Multivariate Mode Hunting: Data Analytic Tools with Measures of Significance

Prabir Burman and Wolfgang Polonik
Department of Statistics
University of California at Davis
Davis, CA 95616, USA

Abstract

Multivariate mode hunting is of increasing practical importance. Only a few such methods exist, however, and there usually is a trade off between practical feasibility and theoretical justification. In this paper we attempt to do both. We propose a method for locating isolated modes (or better, modal regions) in a multivariate data set without pre-specifying their total number. Information on significance of the findings is provided by means of formal testing for the presence of antimodes. Critical values of the tests are derived from large sample considerations. The method is designed to be computationally feasible in moderate dimensions, and it is complemented by diagnostic plots. Since the null-hypothesis under consideration is highly composite the proposed tests involve calibration in order to assure a correct (asymptotic) level. Our methods are illustrated by application to real data sets.

Keywords: Brownian Bridge, distribution free, modality, nearest neighbor method, testing for antimodes, VC-classes.

AMS 2000 subject classification: Primary: 62G99, 62H99 ; secondary: 62G20

Supported by NSF-grant 0079430 (W.P.)

1 Introduction

This paper discusses a computationally feasible statistical tool for nonparametric mode hunting in a multivariate data set. More precisely, we are attempting to find modal regions or clusters without specifying their total number. The methodology presented here is a mixture of data analytic methods and formal statistical tests. Several (local) tests are performed in order to analyze one data set, and we do not present a theory about the overall significance of the final outcome. Still, the local (conditional) testing provides some information about significance of the findings, and the presented theory can be used to derive large sample critical values for those testing procedures. In addition, our formal tests are accompanied by diagnostic plots.

A clustering of data might contain information about unusual or interesting phenomena. Therefore finding clusters in a data set is important in many fields like astronomy, bioinformatics, climatology (see example below), etc. One way to formalize the concept of clustering is to introduce a probabilistic model and to define a cluster as a modal region of the underlying probability density function (Hartigan, 1975). From this point of view finding clusters means finding modal regions. For instance, Jiang (2006) explains how connected components of level sets (at a specific level) correspond to galaxy clusters, and essentially proposes to use a plug-in method by Cuevas et al. (2000) (based on kernel density estimation) to estimate these clusters. (See below for more references to level set methods.) Friedman and Fisher (1999) propose a mode-hunting algorithm PRIM which is meant to find locations of modal regions in high dimensions (see also Polonik and Wang 2006). However, these nonparametric statistical methods lack some theoretical foundation.

The problem of providing quantifications of significance of findings regarding modal regions is not easy. In particular this holds for multivariate situations where the geometry or configuration of modes or modal regions can be very complex. Usually a compromise has to be found between practical feasibility and theoretical justification, and this is also reflected in our method for finding modal regions whose underlying idea is described next.

We assume the data to be sampled form a continuous underlying distribution with isolated modes. The basic idea underlying our method is to first employ a (fast) algorithmic method to find potential candidates representing distinct modal regions (described in more detail below). Then pairwise statistical tests are performed to determine whether our candidates really represent distinct modal regions, and this is the crucial step. The idea behind these tests can be described as follows.

Let x, y be two given candidates modes. Let $x_\alpha = \alpha x + (1-\alpha)y$, $0 \le \alpha \le 1$, denote the point on the line connecting x and y. If x and y represent two distinct modal regions then $f(x_\alpha) < \max(f(x), f(y))$ for at least some values of α. In other words, the intuition is that two modal regions are regarded as distinct if there is an antimode present on the line connecting the two candidates. Equivalently,

$$SB(\alpha) := -\log f(x_\alpha) + \min\{\log f(x), \log f(y)\} > 0 \quad \text{for some } \alpha. \tag{1}$$

Thus an appropriate hypothesis for testing whether x and y represent two different modes is given by

$$H_0: \quad SB(\alpha) \le 0 \ \text{ for all } \alpha \in [0,1]. \tag{2}$$

Now let $X_1, \ldots, X_n$ be an i.i.d. sample of p-dimensional random variables from a continuous distribution F. For $z \in \mathbf{R}^p$ let $\widehat{d}_n(z)$ denote the distance to the k_1-

nearest neighbor to z within the sample using some distance to be specified, like Euclidean distance or city-block distance. Define

$$\widehat{SB}_n(\alpha) = p\,[\log \widehat{d}_n(x_\alpha) - \max\{\log \widehat{d}_n(x), \log \widehat{d}_n(y)\}]. \tag{3}$$

It can be expected (and it is show below) that $\widehat{SB}_n(\alpha)$ converges in an appropriate sense to $SB(\alpha)$. Hence it appears to be reasonable to test H_0 by checking whether $\widehat{SB}_n(\cdot)$ shows a significantly overshoot over zero, or expressed more casually, whether this function shows a 'significant bump'. We propose two types of testing procedures:

(i) Pointwise testing, i.e. rather than testing H_0 we test $H_{0,\alpha} : SB(\alpha) \leq 0$ for some fixed values of α by using $\widehat{SB}_n(\alpha)$ as a test statistic. We reject $H_{0,\alpha}$ at the 95%-level iff $\widehat{SB}_n(\alpha) \geq \sqrt{\frac{2}{k_1}}\,\Phi^{-1}(0.95)$, with Φ denoting the standard normal cdf. The plot of the function $\alpha \to \sqrt{\frac{k_1}{2}}\,\widehat{SB}_n(\alpha)$ is called an SB-plot (cf. Theorem 4.1 for a motivation of the normalizing factor.).

(ii) 'Global' testing, where we test H_0 by essentially testing whether $SB(\cdot)$ is constant on $I := \{\alpha : SB(\alpha) \geq 0\}$. Here an estimator $\widehat{I}_n$ is used, and this estimation of I in effect provides a calibration. Critical values for T_n can be easily found using Theorem 4.2.

For deriving the critical values of the tests in (i) and (ii), the candidates for the modes are regarded as given.

An important question is how to find 'good' candidates for modes which will then be tested as indicated above. We will provide an algorithm designed to return a 'small' number of candidates in order to reduce the number of pairwise testing procedures necessary. Also the pairwise testing procedure will be conducted in a certain iterative fashion to reduce the number of tests to be performed. Details will be spelled out below.

We conclude this introduction with some more discussion. One-dimensional cases usually are less complex than multivariate cases, and a lot of work on investigating modality has been done for one-dimensional data sets (e.g. Haldane (1952), Dalenius (1965), Venter (1967), Lientz (1970), Good and Gaskins (1980), Silverman (1981), Hartigan and Hartigan (1985), Wong (1985), Wong and Schaack (1985), Comparini and Gori (1986), Müller (1989), Müller and Sawitzki (1991), Roeder (1994), Cheng and Hall (1999), Hall and Cheng (1999), Bai and Huang (1999), Walther (2001), Bai et al. (2003)).

Much less work has been done in the area of exploring multidimensional modality, although many of the interesting practical applications in fact are multivariate. One approach of course is to reduce the multidimensional problem to a one-dimensional through some kind of projection idea, and then to apply a univariate method. It is well known, however, that it is far from trivial to find an appropriate dimension reduction without loosing significant information. While dimension reduction seems inevitable in some problems, it appears to be plausible that a reduction to a p-dimensional subspace with $p > 1$ keeps more information than with $p = 1$. Hence, methods which work for moderate dimensions can be quite useful. Only a few of such nonparametric procedures (tests) for multivariate modality have been proposed so far; e.g. Hartigan (1988), Hartigan and Mohatny (1992), Rozal and Hartigan (1994), Minotte (1997), Polonik (1995, 1997), Friedman and Fisher (1999).

Investigating modality via estimating level sets has already been mentioned above as another way to investigate modality. Further work in this area include Tsybakov (1997), Walther (1997), Molchanov (1998), Cuevas et al. (2004), Gayraud and Rousseau (2002), Polonik and Wang (2004).

The approach for mode hunting presented here also is designed to work for moderate dimensions, and it bears some similarities to Minotte (1997) in the sense that *local* testing is used in order to determine if a candidate for a mode actually is a mode or if its presence can be explained by random fluctuation only. The general methodology is different, however. We also would like to point out that Minotte's procedure is for onedimensional situations only. In the following section we explain our method and present a motivating real data example. Simulations studies and some applications to real data sets are presented in Section 3. Section 4 contains some theoretical (large sample) results. Final remarks are given in Section 5. Proofs of main theoretical results can be found in Section 6, and Section 7 contains some interesting miscellaneous results which also are used in the proofs of the main results.

2 The method and real data applications.

Our method can be described as consisting of the three steps:

(I) Selection of initial modal candidates by using an iterative nearest neighbor methods.

(II) Thinning the modal candidates from step (I) by using local parametric tests using a multivariate normal model near a mode.

(III) Deciding whether the candidates from steps (I) and (II) represent *distinct* modal regions via testing for the presence of antimodes along the line segment joining two candidates (cf. introduction). Here the candidates are regarded as fixed.

The use of nearest neighbor instead of a kernel method in step (I) is motivated by the fact that in high density regions, nearest neighbor method yield better estimate of the density than the kernel method (Burman and Nolan, 1992).

Before providing more details, our method is illustrated via on application to a two-dimensional data set which has been used by Corti et al. (1999). The raw data consists of measurements of monthly means of 500-hPa geopotential heights taken on a regular grid with 2.5-degree spacing for the months November to April over the years 1949 to 1994. The two-dimensional data that was finally used for the analysis was obtained by projecting the non-detrended data onto the reduced phase-space spanned by the first two principal components (taken over space) of the detrended data. For further details of the analysis we refer to Corti et al. (1999).

The scatter plot of this data (Fig. 1) clearly shows the problems of a modal

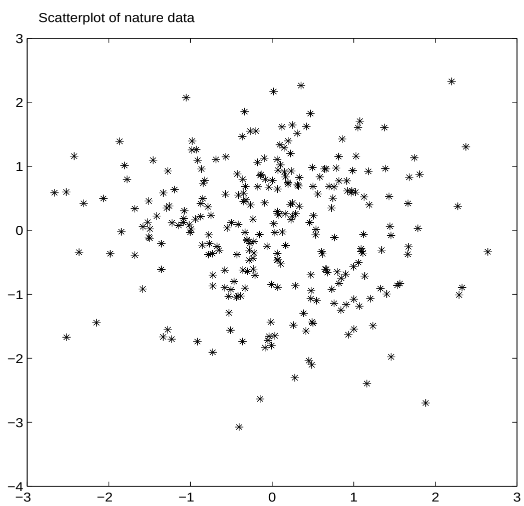

Fig. 1: Climatology data used by Corti et al. 1999

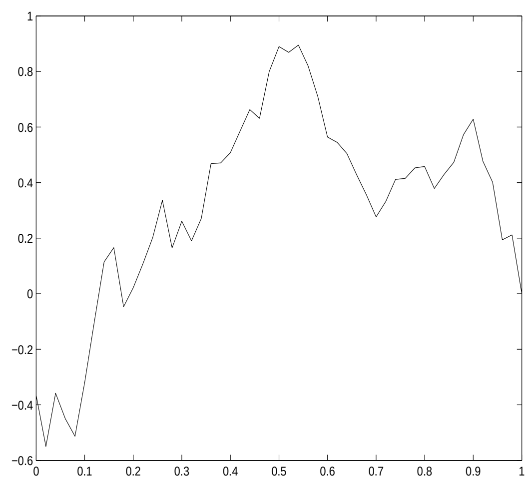

Fig. 2: Most significant SB-plot based on the climatology data. Since the plotted function does not exceed 1.645 (the 95% quantile of the standard normal) there is strong evidence that these candidates belong to the same modal region.

analysis, and supports the necessity for the development of more formal statistical tools that at least provide some indication about significance of the findings. (The data itself is not shown in this *Nature* article.) The existence of several modal regions is crucial for the climatological theory presented in Corti et al. To find such modal regions in their data, these authors used a standard kernel density estimator with a normal kernel. It is claimed that the bandwidth was chosen such that the four modes that can be seen in their estimate are statistically significant. We are not aware of a statistical method that provides such information.

Steps (I) and (II) lead to five potentially different candidates. However, step (III) indicates that they all correspond to the same modal region. Fig. 2 shows the SB-plot (cf. Introduction) corresponding to the two most different candidates (cf. Section 3.2). However, the graph does not show a significant bump, i.e. it does not exceed a threshold given by the $(1-\alpha)$-quantile of the standard normal distribution corresponding to the chosen significance level α, e.g. 1.28 is the approximate threshold at significance 0.1. Hence, the data do not seem to support the hypothesis of several different modes. (Of course this statement is by no means meant to implicitly suggest that the theory presented in Corti et al. regarding the global climate does not hold. We are only saying that it is difficult to claim statistical significance based on their data set.)

In the following we explain the three steps of our procedure in more detail. It is important to point out that we assume the observations $X_1, ..., X_n$ to be pre-standardized, so that we ca assume their mean to be 0 and the variance-covariance matrix to be identity. Intuitively, if the components of X are strongly related then the contours of the probability distribution behave like paraboloids and not like spheres. Consequently, a distance such as the Euclidean or city-block distance may not be appropriate.

Step (I): Selection of initial modal candidates via nearest neighbors. Let k_1 and k_2 be two integers such that $k_2 \leq k_1 < n$. Discussion on the actual choice of values of k_1 and k_2 is postponed to the end of this subsection. Recall that $\widehat{d}_n(x)$ denotes the distance of a point x in $\mathbf{R}^p$ to its k_1^{th} nearest neighbor among the data $\{X_1, ..., X_n\}$. The method presented here repeatedly employs two substeps: a) searching for a modal candidate, and b) elimination of its neighbors. First calculate $\widehat{d}_n(X_j)$, $j = 1, ..., n$. Find the first modal candidate as $W_1 = \text{argmin}_{X_j, j \in \{1,...,n\}} \widehat{d}_n(X_j)$. Next eliminate all those data points which are k_2 nearest neighbors of W_1. Let the remaining data set denoted by D_1. The second modal candidate is then found as $W_2 = \text{argmin}_{X_j \in D_1} \widehat{d}_n(X_j)$. The next elimination substep finds $D_2 \subset D_1$ by removing all the data points which are k_2-nearest neighbors (in the complete data set) of either W_1 or W_2. Then $W_3 = \text{argmin}_{X_j \in D_2} \widehat{d}_n(X_j)$ is our next modal candidate, and so on. The process is continued till no candidate mode can be found.

Step (II): Local parametric tests on initial modal candidates. The class of initial candidates from step (I) by construction is likely to contain candidates which clearly do not lie in modal regions. We will eliminate those which fail to pass this screening test of modality. For each modal candidate W_i we take the k_2 nearest neighbors X_j, $j = 1, ..., k_2$, among the data $X_1, ..., X_n$. Under some smoothness assumptions one can model the distribution locally around a mode as a multivariate normal. In order for W_i to be a modal candidate the mean of this distribution should be equal to W_i. We carry out a Hotelling's test for this at a 0.01 level of significance with this null hypothesis. Thus we thin out the list of modal candidates obtained from step (I). Each candidate is subjected to the test in order to decide if it should be eliminated from the candidate class.

We would like to acknowledge the somewhat heuristic nature of steps (I) and (II). However, the sole purpose of these steps is to reduce the number of pairwise tests to be performed in step (III).

Step (III): Testing and graphical tools for the presence of antimodes. In order to check whether the candidates obtained from steps (I) and (II) really represent different modal regions, we propose to repeatedly perform a crucial local testing procedure which is supported by a diagnostic plot (see Fig. 2 ff.). Two different types of testing procedures with varying degree of complexity are proposed. Given one of these tests we proceed as follows.

Let $W_1, ..., W_m$ be the modal candidates after having gone through steps (I) and (II). Note that by construction $\widehat{d}_n(W_1) \leq \cdots \leq \widehat{d}_n(W_m)$. Let $i = 1$.

Substep (III.1): We test to see whether W_i and W_j, $j = i+1, ..., m$, belong to the same modal regions. If the test indicates that W_i and W_j belong to the same modal region, we remove W_j as a modal candidate. This results in a potentially smaller set of candidates $W_1, \ldots, W_{m_1}$, say, with $\widehat{d}_n(W_1) \leq \cdots \leq \widehat{d}_n(W_{m_1})$, $m_1 \leq m$.

Substep (III.2): Set $i = i + 1$ and repeat substep (III.1) with $W_1, \ldots, W_{m_1}$. These two substeps are iterated till we arrive at distinct modal regions.

Two tests for step (III) are considered in the paper. One is based on pointwise simple z-tests and has already been introduced in the introduction. The other more global "line" test described next.

Global line testing. Our more global test is for $H_0 : \sup_{\alpha \in [0,1]} SB(\alpha) \leq 0$. In fact, we not only consider this H_0 but also the closely related null hypothesis will come into play:

$$H_{0n} : \quad \sup_{\alpha \in [0,1]} SB_n(\alpha) \leq 0., \tag{4}$$

where

$$SB_n(\alpha) := p \log d_n(x_\alpha) - p \max\{\log d_n(x), \log d_n(y)\}, \tag{5}$$

Here $d_n(x_\alpha) := \operatorname{argmin}\{s : F(B(x_\alpha, s)) \geq \frac{k_1}{n}\}$ with k_1 as above, and $B(x_\alpha, s)$ denoting a ball of radius s with midpoint x_α. Our test statistic T_n for these testing problems involves an estimate $\widehat{I}_n$ of $I = \{\alpha \in [0,1] : SB(\alpha) \geq 0\}$ to be specified below. For ease of notation we will assume throughout the paper that both I and its estimate $\widehat{I}_n$ are intervals. (This assumption is not necessary, however. All the arguments in this paper can be extended to the case of I and $\widehat{I}_n$ being unions of finitely many non-degenerate intervals.) In order to define T_n let $\widehat{I}_n = [\widehat{l}_n, \widehat{u}_n]$ and write

$$V_n(t) = \int_{\widehat{l}_n}^{t} p \, \log \widehat{d}_n(x_\alpha) \, d\,\alpha, \qquad t \in [\widehat{l}_n, \widehat{u}_n]. \tag{6}$$

We propose to reject H_0 if T_n is too large, where now

$$T_n = \sup_{t \in [\widehat{l}_n, \widehat{u}_n]} \Big| V_n(t) - \frac{t - \widehat{l}_n}{\widehat{u}_n - \widehat{l}_n} \, V_n(\widehat{u}_n) \Big|.$$

A plot of the function $t \to V_n(t) - \frac{t - \widehat{l}_n}{\widehat{u}_n - \widehat{l}_n} \, V_n(\widehat{u}_n)$; $t \in \widehat{I}_n$ provides another diagnostic plot, of course related to the SB-plot proposed above. If the null-hypothesis holds, then the graph of this function is expected to fluctuate around zero. Significance of deviations can be assessed using Theorem 4.2.

Other statistics which are continuous functions of $V_n(t) - \frac{t - \widehat{l}_n}{\widehat{u}_n - \widehat{l}_n} \, V_n(\widehat{u}_n)$ could also be used, like the Cramér-von-Mises type statistics

$$\int_{\widehat{l}_n}^{\widehat{u}_n} \Big[V_n(t) - \frac{t - \widehat{l}_n}{\widehat{u}_n - \widehat{l}_n} \, V_n(\widehat{u}_n) \Big]^2 \, d\,t. \tag{7}$$

In this paper we will only study T_n. However, using the results presented below it is more or less straightforward to derive distribution theory for (7).

Estimation of the set $I = \{\alpha : SB(\alpha) \geq 0\}$. Notice that I is a set where (under H_0) the function SB is constant. Estimating such sets, or similarly, estimating a function where its derivative is zero, or where it has flat spots, is notoriously difficult. All nonparametric techniques for estimating such sets usually either avoid this situations, or special considerations are needed. Our basic idea is to estimate the set $I(\epsilon_n) := \{\, SB(\alpha) \geq -\epsilon_n \,\}$ with $\epsilon_n \to 0$, and ϵ_n being a level where the SB-function has no flat parts. Of course, this approach is likely to introduce some bias. However, this bias goes into the "right" direction, meaning that it makes our method only a little more conservative.

The estimator we propose is

$$\widehat{I}_n = \widehat{I}_n(\epsilon_n) := \sup_{C \in \mathcal{I}} \int_C (\widehat{SB}_n(\alpha) + \epsilon_n) d\alpha \tag{8}$$

where $\mathcal{I}$:= class of all closed intervals on $[0,1]$, and where in our calculations we chose $\epsilon_n = \frac{\log n}{\sqrt{k_1}}$. (For a motivation of this choice of ϵ_n see Lemma 6.4.)

2.1 Note on neighborhood sizes k_1 and k_2 and the local geometry.

The local geometry, i.e. the specific metric used, enters the covariance structure of the limiting distribution of our test statistic T_n. For a given value x let $A_n(x)$ denote the ball around x of radius $d_n(x)$. Then, for two values x_α, x_β the intersection $A_n(x_\alpha) \cap A_n(x_\beta)$ crucially determines the covariance of $\log \widehat{d}_n(x_\alpha)$ and $\log \widehat{d}_n(x_\beta)$ (cf. proof of Theorem 6.5). Therefore, in order to make our life simple, we chose to define the line connecting the given endpoints x and y of our SB-plot as one of the coordinate axis, and used the city-block distance relative to this axis. Alternatively, we could have selected the rotationally invariant L_2-norm. The computation of the asymptotic variance of T_n then becomes slightly more complicated, however. (See remark after Theorem 6.5.)

Another issue important for practical applications is the choice of the neighborhood sizes. We derive some reasonable values of k_1 and k_2 using a Gaussian reference density assuming that we are interested in all regions except for low density ones. If we obtain a nearest neighbor estimate $\widehat{f}(x)$ of $f(x)$ on the basis of the nearest k_1 observations, then $\widehat{f}(x) = (k_1/n)/vol(\widehat{A}_n(x))$, where $\widehat{A}_n(x)$ is the ball of radius $\widehat{d}_n(x)$ centered at x. We will obtain the optimal value of k_1 for estimating f in all but very low density regions. Let R be a region in $\mathbf{R}^p$, to be determined later, which excludes low density regions. If we denote $L_0 = \int_{\|u\|\leq 1} du$, $L_1 = \int_{\|u\|\leq 1} u_1^2 du$ and the matrix of second partial derivatives of f by $D^2 f$, then straightforward calculations show that the expected mean integrated square error is

$$E \int_R (\hat{f}(x) - f(x))^2 dx \approx (1/k_1)c_1 + (1/4)(k_1/n)^{4/p} c_2$$

where

$$c_1 = \int_R f^2(x)dx, \qquad c_2 = (L_1/L_0^{1+2/p})^2 \int_R \{tr(D^2 f(x))\}^2 f(x)^{-4/p} dx.$$

Clearly this mean integrated square error is minimized when

$$k_1 = n^{4/(p+4)} (pc_1/c_2)^{p/(p+4)}.$$

Hence reasonable estimates of the constants c_1 and c_2 are needed. Let us split c_2 in two component parts $c_2 = c_{21}c_{22}$, where

$$c_{21} = (L_1/L_0^{1+2/p})^2 \text{ and } c_{22} = \int_R \{tr(D^2 f(x))\}^2 f(x)^{-4/p} dx.$$

Elementary calculations show that

$$\sqrt{c_{21}} = \begin{cases} 1/12 & \text{for city-block distance} \\ \{\pi(p+2)\}^{-1}(p/2)^{2/p}(\Gamma(p/2))^{2/p} & \text{for Euclidean distance} \end{cases}$$

Notice that c_1 and c_{22} involve the unknown density. As is done sometimes in the density estimation literature (Chapters 3 and 4 in Simonoff, 1996), we will assume the Gaussian reference density. Since X_i's have mean 0 and variance covariance matrix I, it is not unreasonable to take our reference density to be the pdf f_0 of $N_p(0, I)$, as we are only interested in deriving working values for k_1 and k_2. Since we are not interested in low density regions, we can take the region R to be of the form $R = \{x : x'x \leq \eta\}$ where the probability content of this region under f_0 is 0.95, i.e., $P(\chi_p^2 \leq \eta) = 0.95$. It can be shown that

$$c_1 = \int_R f_0^2(x)dx = (4\pi)^{-p/2} P(\chi_p^2 \leq 2\eta),$$

$$\begin{aligned} c_{22} &= \int_R \{tr(D^2 f_0(x))\}^2 f_0(x)^{-4/p} dx \\ &= (2\pi)^{-1/2+2/p} E[I(\chi_p^2 \leq \eta)(\chi_p^2 - p)^2 \exp(-(1/2 - 2/p)\chi_p^2)]. \end{aligned}$$

Even though c_1 and c_{22} do not have closed form expressions, they can be evaluated rather easily. With these working values of $c_1$and c_{22} we can now obtain a working value of k_1. The value of k_2 is taken to be $\min(\sqrt{n}, k_1)$.

For the nature data and the Swiss bank notes data, the dimension of the observations $p = 2$. However, the sample sizes are different. For the climatology data $n = 270$, leading to $k_1 = 40$ and $k_2 = 17$ using the city-block norm. For the Swiss bank data, the respective values of $(k_1, k_2) = (33, 15)$ with the sample size $n = 200$.

Choice of tuning parameters (k_1, k_2)

	n=100		n=400	
	p=2	p=4	p=2	p=4
city-block	20,10	9,9	50,20	18,18
squared-error norm	21,10	10,10	52,20	20,20

The table above gives (k_1, k_2) for different values of the sample size n and dimension p for the sup-norm and the squared-error norm when the probability content (for the standard multivariate normal) of the region which excludes low density is 0.95.

3 Illustrations

3.1 Application to Swiss bank note data

This data consists of bottom margin and diagonal length of 200 Swiss bank notes out of which 100 are real and the rest forged (e.g. see Simonoff, 1996). The analysis of univariate density estimates given in Simonoff's book (chapter 3.1) suggests that the distribution of diagonal length has two modes indicating existence of two different types of notes (real and forged). Similarly, the distribution of bottom margins has two distinct modes. It may even have a third mode, but it is so faint that it may even be a point of inflection.

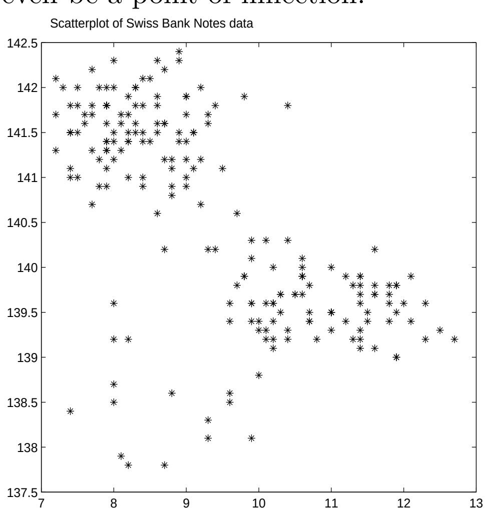

Fig. 3: Scatter plot of Swiss banknote data.

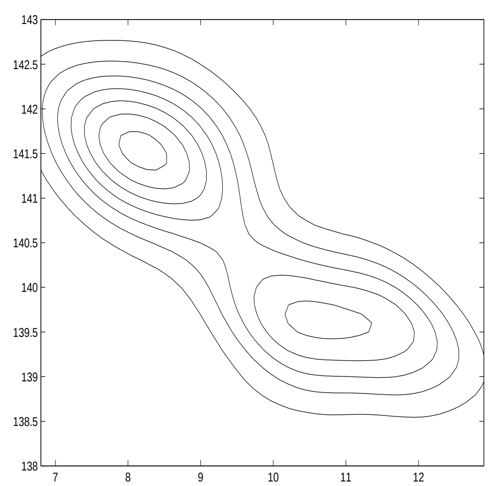

Fig. 4: Contour plot for Swiss banknote data.

We present our analysis on this data taking it to be a bivariate data. Using the approximate method described earlier we get the values of k_1 and k_2 needed for initial selection of modal candidates. Here, we have $(k_1, k_2) = (32, 15)$ for the city-block distance and $(k_1, k_2) = (33, 15)$ for the squared norm.

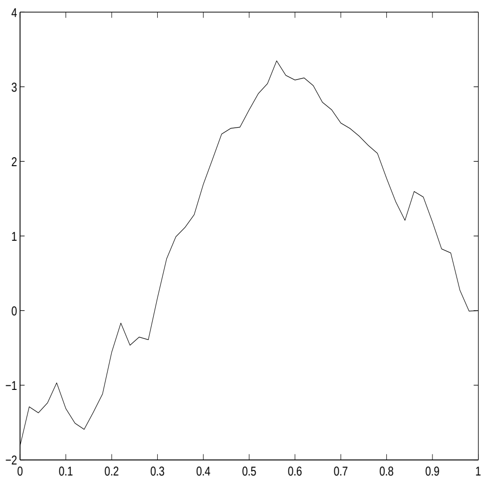

Fig. 5: SB-plot for Swiss bank note data using Euclidean distance.

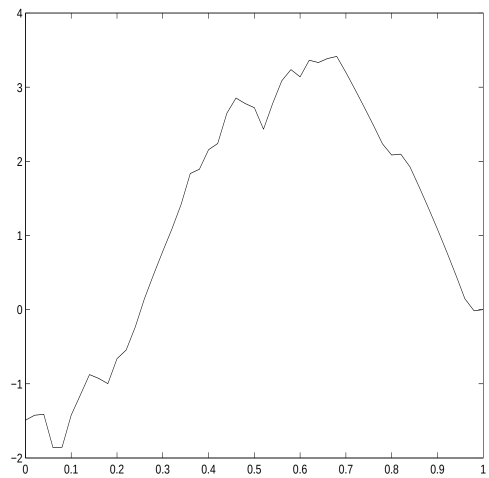

Fig. 6: SB-plot for Swiss bank note data using city-block distance.

Following our analysis described earlier, we find that there are two distinct modal regions for both the city-block distance and the Euclidean distance using either the naive test, i.e. the pointwise test based on $\widehat{SB}_n(\alpha)$, or the asymptotic Kolmogorov-Smirnov (K-S) type test, i.e. the test based on T_n, where we have been using significance level 0.10 in step (III). When the Euclidean distance is used, the modal regions are around (7.9,141.6) and (10.3, 139.7) with the value of the K-S type statistic 3.3476. For the city block distance the modes are around (8.2,141.7) and (10.2, 139.6) and the value of the K-S type statistic is 3.8708. Contour plots are presented below using a normal kernel method with the bandwidths suggested in chapter 4.2 in Simonoff. While the locations of the modes are slightly different for the different measures of distance, the corresponding SB plots convey similar information about the existence of the modes. The plots clearly suggest the distinctiveness of the modal regions.

3.2 Climatology data

This bivariate data set has 270 points, and is discussed in more detail in Section 2 above. It turns out that our approximate method leads to $k_1 = 40$, $k_2 = 17$ for the city-block distance. An application of our procedures for finding modal regions is unable to detect multimodal regions in the data for any distance measure employing any of the test procedures (in step (III)).

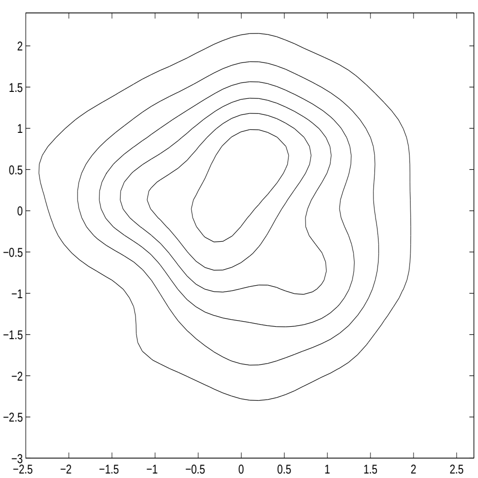

Fig. 7: Contour plot for climatology data.

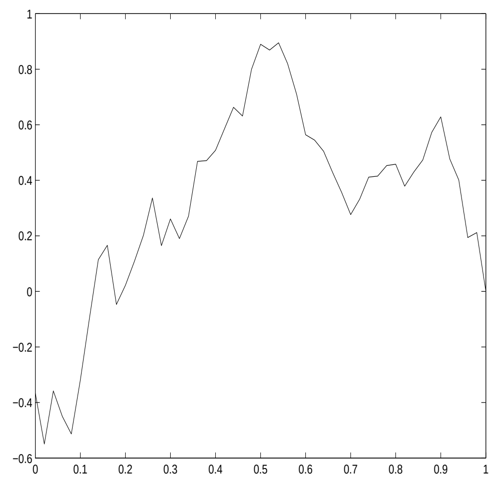

Fig. 8: SB-plot for climatology data using city-block distance.

A contour plot also does not indicate separate modal regions. For the purpose of illustration we have presented earlier the SB plot for the two best candidates for modes. These points are (-0.0921,0.4290) and (-0.3081,-1.1563) and the conclusion from the standardized SB plot is the same as what one would infer from the the contour plot or the plot of density estimate, i.e., there are no separate modes.

4 Large sample results

In this section we study the large sample behavior of the statistics $\widehat{SB}_n(\alpha)$ and T_n under the assumption that $X_1, X_2, \ldots$ are i.i.d. observations. Let us point out again, that in our results we will always assume the endpoints of the SB-plot to be given (i.e. non-random). We first state some more technical assumptions that are used throughout.

Assumptions.

A1. The metric ρ is the supremum distance (or the city-block distance), where for given endpoints x, y of the SB-plot the data have been rotated such that $x - y$ is one of the coordinate axis.

A2. The pdf f of the underlying distribution F is twice continuously differentiable, and for given endpoints x, y of the SB-plot we have with $x_\alpha = \alpha\, y + (1-\alpha)\, x$ that $f(x_\alpha) > 0$ for all $\alpha \in [0, 1]$.

Remark. The geometrical assumptions from $A1$ are made for mathematical convenience. They are not necessary. (See discussion after Theorem 6.5.)

In the following we use the notation

$$r = \frac{k_1}{n}.$$

In other words, r is the fraction of observations in each of the nearest neighbor neighborhoods under consideration. The first result motivates our method of repeated z-tests outlined in the introduction.

Theorem 4.1 *Let $X_1, X_2, \ldots$ be i.i.d. vectors in $\mathbf{R}^p$ with a common density f, and let r be such that $\frac{n\, r}{\log n} \to \infty$, and $n\, r^{1+\frac{4}{p}} \to 0$ as $n \to \infty$. Suppose that $A1$ and $A2$*

hold. For every $\alpha \in (0,1)$ *the following holds as* $n \to \infty$*.*

$$\sqrt{n\,r}\,(\widehat{SB}_n(\alpha) - SB(\alpha)) \to_{\mathcal{D}} \begin{cases} N_1 - \max(N_2,\, N_3) & \text{if } f(x) = f(y) \\ N_1 - N_2 & \text{if } f(x) \neq f(y), \end{cases}$$

where N_1, N_2, N_3 *are independent standard normal random variables.*

Remarks. (i) Note that the distribution of $N_1 - \max\{N_2, N_3\}$ is stochastically smaller than the normal with mean zero and variance two. In our applications we therefore used the critical values from the latter distribution even if $f(x_0) = f(x_1)$, which is a conservative approach.

(ii) As can be seen from the proof of this result, the condition $n\,r^{1+\frac{4}{p}} \to 0$ is needed for the bias to be negligible. Observe, however, that our rule of thumb for the choice of r is such that $n\,r^{1+\frac{4}{p}} \to c > 0$ (cf. section 2.1). This contradiction, in fact, is a common problem in smoothing and it is formally unavoidable if one relies on finding optimal bandwidths using the mean squared error of the density estimator. However, we argue that it still makes sense to use our rule of thumb. Observe that we are not really interested in estimating the underlying pdf f, but we want to test whether two given points lie in distinct model regions of f. Hence, from this point of view it makes sense to use the null hypothesis (4) rather than (2). This, in fact, avoids condition $n\,r^{1+\frac{4}{p}} \to 0$ as replacing $SB(\alpha)$ by $SB_n(\alpha)$ effectively means replacing $\log \frac{r}{C(p)f(x_\alpha)}$ by $p \log d_n(x_\alpha)$, and Theorem 4.1 holds for $\sqrt{n}\,(\widehat{SB}_n(\alpha) - SB_n(\alpha))$ without the condition $n\,r^{1+\frac{4}{p}} \to 0$. (In order to see this observe that the proof of Lemma 6.2 shows that (24) with $\log \frac{r}{C(p)f(x_\alpha)}$ replaced by $p \log d_n(x_\alpha)$ holds without the term $\Delta_n(x_\alpha)/r$, and this is the crucial term in these arguments (cf. proof of Corollary 6.3). Notice that also the $O(r^{4/p})$ term in (24) vanishes in this case, as it is also caused by $\Delta_n(x_\alpha)$.) In other words, we can ignore the bias in estimating f. To some extend these arguments also explain our empirical observation that the rule of thumb choices for k_1 an k_2 appear to work quite reasonably.

The following result concern our more global test statistic T_n. We first introduce some notation. For a given $\epsilon > 0$ let

$$g(\eta) = \sup_{0<\eta<\lambda<\epsilon} \Big|\, \{\alpha \in [0,1] : -\lambda - \eta \leq SB(\alpha) \leq -\lambda + \eta\} \,\Big|, \tag{9}$$

where for a set $A \subset \mathbf{R}$ we let $|A|$ denotes its Lebesgue measure. This function is needed to control bias, and we will need that for small $\epsilon > 0$ we have $g(\eta) \to 0$ as $\eta \to 0$. Notice that this means that $SB(\cdot)$ has no flat parts close to 0. The rate at which $g(\eta)$ convergence to zero if $\eta \to 0$ will become important, too. In fact, we will assume the following.

A3. There exists an $\epsilon_0 > 0$ such that for all $0 < \epsilon \leq \epsilon_0$ we have

$$g(\eta) \leq C\,\eta^\gamma \qquad \text{for some} \quad 0 < C < \infty,\ 0 < \gamma < \infty. \tag{10}$$

Recall that $I(\epsilon_n) := \{\, SB(\alpha) \geq -\epsilon_n \,\}$. Let $\ell(\epsilon_n) := \int_{I(\epsilon_n)} d\alpha$, and $\widehat{\ell}_n(\epsilon_n) := \widehat{u}_n - \widehat{l}_n$.

Theorem 4.2 *Let $X_1, X_2, \dots$ be i.i.d. vectors in $\mathbf{R}^p$ with a common distribution F. Assume that $|I| > 0$, $I(\epsilon_n) \in \mathcal{I}$ for all n, and that A1 - A3 hold. Further let r be such that $n\,r^{1+\frac{3}{p}} \to 0$, $\frac{n\,r^{1+\frac{2}{p}}}{(\log n)^3} \to \infty$, $\sqrt{n\,r^{1-\frac{1}{p}}}\,\epsilon_n^{1+\gamma} \to 0$, and $\sqrt{\frac{\log n}{nr}} = o(\epsilon_n) = o(1)$ as $n \to \infty$. Let $\widehat{a}_n = 2\,r^{-\frac{1}{p}} \int_{\widehat{I}_n(\epsilon_n)} \widehat{d}_n(x_\alpha)\,d\,\alpha$. Then we have under H_0 that as $n \to \infty$*

$$\sqrt{\frac{n\,r^{1-\frac{1}{p}}\,\|\,x - y\,\|}{\widehat{a}_n}}\; T_n \to_{\mathcal{D}} \sup_{t\in[0,1]} |B(t)|, \tag{11}$$

where $\{B(t), t \in [0,1]\}$ denotes a standard Brownian Bridge. If H_0 does not hold then $\sqrt{\frac{n\,r^{1-\frac{1}{p}}}{\widehat{a}_n}}\; T_n \to \infty$ in probability as $n \to \infty$.

Discussion of the assumptions of Theorem 4.2.

a) A heuristic explanation for the order $\sqrt{n\,r^{1-\frac{1}{p}}}$ is given by the fact that the effective number of observations which are used in the above theorem is of the order $n\,r^{1-\frac{1}{p}}$. This is so, because $\widehat{d}_n(x_\alpha)$ is based on the $k_1 = n\,r$ observations in $\widehat{A}_n(x_\alpha)$, and since each of the sets $\widehat{A}_n(x_\alpha)$ is a p-dimensional box with edges length of order $r^{1/p}$ we can find $O(r^{-\frac{1}{p}})$ many disjoint boxes along the line connecting x and y . This results in $O(n\,r\,\ell(\epsilon_n)\,r^{-\frac{1}{p}})$ effective observations. Rescaling of $I(\epsilon_n)$ to $[0,1]$ brings in a factor of $1/\ell(\epsilon_n)$, which goes in quadratically into the variance. Hence the order of the variance can be expected to be $O(\ell(\epsilon_n)/n\,r^{1-\frac{1}{p}})$, and $\ell(\epsilon_n)$ converges to a positive constant.

b) Our choice $\epsilon_n = \frac{\log n}{\sqrt{nr}}$ is compatible with the assumptions of the theorem provided $\gamma > 1/3$.

c) Assumption $nr^{1+\frac{3}{p}} \to 0$ is not needed when considering H_{0n} rather than H_0 (cf. remark (ii) given after Theorem 4.1). Hence, in this case, for any $\gamma > 0$ there exists an r satisfying the conditions of the theorem, and our rule of thumb choice of $r = c\,n^{-\frac{p}{4+p}}$ (or $n\,r^{1+\frac{4}{p}} \to c > 0$) fulfills the conditions provided $\gamma > 1/4$.

5 Proofs of main results

In this section we present the proofs of the theorems given above. These proofs draw on results presented in Section 6, which have been separated out since they not only serve the proofs of the main results, but they appear to be interesting in their own. We denote by F_n the empirical distribution based on $X_1, \ldots, X_n$, so that $F_n(C) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(X_i \in C)$ for $C \subset \mathbf{R}^p$.

Proof of Theorem 4.1. Let $Y_n(\alpha) = \sqrt{n\,r}\ \left(p \log \widehat{d}_n(x_{\alpha_i}) - \log \frac{r}{C(p)\,f(x_{\alpha_i})}\right)$, $\alpha \in [0,1]$. Corollary 6.3 says that for any $0 \le \alpha_1 < \cdots < \alpha_m \le 1$ the random vector $(\,Y_n(\alpha_1), \ldots, Y_n(\alpha_m)\,)$ is asymptotically m-dimensional standard normal. Further, it follows that

$W_n = \sqrt{n\,r}\,\left[\max(p \log \widehat{d}_n(x_0),\, p \log \widehat{d}_n(x_1)) - \max\left(\log \frac{r}{C(p)\,f(x_0)},\, \log \frac{r}{C(p)\,f(x_1))}\right)\right]$ converges in distribution to the maximum of two independent standard normal random variables if $f(x_0) = f(x_1)$, and to a standard normal random variable if $f(x_0) \neq f(x_1)$. Since $\sqrt{n\,r}\,(\widehat{SB}_n(\alpha) - SB(\alpha)) = Y_n(\alpha) - W_n$ the assertion follows.

□

Proof of Theorem 4.2. For an interval $C = [a,b]$ we define

$$\widehat{Z}_n(t,C) := \int_a^t p \log \widehat{d}_n(x_\alpha)\,d\alpha - \frac{t-a}{b-a}\int_a^b p \log \widehat{d}_n(x_\alpha)\,d\alpha, \qquad t \in [0,1]. \tag{12}$$

Similarly, let

$$Z_n(t,C) = \int_a^t \log \frac{r}{C(p)\,f(x_\alpha)}\,d\alpha - \frac{t-a}{b-a}\int_a^b \log \frac{r}{C(p)\,f(x_\alpha)}\,d\alpha, \qquad t \in [0,1]. \tag{13}$$

Observe that

$$T_n = \sup_{t \in \widehat{I}_n} |\widehat{Z}_n(t, \widehat{I}_n)|,$$

and that Proposition 6.6 says that

$$\sqrt{\frac{n\,r^{1-\frac{1}{p}}\,\|x-y\|}{\widehat{\ell}_n(\epsilon_n)}}\ \sup_{t\in[0,1]} |\,(\widehat{Z}_n - Z_n)(t, \widehat{I}_n)| \to_{\mathcal{D}} \sup_{t\in[0,1]} |G_F(t)| \tag{14}$$

for a Gaussian process $\{G_F(t), t \in [0,1]\}$ whose covariance function is given in Proposition 6.6. We will show below that under H_0 we have

$$\sqrt{\frac{nr^{1-\frac{1}{p}}\|x-y\|}{\widehat{\ell}_n(\epsilon_n)}}\ \sup_{t\in[0,1]} |\,Z_n(t, \widehat{I}_n)| = o_P(1) \qquad \text{as } n \to \infty. \tag{15}$$

Assuming that this is true, we obtain together with (14) that

$$\sqrt{\frac{n\,r^{1-\frac{1}{p}}\|x-y\|}{\widehat{\ell}_n(\epsilon_n)}}\,T_n \to_{\mathcal{D}} \sup_{t\in[0,1]} |G_F(t)| \qquad \text{as } n\to\infty.$$

We now show first that this implies the assertion of the theorem, namely that under H_0 we have

$$\sqrt{\frac{n\,r^{1-\frac{1}{p}}\,\|x-y\|}{\widehat{a}_n}}\,T_n \to_{\mathcal{D}} \sup_{t\in[0,1]} |B(t)| \qquad \text{as } n\to\infty,$$

with B being a standard Brownian Bridge. To see this let $I = [u,l]$ and write $z_\gamma = x_l + \gamma\,(x_u - x_l)$, $\gamma\in[0,1]$, where as always $x_\alpha = x + \alpha(y-x)$ for $\alpha\in[0,1]$. Notice that under H_0 we have $f(z_\gamma) = f(x_l)$ for all $\gamma\in[0,1]$. In other words, f is constant on I. It follows that under H_0 the covariance of the limiting Gaussian process G_F from Proposition 6.6 equals

$$2\left(\frac{1}{C(p)\,f(x_l)}\right)^{\frac{1}{p}}\,[\max(t,\,s) - t\,s]\,.$$

Further, we obtain from (30) and (32) that under H_0

$$\left|\,\widehat{a}_n - 2\,(u-l)\,\left(\frac{1}{C(p)\,f(x_l)}\right)^{\frac{1}{p}}\,\right| \;=\; o_P(1). \tag{16}$$

Thus the covariance of the limiting Gaussian process of $\sqrt{\frac{n\,r^{1-\frac{1}{p}}\|x-y\|}{\widehat{a}_n}}\;\widehat{Z}_n(t,\,\widehat{I}_n(\epsilon_n)\,)$ equals $\max(t,\,s) - t\,s,\;\; s,t\in[0,1]$, which is the covariance functions of a standard Brownian Bridge.

It remains to prove (15). Since under H_0 we have that $SB(\alpha) = 0$ for $\alpha\in I = [l,u]$

it follows that

$$\sup_{t\in[0,1]} \Big| \sqrt{nr^{1-\frac{1}{p}}}\, Z_n(t,\widehat{I}_n(\epsilon_n)) \Big|$$

$$\le \sqrt{nr^{1-\frac{1}{p}}} \sup_{t\in[0,1]} \Big[\Big| \int_{\widehat{I}_n(\epsilon_n)\setminus I} SB(\alpha)\,d\,\alpha \Big| + \frac{t-\widehat{l}_n}{\widehat{u}_n-\widehat{l}_n} \Big| \int_{\widehat{I}_n(\epsilon_n)\setminus I} SB(\alpha)\,d\,\alpha \Big| \Big]$$

$$\le \quad 2\sqrt{nr^{1-\frac{1}{p}}} \Big| \int_{\widehat{I}_n(\epsilon_n)\setminus I} SB(\alpha)\,d\,\alpha \Big| = O_P(\sqrt{nr^{1-\frac{1}{p}}}\ \epsilon_n^{1+\gamma}). \tag{17}$$

By assumption $\sqrt{nr^{1-\frac{1}{p}}}\ \epsilon_n^{1+\gamma} = o(1)$ as $n\to\infty$. This completes the proof.

The last equality in (17) can be seen by observing that

$$\int_{\widehat{I}_n(\epsilon_n)\setminus I} |SB(\alpha)|\,d\alpha \le \sup_{\alpha\in\widehat{I}_n(\epsilon_n)} |SB(\alpha)| \int_{\widehat{I}_n(\epsilon_n)\setminus I} d\alpha$$

$$= O_P(\epsilon_n^{1+\gamma}), \tag{18}$$

where we used that Lemma 6.4 implies $\mathrm{Leb}(\widehat{I}_n(\epsilon_n)\setminus I) \le |\,\widehat{I}_n(\epsilon_n)\Delta I\,| = O_P(g(\epsilon_n)) = O_P(\epsilon_n^{\gamma})$.

□

6 Miscellaneous results

Recall the notation $r = k_1/n$, where k_1 has been defined above as the number of observations in a nearest neighbor neighborhood $A_n(x)$ around x of radius $d_n(x)$. Also recall that $\widehat{d}_n(x_\alpha) = \mathrm{argmin}\{s:\ F_n(B(x_\alpha,s)) \ge r\}$ with corresponding nearest neighbor ball $\widehat{A}_n(x_\alpha)$, and $d_n(x_\alpha) = \mathrm{argmin}\{s:\ F(B(x_\alpha,s)) \ge r\}$. Let"$\Delta$" denote set-theoretic symmetric difference, i.e. $A\,\Delta\,B = (A\setminus B)\cup(B\setminus A)$.

Lemma 6.1 *Suppose that $n\,r/\log n \to \infty$. Then*

$$\sup_{\alpha\in[0,1]} F(\widehat{A}_n(x_\alpha)\Delta A_n(x_\alpha)) = O_P\Big(\sqrt{\frac{r}{n}\ \log n}\ \Big).$$

Proof. Observe that because F is continuous we have $|F_n(\widehat{A}_n(x_\alpha)) - F(A_n(x_\alpha))| = |F_n(\widehat{A}_n(x_\alpha)) - r| = O(1/n)$ almost surely uniformly in x and r. Hence, using obvious notation it follows that

$$\begin{aligned} F(\widehat{A}_n(x) \quad \Delta A_n(x)) &= |F(\widehat{A}_n(x_\alpha)) - F(A_n(x_\alpha))| \\ &\leq |(F_n - F)(A_n(x_\alpha))| + |(F_n - F)(\widehat{A}_n(x_\alpha) - A_n(x_\alpha))| + O(1/n). \end{aligned} \tag{19}$$

It follows from (19) that for n large enough (such that $\frac{C}{3}\sqrt{\frac{r}{n}\log n} \geq \frac{1}{n}$) we have with $B_n = \{\sup_{\alpha\in[0,1]} F(\widehat{A}_n(x_\alpha)\Delta A_n(x_\alpha)) > C\sqrt{\frac{r\,\log n}{n}}\}$ that

$$\begin{aligned} P(B_n) \leq \quad & P\Big(\sup_{\alpha\in[0,1]} |(F_n - F)(A_n(x_\alpha))| > \frac{C}{3}\sqrt{\frac{r\,\log n}{n}} \Big) + \\ & + \; P\Big(\sup_{\alpha\in[0,1]} \Big| \frac{(F_n - F)(\widehat{A}_n(x_\alpha) - A_n(x_\alpha))}{F(\widehat{A}_n(x)\Delta A_n(x))} \Big| > \frac{C}{3}\,,\; B_n \Big) \end{aligned} \tag{20}$$

Now note that the sets A_n and $\widehat{A}_n$ are boxes in $\mathbf{R}^p$. The class $\mathcal{B}^p$ of all boxes in $\mathbf{R}^p$ forms a so-called VC-class. A well-known property of the empirical process indexed by the VC-class $\mathcal{B}^p$ is that $\sup_{B\in\mathcal{B}^p;\, F(B)\leq\delta} |(F_n - F)(B)| = O_P(\sqrt{\frac{\delta}{n}\,\log n})$ as long as $\frac{\log n}{n\,\delta} < c$ for some appropriate constant $c > 0$ (e.g., see Alexander, 1984). Since $F(A_n(x_\alpha)) = r$ this implies that the first term on the r.h.s. of (20) can be made arbitrarily small as $n \to \infty$ by choosing C large enough. We now argue that the same holds true for the second term. We can estimate this term by

$$P\Big(\sup_{F(A\Delta B) > C\sqrt{\frac{r\,\log n}{n}}} \Big| \frac{(F_n - F)(A - B))}{F(A\Delta B)} \Big| > \frac{C}{3} \Big) \tag{21}$$

where the supremum is extended over all boxes $A, B \in \mathcal{B}^p$. By identifying boxes with their indicator function the class of all functions of differences $\{A - B;\, A, B \text{ boxes}\}$ also forms a VC-(subgraph)class, where exponential inequalities are readily available. Using them in conjunction with the so-called peeling device we obtain that also the probability in (21) can be made arbitrarily small as $n \to \infty$ by choosing C

large enough. We only briefly outline the peeling device. More details can be found e.g. in van de Geer, 2000. Let $\delta_n = C\sqrt{\frac{r\,\log n}{n}}$. The idea is to write

$$\{F(A\Delta B) > \delta_n\} = \bigcup_{j=1}^{\infty}\{2^{k-1}\delta_n < F(A\Delta B) < 2^k\delta_n\}, \tag{22}$$

in order to estimate (21) through

$$\sum_{j=1}^{\infty} P\Big(\sup_{F(A\Delta B)<2^k\delta_n} |(F_n - F)(A - B))| > \frac{C}{3}2^{k-1}\delta_n\Big). \tag{23}$$

Now good exponential bounds for the probabilities on the r.h.s. can be used to show that the sum becomes small for large C and $n \to \infty$.

$\square$

Lemma 6.2 *Suppose that the assumptions of Theorem 4.1 hold, and let* $\Delta_n(x_\alpha) := \int_{A_n(x_\alpha)}[f(z) - f(x_\alpha)]\,dz$. *There exists a constant* $C(p)$ *such that for* $n\,r/\log n \to \infty$ *we have*

$$\sup_{\alpha\in[0,1]}\Big|\Big(p\,\log\widehat{d}_n(x_\alpha) - \log\frac{r}{C(p)\,f(x_\alpha)}\Big) - \frac{(F - F_n)(A_n(x_\alpha))}{F(A_n(x_\alpha))} + \frac{\Delta_n(x_\alpha)}{r}\Big|$$
$$= O_P\Big(r^{\frac{2}{p}}\sqrt{\frac{\log n}{n\,r}}\Big) + O_P\Big(\Big(\frac{\log n}{n\,r}\Big)^{\frac{3}{4}}\Big) + O\Big(r^{\frac{4}{p}}\Big). \tag{24}$$

Moreover, $\sup_{\alpha\in[0,1]}\Delta_n(x_\alpha) = O\big(r^{1+\frac{2}{p}}\big)$.

Corollary 6.3 *Suppose that the assumptions of Lemma 6.2 hold. If, in addition,* $nr^{1+\frac{4}{p}} = o(1)$ *then we have the following.*

(i) *For any* $(\,\alpha_1,\ldots,\alpha_m\,) \in [0,1]^m$ *as* $n \to \infty$

$$\Big(\sqrt{n\,r}\,\Big(p\,\log\widehat{d}_n(x_{\alpha_i}) - \log\frac{r}{C(p)\,f(x_{\alpha_i})}\Big);\; i = 1,\ldots,m\Big) \to_{\mathcal{D}} \mathcal{N}_m(0, I).$$

(ii)

$$\sup_{\alpha\in[0,1]}\sqrt{\frac{n\,r}{\log n}}\,\Big|p\,\log\widehat{d}_n(x_\alpha) - \log\frac{r}{C(p)\,f(x_\alpha)}\Big| = O_P(1).$$

(iii)

$$\sup_{\alpha\in[0,1]} \sqrt{\frac{n\,r}{\log n}}\,\Big|\,\widehat{SB}_n(\alpha) - SB(\alpha)\Big| = O_P(1).$$

Proof of Corollary 6.3. The main term in (24) is $U_n(\alpha) = \frac{(F-F_n)(A_n(x_\alpha))}{F(A_n(x_\alpha))} = \frac{(F-F_n)(A_n(x_\alpha))}{r}$. Under the present assumptions a standard application of a Lindeberg-Feller CLT hence shows that $\sqrt{n\,r}\,U_n(\alpha_i) \to \mathcal{N}(0,1)$ for each $i = 1,\dots,m$. As for joint convergence observe that $\mathrm{Cov}(\,\sqrt{n\,r}\,U_n(\alpha_i),\,\sqrt{n\,r}\,U_n(\alpha_j)\,) = \frac{1}{r}\,(\,F[\,A_n(\alpha_i)\cap A_n(\alpha_j)\,] - \,F(\,A_n(\alpha_i)\,)\,F(\,A_n(\alpha_j)\,)\,\,) = \frac{1}{r}\,\big(\,0 - r^2\,\big) = -r$ for n large enough, and hence $\mathrm{Cov}(\,U_n(\alpha_i),\,U_n(\alpha_j)\,) \to 0$ as $n\to\infty$. It remains to show that all the remainder terms appearing in (24) tend to zero. Notice that the two O_P-terms in (24) tend to zero by the assumptions on r, and also $\sqrt{n\,r}\;\frac{\Delta_n(x_\alpha)}{r} = o_P(1)$. By using the last assertion of Lemma 6.2, negligibility of the latter term follows since $\sqrt{n\,r}\;\frac{\Delta_n(x_\alpha)}{r} = \sqrt{n\,r}\;O_P\left(r^{\frac{2}{p}}\right) = O_P\left(\sqrt{n\,r^{1+\frac{4}{p}}}\right)$ which tends to zero as $n\to\infty$ by assumption on r. This completes the proof of part (i).

Part (ii) follows similarly by observing that the class of boxes is a so-called VC-class, and hence, by using the fact that $F(A_n(x_\alpha)) = r$ empirical process theory implies that

$$\sup_{\alpha\in[0,1]} \big|\,\sqrt{n}\,(F_n - F)(A_n(x_\alpha))\,\big| = O_P\left(\sqrt{r\,\log\frac{1}{r}}\right).$$

Hence (24) implies the result similarly to the above since the order of the remainder terms are uniform in α.

Part(iii) is an immediate consequence of part (ii) by observing that

$$\sup_{\alpha\in[0,1]} \sqrt{nr}\,(\widehat{SB}_n(\alpha) - SB(\alpha)\,) \le 2\sup_{\alpha\in[0,1]} \sqrt{\frac{n\,r}{\log n}}\,\left|\,p\,\log\widehat{d}_n(x_\alpha) - \log\frac{r}{C(p)\,f(x_\alpha)}\,\right|$$

(cf. proof of Theorem 4.1).

□

Proof of Lemma 6.2. The starting point of this proof is the fact that by using a Taylor expansion we can write

$$p\,\log \widehat{d}_n(x_\alpha) - \log \frac{r}{C(p)\,f(x_\alpha)} = \left(\frac{C(p)\,f(x_\alpha)\,\widehat{d}_n^{\,p}(x_\alpha)}{r} - 1\right) + \text{remainder} \tag{25}$$

We will first concentrate on the main term term on the right-hand side. For a (measurable) set A we will (as above) denote by $|A|$ the volume (or Lebesgue measure) of A. We can write

$$\begin{aligned} F(B(x,t)) &= f(x)\,|B(x,t)| + \Delta(x,t) \\ &= f(x)\,C(p)\,t^p\; + \Delta(x,t), \end{aligned} \tag{26}$$

where $C(p)$ is a constant that depends only on the dimension p and on the particular metric used, and $\Delta(x,t) = \int_{B(x,t)} (f(x) - f(z))\,d\,z$. With $\widehat{\Delta}_n(x_\alpha) = \Delta(x_\alpha, \widehat{d}_n(x_\alpha))$ and $\Delta_n(x_\alpha) = \Delta(x_\alpha, d_n(x_\alpha))$ expansion (26) implies

$$\begin{aligned} \frac{C(p)\,f(x_\alpha)\,\widehat{d}_n^{\,p}(x_\alpha)}{r} - 1 &= \frac{F(\widehat{A}_n(x_\alpha)) - F(A_n(x_\alpha))}{F(A_n(x_\alpha))} - \frac{\widehat{\Delta}_n(x_\alpha)}{r} \\ &= \frac{F(\widehat{A}_n(x_\alpha)) - F(A_n(x_\alpha))}{F(A_n(x_\alpha))} - \frac{\Delta_n(x_\alpha)}{r} + \frac{\widehat{\Delta}_n(x_\alpha) - \Delta_n(x_\alpha)}{r}. \end{aligned} \tag{27}$$

We now consider the last term in (27) and show that

$$\widehat{\Delta}_n(x_\alpha) - \Delta_n(x_\alpha) = O_P\left(r^{\frac{2}{p}}\sqrt{\frac{r\,\log n}{n}}\right) \qquad \text{uniformly in } \ \alpha \in [0,1]. \tag{28}$$

To see this, note that since $\widehat{A}_n(x_\alpha)$ and $A_n(x_\alpha)$ are two neighborhoods with the same midpoint (and hence are nested) it follows that for some constant $C > 0$ we have

$$\begin{aligned} |\widehat{\Delta}_n(x_\alpha) - \Delta_n(x_\alpha)| &= \Big|\int_{\widehat{A}_n(x_\alpha)\Delta A_n(x_\alpha)} [f(x_\alpha) - f(z)]\,d\,z\;\Big| \\ &\le C\,|\,\widehat{A}_n(x_\alpha)\Delta A_n(x_\alpha)\,|\,\big(\max\{\widehat{d}_n(x_\alpha), d_n(x_\alpha)\}\big)^2. \end{aligned} \tag{29}$$

Since by assumption $\sup_{\alpha\in[0,1]} f(x_\alpha) > 0$, Lemma 6.1 implies that $|\,\widehat{A}_n(x_\alpha)\Delta A_n(x_\alpha)\,| = O_P\big(\sqrt{\frac{r\,\log n}{n}}\,\big)$. To complete the proof of (28) we show that (i) $d_n(x_\alpha) = O(r^{\frac{1}{p}})$, and (ii) $O(\max\{\widehat{d}_n(x_\alpha), d_n(x_\alpha)\}) = O_P(d_n(x_\alpha))$. In fact, we will even show that

$$\sup_{\alpha\in[0,1]} \Big|\,\frac{\widehat{d}_n(x_\alpha)}{d_n(x_\alpha)} - 1\,\Big| = o_P(1). \tag{30}$$

First we show (i). As in (29) we have uniformly in $\alpha \in [0,1]$ that

$$|\Delta_n(x_\alpha)| \le C\,|A_n(x_\alpha)|\,d_n^2(x_\alpha) = O(d_n^{p+2}(x_\alpha)). \tag{31}$$

This together with (26) implies that $r = C(p)\,f(x_\alpha)\,d_n^p(x_\alpha)\,[1 + O(d_n^2(x_\alpha))]$, which means that

$$d_n^p(x_\alpha)/r \to \frac{1}{C(p)\,f(x_\alpha)}. \tag{32}$$

We also have shown that

$$\Delta_n(x_\alpha) = O(r^{1+\frac{2}{p}}) \tag{33}$$

uniformly in $\alpha \in [0,1]$, which is the last assertion of the lemma. Now we prove (ii). Using (29) and the fact that (since the neighborhoods $\widehat{A}_n(x_\alpha)$ and $A_n(x_\alpha)$ are nested) we have $|\,\widehat{A}_n(x_\alpha)\,\Delta\,A_n(x_\alpha)\,| = O\big(\widehat{d}_n^p(x_\alpha) - d_n^p(x_\alpha)\big)$, we can write

$$\begin{aligned} F(\widehat{A}_n(x_\alpha)) - F(A_n(x_\alpha)) &= C(p)\,f(x_\alpha)\,(\widehat{d}_n^p(x_\alpha) - d_n^p(x_\alpha)) + \widehat{\Delta}_n(x_\alpha) - \Delta_n(x_\alpha) \\ &= C(p)\,f(x_\alpha)\,(\widehat{d}_n^p(x_\alpha) - d_n^p(x_\alpha))\,[1 + O(\max\{\widehat{d}_n(x_\alpha), d_n(x_\alpha)\})]. \end{aligned}$$

Using (32) (and recall that $F(A_n(x_\alpha)) = r$) it follows that

$$\frac{\widehat{d}_n^p(x_\alpha) - d_n^p(x_\alpha)}{d_n^p(x_\alpha)}(1 + O(\max\{\widehat{d}_n(x_\alpha), d_n(x_\alpha)\}) = \frac{F(\widehat{A}_n(x_\alpha)) - F(A_n(x_\alpha))}{F(A_n(x_\alpha))}. \tag{34}$$

Further, by definition $F_n(\widehat{A}_n(x_\alpha)) \ge r$ and $\widehat{A}_n(x_\alpha)$ it is the smallest ball around x_α with this property. Since F is continuous we obtain that $F_n(\widehat{A}_n(x_\alpha)) = r + O(1/n)$ almost surely. It follows that $F(\widehat{A}_n(x_\alpha)) - F(A_n(x_\alpha)) = (F - F_n)(\widehat{A}_n(x_\alpha)) +$

$O(1/n) = o_P(1)$. The last equation follows from empirical process theory by using the fact that each $\widehat{A}_n(x_\alpha)$ is a rectangle, and that a uniform law of large numbers holds for the class of rectangles. (In fact, the rectangles form a so-called VC-class; see e.g. van der Vaart and Wellner, 1996.) Hence it follows from (34) that $\frac{\widehat{d}_n^p(x_\alpha)}{d_n^p(x_\alpha)} - 1 = o_P(1)$ which is (ii).

We now focus on the first term on the r.h.s. of (27). We have

$$\begin{aligned} F(\widehat{A}_n(x_\alpha)) - F(A_n(x_\alpha)) &= (F - F_n)(\widehat{A}_n(x_\alpha)) + O(1/n) \\ &= -(F_n - F)(A_n(x_\alpha)) - [(F_n - F)(\widehat{A}_n(x_\alpha)) - (F_n - F)(A_n(x_\alpha))] + O(1/n) \\ &= -(F_n - F)(A_n(x_\alpha)) + O_P\Big(\frac{r^{\frac{1}{4}} (\log n)^{\frac{3}{4}}}{n^{\frac{3}{4}}} \Big) \end{aligned} \tag{35}$$

The O_P-term in (35) follows from empirical process theory as follows. First notice that we have for so called VC-classes of sets $\mathcal{C}$ that $\sup_{C \in \mathcal{C}:\, F(C) \le \delta} |\,(F_n - F)(C)\,| = O_P\big(\sqrt{\frac{\delta\, \log(1/\delta)}{n}}\big)$, for $\delta \ge c\, \log n/n$ for some appropriate $c > 0$ (e.g. Alexander 1984.) We already used above that the class of rectangles forms a VC-class. Further it is know that if $\mathcal{C}$ is a VC-class so is $\{\,C \setminus D \,:\, C, C \in \mathcal{C}\,\}$. Hence by writing $(F_n - F)(\widehat{A}_n(x_\alpha)) - (F_n - F)(A_n(x_\alpha)) = (F_n - F)(\widehat{A}_n(x_\alpha) \setminus \widehat{A}_n(x_\alpha)\,) - (F_n - F)(\widehat{A}_n(x_\alpha) \setminus \widehat{A}_n(x_\alpha)\,)$ the asserted rate in (35) follows by using Lemma 6.1.

Finally, notice that the above shows that $\sup_{\alpha \in [0,1]} \Big| \frac{C(p)\, f(x_\alpha)\, \widehat{d}_n^p(x_\alpha)}{r} - 1 \Big| = o_P(1)$. Hence, a Taylor expansion gives

$$\log \widehat{d}_n^p(x_\alpha) - \log \frac{r}{C(p)\, f(x_\alpha)} = \Big(\frac{C(p)\, f(x_\alpha)\, \widehat{d}_n^p(x_\alpha)}{r} - 1 \Big) + O_P\Big(\Big(\frac{C(p)\, f(x_\alpha)\, \widehat{d}_n^p(x_\alpha)}{r} - 1 \Big)^2 \Big) \tag{36}$$

with the O_P-term being uniform in α. Collecting (27), (28), (33) and (35) and plugging them into (36) completes the proof.

$\square$

Lemma 6.4 (Rates of convergence for $\widehat{I}_n(\epsilon_n)$.) *Suppose that the assumptions of Theorem 4.1 hold. In addition assume that* $\sqrt{\frac{\log n}{n\,r}} = o(\epsilon_n)$ *and that (10) holds with* $0 < \gamma \le 1$. *Then we have*

$$d_F(\,\widehat{I}_n(\epsilon_n),\; I(\epsilon_n)\,) = o_P\,(\epsilon_n^\gamma) \qquad \text{as } \; n \to \infty,$$

where $I(\epsilon_n) = \{\,\alpha : SB(\alpha) \ge -\epsilon_n\,\}$. *We also have* $d_F(\,\widehat{I}_n(\epsilon_n),\; I\,) = O_P\,(\epsilon_n^\gamma)$ *as* $n \to \infty$.

Proof. For $\lambda \in \mathbf{R}$ and any (measurable) set C let

$$\begin{aligned} H_{n,\lambda}(C) = &\int_C \left(\log \frac{r}{C(p)\,f(x_\alpha)} - \lambda \right) d\,\alpha \\ \widehat{H}_{n,\lambda}(C) = &\int_C (\,p\,\log \widehat{d}_n(x_\alpha) - \lambda\,)\; d\,\alpha \end{aligned}$$

Observe that with $\widehat{\mu}_n = -\epsilon_n + \max\Big\{\,p\,\log \widehat{d}_n(x),\, p\,\log \widehat{d}_n(y)\Big\}$ and $\mu_n = -\epsilon_n + \max\{\,\log \frac{r}{C(p)\,f(x)},\; \log \frac{r}{C(p)\,f(y)}\}$ we have

$$\widehat{I}_n(\epsilon_n) = \underset{C}{\operatorname{argmax}}\{\widehat{H}_{n,\widehat{\mu}_n}(C)\} \qquad \text{and} \qquad I(\epsilon_n) = \underset{C}{\operatorname{argmax}}\{H_{n,\mu_n}(C)\}.$$

It follows that $0 \le H_{n,\mu_n}(I(\epsilon_n)) - H_{n,\,\mu_n}(\widehat{I}_n(\epsilon_n))$ and also $0 \le \widehat{H}_{n,\widehat{\mu}_n}(\widehat{I}_n(\epsilon_n)) - \widehat{H}_{n,\widehat{\mu}_n}(I(\epsilon_n))$, and hence we obtain

$$\begin{aligned} 0 \le H_{n,\,\mu_n}(I(\epsilon_n)) - H_{n,\,\mu_n}(\widehat{I}_n(\epsilon_n)) &\le (H_{n,\,\mu_n} - \widehat{H}_{n,\,\widehat{\mu}_n})(I(\epsilon_n)) - (H_{n,\,\mu_n} - \widehat{H}_{n,\widehat{\mu}_n})(\widehat{I}_n(\epsilon_n)) \\ &= \int_{\widehat{I}_n(\epsilon_n)} - \int_{I(\epsilon_n)} (\widehat{SB}_n(\alpha) - SB(\alpha))\; d\,\alpha. \qquad (37) \end{aligned}$$

Further we have

$$\begin{aligned} H_{n,\,\mu_n}(I(\epsilon_n)) - H_{n,\,\mu_n}(\widehat{I}_n(\epsilon_n)) &= \int_{\widehat{I}_n(\epsilon_n)\,\Delta\,I(\epsilon_n)} |\mathrm{SB}(\alpha) + \epsilon_n|\; d\,\alpha \\ &\ge \eta \cdot \Big|\,\{\widehat{I}_n(\epsilon_n)\,\Delta\,I(\epsilon_n)\; \cup |\mathrm{SB}(\alpha) + \epsilon_n| \ge \eta\}\,\Big|, \quad (38) \end{aligned}$$

and hence we obtain for $\eta < \epsilon_n$ that

$$
\begin{aligned}
|\,\widehat{I}_n(\epsilon_n)\,\Delta\,I(\epsilon_n)\,| &\le \big|\{|\mathrm{SB}(\alpha)+\epsilon_n| \le \eta\}\big| + \big|\{\widehat{I}_n(\epsilon_n)\,\Delta\,I(\epsilon_n)\ \cup |\mathrm{SB}(\alpha)+\epsilon_n| \ge \eta\}\big| \\
&\le O(\,g(\eta)\,) + \frac{1}{\eta}\,\int_{\widehat{I}_n(\epsilon_n)} - \int_{I(\epsilon_n)} (\widehat{SB}_n(\alpha) - SB(\alpha))\,d\,\alpha \qquad (39) \\
&\le O(\,g(\eta)\,) + \frac{1}{\eta}\,\big|\,\widehat{I}_n(\epsilon_n)\,\Delta\,I(\epsilon_n)\big| \sup_{\alpha\in[0,1]} |\widehat{SB}_n(\alpha) - SB(\alpha)| \qquad (40)
\end{aligned}
$$

Corollary 6.3 part (iii) says that $\sup_{\alpha\in[0,1]} |\widehat{SB}_n(\alpha) - SB(\alpha)| = O_P\Big(\sqrt{\frac{\log n}{n\,r}}\,\Big)$. Hence, for any $\delta > 0$ we can find a constant $C(\delta)$ such that $A_n(\delta) = \{\sup_{\alpha\in[0,1]} |\widehat{SB}_n(\alpha) - SB(\alpha)| \le C(\delta)\sqrt{\frac{\log n}{n\,r}}\,\}$ for n large enough satisfies $P(A_n(\delta)) > 1-\delta$. Consequently, if we choose $\eta = 2\,C(\delta)\sqrt{\frac{\log n}{n\,r}}$ (which, as required, is $\le \epsilon_n$ for n large enough) then (40) implies that on $A_n(\delta)$ we have

$$
\big|\,\widehat{I}_n(\epsilon_n)\,\Delta\,I(\epsilon_n)\big| \le O\Big(\,g\,\Big(\,2\,C(\delta)\sqrt{\frac{\log n}{n\,r}}\,\Big)\,\Big).
$$

In other words, $\big|\,\widehat{I}_n(\epsilon_n)\,\Delta\,I(\epsilon_n)\big| = O_P\big(\,(\,\frac{\log n}{n\,r}\,)^{\gamma/2}\,\big) = o_P(\epsilon_n^\gamma)$ where the last equality holds by assumption on ϵ_n.

□

For some of the above arguments we need to study the asymptotic behavior of a certain empirical process which we study next. For given sequences $x_n, y_n, \in \mathbf{R}^p$ with $x_n \to x$ and $y_n \to y$ we denote points on the line connecting the endpoints x_n and y_n by $x_{n,\alpha} = x_n + \alpha\,(y_n - x_n)$, and points connecting x and y by $x_\alpha = x + \alpha\,(y - x)$, respectively. Further denote for any $t \in [0,1]$

$$
\Psi_t(y) = \int_0^t \mathbf{1}(y \in A_n(x_{n,\alpha}))\,d\,\alpha. \qquad (41)
$$

We now study the process $\{\nu_n(\Psi_t);\ t \in [0,1]\}$ where

$$
\nu_n(\Psi_t) = \sqrt{\frac{n\,\|x_n - y_n\|}{r^{1+\frac{1}{p}}}}\,(F_n - F)(\Psi_t) = \sqrt{\frac{n\,\|x_n - y_n\|}{r^{1+\frac{1}{p}}}}\,\Big(\frac{1}{n}\sum_{i=1}^n \Psi_t(X_i) - r\,t\,\Big). \qquad (42)
$$

Notice that $r\,t = \mathrm{E}\,\Psi_t(X_i)$, because $\mathrm{E}\,\mathbf{1}(X_i \in A_n(x_{n,\alpha})) = F(A_n(x_{n,\alpha})) = r$. Using the notation introduced above we have the following result.

Theorem 6.5 *Let $X_1, X_2, \ldots \sim_{iid} F$ be random vectors in $\mathbf{R}^p$. Assume that assumptions A1 and A2 hold for endpoints x_n, y_n (for all n) and also for x, y. Assume further that $\frac{r^{1/p}}{\|x_n - y_n\|} = o(1)$ as $n \to \infty$. Then we have as $n \to \infty$*

$$\nu_n(\Psi_t) \to_{\mathcal{D}} G_F(t),$$

in the space $C[0,\,1]$ where $\{G_F(t), t \in [0,\,1]\}$ denotes a mean zero Gaussian process with $\mathrm{Cov}(G_F(t), G_F(s)) = D(\min(s,\,t)) - \|x - y\|\,s\,t\,\mathbf{1}(p = 1)$, where $D(\min(s,\,t)) = 2\int_0^{\min(s,t)} \big(\frac{1}{C(p)\,f(x_\alpha)}\big)^{\frac{1}{p}}\,d\,\alpha$.

Remark. Notice that the covariance structure for $p > 1$ is the one of a generalized Brownian motion, and if in addition H_0 holds, i.e. $f(x_\alpha)$ is constant over all $x_\alpha \in I$, the limit is precisely a rescaled Brownian motion. For $p = 1$ the covariance structure of the limit is more similar to the one of a Brownian bridge, however.

The geometric assumptions made in $A2$ determine the covariance structure of ν_n (cf. (45) and (46)). In case we use the Euclidean norm, the function D in the covariance formula has to be replaced by

$$D(s) = \frac{4}{\|x - y\|}\int_0^1 \Psi_\omega(t)\,d\,t\int_0^s d_n(x_\alpha)\,d\,\alpha$$

where $\omega = \frac{x-y}{\|x-y\|}$ and

$$\Psi_\omega(t) = \prod_{i=1}^{p}(1 - |t|\,|\omega_i|\,)_+\,\mathbf{1}(\,|t| \le 1\,).$$

In case we use the city-block but do not rotate the data, and for the Euclidean norm we have

$$\Psi_\omega(t) = \frac{B(\frac{p-1}{2},\,\frac{3}{2};\,1 - t^2) - |t|\,B(\frac{t-1}{2},\,1;\,1 - t^2)}{B(\frac{p-1}{2};\,\frac{3}{2})}\,\mathbf{1}(|t| \le 1).$$

Here $B(\cdot,\, \cdot)$ and $B(\cdot,\, \cdot;\, \cdot)$ denote the complete and incomplete Beta-function,respectively.

Proof of Theorem 6.5 We can write ν_n as a function indexed empirical process as

$$\nu_n(\Psi_t) = \sqrt{n}\,(F_n - F)(h_{t,n}),$$

with index class $\mathcal{H}_n := \{h_{t,n} := \sqrt{\frac{\|x_n - y_n\|}{r^{1+\frac{1}{p}}}}\, \Psi_t \,:\, t \in [0,1]\}$. We will apply Theorem 2.11.22 of van der Vaart and Wellner (1996) which can be used to prove weak convergence for empirical processes indexed by such classes of functions depending on n. We need to verify the conditions of this theorem. First observe that for each n the class of functions $\{h_{t,n} := \sqrt{\frac{\|x_n - y_n\|}{r^{1+\frac{1}{p}}}}\, \Psi_t \,:\, t \in [0,1]\}$ is a totally ordered class of positive functions, and hence it follows that the class of subgraphs $\{H_{t,n} = \{(x,y) \in \mathbf{R}^{p+1};\ 0 < y \le h_{t,n}(x)\};\ t \in [0,1]\,\}$ forms a VC-class of VC-dimension 2. This VC-subgraph property guarantees the metric entropy condition of Theorem 2.11.22 to hold.

It remains to verify condition (2.11.21) of van der Vaart and Wellner. In order to do that first observe that (i) $h_{1,n}$ is a square integrable cover for $\mathcal{H}_n$, because $0 \le h_{t,n}(x) \le h_{1,n}(x)$ and $E(h_{1,n}(X))^2 = 2\int_0^1 \left(\frac{1}{C(p)\, f(x_\alpha)}\right)^{\frac{1}{p}} d\alpha\ (1 + o(1)\,) < \infty$ (cf. (52)). The second condition in (2.11.21) of v.d.Vaart and Wellner is (ii) $E[h_{1,n}^2(X)\mathbf{1}(h_{1,n} > \eta\sqrt{n}\,)] \to 0$ as $n \to \infty$ for each $\eta > 0$. To see this notice that $\{h_{1,n} > \eta\sqrt{n}\} = \{\, \Psi_1 > \eta \sqrt{\frac{n r^{1+\frac{1}{p}}}{\|x_n - y_n\|}}\,\}$. Condition (i) follows immediately by observing that $\sqrt{\frac{n r^{1+\frac{1}{p}}}{\|x_n - y_n\|}} \to \infty$ and $0 \le \Psi_1 \le 1$ and hence $\{h_{1,n} > \eta\sqrt{n}\} = \emptyset$ for n large enough.

The third an final condition of (2.11.21) is (iii) $\sup_{|t-s|<\delta_n} E(h_{t,n}(X) - h_{s,n}(X))^2 \to 0$ as $n \to \infty$ for every sequence $\delta_n \to 0$. This can seen as follows. First observe that for $s < t$ we have $E(h_{t,n}(X) - h_{s,n}(X))^2 = \frac{\|x_n - y_n\|}{r^{1+\frac{1}{p}}} \big[E\Psi_t^2(X) + E\Psi_s^2(X) - 2\,E(\Psi_t(X)\Psi_s(X)) \big]$. Plugging in (52) shows that

$$E[h_{t,n}(X) - h_{s,n}(X)]^2 \leq \; 2 \int_s^t \Big(\frac{1}{C(p)\, f(x_\alpha)}\Big)^{\frac{1}{p}} d\,\alpha \; (\, 1 + o(1)\,) \; \leq \; C\, |t - s|, \quad (43)$$

which shows (iii). Theorem 2.11.22 now implies weak convergence of $\{\nu_n(h_{t,n});\, h_{t,n} \in \mathcal{H}_n\,\}$ to a Gaussian process, provided the covariances $\mathrm{Cov}(h_{t,n}, h_{s,n})$ converge pointwise. The corresponding limit then is the covariance function of the limit process. We have

$$\begin{aligned} &\mathrm{Cov}\,(\,\Psi_s(X_i),\, \Psi_t(X_i)\,) \\ &\quad = \mathrm{E}\,\Big(\int_0^s \mathbf{1}(X_i \in A_n(x_{n,\alpha}))\; d\,\alpha \cdot \int_0^t \mathbf{1}(X_i \in A_n(x_{n,\beta}))\; d\,\beta \Big) - r^2\, s\, t \\ &\quad = \int_0^s \int_0^t F(A_n(x_{n,\alpha}) \,\cap\, A_n(x_{n,\beta}))\; d\,\alpha\; d\,\beta - r^2\, s\, t. \qquad (44) \end{aligned}$$

A key ingredient for further calculations is the equality

$$F(A_n(x_{n,\alpha}) \cap A_n(x_{n,\beta})) = r\, \Big(\frac{d_n(\alpha) + d_n(\beta) - |\alpha - \beta|}{2\, \min\,(d_n(\alpha),\, d_n(\beta)\,)} \Big)^+ \big[\, 1 + O\,\big(\, \min\{\, d_n(\alpha),\, d_n(\beta)\,\}\,\big)\,\big] \qquad (45)$$

where $d_n(\alpha) = d_n(x_\alpha)/\|x_n - y_n\|$ and $a^+ = \max(a, 0)$. Here the term inside $(\quad)^+$ equals

$$\frac{\mathrm{Leb}(\, A_n(x_\alpha) \cap A_n(x_\beta)\,)}{\min\,(\, \mathrm{Leb}(A_n(\alpha)),\, \mathrm{Leb}(A_n(\beta))\,)}. \qquad (46)$$

In other words, (46) equals the fraction the intersection $A_n(x_\alpha) \cap A_n(x_\beta)$ takes up of the smaller of the two single sets. The O-term in (45) stems from the fact that f is only locally constant. In the following we will expand the term in (46) in order to finally show that $\mathrm{Cov}(\,\Psi_s(X_i),\, \Psi_t(X_i)\,)$ converges pointwise.

First observe that $d_n(\alpha)$ is differentiable in α and that

$$\sup_{0 \leq \alpha \leq 1} |d_n'(\alpha)| = O\big(\; r^{1/p} \,/\, \|x_n - y_n\|\big). \qquad (47)$$

Consider the function $H(\alpha, u) = F(B(x_\alpha, u))$. Our assumptions imply that H is differentiable and that $H(\alpha, d_n(x_\alpha)) = r$. Differentiating both side of the latter equation leads to an equation for $\frac{d}{d\alpha} d_n(x_\alpha) = \frac{d_n'(\alpha)}{\|x-y\|}$. Our assumptions then imply the result (47). Further details are omitted.

By writing $d_n(\beta) = d_n(\alpha) + d_n'(\zeta)\,(\beta - \alpha)$ for some ζ between α and β we now write (46) as

$$\begin{aligned}
&\frac{2\,d_n(\alpha) + d_n'(\zeta)\,(\beta-\alpha) - |\alpha-\beta|}{2\,(\,d_n(\alpha) + \min(0,\, d_n'(\zeta)\,(\beta-\alpha)\,)\,)} \\
&\qquad = \Big(1 - \frac{|\alpha-\beta|}{2\,d_n(\alpha)} + \frac{d_n'(\zeta)\,(\beta-\alpha)}{2\,d_n(\alpha)}\Big)\Big(1 - \xi\,\frac{\min(0,\, d_n'(\zeta)\,(\beta-\alpha)\,)}{d_n(\alpha)}\Big) \\
&\qquad = \Big(1 - \frac{|\alpha-\beta|}{2\,d_n(\alpha)}\Big) + \Xi_n(\alpha,\,\beta)
\end{aligned} \tag{48}$$

where $\Xi_n(\alpha,\,\beta)$ is defined by the last equality. ξ is of the form $\frac{1}{(1+x)^2}$ for some x between 1 and $1 + \frac{\min(0,\, d_n'(\zeta)\,(\beta-\alpha)\,)}{d_n(\alpha)}$, and hence $|\xi|$ is bounded uniformly in $\alpha, \beta \in [0,1]$. Thus we can write

$$\begin{aligned}
&\Big(\frac{d_n(\alpha) + d_n(\beta) - |\alpha-\beta|}{2\,\min(d_n(\alpha),\, d_n(\beta)\,)}\Big)^+ \\
&= \Big(\frac{d_n(\alpha) + d_n(\beta) - |\alpha-\beta|}{2\,\min(d_n(\alpha),\, d_n(\beta)\,)}\Big)\,\mathbf{1}\{\,|\alpha-\beta| \le d_n(\alpha) + d_n(\beta)\,\} \\
&= \Big(1 - \frac{|\alpha-\beta|}{2\,d_n(\alpha)}\Big)\,\mathbf{1}\{\,|\alpha-\beta| \le d_n(\alpha) + d_n(\beta)\,\} \\
&\qquad\qquad + \,\Xi_n(\alpha,\,\beta)\,\mathbf{1}\{\,|\alpha-\beta| \le d_n(\alpha) + d_n(\beta)\,\}.
\end{aligned} \tag{49}$$

Notice further that by assumption on r we have $\sup_{\alpha\in[0,1]} |\,d_n'(\alpha)\,| = o(1)$ as $n \to \infty$. It follows that $|\,\mathbf{1}\{\,|\alpha-\beta| \le d_n(\alpha) + d_n(\beta)\,\} - \mathbf{1}\{\,|\alpha-\beta| \le 2\,d_n(\alpha)\,\}\,| \le \mathbf{1}\{\,(2 \mp c_n)\,d_n(\alpha) \le |\alpha-\beta| \le (2 \pm c_n)\,d_n(\alpha)\,\}$ with $0 < c_n = o(1)$. A straightforward calculation now shows that

$$\int_0^t \int_0^s \Big(1 - \frac{|\alpha-\beta|}{2\,d_n(\alpha)} \Big) \mathbf{1}\{\, |\alpha-\beta| \le d_n(\alpha) + d_n(\beta) \,\}\; d\,\alpha\, d\,\beta$$

$$= \int_0^t \int_0^s \Big(1 - \frac{|\alpha-\beta|}{2\,d_n(\alpha)} \Big) \mathbf{1}\{\, |\alpha-\beta| \le 2\,d_n(\alpha) \,\}\; d\,\alpha\, d\,\beta \;\; (1+o(1))$$

$$= \int_0^{\min(s,t)} \big(2\,d_n(\alpha) + O\big(\, d_n^2(\alpha) \,\big) \big)\;\, d\,\alpha \;\; (1+o(1)). \tag{50}$$

Further, on the set $\{\, |\alpha - \beta| \le 2\; d_n(\alpha) \,\}$ we have $\sup_{\alpha,\,\beta} |\, \Xi_n(\alpha,\, \beta)\,| = O(\, \sup_{0\le\alpha\le 1} |d_n'(\alpha)|) = O\big(\;\, r^{1/p} \,/\, \|x_n - y_n\|\big) = o(1)$. Hence we also have

$$\int_0^t \int_0^s \Xi_n(\alpha,\, \beta)\, \mathbf{1}\{\, |\alpha-\beta| \le d_n(\alpha) + d_n(\beta) \,\}\; d\,\alpha\, d\,\beta$$

$$= O(\, r^{\frac{1}{p}} \,/\, \|x_n - y_n\|\,) \; \int_0^t \int_0^s \mathbf{1}\{\, |\alpha-\beta| \le 2\,d_n(\alpha) \,\}\; d\,\alpha\, d\,\beta \;\; (1+o(1))$$

$$= o(1) \; \int_0^{\min(s,t)} \big(2\,d_n(\alpha) + O\big(\, d_n^2(\alpha) \,\big) \big)\;\, d\,\alpha \;\; (1+o(1)). \tag{51}$$

The sum of (50) and (51) equals (46). Hence, by using (32) and again using the fact that by assumption $r^{\frac{1}{p}}/\|x_n - y_n\| = o(1)$ we obtain

$$\mathrm{E}(\, \Psi_s(X_i) \,\cdot\, \Psi_t(X_i)\,) = \int_0^s \int_0^t F(A_n(x_{n,\alpha}) \,\cap\, A_n(x_{n,\beta}))\; d\,\alpha\; d\,\beta$$

$$= 2\, \frac{r}{\|x_n - y_n\|} \Big(\int_0^{\min(s,t)} d_n(x_{n,\alpha}) \;+\; O\left(\frac{d_n^2(x_{n,\alpha})}{\|x_n - y_n\|} \right) \; d\,\alpha \Big) \,(\, 1 + o(1)\,)$$

$$= \;2\, \frac{r^{1+\frac{1}{p}}}{\|x_n - y_n\|} \int_0^{\min(s,t)} \Big(\frac{1}{C(p)\, f(x_\alpha)} \Big)^{\frac{1}{p}} \; d\,\alpha \;(\, 1 + o(1)\,) \tag{52}$$

and hence

$$\mathrm{Cov}(\, h_{s,n}(X_i),\, h_{t,n}(X_i)\,)$$

$$= \;2 \int_0^{\min(s,t)} \Big(\frac{1}{C(p)\, f(x_\alpha)} \Big)^{\frac{1}{p}} \, d\,\alpha \,(\, 1 + o(1)\,) - \|x_n - y_n\|\, r^{1-\frac{1}{p}}\, s\, t. \tag{53}$$

This completes the proof.

□

For the next result recall the definition of $\widehat{Z}_n(t, C)$ and $Z_n(t, C)$, $C \in \mathcal{I}$, given in (12) and (13), respectively.

Proposition 6.6 *Let $X_1, X_2, \ldots$ be i.i.d. vectors in $\mathbf{R}^p$ with a common density f. Suppose that $I = \{\alpha : SB(\alpha) \geq 0\} =: [l, u]$, that A1 and A2 hold, and that g satisfies (10). Further let r be such that $r \to 0$, $n\, r^{1+\frac{3}{p}} \to 0$, and $\frac{n\, r^{1+\frac{2}{p}}}{(\log n)^3} \to \infty$. Also assume that $\epsilon_n \to 0$, and with $\widehat{I}_n(\epsilon_n) =: [\widehat{l}_n, \widehat{u}_n]$ let $\widehat{\ell}_n(\epsilon_n) = \widehat{u}_n - \widehat{l}_n$. Then we have as $n \to \infty$*

$$\sqrt{\frac{n\, r^{1-\frac{1}{p}}\, \|\, x - y\, \|}{\widehat{\ell}_n(\epsilon_n)}} \sup_{t \in \widehat{I}_n(\epsilon_n)} \left| (\widehat{Z}_n - Z_n)(t, \widehat{I}_n(\epsilon_n)) \right| \to_{\mathcal{D}} \sup_{t \in [0,1]} |G_F(t)|.$$

Here $\{\, G_F(t), t \in [0, 1]\,\}$ denotes a mean zero Gaussian process with

$$\mathrm{Cov}(G_F(t), G_F(s)) = D(\min(t, s)) - t\, D(s) - s\, D(t) + s\, t\, D(1),$$

where $D(t) = 2\int_0^t \left(\frac{1}{f(z_\gamma)\, C(p)}\right)^{\frac{1}{p}}\, d\gamma$, and $\{\, z_\gamma,\ \alpha \in [0, 1]\,\}$ is a linear parametrization of I.

Proof. We use the notation $I(\epsilon_n) = [l_n, u_n]$, and $\ell = u - l$. The idea of the proof is as follows. Assume that

$$\sqrt{\frac{n\, r^{1-\frac{1}{p}}\, \|x_n - y_n\|}{\widehat{\ell}_n(\epsilon_n)}} \Big(\sup_{t \in [0,1]} |\, (\widehat{Z}_n - Z_n)(t, \widehat{I}_n(\epsilon_n))\,| - \sup_{t \in [0,1]} \Big|\, (\widehat{Z}_n - Z_n)(t, I)\, \Big| \Big) = o_P(1), \tag{54}$$

and that consequently we only need to consider the non-random interval I. (We will show (54) below.) For each $t \in I(\epsilon_n)$ let

$$W_n(t) := \int_l^t \frac{(F_n - F)(A_n(x_\alpha))}{r}\, d\alpha \; - \frac{t - l}{u - l} \int_l^u \frac{(F_n - F)(A_n(x_\alpha))}{r}\, d\alpha.$$

It will be proven that for $t \in I$

$$(\widehat{Z}_n - Z_n)(t, I) \;=\; W_n(t) \;+\; R_n, \tag{55}$$

where $R_n = o_P\Big(\sqrt{\frac{\widehat{\ell}_n(\epsilon_n)}{n\,r^{1-1/p}}}\Big)$ uniformly in $t \in I$, and hence the asymptotic behavior of $\sqrt{\frac{n\,r^{1-1/p}}{\widehat{\ell}_n(\epsilon_n)}}\ \sup_{t\in I(\epsilon_n)} |\,W_n(t)\,|$ is the same as the one of our target quantity.

The asymptotic distribution of $W_n(t)$ can be obtained from Theorem 6.5 as follows. First rescale again. Let $z_\gamma = x_l - \gamma\,(x_u - x_l)$, where as throughout this paper x_l and x_u denote the endpoints of the interval $\{\,x_\alpha,\ \alpha \in [l,u]\,\}$, where x_α denotes our original parametrization. For $t \in [0,1]$ define

$$\widetilde{W}_n(t) = \int_0^t \frac{(F_n - F)(A_n(z_\gamma))}{r}\, d\gamma\ - t\ \int_0^1 \frac{(F_n - F)(A_n(z_\gamma))}{r}\, d\gamma. \tag{56}$$

Then for $t \in [\,l,\,u\,]$ that $\frac{1}{u-l}\,W_n(t) = \widetilde{W}_n\big(\frac{t-l}{u-l}\big)$, and hence $\sup_{t\in[0,1]} |\,W_n(t)\,| = \sup_{t\in I} |\,W_n(t)\,| = (u-l)\ \sup_{t\in[0,1]} |\,\widetilde{W}_n(t)\,|$.

For $t \in [0,1]$ let $\Psi_t(x) = \int_0^t \mathbf{1}(x \in A_n(z_\gamma))\, d\,\gamma$. Define

$$\nu_n(\Psi_t) = \sqrt{\frac{n\,\|x-y\|\,\ell}{r^{1+\frac{1}{p}}}}\,(F_n - F)(\Psi_t) = \sqrt{\frac{n\,\|x-y\|\,\ell}{r^{1+\frac{1}{p}}}}\,\frac{1}{n}\sum_{j=1}^n \left[\Psi_t(X_i) - \mathrm{E}\,\Psi_t(X_i)\right],$$

and for $t \in [\,l,\,u\,]$ let $t' = \frac{t-l}{u-l}$. Using this notation we have for $t \in [\,l,\,u\,]$

$$\sqrt{\frac{n\,r^{1-\frac{1}{p}}\,\|\,x-y\,\|}{\widehat{\ell}_n(\epsilon_n)}}\ W_n(t) = \sqrt{\frac{n\,r^{1-\frac{1}{p}}\,\ell_n(\epsilon_n)^2\,\|x-y\|}{\widehat{\ell}_n(\epsilon_n)}}\ \ W_n'(t') \tag{57}$$

$$= [\,\nu_n(\Psi_{t'}) - t'\,\nu_n(\Psi_1)\,]\,(1 + o_P(1)). \tag{58}$$

The last equality uses the fact that Lemma 6.4 implies that $\big|\,\frac{\widehat{\ell}_n(\epsilon_n)}{u-l} - 1\,\big| = o_P(1)$ as $n \to \infty$. Hence the behavior of $\nu_n(\Psi_{t'})$ determines the behavior of W_n. The former is studied in Theorem 6.5, and an application of this result implies the asserted asymptotic distribution via an application of the continuous mapping theorem. Calculation of the covariance function of the limit process is straightforward.

Proof of (54), derivation of (55), and estimation of the remainder term in (55).

Fist we show (55). By utilizing (24) we have for $t \in I$ that

$$
\begin{aligned}
&(\widehat{Z}_n - Z_n)(t, I) \\
&= \int_l^t \Big[p \, \log \widehat{d}_n(x_\alpha) - \log \tfrac{r}{C(p)\, f(x_\alpha)}\Big] \; d\,\alpha - \frac{t-l}{u_n - l} \int_l^u \Big[p \, \log \widehat{d}_n(x_\alpha) - \log \tfrac{r}{C(p)\, f(x_\alpha)}\Big] \; d\,\alpha \\
&= \int_l^t \frac{(F_n - F)(A_n(x_\alpha))}{r} \; d\,\alpha - \frac{t-l}{u-l} \int_l^u \frac{(F_n - F)(A_n(x_\alpha))}{r} \, d\alpha \qquad (59) \\
&\qquad + \int_l^t \frac{\Delta_n(x_\alpha)}{r} \, d\alpha \; - \frac{t-l}{u-l} \int_l^u \frac{\Delta_n(x_\alpha)}{r} \, d\alpha \qquad (60) \\
&\qquad + \ell \, O_P\Big(r^{2/p} \sqrt{\frac{\log n}{n\, r}} \; \Big) + \; \ell \, O_P\Big(\Big(\frac{\log n}{n\, r} \Big)^{3/4} \Big) + \ell O\big(r^{\frac{4}{p}}\,\big), \qquad (61)
\end{aligned}
$$

where up to a constant the O_P-terms are the ones from (24). Notice that the term in (59) equals $W_n(t)$. We thus have shown (55) with

$$
R_n \;\; = \;\; R_{1n} \;\; + \;\; R_{2n}
$$

with R_{1n} and R_{2n} denoting the expressions from (60) and (61), respectively. Using the fact that under the present assumptions $\sup_{\alpha \in [0,1]} \Delta_n(x_\alpha) \; = \; O_P\left(r^{1+\frac{2}{p}}\right)$ (see Lemma 6.2), our assumptions on r immediately imply that $\sqrt{\frac{n\, r^{1-1/p}}{\ell}} \; (R_{1n} + R_{2n}) = o_P(1)$. This proves (55).

It remains to prove (54). We have

$$
\begin{aligned}
&\Big| \, (\widehat{Z}_n - Z_n)(t, \widehat{I}_n(\epsilon_n)) \; - \; (\widehat{Z}_n - Z_n)(t, I) \, \Big| \\
&\leq \left| \int_{(\,\widehat{I}_n(\epsilon_n) \setminus I\,) \cap [0,t]} - \int_{(\,I \setminus \widehat{I}_n(\epsilon_n)\,) \cap [0,t]} \Big[\, p \, \log \widehat{d}_n(x_\alpha) - \log \tfrac{r}{C(p)\, f(x_\alpha)} \,\Big] \; d\,\alpha \right| \qquad (62) \\
&\qquad + \frac{|\, \widehat{I}_n(\epsilon_n) \, \Delta \, I \,|}{|\, \widehat{I}_n(\epsilon_n) \,|} \left| \int_{\widehat{I}_n(\epsilon_n) \setminus I} - \int_{I) \setminus \widehat{I}_n(\epsilon_n)} \Big[p \, \log \widehat{d}_n(x_\alpha) - \log \tfrac{r}{C(p)\, f(x_\alpha)}\Big] \; d\,\alpha \right|. \qquad (63)
\end{aligned}
$$

Similar arguments as in the proof of (55) now show that in (62) we can replace $p \, \log \widehat{d}_n(x_\alpha) - \log \frac{r}{C(p)\, f(x_\alpha)}$ by $\frac{(F_n - F)(A_n(x_\alpha))}{r}$ with the error being negligible for our purposes. Writing $I_n = [\widehat{l}_n, \widehat{u}_n]$ and $I = [l, u]$, both the integral terms in (62) and (63) can be estimated by

$$\sqrt{\frac{\widehat{\ell}_n(\epsilon_n)}{nr^{1-\frac{1}{p}}\,\|x-y\|}}\ \big[\,|\nu_n(\Psi_{\widehat{u}_n}) - \nu_n(\Psi_u)| + |\nu_n(\Psi_{\widehat{l}_n}) - \nu_n(\Psi_l)|\,\big].$$

Since $\widehat{u}_n \to u$ and $\widehat{l}_n \to l$ in probability (Lemma 6.4) stochastic equicontinuity of $\{\nu_n(\Psi_t), t \in [0,1]\}$ (Theorem 6.5) essentially implies that both (62) and (63) are $o_P(\sqrt{\frac{\widehat{\ell}_n(\epsilon_n)}{nr^{1-\frac{1}{p}}}})$. (To be precise, since $|\,\frac{\widehat{u}_n - \widehat{l}_n}{u-l} - 1\,| = o_P(1)$ find an interval $[a_n, b_n] \subset [0,1]$ with $|b_n - a_n| = O(|u-l|)$ such that for large n with high probability both $\widehat{I}_n(\epsilon_n)$ and I are subsets of $[a_n, b_n]$. Then consider the process $\nu_n(\Psi_t)$ on $t \in [a_n, b_n]$ and apply Lemma 6.5).)

□

7 References

Alexander, K.S. (1984): Probability inequalities for empirical processes and a law of the iterated logarithm. *Ann. Probab.* **12** 1041-1067; (Correction: **15** 428-430)

Bai, Z.D. and Huang, L.D. (1999): On consistency of the best-r-points-average estimator for the location for the maximizer of a nonparametric regression function. *Sankhya: Indian J. Statist, Ser. A* **61** 208 - 217

Bai, Z., Chen, Z. and Wu, Y. (2003): Convergence rate of the best-r-point-average estimator for the maximizer of a nonparametric regression function. *J. Multivariate Analysis* **84** 319 - 334

Burman, P. and Nolan, D. (1992) Location-adaptive density estimation and nearest neighbor distance. *Journal of Multivariate Analysis* 40:132-157

Cheng, M.-Y. and Hall, P. (1999): Mode testing in difficult cases. *Ann. Statist.* **27** 1294 - 1315

Comparini, A. and Gori, E. (1986): Estimating modes and antimodes of multimodal densities. *Metron.* **44** 307-332

Corti, S., Molteni F. and Palmer P.N. (1999): Signature of recent climate change in frequencies of natural atmospheric circulation regimes. *Nature* **398** 799 - 802

Cuevas, A. Febrero, M. and Fraiman, R. (2000): Estimating the numbers of clusters. *Canad. J. Statist.* **28** 367-382.

Cuevas, A., Gonzáles-Manteiga, W., Rodriguez-Casal, A. (2004): Plug in estimation of general level sets. Manuscript

Dalenius, T. (1965): The mode: a neglected statistical parameter. *J. Roy. Statist. Soc. Ser. A* **128** 110-117

Fisher, N.I., and Marron, J.S. (1999): Mode testing via the excess mass estimate. Unpublished

Friedman, J. and Fisher, N.I. (1999): Bump hunting in high-dimensional data (with discussion). *Computing and Statistics* **9** 123 - 143

Good, I.J. and Gaskins, R.A. (1980): Density estimation and bump-hunting by the penalized maximum likelihood method exemplified by scattering and meteorite data. (with discussion), *J. Amer. Statist. Assoc.* **75** 42-73

Haldane, J.B.S. (1952): Simple tests for bimodality and bitangentiality. *Ann. Eugenics* **14** 309-318

Hall, P. and Cheng, M.Y. (1999): Mode testing in difficult cases. *Ann. Statist.* **27** 1294-1315

Hartigan, J.A. (1975): *Clustering Algorithms.* Wiley, New York

Hartigan, J.A., and Hartigan, P.M. (1985): The dip test of unimodality. *Ann. Statist.* **13** 70-84

Hartigan, J.A. (1987): Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.* **82** 267-270

Hartigan, J.A. (1988): The span test for unimodality. In *Classification and related methods of data analysis* (H.H. Bock, ed.) 229 - 236. North-Holland, Amsterdam

Hartigan, J.A., and Mohatny, S. (1992): The RUNT test for multimodality. *J. Classification* **9** 115 - 132

Jiang, W. (2006); Nonparametric density estimation and clustering in astronomical sky surveys. *Comp. Statist. & Data Analysis* **50** 760-774.

Lientz, B.P. (1970). Results on nonparametric modal intervals. *SIAM J. Appl. Math.* **19** 356-366

Marchette, D.J. and Wegmann, E.J. (1997). The filtered mode tree. *J. Comput. Graph. Statist.* **6** 143-159

Minotte, M.C. (1993): A test of mode existence with applications to multimodality. Dissertation. Rice University

Minotte, M.C. (1997): Nonparametric testing of the existence of modes. *Ann. Statist.* **25** 1646 - 1660

Minnotte, M.C., Marchette, D.J., Wegman, E.J. (1998): The bumpy road to the mode forest. *J. Comput. Graph. Statist.* **7** 239-251

Minotte, M.C. and Scott, D.W. (1993). The mode tree. *J. Comp. Graph. Statist.* **2** 51-68

Molchanov, I. (1998): A limit theorem for solutions of inequalities. *Scand. J. Statist.* **25** 235 - 242

Müller, D.W., and Sawitzki, G. (1991): Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.* **86**, 738-746

Müller, H.-G. (1989): Adaptive nonparametric peak estimation. *Ann. Statist.* **17** 1053 - 1069

Polonik, W. (1995): Measuring mass concentration and estimating density contour clusters: an excess mass approach. *Ann. Statist.* **23** 855-881

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

Polonik, W. and Wang, Z. (2004): Estimation of regression contour clusters: An application of the excess mass approach to regression. To appear in *J. Multivariate Anal.*

Polonik, W. and Wang, Z. (2006): PRIM Analysis. Submitted.

Roeder, K. (1994): A graphical technique for determining the number of components in a mixture of normals. *J. Amer. Statist. Assoc.* **89** 487 - 495

Rozal, P.M., and Hartigan, J.A. (1994): The MAP test for multimodality. *J. Classification* **11** 5- 36

Silverman, B. (1981): Using kernel density estimates to investigate multimodality. *J. Royal. Statist. Soc., Series B* **43** 97-99

Tibshirani, R. Walther, G. and Hastie, T. (2001): Estimating the number of clusters in a dataset via the Gap statistics. *J. Royal. Statist. Soc., Ser. B* **63** 411 - 423

Tsybakov, A.B. (1997): On nonparametric estimation of level sets. *Ann. Statist.* **25** 948 - 969

van de Geer, S. (2000): *Empirical processes in M-estimation.* Cambridge University Press

van der Vaart, A.W. and Wellner, J.A. (1996): *Weak convergence and empirical processes.* Springer, New York.

Venter, J.H. (1967): On estimation of the mode. *Ann. Math. Statist.* **38** 1446-1455

Walther, G. (1997): Granulometric smoothing. *Ann. Statist.* **25** 2273 - 2299

Walther, G. (2001): Multiscale maximum likelihood analysis of a semiparametric model, with applications. *Ann. Statist.* **29** 1297-1319

Wong, M.A. (1985): A bootstrap testing procedure for investigating the number of subpopulations. *J. Statist. Comp. Simul.* **22** 99-112

Wong, M.A. and Schaack, C. (1985): Diagnostics for assessing multimodality. In *Computer science and statistics.* Proceedings of the sixteenth symposium on the interface. L. Billard, ed., 287-296