat{h}_1$	$\hat{h}_2$	$\hat{\sigma}^2_\psi$	$T_{1,n}$
50	44.21	0.002953	0.000037	0.112017	1
100	88.24	-0.001625	-0.001114	0.108880	1
250	220.54	0.003969	-0.004044	0.106807	1
500	441.02	-0.000913	-0.001389	0.105710	1
1000	881.72	0.000345	0.000155	0.105252	1
10000	8809.76	0.000486	-0.000207	0.105095	1

Table 7.3: Empirical mean and empirical power of the estimators in the case $\mathcal{N}(\alpha, 2)$ vs. $\mathcal{N}(\beta, 3)$ and $P = \mathcal{N}(0, 1)$.

Following this simple example, we continue with more complex model comparisons. First, we compare the family of Normal distributions against the family of Laplace distribution by considering the quotient of densities

$$\frac{\frac{1}{\sqrt{2\pi\alpha_1}} \exp\left(-\frac{(x-\alpha_2)^2}{2\alpha_1}\right)}{\frac{1}{2\beta_1} \exp\left(-\frac{|x-\beta_2|}{2\beta_1}\right)}. \tag{7.1}$$

The true distribution P is chosen as a t-distribution with 15 degrees of freedom. Furthermore, we choose the standard maximum likelihood estimators of the respective distributions as estimators of $\alpha^\star$ and $\beta^\star$.

The results are summarized in Table 7.4. The test prefers the family of Normal distributions, since the empirical mean of $\hat{B}_n$ is consistently larger than $n/2$. The empirical means of the estimators $\hat{h}_{1,1}$ and $\hat{h}_{2,1}$ are close to 0, for all sample sizes. Additionally, we observe that the empirical means of the estimators $\hat{h}_{1,2}$ and $\hat{h}_{2,2}$ are stable over all sample sizes. However, the empirical mean of the variance estimator $\hat{\sigma}^2_\psi$ is only stable for sample sizes $n \geq 250$. Table D.4 in Appendix D confirms this observation, since the empirical variance of $\hat{\sigma}^2_\psi$ is strongly decreasing for increasing sample size. The empirical power of $T_{1,n}$ increases with the sample size. For sample size 500, already more than 98% of the tests reject $\mathcal{H}_0$ in favor of the superior family of Normal distributions.

	Empirical Mean Of						Emp. Pow.
n	$\hat{B}_n$	$\hat{h}_{1,1}$	$\hat{h}_{1,2}$	$\hat{h}_{2,1}$	$\hat{h}_{2,2}$	$\hat{\sigma}^2_\psi$	$T_{1,n}$
50	30.29	-0.005121	-0.163571	0.007221	0.276634	0.350380	0.2550
100	60.10	0.004943	-0.161759	0.001601	0.272670	0.330764	0.4070
250	149.91	-0.001462	-0.166385	0.002759	0.272445	0.319871	0.8120
500	300.54	-0.002183	-0.163185	0.000127	0.266858	0.310164	0.9820
1000	599.54	-0.000435	-0.166415	0.001965	0.266085	0.308186	1.0000

Table 7.4: Empirical mean and empirical power of the estimators in the case of Normal vs. Laplace distribution and P equal to a t distribution with 15 degrees of freedom.

As a last example, we compare a Gamma distribution against an Exponential distribution. The Gamma distribution is parametrized as in Section 7.1.2 and the parametrization of the Exponential distribution is as follows:

$$e_\beta(x) = \frac{1}{\beta} \exp^{-\frac{x}{\beta}} .$$

Thus, we consider the quotient of density families

$$\frac{\frac{1}{\alpha_1^{\alpha_2}\Gamma(\alpha_2)} x^{\alpha_2 - 1} \exp\left(-\frac{1}{\alpha_1} x\right)}{\frac{1}{\beta} \exp^{-\frac{x}{\beta}}} \mathbf{1}_{\{x>0\}}. \tag{7.2}$$

The underlying distribution P is chosen to follow an Exponential distribution with mean 1. Fixing the parameter $\alpha_2 \neq 1$ of the Gamma density family ensures that Assumption $B3$ is satisfied. Now, one can easily check that Assumptions $B1 - B6$ are satisfied. For this simulation we choose $\alpha_2 = 2$, which implies that $\hat{\alpha}_n = (n)^{-1} \sum_{i=1}^n X_i/2$ is the same estimator as in Section 7.1.2. The pseudo maximum likelihood estimator of β is given by $\hat{\beta} := \bar{X}$. Note, that the test should clearly prefer the Exponential distribution, since the true distribution can be perfectly approximated by the Exponential density family.

The simulation results are summarized in Table 7.5. We observe that Clarke's test is not able to identify the true underlying Exponential distribution, since the empirical mean of $\hat{B}_n$ is consistently larger than $n/2$. In fact, it consistently prefers the Gamma distribution over all sample sizes. Note that the empirical power of Clarke's test is increasing with the sample size, since the rejections in favor of the Gamma distribution are mathematically correct. Indeed, the median of $\psi(X, \alpha^\star, \beta^\star)$ is greater than $1/2$ in this example, even though the identification of the true model would require a median less than $1/2$. Since Clarke's test is only extracting information about the median of $\psi(X, \alpha^\star, \beta^\star)$, it discards useful information. E.g. a huge positive observation of $\psi(X_i, \hat{\alpha}_n, \hat{\beta}_n)$ has exactly the same influence on the statistic $\hat{B}_n$ as small positive observation of $\psi(X_i, \hat{\alpha}_n, \hat{\beta}_n)$. However, in

the case of a huge observation of $\psi(X_i, \hat{\alpha}_n, \hat{\beta}_n)$ density f is a lot more "likely" than density g, whereas in the case of a small positive observation of $\psi(X_i, \hat{\alpha}_n, \hat{\beta}_n)$ both models f and g are almost equally "likely". This "oversimplification" of the test statistic leads to the drawback, that in similar competing models, the "wrong" model may be chosen. This is the price one needs to pay for simplifying the test statistic in comparison to the test statistic presented in [20].

We have simulated the same example for several other values of α_2 and observed that for $\alpha_2 \geq 3$ and $\alpha_2 \leq 1/2$ Clarke's test consistently prefers the true Exponential distribution. Additionally, we have simulated Vuong's test with $\alpha_2 = 2$ and obtained an empirical power of almost 100% for sample sizes $n \geq 250$. This shows that Vuong's test outperforms Clarke's test for this example, since it consistently prefers the true Exponential distribution. Apart from selecting the "wrong" model, we observe stable estimators across all sample sizes, which is confirmed by the low variability of the estimators, shown in Table D.5 in Appendix D.

	Empirical Mean Of				Emp. Pow.
n	$\hat{B}_n$	$\hat{h}_1$	$\hat{h}_2$	$\hat{\sigma}^2_\psi$	$T_{1,n}$
50	29.41	0.001032	-0.023062	0.255034	0.2510
100	58.65	-0.040359	-0.023301	0.253163	0.4120
250	146.24	-0.039814	-0.013481	0.248716	0.7780
500	291.57	-0.027105	-0.004516	0.246785	0.9680
1000	583.36	-0.017720	-0.006535	0.245561	1.0000

Table 7.5: Empirical mean and empirical power of the estimators in the case of a Gamma vs. Exponential distribution under an Exponential distribution with mean 1.

Supplementary to the examples presented above, a further investigation of the power of $T_{1,n}$ is provided Appendix C.

7.4 Bootstrap Simulations

In this subsection, we investigate the properties of a bootstrap estimation procedure of the asymptotic variance σ^2_ψ and the empirical power of a modified version of $T_{1,n}$. We apply Efron's Bootstrap scheme, corresponding to multinomial weights $\xi_{i,n}$. Obviously, multinomial weights satisfy the conditions stated in Assumptions $W1 - W4$. In addition to Assumptions $B1 - B6$, Assumptions $B7$ and $B9$ are satisfied for all model comparisons considered below. Note that we do not need to verify Assumption $B8$, due to Remark 16.

The study is conducted similarly to the empirical power study in Section 7.3. The only difference is that we replace steps 2. and 3. by the following procedure:

2.′ Calculate $\hat{B}_n$. Draw B bootstrap samples of size n from the sample obtained in step 1. For each bootstrap sample i, calculate $\tilde{B}_{n,i}$ according to (5.6) to obtain B bootstrap replicates of $\hat{B}_n$. Define σ_B^2 as the sample variance of $\left(n^{-1/2}(\tilde{B}_n - \hat{B}_n)\right)_{1\leq i\leq B}$.

3.′ Calculate the test statistic

$$\frac{\hat{B}_n - \frac{n}{2}}{\sqrt{n\hat{\sigma}_B^2}} =: T_{4,n}. \tag{7.3}$$

Note, that the test statistic presented in (7.3) is approximately asymptotically standard Normal distributed. To see the claim, assume that the following assumption is satisfied.

Assumption $B10$ *Define*

$$f_B\left(\frac{\left(\tilde{B}_n^1 - \hat{B}_n\right)}{\sqrt{n}}, ..., \frac{\left(\tilde{B}_n^B - \hat{B}_n\right)}{\sqrt{n}}\right)$$
$$:= \left|\frac{1}{B}\sum_{i=1}^{B}\left(\frac{1}{\sqrt{n}}\left(\tilde{B}_n^i - \hat{B}_n\right) - \frac{1}{B}\sum_{i=1}^{B}\frac{1}{\sqrt{n}}\left(\tilde{B}_n^i - \hat{B}_n\right)\right)^2 - \sigma_\psi^2\right|,$$

where $\tilde{B}_n^i$ is the bootstrapped version of $\hat{B}_n$ calculated from the i-th (independent) bootstrap sample. Assume that for all $\epsilon > 0$ there exists $K \in \mathbb{R}$ such that

$$\limsup_{B,n\to\infty} \mathbb{E}^*_{\mathbb{P}_{XW}}\left[f_B\left(\frac{\left(\tilde{B}_n^1 - \hat{B}_n\right)}{\sqrt{n}}, ..., \frac{\left(\tilde{B}_n^B - \hat{B}_n\right)}{\sqrt{n}}\right)\right.$$
$$\left.\mathbf{1}\left\{f_B\left(\frac{\left(\tilde{B}_n^1 - \hat{B}_n\right)}{\sqrt{n}}, ..., \frac{\left(\tilde{B}_n^B - \hat{B}_n\right)}{\sqrt{n}}\right) > K\right\}\right] < \epsilon.$$

Using this assumption, we can prove the following lemma.

Lemma 7.4.1
Under Assumptions $B1 - B10$, we have

$$\lim_{B\to\infty}\lim_{n\to\infty}\mathbb{P}^*_{XW}\left(\left|\frac{1}{B}\sum_{i=1}^{B}\left(\frac{1}{\sqrt{n}}\left(\tilde{B}_n^i - \hat{B}_n\right) - \frac{1}{B}\sum_{i=1}^{B}\frac{1}{\sqrt{n}}\left(\tilde{B}_n^i - \hat{B}_n\right)\right)^2 - \sigma_\psi^2\right| > \epsilon\right) = 0.$$

The proof of Lemma 7.4.1 can be found in Appendix A.9.

In the following, we will always assume that Assumption $B10$ is satisfied. Additionally, we set $B = 500$. The results of the bootstrap simulations are summarized in tables, which contain the empirical level/power of test statistic $T_{4,n}$ and the empirical mean of σ_B^2, for each sample size n. Tables of the empirical variance of the estimator σ_B^2 can be found in Appendix D.

We start with the simulation of Example 1. The results are summarized in Table 7.6. We observe, that the empirical mean of σ_B^2 is slightly biased or suffers from outliers. Table D.6 in Appendix D shows that the variability of σ_B^2 is small. This indicates that the bootstrap variance estimate σ_B^2 is indeed biased for small sample sizes. However, for increasing sample size the empirical mean of σ_B^2 converges to the true value 1/4. The empirical level of test statistic $T_{4,n}$ is close to 5% for sample sizes $n \geq 100$.

	Empirical Mean Of	Empirical Level Of
n	$\hat{\sigma}_B^2$	$T_{4,n}$
50	0.301268	0.0340
100	0.285053	0.0410
250	0.273778	0.0410
500	0.266621	0.0480
1000	0.261062	0.0450

Table 7.6: Empirical mean and empirical level of the bootstrap estimators for Example 1.

Following Example 1, we present the results for Example 2, which are summarized in Table 7.7. Again, we can observe that the empirical mean of $\hat{\sigma}_B^2$ is slightly biased or suffers from outliers. However, $\hat{\sigma}_B^2$ seems to be better than $\hat{\sigma}_\psi^2$ for Example 2, since the empirical mean of the estimator $\hat{\sigma}_B^2$ is much closer to the true value σ_ψ^2 than the empirical mean of $\hat{\sigma}_\psi^2$. Table D.7 in Appendix D confirms this observation, since the variability of $\hat{\sigma}_B^2$ is very small in comparison to the variability of $\hat{\sigma}_\psi^2$. This again indicates that $\hat{\sigma}_B^2$ is stable, but slightly biased.

The empirical level of the $T_{4,n}$ is close to the approximate asymptotic level of 5%. Therefore, $T_{4,n}$ yields better results than $T_{1,n}$, which is due to the stability of the estimator $\hat{\sigma}_B^2$.

n	Empirical Mean Of $\hat{\sigma}_B^2$	Empirical Level Of $T_{4,n}$
50	0.378511	0.0500
100	0.377113	0.0620
250	0.370148	0.0560
500	0.364260	0.0460
1000	0.362580	0.0500

Table 7.7: Empirical mean and empirical level of the bootstrap estimators for Example 2.

In the following, we present the results for the examples from Section 7.3. We begin with the two competing Normal distributions and compare our results, which are summarized in Table 7.8, with the results of Table 7.3. The results of both approaches are very similar. Both test statistics lead to an empirical power of 100% for each sample size n. Moreover, the empirical mean and variance of the estimators $\hat{\sigma}_B^2$ and $\hat{\sigma}_\psi^2$ are almost identical.

n	Empirical Mean Of $\hat{\sigma}_B^2$	Empirical Power $T_{4,n}$
50	0.114357	1.0000
100	0.113633	1.0000
250	0.111407	1.0000
500	0.109828	1.0000
1000	0.108122	1.0000

Table 7.8: Empirical mean and empirical power of the bootstrap estimators in the case of $\mathcal{N}(0,2)$ vs. $\mathcal{N}(0,3)$ under a standard Normal distribution.

Next, we present the model comparison introduced in (7.1). The results are summarized in Table 7.9. The empirical mean of $\hat{\sigma}_B^2$ is slightly too high for small sample sizes. However, the same behavior can be observed for the estimator $\hat{\sigma}_\psi^2$ in Table 7.4. The empirical power of $T_{4,n}$ increases with the sample size and is very similar to the empirical power of $T_{1,n}$.

n	Empirical Mean Of $\hat{\sigma}_B^2$	Empirical Power $T_{4,n}$
50	0.349600	0.2320
100	0.346448	0.4130
250	0.334506	0.8050
500	0.321654	0.9760
1000	0.315343	1.0000

Table 7.9: Empirical mean and empirical power of the bootstrap estimators in the case of Normal vs. Laplace distribution, where P is equal to a t distribution with 15 degrees of freedom.

As a last example we present the model comparison introduced in (7.2). The results are

summarized in Table 7.10. Once more, we observe that the variance estimate seems to be slightly too high, for small sample sizes. Table D.10 shows that the variability of $\hat{\sigma}_B^2$ is low, again confirming that $\hat{\sigma}_B^2$ is slightly biased. The empirical power of $T_{4,n}$ is slightly lower than the empirical power of $T_{1,n}$, but roughly in the same range.

	Empirical Mean Of	Empirical Power
n	$\hat{\sigma}_B^2$	$T_{4,n}$
50	0.280097	0.2190
100	0.269428	0.3860
250	0.260587	0.7420
500	0.255436	0.9600
1000	0.252380	0.9990

Table 7.10: Empirical mean and empirical power of the bootstrap estimators in the case of a Gamma vs. Exponential distribution under an Exponential distribution with mean 1.

Altogether, we observe that the empirical mean of $\hat{\sigma}_B^2$ is slightly biased for small sample sizes, but converges with increasing sample size. If $\hat{\sigma}_\psi^2$ does not suffer from outliers, it tends to be closer to the true value σ_ψ^2 than $\hat{\sigma}_B^2$. In these cases, the empirical variance of both estimators $\hat{\sigma}_\psi^2$ and $\hat{\sigma}_B^2$ is quite low. However, in comparison to $\hat{\sigma}_\psi^2$, $\hat{\sigma}_B^2$ does not suffer from outliers in Example 2, showing that it can be a more reliable estimator.
The empirical level/power of $T_{4,n}$ is very similar to the empirical level/power of $T_{1,n}$. In case of Example 2, the empirical level of $T_{4,n}$ is closer to 5% than the level of $T_{1,n}$, again indicating that $T_{4,n}$ is not as sensitive to outliers as $T_{1,n}$.

Chapter 8

Summary And Outlook

In this master thesis, we have revisited Clarke's test and derived its correct asymptotic distribution and variance. Based on this result, we have proposed two (approximately) asymptotically standard Normal distributed test statistics for non-nested model selection.

First, we have shown that the claimed Binomial distribution of Clarke's test statistic $\hat{B}_n$ is incorrect. In Chapter 5, we have proven that $n^{-1/2}(\hat{B}_n - 1/2)$ is asymptotically Normal distributed. Additionally, we have shown that the asymptotic variance of $n^{-1/2}(\hat{B}_n - 1/2)$ can be estimated either by Theorem 5.0.5 or via a bootstrap approach. These results can be used to formulate the (approximately) asymptotically standard Normal distributed test statistics $T_{1,n}$ and $T_{4,n}$. As a side result, we have stated sufficient conditions for $\mathcal{F}_\gamma$ to be a *Donsker* class and we have shown that the bootstrapped pseudo maximum likelihood estimator converges to the pseudo-true value in outer probability. Furthermore, we have clarified the asymptotic relation of B_n to $\hat{B}_n$, showing that their scaled difference is asymptotically Normal distributed. The comparison of Vuong's test with the test proposed in this master thesis yields that both approaches require similar assumptions. However, our approach seems to be slightly more general, since it only requires mild moment conditions and is still applicable if a proposed density is not differentiable w.r.t. its parameters. In Chapter 7, we presented two examples of competing densities, which satisfy the null hypothesis $\mathcal{H}_0$ and Assumptions $B1 - B9$. We illustrated these results in a simulation study, showing that the estimator $\hat{\sigma}^2_\psi$ may suffer from outliers. However, the empirical levels of $T_{1,n}$ and $T_{4,n}$ do not strongly deviate from their theoretical level. In the empirical power study, we saw that the proposed test statistics $T_{1,n}$ and $T_{4,n}$ yield similar results in terms of their empirical power. It should be noted that the proposed bootstrap estimator of σ^2_ψ seems to be slightly biased for smaller sample sizes, but does not suffer from outliers in any of the presented examples. Therefore, there is no clear recommendation to prefer one of the proposed variance estimators over the other.

In further investigations, it would be useful to extend our test to multiple competing densities, since we only considered the case of pairwise density comparisons. Furthermore, it is desirable to relax the assumption of strict non-nestedness to the non-nestedness assumption in [20]. Another interesting task is to investigate cases similar to Remark 17, i.e. cases in which the asymptotic variance of $n^{-1/2}(\hat{B}_n - n/2)$ is equal to 0.25. Looking at the results of the empirical power study, it seems to be the case that the partial derivative of h w.r.t. the mean of a proposed elliptical density is zero in case the true underlying distribution is symmetric around its mean. This would simplify some density comparisons since one would not need to estimate the asymptotic variance of Clarke's test statistic. Further, a mathematical justification for the choice of the shrinkage parameter $e(n)$ in the estimation of $\hat{\sigma}^2_\psi$ would prove valuable. Finally, it would be of great interest to find easily verifiable conditions to verify Assumption $B10$.

Appendix A

Proofs

A.1 Extension of our framework to identical marginal distributions of the covariates

Assume that $X = (Y, Z)$. Let $f(Y \mid Z, \alpha^\star)$ and $g(Y \mid Z, \beta^\star)$ be the conditional densities of Y given Z under both alternative models, and define $f_Z(Z, \alpha^\star)$ and $g_Z(Z, \beta^\star)$ as the marginal densities of Z under the "optimal" models for the distribution of X. Now, assume that the marginal distribution of Z is identical in both models, i.e $f_Z(Z, \alpha^\star) = g_Z(Z, \beta^\star) =: h_Z(Z)$. Moreover, the support of $h(\cdot)$ is the support of the true law of Z. Therefore, we have

$$\begin{aligned} &P\left(\log\left(\frac{f(X, \alpha^\star)}{g(X, \beta^\star)}\right) > 0\right) = P\left(\log\left(\frac{f(Y \mid Z, \alpha^\star) f_Z(Z, \alpha^\star)}{g(Y \mid Z, \beta^\star) g_Z(Z, \beta^\star)}\right) > 0\right) \\ &= P\left(\log\left(\frac{f(Y \mid Z, \alpha^\star)}{g(Y \mid Z, \beta^\star)}\right) > 0\right), \end{aligned}$$

which proves that the conditional and unconditional null hypotheses of Clarke's test are equivalent.

A.2 Proof of Theorem 5.0.1

Proof of (i):
Let $\gamma > 0$ such that $\mathcal{F}_\gamma$ is $P - Donsker$ and Assumptions $B1 - B5$ are satisfied. Under the null, we can write the normalized test statistic as

$$\frac{1}{\sqrt{n}}\left(\hat{B}_n - \frac{n}{2}\right) = \sqrt{n}(\mathbb{P}_n \phi_n - P\phi_n) + \sqrt{n}(P\phi_n - P\phi_\star)$$

$$= \mathbb{G}_n\phi_n + \sqrt{n}P(\phi_n - \phi_\star)$$
$$= \mathbb{G}_n(\phi_n - \phi_\star) + \mathbb{G}_n\phi_\star + \sqrt{n}P(\phi_n - \phi_\star).$$

Since $\mathcal{F}_\gamma$ is $P-Donsker$, we can use Corollary 2.3.12 in [19] to show that $\mathbb{G}_n(\phi_n - \phi_\star)$ is $o_{\mathbb{P}}(1)$. We first show that $\rho_P(\phi_n, \phi_\star) := \sqrt{P(\phi_n - \phi_\star)^2} \to 0$ $\mathbb{P}$-almost surely. Note that the pseudometric ρ_P slightly differs from the pseudometric in [19], but it is obvious that both metrics are equivalent in our case. By the continuous mapping theorem, $\psi(x, \hat{\alpha}_n, \hat{\beta}_n) \to \psi(x, \alpha^\star, \beta^\star)$ $\mathbb{P}$-almost surely for any $x \in \mathbb{R}$, since ψ is continuous on E_γ.

Therefore,

$$\mathbb{P}\left(P\left(\lim_{n\to\infty} \psi(X, \hat{\alpha}_n, \hat{\beta}_n) = \psi(X, \alpha^\star, \beta^\star)\right) = 1\right) = 1.$$

Note that $P\left(\lim_{n\to\infty} \psi(X, \hat{\alpha}_n, \hat{\beta}_n) = \psi(X, \alpha^\star, \beta^\star)\right) = \int \mathbf{1}\{\lim_{n\to\infty} \psi(X(\omega), \hat{\alpha}_n, \hat{\beta}_n) = \psi(X(\omega), \alpha^\star, \beta^\star)\} dP(\omega)$ is a random variable, due to the randomness induced by $\hat{\alpha}_n$ and $\hat{\beta}_n$. Furthermore, by the Dominated Convergence Theorem,

$$\begin{aligned}
&\mathbb{P}\left(\lim_{n\to\infty} \int (\phi_n - \phi_\star)^2 dP = 0\right) \\
&\geq \mathbb{P}\left(P\left(\lim_{n\to\infty} \mathbf{1}\{\psi(X, \hat{\alpha}_n, \hat{\beta}_n) > 0\} = \mathbf{1}\left\{\psi(X, \alpha^\star, \beta^\star) > 0\right\}\right) = 1\right) \\
&= \mathbb{P}\Big(P\left(\lim_{n\to\infty} \mathbf{1}\{\psi(X, \hat{\alpha}_n, \hat{\beta}_n) > 0\} = \mathbf{1}\left\{\psi(X, \alpha^\star, \beta^\star) > 0\right\}, \psi(X, \alpha^\star, \beta^\star) \neq 0\right) \\
&+ P\left(\lim_{n\to\infty} \mathbf{1}\{\psi(X, \hat{\alpha}_n, \hat{\beta}_n) > 0\} = \mathbf{1}\left\{\psi(X, \alpha^\star, \beta^\star) > 0\right\}, \psi(X, \alpha^\star, \beta^\star) = 0\right) = 1\Big) \\
&= \mathbb{P}\Big(P\Big(\lim_{n\to\infty} \mathbf{1}\{\psi(X, \hat{\alpha}_n, \hat{\beta}_n) > 0\} = \mathbf{1}\left\{\psi(X, \alpha^\star, \beta^\star) > 0\right\}, \psi(X, \alpha^\star, \beta^\star) \neq 0\Big) = 1\Big) \\
&= 1,
\end{aligned}$$

since $\psi(X, \alpha^\star, \beta^\star)$ has no probability mass at 0, by the definition of strictly non-nested models. Combining the arguments, we get $\rho_P(\phi_n, \phi_\star) \to 0$ $\mathbb{P}$-almost surely.

Next, choose some arbitrary $\epsilon > 0$ and $\nu > 0$. Choose $\delta > 0$ and n large enough such that the equicontinuity condition (2.1.8) from [19] is satisfied with

$$\mathbb{P}\left(\sup_{\rho_P(f-g)\leq\delta} |\mathbb{G}_n(f - g)| > \epsilon\right) \leq \nu,$$

and $\mathbb{P}\left(\rho_P(\phi_n, \phi_\star) > \delta\right) \leq \nu$. This yields

$$\begin{aligned}
\mathbb{P}\left(|\mathbb{G}_n(\phi_n - \phi_\star)| > \epsilon\right) &= \mathbb{P}\left(|\mathbb{G}_n(\phi_n - \phi_\star)| > \epsilon\ , \rho_P(\phi_n, \phi_\star) > \delta\right) \\
&+ \mathbb{P}\left(|\mathbb{G}_n(\phi_n - \phi_\star)| > \epsilon\ , \rho_P(\phi_n, \phi_\star) \leq \delta\right)
\end{aligned}$$

$$\leq \mathbb{P}\left(\rho_P(\phi_n, \phi_\star) > \delta\right) + \mathbb{P}\left(\sup_{\rho_P(f-g)\leq\delta} |\mathbb{G}_n(f-g)| > \epsilon\right)$$
$$\leq 2\nu.$$

Since ν was arbitrary, we get $\mathbb{P}\left(|\mathbb{G}_n(\phi_n - \phi_\star)| > \epsilon\right) \to 0$ for all $\epsilon > 0$, i.e. $\mathbb{G}_n(\phi_n - \phi_\star) = o_{\mathbb{P}}(1)$.

The previous result allows us to solely focus on the convergence of $\mathbb{G}_n\phi_\star + P(\phi_n - \phi_\star)$ in the remaining part of the proof. By a limited expansion of h and under $\mathcal{H}_0$, we have

$$\begin{aligned}
P(\phi_n - \phi_\star) = P\phi_n - \frac{1}{2} &= \int \mathbf{1}_{\{\psi(x,\hat{\alpha}_n,\hat{\beta}_n)>0\}} dP(x) - \frac{1}{2} \\
&= h(\alpha^\star, \beta^\star) + h_1(\alpha^\star, \beta^\star) \cdot (\hat{\alpha}_n - \alpha^\star) + h_2(\alpha^\star, \beta^\star) \cdot (\hat{\beta}_n - \beta^\star) \\
&+ o_{\mathbb{P}}\left(\|\hat{\alpha}_n - \alpha^\star\|\right) + o_{\mathbb{P}}\left(\|\hat{\beta}_n - \beta^\star\|\right) - \frac{1}{2} \\
&= h_1(\alpha^\star, \beta^\star) \cdot (\hat{\alpha}_n - \alpha^\star) + h_2(\alpha^\star, \beta^\star) \cdot (\hat{\beta}_n - \beta^\star) + o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right),
\end{aligned}$$

noting that $h(\alpha^\star, \beta^\star) = \frac{1}{2}$. Thus, this yields

$$\begin{aligned}
\frac{1}{\sqrt{n}}(\hat{B}_n - \frac{n}{2}) &= \mathbb{G}_n(\phi_n - \phi_\star) + \mathbb{G}_n\phi_\star + h_1(\alpha^\star, \beta^\star) \cdot \sqrt{n}(\hat{\alpha}_n - \alpha^\star) \\
&+ h_2(\alpha^\star, \beta^\star) \cdot \sqrt{n}(\hat{\beta}_n - \beta^\star) + o_{\mathbb{P}}(1) \\
&= \mathbb{G}_n\phi_\star + h_1(\alpha^\star, \beta^\star) \cdot \sqrt{n}(\hat{\alpha}_n - \alpha^\star) + h_2(\alpha^\star, \beta^\star) \cdot \sqrt{n}(\hat{\beta}_n - \beta^\star) \\
&+ o_{\mathbb{P}}(1).
\end{aligned}$$

From Assumption $B5$ we get $\sqrt{n}(\hat{\alpha}_n - \alpha^\star) = \sqrt{n}(\mathbb{P}_n s_1 - P s_1) + o_{\mathbb{P}}(1) = \mathbb{G}_n s_1 + o_{\mathbb{P}}(1)$ and $\sqrt{n}(\hat{\beta}_n - \beta^\star) = \sqrt{n}(\mathbb{P}_n s_2 - P s_2) + o_{\mathbb{P}}(1) = \mathbb{G}_n s_2 + o_{\mathbb{P}}(1)$, which allows us to calculate

$$\begin{aligned}
\frac{1}{\sqrt{n}}(\hat{B}_n - \frac{n}{2}) &= \mathbb{G}_n\phi_\star + h_1(\alpha^\star, \beta^\star) \cdot \mathbb{G}_n s_1 + h_2(\alpha^\star, \beta^\star) \cdot \mathbb{G}_n s_2 + o_{\mathbb{P}}(1) \\
&= \mathbb{G}_n(\phi_\star + h_1(\alpha^\star, \beta^\star) \cdot s_1 + h_2(\alpha^\star, \beta^\star) \cdot s_2) + o_{\mathbb{P}}(1).
\end{aligned}$$

Note that the coordinate projections of s_1 and s_2, even multiplied by some constants, are $P-Donsker$. Since finite sums of *Donsker* classes are a *Donsker* class, $\phi_\star + h_1(\alpha^\star, \beta^\star) \cdot s_1 + h_2(\alpha^\star, \beta^\star) \cdot s_2$ is also *Donsker*. Finally,

$$\mathbb{G}_n(\phi_\star + h_1(\alpha^\star, \beta^\star) \cdot s_1 + h_2(\alpha^\star, \beta^\star) \cdot s_2) + o_{\mathbb{P}}(1) \rightsquigarrow \mathbb{G},$$

where $\mathbb{G}$ is Gaussian with mean zero. The variance of $\mathbb{G}$ is equal to

$$\sigma_\psi^2 = \mathrm{Var}\left(\phi_\star(X) + h_1(\alpha^\star, \beta^\star) \cdot s_1(X) + h_2(\alpha^\star, \beta^\star) \cdot s_2(X)\right).$$

Proof of (ii):
By the first part of the proof, we know that

$$\begin{aligned}\frac{1}{\sqrt{n}}\left(\hat{B}_n - \frac{n}{2}\right) &= \mathbb{G}_n(\phi_n - \phi_\star) + \mathbb{G}_n\phi_\star + \sqrt{n}P(\phi_n - \phi_\star) + \sqrt{n}P(\phi_\star - \frac{1}{2}) \\ &= O_{\mathbb{P}}(1) + \sqrt{n}P(\phi_\star - \frac{1}{2}).\end{aligned}$$

If $P(\psi(X, \alpha^\star, \beta^\star) > 0) < \frac{1}{2}$, this yields $\sqrt{n}P(\phi_\star - \frac{1}{2}) \to -\infty$.

Proof of (iii):
Again, by a similar argument, we have

$$\begin{aligned}\frac{1}{\sqrt{n}}\left(\hat{B}_n - \frac{n}{2}\right) &= \mathbb{G}_n(\phi_n - \phi_\star) + \mathbb{G}_n\phi_\star + \sqrt{n}P(\phi_n - \phi_\star) + \sqrt{n}P(\phi_\star - \frac{1}{2}) \\ &= O_{\mathbb{P}}(1) + \sqrt{n}P(\phi_\star - \frac{1}{2}).\end{aligned}$$

If $P(\psi(X, \alpha^\star, \beta^\star) > 0) < \frac{1}{2}$, we deduce $\sqrt{n}P(\phi_\star - \frac{1}{2}) \to \infty$.

A.3 Proof of Proposition 5.0.3

We calculate

$$\begin{aligned}\sigma_h^2 &= \mathrm{Var}\left(h_1(\alpha^\star, \beta^\star) \cdot s_1(X) + h_2(\alpha^\star, \beta^\star) \cdot s_2(X)\right) \\ &= \mathrm{Var}\left(h_1(\alpha^\star, \beta^\star) \cdot s_1(X)\right) + \mathrm{Var}(h_2(\alpha^\star, \beta^\star) \cdot s_2(X)) \\ &+ 2\,\mathrm{Cov}(h_1(\alpha^\star, \beta^\star) \cdot s_1(X), h_2(\alpha^\star, \beta^\star) \cdot s_2(X)) \\ &= \sum_{i=1}^{d_\alpha}\sum_{j=1}^{d_\alpha} \mathrm{Cov}\left(\frac{\partial}{\partial\alpha_i}h(\alpha^\star, \beta^\star)s_{1,i}(X); \frac{\partial}{\partial\alpha_j}h(\alpha^\star, \beta^\star)s_{1,j}(X)\right) \\ &+ \sum_{i=1}^{d_\beta}\sum_{j=1}^{d_\beta} \mathrm{Cov}\left(\frac{\partial}{\partial\beta_i}h(\alpha^\star, \beta^\star)s_{2,i}(X); \frac{\partial}{\partial\beta_j}h(\alpha^\star, \beta^\star)s_{2,j}(X)\right)\end{aligned}$$

$$
\begin{aligned}
&+ 2\sum_{i=1}^{d_\alpha}\sum_{j=1}^{d_\beta} \operatorname{Cov}\left(\frac{\partial}{\partial\alpha_i}h(\alpha^\star,\beta^\star)s_{1,i}(X); \frac{\partial}{\partial\beta_j}h(\alpha^\star,\beta^\star)s_{2,j}(X)\right)\\
&= \sum_{i,j=1}^{d_\alpha} \frac{\partial}{\partial\alpha_i}h(\alpha^\star,\beta^\star)\frac{\partial}{\partial\alpha_j}h(\alpha^\star,\beta^\star)\Sigma_{i,j} + \sum_{i,j=1}^{d_\beta} \frac{\partial}{\partial\beta_i}h(\alpha^\star,\beta^\star)\frac{\partial}{\partial\beta_j}h(\alpha^\star,\beta^\star)\Sigma_{d_\alpha+i,d_\alpha+j}\\
&+ 2\sum_{i=1}^{d_\alpha}\sum_{j=1}^{d_\beta} \frac{\partial}{\partial\alpha_i}h(\alpha^\star,\beta^\star)\frac{\partial}{\partial\beta_j^\star}h(\alpha^\star,\beta^\star)\Sigma_{i,d_\alpha+j}\\
&= h_1(\alpha^\star,\beta^\star)^\intercal\Sigma_f(\alpha^\star,\beta^\star)h_1(\alpha^\star,\beta^\star) + h_2(\alpha^\star,\beta^\star)^\intercal\Sigma_g(\alpha^\star,\beta^\star)h_2(\alpha^\star,\beta^\star)\\
&+ 2h_1(\alpha^\star,\beta^\star)^\intercal\Sigma_{f,g}(\alpha^\star,\beta^\star)h_2(\alpha^\star,\beta^\star)
\end{aligned}
$$

since $\operatorname{Var}(s_1(X)) = \lim_{n\to\infty}\operatorname{Var}(\sqrt{n}(\hat{\alpha}_n - \alpha^\star))$ and $\operatorname{Var}(s_2(X)) = \lim_{n\to\infty}\operatorname{Var}(\sqrt{n}(\hat{\beta}_n - \beta^\star))$ as well as $\operatorname{Cov}(s_1(X); s_2(X)) = \lim_{n\to\infty}\operatorname{Cov}\left(\sqrt{n}(\hat{\alpha}_n - \alpha^\star); \sqrt{n}(\hat{\beta}_n - \beta^\star)\right)$ by the uniform integrability assumption.

A.4 Proof of Lemma 5.0.4

$$
\begin{aligned}
\hat{h}_{1,n,i} =& \frac{1}{2e(n)}\left(h_n(\hat{\alpha}_n + e(n)u_i^\alpha, \hat{\beta}_n) - h_n(\hat{\alpha}_n - e(n)u_i^\alpha, \hat{\beta}_n)\right)\\
&= \frac{1}{2e(n)}\left(\mathbb{P}_n\mathbf{1}\{\psi(\cdot,\hat{\alpha}_n + e(n)u_i^\alpha,\hat{\beta}_n) > 0\} - \mathbb{P}_n\mathbf{1}\{\psi(\cdot,\hat{\alpha}_n - e(n)u_i^\alpha,\hat{\beta}_n) > 0\}\right)\\
&= \frac{1}{2e(n)}\Bigg\{(\mathbb{P}_n - P)\left(\mathbf{1}\{\psi(\cdot,\hat{\alpha}_n + e(n)u_i^\alpha,\hat{\beta}_n) > 0\} - \phi_\star\right)\\
&- (\mathbb{P}_n - P)\left(\mathbf{1}\{\psi(\cdot,\hat{\alpha}_n - e(n)u_i^\alpha,\hat{\beta}_n) > 0\} - \phi_\star\right)\Bigg\}\\
&+ \frac{1}{2e(n)}P\left(\mathbf{1}\{\psi(\cdot,\hat{\alpha}_n + e(n)u_i^\alpha,\hat{\beta}_n) > 0\} - \mathbf{1}\{\psi(\cdot,\hat{\alpha}_n - e(n)u_i^\alpha,\hat{\beta}_n) > 0\}\right)\\
&= \frac{1}{2\sqrt{n}e(n)}\Bigg\{\mathbb{G}_n\left(\mathbf{1}\{\psi(\cdot,\hat{\alpha}_n + e(n)u_i^\alpha,\hat{\beta}_n) > 0\} - \phi_\star\right)\\
&- \mathbb{G}_n\left(\mathbf{1}\{\psi(\cdot,\hat{\alpha}_n - e(n)u_i^\alpha,\hat{\beta}_n) > 0\} - \phi_\star\right)\Bigg\}\\
&+ \frac{1}{2e(n)}P\left(\mathbf{1}\{\psi(\cdot,\hat{\alpha}_n + e(n)u_i^\alpha,\hat{\beta}_n) > 0\} - \mathbf{1}\{\psi(\cdot,\hat{\alpha}_n - e(n)u_i^\alpha,\hat{\beta}_n) > 0\}\right).
\end{aligned}
$$

From the proof of Theorem 5.0.1, we get that $\rho_P(\mathbf{1}\{\psi(\cdot,\hat{\alpha}_n + e(n)u_i^\alpha,\hat{\beta}_n) > 0\}, \phi_\star) \to 0$ $\mathbb{P}$-almost surely and $\rho_P(\mathbf{1}\{\psi(\cdot,\hat{\alpha}_n - e(n)u_i^\alpha,\hat{\beta}_n) > 0\}, \phi_\star) \to 0$ $\mathbb{P}$-almost surely. Using the equicontinuity of $\mathbb{G}_n$ and $\lim_{n\to\infty}\sqrt{n}e(n) > 0$, we get

$$
\frac{1}{2\sqrt{n}e(n)}\Bigg(\mathbb{G}_n\left(\mathbf{1}\{\psi(\cdot,\hat{\alpha}_n + e(n)u_i^\alpha,\hat{\beta}_n) > 0\} - \phi_\star\right) \tag{A.1}
$$

$$-\mathbb{G}_n\left(\mathbf{1}\{\psi(\cdot,\hat{\alpha}_n - e(n)u_i^{\alpha},\hat{\beta}_n) > 0\} - \phi_\star\right)\Big) = o_{\mathbb{P}^*}(1).$$

For the remaining term we use the same Taylor expansion as is the proof of Theorem 5.0.1 to calculate

$$\begin{aligned}
&\frac{1}{2e(n)}P\left(\mathbf{1}\{\psi(\cdot,\hat{\alpha}_n + e(n)u_i^{\alpha},\hat{\beta}_n) > 0\} - \mathbf{1}\{\psi(\cdot,\hat{\alpha}_n - e(n)u_i^{\alpha},\hat{\beta}_n) > 0\}\right)\\
&= \frac{1}{2e(n)}\Big(h_1(\alpha^\star,\beta^\star)\cdot(\hat{\alpha}_n + e(n)u_i^{\alpha} - \alpha^\star) - h_1(\alpha^\star,\beta^\star)\cdot(\hat{\alpha}_n - e(n)u_i^{\alpha} - \alpha^\star) + o_{\mathbb{P}}\left(e(n)\right)\Big)\\
&= \frac{\partial}{\partial\alpha_i}h(\alpha^\star,\beta^\star) + \frac{1}{2e(n)}o_{\mathbb{P}}\left(e(n)\right) = \frac{\partial}{\partial\alpha_i}h(\alpha^\star,\beta^\star) + o_{\mathbb{P}}(1).
\end{aligned}$$

Therefore, $\hat{h}_{1,n,i} \to \frac{\partial}{\partial\alpha_i}h(\alpha^\star,\beta^\star)$ in $\mathbb{P}$-probability. The convergence of $\hat{h}_{2,n,j}$ follows analogously.

A.5 Proof of Theorem 5.0.5

Define $\mathcal{Q}_\gamma := \{f + u_1\cdot s_1 + u_2\cdot s_2 \mid f\in\mathcal{F}_\gamma, (u_1,u_2)\in[h_1(\alpha^\star,\beta^\star)-\gamma, h_1(\alpha^\star,\beta^\star)+\gamma]\times[h_2(\alpha^\star,\beta^\star)-\gamma, h_2(\alpha^\star,\beta^\star)+\gamma]\}$. $\mathcal{Q}_\gamma$ is $P-Glivenko-Cantelli$, since it is a finite sum of $P-Donsker$ classes. Note that for $a<b$, $i\in\{1,2\}$ and $j\in\{1,..,d_{\alpha,\beta}\}$ we have $[a,b]s_{i,j} = \{\lambda_1 a s_{i,j} + (1-\lambda_1)bs_{i,j}\}$ is the convex hull of the *Donsker* class $\{as_{i,j}, bs_{i,j}\}$ and therefore a *Donsker* class. Note that we have invoked the permanence of the *Donsker* property under a convex hull transform (Theorem 2.10.3 in [19]). Additionally, $\mathcal{Q}_\gamma^2$ is also $P-Glivenko-Cantelli$ in $\mathbb{P}$-probability by Lemma 2.10.14 in [19]. It is sufficient to show

$$\mathbb{P}_n\left(\phi_n + \hat{h}_{1,n}\cdot s_1 + \hat{h}_{2,n}\cdot s_2\right)^2 \to P\big(\phi_\star + h_1(\alpha^\star,\beta^\star)\cdot s_1 + h_2(\alpha^\star,\beta^\star)\cdot s_2\big)^2$$

in outer $\mathbb{P}$-probability. Let $\epsilon>0$ and denote $q_n := \phi_n + \hat{h}_{1,n}\cdot s_1 + \hat{h}_{2,n}\cdot s_2$ and $q := \phi_\star + h_1(\alpha^\star,\beta^\star)\cdot s_1 + h_2(\alpha^\star,\beta^\star)\cdot s_2$. First,

$$\begin{aligned}
P(q_n^2 - q^2) &= P(\phi_n^2 - \phi_\star^2) + 2P(\hat{h}_{1,n}\phi_n - h_1(\alpha^\star,\beta^\star)\phi_\star)\cdot s_1\\
&\quad + 2P(\hat{h}_{2,n}\phi_n - h_2(\alpha^\star,\beta^\star)\phi_\star)\cdot s_2 + (\hat{h}_{1,n}^2 - h_1^2(\alpha^\star\beta^\star))\cdot Ps_1^2\\
&\quad + (\hat{h}_{2,n}^2 - h_2^2(\alpha^\star\beta^\star))\cdot Ps_2^2 + 2P\hat{h}_{1,n}\cdot s_1\times\hat{h}_{2,n}\cdot s_2\\
&\quad - 2Ph_1(\alpha^\star,\beta^\star)\cdot s_1\times h_2(\alpha^\star,\beta^\star)\cdot s_2\\
&= P(\phi_n^2 - \phi_\star^2) + 2P(\phi_n - \phi_\star)\hat{h}_{1,n}\cdot s_1 + 2P\phi_\star(\hat{h}_{1,n} - h_1(\alpha^\star,\beta^\star))\cdot s_1\\
&\quad + 2P(\phi_n - \phi_\star)\hat{h}_{2,n}\cdot s_2 + 2P\phi_\star(\hat{h}_{2,n} - h_2(\alpha^\star,\beta^\star))\cdot s_2\\
&\quad + (\hat{h}_{1,n}^2 - h_1^2(\alpha^\star\beta^\star))\cdot Ps_1^2 + (\hat{h}_{2,n}^2 - h_2^2(\alpha^\star\beta^\star))\cdot Ps_2^2
\end{aligned}$$

$$
\begin{aligned}
&+ 2\sum_{i=1}^{d_\alpha}\sum_{j=1}^{d_\beta}\left(\hat{h}_{1,n,i}\hat{h}_{2,n,j} - h_{1,i}(\alpha^\star,\beta^\star)h_{2,j}(\alpha^\star,\beta^\star)\right) Ps_{1,i}s_{2,j} \\
&\leq P(\phi_n^2 - \phi_\star^2) + 2\sum_{i=1}^{d_\alpha}\hat{h}_{1,n,i}\sqrt{P(\phi_n-\phi_\star)^2 Ps_{1,i}^2} \\
&+ 2\sum_{i=1}^{d_\alpha}(\hat{h}_{1,n,i} - h_{1,i}(\alpha^\star,\beta^\star))P\phi_\star \cdot s_{1,i} + 2\sum_{i=1}^{d_\beta}\hat{h}_{2,n,i}\sqrt{P(\phi_n-\phi_\star)^2 Ps_{2,i}^2} \\
&+ 2\sum_{i=1}^{d_\beta}(\hat{h}_{2,n,i} - h_{2,i}(\alpha^\star,\beta^\star))P\phi_\star \cdot s_{2,i} + (\hat{h}_{1,n}^2 - h_1^2(\alpha^\star\beta^\star))\cdot Ps_1^2 \\
&+ 2\sum_{i=1}^{d_\alpha}\sum_{j=1}^{d_\beta}\left(\hat{h}_{1,n,i}\hat{h}_{2,n,j} - h_{1,i}(\alpha^\star,\beta^\star)h_{2,j}(\alpha^\star,\beta^\star)\right)\sqrt{Ps_{1,i}^2 Ps_{2,j}^2} \\
&+ (\hat{h}_{2,n}^2 - h_2^2(\alpha^\star\beta^\star))\cdot Ps_2^2 \\
&\to 0
\end{aligned}
$$

in outer $\mathbb{P}$-probability. The previous statement is valid since from the proof of Theorem 5.0.1 $P(\phi_n - \phi_\star)^2 \to 0$ and by Lemma 5.0.4 $\hat{h}_{1,n,i} \to h_{1,i}(\alpha^\star,\beta^\star)$, $\hat{h}_{2,n,i} \to h_{2,i}(\alpha^\star,\beta^\star)$ as well as $\hat{h}_{1,n,i}\hat{h}_{2,n,j} \to h_{1,i}h_{2,j}$ in outer $\mathbb{P}$-probability.

Therefore,

$$
\begin{aligned}
&\mathbb{P}\left(|\mathbb{P}_n q_n^2 - Pq^2| > 2\epsilon\right) = \mathbb{P}\left(|(\mathbb{P}_n - P)q_n^2 + P(q_n^2 - q^2)| > 2\epsilon\right) \\
&\leq \mathbb{P}\left(|(\mathbb{P}_n - P)q_n^2| > \epsilon\right) + \mathbb{P}\left(|P(q_n^2 - q^2)| > \epsilon\right) \\
&\leq \mathbb{P}\Big(|(\mathbb{P}_n - P)q_n^2| > \epsilon; \|\hat{h}_{1,n} - h_1(\alpha^\star,\beta^\star)\|_2 < \gamma; \|\hat{h}_{2,n} - h_2(\alpha^\star,\beta^\star)\|_2 < \gamma; \\
&\qquad |\phi_n - \phi_\star| < \gamma\Big)\Big) \\
&+ \mathbb{P}\left(\|\hat{h}_{1,n} - h(\alpha^\star,\beta^\star)\|_2 > \gamma\right) + \mathbb{P}\left(\|\hat{h}_{2,n} - h_2(\alpha^\star,\beta^\star)\|_2 > \gamma\right) + \mathbb{P}\left(|\phi_n - \phi_\star| > \gamma\right) \\
&+ \mathbb{P}\left(|P(q_n^2 - q^2)| > \epsilon\right) \\
&\leq \mathbb{P}\left(\sup_{q^2\in\mathcal{Q}_\gamma^2}|(\mathbb{P}_n - P)q^2| > \epsilon\right) + \mathbb{P}\left(\|\hat{h}_{1,n} - h(\alpha^\star,\beta^\star)\|_2 > \gamma\right) \\
&+ \mathbb{P}\left(\|\hat{h}_{2,n} - h_2(\alpha^\star,\beta^\star)\|_2 > \gamma\right) + \mathbb{P}\left(|\phi_n - \phi_\star| > \gamma\right) + \mathbb{P}\left(|P(q_n^2 - q^2)| > \epsilon\right) \\
&\to 0,
\end{aligned}
$$

by the $Glivenko - Cantelli$ property of $\mathcal{Q}_\gamma^2$, the convergence in outer $\mathbb{P}$-probability of $\phi_n \to \phi_\star$, $\hat{h}_{1,n} \to h_1(\alpha^\star,\beta^\star)$, $\hat{h}_{2,n} \to h_2(\alpha^\star,\beta^\star)$ and $P(q_n^2 - q^2) \to 0$.

A.6 Proof of Lemma 5.1.1

We use Theorem 19.5 from [18] to show that for some $\gamma > 0$ $\mathcal{F}_\gamma$ is $P - Donsker$ with envelope $F = 1$. To this goal, we need to show

$$\int_0^1 \sqrt{\log\left(N_{[]}(\tilde{\epsilon}, \mathcal{F}_\gamma, L_2(P))\right)} d\tilde{\epsilon} < \infty$$

to state that $\mathcal{F}_\gamma$ is $P - Donsker$.

For any $f \in \mathcal{F}_\gamma$, we have $0 \leq f \leq 1$ and it suffices to consider $0 < \tilde{\epsilon} < 1$. Thus, let $0 < \tilde{\epsilon} < 1$ and choose $\epsilon \in (0, 1)$ such that $K(\epsilon)\epsilon + \epsilon = \tilde{\epsilon}$.
Let $\langle x_{i,1}, x_{i,2} \rangle$ denote (half-)open or (half-)closed intervals. Under Assumption $B6$, there exist $M_1(\epsilon)$ and $M_2(\epsilon)$ such that, for any $(\alpha, \beta) \in E_\gamma$, $f = \mathbf{1}_{\{\psi(\cdot,\alpha,\beta)>0\}}$ can be written as

$$f(\cdot) = \mathbf{1}_{\{\psi(\cdot,\alpha,\beta)>0\}} \mathbf{1}_{[M_1,M_2]^\complement}(\cdot) + \sum_{i=1}^{\tilde{K}} \mathbf{1}_{\langle x_{i,1}, x_{i,2} \rangle}(\cdot)$$

for some $\tilde{K} \leq \lfloor K(\epsilon) \rfloor$ and $(x_{i,1}, x_{i,2})_{1 \leq i \leq \tilde{K}} \in [M_1, M_2]^2$ with $x_{i,1} < x_{i,2} \leq x_{i+1,1}$.
Note that $\|f \mathbf{1}_{[M_1,M_2]^\complement}\|_{L_2(P)} \leq \epsilon$ and

$$\sum_{i=1}^{\tilde{K}} \mathbf{1}_{\langle x_{i,1}, x_{i,2} \rangle}(\cdot) \in \sum_{1 \leq i \leq \tilde{K}} \mathcal{D} := \left\{ \sum_{i=1}^{\tilde{K}} d_i \Big| \; d_i \in \mathcal{D} \right\}$$

where $\mathcal{D} = \{\mathbf{1}_{\langle a,b \rangle}(\cdot) \big| \; a, b \in \mathbb{R}\}$.

According to [18] Example 19.6, the class of functions $\{\mathbf{1}_{(-\infty,t\rangle} \mid t \in \mathbb{R}\}$ has bracketing numbers that are smaller or equal to $2/\epsilon^2$. Every indicator of the form $\mathbf{1}_{\langle a,b \rangle}$ can be represented as a difference $\mathbf{1}_{(-\infty,b\rangle} - \mathbf{1}_{(-\infty,a\rangle}$. Therefore, the bracketing numbers of $\mathcal{D}$ are smaller or equal to $64/\epsilon^4$.

Now, we know that there exist $d_1, ..., d_{\tilde{K}} \in \mathcal{D}$ such that $\sum_{i=1}^{\tilde{K}} \mathbf{1}_{\langle x_{i,1}, x_{i,2} \rangle} = \sum_{i=1}^{\tilde{K}} d_i$. Choose $d_1^l, ..., d_{\tilde{K}}^l$ and $d_1^u, ..., d_{\tilde{K}}^u$ from the brackets of $\mathcal{D}$ such that $d_i^l \leq d_i \leq d_i^u$ with $\|d_i^u - d_i^l\|_{L_2(P)} \leq \epsilon$. Therefore, $f^l := \sum_{i=1}^{\tilde{K}} d_i^l \leq f \leq \mathbf{1}_{[M_1,M_2]^\complement} + \sum_{i=1}^{\tilde{K}} d_i^u =: f^u$. Using the triangle inequality, we obtain

$$\|f^u - f^l\|_{L_2(P)} = \|\mathbf{1}_{[M_1,M_2]^\complement} + \sum_{i=1}^{\tilde{K}} d_i^u - \sum_{i=1}^{\tilde{K}} d_i^l\|_{L_2(P)} \leq (K(\epsilon) + 1)\epsilon = \tilde{\epsilon}.$$

This shows that for any $f \in \mathcal{F}_\gamma$ we can find a bracket $[f^l, f^u]$ constructed from at most $\lfloor K(\epsilon) \rfloor$ brackets of $\mathcal{D}$ with $\|f^u - f^l\|_{L_2(P)} \leq \tilde{\epsilon}$.

Consequently, $\mathcal{F}_\gamma \mathbf{1}_{[M_1,M_2]} \subset \sum_{1\leq i\leq \lfloor K(\epsilon)\rfloor} \mathcal{D}$ and we get the following bound on the bracketing numbers of $\mathcal{F}_\gamma$:

$$N_{[]}(\tilde{\epsilon}, \mathcal{F}_\gamma, L_2(P)) = N_{[]}((K(\epsilon)+1)\epsilon, \mathcal{F}_\gamma, L_2(P)) \leq N_{[]}(\epsilon, \mathcal{D}, L_2(P))^{\lceil K(\epsilon)+1\rceil}.$$

By Assumption $B6$ we have:

$$\begin{aligned}
&\int_0^\infty \sqrt{\log\left(N_{[]}(\tilde{\epsilon}, \mathcal{F}_\gamma, L_2(P))\right)} d\tilde{\epsilon} = \int_0^1 \sqrt{\log\left(N_{[]}(\tilde{\epsilon}, \mathcal{F}_\gamma, L_2(P))\right)} d\tilde{\epsilon} \\
&= \int_0^a \sqrt{\log\left(N_{[]}(\epsilon(K(\epsilon)+1), \mathcal{F}_\gamma, L_2(P))\right)} (K'(\epsilon)\epsilon + K(\epsilon) + 1) d\epsilon \\
&\leq \int_0^a \sqrt{\log\left(N_{[]}(\epsilon, \mathcal{D}, L_2(P))^{\lceil K(\epsilon)+1\rceil}\right)} (K'(\epsilon)\epsilon + K(\epsilon) + 1) d\epsilon \\
&\leq \int_0^a \sqrt{\log\left(\frac{64^{\lceil K(\epsilon)+1\rceil}}{(\epsilon^4)^{\lceil K(\epsilon)+1\rceil}}\right)} (K'(\epsilon)\epsilon + K(\epsilon) + 1) d\epsilon \\
&= \int_0^a \sqrt{\lceil K(\epsilon)+1\rceil \log(64) - 4\lceil K(\epsilon)+1\rceil \log(\epsilon)} (K'(\epsilon)\epsilon + K(\epsilon) + 1) d\epsilon \\
&< \infty.
\end{aligned}$$

This proves that $\mathcal{F}_\gamma$ is $P - Donsker$.

A.7 Proof of Lemma 5.2.1

Throughout the proof, let $\gamma > 0$ such that Assumption $B2$ and the respective condition $(i), (ii)$ or (iii) is true.

Proof of (i):
Note that $\xi_2(\alpha, \beta)$ takes only values in $\mathbb{R}$ and therefore there exists $t(a, b) \in \mathbb{R}$ such that we can write

$$\mathbf{1}_{\{\xi_1(x) > \xi_2(\alpha,\beta)\}} = \mathbf{1}_{\{\xi_1(x) > t(\alpha,\beta)\}} = 1 - \mathbf{1}_{\{\xi_1(x) \leq t(\alpha,\beta)\}}.$$

Therefore, $\mathbf{1}_{\{\psi(\cdot,\alpha,\beta)>0\}} \in 1 - \{\mathbf{1}_{\{\cdot\leq t\}} \big|\ t \in \mathbb{R}\}$ for all $(\alpha, \beta) \in E_\gamma$. Since $\{\mathbf{1}_{\{\cdot\leq t\}} \big|\ t \in \mathbb{R}\}$ is $Donsker$ for any probability measure, it is also $Donsker$ for $P \circ \xi_1^{-1}$. The linear transformation of a $Donsker$ class is again a $Donsker$ class by [10] Corollary 9.32 and therefore $\mathcal{F}_\gamma$ is $P - Donsker$.

Proof of (ii):

Using that $\mathbf{1}_{\{\cdot>0\}}$ is a monotone function we conclude by Lemma 2.6.18 in [19] that $\{\mathbf{1}_{\{\psi(\cdot,\alpha,\beta)>0\}}\big|\ (\alpha,\beta)\in E_\gamma\}$ is VC-subgraph. Therefore $\mathcal{F}_\gamma$ is $P-Donsker$.

Proof of (iii):
First, note that Lipschitz continuity w.r.t. $\|\cdot\|_r$ implies Lipschitz continuity w.r.t. to the Euclidean norm $\|\cdot\|_2$, by the equivalence of norms on $\mathbb{R}^{d_\alpha+d_\beta}$. Thus, we can assume that L is the upper bound of Lipschitz constants for the Euclidean norm.

By our assumptions, Theorem 2.7.11 in [19] implies that for $\mathcal{F}_\psi := \{\psi(\cdot,\alpha,\beta)\big|$ $(\alpha,\beta)\in E_\gamma\}$:

$$N_{[]}(2L\epsilon,\mathcal{F}_\psi,\|\cdot\|_\infty)\leq N(\epsilon,E_\gamma,\|\cdot\|_2).$$

The latter is $O\left(1/\epsilon^{d_\alpha+d_\beta}\right)$ by Problem 6 on page 94 in [19]. To see this, note that covering numbers are smaller than packing numbers and that any compact subset of $\mathbb{R}^d$ is contained in a Euclidean ball of radius R, for sufficiently large R. Since E_γ is compact, the claim follows.

Now, for any $\psi(\cdot,\alpha,\beta)\in\mathcal{F}_\psi$, there exist two functions ψ_1 and ψ_2, $\psi_1(\cdot)\leq\psi(\cdot,\alpha,\beta)\leq\psi_2(\cdot)$, chosen from the brackets of $\mathcal{F}_\psi$ with $\|\psi_1-\psi_2\|_\infty\leq\epsilon$. Therefore,

$$\mathbf{1}_{\{\psi_1(\cdot)>0\}}\leq\mathbf{1}_{\{\psi(\cdot,\alpha,\beta)>0\}}\leq\mathbf{1}_{\{\psi_2(\cdot)>0\}}.$$

Choose $\bar\epsilon$ such that $\sup_{(\alpha,\beta)\in E_\gamma}P\left(\psi(X,\alpha,\beta)\in[-\epsilon,\epsilon]\right)/\epsilon<A$ for all $\epsilon\leq\bar\epsilon$. For some $\epsilon\leq\bar\epsilon$, we calculate

$$\begin{aligned}\|\mathbf{1}_{\{\psi_1>0\}}-\mathbf{1}_{\{\psi_2>0\}}\|_{L_2(P)}&=\sqrt{P(\psi_2>0,\psi_1<0)}\leq\sqrt{P(\psi(X,\alpha,\beta)\in[-\epsilon,\epsilon])}\\&\leq\sqrt{\sup_{(\alpha,\beta)\in E_\gamma}\frac{P\left(\psi(X,\alpha,\beta)\in[-\epsilon,\epsilon]\right)}{\epsilon}\epsilon}<\sqrt{A\epsilon}.\end{aligned}$$

Therefore,

$$N_{[]}\left(\sqrt{\epsilon},\mathcal{F}_\gamma,\|\cdot\|_{L_2(P)}\right)\leq N_{[]}\left(\frac{\epsilon}{A},\mathcal{F}_\psi,\|\cdot\|_\infty\right)\in O\left(\frac{1}{\epsilon^{d_\alpha+d_\beta}}\right),$$

which is equivalent to

$$N_{[]}\left(\epsilon,\mathcal{F}_\gamma,\|\cdot\|_{L_2(P)}\right)\in O\left(\frac{1}{\epsilon^{2d_\alpha+2d_\beta}}\right).$$

If $\bar{\epsilon} < \epsilon \leq 1$, we have

$$N_{[]}\left(\epsilon, \mathcal{F}_\gamma, \|\cdot\|_{L_2(P)}\right) \leq N_{[]}\left(\bar{\epsilon}, \mathcal{F}_\gamma, \|\cdot\|_{L_2(P)}\right)$$

and for $\epsilon > 1$, we have

$$N_{[]}(\epsilon, \mathcal{F}_\gamma, \|\cdot\|_{L_2(P)}) = 1.$$

Thus,

$$\int_0^\infty \sqrt{log\left(N_{[]}(\epsilon, \mathcal{F}_\gamma, L_2(P))\right)} d\epsilon < \infty$$

and $\mathcal{F}_\gamma$ is $P - Donsker$ by Theorem 8.19 in [10].

A.8 Proof of Theorem 5.3.2

Choose $\gamma > 0$ such that $\mathcal{F}_\gamma$ is $P - Donsker$. We will use the same notation as in the proof of Theorem 5.0.1. Additionally, denote $\tilde{\phi}_n = \mathbf{1}_{\{\psi(\cdot,\tilde{\alpha}_n,\tilde{\beta}_n)>0\}}$. Furthermore, we know that $\tilde{\mathbb{G}}_n \underset{\xi}{\rightsquigarrow} \mathbb{G}$ by Theorem 3.6.13 in [19] since $\mathcal{F}_\gamma$ is $P - Donsker$ and the class $\{f - g \mid f, g \in \mathcal{F}_\gamma\ ,\ \rho_P(f,g)^2 < \delta\}$ is P-measurable by Assumption $B8$.

Proof of (i):

$$\begin{aligned}
\frac{1}{c\sqrt{n}}\left(\tilde{B}_n - \hat{B}_n\right) &= \sqrt{n}c^{-1}\left(\tilde{\mathbb{P}}_n\tilde{\phi}_n - \mathbb{P}_n\phi_n\right) \\
&= \sqrt{n}c^{-1}\Big(\left(\tilde{\mathbb{P}}_n - \mathbb{P}_n\right)\tilde{\phi}_n + \left(\mathbb{P}_n - P\right)\left(\tilde{\phi}_n - \phi_n\right) + P\left(\tilde{\phi}_n - \phi_n\right)\Big) \\
&= \sqrt{n}c^{-1}P\left(\tilde{\phi}_n - \phi_n\right) + \tilde{\mathbb{G}}_n\tilde{\phi}_n + c^{-1}\mathbb{G}_n\left(\tilde{\phi}_n - \phi_n\right) \\
&= \tilde{\mathbb{G}}_n\phi_\star + \sqrt{n}c^{-1}P\left(\tilde{\phi}_n - \phi_n\right) + \tilde{\mathbb{G}}_n\left(\tilde{\phi}_n - \phi_\star\right) \\
&\quad + c^{-1}\mathbb{G}_n\left(\tilde{\phi}_n - \phi_n\right).
\end{aligned}$$

By $B4$, $B5$ and $B7$, we can use a Taylor expansion to write

$$\begin{aligned}
\sqrt{n}c^{-1}P\left(\tilde{\phi}_n - \phi_n\right) &= \sqrt{n}c^{-1}\left(\tilde{\mathbb{P}}_n - \mathbb{P}_n\right)h_1(\alpha^\star, \beta^\star)\cdot s_1 + \sqrt{n}c^{-1}\left(\tilde{\mathbb{P}}_n - \mathbb{P}_n\right)h_2(\alpha^\star, \beta^\star)\cdot s_2 \\
&\quad + o_{\mathbb{P}^*_{XW}}(1) \\
&= \tilde{\mathbb{G}}_n\left(h_1(\alpha^\star, \beta^\star)\cdot s_1 + h_2(\alpha^\star, \beta^\star)\cdot s_2\right) + o_{\mathbb{P}^*_{XW}}(1).
\end{aligned}$$

Therefore,

$$\begin{aligned}&\tilde{\mathbb{G}}_n\phi_\star + \sqrt{n}c^{-1}P\left(\tilde{\phi}_n - \phi_n\right) + \tilde{\mathbb{G}}_n\left(\tilde{\phi}_n - \phi_\star\right) + c^{-1}\mathbb{G}_n\left(\tilde{\phi}_n - \phi_n\right)\\ &= \tilde{\mathbb{G}}_n\Big(\phi_\star + h_1(\alpha^\star,\beta^\star)\cdot s_1 + h_2(\alpha^\star,\beta^\star)\cdot s_2\Big) + \tilde{\mathbb{G}}_n\left(\tilde{\phi}_n - \phi_\star\right) + c^{-1}\mathbb{G}_n\left(\tilde{\phi}_n - \phi_n\right)\\ &+ o_{\mathbb{P}^*_{XW}}(1).\end{aligned}$$

From Lemma B.0.1 we know that $c^{-1}\mathbb{G}_n\left(\tilde{\phi}_n - \phi_n\right)$ and $\tilde{\mathbb{G}}_n\Big(\tilde{\phi}_n - \phi_\star\Big)$ are also $o_{\mathbb{P}^*_{XW}}(1)$. Thus,

$$\begin{aligned}&\tilde{\mathbb{G}}_n\Big(\phi_\star + h_1(\alpha^\star,\beta^\star)\cdot s_1 + h_2(\alpha^\star,\beta^\star)\cdot s_2\Big) + \tilde{\mathbb{G}}_n\left(\tilde{\phi}_n - \phi_\star\right) + c^{-1}\mathbb{G}_n\left(\tilde{\phi}_n - \phi_n\right) + o_{\mathbb{P}^*_{XW}}(1)\\ &= \tilde{\mathbb{G}}_n\Big(\phi_\star + h_1(\alpha^\star,\beta^\star)\cdot s_1 + h_2(\alpha^\star,\beta^\star)\cdot s_2\Big) + o_{\mathbb{P}^*_{XW}}(1).\end{aligned}$$

Proof of (iii):
Obvious by the usual bootstrap convergence.

Proof of (iii):
Obvious by Lemma 3.1 in [4].

A.9 Proof of Lemma 7.4.1

Let $\epsilon > 0$ arbitrary. Define $Z_n^j = \frac{1}{\sqrt{n}}\left(\tilde{B}_n^j - \hat{B}_n\right)$, where $\tilde{B}_n^j$ is the bootstrapped version of $\hat{B}_n$ calculated from the j-th bootstap sample. By Lemma 3.1 b) in [4], we get that for every fixed $B \in \mathbb{N}$ it is true that $(\mathbb{G}_n, \tilde{\mathbb{G}}_n^1, ..., \tilde{\mathbb{G}}_n^B) \rightsquigarrow (\mathbb{G}, \mathbb{G}^1, ..., \mathbb{G}^B)$, where $\tilde{\mathbb{G}}_n^j$ is the bootstrapped empirical process calculated from the j-th bootstrap sample and $\mathbb{G}, \mathbb{G}^1, ..., \mathbb{G}^B$ are independent and identically distributed. Therefore, $(Z_n^1, ..., Z_n^B) \rightsquigarrow (Z^1, ..., Z^B)$, where $Z^1, ..., Z^B$ are independent and Normal distributed with mean zero and variance σ_ψ^2.

Additionally, define

$$f_B : \mathbb{R}^B \to \mathbb{R}\ ;\ x \mapsto \left|\frac{1}{B}\sum_{i=1}^B\left(x_i - \frac{1}{B}\sum_{i=1}^B x_i\right)^2 - \sigma_\psi^2\right|.$$

Note that f_B is continuous and $f(Z_n^1, ..., Z_n^B)$ is uniformly integrable (in n and B) by Assumption $B10$. Therefore, for every $\nu > 0$, we can choose some $K \in \mathbb{R}$ such that for all

n and B large enough $\mathbb{E}^*_{\mathbb{P}_{XW}}\left[f_B(Z_n^1,...,Z_n^B)\mathbf{1}\{f(Z_n^1,...,Z_n^B)>K\}\right]\leq\nu$. We calculate

$$\begin{aligned}\mathbb{P}^*_{XW}\left(f_B(Z_n^1,...,Z_n^B)>\epsilon\right)&\leq\frac{1}{\epsilon}\big(\mathbb{E}^*_{\mathbb{P}_{XW}}\left[f_B(Z_n^1,...,Z_n^B)\mathbf{1}\{f(Z_n^1,...,Z_n^B)>K\}\right]\\&\quad+\mathbb{E}^*_{\mathbb{P}_{XW}}\left[\min\{f_B(Z_n^1,...,Z_n^B),K\}\right]\big)\\&\leq\frac{\nu}{\epsilon}+\frac{1}{\epsilon}\mathbb{E}^*_{\mathbb{P}_{XW}}\left[g_B(Z_n^1,...,Z_n^B)\right],\end{aligned}$$

where $g_B=\min\{f_B,K\}\in C_b(\mathbb{R}^B)$. Therefore,

$$\begin{aligned}\lim_{n\to\infty}\mathbb{P}^*_{XW}\left(f_B(Z_n^1,...,Z_n^B)>\epsilon\right)&\leq\frac{\nu}{\epsilon}+\lim_{n\to\infty}\mathbb{E}^*_{\mathbb{P}_{XW}}\left[g_B(Z_n^1,...,Z_n^B)\right]\\&=\frac{\nu}{\epsilon}+\mathbb{E}_{\mathbb{P}_{XW}}\left[g_B(Z^1,...,Z^B)\right].\end{aligned}$$

Now,

$$\lim_{B\to\infty}\lim_{n\to\infty}\mathbb{P}^*_{XW}\left(f_B(Z_n^1,...,Z_n^B)>\epsilon\right)\leq\frac{\nu}{\epsilon}+\lim_{B\to\infty}\mathbb{E}^*_{\mathbb{P}_{XW}}\left[g_B(Z^1,...,Z^B)\right]=\frac{\nu}{\epsilon},$$

by an application of the Dominated Convergence Theorem, since $\lim_{B\to\infty}g_B(Z^1,...,Z^B)=0$ $\mathbb{P}_{XW}$-almost surely.

Since ν was arbitrary, we have

$$\lim_{B\to\infty}\lim_{n\to\infty}\mathbb{P}^*_{XW}\left(f_B(Z_n^1,...,Z_n^B)>\epsilon\right)=0.$$

Appendix B

Technical Results

Lemma B.0.1
Under the assumptions of Theorem 5.3.2,

(i) for all $\epsilon > 0 : \mathbb{P}_{XW}\left(\|(\tilde{\alpha}_n, \tilde{\beta}_n) - (\alpha^\star, \beta^\star)\|_1 > \epsilon\right) \to 0$, *i.e.* $(\tilde{\alpha}_n, \tilde{\beta}_n) \to (\alpha^\star, \beta^\star)$ *in* $\mathbb{P}_{XW}$*-probability,*

(ii) $\mathbb{G}_n(\mathbf{1}_{\{\psi(\cdot,\tilde{\alpha}_n,\tilde{\beta}_n)>0\}} - \mathbf{1}_{\{\psi(\cdot,\hat{\alpha}_n,\hat{\beta}_n)>0\}}) = o_{\mathbb{P}^*_{XW}}(1)$, *and*

(iii) $\tilde{\mathbb{G}}_n(\mathbf{1}_{\{\psi(\cdot,\tilde{\alpha}_n,\tilde{\beta}_n)>0\}} - \mathbf{1}_{\{\psi(X,\alpha^\star,\beta^\star)>0\}}) = o_{\mathbb{P}^*_{XW}}(1)$.

Proof.
Proof of (i):
Since finite sums of $Glivenko - Cantelli$ functions are again $Glivenko - Cantelli$, we get that $\sum_{i=1}^{d_\alpha} s_{1,i} + \sum_{i=1}^{d_\beta} s_{2,i}$ is $Glivenko - Cantelli$. Define new weights $\tilde{\xi}_{i,n} = n^{-1}\xi_{i,n}$, which remain exchangeable and non-negative. Our assumptions yield $\sum_{i=1}^n \tilde{\xi}_{i,n} = 1$ and $\max_{1\le i\le n} \tilde{\xi}_{i,n} \to 0$ in $\mathbb{P}_W$-probability. Thus, all assumptions of Lemma 3.6.16 in [19] are satisfied. Then, write

$$\mathbb{P}_{XW}\left(\|(\tilde{\alpha}_n, \tilde{\beta}_n) - (\alpha^\star, \beta^\star)\|_1 > \epsilon\right) = \mathbb{E}_{\mathbb{P}_X}\left[\mathbb{P}_W\left(\|(\tilde{\alpha}_n, \tilde{\beta}_n) - (\alpha^\star, \beta^\star)\|_1 > \epsilon\right)\right],$$

because $\tilde{\alpha}_n$ and $\tilde{\beta}_n$ are measurable. Using Assumption $B5$ and $B7$, we get

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_X}\left[\mathbb{P}_W\left(\|(\tilde{\alpha}_n, \tilde{\beta}_n) - (\alpha^\star, \beta^\star)\|_1 > \epsilon\right)\right] \\
&\le \mathbb{E}^*_{\mathbb{P}_X}\left[\mathbb{P}^*_W\left(\sum_{i=1}^{d_\alpha} |\tilde{\mathbb{P}}_n s_{1,i} - P s_{1,i}| + \sum_{i=1}^{d_\beta} |\tilde{\mathbb{P}}_n s_{2,i} - P s_{2,i}| + |o_{\mathbb{P}^*_{XW}}(1)| > \epsilon\right)\right] \\
&\le \sum_{i=1}^{d_\alpha} \mathbb{E}_{\mathbb{P}_X}\left[\mathbb{P}_W\left(|\tilde{\mathbb{P}}_n s_{1,i} - P s_{1,i}| > \frac{\epsilon}{2(d_\alpha + d_\beta)}\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{i=1}^{d_\beta} \mathbb{E}_{\mathbb{P}_X}\left[\mathbb{P}_W\left(|\tilde{\mathbb{P}}_n s_{2,i} - P s_{2,i}| > \frac{\epsilon}{2(d_\alpha + d_\beta)}\right)\right] + \mathbb{E}^*_{\mathbb{P}_X}\left[\mathbb{P}^*_W\left(|o_{\mathbb{P}^*_{XW}}(1)| > \frac{\epsilon}{2}\right)\right] \\
&= \sum_{i=1}^{d_\alpha} \mathbb{E}_{\mathbb{P}_X}\left[\mathbb{P}_W\left(\Big|\sum_{j=1}^{n} \tilde{\xi}_{j,n}\left(\delta_{X_j} - P\right) s_{1,i}\Big| > \frac{\epsilon}{2(d_\alpha + d_\beta)}\right)\right] \\
&+ \sum_{i=1}^{d_\beta} \mathbb{E}_{\mathbb{P}_X}\left[\mathbb{P}_W\left(\Big|\sum_{j=1}^{n} \tilde{\xi}_{j,n}\left(\delta_{X_j} - P\right) s_{2,i}\Big| > \frac{\epsilon}{2(d_\alpha + d_\beta)}\right)\right] + \mathbb{P}^*_{XW}\left(|o_{\mathbb{P}^*_{XW}}(1)| > \frac{\epsilon}{2}\right) \\
&\to 0
\end{aligned}
$$

since, due to Lemma 3.6.16 in [19], every summand goes to 0 outer $\mathbb{P}_X$-almost surely and the outer Dominated Convergence Theorem (Theorem 6.12 in [10]) applies.

Proof of (ii):
First, we show $\mathbb{E}_P\left[\left(\tilde{\phi}_n - \phi_n\right)^2\right] \to 0$ in $\mathbb{P}_{XW}$-probability.

$$
\begin{aligned}
\mathbb{E}_P\left[\left(\tilde{\phi}_n - \phi_n\right)^2\right] &= \mathbb{E}_P\left[|\tilde{\phi}_n - \phi_n|\right] \leq \mathbb{E}_P\left[|\tilde{\phi}_n - \phi_\star|\right] + \mathbb{E}_P\Big[|\phi_\star - \phi_n|\Big] \\
&= \mathbb{E}_P\left[|\phi_n - \phi_\star|\right] + P\left(\psi(X, \alpha^\star, \beta^\star) > 0, \psi(X, \tilde{\alpha}_n, \tilde{\beta}_n) < 0\right) \\
&\quad + P\left(\psi(X, \alpha^\star, \beta^\star) < 0, \psi(X, \tilde{\alpha}_n, \tilde{\beta}_n) > 0\right).
\end{aligned}
$$

From the proof of Theorem 5.0.1, $\mathbb{E}_P\left[|\phi_n - \phi_\star|\right]$ converges to 0 $\mathbb{P}_X$-almost surely, which also implies the $\mathbb{P}_{XW}$-almost sure convergence to 0.
In the following, we show that the two remaining probabilities converge to 0 in $\mathbb{P}_{XW}$-probability. Note that the map $(\tilde{\alpha}_n, \tilde{\beta}_n) \to P(\psi(X, \tilde{\alpha}_n, \tilde{\beta}_n) \in A)$ is measurable for any Borel set A, due to the measurability of the map

$$
(\tilde{\Omega}, \tilde{\mathcal{A}}) \times E_\gamma \to \mathbb{R}, (\tilde{\omega}, \alpha, \beta) \to P\big(\psi(X(\tilde{\omega}), \alpha, \beta) \in A\big),
$$

where $(\tilde{\Omega}, \tilde{\mathcal{A}})$ is an independent copy of the underlying probability space of X.
Choose $\pi \in (0, 1)$ and a sufficiently small $\epsilon > 0$ such that $P(|\psi(X, \alpha^\star, \beta^\star)| \leq \epsilon) \leq \frac{\pi}{2}$. This is possible due to the non-nestedness of the proposed models f and g. Indeed, the non-nestedness asumption implies $0 = P(|\psi(X, \alpha^\star, \beta^\star)| = 0) = \lim_{n\to\infty} P(|\psi(X, \alpha^\star, \beta^\star)| \leq 1/n)$, by the continuity of measures.

Next, for an arbitrary $\nu > 0$, choose N large enough such that

$$
\mathbb{P}_{XW}\left(\|(\tilde{\alpha}_n, \tilde{\beta}_n) - (\alpha^\star, \beta^\star)\|_1 > \delta\right) \leq \nu,
$$

for all $n \geq N$, which is proven to be possible in the first part of the Lemma. Additionally, choose a compact set $A_\pi \subset \mathbb{R}^d$ and $\delta > 0$ with $P(X \in A_\pi) \geq 1 - \pi/4$ and for all $\|(\alpha, \beta) - (\alpha^\star, \beta^\star)\|_1 \leq \delta$ and $x \in A_\pi$, we have $|\psi(x, \alpha, \beta) - \psi(x, \alpha^\star, \beta^\star)| < \epsilon$. Such δ and A_π exist by Assumption $B9$ and the fact that X has a density w.r.t. the Lebesgue measure, which implies the tightness of X. We split the probabilities as follows:

$$
\begin{aligned}
&\mathbb{P}_{XW}\Big(P\left(\psi(X, \alpha^\star, \beta^\star) > 0; \psi(X, \tilde{\alpha}_n, \tilde{\beta}_n) < 0\right) > \pi\Big) \\
&\leq \mathbb{P}_{XW}\Big(P\left(\psi(X, \alpha^\star, \beta^\star) > \epsilon; \psi(X, \tilde{\alpha}_n, \tilde{\beta}_n) < 0\right) > \frac{\pi}{2}\Big) \\
&+ \mathbb{P}_{XW}\Big(P\left(\psi(X, \alpha^\star, \beta^\star) \in (0, \epsilon]; \psi(X, \tilde{\alpha}_n, \tilde{\beta}_n) < 0\right) > \frac{\pi}{2}\Big) \\
&\leq \mathbb{P}_{XW}\Big(P\left(\psi(X, \alpha^\star, \beta^\star) > \epsilon; \psi(X, \tilde{\alpha}_n, \tilde{\beta}_n) < 0\right) > \frac{\pi}{2}; \\
&\qquad \|(\tilde{\alpha}_n, \tilde{\beta}_n) - (\alpha^\star, \beta^\star)\|_1 \leq \delta\Big) + \mathbb{P}_{XW}\left(\|(\tilde{\alpha}_n, \tilde{\beta}_n) - (\alpha^\star, \beta^\star)\|_1 > \delta\right) \\
&\leq \nu + \mathbb{P}_{XW}\Big(P\left(\psi(X, \alpha^\star, \beta^\star) > \epsilon; \psi(X, \tilde{\alpha}_n, \tilde{\beta}_n) < 0; X \in A_\pi\right) > \frac{\pi}{4}; \\
&\qquad \|(\tilde{\alpha}_n, \tilde{\beta}_n) - (\alpha^\star, \beta^\star)\|_1 \leq \delta\Big) \\
&+ \mathbb{P}_{XW}\Big(P\left(\psi(X, \alpha^\star, \beta^\star) > \epsilon; \psi(X, \tilde{\alpha}_n, \tilde{\beta}_n) < 0; X \notin A_\pi\right) > \frac{\pi}{4}; \\
&\qquad \|(\tilde{\alpha}_n, \tilde{\beta}_n) - (\alpha^\star, \beta^\star)\|_1 \leq \delta\Big) \\
&= \nu + \mathbb{P}_{XW}\Big(P\left(\psi(X, \alpha^\star, \beta^\star) > \epsilon; \psi(X, \tilde{\alpha}_n, \tilde{\beta}_n) < 0; X \in A_\pi\right) > \frac{\pi}{4}; \\
&\qquad \|(\tilde{\alpha}_n, \tilde{\beta}_n) - (\alpha^\star, \beta^\star)\|_1 \leq \delta\Big),
\end{aligned}
$$

because $P(X \notin A_\pi) \leq \pi/4$. Finally, we observe that for $X \in A_\pi$ and $\|(\tilde{\alpha}_n, \tilde{\beta}_n) - (\alpha^\star, \beta^\star)\|_1 \leq \delta$, we have $|\psi(X, \hat{\alpha}_n, \hat{\beta}_n) - \psi(X, \alpha^\star, \beta^\star)| < \epsilon$. Therefore,

$$
\begin{aligned}
&\mathbb{P}_{XW}\Big(P\left(\psi(X, \alpha^\star, \beta^\star) > \epsilon; \psi(X, \tilde{\alpha}_n, \tilde{\beta}_n) < 0; X \in A_\pi\right) > \frac{\pi}{4}; \\
&\qquad \|(\tilde{\alpha}_n, \tilde{\beta}_n) - (\alpha^\star, \beta^\star)\|_1 \leq \delta\Big) = 0.
\end{aligned}
$$

Thus,

$$
\mathbb{P}_{XW}\Big(P\left(\psi(X, \alpha^\star, \beta^\star) > 0; \psi(X, \tilde{\alpha}_n, \tilde{\beta}_n) < 0\right) > \pi\Big) \leq \nu.
$$

Since ν was arbitrary, we get

$$P\left(\psi(X,\alpha^\star,\beta^\star)>0,\psi(X,\tilde{\alpha}_n,\tilde{\beta}_n)<0\right)\to 0$$

in $\mathbb{P}_{XW}$-probability. By the same arguments, we also get

$$\mathbb{P}_{XW}\left(\psi(X,\alpha^\star,\beta^\star)<0,\psi(X,\tilde{\alpha}_n,\tilde{\beta}_n)>0\right)\to 0$$

in $\mathbb{P}_{XW}$-probability. Combining the arguments, we have

$$\rho_P^2(\tilde{\phi}_n,\phi_n)=\mathbb{E}_P\left[\left(\tilde{\phi}_n-\phi_n\right)^2\right]\to 0$$

in $\mathbb{P}_{XW}$-probability.

Now, for arbitrary $\zeta,\pi>0$ choose $\lambda>0$ and N large enough such that $\mathbb{P}_X^*\left(\sup_{f,g\in\mathcal{F}_\gamma:\rho_P(f,g)\leq\lambda}\mathbb{G}_n(f-g)>\pi\right)\leq\zeta$ for all $n\geq N$, which is possible due to 2.1.8 in [19]. Additionally, choose N large enough such that $\mathbb{P}_{XW}\left(\rho_P(\tilde{\phi}_n,\phi_n)>\lambda\right)\leq\zeta$ for all $n\geq N$, which is possible due to the previous arguments.

Therefore, we deduce

$$\begin{aligned}
&\mathbb{P}_{XW}\left(|\mathbb{G}_n(\tilde{\phi}_n-\phi_n)|>\pi\right)=\mathbb{P}_{XW}\left(|\mathbb{G}_n(\tilde{\phi}_n-\phi_n)|>\pi;\rho_P(\tilde{\phi}_n,\phi_n)<\lambda\right)\\
&+\mathbb{P}_{XW}\left(|\mathbb{G}_n(\tilde{\phi}_n-\phi_n)|>\pi;\rho_P(\tilde{\phi}_n,\phi_n)\geq\lambda\right)\\
&\leq\mathbb{P}_X^*\left(\sup_{f,g\in\mathcal{F}_\gamma:\rho_P(f,g)<\lambda}|\mathbb{G}_n(f-g)|>\pi\right)+\mathbb{P}_{XW}\left(|\rho_P(\tilde{\phi}_n,\phi_n)|\geq\lambda\right)\\
&\leq 2\zeta.
\end{aligned}$$

Since ζ was arbitrary, we have

$$\mathbb{P}_{XW}\left(|\mathbb{G}_n(\tilde{\phi}_n-\phi_n)|>\pi\right)\to 0,$$

i.e. $\mathbb{G}_n(\tilde{\phi}_n-\phi_n)$ is $o_{\mathbb{P}_{XW}^*}(1)$.

Proof of (iii):
By the previous part of the proof, we get $\rho_p(\tilde{\phi}_n,\phi_\star)\to 0$ in $\mathbb{P}_{XW}$-probability. Mimicking the proof of part 2, let $\pi,\zeta>0$ arbitrary. Choose $\lambda>0$ and N large enough such that

$\mathbb{P}_{XW}\Big(\rho_p(\tilde{\phi}_n, \phi_\star) \geq \lambda\Big) \leq \zeta$ and

$$\mathbb{P}^*_{XW}\left(\sup_{f,g\in\mathcal{F}_\gamma;\rho_P(f,g)<\lambda} |\tilde{\mathbb{G}}_n(f-g)| > \pi\right) \leq \zeta$$

for all $n \geq N$, which is possible due to the equicontinuity of the bootstrapped process $\tilde{\mathbb{G}}_n$. The equicontinuity of the bootstrapped process is obtained in the proof of Theorem 3.6.13 in [19]. This yields

$$\begin{aligned}
&\mathbb{P}_{XW}\left(|\tilde{\mathbb{G}}_n(\tilde{\phi}_n - \phi_\star)| > \pi\right) \\
&\leq \mathbb{P}^*_{XW}\left(\sup_{f,g\in\mathcal{F}_\gamma;\rho_P(f,g)<\lambda} |\tilde{\mathbb{G}}_n(f-g)| > \pi\right) + \mathbb{P}_{XW}\Big(\rho_p(\tilde{\phi}_n, \phi_\star) \geq \lambda\Big) \\
&\leq 2\zeta.
\end{aligned}$$

Thus, $\tilde{\mathbb{G}}_n(\tilde{\phi}_n - \phi_\star)$ is $o_{\mathbb{P}^*_{XW}}(1)$.

□

Appendix C

Further Simulation Results

Consider the quotient of density families

$$\frac{\frac{1}{2\alpha_1}\exp\left(-\frac{|x-\alpha_2|}{2\alpha_1}\right)}{\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}}\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}}.$$

Fix P as a standard Normal distribution. To our knowledge, there exists no estimator of the degrees of freedom of a t-distribution, which satisfies assumption $B6$. This forces us to fix the degrees of freedom heuristically. In the following, we always use the sample size as an estimate of the degrees of freedom of the proposed t-distribution. The results of the simulations are summarized in Table C.1.

	Empirical Mean Of				Emp. Pow.
n	$\hat{B}_n$	$\hat{h}_{1,1}$	$\hat{h}_{1,2}$	$\hat{\sigma}^2_\psi$	$T_{1,n}$
50	21.54	-0.000516	-0.211537	0.444552	0.1420
100	40.27	0.002947	-0.246120	0.497599	0.2860
250	96.68	-0.006917	-0.268540	0.447074	0.7820
500	190.29	-0.013247	-0.271660	0.379970	0.9860
1000	378.23	-0.003495	-0.277170	0.355862	1.0000

Table C.1: Empirical mean and empirical power of the estimators in the case of a Laplace distribution vs. a t-distribution under a standard Normal distribution.

The test consistently prefers the Laplace distribution, since the empirical mean of $\hat{B}_n$ is always less than $n/2$. All estimators seem to be stable, since Table D.11 shows that the empirical variance of all estimators is low. The empirical mean of $\hat{h}_{1,1}$ is close to 0. Notice that, due to changing estimates of the degrees of freedom, the true value of the estimated quantities can vary with the sample size. This explains the changing empirical mean of partial derivative and variance estimates.

The empirical power of $T_{1,n}$ is increasing with the sample size, yielding very good results for $n \geq 500$.

Next, consider the quotient of density families

$$\frac{\frac{1}{\sqrt{2\pi\alpha_1}} \exp\left(-\frac{(x-\alpha_2)^2}{2\alpha_1}\right)}{\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}} \left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}},$$

in case P is chosen as a Laplace distribution with mean 0 and shape parameters 1 and 3. The results are summarized in Tables C.2 and C.3.

	Empirical Mean Of				Emp. Pow.
n	$\hat{B}_n$	$\hat{h}_{1,1}$	$\hat{h}_{1,2}$	$\hat{\sigma}^2_\psi$	$T_{1,n}$
50	17.07	0.001363	-0.126252	0.389593	0.6860
100	31.71	-0.003620	-0.061826	0.227685	0.9800
250	77.87	0.000227	-0.042686	0.169366	1.0000
500	154.63	0.000651	-0.040971	0.162818	1.0000
1000	308.37	-0.000210	-0.040515	0.159139	1.0000

Table C.2: Empirical mean and empirical power of the estimators in the case of Normal distribution vs. a t-distribution under a Laplace distribution with mean 0 and shape 1.

	Empirical Mean Of				Emp. Pow.
n	$\hat{B}_n$	$\hat{h}_{1,1}$	$\hat{h}_{1,2}$	$\hat{\sigma}^2_\psi$	$T_{1,n}$
50	28.00	0.000037	-0.003352	0.342647	0.1170
100	55.89	-0.000255	-0.002553	0.304563	0.1940
250	139.50	0.000277	-0.002595	0.267624	0.4330
500	279.21	-0.000214	-0.002500	0.249998	0.7520
1000	558.13	0.000065	-0.002480	0.238777	0.9770

Table C.3: Empirical mean and empirical power of the estimators in the case of Normal distribution vs. a t-distribution under a Laplace distribution with mean 0 and shape 3.

In general, the results are similar to all other examples. However, there are two peculiarities, which should be noted. First, the variance estimate only seems to be stable for $n \geq 250$, which can be seen in Tables D.12 and D.13. Second, it can be observed that the favored model changes with the parameters of the underlying distribution. This is due to the fact that $\hat{B}_n$ is consistently smaller than $n/2$ in Table C.2 and consistently greater than $n/2$ in Table C.3. Note, that this statement may seem obvious, but no example has been presented yet, which exhibits changing preferences with varying parameters. Both tables show increasing empirical power with increasing sample size. Table C.2 yields an empirical power of 100% for sample sizes $n \geq 250$. Furthermore, the empirical mean of $\hat{h}_{1,1}$ is close to 0 for both tables and every sample size.

Summarizing, we can say that the additional examples presented in Appendix C confirm the previous observations from Section 7.3. For sample sizes 50 and 100 we observed high variances of $\hat{\sigma}_{\psi}^2$, which lead to some instability of $T_{1,n}$.

Appendix D

Empirical Variance Tables

	Empirical Variance Of		
n	$\hat{h}_1$	$\hat{h}_2$	$\hat{\sigma}^2_\psi$
50	0.007421	0.026427	0.001831
100	0.004944	0.017164	0.000560
250	0.002763	0.010071	0.000160
500	0.001774	0.006087	0.000054
1000	0.001091	0.003845	0.000020
10000	0.000237	0.000854	0.000001

Table D.1: Empirical variance of the estimators for Example 1; corresponding to Table 7.1 .

	Empirical Variance Of		
n	$\hat{h}_1$	$\hat{h}_2$	$\hat{\sigma}^2_\psi$
50	0.000497	0.000497	158.962342
100	0.000304	0.000304	15.559060
250	0.000155	0.000155	5.439586
500	0.000096	0.000096	1.448591
1000	0.000065	0.000065	0.876214
10000	0.000013	0.000013	0.049156

Table D.2: Empirical variance of the estimators for Example 2; corresponding to Table 7.2

	Empirical Variance Of		
n	$\hat{h}_{1,1}$	$\hat{h}_{1,2}$	$\hat{\sigma}^2_\psi$
50	0.007503	0.006707	0.001797
100	0.005521	0.005452	0.000791
250	0.004541	0.004233	0.000293
500	0.003189	0.002612	0.000141
1000	0.002087	0.001622	0.000064
10000	0.000602	0.000441	0.000006

Table D.3: Empirical variance of the estimators in the case of $\mathcal{N}(0,2)$ vs. $\mathcal{N}(0,3)$ under a standard Normal distribution; corresponding to Table 7.3

	Empirical Variance Of				
n	$\hat{h}_{1,1}$	$\hat{h}_{1,2}$	$\hat{h}_{2,1}$	$\hat{h}_{2,2}$	$\hat{\sigma}^2_\psi$
50	0.013740	0.013358	0.026687	0.013404	0.018266
100	0.011101	0.008471	0.022317	0.009768	0.006829
250	0.006040	0.005281	0.012629	0.006163	0.002437
500	0.003929	0.003080	0.008147	0.004189	0.000982
1000	0.002507	0.002172	0.005198	0.002914	0.000771

Table D.4: Empirical variance of the estimators in the case of a Normal vs. Laplace distribution under a t distribution with 15 degrees of freedom; corresponding to Table 7.4.

	Empirical Variance Of		
n	$\hat{h}_1$	$\hat{h}_2$	$\hat{\sigma}^2_\psi$
50	0.032906	0.012246	0.001012
100	0.020526	0.009068	0.000608
250	0.013362	0.004931	0.000144
500	0.010338	0.003565	0.000053
1000	0.006971	0.002309	0.000025

Table D.5: Empirical variance of the estimators in the case of a Gamma vs. Exponential distribution under an Exponential distribution with mean equal to 1; corresponding to Table 7.5.

	Empirical Variance Of
n	$\hat{\sigma}^2_B$
50	0.002489
100	0.001278
250	0.000735
500	0.000465
1000	0.000363

Table D.6: Empirical variance of the bootstrap estimators for Example 1; corresponding to Table 7.6.

n	Empirical Variance Of $\hat{\sigma}_B^2$
50	0.025648
100	0.021729
250	0.011075
500	0.007867
1000	0.004854

Table D.7: Empirical variance of the bootstrap estimators for Example 2; corresponding to Table 7.7.

n	Empirical Variance Of $\hat{\sigma}_B^2$
50	0.001653
100	0.000904
250	0.000326
500	0.000195
1000	0.000122

Table D.8: Empirical variance and empirical level of the bootstrap estimators in the case of $\mathcal{N}(0, 2)$ vs. $\mathcal{N}(0, 3)$ under a standard Normal distribution; corresponding to Table 7.8.

n	Empirical Variance Of $\hat{\sigma}_B^2$
50	0.008663
100	0.004874
250	0.008923
500	0.001672
1000	0.001395

Table D.9: Empirical variance of the bootstrap estimators in the case of a Normal vs. Laplace distribution under a t distribution with 15 degrees of freedom; corresponding to Table 7.9.

n	Empirical Variance Of $\hat{\sigma}_B^2$
50	0.001939
100	0.001013
250	0.000487
500	0.000357
1000	0.000300

Table D.10: Empirical variance of the bootstrap estimators in the case of a Gamma vs. Exponential distribution under an Exponential distribution with mean 1; corresponding to Table 7.10.

	Empirical Variance Of		
n	$\hat{h}_{1,1}$	$\hat{h}_{1,2}$	$\hat{\sigma}^2_\psi$
50	0.084446	0.024818	0.027717
100	0.134594	0.013094	0.027402
250	0.093731	0.006675	0.032774
500	0.036978	0.004215	0.005135
1000	0.015369	0.003051	0.001270

Table D.11: Empirical variance of the estimators in the case of Laplace vs. a t distribution under a standard Normal distribution; corresponding to Table C.1.

	Empirical Variance Of		
n	$\hat{h}_{1,1}$	$\hat{h}_{1,2}$	$\hat{\sigma}^2_\psi$
50	0.012123	0.044950	0.428937
100	0.006558	0.012939	0.158472
250	0.003962	0.000677	0.000855
500	0.002454	0.000355	0.000388
1000	0.001566	0.000203	0.000177

Table D.12: Empirical variance of the estimators in the case of a Normal vs. a t distribution under a Laplace distribution with mean 0 shape 1; corresponding to Table C.2.

	Empirical Variance Of		
n	$\hat{h}_{1,1}$	$\hat{h}_{1,2}$	$\hat{\sigma}^2_\psi$
50	0.000436	0.000126	0.247924
100	0.000295	0.000061	0.101232
250	0.000144	0.000034	0.012479
500	0.000090	0.000020	0.003229
1000	0.000059	0.000012	0.000788

Table D.13: Empirical variance of the estimators in the case of Normal vs. a t distribution under a Laplace distribution with mean 0 and shape 3; corresponding to Table C.3.

Appendix E

Calculations For Example 2

We use the following formulas given in [3] and [14]. Assuming that the true distribution follows a generalized Gamma distribution with density $q(x, a, p, d)$ and considering an arbitrary generalized Gamma density $q(x, a_1, p_1, d_1)$, we get

$$\mathbb{E}\left[\log\left(\frac{q(X,a,d,p)}{q(X,a_1,d_1,p_1)}\right)\right] = \log\left(\frac{pa_1^{d_1}\Gamma(d_1/p_1)}{p_1 a^d \Gamma(d/p))}\right) + \left(\frac{\tau(d/p)}{p} + \log(a)\right)(d-d_1) + \frac{\Gamma(\frac{d+p_1}{p})}{\Gamma(d/p)}\left(\frac{a}{a_1}\right)^{p_1} - \frac{d}{p}$$

and

$$\mathbb{E}\left[\log\left(q(X,a,d,p)\right)\right] = \log\left(\frac{p}{a\Gamma(d/p)}\right) - \frac{d}{p} + \frac{d-1}{p}\tau(d/p),$$

where $\tau(y) = \Gamma'(y)/\Gamma(y)$ denotes the digamma function. Exploiting that the Weibull distribution and the Gamma distribution are special cases of a generalized Gamma distribution, we can deduce the following formulas

$$\begin{aligned}\mathbb{E}\left[\log\left(w(X,\alpha,2)\right)\right] &= \mathbb{E}\left[\log\left(q(X,\alpha^{1/2},2,2)\right)\right] \\ &= -\mathbb{E}\left[\log\left(\frac{p(X,a,d,p)}{p(X,\alpha^{1/2},2,2)}\right)\right] + \mathbb{E}\left[\log\left(p(X,a,d,p)\right)\right] \\ &= \log\left(\frac{2a}{\alpha}\right) + \frac{1}{p}\tau\left(\frac{d}{p}\right) - \frac{\Gamma\left(\frac{d+2}{p}\right)}{\Gamma\left(\frac{d}{p}\right)}\frac{a^2}{\alpha}\end{aligned}$$

and

$$\mathbb{E}\left[\log\left(g(X,\beta,2)\right)\right] = \mathbb{E}\left[\log\left(q(X,\beta,2,1)\right)\right]$$

$$
\begin{aligned}
&= -\mathbb{E}\left[\log\left(\frac{p(X,a,p,d)}{p(X,,\beta,2,1)}\right)\right] + \mathbb{E}\left[\log\left(p(X,a,p,d)\right)\right] \\
&= \log\left(\frac{a}{\beta^2}\right) + \frac{1}{p}\tau\left(\frac{d}{p}\right) - \frac{\Gamma\left(\frac{d+1}{p}\right)}{\Gamma\left(\frac{d}{p}\right)}\frac{a}{\beta}.
\end{aligned}
$$

Using the formulas introduced above and that

$$
\mathbb{E}\left[X^s\right] = \frac{a^s\Gamma\left(\frac{s}{p}+\frac{d}{p}\right)}{\Gamma\left(\frac{d}{p}\right)},
$$

we can calculate the asymptotic variance

$$
\begin{aligned}
\sigma_\psi^2 &= \operatorname{Var}\left(\mathbf{1}\{\psi(X,\alpha^\star,\beta^\star)>0\} + h_1(\alpha^\star,\beta^\star)X^2 + \frac{h_2(\alpha^\star,\beta^\star)}{2}X\right) \\
&= \operatorname{Var}\left(\mathbf{1}\{\psi(X,\alpha^\star,\beta^\star)>0\}\right) + h_1^2(\alpha^\star,\beta^\star)\operatorname{Var}\left(X^2\right) + \frac{h_2^2(\alpha^\star,\beta^\star)}{4}\operatorname{Var}\left(X\right) \\
&+ 2\operatorname{Cov}\left(\mathbf{1}\{\psi(X,\alpha^\star,\beta^\star)>0\}, h_1(\alpha^\star,\beta^\star)X^2\right) \\
&+ 2\operatorname{Cov}\left(\mathbf{1}\{\psi(X,\alpha^\star,\beta^\star)>0\}, \frac{h_2(\alpha^\star,\beta^\star)}{2}X\right) \\
&+ 2\operatorname{Cov}\left(h_1(\alpha^\star,\beta^\star)X^2, \frac{h_2(\alpha^\star,\beta^\star)}{2}X\right) \\
&= \frac{1}{4} + h_1^2(\alpha^\star,\beta^\star)\left(a^4\frac{\Gamma\left(\frac{d+4}{p}\right)}{\Gamma\left(\frac{d}{p}\right)} - a^4\frac{\Gamma\left(\frac{d+2}{p}\right)^2}{\Gamma\left(\frac{d}{p}\right)^2}\right) \\
&+ \frac{h_2^2(\alpha^\star,\beta^\star)}{4}\left(a^2\frac{\Gamma\left(\frac{d+2}{p}\right)}{\Gamma\left(\frac{d}{p}\right)} - a^2\frac{\Gamma\left(\frac{d+1}{p}\right)^2}{\Gamma\left(\frac{d}{p}\right)^2}\right) \\
&+ 2h_1(\alpha^\star,\beta^\star)\left(\mathbb{E}\left[X^2\mathbf{1}_{\{X\in[x_1,x_2]^\complement\}}\right] - \frac{a^2}{2}\frac{\Gamma\left(\frac{d+2}{p}\right)}{\Gamma\left(\frac{d}{p}\right)}\right) \\
&+ h_2(\alpha^\star,\beta^\star)\left(\mathbb{E}\left[X\mathbf{1}_{\{X\in[x_1,x_2]^\complement\}}\right] - \frac{a}{2}\frac{\Gamma\left(\frac{d+1}{p}\right)}{\Gamma\left(\frac{d}{p}\right)}\right) \\
&+ h_1(\alpha^\star,\beta^\star)h_2(\alpha^\star,\beta^\star)\left(a^3\frac{\Gamma\left(\frac{d+3}{p}\right)}{\Gamma\left(\frac{d}{p}\right)} - a^3\frac{\Gamma\left(\frac{d+2}{p}\right)\Gamma\left(\frac{d+1}{p}\right)}{\Gamma\left(\frac{d}{p}\right)^2}\right) \\
&\approx 0.3475695.
\end{aligned}
$$

Bibliography

[1] Akaike, H. *Information Theory and an Extension of the Maximum Likelihood Principle*. Springer New York, New York, NY, 1998, pp. 199–213.

[2] Barth, M. E., Gow, I. D., and Taylor, D. J. Why do pro forma and street earnings not reflect changes in gaap? evidence from sfas 123r. *Review of Accounting Studies 17*, 3 (2012), 526–562.

[3] Bauckhage, C. Computing the Kullback-Leibler divergence between two generalized gamma distributions. *CoRR abs/1401.6853* (2014).

[4] Buecher, A., and Kojadinovic, I. A note on conditional versus joint unconditional weak convergence in bootstrap consistency results. *Journal of Theoretical Probability* (Mar 2018).

[5] Chen, X., and Fan, Y. Pseudo-likelihood ratio tests for semiparametric multivariate copula model selection. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique 33*, 3 (2005), 389–414.

[6] Clarke, K. A. A simple distribution-free test for nonnested model selection. *Political Analysis 15*, 3 (2007), 347–363.

[7] Durrett, R. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2010.

[8] Harrison, G. W., Martínez-Correa, J., and Swarthout, J. T. Reduction of compound lotteries with objective probabilities: Theory and evidence. *Journal of Economic Behavior & Organization 119* (2015), 32–55.

[9] Joslin, S., and Sherman, R. P. An equivalence result for vc classes of sets. *Econometric Theory 19*, 6 (2003), 1123–1127.

[10] Kosorok, M. R. *Introduction to empirical processes and semiparametric inference.* Springer, 2008.

[11] Kullback, S., and Leibler, R. A. On information and sufficiency. *Ann. Math. Statist. 22*, 1 (03 1951), 79–86.

[12] Markwat, T., Kole, E., and Van Dijk, D. Contagion as a domino effect in global stock markets. *Journal of Banking & Finance 33*, 11 (2009), 1996–2012.

[13] Martin, L. W., and Vanberg, G. Parties and policymaking in multiparty governments: the legislative median, ministerial autonomy, and the coalition compromise. *American Journal of Political Science 58*, 4 (2014), 979–996.

[14] Morteza, K., and Ahmadabadi, A. Some properties of generalized gamma distribution. *Mathematical Sciences Quarterly Journal 4* (03 2010).

[15] Rudin, W. *Functional Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1991.

[16] Schwarz, G. Estimating the dimension of a model. *Ann. Statist. 6*, 2 (03 1978), 461–464.

[17] Shalev-Shwartz, S., and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[18] Vaart, A. W. v. d. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[19] van der Vaart, A. W., and Wellner, J. A. *Weak Convergence and Empirical Processes*. Springer New York, 1996.

[20] Vuong, Q. H. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica 57*, 2 (1989), 307–333.

[21] White, H. Maximum likelihood estimation of misspecified models. *Econometrica 50*, 1 (1982), 1–25.