Statistical applications of generalized quantiles : nonparametric tolerance regions and P-P plots

Statistical Applications of Generalized Quantiles

Nonparametric Tolerance Regions and P-P Plots

THOMAS STIELTJES INSTITUTE
FOR MATHEMATICS

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN

Mushkudiani, Nino

Statistical applications of generalized quantiles : nonparametric tolerance regions and P-P plots / by Nino Mushkudiani. - Eindhoven : Eindhoven University of Technology, 2000
Proefschrift. - ISBN 90-386-0821-7
NUGI 815
Subject headings : nonparametric statistics / statistical data analysis
2000 Mathematics Subject Classification : 62G15, 62G20, 62H11, 60F05, 62G10, 62-09, 62-07

Printed by Universiteitsdrukkerij Technische Universiteit Eindhoven

Statistical Applications of Generalized Quantiles

Nonparametric Tolerance Regions and P-P Plots

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
Rector Magnificus, prof.dr. M. Rem, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op woensdag 29 november 2000 om 16.00 uur

door

Nino Mushkudiani

geboren te Sagaredjo, Georgië

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr. P. van der Laan
en
prof.dr. E.V. Khmaladze

Copromotor:
dr. J.H.J. Einmahl

Contents

1 Introduction and preliminaries **1**
1.1 Outline 1
1.2 Skorokhod construction 2
1.3 Minimum volume sets 6
1.4 Generalized quantile processes 7

I Nonparametric tolerance regions **9**

2 Brief historical review **11**
2.1 Introduction 11
2.2 Definitions and setup 12
2.3 Tolerance intervals: one dimensional case 13
2.4 Nonparametric tolerance regions 14
2.5 Directional tolerance regions 17

3 Small nonparametric tolerance regions **19**
3.1 Notations and preliminary results 19
3.2 Limit theorems for small tolerance regions 22
3.3 Applications 34
3.4 Appendix 43

4 Small nonparametric tolerance regions for directional data **47**
4.1 Introduction 47
4.2 The setup 48
4.3 Main results 53
4.4 Simulation study and real data example 55

Discussion **59**

II P-P plots **61**

5 Brief review **63**
5.1 Introduction 63
5.2 Probability-probability plots 64

6 Generalized P-P plots **71**
6.1 Introduction and the testing procedure 71
6.2 One-sample problem 72
6.3 Two-sample problem 77
6.4 Numerical results 81
6.5 Proofs 82

Discussion **89**

Bibliography **91**

Index **97**

Samenvatting **99**

Acknowledgments **101**

Curriculum vitae **103**

Chapter 1

Introduction and preliminaries

Here we provide the reader with a short overview of the following chapters and with the concepts of empirical process theory used later on.

1.1 Outline

In Einmahl and Mason (1992) generalized quantile processes are defined and studied. These processes are based on generalized quantile functions, a flexible way to summarize properties of multivariate data or a probability distribution. Statistical applications of the theory of generalized quantiles is the main subject of this thesis.

A special case of generalized quantile functions leads to the idea of minimum volume sets, which inspired a new way of constructing nonparametric multivariate tolerance regions, defined and studied intensely in Part I. We define tolerance intervals as the minimum length intervals, that contain a certain number of observations. In higher dimensions the idea can be extended naturally by defining tolerance regions as minimum volume sets from a general indexing class.

In $\mathbb{R}^k$, $k \geq 1$, for fixed $t_0 \in [0,1]$, $C \in \mathbb{R}$ and $n \in \mathbb{N}$ large enough, define the tolerance region $A_{n,t_0,C}$ as the smallest volume (Lebesgue measure) set from a class of sets $\mathcal{A}$, containing at least $t_0 + \frac{C}{\sqrt{n}}$ observations. The class $\mathcal{A}$ will be specialized to the following classes: all closed

(a) ellipsoids,

(b) hyperrectangles with axes parallel to the coordinate hyperplanes,

(c) convex sets (for $k \leq 2$),

that have probability between 0 and 1. We extend these cases further. For each fixed integer m we consider (a') unions of m closed ellipsoids, (b') unions of m closed parallel hyperrectangles and (c') unions of m closed convex sets, contained in a fixed, large compact set (for $k = 2$).

We use the idea of minimum volume sets to construct tolerance regions for circular and spherical data. For these cases $\mathcal{A}$ will be respectively

(d) the class of arcs;

(e) the class of caps (defined in Chapter 4).

As above we assume here as well that each $A \in \mathcal{A}$ has a probability between 0 and 1.

The asymptotic theory for these tolerance regions is derived under very weak conditions. We show that these tolerance regions are asymptotically minimal with respect to the indexing class and have desirable invariance properties. We also investigate finite sample properties of the tolerance regions through a simulation study and consider real data examples.

In Part II we continue with the application of the theory of generalized quantiles while studying graphical methods for hypothesis testing. Define the generalized P-P plot as

$$m_n(t) := \sup\{P_n(A) \,:\, P_0(A) \le t,\, A \in \mathcal{A}\}, \quad t \in [0,1],$$

here $\mathcal{A}$ is the class of all closed intervals on $\mathbb{R}$. The generalized P-P plot can be considered as a diagnostic plot, which compares the empirical and hypothetical distributions over the class $\mathcal{A}$. Further we define the generalized empirical P-P plot process as

$$M_n(t) := \sqrt{n}\big(\sup\{P_n(A) \,:\, P_0(A) \le t,\, A \in \mathcal{A}\} - t\big), \quad t \in [0,1].$$

This process can be recognized as an inverse of the generalized quantile process defined later. We derive asymptotic behavior of M_n under the null and alternative hypothesis and show that it is asymptotically distribution-free under the null hypothesis. We also study behavior of M_n in case of contiguous alternatives. The two-sample problem is stated and treated similarly.

The rest of this chapter reviews background material from empirical process theory: in Section 1.2 we describe the Skorokhod construction and its applications. Minimum volume sets are introduced in Section 1.3 and finally in Section 1.4 we define the generalized quantile process and present results on this limiting behavior.

Notations introduced in this chapter will be used throughout this thesis.

1.2 Skorokhod construction

Let (S,d) be a metric space with some metric d and let $(S,\mathcal{S})$ denote S with the σ-algebra generated by the open balls. Next suppose that $\xi, \xi_1, \xi_2, \ldots$ is a sequence of random elements defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in $(S,\mathcal{S})$.

Definition 1.1 *We say that the sequence ξ_n, $n \ge 1$, converges weakly to ξ and write $\xi_n \xrightarrow{w} \xi$, iff*

$$\int f dP_n \to \int f dP, \quad as \quad n \to \infty,$$

for every d-continuous, bounded, $\mathcal{S}$-measurable, real-valued function f defined on S, where $P, P_1, P_2, \ldots$ denote the distributions of the random elements $\xi, \xi_1, \xi_2, \ldots$, respectively.

For random elements ξ and η defined on the same probability space, we write that $\mathcal{L}(\xi) = \mathcal{L}(\eta)$ iff ξ and η have the same distribution. The following result which is often used when studying limiting behavior of functionals of the sequences of random elements, can be found for example in Shorack and Wellner (1986).

Theorem 1.1 *Let $\xi_n \xrightarrow{w} \xi$ and let ψ denote a real-valued, $\mathcal{S}$-measurable function on S that is continuous with respect to the metric d, then*

$$\psi(\xi_n) \xrightarrow{w} \psi(\xi), \quad as \quad n \to \infty.$$

Weak convergence of random elements tells us more about these distributions. However when we deal with random processes which have functions as sample paths, we rather have asymptotic results in the metric space of these functions. For this purpose the following Skorokhod-Dudley-Wichura representation theorem (Skorokhod construction) is the right tool.

Theorem 1.2 *Suppose $\xi_n \xrightarrow{w} \xi$, then there exists a probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{I\!P})$ carrying a sequence of random elements $\widetilde{\xi}, \widetilde{\xi}_1, \widetilde{\xi}_2, \ldots$ in $(S, \mathcal{S})$ such that*

$$\begin{gathered} \mathcal{L}(\widetilde{\xi}) = \mathcal{L}(\xi),\ \mathcal{L}(\widetilde{\xi}_n) = \mathcal{L}(\xi_n), \quad for \quad n \geq 1, \\ and \\ d(\widetilde{\xi}_n, \widetilde{\xi}) \to 0 \quad a.s., \quad as \quad n \to \infty. \end{gathered} \tag{1.1}$$

This theorem holds more general: when random elements $\xi_1, \xi_2, \ldots$ are defined on the Borel σ-algebra (see, e.g., Gaenssler (1983)). We apply this construction several times in Chapter 6 in the following setting.

By C denote the class of all continuous functions on $[0, 1]$. Define the supremum norm metric on C as

$$\|f - g\| = \sup_{t \in [0,1]} |f(t) - g(t)|, \quad f, g \in C. \tag{1.2}$$

Then the space C endowed with this metric is a complete, separable metric space and hence we can define $\mathcal{C}$, the Borel σ-algebra (σ-algebra generated by the open sets from C). Further let D be the class of all right-continuous functions on $[0, 1]$, with left-hand limits at each point. Then D with the supremum norm defined in (1.2) is a complete, nonseparable metric space and the Borel σ-algebra of D is too large for our purposes, namely uniform empirical process, defined below, is not measurable with respect to it. To avoid measurability problems, instead of the Borel σ-algebra we consider a smaller σ-algebra: by $\mathcal{D}$ denote the σ-algebra of subsets of D generated by the open balls. Note that equipped with the Skorokhod metric the space of all right-continuous functions defined on [0.1] that have left-hand

limits is complete, separable metric space and its Borel σ-algebra coincides with the σ-algebra generated by the open balls (see, e.g., Billingsley (1968)).

Consider a sequence $U_1, U_2, \ldots$ of i.i.d. uniform-$[0,1]$ random variables defined on some probability space. For each $n \geq 1$, define the uniform empirical process

$$\Gamma_n(t) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \big[I_{[0,t]}(U_i) - t \big], \quad t \in [0,1],$$

where I_A denotes the indicator function of the set A; Γ_n is a random element on $(D, \mathcal{D})$. Let B denote a Brownian bridge, a Gaussian process on $(C, \mathcal{C})$ with

$$I\!E B(t) = 0, \quad \text{and} \quad I\!E[B(s), B(t)] = s \wedge t - st \quad \text{for} \quad 0 \leq s, t \leq 1.$$

Then $\Gamma_n \xrightarrow{w} B$ on $(D, \mathcal{D})$. Given this weak convergence by the Skorokhod-Dudley-Wichura representation theorem we obtain that there exists a probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{I\!P})$ carrying $\widetilde{\Gamma}_1, \widetilde{\Gamma}_2, \ldots$ and a version $\widetilde{B}$ of B in $(C, \mathcal{C})$ with

$$\begin{gathered} \mathcal{L}(\widetilde{B}) = \mathcal{L}(B), \quad \mathcal{L}(\widetilde{\Gamma}_n) = \mathcal{L}(\Gamma_n), \quad \text{for} \quad n \geq 1, \\ \text{and} \\ \sup_{t \in [0,1]} |\widetilde{\Gamma}_n(t) - \widetilde{B}(t)| \to 0 \quad \text{a.s.}, \quad \text{as} \quad n \to \infty. \end{gathered} \tag{1.3}$$

Note that the empirical process $\widetilde{\Gamma}_n$ is based on a triangular array $\widetilde{U}_{n1}, \ldots, \widetilde{U}_{nn}$ of uniform-$[0,1]$ random variables.

Next suppose that $X_1, \ldots, X_n$, $n \geq 1$, are i.i.d. $I\!R$-valued random variables defined on a probability space $(\Omega, \mathcal{F}, I\!P)$ having a common probability distribution P, with corresponding distribution function F. In this case a result similar to (1.3) can be obtained using the F^{-1}-transformation. On $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{I\!P})$ for each $n \geq 1$, define $\widetilde{X}_{ni} = F^{-1}(\widetilde{U}_{ni})$, for $1 \leq i \leq n$, and the empirical distribution function $\widetilde{F}_n$ based on these random variables

$$\widetilde{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,t]}(\widetilde{X}_{ni}), \quad t \in I\!R.$$

Then the empirical process $\widetilde{\alpha}_n(t) := \sqrt{n}(\widetilde{F}_n(t) - F(t)) = \widetilde{\Gamma}_n(F(t))$, $t \in I\!R$, and by (1.3)

$$\sup_{t \in I\!R} |\widetilde{\alpha}_n(t) - \widetilde{B}(F(t))| \to 0 \quad \text{a.s.}, \quad n \to \infty. \tag{1.4}$$

Finally we review the well known results on weak convergence of empirical processes indexed by a class of sets in $I\!R^k$, $k \geq 1$. Let $X_1, \ldots, X_n$, $n \geq 1$, be i.i.d. $I\!R^k$-valued random vectors defined on a probability space $(\Omega, \mathcal{F}, I\!P)$ with common probability distribution P, absolutely continuous with respect to Lebesgue measure and the corresponding distribution function F. Further let $\mathcal{B}$ be the σ-algebra of Borel sets on $I\!R^k$ and let d_0 be the pseudo-metric on $\mathcal{B}$ defined as

$$d_0(B_1, B_2) = P(B_1 \triangle B_2), \quad \text{for } B_1, B_2 \in \mathcal{B}.$$

Denote by P_n the empirical distribution:

$$P_n(B) = \frac{1}{n}\sum_{i=1}^{n} I_B(X_i), \quad B \in \mathcal{B}.$$

For a subclass $\mathcal{A} \subset \mathcal{B}$ define the empirical process indexed by $\mathcal{A}$ as

$$\alpha_n(A) = \sqrt{n}(P_n(A) - P(A)), \quad n \geq 1, \quad A \in \mathcal{A}. \tag{1.5}$$

Then α_n takes values in the space D_0, which is constructed as follows. D_0 is the linear space of all functions $f + g$, where f is the element of the space C_0, of all bounded real functions on $\mathcal{A}$ continuous with respect to the metric d_0 and g is a finite linear combination of point masses. We suppose that the space D_0 is equipped with the supremum norm. D_0 could be considered as an extension of the space D defined above.

Definition 1.2 *A class $\mathcal{A}$ is countably generated (CG), if there exists a countable subclass $\mathcal{G}$ of $\mathcal{A}$, such that for any $A \in \mathcal{A}$ there exists a sequence $\{G_n\}_{n\geq 1}$ such that $I_{G_n}(x) \to I_A(x)$, for all $x \in \mathbb{R}^k$.*

The assumption that an indexing class is CG is often made for measurability purposes; if the class $\mathcal{A}$ is CG then it is empirically measurable class for P (P-EM), that is for all n, the empirical distribution function P_n indexed by $\mathcal{A}$ is a measurable mapping from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(D_0, \mathcal{D}_0)$, where $\mathcal{D}_0$ is the σ-algebra generated by the open balls in D_0.

Definition 1.3 *A P-EM class $\mathcal{A}$ will be called a P-Donsker class if and only if*

$$\alpha_n \xrightarrow{w} B_P, \quad \textit{in} \quad (D_0, \mathcal{D}_0), \quad n \to \infty,$$

where B_P is the P-Brownian bridge, a bounded Gaussian process indexed by $\mathcal{A}$ with zero expectation and covariance $P(A_1 \cap A_2) - P(A_1)P(A_2)$, uniformly continuous with respect to d_0.

Definition 1.4 *We will call a class $\mathcal{A} \subset \mathcal{B}$ a Vapnik-Chervonenkis (VC) class if there exists a polynomial p such that from every set of N points from $\mathbb{R}^k$, the class picks out at most $p(N)$ distinct subsets. Formally, if $\{x_1, \ldots, x_N\} \subset \mathbb{R}^k$, then there are at most $p(N)$ distinct sets of the form $\{x_1, \ldots, x_N\} \cap A$ with $A \in \mathcal{A}$.*

Dudley (1978) extended results of Donsker (1952) and Doob (1949) for a VC class: he showed that under certain measurability conditions, a VC class is a P-Donsker class. A similar result was obtained in Bolthausen (1978) for the class of convex sets in $\mathbb{R}^2$: when the distribution P has a bounded density with respect to Lebesgue measure, the class $\mathcal{A}$ of all open (closed) convex subsets of some compact set $B \subset \mathbb{R}^2$ is a P-Donsker class. Then by the Skorokhod construction we obtain that

$$\sup_{A\in\mathcal{A}} |\widetilde{\alpha}_n(A) - \widetilde{B}_P(A)| \to 0 \ \text{ as } \ \text{as } \ n \to \infty. \tag{1.6}$$

We will apply this result in Chapters 3 and 4 when $\mathcal{A}$ is one of the following classes, (a), (b), (c), (a'), (b'), (c'), (d) or (e).

1.3 Minimum volume sets

Let $\mathcal{A}$ be a class of measurable subsets of $\mathbb{R}^k$, $k \geq 1$ and suppose we have n i.i.d. random vectors taking values in $\mathbb{R}^k$.

Definition 1.5 *For $t \in (0, 1]$, we will call an element A_t of the class $\mathcal{A}$ a minimum volume set (MV-set) or t-minimum volume set iff it is a smallest (with respect to Lebesgue measure) set from $\mathcal{A}$ that contains at least $\lceil nt \rceil$-observations.*

The MV-sets are used in statistics as building blocks for estimators. For this reason consistency and the rates of convergence of these sets are often studied. In robust statistics the MV-sets are used for obtaining estimators for the location and scale parameters. On the line, the smallest interval containing half of the observations (MV-set with $t = 1/2$) is called the shortest half, or "shorth". In Andrews et al. (1972) the arithmetic mean of the observations in the shortest half is considered as an estimator of location. In Rousseeuw and Leroy (1988) the scale estimator for one-dimensional samples is constructed based on the length of the "shorth". The length of the "shorth" was initially introduced and studied in Grübel (1988), where under certain conditions the asymptotic normality of these estimators was proved. In Beirlant et al. (1999) a generalized chi-square quantile plot is defined in terms of minimum volume ellipsoids (MV-sets for class of ellipsoids). Based on this plot, using concepts of generalized quantiles defined in the next section, authors derive tests for multivariate normality. Further in Rousseeuw (1985) minimum volume ellipsoids were used to construct estimates of multivariate locations and dispersion parameters in higher dimensions. Davies (1992) showed that the Rousseeuw's minimal volume estimator is consistent and asymptotically normal under certain differentiability conditions.

Minimum volume sets are considered as well when investigating the properties of the underlying distribution. In Sawitzki (1994) based on the length of the smallest interval, a graphical method is proposed for studying the mass concentration of the distribution. Under the assumptions that the underlying distribution function has a unimodal density f, with continuous first derivative, in Andrews et al. (1972) an estimator of the mode of f is defined as the midpoint of the t-minimum volume interval and the asymptotic distribution of this estimator is derived, when $t \in (0, 1)$ is a fixed constant. Unimodality of the distribution can be observed as well using the similar techniques. For this purpose the excess-mass ellipsoids were introduced and this limit distribution was investigated in, e.g., Muller and Sawitzki (1991), Nolan (1991), etc. Note that the excess-mass ellipsoids are generalizations of the minimum volume ellipsoids.

When the density of the distribution function exists, level sets can be defined and if the class $\mathcal{A}$ contains level sets then these can be approximated by MV-sets. In Chapters 3 and 4 for different indexing classes we define empirical tolerance regions as MV-sets and show under very mild assumptions that these sets are asymptotically optimal (see Lemma 3.2 and 4.2). We show as well that for some choices of the class $\mathcal{A}$ MV-sets exist and are a.s. unique. Since we use empirical process theory to

prove our results Vapnik-Chervonenkis, Glivenko-Cantelli and Donsker classes are natural candidates for $\mathcal{A}$. Minimum volume sets for more general settings are studied intensely in Polonik (1997).

1.4 Generalized quantile processes

The idea of generalizing the ordering of data in higher dimensions is not new. There exist quite a few methods for ordering multidimensional data. Instead, multivariate quantiles introduced in Einmahl and Mason (1992) offer a technique that gives more information on properties of underlying distribution by using an indexing class of sets and a real-valued function defined on this class.

Let $\mathcal{A}$ be a subset of $\mathcal{B}$ and let λ be a real-valued function defined on $\mathcal{A}$. Then the generalized quantile and generalized empirical quantile functions based on P, $\mathcal{A}$ and λ are defined as follows:

$$U(t) = \inf_{A\in\mathcal{A}}\{\lambda(A) : \ P(A) \geq t\}, \quad t \in (0,1),$$

$$U_n(t) = \inf_{A\in\mathcal{A}}\{\lambda(A) : \ P_n(A) \geq t\}, \quad t \in (0,1);$$

set $U(t) = 0$ for $t \leq 0$, and $U(t) = \lim_{s\uparrow 1} U(s)$ for $t \geq 1$. When the function U is differentiable with derivative $g \equiv \widetilde{f} \circ U$, where $\widetilde{f}$ is the derivative of the inverse of U, $\widetilde{f} = (U^{-1})'$, the generalized quantile process β_n is defined as:

$$\beta_n(t) := g(t)\sqrt{n}(U_n(t) - U(t)), \quad t \in (0,1).$$

In Einmahl and Mason (1992) generalized quantiles are studied under very general conditions. This, with the possibility of various choices of $\mathcal{A}$ and λ, makes generalized quantiles a very attractive tool when dealing with multivariate data (see, e.g., Serfling (2000)). Observe that when $\mathcal{A} = \{(-\infty, x] : x \in I\!R\}$ and $\lambda((-\infty, x]) = x$, the functions U and U_n are the classical real-valued quantile and empirical quantile functions and the process β_n is the classical one-dimensional quantile process.

Under certain conditions functions U and U_n have inverse and generalized inverse functions respectively, which can be defined as:

$$\begin{aligned} \widetilde{F}(y) &= \sup_{A\in\mathcal{A}}\{P(A) : \ \lambda(A) \leq y\}, \quad y > 0, \\ \widetilde{F}_n(y) &= \sup_{A\in\mathcal{A}}\{P_n(A) : \ \lambda(A) \leq y\}, \quad y > 0. \end{aligned}$$

These functions could also be called concentration and empirical concentration functions.

Next we generalize the concept of the minimum volume sets defined in the previous section. For $t \in [0,1]$ a set $A_t \in \mathcal{A}$ is a minimum λ set (MV-set) if $\lambda(A_t) = U_n(t)$. Observe that when λ is Lebesgue measure the set A_t is the minimum volume set from the class $\{A : P_n(A) \geq t, A \in \mathcal{A}\}$. Since in a certain sense the generalized

quantile function is a quantile transformation in higher dimensions we can consider the MV-set as the concept corresponding to the quantile.

Following Einmahl and Mason (1992) we sketch assumptions that are required to be fulfilled in order to obtain the limit theorem for the generalized quantile process;

(C_1) Let λ be continuous on $\mathcal{A}$ with respect to d_0.

(C_2) $-$ (C_3) The class $\mathcal{A}$ is CG and P-Donsker.

(C_4) For all $A \in \mathcal{A}$, $0 < P(A) < 1$.

(C_5) For each $t \in (0,1)$ there exists an MV-set from $\mathcal{A}$.

(C_6) For every $\varepsilon > 0$ there exists a $\delta > 0$, such that for $0 \leq t_1, t_2 \leq 1$, with $|t_1 - t_2| < \delta$ and t_1-minimum volume set $A_{t_1} \in \mathcal{A}$ there exists t_2-minimum volume set $A_{t_2} \in \mathcal{A}$ with $d_0(A_{t_1}, A_{t_2}) < \varepsilon$.

(C_7) The function U is strictly increasing on $(0,1)$ having inverse function $\widetilde{F}$ that has a continuous derivative $\widetilde{f}$.

(C_8) For every $\varepsilon > 0$ there exists a $\delta > 0$, such that for $A \in \mathcal{A}$, with $0 < t - \delta < P(A) < t < 1$ and $\lambda(A) < U(t)$ there exists $A' \in \mathcal{A}$ with $\lambda(A') = \lambda(A)$, $P(A') = \widetilde{F}(\lambda(A))$ and $d_0(A, A') < \varepsilon$.

Now we can state the following limit theorems for the generalized quantile process and its inverse process.

Theorem 1.3 *Under assumptions* (C_1) $-$ (C_8) *for all* $0 < a < b < 1$,

$$\sup_{a \leq t \leq b} \Big| \beta_n(t) + \sup_{\substack{P(A)=t \\ \lambda(A)=U(t)}} B_P(A) \Big| \to 0 \quad a.s. \quad as \quad n \to \infty. \tag{1.7}$$

Theorem 1.4 *Under assumptions* (C_1) $-$ (C_8),

$$\sup_{0 \leq t \leq 1} \Big| \sqrt{n} \Big(\sup_{\lambda(A) \leq U(t)} P_n(A) - t \Big) - \sup_{\substack{P(A)=t \\ \lambda(A)=U(t)}} B_P(A) \Big| \to 0 \quad a.s. \quad as \quad n \to \infty. \tag{1.8}$$

Part I

Nonparametric tolerance regions

Chapter 2

Brief historical review

In this chapter we introduce nonparametric tolerance intervals and regions and briefly review the literature.

2.1 Introduction

Predicting a specific event in the future and estimating the probability of this occurrence using the information obtained in the past, is an occupation of different fields of statistics. For example, when testing the life duration of a new product, knowing the survival times of sold products, one would like to make a reasonable statement about the warranty period of a new product. In quality control, the produced article is often considered to be effective or defective depending whether or not the certain characteristics of the product are within previously determined limits tolerated by the manufacturer and the customer. After knowing the amount of defect articles in e.g., 100 batches, manufacturer would like to predict the number of defect products in a new batch. In medical statistics predicting the efficiency of the particular treatment for a new patient or detecting diseased patients are of vital importance. There are many other examples of this kind in, e.g., quality control, reliability statistics, chemistry, etc. (see, e.g., Aitchison and Dunsmore (1975)).

For establishing the problem stated above in statistical terms, usefulness and validity will be desirable for this statistical approach to comply. Then for a sequence $X_1, \ldots, X_n$, $n \geq 1$, of i.i.d. random variables we want to find a measurable set $T = T(X_1, \ldots, X_n)$ that satisfies certain probabilistic conditions. Validity here might be considered as a requirement that the region T will capture the outcome of the future experiment. Usefulness can be described by the statement on the coverage, probability of T with respect to the future experiment. As the region T has to establish tolerated limits for the outcome X_{n+1} of future experiment, it is called a tolerance region. In the ideal case when T will contain X_{n+1} almost surely the coverage is equal to 1. Then the distribution of the coverage will be degenerate at 1. This confirms intuitively the following probabilistic restrictions on

the distribution of the coverage. The first assumption concerns the mean of the distribution of the coverage. It has to be reasonably high, close to one. This leads to mean coverage tolerance regions. The second restriction is that the 'bulk' of the coverage distribution is above some specific value. In other words we want to have a guarantee that a certain portion of the coverage distribution is above this specific value. This restriction is more a condition on the quantile of the coverage distribution rather than on the mean and defines guaranteed coverage tolerance regions (see, e.g., Guttman (1970), Aitchison and Dunsmore (1975)).

Although there is a vast literature on tolerance regions when the underlying distribution belongs to a parametric family (see, e.g., Wilks (1941, 1942), Wald (1942), Guttman (1957, 1970), Aitchison (1966), etc.), we will restrict ourselves by considering only the nonparametric case.

Starting with Wilks (1941), many papers have appeared in the literature on nonparametric tolerance regions. Wilks (1941) introduced distribution free tolerance intervals based on ordered statistics. Wilk's method was extended in Wald (1943) for tolerance regions for two or more dependent variables, for an unknown distribution. These results were extended further in Tukey (1947) for the multivariate case. Using an ordered set of arbitrary real-valued functions, the sample space was partitioned into statistically equivalent blocks. The multivariate tolerance regions are then defined using these blocks. Nonparametric tolerance regions defined by the statistically equivalent blocks were studied further in Fraser (1951, 1953) and more recently in Ackermann (1983, 1985). A totally different approach is presented in Chatterjee and Patra (1980), where a uniformly consistent density estimator is used which yields asymptotically minimal tolerance regions. The monographs Aitchison and Dunsmore (1975) and Guttman (1970) provide thorough overviews of the literature, while extensive bibliographies can be found in Jílek (1981) and Jílek and Ackermann (1989).

2.2 Definitions and setup

As we noted in the previous section there are mainly two types of tolerance regions considered in the literature; guaranteed coverage and mean coverage in the terminology of Aitchison and Dunsmore (1975) or β-content and β-expectation in the terminology of Guttman (1970). Let $X_1, \ldots, X_n$, $n \geq 1$, be a sample on a probability space $(\Omega, \mathcal{F}, I\!P)$ taking values in $I\!R^k$ from a common distribution P.

Definition 2.1 *$T(X_1, \ldots, X_n)$ is a guaranteed coverage tolerance region (β guaranteed coverage tolerance region with confidence level $1-\alpha$) if*

$$I\!P\big\{P\big(T(X_1, \ldots, X_n)\big) \geq \beta\big\} \geq 1 - \alpha.$$

Definition 2.2 *$T(X_1, \ldots, X_n)$ is a mean coverage (β mean coverage) tolerance region if*

$$I\!E\big(P\big(T(X_1, \ldots, X_n)\big)\big) \geq \beta.$$

Here $\mathbb{E}$ is the expectation on Ω defined by $\mathbb{P}$.

The probability content $P\big(T(X_1,\ldots,X_n)\big)$ is called the *coverage* of the tolerance region $T(X_1,\ldots,X_n)$. Since $T(X_1,\ldots,X_n)$ is a random set-function depending on random variables, the coverage $P\big(T(X_1,\ldots,X_n)\big)$ is a random variable as well. Hence the guaranteed coverage tolerance region contains at least 100β percent of the population with probability at least $1-\alpha$.

It is easy to see that a mean coverage tolerance region is actually a prediction region. By definition, a random set $A = A(X_1,\ldots,X_n)$ is a β-prediction region if for a new observation X, independent from the sample $X_1,\ldots,X_n$, with $\mathcal{L}(X) = P$,

$$\mathbb{P}\big(X \in A\big) \geq \beta.$$

However

$$\mathbb{P}\big(X \in A\big) = \mathbb{E}\mathbb{P}\big(X \in A \ : \ X_1,\ldots,X_n\big) = \mathbb{E}P(A).$$

2.3 Tolerance intervals: one dimensional case

Classical nonparametric tolerance intervals were introduced in Wilks (1941), one of the first important papers on tolerance regions. Wilks (1941) defined distribution-free tolerance intervals and investigated the problem of determining the sample size needed to obtain these tolerance intervals, when the parameters β and α are previously determined.

Suppose that X is a random variable defined on $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in $\mathbb{R}$, with an unknown density function f. Let $X_1,\ldots,X_n$, $n \geq 1$, be i.i.d. random variables with $\mathcal{L}(X_i) = \mathcal{L}(X)$, $i \geq 1$, and let $X_{(1)},\ldots,X_{(n)}$ be the order statistics. Define a tolerance interval as

$$T(X_1,\ldots,X_n) = [X_{(r)},\ X_{(n-r+1)}],$$

where r is a positive integer less than $(n+1)/2$ and its value is determined by the values of α and β. Wilks (1941) showed that the coverage of the tolerance region

$$P\big(T(X_1,\ldots,X_n)\big) = \int_{X_{(r)}}^{X_{(n-r+1)}} f(x)dx$$

has the Beta distribution $I(n-2r+1, 2r)$. Let F be the distribution function of X and $U_{(i)} = F(X_{(i)})$, $1 \leq i \leq n$, be the order statistics of n i.i.d. uniform-$[0,1]$ random variables, then

$$\begin{aligned}
&\mathbb{P}\Big\{P\big(T(X_1,\ldots,X_n)\big) \geq \beta\Big\} = \mathbb{P}\Big\{F(X_{(n-r+1)}) - F(X_{(r)}) \geq \beta\Big\} \\
&= \mathbb{P}\Big\{U_{(n-r+1)} - U_{(r)} \geq \beta\Big\} = \mathbb{P}\Big\{U_{(n-2r+1)} \geq \beta\Big\} \\
&\qquad = \frac{n!}{(n-2r)!(2r-1)!}\int_\beta^\infty s^{n-2r}(1-s)^{2r-1}ds.
\end{aligned}$$

Obviously the distribution of the coverage $P(T)$ does not depend on F and hence the tolerance interval is distribution-free. Observe that tolerance intervals can also be defined as

$$T(X_1, \ldots, X_n) = [X_{(r_1)},\ X_{(n-r_2+1)}], \quad r_1 + r_2 = 2r,$$

with truncation that is not necessarily symmetric as above. Note that the coverage of $[X_{(r_1)},\ X_{(r_2)}]$, have the same Beta distribution. However these intervals are most efficient when knowing the shape of the density. In Chapter 3 Tables 3.2 and 3.3 demonstrate that tolerance intervals for asymmetric densities obtained with the classical method are longer then the ones introduced in Chapter 3. It is also important to note that in the classical case it is decided *beforehand* which order statistics will define the interval.

2.4 Nonparametric tolerance regions

Since the classical procedure is based on *order* statistics it was troublesome to extend it to higher dimensions. To overcome this problem "statistically equivalent blocks" and ordering functions were introduced. Generalizing the results of Wilks (1941) and Wald (1943), multivariate tolerance regions were constructed in Tukey (1947, 1948) for continuous and discontinuous distributions, respectively. Before describing the method presented in Tukey (1947), we review the one-dimensional case from a different perspective. Using order statistics in the setting of the previous section, divide $\mathbb{R}$ into $n+1$ blocks: $(-\infty, X_{(1)}],\ (X_{(1)}, X_{(2)}], \ldots, (X_{(n)}, \infty)$. Then as noted above the sum of ℓ, $1 \le \ell \le n$, coverages of these blocks has the Beta distribution $I(n-\ell+1, \ell)$. Note also that

$$\mathbb{E}\big(P(X_{(i)}, X_{(i+1)}]\big) = \frac{1}{n+1}, \quad 0 \le i \le n, \tag{2.1}$$

with $X_{(0)} := -\infty$ and $X_{(n+1)} := \infty$. This explains why these blocks were called statistically equivalent.

To generalize this procedure to higher dimensions, Tukey had to introduce an ordering in $\mathbb{R}^k$. For an i.i.d. sample $X_1, \ldots, X_n$ taking values in $\mathbb{R}^k$ let $\varphi_1, \ldots, \varphi_n$ be measurable (deterministic) real-valued functions of X_1. Using these ordering functions $\mathbb{R}^k$ is divided into disjoint random sets (the statistically equivalent blocks) $S_1, \ldots, S_{n+1}$, with coverages $R_1, \ldots, R_{n+1}$, $(R_i = P(S_i),\ i = 1, \ldots, n+1)$.

The following definition of statistically equivalent blocks is rather complex, therefore it is followed by an example (see also Figure 2.1). By definition, at any stage i, $1 \le i \le n$, the ordering function φ_i may depend only on the values of previously used functions and on the observations defining these functions:

$$\varphi_i = \varphi_i(x\ :\ x_{(1)}^{(j)},\ j = 1, \ldots, i-1),$$

where $x_{(1)}^{(j)}$ is the observation that gives the smallest value of φ_j. The i-th statistically

equivalent block is defined as

$$S_i = \{x \ : \ \varphi_i(x \ : \ x_{(1)}^{(j)}, \ j = 1, \ldots, i-1) < V_{(1)}^{(i)}\} \subset \overline{S}_{i-1},$$

where $V_j^{(i)} = \varphi_i(X_j)$ and $V_{(j)}^{(i)}$ are order statistics of $V_j^{(i)}$ for $1 \leq j \leq n$, $\overline{S}$ denotes the closure of the set S and the observation $x_{(1)}^{(j)}$ may not be used defining subsequent blocks. Tukey (1947) showed that

$$I\!E R_i = \frac{1}{n+1}, \quad i = 1, \ldots, n+1$$

and that the sum of ℓ coverages has the Beta distribution $I(n-\ell+1, \ell)$, hence

$$I\!P\{\sum_{i=1}^{\ell} R_i < \beta\} = I_\beta(n-\ell+1, \ell), \tag{2.2}$$

where

$$I_\beta(n,m) = \frac{\Gamma(n+m)}{\Gamma(n)\Gamma(m)} \int_0^\beta x^{n-1}(1-x)^{m-1} dx$$

is the incomplete beta function, with Γ denoting the gamma function. Note that the method presented in Tukey (1947) uses a fixed sequence of the ordering functions. In Fraser (1953) this method was generalized for randomly chosen sequences of ordering functions and it was proved that the results in Tukey (1947) remain true in this case.

Let us now illustrate this by constructing a tolerance region based on the statistically equivalent blocks (see Figure 2.1). Suppose we have $n = 20$ observations

$$\big((X_1, Y_1), (X_2, Y_2), \ldots, (X_{20}, Y_{20})\big)$$

from a continuous, bivariate distribution and suppose that $r = 7$ block should be cut off. (We take these values of n and r for convenience.) Let the sequence of the ordering functions be as follows:

$$\varphi_1\big((x,y)\big) = x, \ \varphi_2\big((x,y)\big) = y, \ \varphi_3\big((x,y)\big) = -x, \ \varphi_4\big((x,y)\big) = -y,$$

$$\varphi_5\big((x,y)\big) = x - y, \ \varphi_6\big((x,y)\big) = -x + y, \ \varphi_7\big((x,y)\big) = x + y.$$

Then

$$S_1 = \big\{(x,y) \ : \ x < X_{(1)}\big\},$$

where $X_{(1)}$ is the smallest value of the first coordinates of the observations. The second block is

$$S_2 = \big\{(x,y) \in \overline{S}_1 \ : \ y < Y_{(1)}\big\},$$

since

$$V_{(1)}^{(2)} = \min_i \varphi_2((X_i, Y_i)) = \min_i Y_i = Y_{(1)}.$$

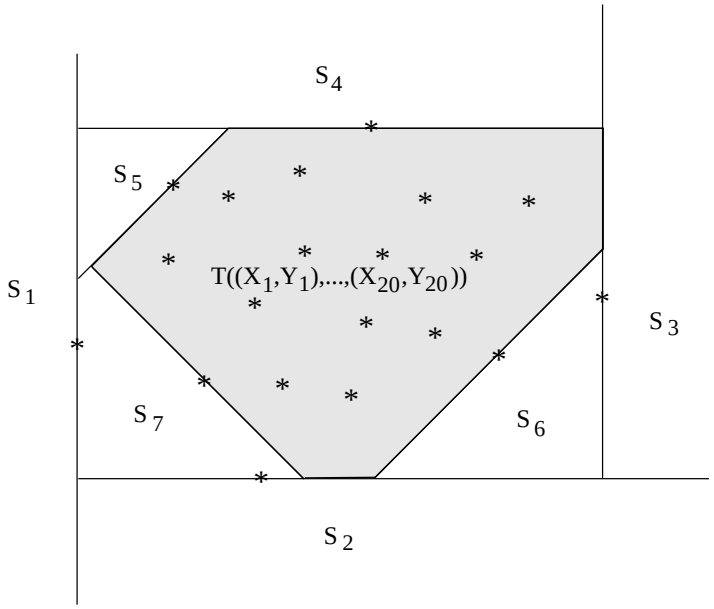

Figure 2.1: Statistically equivalent blocks.

The following five blocks are defined similarly,

$$S_3 = \big\{(x,y) \in \cap_{i=1}^{2} \overline{S}_i \ : \ -x < \min_i -X_i\big\} = \big\{(x,y) \in \cap_{i=1}^{2} \overline{S}_i \ : \ x > \max_i X_i\big\},$$

$$S_4 = \big\{(x,y) \in \cap_{i=1}^{3} \overline{S}_i \ : \ -y < \min_i -Y_i\big\} = \big\{(x,y) \in \cap_{i=1}^{3} \overline{S}_i \ : \ y > \max_i Y_i\big\},$$

$$S_5 = \big\{(x,y) \in \cap_{i=1}^{4} \overline{S}_i \ : \ x - y < \min_i (X_i - Y_i)\big\},$$

$$S_6 = \big\{(x,y) \in \cap_{i=1}^{5} \overline{S}_i \ : \ -x + y < \min_i (-X_i + Y_i)\big\}$$

and

$$S_7 = \big\{(x,y) \in \cap_{i=1}^{6} \overline{S}_i \ : \ x + y < \min_i (X_i + Y_i)\big\}.$$

The procedure ends when seven blocks are cut off. The obtained tolerance region $T\big((X_1,Y_1),\ldots,(X_{20},Y_{20})\big)$ is shaded in Figure 2.1.

Note that in practice, nonparametric tolerance regions should be constructed for at least $n \geq 100$ observations, as for $1-\alpha$ and β, values close to one should be chosen. Generally for these cases r is small. For example, when $n = 100$ for a $\beta = 0.9$ guaranteed content tolerance region with confidence level $1-\alpha = 0.95$, from Definition 2.1 and (2.2) we obtain

$$0.95 = I\!P\Big\{ \sum_{i=1}^{n-r+1} R_i \geq 0.9\Big\} = 1 - I_{0.9}(r, 100-r+1) = I_{0.1}(100-r+1, r),$$

that r is at most 5. In the example above, for $n = 20$ and $r = 7$ for the guaranteed tolerance region with a confidence level of $1 - \alpha = 0.94$ we obtain that $\beta = 0.5$.

Clearly the choice of ordering functions is arbitrary so that by taking different sequences of these functions one can obtain totally different regions with different shapes. Moreover these tolerance regions are not necessarily asymptotically minimal, as defined below. Chatterjee and Patra (1980) defined the concept of asymptotically minimal tolerance regions for the multivariate case and constructed a sequence of tolerance regions that satisfies this criteria. For $n \geq 1$ let $T_n(X_1, \ldots, X_n)$ be β-guaranteed coverage tolerance region with confidence level α_n, for i.i.d. random variables $X_1, \ldots, X_n$ taking values in $\mathbb{R}^k$, having an unknown density f. By assumption α_n may depend on f.

Definition 2.3 *If* $\liminf_{n\to\infty} \alpha_n \geq \alpha$ *for some* $\alpha \in (0, 1]$ *and any* f *then* T_n *represents a sequence of* β *guaranteed coverage tolerance regions with asymptotic confidence level* α.

Definition 2.4 *Assuming that the density* f *has no flat parts, a sequence of* β*-guaranteed coverage tolerance regions* T_n *with asymptotic confidence level* α *is called asymptotically minimal (optimal) if*

$$\lambda(T_n \triangle G_{f,\beta}) \xrightarrow{I\!P} 0 \quad n \to \infty,$$

where λ *denotes Lebesgue measure,*

$$G_{f,\beta} = \{x : f(x) > \gamma_\beta\}$$

is the γ_β*-level set of* f *and* γ_β *is this* $(1 - \beta)$ *quantile.*

Under certain conditions Chatterjee and Patra (1980) showed that for uniformly consistent estimators $\overline{f}_n$ of the density f, there exists a sequence $\overline{\gamma}_{n,\beta}$ such that the sequence of β-guaranteed coverage tolerance regions

$$\overline{T_n} = \{x : \overline{f}_n(x) > \overline{\gamma}_{n,\beta}\},$$

with asymptotic confidence level α is asymptotically minimal, when $\overline{\gamma}_{n,\beta}$ converge in probability to γ_β.

2.5 Directional tolerance regions

Although there is a huge literature on directional data and tolerance regions in general, not much seems to be known on tolerance regions for directional data. Based on the idea of statistically equivalent blocks Ackermann (1985) constructed tolerance regions for circular data.

Suppose $\theta_1, \ldots, \theta_n$, $0 \leq \theta_i < 2\pi$, $n \geq 1$ are i.i.d. circular data measured in angles. Then each θ_i can be identified with a point Z_i on the unit circle. Define statistically equivalent blocks as the arcs

$$S_i = (Z_{(i-1)}, Z_{(i)}], \quad i = 1, \ldots, n,$$

where the $Z_{(i)}$'s are points on the circle that correspond to the order statistics $\theta_{(i)}$ of the θ_i, $i = 1, \ldots, n$ and $Z_{(0)} = Z_{(n)}$. Here and below everywhere a half open arc $(A, B]$ is defined to be the set of all points on the circle that lie between A and B taking anti-clockwise direction and including the point B. Trivially the closed arc $[A, B] = \{A\} \cup (A, B]$.

Based on Tukey (1947) it is shown in Ackermann (1985) that the sum of r coverages, $\sum_{i=1}^{r} P\{S_i\}$ has a Beta distribution. A median direction μ, $0 \leq \mu \leq 2\pi$ for the circular density f is defined by the equation

$$\int_{\mu}^{\mu+\pi} f(\theta)d\theta = \int_{\mu+\pi}^{\mu+2\pi} f(\theta)d\theta = \frac{1}{2},$$

where $f(\mu) > f(\mu + \pi)$ (see, e.g., Mardia (1972)). Suppose n is even. Set $\hat{\mu}$ to be an estimator of the median direction and let $\theta_{(i-1)} < \hat{\mu} \leq \theta_{(i)}$. Thus the block $S_i = (Z_{(i-1)}, Z_{(i)}]$ contains the point on the circle corresponding to the estimator of the median direction $\hat{\mu}$. Then the tolerance region can be defined as a union of r adjacent blocks

$$\mathcal{S} = (Z_{((i-1-r_2+n)(\mathrm{mod}\, n))}, Z_{((i+r_1)(\mathrm{mod}\, n))}],$$

where $r_1 + r_2 + 1 = r \leq n$. Suppose now that n is odd. Set $\theta_{(i)}$ to be the estimated median direction, then

$$\mathcal{S} = (Z_{((i-r_2+n)(\mathrm{mod}\, n))}, Z_{((i+r_1)(\mathrm{mod}\, n))}]$$

is the tolerance region and $r_1 + r_2 = r \leq n$. However the exact or asymptotic behavior of the tolerance regions has not been studied in this setting, but only when the true median direction is known, which is typically not the case in practice.

Chapter 3

Small nonparametric tolerance regions

This chapter is an extended version of Di Bucchianico, Einmahl, and Mushkudiani (2000).

In this chapter a new, natural way of constructing nonparametric multivariate tolerance regions is presented. In the spirit of the shorth (see, e.g., Rousseeuw and Leroy (1988), Grübel (1988)), tolerance intervals are defined as shortest intervals, that contain a certain number of observations. This idea can be extended in a natural way to higher dimensions, by replacing the class of intervals by a general class of indexing sets, which specializes to the classes of ellipsoids, hyperrectangles or convex sets and the classes of finite unions of these sets. Furthermore we show that the procedure presented here is asymptotically correct and the tolerance regions have invariance properties. We also illustrate our approach by computing tolerance regions for leukemia diagnosis from bi- and trivariate observations from blood counts and by investigating the finite sample properties of the tolerance regions through a simulation study.

3.1 Notations and preliminary results

Below we specify our setup and notation. We also state some preliminary results for convenient reference later on. Let $X_1, \ldots, X_n$, $n \geq 1$, be i.i.d. $\mathbb{R}^k$-valued random vectors defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with a common probability distribution P, absolutely continuous with respect to Lebesgue measure, and corresponding distribution function F. Let $\mathcal{B}$ be the σ-algebra of Borel sets on $\mathbb{R}^k$ and let d_0 be the pseudo-metric on $\mathcal{B}$ defined in Chapter 1. Denote by P_n the empirical distribution:

$$P_n(B) = \frac{1}{n} \sum_{i=1}^{n} I_B(X_i), \quad B \in \mathcal{B}.$$

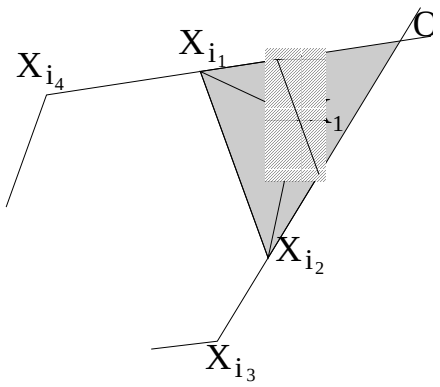

Figure 3.1: Uniqueness of minimum area convex set.

As we mentioned already, tolerance regions defined here are the MV-sets (see Section 1.3). Below we consider several classes of our interest and show that the MV-sets from these classes exist and are a.s. unique.

Let $\mathcal{E}$ be the class of all closed ellipsoids in $\mathbb{R}^k$. Fix $t_0 \in (0,1)$ and $C \in \mathbb{R}$. Set $p_n = t_0 + \frac{C}{\sqrt{n}}$. For n large enough, we need existence and uniqueness of an ellipsoid $A_{n,t_0,C} \in \mathcal{E}$ of minimum volume such that $P_n(A_{n,t_0,C}) \geq p_n$, almost surely. In other words, $A_{n,t_0,C}$ should contain at least $\lceil np_n \rceil$ observations. In the notation of Section 1.3 $A_{n,t_0,C} = A_{t_0+\frac{C}{\sqrt{n}}}$. The sets $A_{n,t_0,C}$ are our candidate *tolerance regions*. The existence and a.s. uniqueness of such an ellipsoid $A_{n,t_0,C}$ was proved in Davies (1992). There are between $k+1$ and $k(k+3)/2$ points on the boundary of $A_{n,t_0,C}$ in dimension k (see Silverman and Titterington (1980)) and hence,

$$t_0 + \frac{C}{\sqrt{n}} \leq P_n(A_{n,t_0,C}) < t_0 + \frac{C}{\sqrt{n}} + \frac{k(k+3)}{2n} \quad \text{a.s. .} \tag{3.1}$$

However with some more effort it can be shown that a minimum volume ellipsoid that contains *at least* m out of n points, contains *exactly* m points, a.s. (see Lemma 3.3). This result seems not to be present in the literature. It yields that

$$P_n(A_{n,t_0,C}) = \frac{1}{n}\left\lceil n\left(t_0 + \frac{C}{\sqrt{n}}\right)\right\rceil \quad \text{a.s. .} \tag{3.2}$$

Let $\mathcal{R}$ be the class of all closed hyperrectangles with faces parallel to the coordinate hyperplanes. It is easy to adapt the proof of Davies (1992) to $\mathcal{R}$. Hence, there exists an a.s. unique smallest volume hyperrectangle $A_{n,t_0,C} \in \mathcal{R}$, with $P_n(A_{n,t_0,C}) \geq p_n$. Since with probability one, all hyperplanes parallel to the coordinate axes contain at most one observation, the equality in (3.2) holds here too.

Consider now the existence and a.s. uniqueness problem of $A_{n,t_0,C}$ for $\mathcal{C}$, the class of all closed convex sets in $\mathbb{R}^2$. It is a well-known fact that the convex hull of

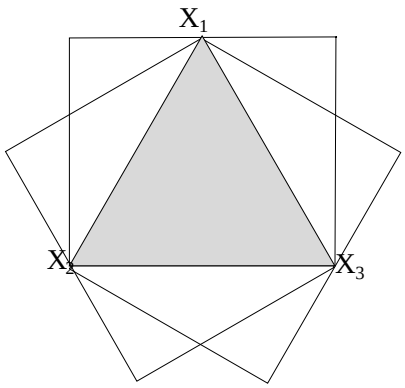

Figure 3.2: Illustration to Remark 1.

$\mathcal{X} = \{X_1, \ldots, X_n\}$ is a bounded polyhedral set in $\mathbb{R}^2$ (i.e., a bounded set which is the intersection of finitely many half-planes, see e.g., Webster (1994), Theorem 3.2.5), and thus a polygon. Since the convex hull of $\mathcal{X}$ is the smallest (with respect to set inclusion) convex set containing $\mathcal{X}$, it follows that the closed convex hull of $\mathcal{X}$ is the a.s. unique smallest area closed convex set containing $\mathcal{X}$. As the number of subsets of $\mathcal{X}$ is finite, the existence of a smallest area convex subset containing $\lceil np_n \rceil$ points from $\mathcal{X}$ is assured. Hence, it is left to show that with probability 1, any two different convex hulls of subsets of the sample will have different areas. Suppose we have two sets of vertices $\{X_{i_1}, \ldots, X_{i_\ell}\}$ and $\{X_{j_1}, \ldots, X_{j_k}\}$, $3 \le \ell$, $k \le n$, with convex hulls A_1 and A_2, respectively. Without loss of generality we assume that X_1 is a vertex of A_1, but not of A_2. If we condition on $\{X_2, \ldots, X_n\}$, then we have to show that for any positive v

$$\mathbb{P}\{X_1 : V(A_1) = v \mid X_2, \ldots, X_n\} = 0, \tag{3.3}$$

where $V(A_1)$ denotes the area of A_1. Since A_1 is convex, X_1 lies in the interior of the triangle $X_{i_1} O X_{i_2}$ (see Figure 3.1), for any neighboring vertices X_{i_1} and X_{i_2}. As the area of A_1 is fixed, X_1 can be only on some interval parallel to $X_{i_1} X_{i_2}$. (Actually, we assumed $5 \le \ell \le n$, but a similar argument works for $\ell = 3$ or $\ell = 4$.) Hence, we see that (3.3) holds. Finally, it is obvious that (3.2) holds for $\mathcal{C}$.

Remark 3.1 *Observe that unlike for the classes above, the minimum volume problem has no unique solution for the case of all hyperrectangles in $\mathbb{R}^k$. Consider a random sample of size n in, e.g., $\mathbb{R}^2$. Then with positive probability, there are 3 sample points that form an acute triangle such that the remaining $n - 3$ sample points are in the interior of that triangle. In this case, there are 3 minimal area rectangles that contain the sample (see Figure 3.2).*

Here are some more definitions and results. By the *Blaschke Selection Principle* (see e.g. Webster (1994), Theorem 2.7.10), every sequence of non-empty compact

convex sets contained in a compact subset of $\mathbb{R}^k$ has a subsequence that converges in the Hausdorff metric to some non-empty compact convex set in $\mathbb{R}^k$. It is easy to show that the Blaschke Selection Principle holds as well for the sequence of finite unions of compact convex sets. By Shephard and Webster (1965), the Hausdorff and the symmetric difference metric $d(A, B) := V(A \triangle B)$, where V denotes volume (Lebesgue measure), are equivalent on the class of all compact convex subsets of $\mathbb{R}^k$ with non-empty interior. Hence, we have convergence in the Hausdorff metric if and only if we have convergence in the symmetric difference metric d.

In the setting of Section 1.4 when λ is Lebesgue measure V we have for an MV-set $A_{n,t_0,C}$ that

$$A_{n,t_0,C} = \text{argmin}\{V(A) : P_n(A) \geq t_0 + \frac{C}{\sqrt{n}}, A \in \mathcal{A}\}$$

and hence

$$\lambda(A_{n,t_0,C}) = V(A_{n,t_0,C}) = U_n(t_0 + \frac{C}{\sqrt{n}}),$$

where U_n in the generalized empirical quantile function. In this chapter we consider the generalized quantile functions only in the case when λ is Lebesgue measure. Then for any class $\mathcal{A} \subset \mathcal{B}$.

3.2 Limit theorems for small tolerance regions

Here we present the asymptotic results on small tolerance regions. Recall the notation of the previous section. Let $\mathcal{A}$ be a class of Borel-measurable subsets of $\mathbb{R}^k$. (We assume that $\mathcal{A}$ is such that no measurability problems occur.)

Theorem 3.1 *Fix $t_0 \in (0, 1)$ and let $C \in \mathbb{R}$. Assume the following conditions are fulfilled:*

C1) $\mathcal{A}$ is P-Donsker class,
C2) There exists an $n_0 \in \mathbb{N}$, such that for all $n \geq n_0$, with probability 1, there exists a unique set $A_{n,t_0,C} \in \mathcal{A}$ with minimum volume and

$$P_n(A_{n,t_0,C}) \geq t_0 + \frac{C}{\sqrt{n}} \ ,$$

C3) There exists a sequence $C_n \downarrow C$, such that for all $n \geq 1$,

$$P_n(A_{n,t_0,C}) \leq t_0 + \frac{C_n}{\sqrt{n}} \quad a.s. \ ,$$

C4) A_{t_0}, the set in $\mathcal{A}$ with minimum volume and $P(A_{t_0}) = t_0$, exists, is unique, and

$$d(A_{n,t_0,C}, A_{t_0}) \xrightarrow{\mathbb{P}} 0 \quad (n \to \infty).$$

Then we have

$$\sqrt{n}(t_0 - P(A_{n,t_0,C})) + C \xrightarrow{d} Z\sqrt{t_0(1-t_0)} \quad (n \to \infty), \tag{3.4}$$

where Z is a standard normal random variable.

Proof For each $n \geq 1$, let α_n be the empirical process indexed by $\mathcal{A}$. Since $\mathcal{A}$ is a P-Donsker class by the Skorokhod construction, in the notation of Chapter 1, we obtain that there exists a probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{I\!P})$ carrying a version $\widetilde{B}_P$ of B_P and versions $\widetilde{\alpha}_n$ of α_n, for all $n \in I\!N$, such that

$$\sup_{A \in \mathcal{A}} |\widetilde{\alpha}_n(A) - \widetilde{B}_P(A)| \to 0 \ \text{ a.s.}, \quad n \to \infty. \tag{3.5}$$

Henceforth, we will drop the tildes from the notation, for notational convenience. By C2) we obtain

$$\sqrt{n}(P_n(A_{n,t_0,C}) - P(A_{n,t_0,C})) - B_P(A_{n,t_0,C}) \to 0 \ \text{ a.s.}, \quad n \to \infty. \tag{3.6}$$

Combining this with C3) yields

$$\sqrt{n}(t_0 - P(A_{n,t_0,C})) + C - B_P(A_{n,t_0,C}) \to 0 \ \text{ a.s.}, \quad n \to \infty. \tag{3.7}$$

From C4) we have that $d_0(A_{n,t_0,C}, A_{t_0}) \xrightarrow{I\!P} 0$ and hence, since B_P is continuous with respect to d_0,

$$B_P(A_{n,t_0,C}) \xrightarrow{I\!P} B_P(A_{t_0}) \ \ n \to \infty. \tag{3.8}$$

From (3.7) and (3.8) we now obtain that

$$\sqrt{n}(t_0 - P(A_{n,t_0,C})) + C \xrightarrow{I\!P} B_P(A_{t_0}) \ \ n \to \infty.$$

Observing that

$$B_P(A_{t_0}) \stackrel{d}{=} Z\sqrt{t_0(1-t_0)},$$

completes the proof. □

The following theorems, which are corollaries to Theorem 3.1, are actually main results about small tolerance regions. In fact, we will show that the sets $A_{n,t_0,C}$, for suitable C, are asymptotic tolerance regions. Theorem 3.2 gives the result for guaranteed coverage tolerance regions, whereas Theorem 3.3 deals with mean coverage tolerance (or prediction) regions. We show that the guaranteed coverage tolerance regions have indeed asymptotically the correct confidence level, whereas the mean coverage tolerance regions have the correct mean coverage with error rate $o(1/\sqrt{n})$. These results are new and of interest in any finite dimension, *including* dimension one. The numbers t_0 and $1-\alpha$ denote the (desired) coverage and confidence level, respectively.

Theorem 3.2 *Fix $\alpha \in (0,1)$ and let $C = C(\alpha)$ be the $(1-\alpha)$-th quantile of the distribution of $Z\sqrt{t_0(1-t_0)}$. Under the conditions of Theorem 3.1 we have*

$$\lim_{n\to\infty} I\!P\{P(A_{n,t_0,C}) \geq t_0\} = 1-\alpha. \tag{3.9}$$

Proof By Theorem 3.1, for all $x \in I\!R$, we have

$$I\!P\{\sqrt{n}(t_0 - P(A_{n,t_0,C})) + C \leq x\} \to I\!P\{Z\sqrt{t_0(1-t_0)} \leq x\}, \quad n \to \infty.$$

Hence, taking $x = C$, we obtain

$$\lim_{n\to\infty} I\!P\{P(A_{n,t_0,C}) \geq t_0\} = I\!P\{Z\sqrt{t_0(1-t_0)} \leq C\} = 1-\alpha.$$

□

Theorem 3.3 *If the conditions of Theorem 3.1 hold and $\sqrt{n}(t_0 - P(A_{n,t_0,0}))$ is uniformly integrable, then*

$$I\!EP(A_{n,t_0,0}) = t_0 + o\left(\frac{1}{\sqrt{n}}\right), \quad n \to \infty. \tag{3.10}$$

Note that $I\!EP(A_{n,t_0,C}) \to t_0$, $n \to \infty$, for *every* $C \in I\!R$.

Proof Theorem 3.1 with $C = 0$ yields

$$\sqrt{n}(t_0 - P(A_{n,t_0,0})) \xrightarrow{d} Z\sqrt{t_0(1-t_0)}, \ n \to \infty. \tag{3.11}$$

By assumption $\sqrt{n}(t_0 - P(A_{n,t_0,0}))$ is uniformly integrable, hence

$$I\!E\sqrt{n}(t_0 - P(A_{n,t_0,0})) \longrightarrow I\!E(Z\sqrt{t_0(1-t_0)}) = 0, \quad n \to \infty,$$

which is the statement of the theorem. □

In the next theorem, we will specialize our general results to three natural and relevant indexing classes, which satisfy the conditions of the above theorems. From the point of view of applications, this is the main result of this chapter. In the sequel, $\mathcal{A}$ will be one of the following classes: all closed

(a) ellipsoids,

(b) hyperrectangles with faces parallel to the coordinate hyperplanes,

(c) convex sets (for $k = 2$)

that have probability strictly between 0 and 1.

These classes of sets are very natural for constructing nonparametric tolerance regions. The class of ellipsoids in (a) is a good choice, since elliptically contoured distributions are considered to be natural and important in probability and statistics. The multivariate normal distribution is of course a prominent example. One should choose the parallel hyperrectangles of (b) as indexing class, if it is desirable, like in many applications, to have a multivariate tolerance region that can be decomposed into (easily interpretable) tolerance intervals for the individual components of the random vectors. The convex sets of (c), which reduce to tolerance regions that are convex polygons, are very natural, since when taking the convex hull of a finite set of data points, one hardly feels the restriction due to the underlying indexing class.

Theorem 3.4 *Fix $t_0 \in (0,1)$. If the density f of the distribution function F is positive on some connected, open set $\mathcal{S} \subset \mathbb{R}^k$ and $f \equiv 0$ on $\mathbb{R}^k \backslash \mathcal{S}$, and if A_{t_0}, the set in $\mathcal{A}$ with minimum volume and $P(A_{t_0}) = t_0$, exists and is unique, then we have for the cases (a) and (b) that (3.4),(3.9) and (3.10) hold.*
If $k = 2$ and, in addition, f is bounded, then (3.4),(3.9) and (3.10) also hold for case (c).

We next present two lemmas. Lemma 3.2 is crucial for the proof of Theorem 3.4, whereas Lemma 3.1 is needed for the proof of Lemma 3.2. Until further notice we shall, for case (c), tacitly *restrict* ourselves to those closed convex sets that are contained in some large circle B (which will be specified later on). In the proof of Theorem 3.4 we will show that this restriction can be removed. For Lemma 3.1, recall the functions U and $\widetilde{F}$, defined in Section 1.4, when λ is Lebesgue measure.

Lemma 3.1 *Under the assumptions of Theorem 3.4 we have for the cases (a), (b) and (c), that the functions U and $\widetilde{F}$ are inverses of each other. Hence, U is continuous on $(0,1)$, $\widetilde{F}$ is continuous on $\mathbb{R}^+$, and they are both strictly increasing.*

Proof We first prove the continuity of U. Note that absolute continuity of P implies that

$$U(t) = \inf_{A \in \mathcal{A}} \{V(A) : \ P(A) > t\}, \text{ for any } t \in (0,1),$$

and

$$\widetilde{F}(y) = \sup_{A \in \mathcal{A}} \{P(A) : \ V(A) < y\}, \quad \text{for any} \quad y \in \mathbb{R}^+.$$

Let us now take an arbitrary decreasing sequence $t_m \downarrow t$, where $t_m, t \in (0,1)$. Consider the sequence of sets

$$D_m = \{V(A) : \ P(A) > t_m, \quad A \in \mathcal{A}\}.$$

It is easy to see that this is a nested sequence of sets, with limit set

$$\bigcup_{m=1}^{\infty} D_m = \{V(A) : \ P(A) > t\}$$

and hence,

$$\lim_{m\to\infty} U(t_m) = \lim_{m\to\infty} \inf D_m = \inf_{A\in\mathcal{A}}\{V(A):\ P(A) > t\} = U(t).$$

In case $t_m \uparrow t$ the proof is analogous. Similar arguments yield continuity of $\widetilde{F}$.

Note that absolute continuity of P also implies that

$$U(t) = \inf_{A\in\mathcal{A}}\{V(A):\ P(A) = t\},\ \text{for any } t \in (0,1), \tag{3.12}$$

and

$$\widetilde{F}(y) = \sup_{A\in\mathcal{A}}\{P(A):\ V(A) = y\},\quad \text{for any}\quad y \in I\!R^+. \tag{3.13}$$

It follows from (3.12) and (3.13) that U is the generalized inverse of $\widetilde{F}$, i.e.

$$U(t) = \inf\{y : \widetilde{F}(y) \geq t\}\ \text{ for any } t \in (0,1).$$

Hence, clearly both U and $\widetilde{F}$ are strictly increasing and continuous. Thus we conclude that they are inverses of each other. □

Note that an in-probability-version of the second lemma, with $k = 1$ and $C = 0$, can be found in Beirlant and Einmahl (1995), Corollary 1; see also Einmahl and Mason (1992).

Lemma 3.2 *Under the assumptions of Theorem 3.4 we have for the cases (a), (b) and (c) that with probability one*

$$d(A_{n,t_0,C}, A_{t_0}) \to 0,$$

and hence $d_0(A_{n,t_0,C}, A_{t_0}) \to 0\ \ (n \to \infty)$.

Proof Since the classes (a), (b) and (c) are P-Donsker (see Section 1.2), (3.5) holds for all three cases. Since B_P is bounded, this yields

$$\sup_{A\in\mathcal{A}} |P_n(A) - P(A)| \to 0\ \text{ a.s.},\quad n \to \infty. \tag{3.14}$$

It now trivially follows from (3.14) and the definitions of $\widetilde{F}_n$ and $\widetilde{F}$ that

$$\sup_{y>0} |\widetilde{F}_n(y) - \widetilde{F}(y)| \to 0\ \text{ a.s.},\quad n \to \infty. \tag{3.15}$$

Let $\ell < 1$ be arbitrary. Since $U(t)$ is continuous, increasing and nonnegative on $(0,1)$ by Lemma 3.1, it is uniformly continuous on $(0,\ell]$, and thus

$$\sup_{t\in(0,\ell]} \left| U\left(t + \frac{C}{\sqrt{n}}\right) - U(t)\right| \to 0,\quad n \to \infty. \tag{3.16}$$

We now want to prove that

$$\sup_{t\in(0,\ell]} |U_n(t) - U(t)| \to 0, \quad n \to \infty. \tag{3.17}$$

For any $\varepsilon > 0$ we have from (3.15) that for n large enough

$$\widetilde{F}(y) - \varepsilon \le \widetilde{F}_n(y) < \widetilde{F}(y) + \varepsilon \ \text{ for all } \ y > 0 \ \text{ a.s. .}$$

By Lemma 3.1, U is the generalized inverse of $\widetilde{F}$. It is easy to see that U_n and $\widetilde{F}_n$ are generalized inverses. Hence, we obtain from the above inequalities that

$$U(t-\varepsilon) \le U_n(t) \le U(t+\varepsilon) \text{ for all } \ t \in (0,1) \ \text{ a.s. .}$$

Since U is uniformly continuous, there exists $\delta > 0$ such that

$$U(t) - \delta \le U(t-\varepsilon) \le U_n(t) \le U(t+\varepsilon) \le U(t) + \delta \quad \text{for any } \ t \in (0,\ell] \ \text{ a.s.,}$$

which immediately yields (3.17). From (3.16) and (3.17) it follows that

$$\sup_{t\in(0,\ell]} \left| U_n\left(t + \frac{C}{\sqrt{n}}\right) - U(t) \right| \to 0 \ \text{ a.s., } \ n \to \infty. \tag{3.18}$$

Now let us return to the sets given in the statement of the lemma:

- $A_{n,t_0,C}$, the a.s. unique MV-set from $\mathcal{A}$ with $P_n(A_{n,t_0,C}) \ge t_0 + \frac{C}{\sqrt{n}}$ (and hence, $V(A_{n,t_0,C}) = U_n(t_0 + \frac{C}{\sqrt{n}})$),
- A_{t_0}, the unique smallest element of $\mathcal{A}$ with $P(A_{t_0}) = t_0$ (and $V(A_{t_0}) = U(t_0)$).

By (3.2),

$$P_n(A_{n,t_0,C}) \to t_0 \ \text{ a.s., } \ n \to \infty,$$

and thus by (3.14)

$$P(A_{n,t_0,C}) \to t_0, \ \text{ a.s., } \ n \to \infty.$$

From (3.18) we have

$$\lim_{n\to\infty} V(A_{n,t_0,C}) = V(A_{t_0}) \ \text{ a.s. .}$$

To apply the Blaschke Selection Principle to the sequence $\{A_{n,t_0,C}\}_{n\ge1}$ we have to show that it is uniformly bounded a.s., i.e. for each $\omega \in \Omega_0$, with $I\!P(\Omega_0) = 1$, there exists a compact set, that contains all $A_{n,t_0,C}(\omega)$'s. Suppose the contrary, that is, for any $\omega \in \Omega'$, with $I\!P(\Omega') > 0$ we have that there exists a subsequence $n_\ell := n_\ell(\omega)$ such that for all $\ell \ge 1$, $A_{n_\ell,t_0,C}$ has an interior point a_{n_ℓ} with $d(a_{n_\ell}, O) \to \infty$, as $\ell \to \infty$, where O denotes the origin. Next since $P(A_{n,t_0,C}) \to t_0$, we will have that for all ℓ large enough $A_{n_\ell,t_0,C}$ has an interior point b_{n_ℓ}, such that $b_{n_\ell} \in A_{n_\ell,t_0,C} \cap B_{O,r}$, where $B_{O,r}$ is a closed ball with the center at the origin O, the radius

r and $P(B_{O,r}) > 1 - t_0$. Thus for ℓ large enough we have that $a_{n_\ell}, b_{n_\ell} \in A_{n_\ell,t_0,C}$, $b_{n_\ell} \in B_{O,r}$ and as $\ell \to \infty$, $d(a_{n_\ell}, O) \to \infty$, which yields

$$\lim_{\ell\to\infty} \text{diam } (A_{n_\ell,t_0,C}) = \infty, \tag{3.19}$$

where for any measurable set A, diam $(A) := \sup_{x,y\in A} d(x,y)$. Let us now recall that $A_{n,t_0,C}$ is an ellipsoid or a parallel hyperrectangle and

$$\lim_{n\to\infty} V(A_{n,t_0,C}) = V(A_{t_0}),$$

then by (3.19)

$$\lim_{\ell\to\infty} V(A_{n_\ell,t_0,C} \cap B_{O,r}) = 0$$

and consequently $\lim_{\ell\to\infty} P(A_{n_\ell,t_0,C} \cap B_{O,r}) = 0$. Hence we obtain that

$$\lim_{\ell\to\infty} P(A_{n_\ell,t_0,C}) = \lim_{\ell\to\infty} P(A_{n_\ell,t_0,C} \cap B_{O,r}) + \lim_{\ell\to\infty} P(A_{n_\ell,t_0,C} \cap B^c_{O,r}) < t_0$$

which is impossible.

Then by the Blaschke Selection Principle the sequence $\{A_{n,t_0,C}\}_{n\geq 1}$ has at least one limit set. So there exists a subsequence $\{A_{n_k,t_0,C}\}_{k\geq 1}$ and a non-empty closed convex set A^* (an element of the indexing class (a), (b) or (c), respectively), such that

$$\lim_{k\to\infty} V(A_{n_k,t_0,C} \triangle A^*) = 0 \quad \text{a.s. .}$$

Hence, $V(A_{n_k,t_0,C}) \to V(A^*)$, and thus $V(A^*) = U(t_0)$ a.s.. Using that P is absolutely continuous with respect to Lebesgue measure, it is easy to see that $P(A^*) = t_0$.

So we have for the limit set A^* that

$$V(A^*) = U(t_0) \quad \text{and} \quad P(A^*) = t_0 \quad \text{a.s.},$$

but by assumption there exists a unique set A_{t_0} satisfying these two equations. Hence, any limit set A^* of the sequence $\{A_{n,t_0,C}\}_{n\geq 1}$ is equal to A_{t_0}, and thus the sequence itself converges to A_{t_0} (a.s.). □

Proof of Theorem 3.4 We will check the conditions C1)-C4) of Theorem 3.1. We first prove (3.4) and (3.9), for the cases (a), (b) and the restricted case (c). As noted in the proof of Lemma 3.2 we have that C1) holds. In Section 3.1 it is shown that C2) holds as well; C3) follows from (3.2). The first part of C4) is an assumption of Theorem 3.4; Lemma 3.2 yields the second part of condition C4). This completes the proof of (3.4) and (3.9) for these cases.

Now consider the unrestricted case (c). We will prove (3.4) and (3.9). Let us first construct a proper circle B, as used in the definition of the restricted class. Let B_{t_0} be a circle with radius r, say, such that $A_{t_0} \subset B_{t_0}$ and $P(B_{t_0}) > t_0 \vee (1 - t_0)$. For sake of notation, any space V_γ between two parallel lines in $\mathbb{R}^2$ at distance γ is

said to be a γ-strip. Note that for a probability measure P with density f we have that

$$\lim_{\gamma \to 0} \sup_{V_\gamma} P(V_\gamma) = 0, \tag{3.20}$$

where each supremum runs over all γ-strips. Therefore there exists a γ_0 satisfying the inequality

$$\sup_{V_{\gamma_0}} P(V_{\gamma_0}) \leq \tfrac{1}{2}\, t_0, \tag{3.21}$$

where the supremum runs over all γ_0-strips. Now choose B to be a circle with the same center as B_{t_0}, but with radius $R > \frac{8}{\gamma_0} U(t_0) + r$, where γ_0 satisfies (3.21).

Next we show that $A_{n,t_0,C} = A^*_{n,t_0,C}$ for large n a.s., where $A^*_{n,t_0,C}$ is defined similarly as $A_{n,t_0,C}$ but for the restricted class. In other words we have to show that for n large enough $A_{n,t_0,C} \subset B$ almost surely. Observe that

$$\begin{aligned} \lim_{n\to\infty} P_n(A_{n,t_0,C}) &= t_0 \ \text{ a.s.}, \\ \lim_{n\to\infty} P_n(B^c_{t_0}) = P(B^c_{t_0}) &< t_0 \wedge (1-t_0) \ \text{ a.s. .} \end{aligned}$$

So, if there exists with positive probability a subsequence $\{A_{n_k,t_0,C}\}_{k\geq 1}$ such that $A_{n_k,t_0,C} \not\subset B$ for all k, then $A_{n_k,t_0,C}$ contains an element of B_{t_0} as well as an element of B^c eventually. Because the γ_0-strips form a VC class, we have that

$$\lim_{n\to\infty} \sup_{V_{\gamma_0}} P_n(V_{\gamma_0}) \leq \tfrac{1}{2}\, t_0 \ \text{ a.s. .}$$

Hence, $A_{n_k,t_0,C}$ eventually contains a triangle with area $\frac{\gamma_0}{4}(R-r) > 2U(t_0)$. However, this can not happen because of the Glivenko-Cantelli theorem. This proves (3.4) and hence (3.9).

Finally we prove (3.10) for all three cases. It suffices to show that $\sqrt{n}(t_0 - P(A_{n,t_0,0}))$ is uniformly integrable. It follows from (3.2) that

$$\begin{aligned} &\left|\sqrt{n}(t_0 - P(A_{n,t_0,0}))\right| \\ &\leq \left|\sqrt{n}(P_n(A_{n,t_0,0}) - P(A_{n,t_0,0}))\right| + \left|\sqrt{n}(t_0 - P_n(A_{n,t_0,0}))\right| \\ &\leq \sup_{A\in\mathcal{A}} |\sqrt{n}(P_n(A) - P(A))| + 1 \ \text{ a.s..} \end{aligned}$$

Therefore it suffices to establish uniform integrability of

$$Y_n := \sup_{A\in\mathcal{A}} |\sqrt{n}(P_n(A) - P(A))|.$$

Note that if Y is a non-negative random variable then

$$I\!E Y = \int_0^\infty I\!P\{Y > y\} dy.$$

Hence,

$$I\!E Y I_{[Y>a]} = \int_0^\infty I\!P\{Y I_{[Y>a]} > y\} dy = a I\!P\{Y > a\} + \int_a^\infty I\!P\{Y > y\} dy.$$

Moreover, for the cases (a) and (b) (as then $\mathcal{A}$ is a VC class), using Theorem 2.11 of Alexander (1984), we have for $\lambda \geq 8$ and $C_1, C_2 \in (0, \infty)$ that

$$\mathbb{P}\{\sup_{A \in \mathcal{A}} |\sqrt{n}(P_n(A) - P(A))| > \lambda\} \leq C_1 \lambda^{C_2} \exp(-2\lambda^2). \tag{3.22}$$

For large enough λ, the right-hand side of (3.22) is less than $\exp(-\lambda^2)$. Let $\varepsilon > 0$. Then for a large enough:

$$\mathbb{E}Y_n I_{[Y_n > a]} = a\mathbb{P}\{Y_n > a\} + \int_a^\infty \mathbb{P}\{Y_n > y\}dy \leq ae^{-a^2} + \int_a^\infty e^{-y^2}dy < \varepsilon.$$

In case (c), using Corollary 2.4 and Example 3 (p. 1045) of Alexander (1984) with $\psi = \psi_3$, we obtain the uniform integrability similarly as above; see also van der Vaart (1996), p. 2134. □

Remark 3.2 *Theorem 3.4 is valid under very mild conditions. In particular, there are no smoothness conditions on the density f. The uniqueness of A_{t_0}, however, is crucial for the results as stated. If it is not satisfied the results can be substantially different. On the other hand, uniqueness of A_{t_0} is a mild condition and holds for many (multimodal) distributions.*

Note that it is well-known, see e.g. Dudley (1982), that for dimension 3 or higher there is no weak convergence of the empirical process indexed by closed convex sets, since this class of sets has a too large entropy. (Actually the supremum of the absolute value of this empirical process tends to infinity, in probability, as $n \to \infty$.) This means that for this case Theorem 3.4, if true at all, can not be proved with the methods presented in this thesis.

Remark 3.3 *Since our general tolerance regions $A_{n,t_0,C}$ converge in probability to A_{t_0}, they are asymptotically minimal with respect to the chosen indexing class. That means, e.g. for case (a), that no tolerance ellipsoids can be found the volume of which converge to a number smaller than $V(A_{t_0})$. However under weak additional conditions (see Section 2.4) there exists a region of the form $\{x \in \mathbb{R}^k : f(x) \geq c\}$, for some $c > 0$, that has probability t_0 and minimal Lebesgue measure. Such a minimal region is unique up to sets of Lebesgue measure 0. If the above level set belongs to the indexing class we use, then our tolerance regions are asymptotically minimal (with respect to all Borel-measurable sets).*

At a finer scale, it seems possible to prove (under additional conditions) along the lines of Einmahl and Mason (1992) that in fact $V(A_{n,t_0,C}) = V(A_{t_0}) + O_{\mathbb{P}}(n^{-1/2})$.

Remark 3.4 *It is rather easy to show that the tolerance regions of Theorem 3.4 have desirable invariance properties. For cases (a) and (c) the tolerance region $A_{n,t_0,C}$ is affine equivariant, i.e. for a nonsingular $k \times k$ matrix M and a vector v in $\mathbb{R}^k$, we have that $MA_{n,t_0,C} + v$ is the tolerance region corresponding to the $MX_i + v$. (Here $MA_{n,t_0,C} = \{Mx : x \in A_{n,t_0,C}\}$.) Since case (b) deals with parallel hyperrectangles,*

this property does not hold in full generality for this case, but it does hold when M is a nonsingular diagonal matrix, which means that we allow affine transformations of the coordinate axes.

Finally, we will extend Theorem 3.4 to more general classes of sets: let $m > 1$ be an integer and let $\mathcal{A} \subset \mathcal{B}$ be the class consisting of
(a') unions of m closed ellipsoids,
(b') unions of m closed parallel hyperrectangles, or
(c') unions of m closed convex sets, contained in a fixed, large compact set (for $k = 2$),
with probability strictly between 0 and 1, respectively. Then an MV-set $A_{n,t_0,C}$ from the class (a'), (b') or (c') consists of at most m 'components' and some of these components may have an empty interior. Since now with positive probability there exists more than one MV-set $A_{n,t_0,C}$ from $\mathcal{A}$ by $A_{n,t_0,C}$ denote any member of the class

$$\mathcal{A}_{n,t_0,C} := \left\{ A \in \mathcal{A} \ : \ P_n(A) \geq t_0 + \frac{C}{\sqrt{n}},\ V(A) = U_n\Big(t_0 + \frac{C}{\sqrt{n}}\Big) \right\}. \tag{3.23}$$

Note that Remark 3.4, mutatis mutandis, holds for the classes (a'), (b') and (c').

Theorem 3.5 *Fix $t_0 \in (0,1)$. If the density f of the distribution function F is positive on some connected, open set $\mathcal{S} \subset \mathbb{R}^k$ and $f \equiv 0$ on $\mathbb{R}^k \backslash \mathcal{S}$, and if A_{t_0}, the set in $\mathcal{A}$ with minimum volume and $P(A_{t_0}) = t_0$, exists and is unique, then we have for the cases (a') and (b') that (3.4),(3.9) and (3.10) hold.*
If $k = 2$ and, in addition, f is bounded, then (3.4),(3.9) and (3.10) also hold for case (c').

Proof First we show that for fix $t_0 \in (0,1)$, $C \in \mathbb{R}$ and $n \geq 1$, the class $\mathcal{A}_{n,t_0,C}$ is not empty when $\mathcal{A}$ is the class of unions of two closed ellipsoids. For the other cases the proof will be similar.

We have to show that there exists the minimum volume set $A_{n,t_0,C} \in \mathcal{A}$ that contains at least $\lceil np_n \rceil$ observations. Without loss of generality we can assume that the class $\mathcal{A}_{\lceil np_n \rceil}$ defined as

$$\mathcal{A}_{\lceil np_n \rceil} := \{ A \in \mathcal{A} \ : \ P_n(A) \geq p_n,\ U_n(p_n) \geq V(A) \geq 2U_n(p_n) \}$$

is uniformly bounded. Observe that $\mathcal{A}_{n,t_0,C} \subset \mathcal{A}_{\lceil np_n \rceil}$. Since any $A \in \mathcal{A}_{\lceil np_n \rceil}$, can be represented as the union of two ellipsoids $A = A_1 \cup A_2$, we will obtain two classes of ellipsoids $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$. Then by the Blaschke Selection Principle there exists a sequence $\{A_n^{(1)}\}_{n\geq 1}$ from $\mathcal{A}_1$ that converges in the metric d to a nonempty ellipsoid $A^{(1)}$ and this corresponding sequence $\{A_n^{(2)}\}_{n\geq 1}$ from $\mathcal{A}_2$ has a subsequence $\{A_{n_k}^{(2)}\}_{k\geq 1}$ that converges in the metric d to a nonempty ellipsoid $A^{(2)}$. Hence any sequence $\{A_n^{(1)} \cup A_n^{(2)}\}_{n\geq 1}$ from $\mathcal{A}_{\lceil np_n \rceil}$ has at least one limit set $A^{(1)} \cup A^{(2)}$. It is easy to show that $A^{(1)} \cup A^{(2)} \in \mathcal{A}_{\lceil np_n \rceil}$. Then $\mathcal{A}_{\lceil np_n \rceil}$ with the metric d is a

compact. Note that the same argument can be used for extending the Blaschke selection Principle for the case of the uniformly bounded sequence of m unions of non-empty compact convex sets. Next consider a real-valued function $f(A) = V(A)$, $A \in \mathcal{A}_{\lceil np_n \rceil}$. Then, since $\mathcal{A}_{\lceil np_n \rceil}$ is a compact and f is a continuous mapping on it, f will reach its upper and lower bounds on $\mathcal{A}_{\lceil np_n \rceil}$. Hence there exists an MV-set from $\mathcal{A}_{\lceil np_n \rceil}$.

Let again $\mathcal{A}$ be the class (a') when $m = 2$. We will show that an MV-set $A_{n,t_0,C} = A_1 \cup A_2 \in \mathcal{A}_{n,t_0,C}$ will contain at most $\lceil np_n \rceil + \frac{k(k+1)}{2}$ observations. Suppose the contrary, that $A_{n,t_0,C}$ contains $\lceil np_n \rceil + \frac{k(k+1)}{2} + \ell$, $\ell > 0$ observations. As we already mentioned above A_1 will have at most $\frac{k(k+3)}{2}$ points on this boundary, among which there is at least one point, say X_1 that does not belong to A_2. Then we can "peel" these boundary points and construct the ellipsoid $A_{1\varepsilon} \subset A_1$ that contains all observations in A_1 except its boundary observations. Then $A_{1\varepsilon} \cup A_2$ will have the volume smaller than $A_{n,t_0,C}$ and will contain at least $\lceil np_n \rceil + \ell$ and at most $\lceil np_n \rceil + \frac{k(k+1)}{2} + \ell - 1$ observations, but this is impossible, hence we obtain that (3.1) holds true. It is easy to show that (3.1) remains true for cases when $m > 2$. Similarly can be shown that (3.2) holds for the classes (b') and (c').

Let us now consider the conditions C1)-C4) of Theorem 3.1. When it is not mentioned otherwise $\mathcal{A}$ will denote below any of the classes (a'), (b') or (c').

C1). The classes (a') and (b') are VC classes and are satisfying required measurability conditions, hence they are P-Donsker. For the class (c') for $m = 2$, when $A_1 \cup A_2 \in \mathcal{A}$ we have that

$$\begin{aligned} &\alpha_n(A_1 \cup A_2) = \\ &\alpha_n(A_1) + \alpha_n(A_2) - \alpha_n(A_1 \cap A_2) \xrightarrow{I\!P} B_P(A_1) + B_P(A_2) - B_P(A_1 \cap A_2) \\ &\qquad = B_P(A_1 \cup A_2), \end{aligned} \tag{3.24}$$

since (c) is a P-Donsker class (see Section 1.2). For the case when $m > 2$ the weak convergence can be obtained using the induction.

C2). We already showed existence of MV-sets.

C3). It follows from (3.1) for the class (a') and from (3.2) for the classes (b') and (c').

C4). We have to show that Lemma 3.2 holds for $\mathcal{A}$. Observe that

$$P_n(A_{n,t_0,C}) \to t_0 \quad \text{a.s.} \quad \text{as } n \to \infty.$$

However C1) implies that

$$\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| \to 0, \quad \text{a.s.} \quad \text{as } n \to \infty$$

and hence

$$P(A_{n,t_0,C}) \to t_0 \quad \text{a.s.} \quad \text{as } n \to \infty.$$

Then following the lines of the proof of Lemma 3.2 we obtain that (3.18) holds for $\mathcal{A}$ and thus

$$\lim_{n \to \infty} V(A_{n,t_0,C}) = V(A_{t_0}) \quad \text{a.s.}.$$

Next we want to show that

$$\lim_{n\to\infty} V(A_{n,t_0,C} \triangle A_{t_0}) = 0 \ \text{ a.s..} \tag{3.25}$$

Let us first prove that the sequence $\{A_{n,t_0,C}\}_{n\geq 1}$ is essentially uniformly bounded, that is there exists a compact set $B = B(\omega)$ such that

$$I\!P\big\{V(A_{n,t_0,C}\backslash B) \to 0\big\} = 1. \tag{3.26}$$

Suppose the contrary, that for any $\omega \in \Omega'$, with $I\!P(\Omega') > 0$,

$$V\big(A_{n,t_0,C}(\omega)\backslash B(\omega)\big) \not\to 0, \tag{3.27}$$

for all closed compact sets $B(\omega) \in \mathcal{B}$. Fix $\omega \in \Omega'$. There exists at least one sequence $\{K_n\}_{n\geq 1}$ of the components of $\{A_{n,t_0,C}\}_{n\geq 1}$ such that

$$V\big(K_n\backslash B\big) \not\to 0, \ \text{ as } \ n \to \infty. \tag{3.28}$$

Then there exists a subsequence of points $a_n \in K_n\backslash B$, such that $d(O, a_n) \to \infty$, where $n \geq 1$ denotes its subsequence for the notational convenience. Then there can be two cases, when $\liminf_{n\to\infty} d(O, K_n) = \infty$ and when it is not. The first case yields that there exists a subsequence $K_{n'}$, with $P(K_{n'}) \to 0$, however we can construct a sequence $D_{n'}$, such that

$$P(K_{n'}) < P(D_{n'}) \ \text{ and } \ V(K_{n'}) > V(D_{n'}),$$

but this is impossible, since $K_{n'}$ is a component of the MV-set $A_{n',t_0,C}$. The second case yields that there exists a subsequence n', such that

$$\text{diam } (K_{n'}) \to \infty, \ \text{ as } \ n' \to \infty.$$

Here again two cases are possible, first suppose that there exists a subsequence n'' such that $P(K_{n''}) \to 0$. Then again consider a sequence $D_{n''}$, with the properties as above $P(K_{n''}) < P(D_{n''})$ and $V(K_{n''}) > V(D_{n''})$, which will lead to contradiction. Therefore $\liminf_{n'\to\infty} P(K_{n'}) > 0$, which yields that there exists a subsequence n'' such that

$$\lim_{n''\to\infty} P(K_{n''}) = \varepsilon,$$

where ε is a positive constant. Note that (3.20) can be extended to the case when $k > 2$ and then it will yield that there exists γ small enough, such that $K_{n''} \not\subset V_\gamma$, $n'' \geq 1$. But since diam $(K_{n''}) \to \infty$ we obtain that $V(K_{n''}) \to \infty$, which gives a contradiction. Hence (3.26) holds true.

Let $\{A^*_{n,t_0,C}\}_{n\geq 1}$ denote the sequence defined similarly as $\{A_{n,t_0,C}\}_{n\geq 1}$ but from the restricted class. By the extended Blaschke Selection Principle the sequence $\{A^*_{n'',t_0,C}\}_{n''\geq 1}$ has a subsequence $\{A^*_{k,t_0,C}\}_{k\geq 1}$ such that

$$\lim_{k\to\infty} V(A^*_{k,t_0,C} \triangle A_{t_0}) = 0 \ \text{ a.s..}$$

But by (3.26) we obtain that

$$\lim_{k\to\infty} V(A^*_{k,t_0,C} \triangle A_{k,t_0,C}) = 0 \quad \text{a.s..}$$

Finally (3.25) follows from these equations.

To complete the proof note that along the lines of the proof of Theorem 3.4 we can show the uniform integrability of the sequence $\sqrt{n}(t_0 - P(A_n, t_0, 0))$, $n \geq 1$ using that the classes (a') and (b') are VC classes and that

$$\sup_{A\in\mathcal{A}_m} |\alpha_n(A)| \leq C(m) \sup_{A\in\mathcal{A}} |\alpha_n(A)|,$$

for the class (c'), where $\mathcal{A}_m$ denotes the class (c'), $\mathcal{A}$ the class (c) and $C(m)$ is a constant depending on m. □

3.3 Applications

All simulations performed in this section consist of 1000 replications.

3.3.1 Comparison of classical and small tolerance intervals

The asymptotic behavior of small tolerance regions does not change if we vary the number of observations in the tolerance regions within $o(\sqrt{n})$. However, even for the classical nonparametric tolerance *intervals* (see Section 1.2), the finite sample behavior is very sensitive to the actual number of used order statistics (see Table 3.1).

number of order statistics	93	94	95	96	97
confidence level	67.9%	79.3%	88.3%	94.2%	97.6%

Table 3.1: Sensitivity of classical 90% guaranteed coverage tolerance intervals with $n = 100$.

Simulations showed a similar sensitivity for small tolerance regions. Moreover, including exactly $\lceil np_n \rceil$ observations we obtained slightly too low coverages, resulting in too low simulated confidence levels. Since the boundary of a tolerance region has probability zero, we decided to add the number of points on the boundary of our tolerance regions to $\lceil np_n \rceil$.

For the classical tolerance intervals, we of course used an exact calculation, based on the beta distribution, for the number of observations to be included. These intervals were chosen in such a way that the indices of the order statistics that serve as endpoints are (almost) symmetric around $(n+1)/2$. We thus expect *small* tolerance intervals to be substantially shorter for skewed distributions, as they automatically scan for the interval with highest mass concentration. As mentioned above, we

added 2 observations when constructing our tolerance intervals. Tables 3.2 and 3.3 contain our simulation results for guaranteed coverage and mean coverage tolerance intervals. These tables show very good behavior of small tolerance intervals. In particular, for the highly skewed distributions they perform much better with respect to length; e.g., for the Pareto distribution the length is reduced with 50%. In general, we see that the asymptotic theory works well.

3.3.2 Tolerance hyperrectangles

Here we performed simulations for tolerance hyperrectangles for $k = 2$ and 3. Table 3.4 gives simulation results for mean coverage rectangles with sides parallel to the coordinate axes. We included 4 extra observations in all cases, i.e. we used 274 observations for $n = 300$ and 904 for $n = 1000$. We simulated from the following distributions:

- bivariate standard normal with mean $\binom{0}{0}$ and covariance matrix $\left(\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right)$,
- bivariate half-normal with density $f(x,y) = \frac{2}{\pi}\, e^{-\frac{1}{2}\left(x^2+y^2\right)}$, $x, y \geq 0$
- bivariate Cauchy distribution with density $f(x,y) = \frac{1}{2\pi}\left(1+x^2+y^2\right)^{-3/2}$,
- bivariate exponential (1,1) distribution with density $f(x,y) = e^{-(x+y)}$, $x, y \geq 0$,
- bivariate pyramid distribution with density $f(x,y) = \frac{1}{8(|x|\vee|y|)}\, e^{-(|x|\vee|y|)}$; see Figure 3.3 below.

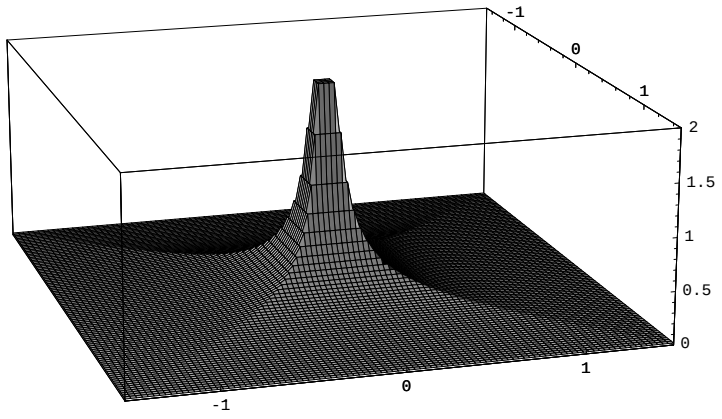

Figure 3.3: Bivariate pyramid density.

distribution	sample size	simulated confidence level		average length	
		classical	new	classical	new
standard normal	300	95.7%	92.2%	3.61	3.57
	1000	95.3%	90.5%	3.46	3.43
standard Cauchy	300	95.8%	94.2%	18.9	18.3
	1000	96.4%	93.2%	15.3	14.9
exponential(1)	300	95.3%	96.8%	3.31	2.73
	1000	94.6%	96.9%	3.11	2.50
Pareto(1)	300	96.2%	97.5%	28.4	14.6
	1000	95.1%	96.0%	22.8	11.2
chi-square(5)	300	95.8%	94.3%	11.0	10.0
	1000	95.2%	92.6%	10.4	9.42
logistic	300	97.9%	93.2%	6.76	6.57
	1000	95.4%	88.9%	6.27	6.18
Student-$t(5)$	300	97.4%	93.7%	4.69	4.54
	1000	95.8%	91.5%	4.33	4.27

Table 3.2: 90% guaranteed coverage tolerance intervals with confidence level 95%.

distribution	sample size	simulated coverage		average length	
		classical	new	classical	new
standard normal	300	90.0%	89.0%	3.31	3.24
	1000	90.0%	89.5%	3.29	3.26
standard Cauchy	300	90.1%	89.5%	13.6	12.7
	1000	90.0%	89.6%	12.9	12.4
exponential(1)	300	90.0%	90.0%	2.98	2.35
	1000	90.0%	90.0%	2.96	2.32
Pareto(1)	300	90.1%	90.1%	20.5	9.71
	1000	90.1%	90.0%	19.5	9.22
chi-square(5)	300	90.0%	89.4%	10.0	8.91
	1000	90.0%	89.7%	9.97	8.92
logistic	300	90.1%	89.1%	5.97	5.80
	1000	90.0%	89.5%	5.90	5.83
Student-t(5)	300	90.0%	89.1%	4.08	3.97
	1000	90.0%	89.6%	4.05	3.99

Table 3.3: 90% mean coverage tolerance intervals.

distribution	sample size	
	300	1000
bivariate normal	87.7%	88.7%
bivariate half-normal	88.3%	88.9%
bivariate Cauchy	86.2%	86.3%
bivariate exponential	88.5%	89.0%
bivariate pyramid	86.4%	87.1%

Table 3.4: Simulated coverages of 90% mean coverage tolerance rectangles.

From this table, we again see that small tolerance regions perform well: the coverages are close to 90%, but slightly too low. This effect is caused by the minimum area property of small tolerance regions, and has a drastic impact on the confidence level of guaranteed coverage tolerance rectangles. Therefore, we do not present simulation results for those rectangles. However, a better performance of the mean coverage tolerance rectangles is possible by including more observations.

For tolerance hyperrectangles in $\mathbb{R}^3$, simulations were performed from the following trivariate distributions:

- trivariate standard normal with mean $\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ and covariance matrix $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$,
- trivariate half-normal with density $f(x,y,z) = \left(\frac{2}{\pi}\right)^{3/2} e^{-\frac{1}{2}\left(x^2+y^2+z^2\right)}$, $x, y, z \geq 0$,
- trivariate Cauchy distribution with density $f(x,y,z) = \frac{1}{\pi^2}\left(1+x^2+y^2+z^2\right)^{-2}$,
- trivariate exponential distribution with density $f(x,y,z) = e^{-(x+y+z)}$, $x, y, z \geq 0$.

In Table 3.5 simulation results for the mean coverage hyperrectangles for $n = 300$ are presented. Here we included 6 extra points. Hence for the 95% mean coverage tolerance regions 291 data points were included. As is clear from this table the results are again very good. Replacing 90% (Table 3.5) by 95% seems to improve the asymptotics, as could be expected. We chose 95% here, not to improve on the coverage, but to speed up the computations; now the number of points that have to be excluded is substantially less (9 against 24).

3.3.3 MV-hyperrectangle algorithm

Here we give a description of the algorithm that was used for computing the minimum volume parallel hyperrectangles for $k = 3$, which led to Table 3.5 (see also Figure 3.6). This algorithm can be easily extended to $k > 3$; the same idea was used for Table 3.4 for $k = 2$.

distribution	simulated coverage
trivariate normal	93.6%
trivariate half-normal	94.1%
trivariate Cauchy	94.8%
trivariate exponential	94.2%

Table 3.5: Simulated coverages of 95% mean coverage tolerance hyperrectangles.

Suppose $\mathfrak{X} = \{(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)\}$ are given n observations in $\mathbb{R}^3$. We want to construct the minimum volume hyperrectangle (*MVH*) that contains at least $\lceil np_n \rceil$-points from $\mathfrak{X}$. The basic idea of the procedure is that since tolerance regions typically have a coverage of 90% or 95%, it is the outermost points that determine the minimum area rectangle. As we have to find the smallest rectangle over $\lceil np_n \rceil$ observations from $\mathfrak{X}$ we 'peel' our data $r+1$ times, where $r := n - \lceil np_n \rceil$. The 'peeling' consists in removing the boundary observations that determine the smallest hyperrectangle. Since a hyperplane parallel to the coordinate hyperplanes contains at most one observation with probability one, we assume that each face of the *MVH* contains one boundary observation from $\mathfrak{X}$. For the notational convenience let H_{n-r} be the *MVH* over $n-r$ points. Hence we want to construct H_{n-r}.

Procedure: I. To simplify our procedure we first peel the data $r+1$ times and drop the points from $\mathfrak{X}$ which are $r+2$ level and deeper. Let H_n be the *MVH* over $\mathfrak{X}$. By V_1 denote the set of observations from $\mathfrak{X}$ that define H_n (the boundary observations of H_n), obviously $\#\{V_1\} \leq 6$. Further consider a new set of observations $\mathfrak{X} \setminus V_1$ and again construct *MVH* over $\mathfrak{X} \setminus V_1$. The set of this boundary observations denote by V_2. Repeat this procedure $r+1$ times. Set $\mathfrak{X}^* := \cup_{i=1}^{r+1} V_i$ and denote by ℓ the cardinality of $\mathfrak{X}^*$ ($\ell := \#\{\mathfrak{X}^*\}$).
Note that since we have to find at most six points (or six faces) that define H_{n-r}, in the procedure there will be five free indexes.

II. *a*) Order (by increasing) z-coordinates of the elements of $\mathfrak{X}^*$ and denote them by $Z_{j:\ell}$, $j = 1, \ldots, \ell$. The horizontal sides of the H_{n-r} can lie only on the planes $z = Z_{j:\ell}$ and $z = Z_{\ell-r+i-1:\ell}$, where $j = 1, \ldots, r+1$ and $i = j, \ldots, r+1$. For each fixed i and j there are $j - 1 + \big(\ell - (\ell - r + i - 1)\big) = r - (i - j)$ points outside of the planes $z = Z_{j:\ell}$ and $z = Z_{\ell-r+i-1:\ell}$ (recall that $i \geq j$). Hence more $i-j$-points should be removed.

b) Fix i and j. Consider the elements of $\mathfrak{X}^*$ that lie between the planes: $Z_{j:\ell}$ and $Z_{\ell-r+i-1:\ell}$. In other words, take the points of $\mathfrak{X}^*$ with the third coordinates satisfying the inequality: $Z_{j:\ell} \leq z \leq Z_{\ell-r+i-1:\ell}$. Denote this set of points by $\mathfrak{X}^*_{i,j}$. Since $\mathfrak{X}^*$ contains ℓ-points and $r - i + j$-points have been dropped, we have that $\#\{\mathfrak{X}^*_{i,j}\} = \ell - r - j + i$. Further order the y-coordinates of the elements of $\mathfrak{X}^*_{i,j}$ and denote them by $Y_{k:m}$, $k = 1, \ldots, m$, where $m := \ell - r - j + i$. The y-faces of H_{n-r} can only be on the planes $y = Y_{k:m}$ and $y = Y_{m-(i-j)+p-1:m}$, where $k = 1, \ldots, i-j+1$ and $p = k, \ldots, i-j+1$. Hence for fixed i, j, k and p we removed

additional $k-1+m-(m-(i-j)+p-1)=(i-j)-(k-p)$-points. Thus the total amount of the points that have been removed is $r-(i-j)+(i-j)-(p-k)=r-(p-k)$. Hence $p-k$ points should be dropped at the next step.

c) Fix i,j,k and p, take the observations from $\mathfrak{X}^*_{i,j}$ that lie between the planes $y=Y_{k:m}$ and $y=Y_{m-(i-j)+p-1:m}$ and denote the set of this observations by $\mathfrak{X}^*_{i,j,k,p}$. Order x-coordinates of the elements of $\mathfrak{X}^*_{i,j,k,p}$ and denote them by $X_{q:h}$, $q=1,\ldots,h$, with $h:=\ell-(r-(p-k))$. Then the x-faces of H_{n-r} can only lie on the planes $x=X_{q:h}$ and $x=X_{h-(p-k)+q-1:h}$, where $q=1,\ldots,p-k+1$.

Finally for each fixed i,j,k,p and q the volume of the *MVH* ($H(i,j,k,p,q)$), with the faces on the planes $z=Z_{j:\ell}$, $z=Z_{\ell-r+i-1:\ell}$, $y=Y_{k:m}$, $y=Y_{m-(i-j)+p-1:m}$, $x=X_{q:h}$ and $x=X_{h-(p-k)+q-1:h}$ will be

$$V\{H(i,j,k,p,q)\}=(Z_{\ell-r+i-1:\ell}-Z_{j:\ell})\times(Y_{m-(i-j)+p-1:m}-Y_{k:m})\times(X_{h-(p-k)+q-1:h}-X_{q:h})$$

and the final *MVH* is the hyperrectangle with the smallest volume among $H(i,j,k,p,q)$'s:

$$H_{n-r}=\operatorname{argmin}\big\{V\{H(i,j,k,p,q)\}\ :\ j=1,\ldots,r+1,\quad i=j,\ldots,r+1,\quad k=1,\ldots,i-j+1,\quad p=k,\ldots,i-j+1,\quad q=1,\ldots,p-k+1\big\}.$$

3.3.4 'Smoothed' tolerance intervals

Given the discrete nature of the empirical measure and the aforementioned sensitivity of tolerance regions it can be, in particular when the density f is smooth, that a smoothed version of the empirical measure yields somewhat better tolerance regions than the ones presented in Section 3.1. We will briefly consider this here and will restrict ourselves to the one dimensional situation and guaranteed coverage tolerance intervals. It can be shown, see e.g. Azzalini (1981), Shorack and Wellner (1986), Section 23.2, and van der Vaart (1994), that an integrated kernel density estimator ($\hat{P}_n$, say) as an estimator for the probability measure yields the same limiting behavior as in Section 3.2, when the bandwidth is chosen to be $K/n^{1/3}$, $K\in(0,\infty)$. So asymptotically, in first order, there is no difference between the two procedures, i.e. Theorem 3.2 holds true, when $A_{n,t_0,C}$ is based on $\hat{P}_n$ instead of on P_n. However, for finite n it may be that a 'smoothed procedure' works better. We investigated this through a simulation. Table 3.6 gives the results. We chose the Epanechnikov kernel (with support $[-1,1]$) and $K=\frac{1}{2}\sqrt{5}S$, with S the sample standard deviation, as suggested in Azzalini (1981). Since $\hat{P}_n$ is absolutely continuous we did not add the 2 observations as indicated above.

This table shows excellent behavior of the 'smoothed' tolerance intervals. We see indeed that there is some evidence that, when the underlying density is smooth, our procedures can be somewhat improved by properly smoothing the empirical.

distribution	sample size	simulated conf. level	average length
standard normal	300	92.6%	3.58
	1000	92.7%	3.44
chi-square(5)	300	96.4%	9.98
	1000	96.8%	9.50
beta(5,10)	300	94.5%	0.42
	1000	94.3%	0.40
logistic	300	93.4%	6.51
	1000	93.1%	6.23
Student-$t(5)$	300	93.6%	4.52
	1000	92.7%	4.28

Table 3.6: 'Smoothed' 90% guaranteed coverage tolerance intervals with confidence level 95%.

All simulations presented in this section were performed on a SunSparc5 and SunUltra10. Simulations in dimensions one and three were performed using the statistical packages of the computer algebra system Mathematica. The (two-dimensional) rectangles algorithm was implemented in $C++$, which was linked with a Mathematica notebook where data were generated and coverages were computed. The computation for one replication (including the coverage computation) with $n = 1000$ took at most 6 seconds.

3.3.5 Medical data example

As mentioned before medical statistics is one of the fields where tolerance regions are used. Here we illustrate our theory with an application to Leukemia diagnosis. Leukemia is a cancer of blood-forming tissue such as bone marrow. The diagnosis of Leukemia is based on the results of both blood and bone marrow tests. There are only three major types of blood cells: red blood cells, white blood cells and platelets. These cells are produced in the bone marrow and circulate through the blood stream in a liquid called plasma. When the bone marrow is functioning normally the count of blood cells remains stable. In the case of this disease the number of blood cells changes drastically and is therefore easy to detect with tolerance regions. We now construct a 95% mean coverage tolerance ellipse and two 95% mean coverage tolerance (hyper)rectangles (for dimension $k = 2$ and $k = 3$) for blood count data kindly provided by Blood bank de Meierij, Eindhoven. Blood samples were taken from 1000 adult, supposedly healthy potential blood donors. Among the measured variables were the total number of white blood cells (WBC), red blood cells (RBC), and platelets (PLT) in one nanoliter, picoliter, and nanoliter, respectively, of whole blood. We computed tolerance regions (ellipse, rectangle, hyperrectangle) for the following combinations of variables: (WBC, PLT), (WBC, RBC) and (WBC, RBC, PLT), for 500, 1000 and 500 observations, respectively (see Figures 3.4,

3.5 and 3.6 below).

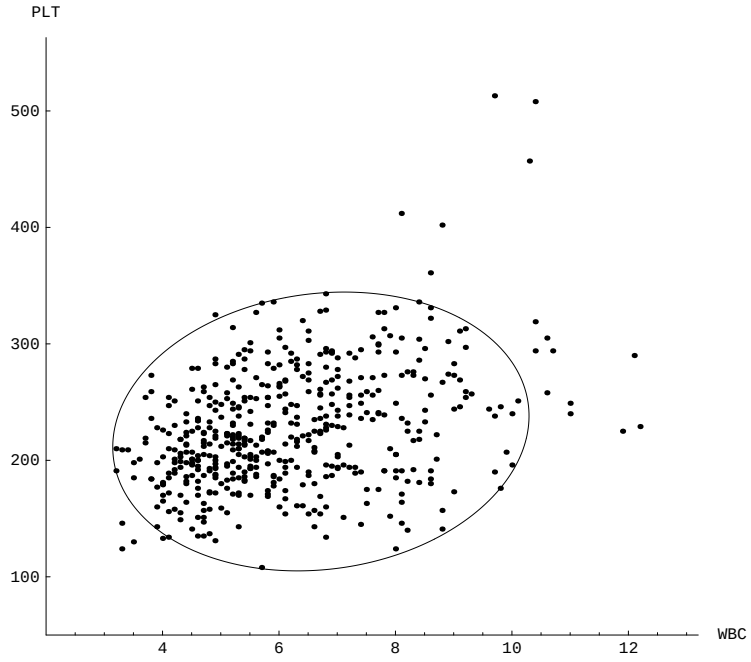

Figure 3.4: 95% mean coverage tolerance ellipse.

Comparing the tolerance regions in Figures 3.4, 3.5 and 3.6 with the in practice used one-dimensional 'reference' or 'normal' values for WBC, RBC, and PLT (which we do not record here), it can be seen that our procedures work nicely. Due to the fact that the one-dimensional distributions of WBC and PLT are somewhat skewed to the right our procedures tend to give smaller regions (when these variables are involved), than those constructed (in one way or another) from the one-dimensional reference values. This is the same effect as seen in Tables 2 and 3 for the skewed distributions there. Moreover, our tolerance regions are somewhat shifted to the 'left' because of this skewness of the distributions of these variables. It is obvious, but it can be important, that in Figure 3.4, the tolerance ellipse does not include certain bivariate values, which would be included when forming two intervals by projecting the ellipse on the horizontal and vertical axes. For Acute Leukemia, newly diagnosed, adult patients very often have WBC values considerably over 10 (in many cases even above 100(!)) or RBC values around 3 or PLT values below 100. Clearly these values can be easily detected by the depicted tolerance regions.

Finally we give some references on computing minimum volume ellipsoids and minimum area planar convex sets (which we did not compute in this section). An

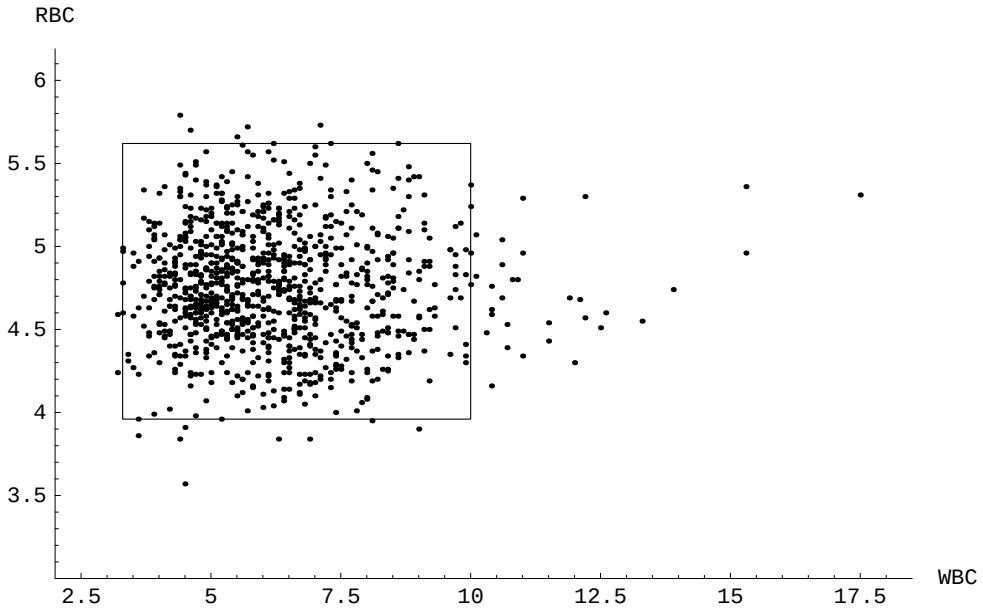

Figure 3.5: 95% mean coverage tolerance rectangle.

algorithm for computing the minimum volume ellipsoid containing *all* data points is presented in Silverman and Titterington (1980). Algorithms for computing *approximate* minimum area ellipsoids containing m ($< n$) points are given in Nolan (1991) and Rousseeuw and van Zomeren (1991) and the *exact* algorithm we used for the minimum volume ellipse containing m ($< n$) points was developed in Agulló (1996). The computer code of this algorithm was kindly placed to our disposal by the author; it also works in higher dimensions (up to 10). As we noted in Section 3.1, the minimum area planar convex set containing m ($< n$) sample points is a polygon. *Exact* algorithms for computing such sets can be found in Eppstein et al. (1992) and Eppstein (1992).

3.4 Appendix

Recall the notation of Section 3.1, in particular let $X_1, \ldots, X_n$ and $\mathcal{E}$ be as in that section. Denote with $E_1 \in \mathcal{E}$ the almost surely unique ellipsoid of minimum volume containing *at least* $m \in \{k+1, \ldots, n\}$ (data) points.

Lemma 3.3 *E_1 contains exactly m points, almost surely.*

Proof Assume that E_1 contains $\ell > m$ points and t ($k+1 \leq t \leq k(k+3)/2$ a.s.) of these points are on its boundary. Note that the smallest ellipsoid containing these

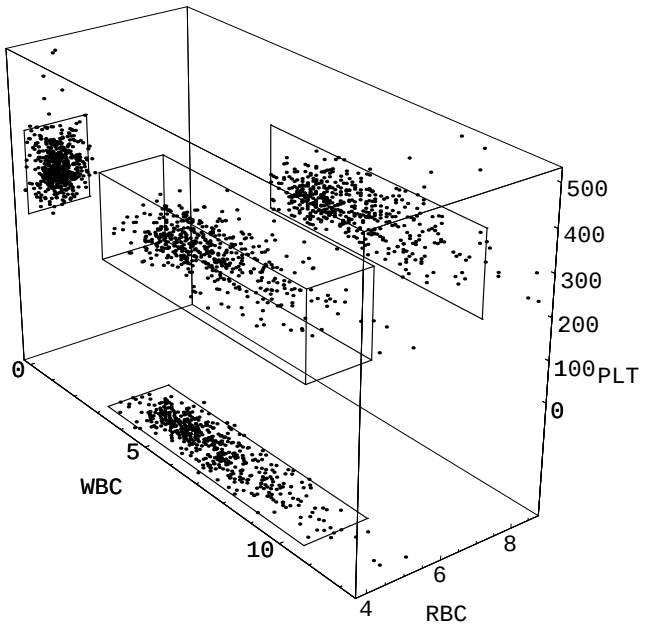

Figure 3.6: 95% mean coverage tolerance hyperrectangle.

t boundary points is equal to E_1, see Silverman and Titterington (1980). Consider $t-1$ of the t boundary points (call this set B) and let E_0 be the smallest ellipsoid containing B. Denote the remaining t-th boundary point of E_1 with Y_1. Observe that $Y_1 \notin E_0$. It follows from a conditioning argument that for any subset of size $r > 1$ of the n points, we have a.s. that none of the remaining $n-r$ points is on the boundary of the smallest ellipsoid containing these r points. This yields that a.s. $V(E_0) < V(E_1)$.

Note that the smallest ellipsoid containing a finite set is equal to the smallest ellipsoid containing the convex hull of that set. Denote with Y_0 a point on the boundary of E_0 such that the line through Y_0 and Y_1 intersects the convex hull of B and such that the open interval from Y_0 to Y_1 has an empty intersection with E_0. Set $Y_\lambda = (1-\lambda)Y_0 + \lambda Y_1$, $\lambda \in [0,1]$. Let C_λ be the convex hull of $B \cup \{Y_\lambda\}$. Note that for $\lambda < \lambda'$ we have that $C_\lambda \subset C_{\lambda'}$. Let E_λ be the smallest ellipsoid containing C_λ. So $V(E_\lambda) \leq V(E_{\lambda'})$ for $\lambda \leq \lambda'$.

From the Blaschke Selection Principle it follows that there exists a sequence λ_j (< 1), $j \in \mathbb{N}$, converging to 1 and such that

$$\lim_{j\to\infty} V(E_{\lambda_j} \triangle E^*) = 0$$

for some $E^* \in \mathcal{E}$. We have $V(E^*) \leq V(E_1)$, since $V(E_{\lambda_j}) \leq V(E_1)$, $j \in \mathbb{N}$. But $C_1 \subset E^*$, so $V(E_1) \leq V(E^*)$. Hence $V(E^*) = V(E_1)$ and E^* and E_1 both contain C_1. But, with probability 1, E_1 is unique, so $E^* = E_1$ and hence

$$\lim_{j\to\infty} V(E_{\lambda_j} \triangle E_1) = 0.$$

So there exists a large j (denote the corresponding λ_j with η) such that E_η contains all the $\ell - t$ points in the interior of E_1 and the points of B and does not contain the $n-\ell$ points in the complement of E_1. If $Y_1 \in E_\eta$, then Y_η is in the interior of E_η, so according to Silverman and Titterington (1980), $E_\eta = E_0$ and hence $V(E_\eta) = V(E_0) < V(E_1)$ a.s., but this can not happen since $C_1 \subset E_\eta$. This yields that $Y_1 \notin E_\eta$. We now see that E_η contains $\ell - 1$ $(\geq m)$ points and $V(E_\eta) \leq V(E_1)$. Since E_1 is the minimum volume ellipsoid containing at least m points, we have that $V(E_\eta) = V(E_1)$. Since $E_\eta \neq E_1$ this contradicts the a.s. uniqueness of the minimum volume ellipsoid. □

Chapter 4

Small nonparametric tolerance regions for directional data

This chapter is an extended version of Mushkudiani (2000).

Continuing the study of tolerance regions here we construct directional tolerance regions based on the method presented in Chapter 3. Tolerance regions for circular and spherical data are defined as the MV-sets from the classes of arcs and caps, respectively. Asymptotic results on these tolerance regions are presented. The tolerance regions investigated in this chapter are asymptotically minimal under some very mild conditions. The method is applied to real data of wind directions.

4.1 Introduction

Data that represent directions in space of any number of dimensions are called directional data. In practice the cases that are considered and studied are of directional data in two and three dimensional spaces. Then these data are called circular and spherical data respectively.

Circular and spherical data points occur in many applications in biology, geology, meteorology, geography, medicine and physics. The corresponding statistical theory is studied intensely in e.g., Mardia (1972), Batschelet (1981), Fisher, Lewis and Embleton (1987), Fisher (1993), etc. These monographs contain vast data examples obtained from different areas. Typical directional data sources are e.g., bird or animal orientation and navigation data that one normally encounters in biology, while exploring homing, migration or other activities. In meteorology wind and ocean directions, thunderstorm and rainfall data are the prominent examples when directional statistics are natural to apply. More examples of directional data are

orientations of cross-beddings or fractures and fabric elements in deformed rocks, micro seismic and earthquake directions in a certain region, etc. in geology and geography.

Circular data can be represented as the angles θ, $0 \leq \theta < 2\pi$ measured from the X-axis in the anti-clockwise direction or as the points on the unit circle that correspond to θ. Similarly, spherical observations are identified with directions in space and hence can be represented by two angles (θ, φ). These angles can be defined in various ways. Directions in space can be also identified with the points on the unit sphere.

In this chapter we construct tolerance regions for circular and spherical data based on minimum volume sets and the techniques presented in Chapter 3. The directional data below are represented by the points on the unit circle and on the unit sphere in $\mathbb{R}^2$ and $\mathbb{R}^3$, respectively. To construct the MV-sets we consider the following indexing classes: the class of arcs defined on the unit circle (see Section 2.5) and the class of caps defined on the unit sphere (see Section 4.2). Then the tolerance regions are certain MV-sets from these classes. We establish the limiting behavior for these regions and show that they are asymptotically minimal with respect to the indexing class.

4.2 The setup

In this and the next section we will deal only with spherical data. However, the results obtained below also hold for circular data, with slight modifications, taking into account that the analogue of the class of caps $\mathfrak{C}$ defined below, is the class of arcs on the circle.

As we have mentioned above spherical data can be specified in different ways. The one we will need here is as follows. Take $L = (x, y, z) \in \mathbb{R}^3$ and set O to be the origin. Suppose $L \neq O$ and let L' be the point in which the vector OL cuts the surface of the unit ball $B_{(O,1)}$ with center in O. The direction of OL can be identified with the point L'.

Let $X_1, \ldots, X_n$, $n \geq 1$, be i.i.d. random vectors with values on the unit sphere $\mathcal{S}^2$ (the surface of $B_{(O,1)}$) defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, from a common distribution P (see e.g., Mardia (1972), Fisher et al. (1987)). Denote the σ-algebra of Borel sets on $\mathbb{R}^3$ with $\mathcal{B}$ and let d_0 be the pseudo-metric defined in Section 1.2. Note that since whole mass of the probability measure P is concentrated on $\mathcal{S}^2$, $P(B) = P(B \cap \mathcal{S}^2)$ for any $B \in \mathcal{B}$. Let P_n denote the empirical distribution of the sample $X_1, \ldots, X_n$ indexed by $\mathcal{B}$.

Set $\mathfrak{C} \subset \mathcal{B}$ to be the class of caps C, defined as follows

$$C = \{(x, y, z) \ : \ x^2 + y^2 + z^2 = 1 \text{ and } ax + by + cz + d \geq 0\},$$

where $a, b, c, d \in \mathbb{R}$ (see also Ruymgaart (1989)). In other words a set C from $\mathfrak{C}$ is the intersection of the half-space $ax+by+cz+d \geq 0$ with $\mathcal{S}^2$. The circle with center B, created by the intersection will be called the boundary circle (see Figure 4.2). The perpendicular line to the boundary circle at B goes through the cap and 'cuts'

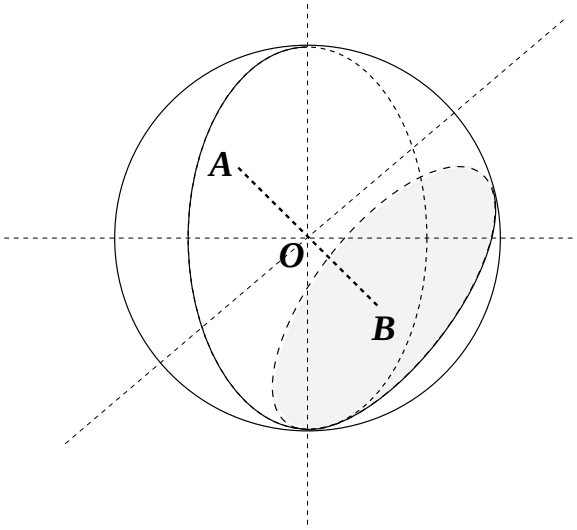

Figure 4.1: The cap with the center A and the boundary circle centered at B.

the sphere $\mathcal{S}^2$ at the point A. Point A will be called the center of the cap and $|AB|$ its height, with

$$|AB| = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

for $A = (x_1, y_1, z_1)$, $B = (x_2, y_2, z_2) \in \mathbb{R}^3$. To avoid technical inconveniences, from now on let $\mathfrak{C}$ be the class of caps with $0 < P(C) < 1$. One of the engaging properties of the elements of $\mathfrak{C}$ is that they can be very easily parametrized. Any set $C \in \mathfrak{C}$ is uniquely determined by its center η_C and height ℓ_C. Hence to each $C \in \mathfrak{C}$ corresponds a point $(\eta_C, \ell_C) \in \mathcal{S}^2 \times (0, 2)$. Take a sequence $\{C_n\}_{n\geq 1}$ from $\mathfrak{C}$, denote the sequence of the corresponding parameters by $\{(\eta_n, \ell_n)\}_{n\geq 1}$. Since the sequence $\{(\eta_n, \ell_n)\}_{n\geq 1}$ is bounded there exists a subsequence $\{(\eta_{n_k}, \ell_{n_k})\}_{k\geq 1}$ that converges coordinate-wise to some point $(\eta^*, \ell^*) \in \mathcal{S}^2 \times [0, 2]$. Since for $\eta_{n_k} = (x_{n_k}, y_{n_k}, z_{n_k})$ we can write that

$$1 = \lim_{k\to\infty} [x_{n_k}^2 + y_{n_k}^2 + z_{n_k}^2] = x^{*2} + y^{*2} + z^{*2},$$

where $(x^*, y^*, z^*) = \eta^*$, it is clear that there exists a cap corresponding to (η^*, ℓ^*) and $C^* \in \mathfrak{C}$ unless when ℓ^* is 0 or 2. It is easy to see that the following equation holds as well

$$\lim_{k\to\infty} V(C_{n_k} \triangle C^*) = 0 \quad \text{a.s.}, \tag{4.1}$$

where V denotes the area (Lebesgue measure) on $\mathcal{S}^2$. Similar results for ellipsoids can be found in Nolan (1991).

Suppose now that P is absolutely continuous with respect to Lebesgue measure on $\mathcal{S}^2$. Define the MV-sets based on the indexing class $\mathfrak{C}$ as follows. For any fixed $t \in (0,1)$ and $q \in I\!R$ denote by $C_{n,t,q}$ a MV-set from $\mathfrak{C}$ with empirical measure at least $t_n = t + \frac{q}{\sqrt{n}}$, thus

$$C_{n,t,q} = \text{argmin}_{C \in \mathfrak{C}} \big\{ V(C) \ : \ P_n(C) \geq t_n \big\}.$$

Set $C_{n,t} = C_{n,t,0}$. The sets $C_{n,t,q}$ and $C_{n,t}$ are the candidate guaranteed content and mean content tolerance caps respectively

(see also Section 3.1).

Lemma 4.1 *Suppose $X_1, \ldots, X_n$, $n \geq 1$, are i.i.d. random vectors with values in $\mathcal{S}^2$ from the common distribution P, that is absolutely continuous with respect to Lebesgue measure on $\mathcal{S}^2$. Then the following hold:*

(a) An MV-set $C_{n,t,q}$ from $\mathfrak{C}$ exists and is a.s. unique.

(b) The MV-set $C_{n,t,q}$ will contain exactly $\lceil nt_n \rceil$ observations from $X_1, \ldots, X_n$, with probability one.

Proof *(a)* We first prove the existence and a.s. uniqueness of the MV-cap $C_{\mathcal{X}}$ (MV-set from $\mathfrak{C}$) that contains $\mathfrak{X} = \{X_1, \ldots, X_n\}$.

Trivially $P_n(C_{\mathcal{X}}) = 1$. Let $\mathfrak{C}_{\mathfrak{X}} \subset \mathfrak{C}$ be the class of all caps that contain $\mathfrak{X}$. We will prove the existence of $C_{\mathcal{X}}$ by using the parametrization argument described above. From the definition of $\mathfrak{C}$ it is clear that

$$\mathcal{L}_{\mathfrak{X}} := \{\ell_C \ : \ C \in \mathfrak{C}_{\mathfrak{X}}\} \subset (0,2).$$

Set $\ell^* := \inf \mathcal{L}_{\mathfrak{X}}$. Further take a sequence $\{\ell_n\}_{n \geq 1}$ from $\mathcal{L}_{\mathfrak{X}}$ with $\ell_n \downarrow \ell^*$. Denote by $\{\eta_n\}_{n \geq 1}$ the sequence of η's corresponding to $\{\ell_n\}_{n \geq 1}$. By the same argument as above there exists a subsequence $\{\eta_{n_k}, \ell_{n_k}\}_{k \geq 1}$ that converges to some point $(\eta^*, \ell^*) \in \mathcal{S}^2 \times (0,2)$, with $\ell^* = \inf \mathcal{L}_{\mathfrak{X}}$. Then there exists $C^* \in \mathfrak{C}$ that corresponds to (η^*, ℓ^*) and a sequence $\{C_{n_k}\}_{k \geq 1}$ of caps such that (4.1) holds. To complete the existence proof we have to show that $C^* \in \mathfrak{C}_{\mathfrak{X}}$ or that $\mathfrak{X} \in C^*$. Suppose there exists X_i with $X_i \notin C^*$ then there exists k_0 such that for any $k > k_0$, C_{n_k} will not contain X_i, which is impossible. Hence $C^* = C_{\mathcal{X}}$ is a MV-cap.

Now we prove a.s. uniqueness of $C_{\mathcal{X}}$. Note that with probability one, there can be at most three observations on any circle on $\mathcal{S}^2$ and any two circles that pass through different sets of three observations will have different radii. By $\overline{\mathfrak{C}_{\mathfrak{X}}}$ denote the class of sets that are obtained by taking convex hulls (in $I\!R^3$) of the elements of $\mathfrak{C}_{\mathfrak{X}}$. It is easy to show that an MV-set from $\overline{\mathfrak{C}_{\mathfrak{X}}}$ corresponds to $C_{\mathcal{X}}$. Construct the polyhedron $H_{\mathfrak{X}}$ with n vertices from $\mathfrak{X}$. Clearly each face of $H_{\mathfrak{X}}$ is a triangle a.s.. It can be shown by induction that for $n \geq 4$, $H_{\mathfrak{X}}$ will have $2n-4$ faces. Since $H_{\mathfrak{X}}$ is the smallest convex set containing $\mathcal{X}$, each element of $\overline{\mathfrak{C}_{\mathfrak{X}}}$ will contain $H_{\mathcal{X}}$. There can be two kinds of polyhedra $H_{\mathfrak{X}}$, those that contain the origin and those that do not contain it. We will treat these cases separately.

I. $O \in H_{\mathfrak{X}}$. Since each convex hull from the class $\overline{\mathfrak{C}_{\mathfrak{X}}}$, they will also contain the origin. Then an MV-set from this class will also contain the origin and will have the boundary circle with the biggest radius (in comparison to the elements of $\overline{\mathfrak{C}_{\mathfrak{X}}}$). Hence, its boundary circle will lie in the plane of one of the faces of the polyhedron and will pass through three points from $\mathcal{X}$. Assume there exist two different MV-sets C_1 and C_2 from $\overline{\mathfrak{C}_{\mathfrak{X}}}$. Hence, their areas are equal. Then the radii of their boundary circles are equal as well. Since both boundary circles pass through three observations it is impossible with probability one.

II. $O \notin H_{\mathfrak{X}}$. The polyhedron $H_{\mathcal{X}}$ is the intersection of the half-spaces created by the planes of its faces. Hence, there exists at least one of these half-spaces, that does not contain the origin. Therefore, an MV-set from $\overline{\mathfrak{C}_{\mathfrak{X}}}$ will not contain the origin either and hence will have the boundary circle with the smallest radius (in comparison to the elements of $\overline{\mathfrak{C}_{\mathfrak{X}}}$ that do not contain the origin) and the smallest height (in comparison to all elements of $\overline{\mathfrak{C}_{\mathfrak{X}}}$). It is easy to show that the boundary circle of an MV-cap $C_{\mathcal{X}}$ will pass through three or two points from $\mathcal{X}$. However, it will pass through two points only in case it is the smallest circle in $\mathcal{S}^2$ passing through these two points and if so there will be third observation on this circle with probability zero. Suppose that C_1 and C_2 are two MV-caps from $\mathfrak{C}_{\mathfrak{X}}$, hence the radii of their boundary circles are equal. The case when both boundary circles pass through three observations can be treated similarly as in I. Assume that the boundary circle of C_1 passes through three points $\{X_{i_1}, X_{i_2}, X_{i_3}\}$, while the boundary circle of C_2 through two points $\{X_{j_1}, X_{j_2}\}$. Without loss of generality we can assume that $X_1 \in \{X_{i_1}, X_{i_2}, X_{i_3}\} \backslash \{X_{j_1}, X_{j_2}\}$. If we condition on $\{X_2, \ldots, X_n\}$, then it is left to show that for any $r \in (0, 1)$

$$I\!P\{X_1 \,:\, R(C_1) = r \;:\; X_2, \ldots, X_n\} = 0,$$

where $R(C)$ stands for the radius of the boundary circle of the cap C. This is trivial since $R(C_1) = r$, implies that X_1 can lie only on at most two prescribed circles. The case when C_1 and C_2 have two points on their boundary circles can be treated analogically.

Using the same arguments as above now we can prove the existence and a.s. uniqueness of the MV-cap $C_{n,t,q}$. Clearly an MV-cap $C_{n,t,q}$ should contain at least $\lceil nt_n \rceil$ observations from $\mathcal{X}$. Since there are finitely many $\lceil nt_n \rceil$-element subsets of $\mathcal{X}$ and we can construct the MV-cap for each subset, the existence of $C_{n,t,q}$ is trivial. Now we prove uniqueness. Suppose there exist two MV-caps $C_{n,t,q}$ and $C^*_{n,t,q}$, then the boundary circles of these caps will pass through two or three observations from $\mathcal{X}$. However, we already discussed these cases above. Hence

$$I\!P\{C_{n,t,q} = C^*_{n,t,q}\} = 1.$$

(b) Suppose in contrary that the MV-cap $C_{n,t,q}$ contains m observations

$$\mathcal{X}_m := \{X_{i_1}, \ldots, X_{i_m}\} \subset \mathcal{X},$$

where $m > \lceil nt_n \rceil$. Again consider two cases: when $O \in H_{\mathfrak{X}}$ and when $O \notin H_{\mathfrak{X}}$.

I. Since $O \in H_{\mathfrak{X}}$, the boundary circle of the cap $C_{n,t,q}$ will pass through three observations from $\mathcal{X}_m$, say $\{X_{i_1}, X_{i_2}, X_{i_3}\}$. Without loss of generality we can assume that $\overline{X_{i_1}X_{i_2}}$ is the smallest side of the triangle with vertices X_{i_1}, X_{i_2} and X_{i_3}. Let $C_{X_{i_1},X_{i_2}}$ be the smallest cap containing X_{i_1} and X_{i_2}. Obviously $C_{X_{i_1},X_{i_2}}$ will not contain X_{i_3} (see Figure 4.2, I). Since we want to show that there exists a cap that contains $m-1$ points from $\mathcal{X}$ and has a smaller area than $C_{n,t,q}$, it will be sufficient to construct a cap that contains only $\{X_{i_1}, \ldots, X_{i_m}\}\backslash\{X_{i_3}\}$ and show that this boundary circle has radius greater than the one of $C_{n,t,q}$. To drop the point X_{i_3} one can rotate the plane of the boundary circle of $C_{n,t,q}$ around the axis $\{X_{i_1}, X_{i_2}\}$ with some small angle ε. Call the cap obtained by the rotation $C^{\varepsilon}_{n,t,q}$. Since $X_{i_3} \notin C_{X_{i_1},X_{i_2}}$ one will have to rotate the boundary circle of the cap $C_{n,t,q}$ away from the boundary circle of $C_{X_{i_1},X_{i_2}}$. Therefore there exists an $\varepsilon > 0$ small enough such that $C^{\varepsilon}_{n,t,q}$ will contain $m-1$ observations and the radius of its boundary circle will be greater than the radius of the boundary circle of $C_{n,t,q}$, which is impossible since $C_{n,t,q}$ is the MV-cap containing at least $\lceil nt_n \rceil$ observations from $\mathcal{X}$.

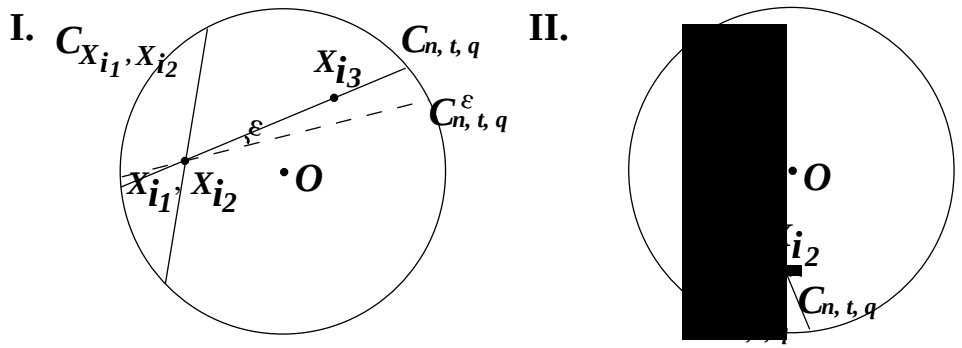

Figure 4.2: The cross section of $B_{(O,1)}$ cut on the plane: passing through the origin O and perpendicular to $\overline{X_{i_1}X_{i_2}}$ *(I)*; passing through the origin O, parallel to $\overline{X_{i_1}, X_{i_2}}$ and perpendicular to the boundary circle of $C_{n,t.q}$ *(II)*.

II. When $O \notin H_{\mathfrak{X}}$, the boundary circle of $C_{n,t,q}$ will pass through either two or three observations from $\mathcal{X}$. In case when it passes through three points we can obtain a contradiction similarly as above. Suppose that the boundary circle of $C_{n,t,q}$ passes through two observations $\{X_{i_1}, X_{i_2}\}$. As in I we want to construct a cap smaller than $C_{n,t,q}$ that contains only $m-1$ points. Without loss of generality we can assume that these $m-1$ points are $\{X_{i_1}, \ldots, X_{i_m}\}\backslash\{X_{i_2}\}$. To obtain such a cap rotate the plane of the boundary circle of $C_{n,t,p}$ around the point X_{i_1}, with some angle $\varepsilon > 0$ in the direction of away from X_{i_2} (see Figure 4.2, II). Clearly there exists a small enough $\varepsilon > 0$ such that the cap $C^{\varepsilon}_{n,t,q}$ obtained by the rotation will have an area smaller then $C_{n,t,q}$ and will contain at least $\lceil nt_n \rceil$ observations from

$\mathcal{X}$, which gives a contradiction.

Hence we have proved that the MV-cap $C_{n,t,q}$ will contain exactly $\lceil nt_n \rceil$ observations, which trivially implies

$$t_0 + \frac{q}{\sqrt{n}} \leq P_n(C_{n,t,q}) < t_0 + \frac{q}{\sqrt{n}} + \frac{1}{n} \text{ a.s.}. \tag{4.2}$$

□

Recall that similar results were obtained for the classes of all closed ellipsoids, hyperrectangles and convex sets in (3.2) and Lemma 3.3.

4.3 Main results

In this section we use the setting and the notation introduced in the previous section. Suppose that P has a density f which is absolutely continuous with respect to Lebesgue measure on $\mathcal{S}^2$ and that f is strictly positive on some connected open set $A \subset \mathcal{S}^2$ ($f \equiv 0$ on $\mathcal{S}^2 \setminus A$).

Theorem 4.1 *Fix $t_0 \in (0,1)$. If the minimum volume set C_{t_0} from $\mathfrak{C}$ with $P(C_{t_0}) = t_0$ exists and is unique, then for every $q \in \mathbb{R}$*

$$\sqrt{n}(t_0 - P(C_{n,t_0,q})) + q \xrightarrow{d} Z\sqrt{t_0(1-t_0)} \quad (n \to \infty), \tag{4.3}$$

where Z is a standard normal random variable.

Observe that Theorem 4.1 can be proved similarly as Theorem 3.1. Hence we will have to show that in the setting of this chapter the conditions C1)-C4) of Theorem 3.1 are satisfied for the class $\mathfrak{C}$. We will need the following lemma (see also Lemma 3.2 in Section 3.2).

Lemma 4.2 *Under the assumptions of Theorem 4.1 we have with probability one that*

$$d(C_{n,t_0,q}, C_{t_0}) \to 0,$$

and hence $d_0(C_{n,t_0,q}, C_{t_0}) \to 0 \quad (n \to \infty)$.

For proving Lemma 4.2 one does not need to make any crucial changes in the proof of Lemma 3.2 in Section 3.2. However, the parametrization from Section 4.2 could be used instead of the Blaschke Selection Principle.

Proof of Theorem 4.1 For each $n \geq 1$, define the empirical process indexed by $\mathfrak{C}$ to be

$$\alpha_n(C) = \sqrt{n}(P_n(C) - P(C)), \quad C \in \mathfrak{C}.$$

The class $\mathfrak{C}$ is a VC class, since it is obtained by the intersection of half-spaces and the unit sphere $\mathcal{S}^2$ and it satisfies the required measurability conditions. Hence $\mathfrak{C}$ is a P-Donsker class (see Section 1.2) and hence by the Skorokhod construction there

exists a probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{I\!P})$ carrying a version $\widetilde{B}_P$ of B_P and versions $\widetilde{\alpha}_n$ of α_n, for all $n \in I\!N$, such that

$$\sup_{C \in \mathfrak{C}} |\widetilde{\alpha}_n(C) - \widetilde{B}_P(C)| \to 0 \quad \text{a.s.} \quad n \to \infty. \tag{4.4}$$

For convenience, we will drop the tildes from the notation:

$$\sup_{C \in \mathfrak{C}} |\sqrt{n}(P_n(C) - P(C)) - B_P(C)| \to 0 \quad \text{a.s.} \quad n \to \infty. \tag{4.5}$$

Then by the existence and a.s. uniqueness of the MV-cap $C_{n,t_0,q}$ in Lemma 4.1 we have that

$$\sqrt{n}(P_n(C_{n,t_0,q}) - P(C_{n,t_0,q})) - B_P(C_{n,t_0,q}) \to 0 \quad \text{a.s.} \quad n \to \infty. \tag{4.6}$$

Using (4.2) and (4.6) we obtain

$$\sqrt{n}(t_0 - P(C_{n,t_0,q})) + q - B_P(C_{n,t_0,q}) \to 0 \quad \text{a.s.} \quad n \to \infty. \tag{4.7}$$

Using Lemma 4.2 and the uniform continuity of B_P on $\mathfrak{C}$ for the pseudo metric d_0, we will get that

$$B_P(C_{n,t_0,q}) \to B_P(C_{t_0}) \quad \text{a.s.} \quad n \to \infty. \tag{4.8}$$

Further it trivially follows from (4.7) and (4.8) that

$$\sqrt{n}(t_0 - P(C_{n,t_0,q})) + q - B_P(C_{t_0}) \to 0 \quad \text{a.s.} \quad n \to \infty.$$

And at last using that

$$B_P(C_{t_0}) \stackrel{d}{=} Z\sqrt{t_0(1-t_0)},$$

we obtain our result

$$\sqrt{n}(t_0 - P(C_{n,t_0,q})) + q \stackrel{d}{\to} Z\sqrt{t_0(1-t_0)} \quad n \to \infty.$$

□

The following limit theorems, similarly to Theorems 3.2 and 3.3, immediately follow from Theorem 4.1 and are the main results of this section. Set q_α to be the $(1-\alpha)$-th quantile of the distribution of $Z\sqrt{t_0(1-t_0)}$. Then by the following theorems MV-caps C_{n,t_0,q_α} and C_{n,t_0} are asymptotic t_0-guaranteed coverage tolerance regions with confidence level $1-\alpha$ and t_0-mean coverage tolerance regions respectively. Theorem 4.2 below deals with the asymptotic behavior of C_{n,t_0,q_α}, whereas in Theorem 4.3 the limit result for C_{n,t_0} can be found.

Theorem 4.2 *Fix $\alpha \in (0,1)$, then under the conditions of Theorem 4.1 we have*

$$\lim_{n\to\infty} I\!P\{P(C_{n,t_0,q_\alpha}) \geq t_0\} = 1 - \alpha.$$

Theorem 4.3 *Under the conditions of Theorem 4.1*

$$\mathbb{E}P(C_{n,t_0}) = t_0 + o(n^{-1/2}), \quad n \to \infty.$$

Note that for *every* $q \in \mathbb{R}$

$$\mathbb{E}P(C_{n,t_0,q}) \to t_0, \quad n \to \infty.$$

The proof of Theorem 4.2 is mainly the same as that for Theorem 3.2. Theorem 4.3 can be obtained from Theorem 3.3 and from the fact that the sequence of random variables $\sqrt{n}(t_0 - P(C_{n,t_0}))$ is uniformly integrable. Following the lines of the proof of Theorem 3.4 uniform integrability of $\sqrt{n}(t_0 - P(C_{n,t_0}))$ can be shown using that $\mathfrak{C}$ is a VC class.

Remark Notice that the assumptions under which the results are proved are very mild, in particular, there are no smoothness conditions on the density f.

As we have already mentioned above, t_0-content and t_0-expectation tolerance regions for n *circular* data can be defined as the MV-sets from the class of arcs with empirical measure $t_0 + \frac{q_\alpha}{\sqrt{n}}$ and t_0, respectively.

Theorem 4.4 *Theorems 4.2 and 4.3 remain true, mutatis mutandis, for circular data and the class of arcs.*

4.4 Simulation study and real data example

Here we present simulation results for tolerance arcs based on circular data. The number of replications for the performed simulations is 1000. The distributions from which we sampled data satisfy our conditions: the support of the density f is connected and there exists a unique shortest arc (α, β) with coverage $\int_\alpha^\beta f(\varphi)d\varphi = t_0$. Note that the density h defined below (see also Figure 4.3) is bimodal, however the conditions are still satisfied since t_0 is close to 1 and this is the case of interest in practice. The tolerance region for n circular data is the shortest arc that contains at least $\lceil nt_n \rceil$ observations, where $t_n = t_0 + \frac{q_\alpha}{\sqrt{n}}$. As we already have seen in Section 3.3, the finite sample behavior of our tolerance regions is very sensitive to the number of observations included. For example 90% guaranteed coverage tolerance arcs with $n = 300$ simulated from the von Mises $(\pi, 3)$ distribution had confidence levels: 80.4%, 85.1%, 88.7%, 92.9% and 95.2% when we included 278, 279, 280, 281 and 282 points respectively, while $\lceil nt_n \rceil = 279$ (see also Table 3.1). Hence as in Section 3.1, we have increased the number of points in the tolerance regions with the number of points on this boundary. Thus the tolerance arcs we constructed contain $\lceil nt_n \rceil + 2$ observations.

We simulated from the following circular distributions (see e.g. Batschelet (1981)):

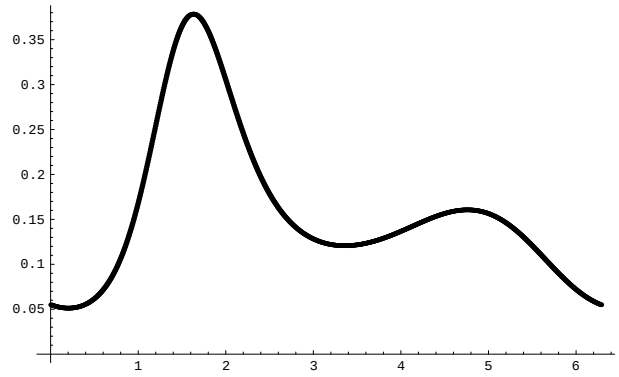

Figure 4.3: Linear plot of the bimodal circular distribution h.

distribution	sample size	simulated conf. level	simulated coverage
von Mises$(\pi, 3)$	300	92.9%	89.1%
	1000	92.1%	89.6%
von Mises$(\pi, 8)$	300	94.3%	89.1%
	1000	92.2%	89.5%
$g(\varphi)$	300	92.1%	89.0%
	1000	89.9%	89.5%
$h(\varphi)$	300	90.8%	89.0%
	1000	90.2%	89.4%

Table 4.1: Simulated confidence level for 90% guaranteed coverage tolerance arcs with confidence level 95% and simulated coverage for 90% mean coverage tolerance arcs.

- von Mises distribution with parameters $(\pi, 3)$ and $(\pi, 8)$ respectively;
- $g(\varphi) = \frac{1}{2\pi} + \frac{k}{2\pi}\sin(\varphi + \nu \sin\varphi)$ with parameters $k = 1$ and $\nu = \pi/3$, where $\varphi \in [0, 2\pi]$;
- $h(\varphi) = c\exp[k\cos(3.4 + \varphi + \mu\cos(3.4 + \varphi))]$ with $c = 0.139236$, $k = 1$ and $\mu = \frac{5}{12}\pi$, where $\varphi \in [0, 2\pi]$.

In Table 4.1 the simulation results for the guaranteed coverage and mean coverage tolerance arcs are presented. For the guaranteed coverage tolerance arcs we computed the empirical confidence level: the percentage of tolerance arcs with a coverage greater than or equal to 90%. If we take into account that the coverage of the tolerance regions is extremely sensitive to the number of points included, then the simulation results are indeed very satisfactory.

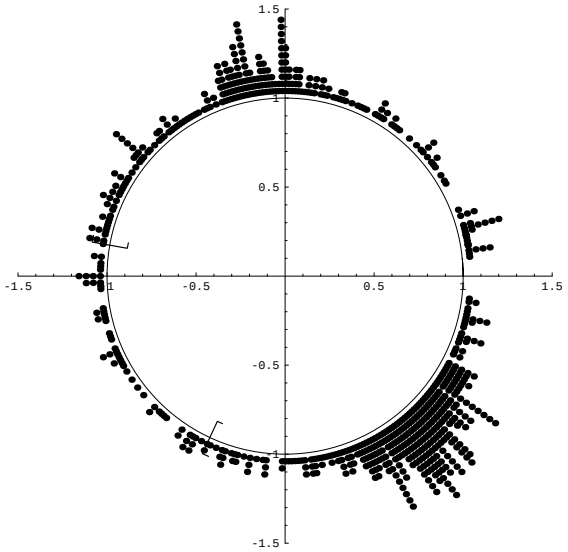

Figure 4.4: Tolerance arc for wind directions at Pt. Conception, CA.

Following the 'smoothing' procedure presented in Section 3.3 (using the Epanechnikov kernel), we simulated 'smoothed' mean coverage and 'smoothed' guaranteed coverage tolerance arcs. Simulation results are presented in Table 4.2. Similarly as in Table 3.6, here we see again that by using the density estimation procedure the results have been improved considerably. For example, for $n = 1000$ a simulated confidence level of the 90%-guaranteed coverage tolerance arc with confidence level 95% from the distribution g, based on the empirical measure is 89.9%, whereas the simulated confidence level for the 'smoothed' tolerance arc with the same parameters is 95.4%.

Next we construct a guaranteed coverage tolerance arc for wind direction data ($n = 694$) obtained from the U.S. National Weather Service at weather station Pt. Conception, CA, USA; these observations are measured in degrees (see Figure 4.4). Clearly the underlying density has a connected support, is bimodal and not symmetrical in any direction. Hence we can assume the uniqueness of MV-arc and apply our procedure to this data set. For the tolerance arc a coverage of 90% and a confidence level of 95% were chosen. The number of observations to be included in the arc is equal to $\lceil nt_n \rceil + 2 = 640$. Then the guaranteed coverage tolerance arc is $[X_{325:694} = 245°,\ X_{270:694} = 170°]$.

Tolerance regions for wind directions can be applied for example in architectural aerodynamics, the study of relationships between wind and buildings. To survey

distribution	sample size	simulated conf. level	simulated coverage
von Mises$(\pi, 3)$	300	93.7%	89.5%
	1000	93.6%	89.9%
von Mises$(\pi, 8)$	300	94.7%	89.4%
	1000	93.3%	89.8%
$g(\varphi)$	300	95.6%	89.6%
	1000	95.4%	89.8%
$h(\varphi)$	300	95.2%	89.6%
	1000	95.9%	89.9%

Table 4.2: Simulated confidence level for 'smoothed' 90% guaranteed coverage tolerance arcs with confidence level 95% and simulated coverage for 'smoothed' 90% mean coverage tolerance arcs.

this relation two factors, direction and speed of wind, can be observed. Knowledge of the wind speed distribution and the most frequent wind directions is very crucial for choosing wind turbines and locating them. Tolerance arcs for wind directions can be used for example for choosing directions of wind turbines.

Discussion

To argue about the novelty and advantages of the method presented in Chapters 3 and 4, let us first consider the one-dimensional case. As we have mentioned already the classical nonparametric tolerance intervals defined in Section 2.3 have the same asymptotic behavior as the small nonparametric tolerance intervals, introduced in Chapter 3. However the indices of the order statistics that define the classical tolerance intervals are chosen beforehand and in the case of skew, asymmetric distributions like, e.g., Pareto or exponential distributions the length of the classical nonparametric tolerance intervals may be much greater than of the small nonparametric tolerance intervals (see Tables 3.2 and 3.3). In addition since the tolerance intervals obtained by the new method are the shortest intervals that contain a certain number or order statistics, it is obvious that they will never be longer than the classical tolerance intervals. Furthermore it is difficult and unnatural to extend the classical procedure to higher dimensions, since an ordering has to be introduced on $\mathbb{R}^k$, $k > 1$. In contrast to this, the new method can be extended naturally to higher dimensions by using minimum volume sets.

The common procedures for multivariate nonparametric tolerance regions in $\mathbb{R}^k$, $k \geq 1$, are based on statistically equivalent blocks or on density estimation (see, Section 2.4). The method based on statistically equivalent blocks, depends on auxiliary ordering functions and is essentially a one-dimensional procedure. The tolerance regions obtained by this method are exact, however they are not always asymptotically minimal and have a shape that is difficult to work with. The other approach, that is based on density estimation is more attractive and yields asymptotically minimal tolerance regions. Though it is (very) conservative since for the tolerance regions it is only required that

$$\liminf_{n\to\infty} \alpha_n \geq \alpha,$$

(see Definition 2.3). Note as well that since the method is based on a density estimator some regularity conditions have to be satisfied.

In contrast to these methods, the new method is based on an indexing class and therefore the shape of the tolerance regions can be chosen conveniently. Furthermore when this indexing class includes the class of level sets the small tolerance regions

are asymptotically minimal, hence best possible. Generally, these tolerance regions are asymptotically correct, have better convergence rate than the tolerance regions based on the density estimation from Section 2.4 and are affine equivariant. Note that in principal method based on density estimation can be very smooth, this was shown by the simulation results in Sections 3.3 and 4.4, however it is important to choose the proper bandwidth. Note as well that the results obtained in Chapters 3 and 4 hold under very mild conditions.

Part II

P-P plots

Chapter 5

Brief review

5.1 Introduction

Graphical methods in nonparametric statistics have a long history and are nowadays commonly used for analyzing data. Recent developments in computer science and its interaction with statistics and in particular with the theory of nonparametric statistics made the practical applications of the graphical methods even more possible. As a result of this interaction, highly developed statistical packages are used in almost all scientific fields that deal with large quantities of raw, empirical data and graphical methods are used to visualize the performed analysis. The graphical methods are generally applied while investigating location, scale, skewness, kurtosis or other differences in two-sample problems; symmetry or goodness of fit-problems for one sample; analysis of covariance, k-sample or other multivariate procedures. An extensive review and bibliography of graphical methods in nonparametric statistics can be found in, e.g., Doksum (1977), Gnanadesikan (1977), Fisher (1983), Sawitzki (1994).

Most of the graphical methods are based on diagnostic plots. The simplest examples of diagnostic plots: histograms, empirical distribution functions or boxplots, can be found in every elementary statistical textbook. The plots are often used for detecting validity of the model or analyzing data in an already defined model. Therefore fitting diagnostic or other plots is a necessary step during the data analysis. It is usually required that strong theory supports these fitting procedures and often features as power or reliability of the diagnostic plots are considered.

Citing Fisher (1983) "in nonparametric statistics probably the most powerful and useful graphical methods are those based on comparison of the sample distribution functions". The most prominent example of those methods are based on probability-probability (P-P) and related plots as quantile-quantile (Q-Q) plots, pair charts, receiver operating characteristic (ROC) curves, proportional hazards plots, etc.. In this and the following chapter we will study P-P and related plots.

5.2 Probability-probability plots

5.2.1 Definitions

The classical P-P plot $\{(F(x), G(x)) : x \in I\!R\}$ is based on the univariate distribution functions F and G and is generally used to compare them (see Figure 5.1). Clearly, the plot begins at $(0, 0)$ and ends at $(1, 1)$, and if the distribution functions are identical, the plot is a straight line, otherwise we will obtain a curve with various shapes depending on the difference of F and G. It is visible from the Figure 5.1 that the plot is much more sensitive to differences between these two functions in the center of their mass, then in their tails. The P-P plot can also be identified with the graph of the function

$$\{F \circ G^{-1}(y) : y \in (0, 1)\}, \tag{5.1}$$

where G^{-1} is the inverse function of G for a strictly increasing, continuous G, otherwise it is presumed to be the generalized inverse function defined as

$$G^{-1}(y) = \inf\{x : G(x) \geq y\}, \quad y \in (0, 1). \tag{5.2}$$

The receiver operating characteristic (ROC) curve defined as

$$\{1 - F \circ G^{-1}(1 - y) : y \in (0, 1)\} \tag{5.3}$$

is closely related to the P-P plot and hence results for this curve are immediately obtained using the corresponding results for P-P plots. Generally, the ROC curve is used in signal theory, psychology, radiology, medicine, etc. (see, e.g., Swets and Pickett (1982), Li et al. (1996), Hsieh and Turnbull (1996)).

For testing the null hypothesis $H_0 : F \equiv G$ in the one-sample case, when a sequence of i.i.d. random variables $X_1, X_2, \ldots$ has a common unknown distribution function F, the classical P-P plot can be estimated by the empirical version

$$\{F_n \circ G^{-1}(y) : y \in [0, 1]\}, \tag{5.4}$$

where F_n is the empirical distribution function based on $X_1, \ldots, X_n$. Figure 5.2 shows the empirical P-P plots for $n = 100$, corresponding to the theoretical plots of Figure 5.1.

For the two-sample problem, where two sequences $X_1, X_2, \ldots$ and $Y_1, Y_2, \ldots$ of univariate random variables, from unknown distribution functions F and G are given, we want to compare these samples. For $n, m \geq 1$, define the empirical P-P plot as

$$\{F_n \circ G_m^{-1}(y) : y \in (0, 1)\}, \tag{5.5}$$

where $G_m^{-1}(y) = \inf\{x : G_m(x) \geq y\}$, for $y \in (0, 1)$, F_n is the empirical distribution function based on a sample $X_1, \ldots, X_n$, and G_m is the empirical distribution functions based on a pooled sample $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ or on the second sample $Y_1, \ldots, Y_m$.

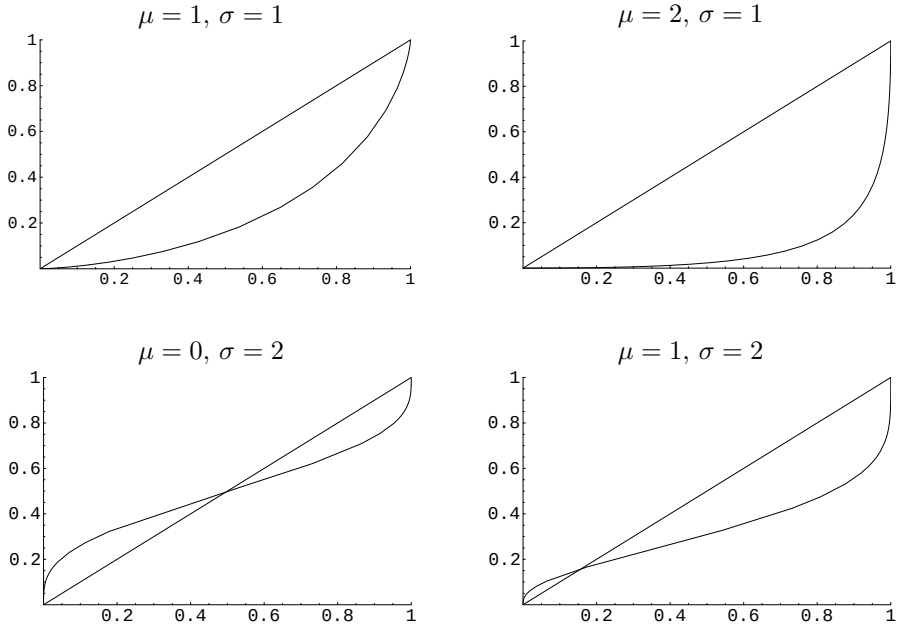

Figure 5.1: P-P plots of Normal(μ, σ) against Normal(0,1) distribution.

From the plots related to the P-P plots the quantile-quantile (Q-Q) plots have been studied most frequently. The Q-Q plots were first used by Lorenz (1905) for comparing two independent samples. In Gnanadesikan (1977), Fisher (1983), Aly (1986b), etc. the properties of these plots are studied profoundly. For the continuous, univariate distribution functions F and G for each $y \in (0,1)$, $F^{-1}(y) = G^{-1}(y)$, if these distribution functions are identical. Then for arbitrary univariate distribution functions F and G, the Q-Q plot is defined as $\{(F^{-1}(y), G^{-1}(y)) : y \in (0,1)\}$, where these inverse functions are defined as in (5.2). Similarly to P-P plots, Q-Q plots can also be represented as $\{F^{-1} \circ G(x) : x \in I\!R\}$. Again when F is not assumed to be continuous, F^{-1} will denote this generalized inverse function. The empirical Q-Q plot is defined as

$$\{F_n^{-1} \circ G(x) \ : \ x \in I\!R\} \tag{5.6}$$

and

$$\{F_n^{-1} \circ G_m(x) \ : \ x \in I\!R\}, \tag{5.7}$$

respectively. When F and G are identical, the Q-Q plot will be a straight line with slope one. Note that the linearity will not change when the functions F and G only differ in intercept or scale, however the plot will have a different location and slope. Note also that a Q-Q plot is more sensitive to differences between F and G in the tails of the distributions than in the centers.

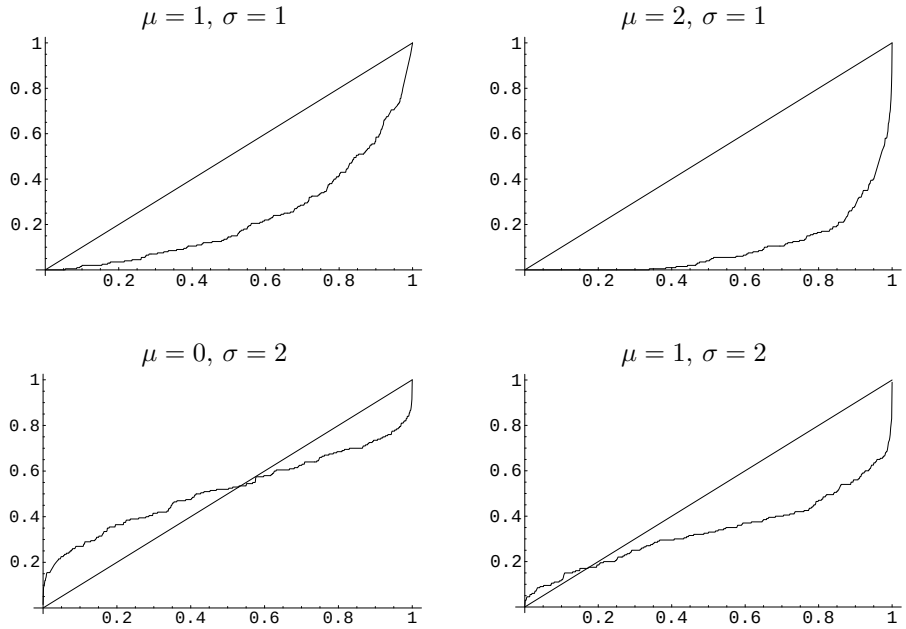

Figure 5.2: Empirical P-P plots of Normal(μ, σ) against Normal(0,1) distribution.

Nonparametric statistics based on P-P plots and modifications have been studied widely (see, e.g., Gnanadesikan (1977), Parzen (1993), Beirlant and Deheuvels (1990), Sawitzki (1994), Deheuvels and Einmahl (1992), Polonik (1999), Nair (1981, 1982), Aly (1986a), Girling (2000), etc.).

5.2.2 Approximations of P-P and Q-Q plot processes

We maintain the notations introduced above. Consider the classical two-sample comparison problem in nonparametric statistics. Suppose $X_{11}, \ldots, X_{1n_1}$, $n_1 \geq 1$, and $X_{21}, \ldots, X_{2n_2}$, $n_2 \geq 1$, are two independent samples from the distribution functions F_1 and F_2, respectively, and let F_{jn_j} be the empirical distribution function based on the sample $X_{j1}, \ldots, X_{jn_j}$, for $j = 1, 2$. When F_j and F_j^{-1}, for $j = 1, 2$, are continuous, by the Glivenko-Cantelli theorem we obtain that for any $\varepsilon > 0$ and n_2 large enough

$$F_2(x) - \varepsilon \leq F_{2n_2}(x) < F_2(x) + \varepsilon \quad \text{for all} \quad x \in I\!R \quad \text{a.s..}$$

Since $F_{2n_2}^{-1}$ is the generalized inverse of F_{2n_2} we obtain that

$$F_2^{-1}(y - \varepsilon) \leq F_{2n_2}^{-1}(y) < F_2^{-1}(y + \varepsilon) \quad \text{for all} \quad y \in (0, 1) \quad \text{a.s.},$$

then for each $y \in (0,1)$ there exists $\delta > 0$, such that

$$F_2^{-1}(y) - \delta \le F_2^{-1}(y-\varepsilon) \le F_{2n_2}^{-1}(y) < F_2^{-1}(y+\varepsilon) \le F_2^{-1}(y) + \delta \quad \text{a.s..}$$

Hence for each fixed $y \in (0,1)$,

$$\begin{aligned} &|F_{1n_1}(F_{2n_2}^{-1}(y)) - F_1(F_2^{-1}(y))| \le |F_{1n_1}(F_{2n_2}^{-1}(y)) - F_1(F_{2n_2}^{-1}(y))| \\ &+ |F_1(F_2^{-1}(y)) - F_1(F_2^{-1}(y))| \to 0, \quad \text{a.s.} \quad n_1, n_2 \to \infty. \end{aligned} \tag{5.8}$$

Similarly for each $x \in (a,b)$, with

$$a := \inf\{x : F_2(x) > 0\} \quad \text{and} \quad b := \sup\{x : F_2(x) < 1\},$$

we obtain that

$$|F_{1n_1}^{-1}(F_{2n_2}(x)) - F_1^{-1}(F_2(x))| \to 0, \quad \text{a.s.} \quad n_1, n_2 \to \infty. \tag{5.9}$$

Hence, empirical P-P and Q-Q plots converge pointwise to their theoretical counterparts with probability one. The next step is to investigate the rates of convergence of the weighted difference of the empirical and theoretical plots.

Assuming that F_1 has density f, define P-P and Q-Q plot processes based on P-P and Q-Q plots as follows

$$\Delta_{n_1 n_2}(y) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\big(F_{1n_1}(F_{2n_2}^{-1}(y)) - F_1(F_2^{-1}(y))\big), \quad y \in (0,1). \tag{5.10}$$

and

$$\Gamma_{n_1 n_2}(x) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} f\big(F_1^{-1}(F_2(x))\big)\big(F_{1n_1}^{-1}(F_{2n_2}(x)) - F_1^{-1}(F_2(x))\big), \quad x \in (a_m, b_m). \tag{5.11}$$

It is proved in Beirlant and Deheuvels (1990) that under certain conditions, under the null hypothesis $H_0 : F_1 \equiv F_2$, the P-P plot process $\Delta_{n_1 n_2}$ is distribution-free, and so is the Q-Q plot process $\Gamma_{n_1 n_2}$ after changing its scale $x = F_2^{-1}(y)$. They showed as well that these processes converge in distribution to Brownian bridges (with the rate $n^{-1/4}(\log n)^{1/2}(\log\log n)^{1/4}$, where $n_j = n_j(n) \to \infty$ as $n \to \infty$, $j = 1, 2$, which is optimal in the setting of Beirlant and Deheuvels (1990)).

5.2.3 Applications of P-P plots

Observe that testing procedures based on the P-P plot are rather straightforward. In general, for a given random sample $X_1, \ldots, X_n$, $n \ge 1$, having a common unknown distribution function F, one has to define the empirical distribution function F_n to test the null hypothesis $H_0 : F = F_0$, using P-P plots. Here, F_0 is a given distribution function. In this case, the limiting distribution of the test statistic defined in terms of the empirical P-P plot, can be derived when the asymptotic behavior of F_n is known. Consider, for example, the goodness-of-fit problem for randomly

censored data in the i.i.d. case; suppose that $X_1, X_2, \ldots$ and $U_1, U_2, \ldots$ are two independent sequences of nonnegative i.i.d. random variables with continuous distribution functions F and H, where X_i, $i \geq 1$, denote uncensored random variables and U_i, $i \geq 1$, these random censoring times. Suppose also that the functions F and H are unknown, and that the null hypothesis we want to test is $H_0 : F = F_0$, where F_0 is a given distribution function. Then F_n can be defined as the Kaplan-Meier estimator of the true lifetime distribution function F. Nair (1981) introduced the test statistics based on the weighted difference of the F_0 and F_n functions and derived the limiting distributions of these statistics under the null hypothesis, using the fact that $\sqrt{n}(F_n(x) - F(x))$, $x \in (0, T)$ with $0 < F(T)H(T) < 1$, converges in distribution to a Gaussian process. He also constructed the simultaneous confidence bands for P-P, Q-Q and hazard plots (see Nair (1981)). Similarly, for the two-sample problem in the case of censorship, the Kaplan-Meier product-limit estimators of lifetime distributions can be used to define the P-P plots and to establish strong approximations of the empirical P-P plot process based on these plots (see Deheuvels and Einmahl (1992)).

In the other application considered below, the P-P plot is used as an aid in calculating the test statistics for the Wald and Wolfowitz runs test, the Wilcoxon/Mann-Whitney test and modifications. Suppose that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are two independent samples, then the Mann-Whitney-Wilcoxon statistic M_{WW} can be represented as the area of the region bounded by the empirical P-P plot (5.5) and the lines $x = 1$ and $y = 0$. Hence

$$M_{WW} := \frac{1}{nm} \sum_i \sum_j \delta_{ij} = \int_{R_{nm}} dxdy, \tag{5.12}$$

where

$$\delta_{ij} = \begin{cases} 0 & \text{if } X_i < Y_j, \\ \frac{1}{2} & \text{if } X_i = Y_j, \\ 1 & \text{if } X_i > Y_j. \end{cases}$$

and R_{nm} in the region under the empirical P-P plot. Other versions of the Mann-Whitney-Wilcoxon rank statistics, also the trimmed and censored ones can be represented in a similar way, using the region under the empirical P-P plot or ROC curve (see, e.g., Quade (1973), Girling (2000)).

5.2.4 Multivariate P-P and Q-Q plots

The plots considered until now were defined in the one-dimensional case. Since these plots are based on the quantile transformation, there is no straightforward way of generalizing these fitting procedures into higher dimensions. However, one can define multivariate P-P and Q-Q plots based on the generalized quantile functions defined in Section 1.4. In Polonik (1999) multivariate P-P and Q-Q plots are defined and the procedure for a goodness-of-fit test based on these plots is proposed for the multivariate case. However, the author calls these plots C-C plots, since they contain

information on the concentrations of the comparing distributions. The C-C plots are defined in terms of generalized MV-sets. Recall that MV-sets defined in Section 1.3 can be represented using empirical generalized quantile functions, when the real-valued function $\lambda \equiv V$ is the Lebesgue measure. In Polonik (1999) generalized MV-sets and generalized empirical MV-sets are defined similarly to the MV-sets, for the case when λ is an arbitrary real-valued function.

Suppose $X_1, X_2, \ldots$ is a sequence of random vectors taking values in $\mathbb{R}^k$, $k \geq 1$, from a distribution P. In the notation of Chapter 1 for a class $\mathcal{A} \subset \mathcal{B}$ and a fixed $t \in [0,1]$, sets $A_{P,t}$, $A_{P_n,t} \in \mathcal{A}$ are called the generalized MV-set and generalized empirical MV-set, respectively, iff

$$A_{P,t} \in \operatorname*{argmin}_{A \in \mathcal{A}} \{\lambda(A) \ : \ P(A) \geq t\}$$

and

$$A_{P_n,t} \in \operatorname*{argmin}_{A \in \mathcal{A}} \{\lambda(A) \ : \ P_n(A) \geq t\},$$

where P_n is the empirical distribution based on sample $X_1, \ldots, X_n$. Then under the assumption that the generalized MV-sets are determined uniquely up to λ-nullsets, for testing the null hypothesis $H_0 : P \equiv P_0$, the C-C plot is defined as the combination of the multivariate P-P and Q-Q plots, respectively, $\{P(A_{P_0,t}) : t \in [0,1]\}$ and $\{P(A_{P_0,t}) : t \in [0,1]\}$. The empirical versions of these plots are defined as $\{P_n(A_{P_0,t}) : t \in [0,1]\}$ and $\{P_0(A_{P_n,t}) : t \in [0,1]\}$. A diagnostic plot consisting of these two graphs is called an empirical C-C plot. Polonik (1999) showed that both graphs in the theoretical C-C plot are straight lines through the origin with slope one iff $P \equiv P_0$. In addition, he showed that certain test statistics based on the P-P and Q-Q plot processes are asymptotically distribution-free.

Chapter 6

Generalized P-P plots

6.1 Introduction and the testing procedure

As mentioned in the previous chapter the P-P plot is a commonly used graphical method in hypothesis testing. In this chapter we introduce the generalized P-P plot and the generalized P-P plot process for testing goodness-of fit and two-sample problems for fixed and contiguous alternatives.

Suppose $X_1, \dots X_n$, $n \geq 1$, is a given random sample with values in $\mathbb{R}$, having a unknown common distribution P, that is absolutely continuous with respect to Lebesgue measure. To test the null hypothesis $H_0 : P \equiv P_0$, with P_0 a given distribution, absolutely continuous with respect to Lebesgue measure, define the generalized P-P plot process

$$M_n(t) = \sup_{\substack{P_0(A)=t \\ A\in\mathcal{A}}} \sqrt{n}\big(P_n(A) - t\big), \quad t \in [0,1],$$

where $\mathcal{A}$ is the class of the closed intervals (defined in the next section), $V(A)$, $A \in \mathcal{A}$ is the Lebesgue measure of the set A and P_n is the empirical distribution function based on the sample $X_1, \dots, X_n$. Based on the P-P plot process M_n, construct the test statistic T_n defined as

$$T_n = \sup_{t\in[0,1]} M_n(t).$$

For the testing procedure use the critical region $[C_\alpha, \infty)$, where C_α is the solution of the following equation

$$\mathbb{P}\{\sup_{\substack{V(A)=t \\ A\in\mathcal{A}}} B(A) > C_\alpha\} = \alpha,$$

and B is a Brownian bridge indexed be the class $\mathcal{A}$. The simulated values of C_α for $\alpha = 0.05$ are presented in Table 6.1 below. In the following section we show that the proposed test

- is distribution-free under the null hypothesis;
- is consistent against all fixed alternatives;
- is easy to visualize due to the P-P plots.

We also consider contiguous alternatives and derive the limiting distribution on the generalized P-P plot process for this case. The corresponding two-sample problem is also considered and treated similarly.

6.2 One-sample problem

6.2.1 Fixed alternatives

Let $X_1, \ldots, X_n$, $n \geq 1$, be a sequence of i.i.d. random variables defined on a probability space $(\Omega, \mathcal{F}, I\!P)$ taking values in $I\!R$, from an unknown common distribution P, that is absolutely continuous with respect to Lebesgue measure and has the corresponding distribution function F. Let $\mathcal{B}$ denote the Borel σ-algebra on $I\!R$, d_0 the pseudo-metric on $\mathcal{B}$ and P_n the empirical distribution of the sample $X_1, \ldots, X_n$, $n \geq 1$, as defined in Section 1.2.

We want to test the null hypothesis $H_0 : P = P_0$ against the alternative $H_1 : P \neq P_0$, where P_0 is a given probability measure, absolutely continuous with respect to Lebesgue measure. Let F_0 denote the distribution function corresponding to P_0.

Set $\mathcal{A} \subset \mathcal{B}$ to be the class of all closed and half open intervals $A = [x, y]$, $(-\infty, y]$ or $[x, \infty)$, with $x, y \in I\!R$ such that $0 < P_0(A) < 1$.

Define the main object of our interest, the generalized empirical P-P plot as

$$m_n(t) := \sup\{P_n(A) \,:\, P_0(A) \leq t,\, A \in \mathcal{A}\}, \quad t \in [0,1].$$

Figures 6.1 and 6.2 show the examples of the theoretical and empirical versions of the generalized P-P plot. Observe that when $\mathcal{A}$ would be $\{(-\infty, y] \,:\, y \in I\!R\}$ we would obtain that m_n is the classical P-P plot. Define the generalized empirical P-P plot process as

$$M_n(t) := \sqrt{n}\big(\sup\{P_n(A) \,:\, P_0(A) \leq t,\, A \in \mathcal{A}\} - t\big), \quad t \in [0,1].$$

Then under the null hypothesis

$$\begin{aligned} M_n(t) &= \sup\{\sqrt{n}(P_n(A) - P_0(A)) \,:\, P_0(A) \leq t,\, A \in \mathcal{A}\} \\ &= \sup\{\alpha_n(v) - \alpha_n(u) \,:\, P_0([u,v]) = t,\, 0 \leq u < v \leq 1\} \\ &= \sup\{\Gamma_n(v) - \Gamma_n(u) \,:\, v - u = t,\, 0 \leq u < v \leq 1\} \\ &= \sup\{\Gamma_n(A) \,:\, V(A) = t,\, A \in \mathcal{A}_{[0,1]}\}, \end{aligned}$$

where $\alpha_n(t)$, $t \in [0,1]$, is the empirical process and $\Gamma_n(A) := \Gamma_n(v) - \Gamma_n(u-)$, $A = [u, v]$ is the uniform empirical process indexed by the class $\mathcal{A}_{[0,1]}$, that is the restriction of $\mathcal{A}$ on $[0,1]$. This implies that the process M_n is distribution-free under the null hypothesis.

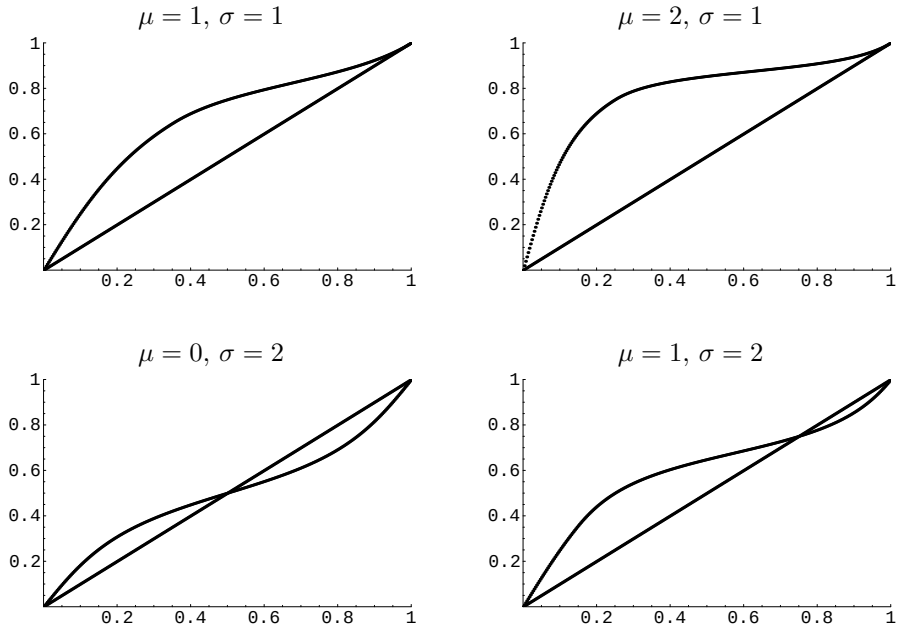

Figure 6.1: Generalized P-P plots of Cauchy(μ, σ) against Cauchy(0,1) distribution.

Let us now recall the Skorokhod construction for $\mathbb{R}$-valued random variables from Section 1.2; there exists a probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{I\!P})$, carrying processes $\widetilde{\alpha}_n(t) = \sqrt{n}(\widetilde{F}_n(t) - F(t))$, $n \geq 1$, and $\widetilde{B}(F(t))$, $t \in \mathbb{R}$, such that

$$\sup_{t \in \mathbb{R}} |\widetilde{\alpha}_n(t) - \widetilde{B}(F(t))| \to 0 \quad \text{a.s.}, \quad n \to \infty. \tag{6.1}$$

Define processes $\widetilde{\alpha}_n$, $n \geq 1$, and $\widetilde{B}_P$ indexed by the class $\mathcal{A}$ as

$$\begin{aligned} &\widetilde{\alpha}_n(A) := \widetilde{\alpha}_n(t) - \widetilde{\alpha}_n(s-), \\ &\widetilde{B}_P(A) := \widetilde{B}(F(t)) - \widetilde{B}(F(s)), \quad A = [s,t] \in \mathcal{A}. \end{aligned} \tag{6.2}$$

The process $\widetilde{B}_P$ is P-Brownian bridge, indexed by $\mathcal{A}$, and (6.1) implies that

$$\sup_{A \in \mathcal{A}} |\widetilde{\alpha}_n(A) - \widetilde{B}_P(A)| \to 0 \quad \text{a.s.}, \quad n \to \infty. \tag{6.3}$$

From (6.3) it is easy to obtain the following statement. We will drop the tildes for notational convenience.

Theorem 6.1 *When $P = P_0$, we have as $n \to \infty$,*

$$\sup_{t \in [0,1]} \Big| M_n(t) - \sup_{\substack{V(A)=t \\ A \in \mathcal{A}}} B(A) \Big| \to 0 \quad \textit{a.s.}. \tag{6.4}$$

Observe that Theorem 6.1 is the special case of Theorem 6.2 when $H_n \equiv 0$.

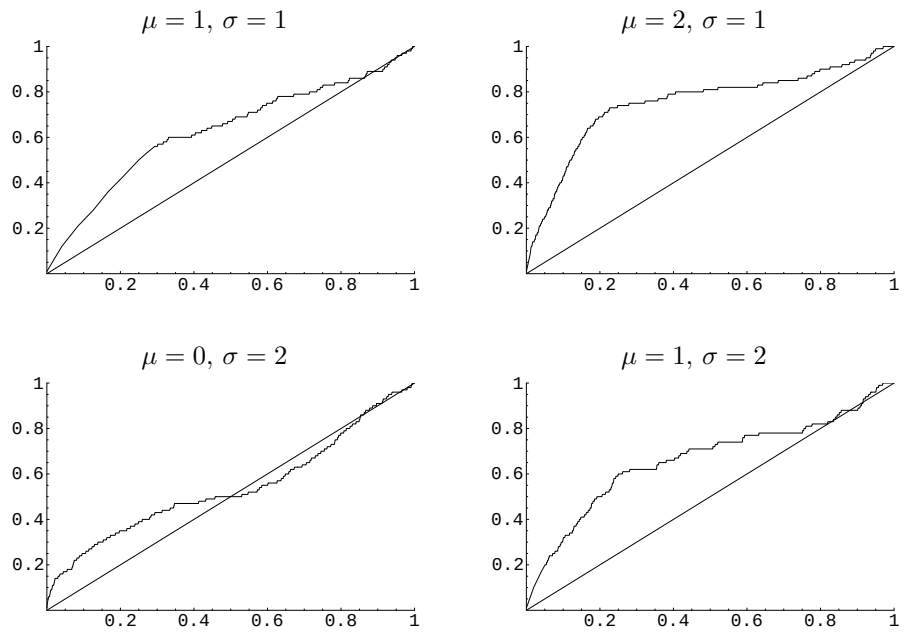

Figure 6.2: Generalized Empirical P-P plots of Cauchy(μ,σ) against Cauchy$(0,1)$ distribution for $n = 100$.

Then from Theorem 1.1 we obtain that for any function $\psi : D \to \mathbb{R}$ that is $(\mathcal{D}, \mathcal{B})$-measurable and continuous on C for the supremum metric,

$$\psi(M_n(t)) \xrightarrow{w} \psi\Big(\sup_{\substack{V(A)=t \\ A\in\mathcal{A}}} B(A)\Big), \quad \text{as} \quad n \to \infty \quad t \in [0,1],$$

hence

$$T_n = \sup_{t\in[0,1]} M_n(t) \xrightarrow{d} \sup_{A\in\mathcal{A}} B(A) = \sup_{0\le u<v\le 1} (B(v) - B(u)). \tag{6.5}$$

Let us now show that T_n is consistent (against all fixed alternatives). Observe that when $P \neq P_0$, there exists $t_0 \in (0,1)$, such that for some $A^* \in \mathcal{A}$, $P_0(A^*) = t_0$ and $P(A^*) > t_0$, then suppose $P(A^*) = t_0 + \varepsilon$, for some $\varepsilon > 0$. Then trivially for n large enough with probability one

$$\begin{aligned} M_n(t_0) = &\sup_{\substack{P_0(A)=t_0 \\ A\in\mathcal{A}}} \big(\alpha_n(A) + \sqrt{n}(P(A) - P_0(A))\big) \\ \ge &\alpha_n(A^*) + \sqrt{n}\varepsilon, \end{aligned} \tag{6.6}$$

and hence

$$\sup_{t\in[0,1]} M_n(t) \to \infty \quad \text{a.s.}, \quad n \to \infty.$$

6.2.2 Contiguous alternatives

The concept of absolute continuity of two measures in context of asymptotic theory has been generalized by the concept of contiguity of two sequences of measures. By definition the sequence of probability measures $Q^{(n)}{}_{n\geq 1}$ is contiguous with respect to the sequence of probability measures $P^{(n)}{}_{n\geq 1}$ if for any sequence of measurable sets A_n, $\lim\limits_{n\to\infty} P^{(n)}(A_n) = 0$ implies $\lim\limits_{n\to\infty} Q^{(n)}(A_n) = 0$. Contiguity of probability measures has been studied widely in many applications. Contiguity of two sequences of measures implies their closeness in Hellinger distance, which is strongly related to the total variational distance. The last plays an important role in defining the optimality of goodness-of-fit or two-sample tests (see, e.g., Oosterhoff and van Zwet (1979)).

Let again $X_1, \ldots, X_n$, $n \geq 1$, be a sequence of i.i.d. random variables defined on a probability space $(\Omega, \mathcal{F}, I\!P)$ taking values in $I\!R$, from an unknown common distribution P, that is absolutely continuous with respect to Lebesgue measure and has the corresponding distribution function F.

Our goodness-of-fit test for contiguous alternatives will consist of the following; suppose that under the null hypothesis each X_i has a known distribution P_0 ($H_0 : P = P_0$), that is absolutely continuous with respect to Lebesgue measure and its corresponding distribution function F_0 is continuous. Suppose under the alternative hypothesis each X_i has a distribution $P_1^{(n)}$ ($H_1 : P = P_1^{(n)}$) defined by

$$\left(\frac{dP_1^{(n)}}{dP_0}(x)\right)^{1/2} = 1 + \frac{1}{2\sqrt{n}} h_n(x), \tag{6.7}$$

and let $F_1^{(n)}$ denote this corresponding distribution function. Clearly $P_1^{(n)}$, $n \geq 1$, is absolutely continuous with respect to P_0. The necessary and sufficient conditions for contiguity that is of interest to us are as follows

$$\overline{\lim_{n\to\infty}} \int_{I\!R} h_n^2(x) dP_0(x) < \infty, \tag{**i**}$$

$$n I\!P_1\left\{X_i \in \left\{x : \frac{dP_1^{(n)}}{dP_0}(x) > K_n\right\}\right\} \to 0, \quad \text{for any sequence} \quad K_n \to \infty \tag{**ii**}$$

(see e.g. Oosterhoff and van Zwet (1979)). Here $I\!P_1$ denotes the probability measure on $(\Omega, \mathcal{F})$ when $P = P_1^{(n)}$. Note that when $h_n \equiv 0$ from (6.7) we obtain that these conditions remain true. Hence the alternative hypothesis implies the null hypothesis. Therefore while dealing with the testing procedures, assume that under the alternative $h_n \not\equiv 0$.

In the literature often a stronger condition than (i) and (ii) is considered. Sometimes we will assume that there exists a function h such that

$$0 < \int_{I\!R} h^2(x) dP_0(x) < \infty \quad \text{and} \quad \int_{I\!R} (h_n(x) - h(x))^2 dP_0(x) \to 0, \quad n \to \infty. \quad \textbf{(iii)}$$

Next define a sequence of functions indexed by the class of sets $\mathcal{A} \in \mathcal{B}$, where $\mathcal{A}$ and $\mathcal{B}$ were defined above

$$H(A) := \int_A h(x) dP_0(x), \quad H_n(A) := \int_A h_n(x) dP_0(x) \tag{6.8}$$

and

$$\big\|h_n\big\|_A := \left[\int_A h_n^2(x) dP_0(x) \right]^{\frac{1}{2}} \quad A \in \mathcal{A}.$$

Let us now investigate the limiting behavior of the generalized P-P plot process when (6.7), (i) and (ii) hold. Again apply Skorokhod construction, then there exists a probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{I\!P})$ carrying processes $\widetilde{\alpha}_n(t)$, $n \geq 1$, $\widetilde{B}(F_1^{(n)}(t))$, $n \geq 1$, and $\widetilde{B}(F_0(t))$, with $t \in I\!R$, such that

$$\sup_{t \in [0,1]} |\widetilde{\alpha}_n(t) - \widetilde{B}(F_1^{(n)}(t))| \to 0 \ \text{ a.s.}, \quad n \to \infty. \tag{6.9}$$

Define processes $\widetilde{\alpha}_n$, $n \geq 1$, $\widetilde{B}_{P_1}$, $n \geq 1$, and $\widetilde{B}_{P_0}$ all indexed by the class $\mathcal{A}$

$$\begin{aligned} \widetilde{\alpha}_n(A) &:= \widetilde{\Gamma}_n(F_1^{(n)}(t)) - \widetilde{\Gamma}_n(F_1^{(n)}(s-)), \\ \widetilde{B}_{P_1}(A) &:= \widetilde{B}(F_1^{(n)}(t)) - \widetilde{B}(F_1^{(n)}(s)), \\ \widetilde{B}_{P_0}(A) &:= \widetilde{B}(F_0(t)) - \widetilde{B}(F_0(s)), \quad A = [s,t] \in \mathcal{A}. \end{aligned} \tag{6.10}$$

The processes $\widetilde{B}_{P_0}$ and $\widetilde{B}_{P_1}$ are P_0- and $P_1^{(n)}$-Brownian bridges, indexed by $\mathcal{A}$. Note that since for all $n \geq 1$, $P_1^{(n)}$ is absolutely continuous with respect to P_0, the process $\widetilde{B}_{P_1}$ will be uniformly continuous in d_0 on $\mathcal{A}$. Then (6.9) and (6.10) imply that

$$\sup_{A \in \mathcal{A}} |\widetilde{\alpha}_n(A) - \widetilde{B}_{P_1}(A)| \to 0 \ \text{ a.s.}, \quad n \to \infty. \tag{6.11}$$

Henceforth, we will drop the tildes for the notational convenience.

Theorem 6.2 *In the setting above when (6.7) holds, under conditions (i) and (ii) we have that*

$$\sup_{t \in [0,1]} \Big| M_n(t) - \sup_{\substack{P_0(A) = t \\ A \in \mathcal{A}}} \big(B_{P_0}(A) + H_n(A) \big) \Big| \to 0 \quad \textit{a.s.} \ \ n \to \infty. \tag{6.12}$$

Note that as Theorem 6.2 is stated for the alternative hypothesis, it implies the corresponding statement for the null hypothesis, when $h_n \equiv 0$.

It is easy to see that the condition (iii) implies (i) and (ii) and hence the following corollary to Theorem 6.2 holds true.

Corollary 6.1 *Theorem 6.2 remains true when condition (i) is replaced by condition (iii).*

Second corollary to Theorem 6.2 deals with the case of random-size samples, which occurs often in practice. Let N_n, $n \geq 1$, be a sequence of random variables defined on the probability space $(\Omega, \mathcal{F}, I\!P)$, taking values in $I\!N$ and independent from the sample $X_1, X_2, \ldots$. Suppose also that

$$N_n \xrightarrow{I\!P} \infty.$$

Then by the Skorokhod construction as above there exists a probability space, carrying independent versions of α_n, $n \geq 1$, and N_n, $n \geq 1$, and the version of the Brownian bridge. We skip tildes for convenience. Then

$$N_n \to \infty \text{ a.s. } n \to \infty$$

and the following result follows immediately from Theorem 6.2.

Corollary 6.2 *Suppose conditions (i) and (ii) hold, then*

$$\sup_{t\in[0,1]} \Big| M_{N_n}(t) - \sup_{\substack{P_0(A)=t \\ A\in\mathcal{A}}} \big(B_{P_0}(A) + H_{N_n}(A) \big) \Big| \to 0 \quad a.s. \ \ n \to \infty, \tag{6.13}$$

where

$$M_{N_n}(t) := \sqrt{N_n}\big(\sup\{ P_{N_n}(A) \,:\, P_0(A) \leq t,\, A \in \mathcal{A}\} - t \big), \quad t \in [0,1].$$

Remark 6.1 *(Scan statistics.) Generally, the scan statistic is defined in terms of continuous scanning with a window of fixed length. Since the scan statistic searches for the maximum mass it can be used for testing for uniformity (see, e.g., Dijkstra et al. (1984)). The test statistic T_n defined above is an analogue of the scan statistic, though the length of its scanning window varies and this makes possible to detect clusters of small unknown size.*

6.3 Two-sample problem

6.3.1 Fixed alternatives

Similar to the previous section, first consider the two-sample problem for fixed alternatives. Let $X_{11} \ldots, X_{1n_1}$, $n_1 \geq 1$, and $X_{21} \ldots, X_{2n_2}$, $n_2 \geq 1$, be two independent random samples defined on a probability space $(\Omega, \mathcal{F}, I\!P)$, taking values in $I\!R$, from

unknown distributions P_1 and P_2, respectively. Suppose as well that for $j = 1, 2$, P_j is absolutely continuous with respect to Lebesgue measure and has distribution function F_j. Our two-sample problem consists of testing the null hypothesis $H_0 : P_1 = P_2$ against the alternative $H_1 : P_1 \neq P_2$. Let $\mathcal{B}, \mathcal{A}$ and d_0 be as defined in the previous section.

Define the generalized empirical P-P plot as

$$m_{n_1 n_2}(t) := \sup\{P_{n_1}(A) \,:\, P_{n_2}(A) \leq t,\, A \in \mathcal{A}\}, \quad t \in [0,1].$$

(Observe that here as well when $\mathcal{A}$ would be $\{(-\infty, y] \,:\, y \in I\!R\}$ we would get the classical empirical P-P plot.) Then the generalized P-P plot process can be defined as

$$M_{n_1 n_2}(t) := \sqrt{\frac{n_1 n_2}{n}}\Big(\sup\{P_{n_1}(A) \,:\, P_{n_2}(A) \leq t,\, A \in \mathcal{A}\} - t\Big), \quad t \in [0,1].$$

Similar to the one-sample case, by using the probability integral transform we obtain that $M_{n_1 n_2}$ is distribution-free under the null hypothesis. The following statement holds true on some probability space $\{\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{I\!P}\}$ and is the special case of Theorem 6.4 below when $H_{n_1 n_2} \equiv 0$.

Theorem 6.3 *When $P_1 = P_2$, we have*

$$\sup_{t \in [0,1]} \Big| M_{n_1 n_2}(t) - \sup_{\substack{V(A) = t \\ A \in \mathcal{A}}} B(A) \Big| \to 0 \quad a.s. \quad n \to \infty. \tag{6.14}$$

Note that it is again easy to show that $M_{n_1 n_2}$ is consistent (for all fixed alternatives).

6.3.2 Contiguous alternatives

Let $\{X_{1i}\}_{i=1}^{n_1}$, $n_1 \geq 1$, and $\{X_{2i}\}_{i=1}^{n_2}$, $n_2 \geq 1$, be two independent sequences of i.i.d. random variables on a probability space $(\Omega, \mathcal{F}, I\!P)$.

The generalization of one-sample problem to two-sample problem is rather evident. For the distributions $P_1^{(n)}$ and $P_2^{(n)}$ of X_{1i} and X_{2i}, respectively, we have that under the null hypothesis $P_1^{(n)} = P_2^{(n)} = P_0$, where P_0 is a probability measure absolutely continuous with respect to Lebesgue measure. While under the alternative $P_1^{(n)}$ and $P_2^{(n)}$ depend on n_1 and n_2 and satisfy the following equations

$$\left(\frac{dP_1^{(n)}}{dP_0}(x)\right)^{\frac{1}{2}} = 1 + \frac{1}{2}\sqrt{\frac{n_2}{n_1 n}} h_{1n}(x), \tag{6.15}$$

$$\left(\frac{dP_2^{(n)}}{dP_0}(x)\right)^{\frac{1}{2}} = 1 - \frac{1}{2}\sqrt{\frac{n_1}{n_2 n}} h_{2n}(x), \tag{6.16}$$

$$\text{with } n = n_1 + n_2, \quad n_1 = n_1(n) \text{ and } n_1 \to \infty \text{ if } n \to \infty;$$
$$n_2 = n_2(n) \text{ and } n_2 \to \infty \text{ if } n \to \infty.$$

Here as well we assume that these measures are contiguous and state the necessary and sufficient conditions for contiguity:

$$n_j I\!P_j\left\{X_i \in \left\{x : \frac{dP_j^{(n)}}{dP_0}(x) > K_n\right\}\right\} \to 0, \text{ for any sequence } K_n \to \infty; \qquad (\mathbf{iv})$$

$$\overline{\lim_{n\to\infty}} \int_{I\!R} h_{jn}^2(x) dP_0(x) < \infty \text{ for } j = 1, 2, \qquad (\mathbf{v})$$

$$\int_{I\!R} \big(h_{1n}(x) - h_{2n}(x)\big)^2 dP_0(x) \to 0, \quad n \to \infty, \qquad (\mathbf{vi})$$

where for $j = 1, 2$ $I\!P_j$ denotes the probability measure on $(\Omega, \mathcal{F})$ when $P = P_j^{(n)}$. As in the one-sample case the alternative hypothesis implies the null hypothesis here as well. Hence assume that $h_{jn} \not\equiv 0$, $j = 1, 2$, while performing the testing procedure. Note that the condition (vi) together with (v) for $j = 1$, implies (v) for $j = 2$. For each $n \geq 1$, define the sequence of functions $H_{n_1 n_2}$ indexed by the class of closed intervals $\mathcal{A}$,

$$H_{n_1 n_2}(A) := \frac{n_2}{n} \int_A h_{1n}(x) dP_0(x) + \frac{n_1}{n} \int_A h_{2n}(x) dP_0(x). \qquad (6.17)$$

Observe that if $h_{1n} \equiv h_{2n} =: h_n$, we obtain that $H_{n_1 n_2} \equiv H_n$, with H_n as in the previous section. Set

$$\big\|h_{jn}\big\|_A := \left[\int_A h_{jn}^2(x) dP_0(x)\right]^{\frac{1}{2}} \text{ for } j = 1, 2. \qquad (6.18)$$

Let P_{jn_j} be the empirical distribution of the sample $X_{j1}, \ldots, X_{jn_j}$, $n_j \geq 1$, $j = 1, 2$

$$P_{jn_j}(B) = \frac{1}{n_j} \sum_{i=1}^{n_j} I_B(X_{ji}), \quad B \in \mathcal{B}.$$

For each $t \in [0, 1]$ and $n_1, n_2 \geq 1$, define the two-sample generalized P-P plot process

$$M_{n_1,n_2}(t) := \sqrt{\frac{n_1 n_2}{n}}\Big(\sup\{P_{1n_1}(A) \,:\, P_{2n_2}(A) \leq t,\, A \in \mathcal{A}\} - t\Big). \qquad (6.19)$$

As we proceed behavior of this process will be our main interest. Similarly to the one-sample case we will construct Gaussian processes B_{jP_j} and B_{jP_0} indexed by the class of closed intervals $\mathcal{A}$. Consider two independent sequences $U_{j1}, \ldots, U_{jn_j}$, $n_j \geq 1$, for $j = 1, 2$ of i.i.d. uniform random variables defined on some probability space $(\Omega', \mathcal{F}', I\!P')$ with values in $[0, 1]$. Let Γ_{jn_j} be the uniform empirical process

$$\Gamma_{jn_j}(t) = \frac{1}{\sqrt{n_j}} \sum_{i=1}^{n_j} \big[I_{[0,t]}(U_{ji}) - t\big], \quad t \in [0, 1], \quad n_j \geq 1, \quad \text{for } j = 1, 2.$$

The process Γ_{jn_j} converges weakly to a Brownian bridge B_j in $(D, \mathcal{D})$ and the processes B_1 and B_2 are independent. Given this weak convergence by the Skorokhod construction there exists a probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{I\!P})$ and a sequence of the random processes $\widetilde{\Gamma}_{j1}, \widetilde{\Gamma}_{j2}, \ldots$ in $(D, \mathcal{D})$ and independent processes $\widetilde{B}_j$ in $(C, \mathcal{C})$ all defined on the same probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{I\!P})$, with

$$\widetilde{B}_j \stackrel{d}{=} B_j, \ \widetilde{\Gamma}_{jn_j} \stackrel{d}{=} \Gamma_{jn_j}, \quad n_j \geq 1,$$

and

$$\sup_{t\in[0,1]} |\widetilde{\Gamma}_{jn_j}(t) - \widetilde{B}_j(t)| \to 0 \ \text{ a.s.}, \quad n_j \to \infty, \quad j = 1, 2. \tag{6.20}$$

For $j = 1, 2$ define the processes $\widetilde{\alpha}_{jn_j}$, $n_j \geq 1$, $\widetilde{B}_{jP_j}$, $n_j \geq 1$, and $\widetilde{B}_{jP_0}$ indexed by the class $\mathcal{A}$

$$\begin{aligned}
\widetilde{\alpha}_{jn_j}(A) &:= \widetilde{\Gamma}_{jn_j}(F_j^{(n)}(t)) - \widetilde{\Gamma}_{jn_j}(F_j^{(n)}(s-)), \\
\widetilde{B}_{jP_j}(A) &:= \widetilde{B}_j(F_j^{(n)}(t)) - \widetilde{B}_j(F_j^{(n)}(s)), \\
\widetilde{B}_{jP_0}(A) &:= \widetilde{B}_j(F_0(t)) - \widetilde{B}_j(F_0(s)), \quad A = [s, t] \in \mathcal{A},
\end{aligned} \tag{6.21}$$

where $F_j^{(n)}$ is the distribution function corresponding to $P_j^{(n)}$, for $j = 1, 2$. Note that

$$\widetilde{\alpha}_{jn_j}(A) = \sqrt{n_j}\big(\widetilde{P}_{jn_j}(A) - P_j^{(n)}(A)\big), \text{ for } j = 1, 2, \tag{6.22}$$

where $\widetilde{P}_{jn_j}$ is the empirical measure of the triangular array of the sequence of random variables $\big\{(F_j^{(n)})^{-1}(\widetilde{U}_{ji})\big\}_{i\geq 1}$. Then (6.20) yields

$$\sup_{A\in\mathcal{A}} \big|\widetilde{\alpha}_{jn_j}(A) - \widetilde{B}_{jP_j}(A)\big| \to 0 \ \text{ a.s.}, \quad n_j \to \infty, \quad j = 1, 2. \tag{6.23}$$

The processes $\widetilde{B}_{jP_j}$ and $\widetilde{B}_{jP_0}$ are $P_j^{(n)}$- and P_0-Brownian bridges, respectively, and $\widetilde{B}_{1P_1}$ and $\widetilde{B}_{2P_2}$ are independent as are $\widetilde{B}_{1P_0}$ and $\widetilde{B}_{2P_0}$ processes.

For each $n_1, n_2 \geq 1$ construct the process $\widetilde{B}_{P_0}^{(n)}$ as follows

$$\widetilde{B}_{P_0}^{(n)}(A) := \sqrt{\frac{n_2}{n}}\widetilde{B}_{1P_0}(A) - \sqrt{\frac{n_1}{n}}\widetilde{B}_{2P_0}(A), \quad A \in \mathcal{A}. \tag{6.24}$$

Note that for each fixed n, $\widetilde{B}_{P_0}^{(n)}$ is a P_0-Brownian bridge.

From now on we will drop the tildes, for notational convenience.

Theorem 6.4 *Assume that the probability measures $P_1^{(n)}$ and $P_2^{(n)}$ defined by (6.15) and (6.16) satisfy conditions (iv) and (v), then with probability one*

$$\sup_{t\in[0,1]} \Big| M_{n_1,n_2}(t) - \sup_{\substack{P_0(A)=t \\ A\in\mathcal{A}}} \big(B_{P_0}^{(n)}(A) + H_{n_1n_2}(A)\big)\Big| \to 0, \quad n \to \infty. \tag{6.25}$$

6.4 Numerical results

In this section we use the setting and notation introduced in Section 6.2. Let us first give a brief description of an algorithm for simulating the test statistic T_n. For a given sample $X_1, \ldots, X_n$, $n \geq 1$, rewrite m_n as

$$m_n(t) = \sup_{\substack{v-u=t \\ 0 \leq u < v \leq 1}} \overline{P}_n([u, v]),$$

where $\overline{P}_n([u, v])$ is the empirical measure of the interval $[u, v]$, of the transformed sample $F_0(X_1), \ldots, F_0(X_n)$. It is easy to see that m_n is right continuous step-function taking values $1/n, 2/n, \ldots, 1$, hence we have to obtain its jumping points. Since each observation can be covered by the closed interval of length 0,

$$m_n(t) = 1/n \quad \text{for} \quad 0 \leq t < \min_{1 \leq i \leq n-1} \{Y_{(i+1)} - Y_{(i)}\},$$

where $Y_{(i)}$'s are the ordered statistics of $Y_i = F_0(X_i)$, $1 \leq i \leq n$. Then for k, $0 \leq k \leq n-1$,

$$m_n(t) = \frac{k+1}{n}, W_k \leq t < W_{k+1},$$

where $W_0 = 0$ and $W_k = \min_{1 \leq i \leq n-k} \{Y_{(i+k)} - Y_{(i)}\}$, $1 \leq k \leq n-1$, are the jumping points of m_n. Now computing T_n is trivial.

sample size	10	20	50	100	∞
critical value	1.58	1.6	1.59	1.598	1.64

Table 6.1: Critical values of the statistic T_n when $\alpha = 0.05$.

Each performed simulation presented here consists of 10 000 replications. In Table 6.1 the simulated critical values, corresponding to $\alpha = 0.05$, for the test statistic T_n are given. These critical values can be used to obtain the empirical power of the test statistic. In Table 6.2 simulated values of the empirical power of T_n are presented for:

(a) testing uniformity against the alternative with the density $g(x) = \frac{1}{2\sqrt{x}}$, $x \in [0, 1]$;

(b) goodness-of-fit test: alternative Cauchy (1,1) against null distribution Cauchy (0,1);

(c) goodness-of-fit test: alternative Beta (2,1) against null distribution Normal $(\frac{2}{3}, \frac{1}{3\sqrt{2}})$,

In case (c) the parameters of the Normal distribution were chosen such that it has the same μ and σ as the Beta (2,1). As we mentioned above the test statistic T_n

resembles the scan statistic with its structure, and hence can be used for testing uniformity against spike alternatives (test (a)). Indeed, Table 6.2 shows that the test is highly powerful in the case (a) as well as in the cases (b) and (c).

sample size	10	20	50	100	∞
test (a)	0.12	0.26	0.64	0.93	1
test (b)	0.37	0.89	0.99	1	
test (c)	0.68	1			

Table 6.2: Empirical powers of the goodness-of-fit tests.

In Table 6.3 simulated empirical power when testing uniformity against the contiguous alternatives are presented. Suppose that in (6.7) $h_n \equiv g$, and consider two examples of the function g:

(1) $g_1(x) = 9\, I_{[0,\frac{1}{10}]}(x) + (-1) I_{(\frac{1}{10},1]}(x)$, for $x \in [0,1]$;

(2) $g_2(x) = -2\, I_{[0,\frac{1}{2}]}(x) + 2\, I_{(\frac{1}{2},1]}(x)$, for $x \in [0,1]$.

Simulation results show that the test is not equally powerful for these alternatives, though for all cases the simulated empirical power is greater than α.

sample size	10	20	50	100
g_1	0.06	0.07	0.14	0.24
g_2	0.29	0.35	0.36	0.36

Table 6.3: Empirical powers for contiguous alternatives.

6.5 Proofs

Here we present the proofs of Sections 6.2 and 6.3

Proof of Theorem 6.2 From (6.7) we have that

$$\frac{dP_1^{(n)}}{dP_0}(x) = 1 + \frac{1}{\sqrt{n}} h_n(x) + \frac{1}{4n} h_n^2(x).$$

Then for $A \in \mathcal{A}$

$$\begin{aligned} P_1^{(n)}(A) &= P_0(A) + \frac{1}{\sqrt{n}} \int_A h_n(x) dP_0(x) + \frac{1}{4n} \int_A h_n^2(x) dP_0(x) \\ &= P_0(A) + \frac{1}{\sqrt{n}} H_n(A) + \frac{1}{4n} \|h_n\|_A^2. \end{aligned} \tag{6.26}$$

By the absolute continuity of P_0 we obtain from (6.26)

$$\begin{aligned} &M_n(t)\\ &= \sqrt{n}\sup\{P_n(A) - t \,:\, P_0(A) = t,\ A \in \mathcal{A}\}\\ &= \sup\Big\{\sqrt{n}\big(P_n(A) - P_1^{(n)}(A)\big) + \sqrt{n}\big(P_1^{(n)}(A) - P_0(A)\big) : P_0(A) = t,\, A \in \mathcal{A}\Big\}\\ &= \sup\Big\{\alpha_n(A) + H_n(A) + \frac{1}{4\sqrt{n}}\big\|h_n\big\|_A^2 \,:\, P_0(A) = t,\ A \in \mathcal{A}\Big\}, \end{aligned} \tag{6.27}$$

which yields

$$\begin{aligned} &\sup_{t\in[0,1]}\Big|M_n(t) - \sup_{\substack{P_0(A)=t\\ A\in\mathcal{A}}}\big(B_{P_0}(A) + H_n(A)\big)\Big|\\ &= \sup_{t\in[0,1]}\Big|\sup_{\substack{P_0(A)=t\\ A\in\mathcal{A}}}\big(\alpha_n(A) + H_n(A) + \frac{1}{4\sqrt{n}}\|h_n\|_A^2\big)\\ &- \sup_{\substack{P_0(A)=t\\ A\in\mathcal{A}}}\big(B_{P_0}(A) + H_n(A)\big)\Big| \le \sup_{t\in[0,1]}\Big[\sup_{\substack{P_0(A)=t\\ A\in\mathcal{A}}}\big|\alpha_n(A) - B_{P_1}(A)\big|\\ &+ \sup_{\substack{P_0(A)=t\\ A\in\mathcal{A}}}\big|B_{P_1}(A) - B_{P_0}(A)\big| + \sup_{\substack{P_0(A)=t\\ A\in\mathcal{A}}}\frac{1}{4\sqrt{n}}\big\|h_n\big\|_A^2\Big] \end{aligned} \tag{6.28}$$

$$\le \sup_{A\in\mathcal{A}}\big|\alpha_n(A) - B_{P_1}(A)\big| + \sup_{A\in\mathcal{A}}\big|B_{P_1}(A) - B_{P_0}(A)\big| + \frac{1}{4\sqrt{n}}\big\|h_n\big\|_{I\!R}^2 \tag{6.29}$$

To complete our proof we have to show that with probability one each term in (6.29) converges to zero. By (6.26) and condition (i) it remains to show that

$$\sup_{A\in\mathcal{A}}|B_{P_0}(A) - B_{P_1}(A)| \to 0 \ \text{ a.s.}, \quad n \to \infty. \tag{6.30}$$

By the uniform continuity of B this follows from

$$\sup_{A\in\mathcal{A}}|P_0(A) - P_1^{(n)}(A)| \to 0, \qquad n \to \infty. \tag{6.31}$$

However, this is equivalent to the following

$$\sup_{A\in\mathcal{A}}\Big|\frac{1}{\sqrt{n}}H_n(A) + \frac{1}{4n}\big\|h_n\big\|_A^2\Big| \to 0, \quad n \to \infty. \tag{6.32}$$

Observe that

$$\begin{aligned} &\sup_{A\in\mathcal{A}}\left|\frac{1}{\sqrt{n}}\int_A h_n(x)dP_0(x) + \frac{1}{4n}\int_A h_n^2(x)dP_0(x)\right|\\ &\qquad \le \sup_{A\in\mathcal{A}}\left|\frac{1}{\sqrt{n}}\int_A h_n(x)dP_0(x)\right| + \sup_{A\in\mathcal{A}}\frac{1}{4n}\int_A h_n^2(x)dP_0(x). \end{aligned} \tag{6.33}$$

Further by the Cauchy-Schwarz inequality for any $A \in \mathcal{A}$

$$\overline{\lim_{n\to\infty}} \int_A |h_n(x)| dP_0(x) \le \overline{\lim_{n\to\infty}} \sqrt{\int_{I\!R} h_n^2(x) dP_0(x) \int_{I\!R} I_A^2(x) dP_0(x)} \tag{6.34}$$

$$= \overline{\lim_{n\to\infty}} \sqrt{P_0(A) \int_{I\!R} h_n^2(x) dP_0(x)} \le \overline{\lim_{n\to\infty}} \sqrt{\int_{I\!R} h_n^2(x) dP_0(x)} < \infty.$$

And finally (6.33), (6.34) and condition (i) imply (6.32) and consequently (6.31). □

Proof of Theorem 6.4 From (6.15) and (6.16) we obtain that for $A \in \mathcal{A}$

$$P_1^{(n)}(A) = P_0(A) + \sqrt{\frac{n_2}{n_1 n}} \int_A h_{1n}(x) dP_0(x) + \frac{n_2}{4n_1 n} \int_A h_{1n}^2(x) dP_0(x), \tag{6.35}$$

$$P_2^{(n)}(A) = P_0(A) - \sqrt{\frac{n_1}{n_2 n}} \int_A h_{2n}(x) dP_0(x) + \frac{n_1}{4n_2 n} \int_A h_{2n}^2(x) dP_0(x). \tag{6.36}$$

Then trivially

$$\begin{aligned} &\sqrt{\frac{n_1 n_2}{n}} \big(P_1^{(n)}(A) - P_2^{(n)}(A)\big) \\ &\qquad = H_{n_1 n_2}(A) + \sqrt{\frac{n_2^3}{16 n_1 n^3}} \|h_{1n}\|_A^2 - \sqrt{\frac{n_1^3}{16 n_2 n^3}} \|h_{2n}\|_A^2. \end{aligned} \tag{6.37}$$

Rewrite M_{n_1,n_2} as follows,

$$M_{n_1,n_2}(t) \overset{\text{a.s.}}{=} \sqrt{\frac{n_1 n_2}{n}} \sup \Big\{ \big(P_{1n_1}(A) - t\big) \,:\, P_{2n_2}(A) = \frac{\lfloor n_2 t \rfloor}{n_2},\, A \in \mathcal{A} \Big\}.$$

Using equations in (6.37) and (6.22) obtain that

$$\begin{aligned} M_{n_1,n_2}(t) \overset{\text{a.s.}}{=} \sup_{\substack{P_{2n_2}(A) = \overline{t} \\ A \in \mathcal{A}}} \Big[&\sqrt{\frac{n_2}{n}} \alpha_{1n_1}(A) - \sqrt{\frac{n_1}{n}} \alpha_{2n_2}(A) + H_{n_1 n_2}(A) \\ &+ \sqrt{\frac{n_2^3}{16 n_1 n^3}} \|h_{1n}\|_A^2 - \sqrt{\frac{n_1^3}{16 n_2 n^3}} \|h_{2n}\|_A^2 + \sqrt{\frac{n_1 n_2}{n}} \big(\overline{t} - t\big) \Big], \end{aligned} \tag{6.38}$$

with $\overline{t} := \frac{\lfloor n_2 t \rfloor}{n_2}$. Then (6.38) imply that with probability one

$$\begin{aligned}
&\sup_{t\in[0,1]}\Big|M_{n_1,n_2}(t) - \sup_{\substack{P_0(A)=t\\ A\in\mathcal{A}}}\big(B^{(n)}_{P_0}(A) + H_{n_1n_2}(A)\big)\Big| \\
&\quad= \sup_{t\in[0,1]}\Big|\sup_{\substack{P_{2n_2}(A)=\overline{t}\\ A\in\mathcal{A}}}\Big[\sqrt{\frac{n_2}{n}}\alpha_{1n_1}(A) - \sqrt{\frac{n_1}{n}}\alpha_{2n_2}(A) + H_{n_1n_2}(A) \\
&\qquad+ \sqrt{\frac{n_2^3}{16n_1n^3}}\big\|h_{1n}\big\|_A^2 - \sqrt{\frac{n_1^3}{16n_2n^3}}\big\|h_{2n}\big\|_A^2 + \sqrt{\frac{n_1n_2}{n}}\big(\overline{t}-t\big)\Big] \\
&\qquad\qquad- \sup_{\substack{P_0(A)=t\\ A\in\mathcal{A}}}\big(B^{(n)}_{P_0}(A) + H_{n_1n_2}(A)\big)\Big|.
\end{aligned} \tag{6.39}$$

For the sake of readability, introduce some notations

$$\mathcal{A}_{0t} := \{A \ : \ P_0(A) = t, \ A\in\mathcal{A}\},$$

$$\mathcal{A}^{(n)}_{2t} := \{A \ : \ P_{2n_2}(A) = \overline{t}, \ A\in\mathcal{A}\},$$

$$W_t^{(n_1n_2)}(A) := \sqrt{\frac{n_2^3}{16n_1n^3}}\big\|h_{1n}\big\|_A^2 - \sqrt{\frac{n_1^3}{16n_2n^3}}\big\|h_{2n}\big\|_A^2 + \sqrt{\frac{n_1n_2}{n}}\big(\overline{t}-t\big).$$

Split the absolute value in (6.39) into two parts. First consider

$$\begin{aligned}
\sup_{t\in[0,1]}\Big\{\sup_{A\in\mathcal{A}^{(n)}_{2t}}\Big[\sqrt{\frac{n_2}{n}}\alpha_{1n_1}(A) - \sqrt{\frac{n_1}{n}}\alpha_{2n_2}(A) + H_{n_1n_2}(A) + W_t^{(n_1n_2)}(A)\Big] \\
- \sup_{A\in\mathcal{A}_{0t}}\big(B^{(n)}_{P_0}(A) + H_{n_1n_2}(A)\big)\Big\},
\end{aligned} \tag{6.40}$$

which, by (6.24), is equal to

$$\begin{aligned}
&= \sup_{t\in[0,1]}\sup_{A_2\in\mathcal{A}^{(n)}_{2t}}\inf_{A_0\in\mathcal{A}_{0t}}\Big\{\sqrt{\frac{n_2}{n}}\big(\alpha_{1n_1}(A_2) - B_{1P_1}(A_2)\big) \\
&\quad+ \sqrt{\frac{n_1}{n}}\big(B_{2P_2}(A_2) - \alpha_{2n_2}(A_2)\big) + \sqrt{\frac{n_2}{n}}\big(B_{1P_1}(A_2) - B_{1P_0}(A_2)\big) \\
&\quad+ \sqrt{\frac{n_1}{n}}\big(B_{2P_0}(A_2) - B_{2P_2}(A_2)\big) + \big(B^{(n)}_{P_0}(A_2) + H_{n_1n_2}(A_2)\big) \\
&\quad+ W_t^{(n_1n_2)}(A_2) - \big(B^{(n)}_{P_0}(A_0) + H_{n_1n_2}(A_0)\big)\Big\}.
\end{aligned} \tag{6.41}$$

Observe that (6.41) is bounded from above by

$$\begin{aligned}
&\leq \sup_{A\in\mathcal{A}} \sqrt{\frac{n_2}{n}}\Big|\alpha_{1n_1}(A) - B_{1P_1}(A)\Big| + \sup_{A\in\mathcal{A}} \sqrt{\frac{n_1}{n}}\Big|B_{2P_2}(A) - \alpha_{2n_2}(A)\Big| \\
&+ \sup_{A\in\mathcal{A}} \sqrt{\frac{n_2}{n}}\Big|B_{1P_1}(A) - B_{1P_0}(A)\Big| + \sup_{A\in\mathcal{A}} \sqrt{\frac{n_1}{n}}\Big|B_{2P_2}(A) - B_{2P_0}(A)\Big| \\
&\qquad + \sup_{A\in\mathcal{A}} \Big|W_t^{(n_1n_2)}(A)\Big| \\
&+ \sup_{t\in[0,1]} \left| \sup_{A\in\mathcal{A}_{2t}^{(n)}} \Big(B_{P_0}^{(n)}(A) + H_{n_1n_2}(A)\Big) - \sup_{A\in\mathcal{A}_{0t}} \Big(B_{P_0}^{(n)}(A) + H_{n_1n_2}(A)\Big)\right|.
\end{aligned} \tag{6.42}$$

Similarly it can be shown that the other part of the absolute value in (6.39) is also bounded by the expression in (6.42). Using the same arguments as in the previous section and the uniform continuity of B_{jP_j} and B_{jP_0}, respectively, for $j = 1, 2$, we obtain

$$\sup_{A\in\mathcal{A}} |B_{jP_0}(A) - B_{jP_j}(A)| \to 0 \quad \text{a.s.}, \quad n_j \to \infty, \quad \text{for} \quad j = 1, 2. \tag{6.43}$$

Hence by (6.23) and condition (v) it remains to show that

$$\begin{gathered}
\sup_{t\in[0,1]} \left| \sup_{A\in\mathcal{A}_{2t}^{(n)}} \Big(B_{P_0}^{(n)}(A) + H_{n_1n_2}(A)\Big) - \sup_{A\in\mathcal{A}_{0t}} \Big(B_{P_0}^{(n)}(A) + H_{n_1n_2}(A)\Big)\right| \to 0 \\
\text{a.s.}, \quad n \to \infty.
\end{gathered} \tag{6.44}$$

On the other hand, since $B_{P_0}^{(n)} + H_{n_1n_2}$ is d_0-uniformly equicontinuous with probability one (see Lemma 6.1 below), using a similar argument as in (6.41) we have to show that

$$\sup_{t\in[0,1]} \sup_{A_2\in\mathcal{A}_{2t}^{(n)}} \inf_{A_0\in\mathcal{A}_{0t}} d_0(A_2, A_0) \to 0 \quad \text{a.s.}, \quad n \to \infty \tag{6.45}$$

and

$$\sup_{t\in[0,1]} \sup_{A_0\in\mathcal{A}_{0t}} \inf_{A_2\in\mathcal{A}_{2t}^{(n)}} d_0(A_2, A_0) \to 0 \quad \text{a.s.}, \quad n \to \infty. \tag{6.46}$$

We can also state (6.45) as follows: for every $\varepsilon > 0$ we can choose $N_\varepsilon \geq 1$ such that for $n \geq N_\varepsilon$ and for all $t \in [0, 1]$, $A_2 \in \mathcal{A}_{2t}^{(n)}$ there exists a set $A_0 = A_0(A_2, \varepsilon, t) \in \mathcal{A}_{0t}$ and

$$P_0(A_2 \triangle A_0) < \varepsilon \quad \text{a.s..}$$

Take an arbitrary $\varepsilon > 0$. Observe that there exists $N_\varepsilon^{(1)} \geq 1$ such that for $n \geq N_\varepsilon^{(1)}$ and for all $t \in [0, 1]$ and all $A_2 \in \mathcal{A}_{2t}^{(n)}$, $|P_2^{(n)}(A_2) - t| < \frac{\varepsilon}{2}$ a.s.. Next choose $N_\varepsilon^{(2)} \geq 1$ such that for $n \geq N_\varepsilon^{(2)}$ and for all $A_2 \in \mathcal{A}_{2t}^{(n)}$, $|P_0(A_2) - P_2^{(n)}(A_2)| < \frac{\varepsilon}{2}$. Set $N_\varepsilon := max(N_\varepsilon^{(1)},\ N_\varepsilon^{(2)})$. Then trivially for $n \geq N_\varepsilon$ and for all $A_2 \in \mathcal{A}_{2t}^{(n)}$

$$|P_0(A_2) - t| < \varepsilon \quad \text{a.s..}$$

However, since P_0 is absolutely continuous with respect to Lebesgue measure, there exists a set A_0, with $P_0(A_0) = t$ and $A_0 \subset A_2$ or $A_0 \supset A_2$ and hence

$$P_0(A_2 \triangle A_0) < \varepsilon \ \text{ a.s..}$$

Note that (6.46) can be treated similarly. Hence (6.44) is true and thus we have proved the theorem. □

Lemma 6.1 *The collection of continuous functions* $\{B^{(n)}_{P_0} + H_{n_1 n_2} \ : \ n \in I\!N\}$ *is* d_0*-uniformly equicontinuous.*

Proof By definition a collection of continuous functions $\mathcal{F}$ from some metric space (S, e) into other metric space (X, d) is d-uniformly equicontinuous if for every $\varepsilon > 0$ there is a $\delta > 0$ such that $e(x, y) < \delta$ implies $d(f(x), f(y)) < \varepsilon$ for all x and y in S and all f in $\mathcal{F}$. We prove the statement using the well-known fact on a modulus of continuity of a standard Brownian bridge B, (see, e.g., Shorack and Wellner (1986))

$$\lim_{a \downarrow 0} \frac{\sup_{|t-s| \le a} |B(t) - B(s)|}{\sqrt{2a \log(1/a)}} = 1 \ \text{ a.s..}$$

Then by a simple transformation the similar result for P_0-Brownian bridge is true

$$\lim_{a \downarrow 0} \sup_{\substack{d_0(A_1, A_2) \le a \\ A_1, A_2 \in \mathcal{A}}} |B_{P_0}(A_1) - B_{P_0}(A_2)| \to 0 \ \text{ a.s..}$$

Since almost sure convergence yields convergence in probability and the sequence of P_0-Brownian bridges have the same distribution, following holds trivially

$$\lim_{a \downarrow 0} \sup_{\substack{d_0(A_1, A_2) \le a \\ A_1, A_2 \in \mathcal{A}}} |B^{(n)}_{P_0}(A_1) - B^{(n)}_{P_0}(A_2)| \xrightarrow{d} 0. \tag{6.47}$$

By the Skorokhod-Dudley-Wischura theorem there exists a probability space where the result corresponding to (6.47) holds almost surely. For convenience we will not change our notations. Hence for any $\varepsilon > 0$ there is small a such that for all $n \ge 1$

$$\sup_{\substack{d_0(A_1, A_2) \le a \\ A_1, A_2 \in \mathcal{A}}} |B^{(n)}_{P_0}(A_1) - B^{(n)}_{P_0}(A_2)| < \varepsilon \ \text{ a.s.}$$

and this implies that $\{B^{(n)}_{P_0} \ : \ n \in I\!N\}$ is d_0-uniformly equicontinuous.

Let $A_1, A_2 \in \mathcal{A}$. Consider

$$\begin{aligned} &\left| H_{n_1 n_2}(A_1) - H_{n_1 n_2}(A_2) \right| \\ &\qquad \le \left| \frac{n_2}{n} \int_{I\!R} I_{A_1}(x) h_{1n}(x) dP_0(x) - \frac{n_2}{n} \int_{I\!R} I_{A_2}(x) h_{1n}(x) dP_0(x) \right| \\ &\qquad + \left| \frac{n_1}{n} \int_{I\!R} I_{A_1}(x) h_{2n}(x) dP_0(x) - \frac{n_1}{n} \int_{I\!R} I_{A_2}(x) h_{2n}(x) dP_0(x) \right| \end{aligned}$$

$$\leq \frac{n_2}{n}\int_{I\!R} I_{A_1\triangle A_2}(x)\big|h_{1n}(x)\big|dP_0(x) + \frac{n_1}{n}\int_{I\!R} I_{A_1\triangle A_2}(x)\big|h_{2n}(x)\big|dP_0(x). \qquad (6.48)$$

By the Cauchy-Schwarz inequality since $d_0(A_1, A_2) = P_0(A_1\triangle A_2)$

$$\int_{I\!R} I_{A_1\triangle A_2}(x)\big|h_{jn}(x)\big|dP_0(x) \leq \sqrt{\int_{I\!R} h_{jn}^2(x)dP_0(x)\int_{I\!R} I_{A_1\triangle A_2}(x)dP_0(x)}$$
$$= \sqrt{d_0(A_1, A_2)}\sqrt{\int_{I\!R} h_{jn}^2(x)dP_0(x)}.$$

However, by the condition (v) the sequence $\|h_{jn}\|_{I\!R}$, $n \geq 1$ for $j = 1, 2$, is uniformly bounded, hence for any $\varepsilon > 0$ there exists a $\delta > 0$ such that for all $n_1, n_2 \in I\!N$ and any $A_1, A_2 \in \mathcal{A}$, with $d_0(A_1, A_2) < \delta$ we will have that

$$|H_{n_1n_2}(A_1) - H_{n_1n_2}(A_2)| < \varepsilon.$$

Thus $H_{n_1n_2}$ is d_0-uniformly equicontinuous as well. □

Discussion

The generalized P-P plots introduced in Chapter 6 globally preserve the properties of the classical P-P plots. The presented one- and two-sample tests, based on these plots, are general and flexible. The test statistic is easy to compute, using the simple algorithm from Section 6.4. Since the P-P plot process and, consequently, the test statistic is distribution-free under the null hypothesis, we were able to simulate critical values for the case of a finite sample (see Table 6.1). The testing procedure is consistent against all fixed alternatives. The values of the empirical power are indeed quite satisfactory (see Table 6.2).

The proposed test statistic resembles the classical scan statistic by its structure, however it has a window with varying length. The spike alternatives are therefore natural to consider. Although the considered indexing class is the class of closed intervals, thus allowing the detection of only one spike, the procedure can be generalized for the class of finite unions of closed intervals.

The case of the contiguous alternatives is studied and the limiting distribution of the test statistic is derived. Some numerical results for the empirical power for these alternatives are given in Table 6.3.

Bibliography

Ackermann, H. (1983). Multivariate nonparametric tolerance regions: a new construction technique. *Biometrical J.* **25**, 351–359.

Ackermann, H. (1985). Verteilungsfreie Toleranzbereiche fur zirkulare Daten. *EDV in Medizin und Biologie* **16**, 97–99.

Agulló, J. (1996). Exact iterative computation of the multivariate minimum volume ellipsoid estimator with a branch and bound algorithm. In A. Prat (Ed.), *COMPSTAT '96 Barcelona*, Heidelberg, pp. 175–180. Physica-Verlag.

Aitchison, J. (1966). Expected-cover and linear-utility tolerance intervals. *J. Roy. Statist. Soc. Ser.* **B 28**, 57–62.

Aitchison, J. and I. R. Dunsmore (1975). *Statistical prediction analysis*. Cambridge University Press.

Alexander, K. S. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Probab.* **12**, 1041–1067.

Aly, E.-E. A. A. (1986a). Quantile-quantile plots under random censorship. *J. Statist. Plann. Inference* **15**, 123–128.

Aly, E.-E. A. A. (1986b). Strong approximations of the Q-Q process. *J. Multivariate Anal.* **20**, 114–128.

Andrews, D. F., P. J. Bickel, P. J. Hampel, W. H. Rodgers, and J. W. Tukey (1972). *Robust estimation of location: Survey and advances*. Princeton Univ. Press, Princeton, NJ.

Azzalini, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika* **68**, 326–328.

Batschelet, E. (1981). *Circular statistics in biology*. Academic Press, London.

Beirlant, J. and P. Deheuvels (1990). On the approximation of P-P and Q-Q plot processes by Brownian bridges. *Statist. Probab. Lett.* **9**, 241–251.

Beirlant, J. and J. H. J. Einmahl (1995). Asymptotic confidence intervals for the length of the shortt under random censoring. *Statist. Neerlandica* **49**, 1–8.

Beirlant, J., D. M. Mason, and C. Vynckier (1999). Goodness-of-fit analysis for multivariate normality based on generalized quantiles. *Comput. Statist. Data Anal.* **30**, 119–142.

Billingsley, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

Bolthausen, E. (1978). Weak convergence of an empirical process indexed by the closed convex subsets of I^2. *Z. Wahrsch. verw. Gebiete* **43**, 173–181.

Chatterjee, S. K. and N. K. Patra (1980). Asymptotically minimal multivariate tolerance sets. *Calcutta Statist. Assoc. Bull.* **29**, 73–93.

Davies, L. (1992). The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *Ann. Statist.* **20**, 1828–1843.

Deheuvels, P. and J. H. J. Einmahl (1992). Approximations and two-sample tests based on P-P and Q-Q plots of the Kaplan-Meier estimators of lifetime distributions. *J. Multivariate Anal.* **43**, 200–217.

Di Bucchianico, A., J. H. J. Einmahl, and N. A. Mushkudiani (2000). Small nonparametric tolerance regions. *Report 2000-011, Eurandom, Eindhoven, The Netherlands*.

Dijkstra, J. B., T. J. M. Rietjens, and F. W. Steutel (1984). A simple test for uniformity. *Statist. Neerlandica* **1**, 33–44.

Doksum, K. (1977). Some graphical methods in statistics. A review and some extensions. *Statist. Neerlandica* **31**, 53–68.

Donsker, M. D. (1952). Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Statist.* **23**, 277–281.

Doob, J. L. (1949). Heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Statist.* **20**, 393–403.

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probab.* **6**, 899–929.

Dudley, R. M. (1982). Empirical and Poisson processes on classes of sets or functions too large for central limit theorems. *Z. Wahrsch. Verw. Gebiete* **61**, 355–368.

Einmahl, J. H. J. and D. M. Mason (1992). Generalized quantile processes. *Ann. Statist.* **20**, 1062–1078.

Eppstein, D. (1992). New algorithms for minimum area k-gons. In *Proceedings of the Third Annual ACM-SIAM Symposium on Discrete Algorithms*, New York, pp. 83–88. ACM.

Eppstein, D., M. Overmars, G. Rote, and G. Woeginger (1992). Finding minimum area k-gons. *Discrete Comput. Geom.* **7**, 45–58.

Fisher, N. I. (1983). Graphical methods in nonparametric statistics: a review and annotated bibliography. *Internat. Statist. Rev.* **51**, 25–58.

Fisher, N. I. (1993). *Statistical analysis of circular data*. Cambridge University Press, Cambridge.

Fisher, N. I., T. Lewis, and B. J. J. Embleton (1987). *Statistical analysis of spherical data*. Cambridge University Press, Cambridge-New York.

Fraser, D. A. S. (1951). Sequentially determined statistically equivalent blocks. *Ann. Math. Statist.* **22**, 372–381.

Fraser, D. A. S. (1953). Nonparametric tolerance regions. *Ann. Math. Statist.* **24**, 44–55.

Gaenssler, P. (1983). *Empirical Processes*. IMS, Hayward, California.

Girling, A. J. (2000). Rank statistics expressible as integrals under P-P-plots and receiver operating characteristic curves. *J. R. Stat. Soc. Ser.* **B 62**, 367–382.

Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations*. John Wiley & Sons, New York-London-Sydney. A Wiley Publication in Applied Statistics.

Grübel, R. (1988). The length of the shorth. *Ann. Statist.* **16**, 619–628.

Guttman, I. (1957). On the power of optimum tolerance regions when sampling from normal distributions. *Ann. Math. Statist.* **28**, 773–778.

Guttman, I. (1970). *Statistical tolerance regions: classical and Bayesian*. London: Charles Griffin.

Hsieh, F. and B. W. Turnbull (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *Ann. Statist.* **24**, 25–40.

Jílek, M. (1981). A bibliography of statistical tolerance regions. *Math. Operationsforsch. Statist. Ser. Statist.* **12**, 441–456.

Jílek, M. and H. Ackermann (1989). A bibliography of statistical tolerance regions. II. *Statistics* **20**, 165–172.

Li, G., R. C. Tiwari, and M. T. Wells (1996). Quantile comparison functions in two-sample problems, with application to comparisons of diagnostic markers. *J. Amer. Statist. Assoc.* **91**, 689–698.

Lorenz, M. O. (1905). Methods on measuring the concentration of wealth. *J. Amer. Statist. Assoc.* **70**, 209–219.

Mardia, K. V. (1972). *Statistics of directional data*. Academic Press, London-New York. Probability and Mathematical Statistics, No. 13.

Muller, D. W. and G. Sawitzki (1991). Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.* **86**, 738–746.

Mushkudiani, N. A. (2000). Small nonparametric tolerance regions for directional data. *SPOR-Report 2000-06, Eindhoven University of Technology, The Netherlands. To appear in J. Statist. Plann. Inference*.

Nair, V. N. (1981). Plots and tests for goodness of fit with randomly censored data. *Biometrika* **68**, 99–103.

Nair, V. N. (1982). Q-Q plots with confidence bands for comparing several populations. *Scand. J. Statist.* **9**, 193–200.

Nolan, D. (1991). The excess-mass ellipsoid. *J. Multivariate Anal.* **39**, 348–371.

Oosterhoff, J. and W. R. van Zwet (1979). A note on contiguity and Hellinger distance. In J. Jurechkova (Ed.), *Contribution to Statistics, Jaroslav Hajek Memorial Volume*. Dordrecht: Reidel.

Parzen, E. (1993). Change PP plot and continuous sample quantile function. *Comm. Statist. Theory Methods* **22**, 3287–3304.

Polonik, W. (1997). Minimum volume sets and generalized quantile processes. *Stochastic Process. Appl.* **69**, 1–24.

Polonik, W. (1999). Concentration and goodness-of-fit in higher dimensions: (asymptotically) distribution-free methods. *Ann. Statist.* **27**, 1210–1229.

Quade, D. (1973). The pair chart. *Statist. Neerlandica* **27**, 29–45.

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In W. Grossman, G. Pflug, I. Vincze, and W. Wertz (Eds.), *Mathematical Statistics with Applications*, Volume **B**, pp. 283–297. Reidel, Dordrecht.

Rousseeuw, P. and A. Leroy (1988). A robust scale estimator based on the shortest half. *Statist. Neerlandica* **42**, 103–116.

Rousseeuw, P. and B. C. van Zomeren (1991). Robust distances: simulations and cutoff values. In W. Stahel and S. Weisberg (Eds.), *Directions in Robust Statistics and Diagnostics, part II*, Volume 34 of *The IMA Volumes in Mathematics and Its Applications*, pp. 195–203. New York: Springer Verlag.

Ruymgaart, F. H. (1989). Strong uniform convergence of density estimators on spheres. *J. Statist. Plann. Inference* **23**, 45–52.

Sawitzki, G. (1994). Diagnostic plots for one-dimensional data. In P. Dirschedl and R. Ostermann (Eds.), *Papers collected on the Occasion of the 25th Conf. on Statistical Computing at Schloss Reisensburg*. Physica, Heidelberg.

Serfling, R. (2000). Generalized quantile processes based on multivariate depth functions, with applications in nonparametric multivariate analysis. Submitted.

Shephard, G. C. and R. J. Webster (1965). Metric sets of convex bodies. *Mathematika* **12**, 73–88.

Shorack, G. R. and J. A. Wellner (1986). *Empirical Processes with Applications to Statistics*. New York: Wiley.

Silverman, B. W. and D. M. Titterington (1980). Minimum covering ellipses. *SIAM J. Sci. Stat. Comput.* **1**, 401–409.

Swets, I. A. and R. M. Pickett (1982). *Evaluation of Diagnostic Systems: Methods From Signal Detection Theory*. New York: Academic Press.

Tukey, J. W. (1947). Nonparametric estimation, II. Statistical equivalent blocks and tolerance regions - the continuous case. *Ann. Math. Statist.* **18**, 529–539.

Tukey, J. W. (1948). Nonparametric estimation, III. Statistical equivalent blocks and multivariate tolerance regions - the discontinuous case. *Ann. Math. Statist.* **19**, 30–39.

van der Vaart, A. W. (1994). Weak convergence of smoothed empirical processes. *Scand. J. Statist.* **4**, 501–504.

van der Vaart, A. W. (1996). New Donsker classes. *Ann. Probab.* **24**, 2128–2140.

Wald, A. (1942). Setting of tolerance limits when the sample is large. *Ann. Math. Statist.* **13**, 389–399.

Wald, A. (1943). An extension of Wilks' method for setting tolerance limits. *Ann. Math. Statist.* **14**, 45–55.

Webster, R. (1994). *Convexity*. Oxford University Press.

Wilks, S. S. (1941). Determination of sample sizes for setting tolerance limits. *Ann. Math. Statist.* **12**, 91–96.

Wilks, S. S. (1942). Statistical prediction with special reference to the problem of tolerance limits. *Ann. Math. Statist.* **13**, 400–409.

Index

P-Donsker class, **5**, 7, 8, 22, 32, 53
γ-strip, **29**, 33

Alternatives
  contiguous, 2, 71, **75**, **78**
  fixed, 71, **72**, 74, **77**
Arc, **18**, 55
  tolerance, 55

Blaschke Selection Principle, **21**, 27, 28, 31, 53
Brownian bridge, **4**, 71, 80
  P-, **5**, 73, 80

Class of caps, **48**
Contiguity, 75
Countably generated (CG) class, **5**, 8

Data
  circular, **48**, 55
  directional, **48**
  spherical, **48**

Empirical distribution, **5**
Empirically measurable class for P (P-EM), **5**
Essentially uniformly bounded, **33**

Function
  empirical distribution, **4**
  generalized empirical quantile, **7**, 22
  generalized quantile, 1, **7**
  incomplete beta, **15**

Glivenko-Cantelli class, 7
Glivenko-Cantelli theorem, 29

Hausdorff metric, **22**

Median direction, **18**
Minimum volume
  cap (MV-cap), **50**, 51
  ellipsoid, 20, 43
  hyperrectangle, 38
  set (MV-set), 1, **6**, 7, 20, 22, 31, 50

Plot
  probability-probability (P-P), **64**
    empirical, **64**, 67
    generalized, 2, 71
    generalized empirical, **72**, 78
    multivariate, 68, **69**
  quantile-quantile (Q-Q), **65**
    empirical, **65**, 67
    multivariate, 68, **69**
Prediction region, **13**
Process
  empirical, **4**, 72
  generalized empirical P-P plot, 2, **72**, 76, 78
  generalized quantile, 1, **7**
  P-P plot, **67**, 69
    generalized, 71
  Q-Q plot, **67**, 69
  two-sample generalized P-P plot, **79**
  uniform empirical, **4**, 72, 79

Receiver operating characteristic (ROC) curve, **64**

Skorokhod construction, 2, **3**, 23, 53, 73, 80

Skorokhod-Dudley-Wichura representation theorem, *see* Skorokhod construction
Statistic
  consistent, 74, 78
  scan, **77**
Statistically equivalent blocks, 12, **14**
Supremum metric, **3**, 74
Symmetric difference metric, **22**

Test
  goodness-of-fit, 71, 75
  two-sample, 71, 77
Tolerance cap
  guaranteed content, **50**
  mean content, **50**
Tolerance interval, 1
  classical nonparametric, 13, 34
  smoothed, 40
Tolerance region, 11
  nonparametric, **14**
  affine equivariant, 30
  asymptotically minimal, 12, **17**, 30
  coverage of, **13**
  directional, **17**, 47
  for circular data, **48**
  for spherical data, **48**
  guaranteed coverage, **12**, 23, 35
  mean coverage, **12**, 23, 35
  nonparametric, 1, 16, 19
  small, **20**, 22, 23, 34

Uniformly
  bounded a.s., **27**
  equicontinuous, **87**

Vapnik-Chervonenkis (VC) class, **5**, 7, 29, 30, 32, 53

Weak convergence, **3**
  of uniform empirical process, 4
  criteria for, **3**

Samenvatting

Beschouw n onafhankelijke, identiek verdeelde stochastische vectoren met waarden in $\mathbb{R}^k$, $k \geq 1$. Om eigenschappen van multivariate data te onderzoeken wordt vaak een ordening van de data gebruikt. In Einmahl and Mason (1992) worden hiervoor echter univariate functies van het quantiel-type geïntroduceerd, zonder dat een ordening nodig is. Naast voorbereidende resultaten en definities (o.a. de Skorokhodconstructie en minimum-volume verzamelingen) wordt in Hoofdstuk 1 het begrip gegeneraliseerde quantielen gedefineerd en kort toegelicht. De diverse keuzes van index-klassen $\mathcal{A}$ en reële functies λ, gedefinieerd op $\mathcal{A}$, die de gegeneraliseerde quantielen definiëren, maken het mogelijk om deze functies te gebruiken in verscheidene niet-parametrische statistische procedures.

In dit proefschrift worden twee toepassingen van gegeneraliseerde quantielen bestudeerd. In deel I introduceren we een nieuwe methode voor het construeren van niet-parametrische multivariate tolerantiegebieden, terwijl in deel II een geheel ander gebied van de niet-parametrische statistiek wordt behandeld. Hier bestuderen we één- en twee-steekproevenproblemen met behulp van gegeneraliseerde Probability-Probability (P-P) plots.

Een speciaal geval van gegeneraliseerde quantielfuncties brengt ons tot minimum-volume verzamelingen, welke leiden tot een nieuwe benadering van niet-parametrische multivariate tolerantiegebieden. Klassieke niet-parametrische tolerantieintervallen zijn intervallen met twee order statistics als eindpunten, waarbij op voorhand wordt besloten welke order statistics worden gebruikt. Omdat de klassieke methode is gebaseerd op ordening, is er geen natuurlijke wijze om deze methode uit te breiden naar hogere dimensies. Om dit probleem op te lossen werden "statistisch equivalente blokken" en ordeningsfuncties geïntroduceerd. Echter, omdat deze ordeningsfuncties willekeurig gekozen kunnen worden, kan men gebieden met volkomen verschillende vormen krijgen. Bovendien zijn deze gebieden niet noodzakelijk asymptotisch minimaal. De andere benadering uit de literatuur voor het construeren van multivariate tolerantiegebieden is gebaseerd op dichtheidsschatten en geeft asymptotisch minimale tolerantiegebieden. Deze en andere begrippen uit de literatuur m.b.t. niet-parametrische tolerantiegebieden worden gepresenteerd in Hoofdstuk 2.

In Hoofdstukken 3 en 4 definiëren we niet-parametrische tolerantiegebieden als

minimum-volume verzamelingen voor respectievelijk Euclidische en directionele data. In tegenstelling tot de klassieke methode, definiëren we het tolerantiegebied als het kleinste gebied dat een bepaald aantal waarnemingen bevat. We breiden dit idee uit naar hogere dimensies door middel van het definiëren van tolerantiegebieden als de minimum-volume verzamelingen voor een zekere index-klasse, die kan worden gespecialiseerd tot de klasse van verenigingen van een constant eindig aantal ellipsoïden, hyperrechthoeken of convexe verzamelingen voor Euclidische data en tot de klasse van cirkelbogen en bolsegmenten voor directionele data. We leiden een asymptotische theorie af voor deze tolerantiegebieden onder zeer zwakke voorwaarden en laten zien dat ze asymptotisch minimaal zijn met betrekking tot de index-klasse. Een simulatiestudie en toepassingen op echte data worden gepresenteerd aan het eind van deze hoofdstukken.

Zoals reeds opgemerkt, bestuderen we in het tweede deel van dit proefschrift, één- en twee-steekproevenproblemen voor identieke, onafhankelijk verdeelde reëelwaardige stochasten met behulp van P-P plots. Grafische methoden zoals P-P plots worden dikwijls toegepast in de niet-parametrische statistiek. In Hoofdstuk 5 presenteren we een beknopt literatuuroverzicht voor klassieke P-P plots en modificaties, terwijl we in Hoofdstuk 6 gegeneraliseerde P-P plots introduceren. Deze plots zijn gedefinieerd via indicering met intervallen en behouden, globaal gesproken, de eigenschappen van de klassieke P-P plots. Met behulp van deze gegeneraliseerde P-P plots, bestuderen we het één- en twee-steekproevenprobleem voor vaste en naburige alternatieven. We laten zien dat het gegeneraliseerde P-P plot proces, en dus ook de hiervan afgeleide toetsingsgrootheid, verdelingsvrij is onder de nulhypothese en dat de procedure consistent is (voor alle vaste alternatieven). In het geval van naburige alternatieven leiden we de limietverdeling van de toetsingsgrootheid af. Bovendien worden enkele numerieke resultaten over onderscheidingsvermogen gepresenteerd.

Acknowledgments

Some of the time I spent in Eindhoven as a PhD student was quite difficult, and without the support of the people around me, I would not have been able to accomplish this project.

I am very grateful to my supervisor, Dr. John Einmahl for his continuous support and guidance. He was never too busy to answer my questions and never got tired of carefully reading my manuscripts. Due to his efforts I became more careful about making quick, intuitive, statements. Next I would like to thank Prof. Paul van der Laan. He was always present in the background during the years of research. In addition to these two people, the third person who made possible for me to come to The Netherlands is Prof. Estate Khmaladze. I would like to thank him for what he has done for me.

I would also like to express my gratitude to Dr. Ingrid van Keilegom. Our reading seminars were most motivating for me. I am grateful to Dr. Alessandro Di Bucchianico for our collaboration, which was very educational.

I would like to thank my family, I missed their presence a lot. Short phone calls and e-mail messages were most valuable. I discovered lots of friends here, in Eindhoven. I thank them for their love and support.

Curriculum vitae

Nino Mushkudiani was born on 14 January 1973, in Sagaredjo, Georgia. From 1989 till 1994 she studied mathematics at Tbilisi State University, from which she graduated (with honor) in June 1994. In January 1995 Nino started her PhD studies at the A.M. Razmadze Mathematical Institute of the Georgian Academy of Sciences, which she dropped in August of the same year. 'Escaping' the economical crisis in by then independent Georgia, she started working as a PhD student at the department of mathematics of Eindhoven University of Technology in 1996. Here she worked in nonparametric statistics under supervision of Dr. J.H.J. Einmahl. Currently Nino is an assistant professor in the same department in Eindhoven University of Technology.