

## Smallest nonparametric tolerance regions

**Citation for published version (APA):**

Di Bucchianico, A., Einmahl, J. H. J., & Mushkudiani, N. A. (2001). Smallest nonparametric tolerance regions. The Annals of Statistics, 29(5), 1320-1343. DOI: 10.1214/aos/1013203456

**DOI:**

[10.1214/aos/1013203456](https://doi.org/10.1214/aos/1013203456)

**Document status and date:**

Published: 01/01/2001

**Document Version:**

Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

**Please check the document version of this publication:**

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

**General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

[www.tue.nl/taverne](http://www.tue.nl/taverne)

**Take down policy**

If you believe that this document breaches copyright please contact us at:

[openaccess@tue.nl](mailto:openaccess@tue.nl)

providing details and we will investigate your claim.

## SMALLEST NONPARAMETRIC TOLERANCE REGIONS<sup>1</sup>

BY ALESSANDRO DI BUCCHIANICO, JOHN H. J. EINMAHL<sup>2</sup>  
AND NINO A. MUSHKUDIANI

*Eindhoven University of Technology and EURANDOM, Tilburg University  
and Eindhoven University of Technology*

We present a new, natural way to construct nonparametric multivariate tolerance regions. Unlike the classical nonparametric tolerance intervals, where the endpoints are determined by beforehand chosen order statistics, we take the shortest interval, that contains a certain number of observations. We extend this idea 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. The asymptotic behavior of our tolerance regions is derived using empirical process theory, in particular the concept of generalized quantiles. Finite sample properties of our tolerance regions are investigated through a simulation study. Real data examples are also presented.

**1. Introduction.** Several practical statistical problems require information on the distribution itself rather than on functionals of the distribution, like mean and variance. For example, in life testing of new products it is required that a certain percentage of sold products will not fail before the end of the warranty period. There are many other examples of this kind in various fields, such as reliability theory, medical statistics, chemistry, quality control, etc. (see, e.g., [3]). The statistical literature provides tolerance intervals and regions as a solution to these problems. Starting with [40], many papers on this topic have appeared. The monographs [3] and [19] provide thorough overviews of the literature, while extensive bibliographies can be found in [20] and [21]. Although there is a vast literature on the two types of tolerance regions (guaranteed coverage and mean coverage in the terminology of [3] or  $\beta$ -content and  $\beta$ -expectation in the terminology of [19]), statistics text books, both the mathematically and the engineering oriented ones, hardly deal with this topic explicitly. This is surprising since prediction regions are in fact  $\beta$ -expectation/mean coverage tolerance regions. We refer to the introduction of [8] for useful remarks on this issue, in particular on when to use which type of tolerance region. In case tolerance regions are mentioned in textbooks, the treatment is often confined to tolerance *intervals* for the normal distribution. In practice, however, one often encounters situations where the data are not normally distributed or univariate. In order to deal with the first problem,

---

Received August 1998; revised March 2001.

<sup>1</sup>Supported in part by European Union HCM Grant ERB CHRX-CT 940693.

<sup>2</sup>Research performed at Eindhoven University of Technology and EURANDOM.

AMS 2000 *subject classifications*. Primary 62G15, 62G20, 62G30, 60F05.

*Key words and phrases*. Nonparametric tolerance region, prediction region, empirical process, asymptotic normality, minimum volume set.

nonparametric tolerance intervals are used. The idea, which first appeared in the seminal paper [40], is to consider intervals with two order statistics as endpoints. It is important to note that it is decided *beforehand* which order statistics to take.

In the spirit of the shorth (see, e.g., [28], [18]), we propose a new approach to nonparametric tolerance intervals by taking the shortest interval that contains a certain number of order statistics. Surprisingly, the asymptotic theory concerning content (or coverage) is the same as for the classical procedure, although obviously by definition our intervals are not longer, and often much shorter. A problem with nonparametric techniques in higher dimensions is that there is no canonical ordering. In order to overcome this problem, essentially one-dimensional procedures such as statistically equivalent blocks were developed to construct multivariate tolerance regions (see [38], [34], [35], [16] and more recently [1]). From a statistical point of view, there is much arbitrariness in these procedures, since they depend on auxiliary ordering functions. Moreover, they are not necessarily asymptotically minimal (see [9]). Instead, one would like to have a genuine multivariate procedure, that is not based on ordering the data. In [9] a procedure is presented based on nonparametric density estimation, which yields asymptotically minimal tolerance regions. Our procedure is inspired by empirical process theory and extends to higher dimensions in a natural way. It avoids the choices that have to be made when estimating densities and it does not require any smoothness of the underlying density. On the other hand, we have to choose an indexing class to parametrize our empirical process, which however has the advantage that we can choose the shape of the tolerance region. We will show that our procedures are asymptotically *correct*, in contrast to those in [9] where only asymptotic conservatism is shown. Our tolerance regions are asymptotically minimal with respect to the indexing class and have desirable invariance properties. For their actual computation, which is non-trivial in higher dimensions, algorithms and software are available. A related paper dealing with directional data is [25]. In medical statistics, multivariate tolerance regions based on data from, for example, blood counts, can be used for screening of patients. In this paper, we will illustrate our approach by computing tolerance regions for bi- and trivariate observations of blood counts for Leukemia and AIDS. Multivariate tolerance regions can be applied in several other fields. For example, in statistical process control a multivariate approach to capability studies (which, if properly conducted, should be based on tolerance regions) is highly desirable, when various quality characteristics are taken into account.

This paper is organized as follows. In Section 2 we present the main results. In Section 3 we study the finite sample properties of our tolerance regions through simulations and apply the methods to real data examples. Section 4 contains the proofs of the results in Section 2.

**2. Main results.** In this section we present the asymptotic results for our tolerance regions. Let  $X_1, \dots, X_n$ ,  $n \geq 1$ , be i.i.d.  $\mathbb{R}^k$ -valued random vectors defined on a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  with a common probability distribu-

tion  $P$ , absolutely continuous with respect to Lebesgue measure, and corresponding distribution function  $F$ . Let  $\mathcal{B}$  be the  $\sigma$ -algebra of Borel sets on  $\mathbb{R}^k$  and define the pseudo-metric  $d_0$  on  $\mathcal{B}$  by

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

and similarly the related pseudo-metric  $d(B_1, B_2) := V(B_1 \Delta B_2)$ , where  $V$  denotes volume (Lebesgue measure). Denote by  $P_n$  the empirical distribution:

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

where  $I_B$  is the indicator function of the set  $B$ . 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 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:  $\sqrt{n}(P_n - P)$  converges weakly on  $\mathcal{A}$  (in the sense of [11]) to a bounded, mean zero Gaussian process  $B_P$ ; the process  $B_P$  is uniformly continuous on  $(\mathcal{A}, d_0)$  and has covariance function  $P(A_1 \cap A_2) - P(A_1)P(A_2)$ ,  $A_1, A_2 \in \mathcal{A}$ .*

(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  $n \geq 1$ ,*

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

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

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

*Then we have*

$$(2.1) \quad \sqrt{n}(t_0 - P(A_{n,t_0,C})) + C \xrightarrow{d} Z\sqrt{t_0(1-t_0)}, \quad n \rightarrow \infty,$$

*where  $Z$  is a standard normal random variable.*

The following theorems, which are corollaries to Theorem 1, are actually our main general results about tolerance regions. In fact, we will show that the sets  $A_{n,t_0,C}$ , for suitable  $C$ , are asymptotic tolerance regions. Theorem 2 gives the result for guaranteed coverage tolerance regions, whereas Theorem 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. Note that surprisingly the results are asymptotically distribution-free. The numbers  $t_0$  and  $1 - \alpha$  denote the (desired) coverage and confidence level, respectively.

**THEOREM 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 1 we have*

$$(2.2) \quad \lim_{n \rightarrow \infty} \mathbb{P}\{P(A_{n,t_0,C}) \geq t_0\} = 1 - \alpha.$$

**THEOREM 3.** *If the conditions of Theorem 1 hold and  $\sqrt{n}(t_0 - P(A_{n,t_0,0}))$  is uniformly integrable, then*

$$(2.3) \quad \mathbb{E}P(A_{n,t_0,0}) = t_0 + o\left(\frac{1}{\sqrt{n}}\right), \quad n \rightarrow \infty.$$

Note that  $\mathbb{E}P(A_{n,t_0,C}) \rightarrow t_0$ ,  $n \rightarrow \infty$ , for every  $C \in \mathbb{R}$ .

In the final 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 the paper. 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. The latter choice of the indexing class might seem unnecessary, as there are some interesting and deep geometrical results relating convex sets and ellipsoids; see, for example, [30]. However, we have observed that, for a large class of distributions the choice of an indexing class should be made with delicacy, as the tolerance regions defined here are highly sensitive to the number of points included (see Section 3).

In order to present the theorem in an unambiguous way we present some preliminaries, which guarantee that the conditions (C2) and (C3) of Theorem 1 are satisfied for the classes above. Let  $\mathcal{E}$  be the class of all closed ellipsoids  $A$

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. The existence and a.s. uniqueness of such an ellipsoid  $A_{n,t_0,C}$  was proved in [10]. There are between  $k+1$  and  $k(k+3)/2$  points on the boundary of  $A_{n,t_0,C}$  in dimension  $k$  (see, e.g., [33]) and hence,

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

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 at the end of Section 4). This result seems not to be present in the literature. It yields that

$$(2.4) \quad P_n(A_{n,t_0,C}) = \frac{1}{n} \left\lceil n \left( t_0 + \frac{C}{\sqrt{n}} \right) \right\rceil \quad \text{a.s.}$$

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 [10] 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 hyperplanes contain at most one observation, the equality in (2.4) 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  $\mathcal{X} = \{X_1, \dots, X_n\}$  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., [39], 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  $\{X_{i_1}, \dots, X_{i_\ell}\}$  and  $\{X_{j_1}, \dots, X_{j_k}\}$ ,  $3 \leq \ell, k \leq n$  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  $\{X_2, \dots, X_n\}$ , then we have to show that for any positive  $v$

$$(2.5) \quad \mathbb{P}\{X_1 : V(A_1) = v \mid X_2, \dots, X_n\} = 0,$$

where  $V(A_1)$  denotes the area of  $A_1$ . Since  $A_1$  is convex,  $X_1$  lies in the interior of the triangle  $X_{i_1}OX_{i_2}$  (see Figure 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  $5 \leq \ell \leq n$ , but a similar argument works for  $\ell = 3$  or  $\ell = 4$ .) Hence, we see that (2.5) holds. Finally, it is obvious that (2.4) holds for  $\mathcal{C}$ .

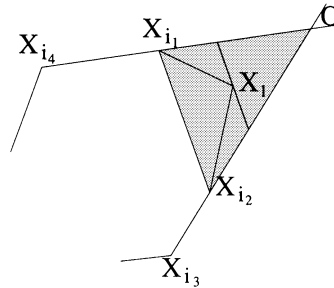


FIG. 1. Uniqueness of minimum area convex set.

**THEOREM 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  $\mathbb{R}^k \setminus \mathcal{S}$ , 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 cases (a) and (b) that (2.1), (2.2) and (2.3) hold.

If  $k = 2$  and, in addition,  $f$  is bounded, then (2.1), (2.2) and (2.3) also hold for case (c).

**REMARK 1.** Theorem 4 is valid under very mild conditions. In particular, there are no smoothness conditions on the density  $f$ , as in [9]. The uniqueness of  $A_{t_0}$ , however, is crucial for the results as stated. If it is not satisfied, then the results may be substantially different. On the other hand, uniqueness of  $A_{t_0}$  is a mild condition and holds for many (multimodal) distributions.

In [13] generalized quantile functions and processes (see Section 4) are introduced and studied. Generalized quantiles and related concepts play an important role in this paper and are further investigated in, for example, [27], where also a quantity closely related to the left-hand-side of (2.1) is studied. Mainly due to the fact that the results in both papers are uniform (in our  $t_0$ ), the conditions in both papers are much stronger than in the present paper.

Note that it is well-known (see, e.g., [12]) that for dimension 3 or higher there is no weak convergence of the empirical process indexed by closed convex sets, since the entropy of this class of sets is too large. (Actually the supremum of the absolute value of this empirical process tends to infinity, in probability, as  $n \rightarrow \infty$ .) This means that for this case Theorem 4, if true at all, cannot be proved with the methods presented in this paper.

**REMARK 2.** 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, for example, for case (a), that no tolerance ellipsoids can be found the volume of which converge to a number smaller than  $V(A_{t_0})$ . It is well-known that under weak additional conditions (see, e.g., [9]) 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 min-

imal (with respect to *all* Borel-measurable sets). Since such a minimal region contains just those points  $x$  where the density exceeds a certain level, “minimality” is a desirable property, because the minimal region contains those  $x$ 's which are “most likely” under  $f$ .

REMARK 3. It is rather easy to show that the tolerance regions of Theorem 4 have desirable invariance properties. For cases (a) and (c) the tolerance region  $A_{n,t_0,C}$  is affine equivariant, that is, 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.

REMARK 4. 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.

Note that a minimum volume set  $A_{n,t_0,C}$  consists of at most  $m$  ‘components’ and that some of these components may have an empty interior. Note as well that now a minimum volume set  $A_{n,t_0,C}$  need not be almost surely unique, hence (C2) is not satisfied, but we still have the second part of (C4) of Theorem 1, which yields “asymptotic uniqueness.” Since also (C1), the “existence part” of (C2), and (C3) are satisfied we see that Theorem 4 remains true when replacing the cases (a), (b) and (c) by (a'), (b') and (c'), respectively. This can be relevant for multimodal distributions. However, often the indexing classes of cases (a), (b) and (c) suffice, since in many (multimodal) situations the smallest closed set having probability  $t_0$ , is a “nice” connected set, because  $t_0$  is typically close to 1. Note that Remark 3, *mutatis mutandis*, holds true for the classes defined in (a'), (b') and (c'). For more details and a proof of the statements in this remark, see [24], Theorem 3.5.

REMARK 5. In order to argue about the novelty and advantages of the method presented in this paper, let us first consider the one-dimensional case. As we have mentioned already, classical nonparametric tolerance intervals have, concerning content, the same asymptotic behavior as the smallest nonparametric tolerance intervals. (This essentially follows from the weak convergence of the classical uniform quantile process to a Brownian bridge.) 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, for example, Pareto or exponential distributions the length of the classical nonparametric tolerance intervals may be much larger than that of our tolerance intervals (see Tables 1 and 2 below). In addition, since the tolerance intervals obtained by the new method are the shortest intervals that contain a



TABLE 1  
 90% guaranteed coverage tolerance intervals with confidence level 95%

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

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 arbitrarily chosen 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 are based on statistically equivalent blocks or on density estimation. The method based on statistically equivalent blocks (see [34] and [16] for a description), depends on arbitrary, auxiliary ordering functions and is essentially a one-dimensional procedure. The tolerance regions obtained by this method are exact, however they are in general not asymptotically minimal and have a shape that is difficult to work with and which very much depends on the chosen ordering functions. The other approach, that is based on density estimation [9], is more attractive and yields asymptotically minimal tolerance regions, but it is (very) conservative. Note as well that since the method is based on density estimation some regularity conditions have to be satisfied. Note however, that in principle the method based on density estimation can perform very well if a proper bandwidth is chosen, see the simulation results in Section 3.

In contrast to these methods, our 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 our tolerance regions are asymptotically minimal, hence best possible.

Another possible approach could be to construct tolerance regions using the concept of data depth; for discussions on the concept, theory and applications of depth in univariate and multivariate cases; see, for example, [22] and [23]. This would lead to regions based on the 'central' part of the data. For skewed distributions, these regions would be typically larger than the ones studied in this paper.

TABLE 2  
90% mean coverage tolerance intervals

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

**3. Simulation study and real data examples.** First we present results on the finite sample behavior of our tolerance regions through simulations. Each simulation consisted of 1000 replications. Note that the asymptotic behavior of our 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 tolerance *intervals*, the finite sample behavior is very sensitive to the actual number of order statistics used. For example, when  $n = 100$ , inclusion of 93, 95 or 97 order statistics leads for 90 % guaranteed coverage tolerance intervals with confidence level 67.9 %, 88.3 % and 97.6 %, respectively. Simulations showed a similar sensitivity for our 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  $(n + 1)/2$ . As mentioned above, we added 2 observations when constructing *our* tolerance intervals. Tables 1 and 2 contain our simulation results for guaranteed coverage and mean coverage tolerance intervals. These tables show very good behavior of our tolerance intervals. In particular, for the highly skewed distributions they perform much better with respect to length; for example, for the Pareto distribution the length is reduced with 50%. In general, we see that the asymptotic theory works well.

Table 3 gives simulation results for mean coverage rectangles with sides parallel to the coordinate axes. We included 4 extra observations in all cases, that is, we used 274 observations for  $n = 300$  and 904 for  $n = 1000$ . We simulated from the following distributions:

TABLE 3  
*Simulated coverages of 90% mean coverage tolerance rectangles*

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%

- bivariate standard normal with mean  $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$  and covariance matrix  $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
- bivariate half-normal with density  $f(x, y) = \frac{2}{\pi} e^{-\frac{1}{2}(x^2+y^2)}$ ,  $x, y \geq 0$
- bivariate Cauchy distribution with density  $f(x, y) = \frac{1}{2\pi} (1 + x^2 + y^2)^{-3/2}$
- bivariate exponential (1,1) distribution with density  $f(x, y) = e^{-(x+y)}$ ,  $x, y \geq 0$
- bivariate pyramid distribution with density  $f(x, y) = \frac{1}{8(|x|+|y|)} e^{-(|x|+|y|)}$ .

From this table, we again see that our tolerance regions perform well: the coverages are close to 90%, but slightly too low. This effect is caused by the minimum area property of our tolerance regions.

We have also performed simulations for tolerance hyperrectangles in  $\mathbb{R}^3$ , 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  $f(x, y, z) = \left(\frac{2}{\pi}\right)^{3/2} e^{-\frac{1}{2}(x^2+y^2+z^2)}$ ,  $x, y, z \geq 0$
- trivariate Cauchy distribution with density  $f(x, y, z) = \frac{1}{\pi^2} (1 + x^2 + y^2 + z^2)^{-2}$
- trivariate exponential distribution with density  $f(x, y, z) = e^{-(x+y+z)}$ ,  $x, y, z \geq 0$ .

In Table 4 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) by 95% seems to improve the asymptotics. 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).

Note that we did not perform simulations for the classical tolerance regions, based on statistically equivalent blocks, in higher dimensions, since the results

TABLE 4

*Simulated coverages of 95% mean coverage tolerance hyperrectangles*

Distribution	Simulated coverage
trivariate normal	93.6%
trivariate half-normal	94.1%
trivariate Cauchy	94.8%
trivariate exponential	94.2%

will depend very much on the choice of the ordering functions. Nevertheless, in general, the results would give the same picture as in the one-dimensional case.

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 2. 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., [5], [32], Section 23.2, and [36]) 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 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, that is, Theorem 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 5 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 [5]. Since  $\hat{P}_n$  is absolutely continuous we did not add the 2 observations as indicated above.

TABLE 5

*“Smoothed” 90% guaranteed coverage tolerance intervals with confidence level 95%*

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%	.415
	1000	94.3%	.400
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 5 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.

All simulations 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. Our simulations procedures for parallel hyperrectangles can easily be extended to dimensions 4 and higher.

As mentioned before medical statistics is one of the fields where tolerance regions are used. Here we illustrate our theory with applications to leukemia and AIDS diagnoses. 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 dimensions  $k = 2$  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 2, 3 and 4 below).

Comparing the tolerance regions in Figures 2, 3 and 4 with the one-dimensional “reference” or “normal” values for WBC, RBC, and PLT used in practice (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 1 and 2 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 2, 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

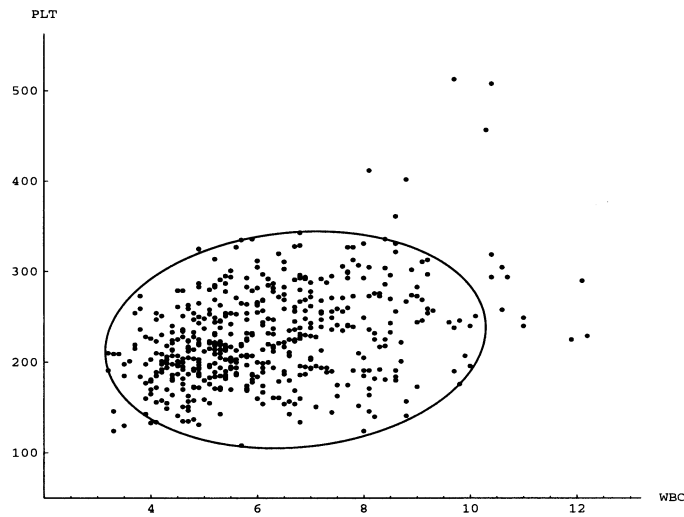


FIG. 2. 95% mean coverage tolerance ellipse.

above 100(!)] or RBC values around 3 or PLT values below 100. Clearly these values can be easily detected by the depicted tolerance regions.

The second application of our methods is on data from the Dutch ATHENA study on HIV/AIDS. HIV is a retrovirus that infects several kinds of cells in the body, the most important of which is a type of white blood cell called the CD4 lymphocyte. The CD4 cell is a major component of the human immune system that helps keep people free from many infections and some cancers; the so called CD8 cell is a very similar type of cell. HIV can effectively disable the body's immune system, and destroy its ability to fight diseases. The two two-dimensional data sets we use consist of CD4 and CD8 counts (both per microliter) and CD4 and  $^{10}\log$ -ged HIV counts (per milliliter), respectively; the sample sizes are 119 and 114. Each of the data points represents a deceased HIV-infected person who died of AIDS, meaning that his death was caused by a CDC-C event, that is, one of several diseases or cancers, including: Kaposi's Sarcoma (skin cancer), Tuberculosis, Toxoplasmosis, PCP, wasting syndrome (involuntary weight loss), Candidiasis, HIV dementia (memory impairment). The blood measurements were taken relatively shortly (at most 100 days) before death. All the patients died under Highly Active Anti-Retroviral Therapy (combination therapy). Typically, for AIDS patients CD4 and CD8 counts are relatively low and HIV counts are relatively high. In Figures 5 and 6 below we computed a 90% mean coverage tolerance ellipse and tolerance rectangle, respectively, based on the just described data. These tolerance regions can be very helpful in determining whether a deceased HIV-infected person did or did not die because of AIDS, since in practice it is not always possible to determine if a CDC-C event was present and caused death or not. A bivariate observation, from a deceased HIV-infected person under HAART therapy, outside the tolerance region indicates that the person does have counts that are

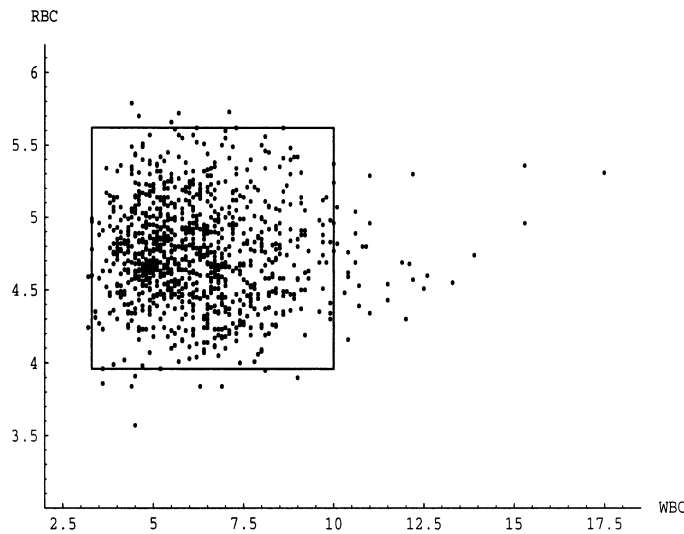


FIG. 3. 95% mean coverage tolerance rectangle.

atypical for a person who died of AIDS. Observe that most of the 'statistical' remarks made when discussing the tolerance regions for the Leukemia application are valid here too. Note also that an earlier (until 1993) definition of AIDS is having a CD4 count of 200 or less. It is very clear from the tolerance regions in Figures 5 and 6 that this (old) definition is very different from the definition being presently in force.

Finally we give some references on computing minimum volume ellipsoids, minimum volume hyperrectangles and minimum area planar convex sets (which we did not compute in this section). An algorithm for computing the minimum volume ellipsoid containing *all* data points is presented in [33]. Algorithms for computing *approximate* minimum area ellipsoids containing  $m$  ( $< n$ ) points are given in [26] and [29] and the *exact* algorithm we used for the minimum volume ellipse containing  $m$  ( $< n$ ) points was developed in [2]. The computer code of this algorithm was kindly placed at our disposal by the author; it also works in higher dimensions (up to 10). A description of the algorithm we used for computing minimum volume rectangles and hyperrectangles can be found in [24], Section 3.3.3. As we noted in Section 2, the minimum area planar convex set containing  $m$  ( $< n$ ) sample points is a polygon. *Exact* algorithms for computing such sets are described in [14] and [15].

**4. Proofs.** Here we present the proofs of the theorems of Section 2.

PROOF OF THEOREM 1. For each  $n \geq 1$ , define the empirical process indexed by  $\mathcal{A}$  to be

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

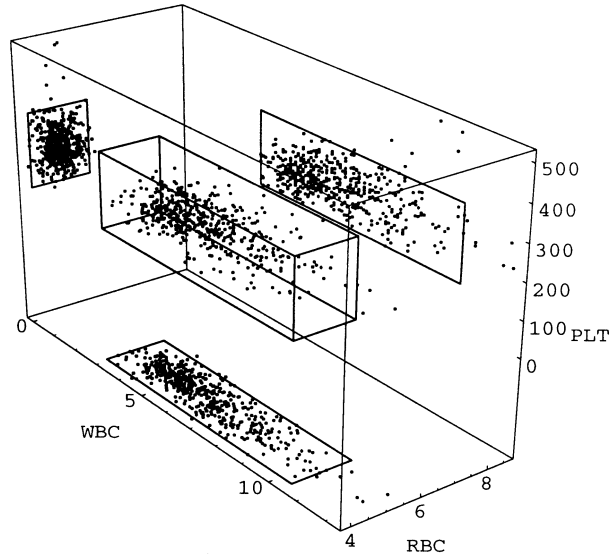


FIG. 4. 95% mean coverage tolerance hyperrectangle.

Because of (C1) and the Skorohod-Dudley-Wichura representation theorem (see, e.g., [17], page 82), there exists a probability space  $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$  carrying a version  $\tilde{B}_P$  of  $B_P$  and versions  $\tilde{\alpha}_n$  of  $\alpha_n$ , for all  $n \in \mathbb{N}$ , such that

$$(4.1) \quad \sup_{A \in \mathcal{A}} |\tilde{\alpha}_n(A) - \tilde{B}_P(A)| \rightarrow 0 \quad \text{a.s., } n \rightarrow \infty.$$

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

$$(4.2) \quad \sqrt{n}(P_n(A_{n,t_0,C}) - P(A_{n,t_0,C})) - B_P(A_{n,t_0,C}) \rightarrow 0 \quad \text{a.s., } n \rightarrow \infty.$$

Combining this with (C3) yields

$$(4.3) \quad \sqrt{n}(t_0 - P(A_{n,t_0,C})) + C - B_P(A_{n,t_0,C}) \rightarrow 0 \quad \text{a.s., } n \rightarrow \infty.$$

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

$$(4.4) \quad B_P(A_{n,t_0,C}) \xrightarrow{\mathbb{P}} B_P(A_{t_0}), \quad n \rightarrow \infty.$$

From (4.3) and (4.4) we now obtain that

$$\sqrt{n}(t_0 - P(A_{n,t_0,C})) + C \xrightarrow{\mathbb{P}} B_P(A_{t_0}), \quad n \rightarrow \infty.$$

Since

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

the proof is complete.  $\square$



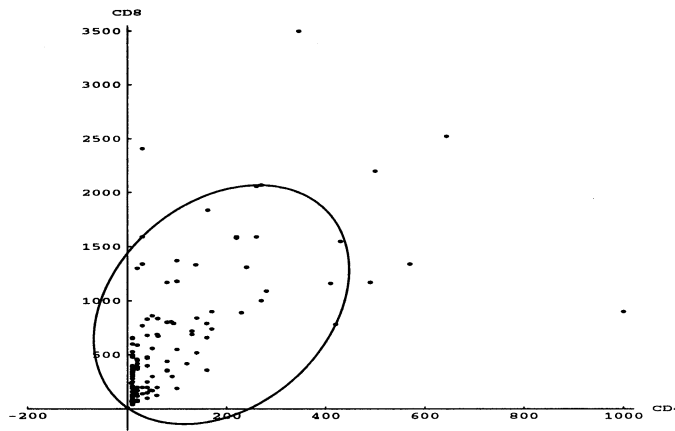


FIG. 5. 90% mean coverage tolerance ellipse.

PROOF OF THEOREM 2. By Theorem 1, for all  $x \in \mathbb{R}$ , we have

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

Hence, taking  $x = C$ , we obtain

$$\lim_{n \rightarrow \infty} \mathbb{P}\{P(A_{n,t_0,C}) \geq t_0\} = \mathbb{P}\{Z\sqrt{t_0(1-t_0)} \leq C\} = 1 - \alpha. \quad \square$$

PROOF OF THEOREM 3. Theorem 1 with  $C = 0$  yields

$$(4.5) \quad \sqrt{n}(t_0 - P(A_{n,t_0,0})) \xrightarrow{d} Z\sqrt{t_0(1-t_0)}, \quad n \rightarrow \infty.$$

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

$$\mathbb{E}\sqrt{n}(t_0 - P(A_{n,t_0,0})) \rightarrow \mathbb{E}(Z\sqrt{t_0(1-t_0)}) = 0, \quad n \rightarrow \infty,$$

which is the statement of the theorem.  $\square$

We next present two lemmas. Lemma 2 is crucial for the proof of Theorem 4, whereas Lemma 1 is needed for the proof of Lemma 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 4 we will show that this restriction can be removed. For Lemma 1 we need some more notation. Define for any class  $\mathcal{A} \subset \mathcal{B}$ :

$$\tilde{F}_n(y) = \sup_{A \in \mathcal{A}} \{P_n(A) : V(A) \leq y\},$$

$$\tilde{F}(y) = \sup_{A \in \mathcal{A}} \{P(A) : V(A) \leq y\}, \quad y > 0,$$

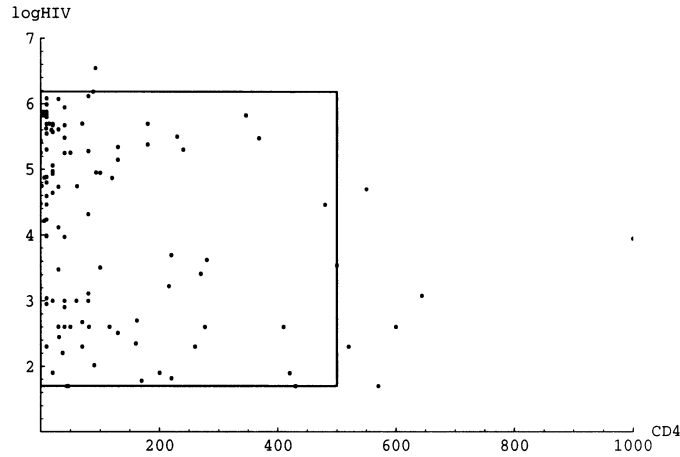


FIG. 6. 90% mean coverage tolerance rectangle.

and introduce as in [13] the generalized empirical quantile and quantile functions, based on  $P$ ,  $V$  and  $\mathcal{A}$  by

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

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

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

LEMMA 1. Under the assumptions of Theorem 4 we have for the cases (a), (b) and (c), that the functions  $U$  and  $\tilde{F}$  are inverses of each other. Hence,  $U$  is continuous on  $(0, 1)$ ,  $\tilde{F}$  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

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

and

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

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

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

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

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

Hence,

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

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

Note that absolute continuity of  $P$  also implies that

$$(4.6) \quad U(t) = \inf_{A \in \mathcal{A}} \{V(A) : P(A) = t\} \quad \text{for any } t \in (0, 1)$$

and

$$(4.7) \quad \tilde{F}(y) = \sup_{A \in \mathcal{A}} \{P(A) : V(A) = y\} \quad \text{for any } y \in \mathbb{R}^+.$$

It follows from (4.6) and (4.7) that  $U$  is the generalized inverse of  $\tilde{F}$ , that is,

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

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

LEMMA 2. *Under the assumptions of Theorem 4 we have for the cases (a), (b) and (c) that with probability one*

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

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

Note that an in-probability-version of this lemma, with  $k = 1$  and  $C = 0$ , can be found in [6], Corollary 1; see also [13] and [27].

PROOF OF LEMMA 2. Since for the cases (a) and (b)  $\mathcal{A}$  is a Vapnik-Chervonenkis (VC) class we have that (C1) of Theorem 1 holds. The (restricted) class of convex sets is not a VC class, but we still have (C1); see [7] and [11], page 918. Hence, we have (4.1) for all three cases. Since  $B_P$  is bounded, this yields

$$(4.8) \quad \sup_{A \in \mathcal{A}} |P_n(A) - P(A)| \rightarrow 0 \quad \text{a.s., } n \rightarrow \infty.$$

It now trivially follows from (4.8) and the definitions of  $\tilde{F}_n$  and  $\tilde{F}$  that

$$(4.9) \quad \sup_{y > 0} |\tilde{F}_n(y) - \tilde{F}(y)| \rightarrow 0 \quad \text{a.s., } n \rightarrow \infty.$$

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

$$(4.10) \quad \sup_{t \in (0, \ell]} \left| U\left(t + \frac{C}{\sqrt{n}}\right) - U(t) \right| \rightarrow 0, \quad n \rightarrow \infty.$$

We now want to prove that

$$(4.11) \quad \sup_{t \in (0, \ell]} |U_n(t) - U(t)| \rightarrow 0, \quad n \rightarrow \infty.$$

For any  $\varepsilon > 0$  we have from (4.9) that for  $n$  large enough

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

By Lemma 1,  $U$  is the generalized inverse of  $\tilde{F}$ . It is easy to see that  $U_n$  and  $\tilde{F}_n$  are generalized inverses. Hence, we obtain from the above inequalities that

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

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

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

which immediately yields (4.11). From (4.10) and (4.11) it follows that

$$(4.12) \quad \sup_{t \in (0, \ell]} \left| U_n \left( t + \frac{C}{\sqrt{n}} \right) - U(t) \right| \rightarrow 0 \quad \text{a.s., } n \rightarrow \infty.$$

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

- $A_{n,t_0,C}$ , the a.s. unique smallest element of  $\mathcal{A}$  with  $P_n(A_{n,t_0,C}) \geq t_0 + \frac{C}{\sqrt{n}}$  (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 (2.4),

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

and thus by (4.8)

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

From (4.12) we have

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

The sequence  $\{A_{n,t_0,C}\}_{n \geq 1}$  is uniformly bounded a.s., that is, for each  $\omega \in \Omega_0$ , with  $\mathbb{P}(\Omega_0) = 1$ , there exists a compact set  $\mathcal{M}_\omega$ , that contains all the  $A_{n,t_0,C}$ 's (for details see [24], pages 27–28). By the Blaschke Selection Principle (see, e.g., [39], Theorem 2.7.10), and the fact that the Hausdorff and the pseudo-metric  $d(A, B) = V(A \Delta B)$  are equivalent on the class of all compact convex subsets of  $\mathbb{R}^k$  with non-empty interior (see [31]), 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

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

Hence,  $V(A_{n_k,t_0,C}) \rightarrow V(A^*)$ , 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.).  $\square$

PROOF OF THEOREM 4. We will check the conditions (C1)-(C4) of Theorem 1. We first prove (2.1) and (2.2), for the cases (a), (b) and the restricted case (c). As noted in the proof of Lemma 2 we have that (C1) holds. In Section 2 it is shown that (C2) holds; (C3) follows from (2.4). The first part of C4) is an assumption of Theorem 4; Lemma 2 yields the second part of condition (C4). This completes the proof of (2.1) and (2.2) for these cases.

Now consider the unrestricted case (c). We will prove (2.1) and (2.2). 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  $V_\gamma$  between two parallel lines in  $\mathbb{R}^2$  at distance  $\gamma$  is said to be a  $\gamma$ -strip. Note that for a probability measure  $P$  with density  $f$  we have that

$$\limsup_{\gamma \rightarrow 0} \sup_{V_\gamma} P(V_\gamma) = 0,$$

where each supremum runs over all  $\gamma$ -strips. Therefore there exists a  $\gamma_0 > 0$  satisfying the inequality

$$(4.13) \quad \sup_{V_{\gamma_0}} P(V_{\gamma_0}) \leq \frac{1}{2}t_0,$$

where the supremum runs over all  $\gamma_0$ -strips. Now choose  $B$  to be a circle with the same centre as  $B_{t_0}$ , but with radius  $R > \frac{\delta}{\gamma_0}U(t_0) + r$ , where  $\gamma_0$  satisfies (4.13).

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

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

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  $B_{t_0}^c$  as well as an element of  $B^c$  eventually. Because the  $\gamma_0$ -strips form a VC class, we have that

$$\lim_{n \rightarrow \infty} \sup_{V_{\gamma_0}} P_n(V_{\gamma_0}) \leq \frac{1}{2}t_0 \quad \text{a.s.}$$

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 (2.1) and hence (2.2).

Finally we prove (2.3) 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 (2.4) that

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

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

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

Hence,

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

Moreover, for the cases (a) and (b) (as then  $\mathcal{A}$  is a VC class), using Theorem 2.11 of [4], we have for  $\lambda \geq 8$  and  $C_1, C_2 \in (0, \infty)$  that

$$(4.14) \quad \mathbb{P}\left\{\sup_{A \in \mathcal{A}} |\sqrt{n}(P_n(A) - P(A))| > \lambda\right\} \leq C_1 \lambda^{C_2} \exp(-2\lambda^2).$$

For large enough  $\lambda$ , the right-hand side of (4.14) is less than  $\exp(-\lambda^2)$ . Let  $\varepsilon > 0$ . Then for  $a$  large enough:

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

In case (c), using Corollary 2.4 and Example 3 (page 1045 of [4]) with  $\psi = \psi_3$ , we obtain the uniform integrability similarly as above; see also [37], page 2134.  $\square$

Recall the notation of Section 2, in particular let  $X_1, \dots, X_n$  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  $m \in \{k + 1, \dots, n\}$  (data) points.

LEMMA 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  $t$  boundary points is equal to  $E_1$ ; see [33]. 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  $E_1$  by  $Y_1$ . 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 are 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 by  $Y_0$  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  $C_\lambda$  be the convex hull of  $B \cup \{Y_\lambda\}$ . Note that for  $\lambda < \lambda'$  we have that  $C_\lambda \subset C_{\lambda'}$ . Let  $E_\lambda$  be the smallest ellipsoid containing  $C_\lambda$ . So  $V(E_\lambda) \leq V(E_{\lambda'})$  for  $\lambda \leq \lambda'$ .

It follows from the Blaschke Selection Principle that there exists a sequence  $\{\lambda_j\}_{j \in \mathbb{N}}$ ,  $0 < \lambda_j < 1$ , converging to 1 and such that

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

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 both  $E^*$  and  $E_1$  contain  $C_1$ . But, with probability 1,  $E_1$  is unique, so  $E^* = E_1$  and thus

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

So there exists a large  $j$  (denote the corresponding  $\lambda_j$  by  $\eta$ ) such that  $E_\eta$  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  $Y_\eta$  is in the interior of  $E_\eta$ , so according to [33],  $E_\eta = E_0$  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  $E_\eta$  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.  $\square$

**Acknowledgments.** We are grateful to an Associate Editor and a referee for a careful reading of the manuscript and insightful remarks which led to substantial improvements of the paper. Thanks are also due to Jose Agulló for kindly making available to us the computer code for the minimum volume ellipsoid algorithm, to Bart Hartgers for implementing our rectangles algorithm in C++ and linking it to our Mathematica notebook, and to Marko Boon for help with Figures 2, 4 and 5. Finally, we want to express our gratitude to Blood bank de Meierij, Eindhoven, in particular to Harry von Hegedus, Paul van Noord, Anja Verheijden, and Anouk Wesselink, for generously providing us with medical expertise and the blood count data set and to Mark van de Wiel and Frank de Wolf for making accessible to us the data from the national Dutch ATHENA study that we used in the AIDS example.

## REFERENCES

- [1] ACKERMANN, H. (1983). Multivariate nonparametric tolerance regions: a new construction technique. *Biometrical J.* **25** 351–359.

- [2] AGULLÓ, J. (1996). Exact iterative computation of the multivariate minimum volume ellipsoid estimator with a branch and bound algorithm. In *COMPSTAT '96* (A. Prat, ed.) 175–180. Physica, Heidelberg.
- [3] AITCHISON, J., and DUNSMORE, I. R. (1975). *Statistical Prediction Analysis*. Cambridge Univ. Press.
- [4] ALEXANDER, K. S. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Probab.* **12** 1041–1067.
- [5] AZZALINI, A. (1981). A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika* **68** 326–328.
- [6] BEIRLANT, J. and EINMAHL, J. H. J. (1995). Asymptotic confidence intervals for the length of the shortest under random censoring. *Statist. Neerlandica* **49** 1–8.
- [7] 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.
- [8] CARROLL, R. J. and RUPPERT, D. (1991). Prediction and tolerance intervals with transformation and/or weighting. *Technometrics* **33** 197–210.
- [9] CHATTERJEE, S. K. and PATRA, N. K. (1980). Asymptotically minimal multivariate tolerance sets. *Calcutta Statist. Assoc. Bull.* **29** 73–93.
- [10] DAVIES, L. (1992). The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. *Ann. Statist.* **20** 1828–1843.
- [11] DUDLEY, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probab.* **6** 899–929.
- [12] 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.
- [13] EINMAHL, J. H. J. and MASON, D. M. (1992). Generalized quantile processes. *Ann. Statist.* **20** 1062–1078.
- [14] EPPSTEIN, D. (1992). New algorithms for minimum area  $k$ -gons. In *Proceedings of the Third Annual ACM-SIAM Symposium on Discrete Algorithms* 83–88. ACM, New York.
- [15] EPPSTEIN, D., OVERMARS, M. ROTE, G. and WOEGERING, G. (1992). Finding minimum area  $k$ -gons. *Discrete Comput. Geom.* **7** 45–58.
- [16] FRASER, D. A. S. (1953). Nonparametric tolerance regions. *Ann. Math. Statist.* **24** 44–55.
- [17] GAENSSLER, P. (1983). *Empirical Processes*. IMS, Hayward, CA.
- [18] GRÜBEL, R. (1988). The length of the shortest. *Ann. Statist.* **16** 619–628.
- [19] GUTTMAN, I. (1970). *Statistical Tolerance Regions: Classical and Bayesian*. Griffin, London.
- [20] JÍLEK, M. (1981). A bibliography of statistical tolerance regions. *Math. Operations. Statist. Ser. Statist.* **12** 441–456.
- [21] JÍLEK, M. and ACKERMANN, H. (1989). A bibliography of statistical tolerance regions. II. *Statistics* **20** 165–172.
- [22] LIU, R. Y. (1990). On a notion of data depth based on random simplices. *Ann. Statist.* **18** 405–414.
- [23] LIU, R. Y., PARELIUS, J. M. and SINGH, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference (with discussion). *Ann. Statist.* **27** 783–858.
- [24] MUSHKUDIANI, N. (2000). *Statistical applications of generalized quantiles: nonparametric tolerance regions and P-P plots*. Eindhoven Univ. Technology.
- [25] MUSHKUDIANI, N. (2002). Small nonparametric tolerance regions for directional data. *J. Statist. Plann. Inference* **100** 67–80.
- [26] NOLAN, D. (1991). The excess-mass ellipsoid. *J. Multivariate Anal.* **39** 348–371.
- [27] POLONIK, W. (1997). Minimum volume sets and generalized quantile processes. *Stochastic Process. Appl.* **69** 1–24.
- [28] ROUSSEEUW, P. and LEROY, A. (1988). A robust scale estimator based on the shortest half. *Statist. Neerlandica* **42** 103–116.
- [29] ROUSSEEUW, P. and VAN ZOMEREN, B. C. (1991). Robust distances: simulations and cutoff values. In *Directions in Robust Statistics and Diagnostics II* (W. Stahel and S. Weisberg, eds.) 195–203. Springer, New York.
- [30] SCHNEIDER, R. (1993). *Convex Bodies: the Brunn-Minkowski Theory*. Cambridge Univ. Press.



- [31] SHEPHARD, G. C. and WEBSTER, R. J. (1965). Metric sets of convex bodies. *Mathematika* **12** 73–88.
- [32] SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.
- [33] SILVERMAN, B. W. and TITTERINGTON, D. M. (1980). Minimum covering ellipses. *SIAM J. Sci. Statist. Comput.* **1** 401–409.
- [34] TUKEY, J. W. (1947). Nonparametric estimation, II. Statistical equivalent blocks and tolerance regions - the continuous case. *Ann. Math. Stat.* **18** 529–539.
- [35] TUKEY, J. W. (1948). Nonparametric estimation, III. Statistical equivalent blocks and multivariate tolerance regions - the discontinuous case. *Ann. Math. Stat.* **19** 30–39.
- [36] VAN DER VAART, A. W. (1994). Weak convergence of smoothed empirical processes. *Scand. J. Statist.* **4** 501–504.
- [37] VAN DER VAART, A. W. (1996). New Donsker classes. *Ann. Probab.* **24** 2128–2140.
- [38] WALD, A. (1943). An extension of Wilks' method for setting tolerance limits. *Ann. Math. Statist.* **14** 45–55.
- [39] WEBSTER, R. (1994). *Convexity*. Oxford Univ. Press.
- [40] WILKS, S. S. (1941). Determination of sample sizes for setting tolerance limits. *Ann. Math. Statist.* **12** 91–96.

A. DI BUCCHIANICO  
EURANDOM AND  
DEPARTMENT OF TECHNOLOGY MANAGEMENT  
EINDHOVEN UNIVERSITY OF TECHNOLOGY  
P.O. BOX 513  
5600 MB EINDHOVEN  
THE NETHERLANDS  
E-MAIL: A.d.Bucchianico@tm.tue.nl

J. H. J. EINMAHL  
DEPARTMENT OF ECONOMETRICS  
TILBERG UNIVERSITY  
P.O. BOX 90153  
5000 LE TILBURG  
THE NETHERLANDS  
E-MAIL: J.H.J.Einmahl@kub.nl

N. A. MUSHKUDIANI  
DEPARTMENT OF MATHEMATICS  
AND COMPUTING SCIENCE  
EINDHOVEN UNIVERSITY OF TECHNOLOGY  
P.O. BOX 513  
5600 MB EINDHOVEN  
THE NETHERLANDS  
E-MAIL: N.A.Mushkudiani@tue.nl