Small nonparametric tolerance regions for directional data

TU/e technische universiteit eindhoven

SPOR-Report 2000-06

Small nonparametric tolerance regions for directional data

N.A. Mushkudiani

Eindhoven, March 2000
The Netherlands

department of mathematics and computing science

Small nonparametric tolerance regions for directional data

Nino A. Mushkudiani

March 27, 2000

Department of Mathematics and Computing Science
Eindhoven University of Technology
5600 MB Eindhoven, The Netherlands

Abstract

We present a natural approach, based on minimum volume sets, for constructing nonparametric tolerance regions for directional data. The tolerance regions have desirable features like invariance and are asymptotically minimal under certain conditions. We establish the asymptotic correctness of our tolerance regions by using the theory of empirical processes and generalized quantiles. The results are obtained under minimal conditions. In case of circular data, the finite sample properties of the tolerance arcs are studied through simulations. The method is also applied to a real data example.

Mathematics Subject Classification (1991) : 62G15, 62H11, 62G30, 60F05.

Key words : nonparametric tolerance region, circular data, spherical data, empirical process, minimum volume set, asymptotic normality.

1 Introduction

Starting from the early forties many publications have appeared in the literature on tolerance intervals and regions (e.g., Wilks (1941), Wald (1943), Tukey (1947, 1948), Ackermann (1983)). They dealt with both parametric and nonparametric cases and considered two types of tolerance regions (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)). Classical tolerance intervals introduced by Wilks (1941) are intervals with order statistics as endpoints. 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 are constructed in Tukey (1947, 1948) for continuous and discontinuous distributions, respectively. For an i.i.d. sample $W_1, \ldots W_n$ in $\mathbb{R}^k$ using the ordering functions, divide $\mathbb{R}^k$ into disjoint random sets (the statistically equivalent blocks) $S_1, \ldots, S_{n+1}$, with coverages $U_1, \ldots, U_{n+1}$, $(U_i = P\{S_i\},\ i = 1, \ldots, n+1)$. It is shown in Tukey (1947) that

$$\mathbb{E}U_i = \frac{1}{n+1}, \quad i = 1, \ldots, n+1$$

and

$$\mathbb{P}\{\sum_{i=1}^{r} U_i < t\} = I_t(n-r+1, r),$$

where $I_t(n,m) = \frac{\Gamma(n+m)}{\Gamma(n)\Gamma(m)} \int_0^t x^{n-1}(1-x)^{m-1}dx$ is the incomplete beta function, with Γ denoting the gamma function. Then the t-content tolerance region $\mathcal{S}$ at confidence level $1-\alpha$ is composed of r blocks, such that $n-r+1$ blocks define the region $\overline{\mathcal{S}}$ outside the tolerance region and r is determined by the following equation

$$\mathbb{P}\{\sum_{i=1}^{n-r+1} U_i < 1-t\} = I_{1-t}(r, n-r+1) = 1-\alpha.$$

The topic of this paper is the construction of tolerance regions for directional data. Such data points occur in many applications in biology, geology, meteorology, geography, medicine and physics. Vast data examples obtained from these areas are given in Mardia (1972), Batschelet (1981), Fisher, Lewis and Embleton (1987), Fisher (1993), etc. Typical directional data sources are bird or animal orientation and navigation with homing, migration or other activity, wind and ocean directions, orientations of cross-beddings or fractures and fabric elements in deformed rocks, micro seismic and earthquake directions in a certain region, etc. 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 $\theta_1, \ldots, \theta_n$, $0 \le \theta_i < 2\pi$, $n \ge 1$ 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 a set of all points on the circle that lie between A and B taking counter clockwise direction and including 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 the 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.

From a statistical point of view, there is much arbitrariness in procedures based on statistically equivalent blocks, since they depend on auxiliary ordering functions. An alternative way of constructing tolerance regions is presented in Di Bucchianico, Einmahl and Mushkudiani (1998), where in contrast to the classical procedure, tolerance intervals are defined as the shortest intervals, that contain a certain number of observations. This idea naturally extends to higher dimensions by considering classes of sets (ellipsoids, hyperrectangles, convex sets) and defining the tolerance region as the minimum volume set from the chosen class that again contains a certain number of observations. The asymptotic behavior of these tolerance regions is established using empirical process theory and generalized quantiles. It is also shown that the presented tolerance regions are asymptotically minimal with respect to the chosen indexing class and have desirable invariance properties.

Based on minimum volume sets and the techniques presented in Di Bucchianico et al. (1998), here we propose a new way of constructing tolerance regions for circular and spherical data. The limiting behavior for these regions is established and it is shown that they are asymptotically minimal with respect to the indexing class.

The paper is organized as follows; Section 2 contains a formal definition of our tolerance regions on the class of so-called caps. In Section 3 the main results are presented and finally in Section 4 we construct a tolerance region for wind direction data and study finite sample properties through simulations.

2 The setup

In this and the next section we will assume that our data are spherical. 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.

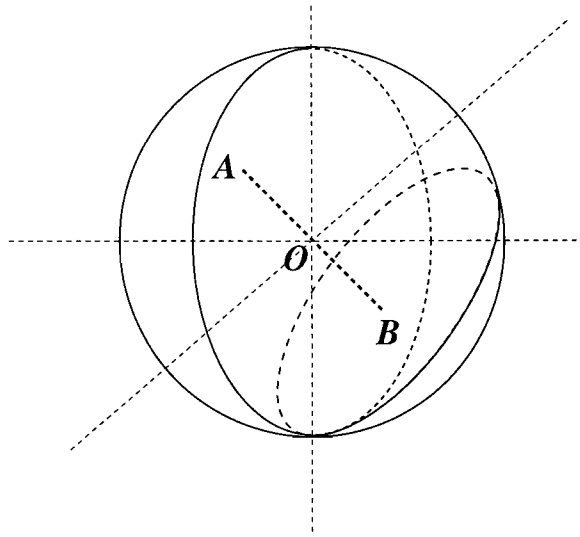

Figure 1: The cap with the center A and the boundary circle centered at B

Three dimensional directional observations 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 sphere $\mathcal{S}_{(0,1)}$ with center in O. The direction of OL can be identified with the point L'. Hence we assume the spherical data $X_1, \ldots, X_n$, $n \geq 1$, to be i.i.d. random vectors with values in $\partial\mathcal{S}_{(0,1)}$ (the surface of $\mathcal{S}_{(0,1)}$) 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 define the pseudo-metric d_0 on $\mathcal{B}$ by

$$d_0(B_1, B_2) = P(B_1 \triangle B_2),$$

where $B_1, B_2 \in \mathcal{B}$. Note that for any $B \in \mathcal{B}$, $P(B) = P(B \cap \partial\mathcal{S}_{(0,1)})$. Let P_n denote the empirical distribution:

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

where I_B is the indicator function of the set 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 $\partial\mathcal{S}_{(0,1)}$. The circle with center B, created by the intersection will be called the boundary circle (see Figure 2). The perpendicular line to the boundary circle at B goes through the cap 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 some technical inconveniences, from now on let $\mathfrak{C}$ be the class of caps with $0 < P(C) < 1$. One of the 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 $(\eta_C, \ell_C) \in \partial\mathcal{S}_{(0,1)} \otimes (0,2)$. 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 $(\eta^*, \ell^*) \in \partial\mathcal{S}_{(0,1)} \otimes [0,2]$. 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

$$\lim_{k\to\infty} V(C_{n_k} \triangle C^*) = 0 \quad \text{a.s.}, \tag{1}$$

where V denotes the area (Lebesgue measure) on $\partial\mathcal{S}_{(0,1)}$. Similar results for ellipsoids can be found in Nolan (1991).

Further consider the generalized empirical quantile and generalized quantile functions introduced in Einmahl and Mason (1992) based on the class $\mathfrak{C}$

$$U_n(t) = \inf_{C\in\mathfrak{C}} \{V(C) : \ P_n(C) \geq t\},$$

$$U(t) = \inf_{C\in\mathfrak{C}} \{V(C) : \ P(C) \geq t\}, \quad t \in (0,1);$$

set $U(t) = 0$ for $t \leq 0$, and $U(t) = \lim_{s\uparrow 1} U(s)$ for $t \geq 1$.

In general, when the generalized quantile functions are based on some class of sets $\mathcal{A}$ on $\mathbb{R}^d$, $d \geq 1$ one can define the minimum volume sets (MV-sets) as follows (see also Polonik (1997)). For any $t \in (0,1)$ call $A(t) \in \mathcal{A}$ an MV-set if $V(A(t)) = U(t)$. Similarly define the empirical MV-sets $A_n(t)$, $V(A_n(t)) = U_n(t)$. In a certain sense MV-sets are higher dimensional quantiles. If the choice of $\mathcal{A}$ is appropriate, using the MV-sets one can determine various properties of the underlying distribution. When for example all level sets of the underlying distribution are in $\mathcal{A}$ then the MV-sets are the level sets and can be approximated by the empirical MV-sets.

Let us now go back to our initial notations. Suppose P is absolutely continuous with respect to Lebesgue measure on $\partial\mathcal{S}_{(0,1)}$. Define the MV-sets based on the indexing class $\mathfrak{C}$ as follows. For any fixed $t \in (0,1)$ and $q \in \mathbb{R}$ 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 $V(C_{n,t,q}) = U_n(t_n)$. Set $C_{n,t} = C_{n,t,0}$.

Lemma 1 *Suppose $X_1, \ldots, X_n$, $n \geq 1$, are i.i.d. random vectors with values in $\partial\mathcal{S}_{(0,1)}$ from the common distribution P, that is absolutely continuous with respect to Lebesgue measure on $\partial\mathcal{S}_{(0,1)}$. 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 $\mathcal{X} = \{X_1, \ldots, X_n\}$.
Trivially $P_n(C_{\mathcal{X}}) = 1$. Let $\mathfrak{C}_{\mathcal{X}} \subset \mathfrak{C}$ be the class of all caps that contain $\mathcal{X}$. 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 $\mathcal{L}_{\mathcal{X}} := \{\ell_C : C \in \mathfrak{C}_{\mathcal{X}}\} \subset (0,2)$. Set $\ell^* := \inf \mathcal{L}_{\mathcal{X}}$. Further take a sequence $\{\ell_n\}_{n\geq 1}$ from $\mathcal{L}_{\mathcal{X}}$ 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 $(\eta^*, \ell^*) \in \partial\mathcal{S}_{(0,1)} \otimes (0,2)$, with $\ell^* = \inf \mathcal{L}_{\mathcal{X}}$. Then there exists $C^* \in \mathfrak{C}$ that corresponds to (η^*, ℓ^*) and a sequence $\{C_{n_k}\}_{k\geq 1}$ such that (1) holds. To complete the existence proof we have to show that $C^* \in \mathfrak{C}_{\mathcal{X}}$ or that $\mathcal{X} \in C^*$. 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 there can be at most three observations a.s. on any circle in $\partial\mathcal{S}_{(0,1)}$ and any two circles that pass through different sets of three observations will have different radii with probability one. By $\overline{\mathfrak{C}_{\mathcal{X}}}$ denote the class of sets that are obtained by taking convex hulls (in $\mathbb{R}^3$) of the elements of $\mathfrak{C}_{\mathcal{X}}$. It is easy to show that an MV-set from $\overline{\mathfrak{C}_{\mathcal{X}}}$ corresponds to $C_{\mathcal{X}}$. Construct the polyhedron $H_{\mathcal{X}}$ with n vertices from $\mathcal{X}$. Clearly each face of $H_{\mathcal{X}}$ is a triangle a.s.. It can be shown by induction that for $n \geq 4$, $H_{\mathcal{X}}$ will have $2n-4$ faces. Since $H_{\mathcal{X}}$ is the smallest convex set containing $\mathcal{X}$, each element of $\overline{\mathfrak{C}_{\mathcal{X}}}$ will contain $H_{\mathcal{X}}$. There can be two kinds of polyhedra $H_{\mathcal{X}}$, those that contain the origin and those that do not contain it. We will treat these cases separately.

I. $O \in H_{\mathcal{X}}$. Then an MV-set from $\overline{\mathfrak{C}_{\mathcal{X}}}$ will also contain the origin and will have the boundary circle with the biggest radius (in comparison to the elements of $\overline{\mathfrak{C}_{\mathcal{X}}}$). 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 $\overline{\mathfrak{C}_{\mathcal{X}}}$. 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. $O \notin H_{\mathcal{X}}$. 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 half-space that does not contain the origin. Therefore an MV-set from $\overline{\mathfrak{C}_{\mathcal{X}}}$ will not contain the origin either and hence will have the boundary circle with the smallest radius (in comparison to the elements of $\overline{\mathfrak{C}_{\mathcal{X}}}$ that do not contain the origin). 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 $\partial\mathcal{S}_{(0,1)}$ 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 $\mathfrak{C}_{\mathcal{X}}$, 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 $r \in (0,1)$

$$\mathbb{P}\{X_1 : R(C_1) = r \mid X_2, \ldots, X_n\} = 0,$$

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 the existence and a.s. uniqueness of the MV-cap $C_{n,t,q}$ can be proved: 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 $O \in H_{\mathcal{X}}$ and when $O \notin H_{\mathcal{X}}$.

I. Since $O \in H_{\mathcal{X}}$, 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 $X_{i_1}, X_{i_2}, X_{i_3}$. 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 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 $\mathcal{X}$.

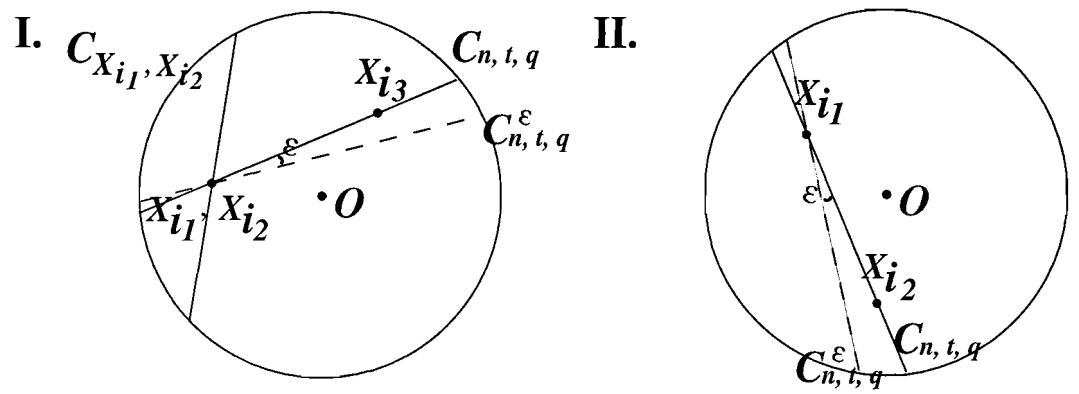

Figure 2: The cross section of $S_{(0,1)}$ cut on the plane: passing through the origin O and perpendicular to $\overline{X_{i_1}X_{i_2}}$ *(I)*; passing through the origin O, parallel to $\overline{X_{i_1}, X_{i_2}}$ and perpendicular to the boundary circle of $C_{n,t.q}$ *(II)*.

II. When $O \notin H_{\mathcal{X}}$, 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 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

$$t_0 + \frac{q}{\sqrt{n}} \leq P_n(C_{n,t,q}) < t_0 + \frac{q}{\sqrt{n}} + \frac{1}{n} \quad \text{a.s.}. \tag{2}$$

□

3 Main Results

In this section we use the settings 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 $\partial\mathcal{S}_{(0,1)}$ and that f is strictly positive on some connected open set $A \subset \mathcal{S}_{(0,1)}$ ($f \equiv 0$ on $\mathcal{S}_{(0,1)} \setminus A$).

Theorem 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 $q \in \mathbb{R}$*

$$\sqrt{n}(t_0 - P(C_{n,t_0,q})) + q \xrightarrow{d} Z\sqrt{t_0(1-t_0)} \quad (n \to \infty), \tag{3}$$

where Z is a standard normal random variable.

To prove Theorem 1 we will need the following result. Set $d(C_1, C_2) := V(C_1 \triangle C_2)$ to be the symmetric difference metric.

Lemma 2 *Under the assumptions of Theorem 1 we have with probability one that*

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

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

For proving Lemma 2 one does not need to make any crucial changes in the proof of the similar result in Di Bucchianico et al. (1998); however the parametrization from Section 2 could be used instead of the Blaschke Selection Principle.

Proof of Theorem 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 process α_n converges weakly (in the sense of Dudley (1978)) to a bounded, mean zero Gaussian process B_P indexed by $\mathfrak{C}$, since $\mathfrak{C}$ is a Vapnik-Chervonenkis (VC) class. The process B_P is uniformly continuous on $(\mathfrak{C}, d_0)$ and has covariance function $P(C_1 \cap C_2) - P(C_1)P(C_2)$, $C_1, C_2 \in \mathfrak{C}$.

By the Skorohod-Dudley-Wichura representation theorem (see e.g., Gänssler (1983, p. 82)), there exists a probability space $(\widetilde{\Omega}, \widetilde{\mathcal{F}}, \widetilde{P})$ carrying a version $\widetilde{B}_P$ of B_P and versions $\widetilde{\alpha}_n$ of α_n, for all $n \in \mathbb{N}$, such that

$$\sup_{C \in \mathfrak{C}} |\widetilde{\alpha}_n(C) - \widetilde{B}_P(C)| \to 0 \quad \text{a.s.} \quad n \to \infty. \tag{4}$$

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

$$\sup_{C\in\mathfrak{C}} |\sqrt{n}(P_n(C) - P(C)) - B_P(C)| \to 0 \quad \text{a.s.} \quad n \to \infty. \tag{5}$$

Then by the existence and a.s. uniqueness of the MV-cap $C_{n,t_0,q}$ in Lemma 1 we have that

$$\sqrt{n}(P_n(C_{n,t_0,q}) - P(C_{n,t_0,q})) - B_P(C_{n,t_0,q}) \to 0 \quad \text{a.s.} \quad n \to \infty. \tag{6}$$

Using (2) and (6) we obtain

$$\sqrt{n}(t_0 - P(C_{n,t_0,q})) + q - B_P(C_{n,t_0,q}) \to 0 \quad \text{a.s.} \quad n \to \infty. \tag{7}$$

By Lemma 2 and the continuity of B_P we will get that

$$B_P(C_{n,t_0,q}) \to B_P(C_{t_0}) \quad \text{a.s.} \quad n \to \infty. \tag{8}$$

Further it trivially follows from (7) and (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 are the main results of this paper, though they immediately follow from Theorem 1. Set q_α to be the $(1-\alpha)$-th quantile of the distribution of $Z\sqrt{t_0(1-t_0)}$. Then by Theorem 1, 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 2 below deals with the asymptotic behavior of guaranteed coverage tolerance regions C_{n,t_0,q_α}, while in Theorem 3 results for the mean coverage tolerance regions, C_{n,t_0} can be found.

Theorem 2 *Fix $\alpha \in (0,1)$, then under the conditions of Theorem 1 we have*

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

Proof By Theorem 1, for all $x \in I\!R$, we have

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

Hence, taking $x = q_\alpha$, we obtain

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

□

Theorem 3 *Under the conditions of Theorem 1*

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

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

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

Proof Let us first show that the sequence of random variables $\sqrt{n}(t_0 - P(C_{n,t_0}))$ is uniformly integrable. However, as

$$\begin{aligned} \left|\sqrt{n}(t_0 - P(C_{n,t_0}))\right| &\leq \left|\sqrt{n}(P_n(C_{n,t_0}) - P(C_{n,t_0}))\right| + \left|\sqrt{n}(t_0 - P_n(C_{n,t_0}))\right| \\ &\leq \sup_{C\in\mathfrak{C}} |\sqrt{n}(P_n(C) - P(C))| + 1/\sqrt{n} \text{ a.s.}, \end{aligned}$$

where we used (2) for $q = 0$, it suffices to show that

$$Y_n := \sup_{C\in\mathfrak{C}} |\sqrt{n}(P_n(C) - P(C))|$$

is uniformly integrable. Hence we have to derive that for any $\varepsilon > 0$ there exists a large enough such that

$$\sup_{n\geq 1} \mathbb{E}(Y_n I_{\{Y_n \geq a\}}) < \varepsilon.$$

Note that for any non-negative random variable Y it is true that

$$\mathbb{E}Y I_{[Y>a]} = \int_0^\infty \mathbb{P}\{Y I_{[Y>a]} > y\}dy = a\mathbb{P}\{Y > a\} + \int_a^\infty \mathbb{P}\{Y > y\}dy. \tag{9}$$

Furthermore, as $\mathfrak{C}$ is a VC class, from Alexander (1984, Theorem 2.11) we obtain that for $\lambda \geq 8$ and $K_1, K_2 \in (0, \infty)$

$$\mathbb{P}\{\sup_{C\in\mathfrak{C}} |\sqrt{n}(P_n(C) - P(C))| > \lambda\} \leq K_1\lambda^{K_2}\exp(-2\lambda^2) \leq \exp(-\lambda^2), \tag{10}$$

where the last inequality holds for λ large enough. Then from (9) it follows that for any $\varepsilon > 0$ there exists a large enough such that:

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

Hence indeed $\sqrt{n}(t_0 - P(C_{n,t_0}))$ is uniformly integrable.
For $q = 0$, Theorem 1 yields that

$$\sqrt{n}(t_0 - P(C_{n,t_0})) \xrightarrow{d} Z\sqrt{t_0(1-t_0)}, \ n \to \infty. \tag{11}$$

Thus as the left-hand side of (11) is uniformly integrable we obtain

$$\mathbb{E}\sqrt{n}(t_0 - P(C_{n,t_0})) \longrightarrow \mathbb{E}(Z\sqrt{t_0(1-t_0)}) = 0, \quad n \to \infty.$$

which is equivalent to the statement in the theorem. □

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 *Theorems 2 and 3 remain true, mutatis mutandis, for circular data and the class of arcs.*

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 3) is bimodal, however our 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 $t_n = t_0 + \frac{q_\alpha}{\sqrt{n}}$. Note that 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 $\lceil nt_n \rceil = 279$. Since our asymptotic results remain true if we change the number of observations in the tolerance region within the range $o(\sqrt{n})$ and in addition the boundary of the tolerance regions has probability zero, 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.

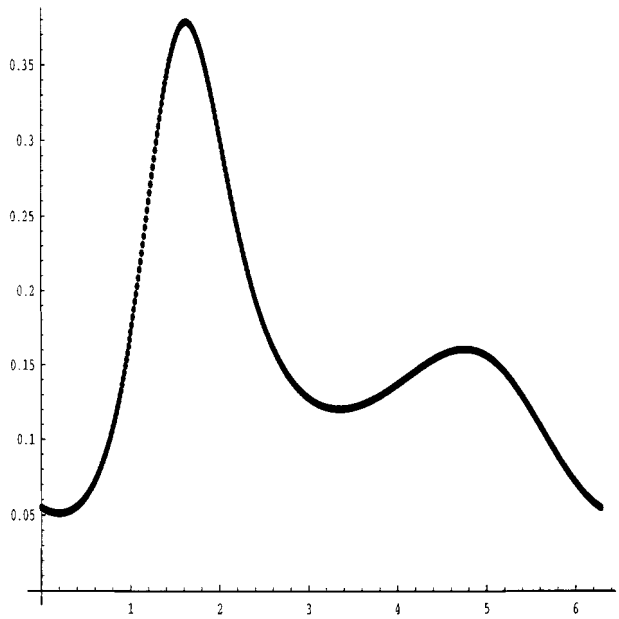

Figure 3: Linear plot of the bimodal circular distribution h.

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

- 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 $\varphi \in [0, 2\pi]$;
- $h(\varphi) = c\exp[k\cos(\varphi + \mu\cos\varphi)]$ with $c = 0.139236$, $k = 1$ and $\mu = \frac{5}{12}\pi$, where $\varphi \in [0, 2\pi]$.

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

Table 1: simulated confidence level for 90% guaranteed coverage tolerance arcs with confidence level 95% and simulated coverage for 90% mean coverage tolerance arcs.

In Table 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.

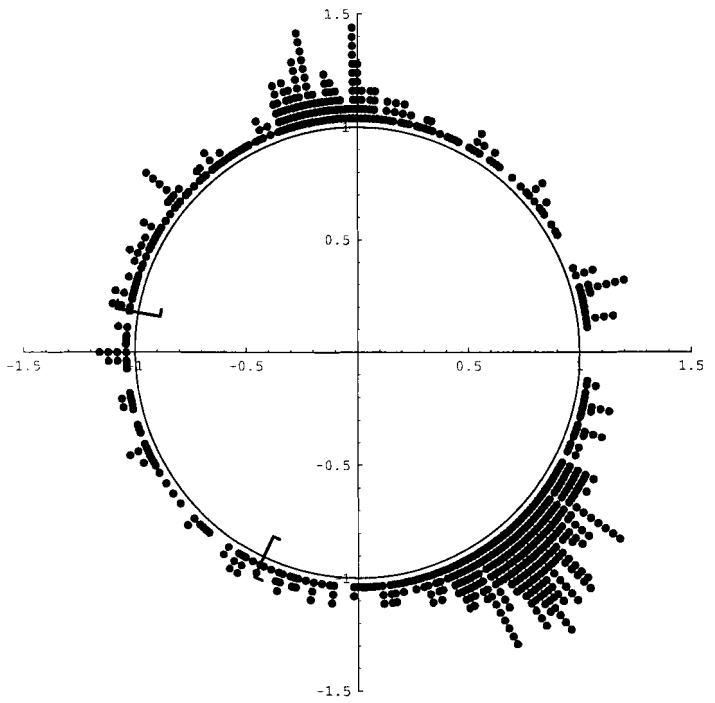

Figure 4: Tolerance arc for wind directions at Pt. Conception, CA.

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). 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 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.

References

Ackermann, H. (1983). Multivariate non-parametric 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.

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

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

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

Di Bucchianico, A., J. H. J. Einmahl, and N. A. Mushkudiani (1998). Small nonparametric tolerance regions. *Memorandum COSOR 98-16, Eindhoven University of Technology, The Netherlands.*

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

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

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.

Gänssler, P. (1983). *Empirical processes.* Institute of Mathematical Statistics, Hayward, Calif.

Guttman, I. (1970). *Statistical tolerance regions: classical and Bayesian.* Charles Griffin, London.

Mardia, K. V. (1972). *Statistics of directional data.* Academic Press, London-New York. Probability and Mathematical Statistics, No. 13.

Nolan, D. (1991). The excess-mass ellipsoid. *J. Multivariate Anal.* **39**, 348–371.

Polonik, W. (1997). Minimum volume sets and generalized quantile processes. *Stochastic Process. Appl.* **69**, 1–24.

Ruymgaart, F. H. (1989). Strong uniform convergence of density estimators on spheres. *J. Statist. Plann. Inference* **23**, 45–52.

Tukey, J. W. (1947). Nonparametric estimation, II. Statistical equivalent blocks and tolerance regions - the continuous case. *Ann. Math. Stat.* **18**, 529–539.

Tukey, J. W. (1948). Nonparametric estimation, III. Statistical equivalent blocks and multivariate tolerance regions - the discontinuous case. *Ann. Math. Stat.* **19**, 30–39.

Wald, A. (1943). An extension of Wilks' method for setting tolerance limits. *Ann. Math. Stat.* **14**, 45–55.

Wilks, S. S. (1941). Determination of sample sizes for setting tolerance limits. *Ann. Math. Stat.* **12**, 91–96.