Measuring mass concentrations and estimating density contour clusters - an excess mass approach

by Wolfgang Polonik

Universität Heidelberg

By using empirical process theory we study a method addressed to testing for multimodality and estimating density contour clusters in higher dimensions. The method is based on the so-called excess mass over $\mathbb{C}$. Given a probability measure F and a class of sets $\mathbb{C}$ in the d-dimensional Euclidean space, the excess mass over $\mathbb{C}$ at a level $\lambda \geq 0$, denoted by $E_{\mathbb{C}}(\lambda)$, is defined as the maximal difference between the F-measure and λ times the Lebesgue measure of sets in $\mathbb{C}$. $E_{\mathbb{C}}(\lambda)$ can be estimated by replacing F by the empirical measure. Those sets which maximize the corresponding difference of empirical measure and λ times the Lebesgue measure over sets in $\mathbb{C}$ can be used for estimating density contour clusters. Comparing excess masses over different classes $\mathbb{C}$ yields information about the modality of the underlying probability measure. This can be used to construct tests for multimodality. The asymptotic behaviour of the considered estimators and test statistics is studied for general classes $\mathbb{C}$, including the classes of balls, ellipsoids and convex sets.

1. Introduction

The excess mass approach which will be studied in this paper by means of empirical process theory has first been considered independently by Müller and Sawitzki (1987) and Hartigan (1987). It yields a method for testing for multimodality and estimating density contour clusters in higher dimensions. For a distribution F on $\mathbf{R}^d$ with Lebesgue density f the *density contour cluster (of f) at a level $\lambda \geq 0$* is defined as the set $C(\lambda) = C_f(\lambda) = \{x: f(x) \geq \lambda\}$. Note that Hartigan (1975) used the notion density contour cluster for the connected components of $C(\lambda)$, whereas here the sets $C(\lambda)$ need not be connected. Müller and Sawitzki defined the *excess mass functional* as $\lambda \rightarrow E(\lambda) = F(C(\lambda)) - \lambda\, \mathrm{Leb}(C(\lambda))$, where Leb denotes the Lebesgue measure (cf. Fig. 2.1

Research supported by the Sonderforschungsbereich 123 and the Deutsche Forschungsgemeinschaft

American Mathematical Society 1990 *subject classifications.* Primary 62G; secondary 62H

Key words and phrases. Excess mass, density contour cluster, multimodality, empirical process theory, support estimation, convex hull

below, which motivates the name “excess mass”). Define H_λ = F - λ Leb, then the excess mass functional which defines a concentration function can be rewritten as E(λ) = sup{H_λ(C), C $\subset$ $\mathbf{R}^d$ measurable}. Replacing F by the empirical measure F_n of n i.i.d. observations $X_1,...,X_n$ drawn from F leads to an estimator of the excess mass functional. However, the supremum of $H_{n,\lambda}(C)$ = $F_n(C)$ - λ Leb(C) over all measurable sets equals 1 (take C = {$X_1,...,X_n$}). Hence one has to restrict the class of sets over which the supremum is extended. Let $\mathbb{C}$ denote a class of measurable subsets of $\mathbf{R}^d$. Generalizing the excess mass functional we define the *excess mass over* $\mathbb{C}$ *at a level* $\lambda \geq 0$ by

$$E_{\mathbb{C}}(\lambda) \;=\; sup\, \{\, H_\lambda(C) : C \in \mathbb{C} \,\},$$

Those sets which maximize H_λ over the class $\mathbb{C}$ are called *generalized λ-clusters in* $\mathbb{C}$, i.e. every set $\Gamma_{\mathbb{C}}(\lambda) \in \mathbb{C}$ with

$$E_{\mathbb{C}}(\lambda) = H_\lambda(\Gamma_{\mathbb{C}}(\lambda))$$

is a generalized λ-cluster in $\mathbb{C}$. We always have $E_{\mathbb{C}}(\lambda) \leq E(\lambda)$ and if $C(\lambda) \in \mathbb{C}$, then C($\lambda$) is a generalized λ-cluster and $E_{\mathbb{C}}(\lambda) = E(\lambda)$. Differences of excess masses over different classes yield information about modality (Müller and Sawitzki (1987) and Hartigan (1987)). To see this consider the following univariate situation: Assume that F is a (sufficiently smooth) symmetric distribution on the real line such that F has exactly m modes. In this case the density contour clusters C(λ) all lie in $\mathcal{I}_m$, m $\in$ $\mathbf{N}$, the class of unions of at most m intervals. Hence, $E(\lambda) = E_{\mathcal{I}_m}(\lambda)$ for all $\lambda \geq 0$, and for any k < m there exist levels λ with $E(\lambda) > E_{\mathcal{I}_k}(\lambda)$. Therefore the maximal difference $E_{\mathcal{I}_m}(\lambda) - E_{\mathcal{I}_k}(\lambda)$ is strictly bigger than zero for all k < m and is equal to zero for k $\geq$ m.

The excess mass over $\mathbb{C}$ and the generalized λ-clusters can be estimated from the data. The corresponding estimators, $E_{n,\mathbb{C}}(\lambda)$ and $\Gamma_{n,\mathbb{C}}(\lambda)$, respectively, will be defined by analogy to the corresponding theoretical quantities by replacing F by the empirical distribution of n i.i.d. observations (Section 2). $\Gamma_{n,\mathbb{C}}(\lambda)$ defines an estimator for the density contour clusters and the maximal difference of $E_{n,\mathbb{D}}(\lambda) - E_{n,\mathbb{C}}(\lambda)$ over appropriately chosen classes $\mathbb{C}$ and $\mathbb{D}$ will be used as a test statistic for multimodality.

In this paper we do not restrict ourselves to specific classes $\mathbb{C}$ and $\mathbb{D}$, respectively. Using empirical process theory we study how the asymptotic properties of the considered empirical quantities depend on the classes under consideration. All the standard classes, such as the classes of balls, ellipsoids or convex set, are included in our study. The asymptotic results can be used as hints on how to choose appropriate classes for special problems. One will have to balance between the richness of the model (which means richness of the classes under consideration, see below), desirable statistical properties and the time needed for calculation.

The assumption $C(\lambda) \in \mathbb{C}$ for all $\lambda \geq 0$, or, for short, the choice of a class $\mathbb{C}$, may be interpreted as the choice of a nonparametric statistical model: the class of all distributions dominated by Lebesgue measure whose density contour clusters lie in $\mathbb{C}$. In contrast to defining models through smoothness assumptions on the density it is possible to model certain *qualitative* aspects, such as modality, of the underlying distribution through appropriate choices of $\mathbb{C}$. As already mentioned, the class $\mathcal{I}_m$, $m \in \mathbf{N}$, of unions of at most m intervals, corresponds to a onedimensional distribution with at most m modes. Hence, the assumption "$C(\lambda) \in \mathcal{I}_m$ for all $\lambda \geq 0$" defines the model of all univariate distributions with at most m modes. Below we also give multivariate analogs.

This one-dimensional setup has been considered in Müller and Sawitzki. Hartigan considered the two-dimensional case. In our terminology he used the excess mass over the class of closed convex sets in $\mathbf{R}^2$, denoted by $\mathcal{C}^2$, and compared it with the excess mass over those convex sets lying exterior to $\Gamma_{n,\mathcal{C}^2}(\lambda)$. In a more parametric setup Nolan (1991) considered the case $\mathbb{C} = \mathcal{E}^d$, the class of all closed ellipsoids in $\mathbf{R}^d$. In all these papers it is assumed that the underlying distribution has density contour clusters $C(\lambda)$ lying in the class $\mathbb{C}$ under consideration.

The density contour clusters $C(\lambda)$ themselves contain information about the location of mass concentration. If $C(\lambda) \in \mathbb{C}$, or in other words, if the chosen model which corresponds to the choice of $\mathbb{C}$ is correct, then the sets $C(\lambda)$ can be estimated from the data. This could also be done by first estimating the density by a kernel estimator and then estimating the density contour clusters by the corresponding density contour clusters of the kernel estimator. The resulting estimator will be consistent under appropriate smoothness assumptions. However, the kernel estimator approach does not allow to enclose a prior knowledge about the shape of the density contour clusters (such as convexity). Furthermore, although one never knows in practice that the density contour clusters lie in $\mathbb{C}$, the interpretation of the empirical generalized λ-clusters $\Gamma_{n,\mathbb{C}}(\lambda)$ as sets maximizing the excess mass still holds and therefore they might contain useful information even for finite n.

Our paper is organized as follows: The behaviour of $E_{n,\mathbb{C}}(\cdot)$ is studied in Section 2 and Section 4. We show that $E_{n,\mathbb{C}}(\cdot)$ is a consistent estimator for $E_{\mathbb{C}}(\cdot)$ and prove asymptotic normality. In the case where $C(\lambda) \in \mathbb{C}$ and "f has no flat part", i.e. $F\{x: f(x) = \lambda\} = 0 \ \forall \ \lambda \geq 0$, the limit process is a Brownian Bridge with transformed time scale (Theorem 4.3). The asymptotic behaviour of the sets $\Gamma_{n,\mathbb{C}}(\lambda)$ is studied in Section 3. We show the consistency of $\Gamma_{n,\mathbb{C}}(\lambda)$ as an estimator of $\Gamma_{\mathbb{C}}(\lambda)$ (Theorem 3.2 and Theorem 3.5) and in the case where $C(\lambda) \in \mathbb{C}$ we also give rates of convergence (Theorems 3.6 and 3.7). As a special case ($\lambda = 0$ and $\mathbb{C} = \mathbf{C}^d$) we obtain rates of convergence for the convex hull of the sample as a by-product (Proposition 3.8). In Section 5 we address questions related to testing. Given two nested classes $\mathbb{C}$ and $\mathbb{D}$, the maximal difference of the empirical excess masses over $\mathbb{C}$ and $\mathbb{D}$, i.e. $\sup_{\lambda\geq 0}(E_{n,\mathbb{D}}(\lambda) - E_{n,\mathbb{C}}(\lambda))$, can be used for testing the hypothesis that the density contour clusters lie in the smaller classes $\mathbb{C}$ against the alternative that they lie in $\mathbb{D}\backslash\mathbb{C}$. For special choices of $\mathbb{C}$ and $\mathbb{D}$ this leads to tests for unimodality, as proposed by Müller and Sawitzki and by Hartigan (see above). The asymptotic distribution of the maximal difference of the empirical excess masses over $\mathbb{C}$ and $\mathbb{D}$ is known only for the special case of an underlying uniform distribution (Theorem 5.4). However, under the null-hypotheses we derive rates of convergence for general F (Theorem 5.2, Theorem 5.3) which in some special univariate situations are known to be the exact rates. Section 6 contains the proofs of all the results given in the previous sections.

We close the introduction by giving some related work from the literature. There exist some other nonparametric approaches to measuring mass concentrations and investigating the modality of the underlying distribution in the literature, which also are based on the idea of comparing the volume of a certain class of sets with the mass carried by these sets:

In a fundamental paper Chernoff (1964) considered (in the univariate case) the midpoint x of an interval with given length *l* which carries maximal mass among all intervals with the same length *l*. If *l* tends to zero and the distribution is dominated by the Lebesgue measure, then (in regular cases) x converges to the mode of the density. However, if *l* is not too small, the midpoint indicates a location around which a non-negligible portion of the mass is concentrated. Considered as a function of *l* the maximal mass $\alpha = \alpha(l)$ becomes the well known concentration function.

Alternatively, one can consider the inverse problem: Fix the mass α and ask for the interval with minimal length among all intervals carrying (at least) mass α. Such intervals are called minimal volume intervals or modal intervals (cf. Lientz (1970), Andrews et al. (1972), Robertson and

Cryer (1974), Grübel (1988)).

In these one-dimensional situations it is a "natural" decision to use intervals. However, strictly speaking, the choice of intervals is natural only if the underlying distribution has a unimodal Lebesgue density, f. For in this case (under some regularity conditions) the density contour clusters of f are intervals and maximize the (theoretical) functions in the procedures given above. This corresponds to the situation “$C(\lambda) \in \mathbb{C}$” in the context of excess masses (see above).

For generalizing the above mentioned procedures to higher dimensions there is no "natural" choice of a class of sets, even in the unimodal case. One might for example use the classes of all balls, ellipsoids or convex sets. Sager (1979) for example generalized the method of Robertson and Cryer by replacing the class of intervals by the class of convex sets. The problem of how to choose an appropriate class $\mathbb{C}$ (especially in higher dimensions) of course also exists in the excess mass approach. However, as mentioned earlier, since the results in the present paper are given for an unspecified class $\mathbb{C}$, they can be used as hints on how to choose $\mathbb{C}$ in a special problem.

2. The empirical excess mass over $\mathbb{C}$

Let (Ω, P) denote the underlying probability space and let $X_1, X_2,..,X_n,..$ be i.i.d. random vectors in $\mathbf{R}^d$ with distribution F. In order to obtain an estimator of the excess mass over $\mathbb{C}$ we replace the unknown distribution F by F_n, the empirical distribution of $X_1,..,X_n$. This leads to the *empirical* excess mass over $\mathbb{C}$, defined by

$$E_{n,\mathbb{C}}(\lambda) := \sup \{ F_n(C) - \lambda \, Leb(C) : C \in \mathbb{C} \}, \quad \lambda \geq 0.$$

Let $H_{n,\lambda} = F_n - \lambda \, Leb$, $\lambda \geq 0$. A set $\Gamma_{n,\mathbb{C}}(\lambda) \in \mathbb{C}$ such that

$$E_{n,\mathbb{C}}(\lambda) = H_{n,\lambda}(\Gamma_{n,\mathbb{C}}(\lambda)),$$

is called *empirical generalized* λ*-cluster.*

F i g u r e 2.1

Since the "excess" $E_{n,\mathbb{C}}(\lambda)$ (and also $E_{\mathbb{C}}(\lambda)$) should be non-negative we always assume that $\emptyset \in \mathbb{C}$. In the following proposition some elementary properties of $E_{n,\mathbb{C}}$ are summarized.

Proposition 2.1: *Let $\emptyset \in \mathbb{C}$. Then we have:*

(i) $0 \leq E_{n,\mathbb{C}}(\lambda) \leq 1$ *for all* $\lambda \geq 0$.

(ii) $E_{n,\mathbb{C}}(\cdot)$ *is monotone decreasing and convex in* $[0,\infty)$.

(iii) $\lambda \to E_{n,\mathbb{C}}(\lambda)$ *is piecewise linear with at most* $n + 1$ *changes of slope.*

For every distribution F the properties (i) and (ii) also hold for $E_{\mathbb{C}}(\cdot)$.

<u>*Consistency:*</u>

First note, that there exist situations where $E_{n,\mathbb{C}}(\lambda)$ contains no information about the underlying distribution (cf. Hartigan (1987)). Define $A = \{X_1,\ldots,X_n\}$, then $F_n(A) = 1$ and $Leb(A) = 0$. Hence, if $A \in \mathbb{C}$, then $E_{n,\mathbb{C}}(\lambda) = 1$ for all $\lambda \geq 0$, independent of F. Therefore $E_{n,\mathbb{C}}$ is in general not a consistent estimator of $E_{\mathbb{C}}$. The Consistency Lemma 2.2 given below shows that $E_{n,\mathbb{C}}$ is consistent if $\mathbb{C}$ is a *Glivenko-Cantelli-class for F*. Given a class $\mathbb{C}$ denote $\| F_n - F \|_{\mathbb{C}} = \sup \{ \,|\, (F_n - F)(C) \,|\, : C \in \mathbb{C} \}$.
To avoid measurability considerations we define for any function $f : \Omega \to \mathbf{R}$ the *measurable cover function f** as the smallest measurable function from Ω to $\mathbf{R}$ lying everywhere above f (see Dudley (1984)). Furthermore, let P* denote *outer probability*.

Definition: *A <u>Glivenko-Cantelli-class</u> (GC-class) for a distribution F, or for short a GC(F)-class, is a class $\mathbb{C}$ of measurable sets such that with probability 1*

$$\| F_n - F \|_{\mathbb{C}}^{*} \to 0 \qquad \text{as } n \to \infty.$$

Consistency Lemma 2.2: *For any class* $\mathbb{C}$ *we have*

$$sup_{\lambda\geq 0}\, |\, E_{n,\mathbb{C}}(\lambda) - E_{\mathbb{C}}(\lambda)\, | \;\; \leq \;\; ||\, F_n - F\, ||_{\mathbb{C}}.$$

Hence, if $\mathbb{C}$ *is a GC-class for F, then we have that with probability 1*

$$sup_{\lambda\geq 0}\, |\, E_{n,\mathbb{C}}(\lambda) - E_{\mathbb{C}}(\lambda)\, |^* \; \to \; 0 \qquad as \;\; n \to \infty.$$

It is well known that Vapnic-Cervonencis (VC)- classes are GC-classes for *all* F if in addition they satisfy some measurability assumptions. Examples for such classes are the class of all d-dimensional closed balls, $\mathbb{B}^d$, and the class of all d-dimensional closed ellipsoids, $\mathbb{E}^d$. There also exist interesting classes which are GC-classes (for certain F) but no VC-classes, as for example the class of all closed convex sets in $\mathbf{R}^d$, $d \geq 2$, denoted by $\mathbb{C}^d$. These are GC-class for all distributions F which have a bounded Lebesgue density (see Eddy & Hartigan (1977) for a characterization of the GC-property of $\mathbb{C}^d$).

\- -

F i g u r e 2.2

\- -

As mentioned in the introduction, we identify the choice of $\mathbb{C}$ with the choice of a statistical model, which consists of those distributions whose density contour clusters lie in $\mathbb{C}$. In order to model multimodality, we make the following construction: Given a class $\mathbb{C}$ of closed subsets of $\mathbf{R}^d$ let $\mathbb{C}_k$, $k \in \mathbf{N}$ denote the class of sets which can be written as a union of k (possibly empty) sets in $\mathbb{C}$, and let

$$\mathbb{N}_{m,k}(\mathbb{C}) \;:=\; \{\bigcup_{j=1}^{m} \; (C_j \backslash \mathring{D}_j\,),\;\; C_j \in \mathbb{C},\, \mathring{D}_j \;\in \mathbb{C}_k,\, j \;=\; 1,..,m\}, \quad m,\, k \in \mathbf{N},$$

where $\mathring{D}_j$ denotes the open kernel of D_j. Note that the sets in $\mathbb{N}_{m,k}(\mathbb{C})$ are closed by definition, and that $\mathbb{C}_m \subset \mathbb{N}_{m,k}(\mathbb{C})\ \forall\ m \geq 1$. The classes $\mathbb{N}_{m,k}(\mathbb{C}^d)$ seem to be appropriate to model for example an underlying mixture of normal distributions (cf. Fig. 2.2 and Fig. 2.3).

\- -

F i g u r e 2.3

\- -

The classes $\mathbb{N}_{m,k}(\mathbb{C})$ are special cases of GC-classes which we call <u>*k-constructible*</u> (Alexander (1984) used this terminology for VC-classes): A class $\mathbb{C}$ in a measurable space (X,A) is called *k-constructible* from a GC-class $\mathbb{D}$, if there exists a function φ from $\mathbb{D}^k$ to A such that $\mathbb{C} \subset \varphi(\mathbb{D}^k)$. For example, the class $\mathbb{C} \setminus \mathbb{C} = \{ C \setminus D\colon C, D \in \mathbb{C}\}$ is 2-constructible from $\mathbb{C}$. More generally, the classes $\mathbb{N}_{m,k}(\mathbb{C})$ are m(k+1)-constructible from $\mathbb{C}$.

If $\mathbb{C}$ is a VC-class then classes which are k-constructible from $\mathbb{C}$ also are VC-classes, i.e. the VC-property of $\mathbb{C}$ carries over to the classes $\mathbb{N}_{m,k}(\mathbb{C})$. This is well known (Dudley (1978)). The analogous property also holds for GC-classes (e.g., see Pollard (1984), Theorem 21 and its proof). Hence we have the following corollary:

Corollary 2.3: *Let $\mathbb{C}$ be a GC-class for F. Then we have for every m, k $\in$ **N**, that with probability 1*

$$sup_{\lambda \geq 0} \,|\, E_{n,\mathbb{N}_{m,k}(\mathbb{C})}(\lambda) - E_{\mathbb{N}_{m,k}(\mathbb{C})}(\lambda) \,|^* \;\to\; 0 \qquad as\;\; n \to \infty.$$

3. The empirical generalized λ-clusters

The asymptotic behaviour of the empirical generalized λ-clusters $\Gamma_{n,\mathbb{C}}(\lambda)$ will be studied in this section. As a measure of distance we use the pseudometric

$$d_F(C,D) := F(C \,\Delta\, D), \qquad C, D \in \mathbb{C},$$

where Δ denotes the symmetric difference. The empirical generalized λ-clusters exist for interesting classes $\mathbb{C}$ which consist of *closed* sets, as for example for the classes $\mathbb{C} = \mathbb{B}^d$, $\mathbb{E}^d$ or $\mathbb{C}^d$ and for the corresponding classes $\mathbb{N}_{m,k}(\mathbb{C})$ (defined in Section 2). Therefore we shall assume in all of that what follows that

$\mathbb{C}$ consists of *closed* sets.

In addition we assume that

$$(3.1) \qquad Leb\{\, \overline{C(\lambda)} \setminus C(\lambda) \,\} \;=\; 0 \qquad for\ all\;\; \lambda \geq 0,$$

and only consider

$$\Gamma(\lambda) := \overline{C(\lambda)}$$

the *closure* of the density contour cluster. Because of (3.1) one can still think of $\Gamma(\lambda)$ as the density contour cluster. (3.1) is trivially satisfied for all upper semicontinuous densities, but of course many other densities also have this property.

In the sequel the following assumptions are assumed to hold unless stated otherwise:

General assumptions:

(A1) *For all $\lambda \geq 0$ there exists a generalized and an empirical generalized λ-cluster.*

(A2) *The underlying distribution F on $\mathbf{R}^d$ has a Lebesgue density f with $\max\{f(x)\} = M < \infty$. Furthermore, (3.1) holds.*

(A3) *All classes $\mathbb{C}$ under consideration are GC(F)-classes and consist of closed sets. Furthermore we assume that $\emptyset \in \mathbb{C}$.*

As already mentioned, the existence of a set $\Gamma_{\mathbb{C}}(\lambda)$ is guaranteed if $\Gamma(\lambda) \in \mathbb{C}$. But this assumption it is not necessary. For every distribution G which has a strictly positive Lebesgue density and every fixed $\lambda \geq 0$ the function $C \to H_\lambda(C)$ is upper semicontinuous on $(\mathbb{C}, d_G)$ (Lemma 6.1). Hence, if the space $(\mathbb{C}, d_G)$ is compact, then a generalized λ-cluster $\Gamma_{\mathbb{C}}(\lambda)$ exists. If in addition we know a priori that the sets $\Gamma_{\mathbb{C}}(\lambda)$ are compact, then for every fixed level λ there exists a compact set $K \subset \mathbf{R}^d$, such that we can restrict ourselves to $\mathbb{C}(K) = \{C \in \mathbb{C}, C \subset K\}$. In this situation the existence of a generalized λ-cluster is guaranteed if the space $(\mathbb{C}(K), d_G)$ is compact. The latter situation holds for example for $\mathbb{C} = \mathbb{B}^d$, $\mathbb{E}^d$ or $\mathbb{C}^d$.

<u>*Consistency:*</u>

First we consider the case where $\Gamma(\lambda)$ is not necessarily assumed to lie in $\mathbb{C}$, or in other words, we consider a situation where the corresponding model (see above) need not necessarily be correct. Note that the sets $\Gamma_{\mathbb{C}}(\lambda)$ and $\Gamma_{n,\mathbb{C}}(\lambda)$ need not to be unique. The non-uniqueness of $\Gamma_{n,\mathbb{C}}(\lambda)$ is not crucial, and the results given below hold for every choice of $\Gamma_{n,\mathbb{C}}(\lambda)$. This will not be mentioned further in the formulation of the results.

Theorem 3.2: *Let $\Lambda \subset [0,\infty)$. Suppose that the following two conditions hold:*

- *(i) For a distribution G with strictly positive Lebesgue density the space $(\mathbb{C},d_G)$ is quasicompact.*
- *(ii) For every $\lambda \in \Lambda$ the generalized λ-cluster is unique up to F-nullsets.*

Then we have with probability 1 that

$$\sup_{\lambda \in \Lambda} d_F(\Gamma_{\mathbb{C}}(\lambda),\Gamma_{n,\mathbb{C}}(\lambda)) \;{*}\!\!\to 0 \qquad \text{as } n \to \infty.$$

Remark: For the class of all closed convex sets with non empty interior in $\mathbf{R}^2$, the consistency of the empirical generalized λ-cluster (in the Hausdorff metric) was shown by Hartigan (1987) for fixed λ. Müller & Sawitzki (1991) proved uniform consistency in the one-dimensional case with $\mathbb{C} = \mathcal{I}_k$, where they assumed in addition $\Gamma(\lambda) \in \mathbb{C}$. Nolan (1991) considered the case $\mathbb{C} = \mathcal{E}^d$ in a more parametric setup.

There are interesting situations where the generalized λ-clusters are not unique. Assume for example that F is a (smooth) bimodal univariate distribution with density f, symmetric around zero, where a mode is defined to be a local maximum of f. Then, for some λ large enough, the density contours cluster is a union of two nonempty intervals, I_1 and I_2, say. If we choose $\mathbb{C}$ as the class of all intervals, then I_1 and I_2 both are generalized λ-clusters.

Let $\mathcal{M}_{\mathbb{C}}(\lambda)$ denote the class of all sets at which H_λ attains the supremum over $\mathbb{C}$, i.e.

$$\mathcal{M}_{\mathbb{C}}(\lambda) := \{\, \Gamma \in \mathbb{C} : \; H_\lambda(\Gamma) = E_{\mathbb{C}}(\lambda) \,\}.$$

Note that by (A1) we have $\mathcal{M}_{\mathbb{C}}(\lambda) \neq \emptyset$ for all $\lambda \geq 0$.

Theorem 3.3: *Suppose that assumption (i) of Theorem 3.2 holds. Then we have for every $\lambda \geq 0$ that with probability 1*

$$\inf_{\Gamma \in \mathcal{M}_{\mathbb{C}}(\lambda)} \{d_F(\Gamma_{n,\mathbb{C}}(\lambda), \Gamma)\}^* \to 0 \qquad \text{as } n \to \infty.$$

In the following we shall assume that "$\Gamma(\lambda) \in \mathbb{C}$". In contrast to the more general case considered above, this additional assumption allows us to derive *explicit* upper bounds for $d_F(\Gamma(\lambda),\Gamma_{n,\mathbb{C}}(\lambda))$, which are the key to derive consistency results and rates of convergence in the case "$\Gamma(\lambda) \in \mathbb{C}$".

Proposition 3.4: *Let $\lambda \geq 0$ be fixed and assume that $\Gamma(\lambda) \in \mathbb{C}$. Then the following inequalities hold for every $\eta > 0$:*

$$\begin{aligned} d_F(\Gamma(\lambda),\Gamma_{n,\mathbb{C}}(\lambda)) \quad &\leq \quad F\{x: | f(x) - \lambda | < \eta\} \\ &\quad + \eta^{-1} M [(F_n - F) (\Gamma_{n,\mathbb{C}}(\lambda)) - (F_n - F) (\Gamma(\lambda))]. \end{aligned} \tag{3.2a}$$

Furthermore we have for $\lambda = 0$ that

$$d_F(\Gamma(0),\Gamma_{n,\mathbb{C}}(0)) \quad \leq \quad (F_n - F) (\Gamma_{n,\mathbb{C}}(0)) - (F_n - F) (\Gamma(0)). \tag{3.2b}$$

The proof of the next theorem follows immediately from (3.2a) together with (A1):

Theorem 3.5: *Let Λ be a closed subset of the real line such that $\Gamma(\lambda) \in \mathbb{C}$ for all $\lambda \in \Lambda$ and suppose that*

$$\sup_{\lambda \in \Lambda} F\{x: | f(x) - \lambda | < \eta\} \to 0 \qquad \text{as } \eta \to 0. \tag{3.3}$$

Then we have with probability 1 that

$$\sup_{\lambda \in \Lambda} d_F(\Gamma_{\mathbb{C}}(\lambda),\Gamma_{n,\mathbb{C}}(\lambda))^* \to 0 \qquad \text{as } n \to \infty.$$

Remark: Condition (3.3) says, that "F has no flat part in Λ", i.e. $F\{x : f(x) = \lambda\} = 0$ for all $\lambda \in \Lambda$. Another equivalent formulation of (3.3) is to say that $\lambda \to \Gamma(\lambda)$ is continuous in Λ for the d_F-pseudometric. This follows from $F\{x: | f(x) - \lambda | < \eta\} = F((\Gamma(\lambda - \eta)) - F(\Gamma(\lambda + \eta)) - F\{x : f(x) = \lambda - \eta\}$.

<u>*Rates of Convergence*</u>

Our two main results on rates of convergence are Theorem 3.6 and Theorem 3.7. The first one deals with VC-classes $\mathbb{C}$. In the second we also allow more richer classes, where the richness is measured in terms of the <u>*metric entropy with inclusion*</u> of $\mathbb{C}$ with respect to F, which is defined as follows: let

$$N_I(\varepsilon, \mathbb{C}, F) := \inf\Big\{ m \in \mathbf{N} : \exists\, C_1,...C_m \textit{ measurable, such that for every } C \in \mathbb{C} \textit{ there exist } i, j \in \{1,..., m\} \textit{ with } C_i \subset C \subset C_j \textit{ and } F(C_j \setminus C_i) < \varepsilon \Big\},$$

then $\log N_I(\varepsilon, \mathbb{C}, F)$ is called metric entropy with inclusion of $\mathbb{C}$ with respect to F.

For the proofs of the theorems below we shall use results of Alexander (1984) about the behaviour of the empirical process. For that reason we shall also use some of his terminology. Alexander considered VC-classes which satisfy a certain measurability condition which he called "n-deviation measurable". Here we shall not give this definition and the underlying construction of the empirical measure, because all the standard VC-classes which we are interested in (the classes of balls, ellipsoids and finite unions and differences of them) satisfy this measurability condition. Furthermore, we call $\mathbb{C}$ a *(v,m)-constructible VC-class,* if $\mathbb{C}$ is m-constructible (as defined in Section 2) from a VC-class $\mathbb{D}$ whose index is smaller than or equal to k. The index of a VC-class is defined as the smallest integer k, such that $\mathbb{D}$ "shatters" no set which consists of v points. And $\mathbb{D}$ "shatters" a finite set C, iff every $B \subset C$ is of the form $C \cap D$ for some $D \in \mathbb{D}$.

Theorem 3.6: *Let $\mathbb{C}$ be a "n-deviation measurable" (v,m)-constructible VC-class and suppose that for a closed subset Λ of $[0,\infty)$ there exist constants γ, $C \geq 0$ such that*

$$(3.5) \qquad \sup_{\lambda\in\Lambda} F\{\, x : |\, f(x) - \lambda \,| < \eta \,\} \;\leq\; C\, \eta^{\gamma}.$$

If $\Gamma(\lambda) \in \mathbb{C}$ $\forall\, \lambda \in \Lambda$ then there exists a constant $K = K(M,C,\gamma,\mathbb{C})$ such that

$$P^*\Big(\sup_{\lambda\in\Lambda} d_F(\Gamma(\lambda),\Gamma_{n,\mathbb{C}}(\lambda)) > K\, (\, n/\log(n)\,)^{-\gamma/(2+\gamma)} \Big) \;\to\; 0 \quad \text{as} \quad n \to \infty.$$

Examples: Consider a fixed level $\lambda > 0$. For levels λ where $\|$ grad $f(x) \|$ is bounded away from zero in a neighbourhood of $\{x: f(x) = \lambda\}$ we have $\gamma = 1$. Let F be a smooth unimodal distribution. Then, for d = 1, the density contour clusters are intervals so that $\mathbb{C} = \mathfrak{I}_1$ is an appropriate choice. For $d \geq 2$ we assume the density contour clusters to be balls or ellipsoids, i.e. we take $\mathbb{C} = \mathfrak{B}^d$ or $\mathfrak{E}^d$. For these situations we obtain from Theorem 3.6 that

$$d_F(\Gamma(\lambda),\Gamma_{n,\mathbb{C}}(\lambda)) = O_{P^*}(\, n^{-1/3}(\log n)^{1/3}\,).$$

Levels λ where $\gamma < 1$ are called *critical levels*. If for example f has a unique maximum λ_0 at the mode x_0 and behaves like a parabola in a neighbourhood of x_0, then it can be shown that

$$F\{\, x : |\, f(x) - \lambda_0 \,| < \eta \,\} = \begin{cases} O(\eta^{1/2}), & \text{for } d = 1 \\ O(\eta), & \text{for } d \geq 2. \end{cases}$$

Hence, if f has no other critical levels then we have for $\Lambda = [\delta,\infty)$, $0 < \delta < \lambda_0$, that $\gamma = 1/2$ for d = 1 and $\gamma = 1$ for $d > 1$. If we consider the same VC-classes as above, i.e. $\mathbb{C} = \mathfrak{I}_1$ for d = 1 and $\mathbb{C} = \mathfrak{B}^d$ or $\mathfrak{E}^d$ for $d \geq 2$, then

$$\sup_{\lambda\geq\delta} d_F(\Gamma(\lambda),\Gamma_{n,\mathbb{C}}(\lambda)) = \begin{cases} O_{P^*}(n^{-1/5}(\log n)^{1/5}), & \text{for } d = 1 \\ O_{P^*}(n^{-1/3}(\log n)^{1/3}), & \text{for } d \geq 2. \end{cases}$$

If we want to include $\delta = 0$, then additional conditions on the tail behaviour of f are necessary to control $\sup_{\lambda\geq 0} F\{x : |\, f(x) - \lambda \,| < \eta\}$ as $\eta \to 0$.

Theorem 3.7: *Let $\mathbb{C}$ be such that there exist constants A, r > 0 with*

$$\text{(3.6)} \qquad \log N_I(\varepsilon, \mathbb{C}, F) \le A\,\varepsilon^{-r} \qquad \forall\, \varepsilon > 0.$$

Suppose that there exists a closed subset Λ of $[0,\infty)$ such that $\Gamma(\lambda)\in \mathbb{C}$ for all $\lambda \in \Lambda$ and that (3.5) holds. Then there exist positive constants $L(r) = L(r,A,M)$ such that with probability tending to one as $n \to \infty$.

$$sup_{\lambda\in\Lambda}\, d_F(\Gamma(\lambda),\Gamma_{n,\mathbb{C}}(\lambda))^* \le \begin{cases} L(r)\, n^{-\gamma/(2+(1+r)\gamma)}, & r < 1 \\ L(r)\, n^{-\gamma/2(\gamma+1)}\, L(n), & r = 1 \\ L(r)\, n^{-\gamma/(\gamma+1)(r+1)}, & r > 1 \end{cases}$$

Example: Let $\mathbb{C} = \mathbf{C}^2$ and assume that the sets $\Gamma(\lambda)$, $\lambda \in \Lambda$, all lie in a compact set K. Then we have r = 1/2 (Dudley (1984). Hence, we obtain from Theorem 3.7 for regular situations where γ = 1 (see above) that $d_F(\Gamma(\lambda),\Gamma_{n,\mathbb{C}}(\lambda)) = O_{P*}(n^{-2/7})$. Hartigan (1987) conjectured that for such cases the rate is $O_{P*}(n^{-2/7}(\log n)^{2/7})$ in the Hausdorff-distance. If such a compact set K does not exist (for example for a distribution with unbounded support and $0 \in \Lambda$), then one in addition needs conditions on the tail behaviour of F to ensure that r = 1. A sufficient condition is that there exist constants $0 \le \eta$, c, k< ∞ such that $f(x)\, ||x||^{\eta} \le c$ for $||x|| > k$. This is shown in Polonik (1992).

<u>*Estimating the support of a density and the case of an underlying uniform distribution:*</u>

Estimating the density contour clusters of a uniform distribution U (for λ bounded away from the maximum of the density) means estimating the support of U. Since in this situation the quantity $F\{x: |f(x) - \lambda| < \eta\}$ which appears in (3.2a) is zero for η small enough, we formally have the same basic inequality as in the case of estimating the support of an arbitrary distribution F (cf. (3.2b)). Therefore we summarize the results concerning these both cases in Proposition 3.8 below. The assertion of Proposition 3.8 formally follows from Theorem 3.6 and Theorem 3.7, respectively, by taking $\gamma = \infty$.

As mentioned earlier, the support of f, supp{f}, is a generalized 0-cluster if it lies in $\mathbb{C}$. For $\mathbb{C}$ =

$\mathbf{C}^d$, $d \geq 2$, the convex hull of the sample $X_1,.., X_n$, denoted by $conv_n$, is an empirical generalized 0-cluster.

Proposition 3.8: *The given results hold with probability tending to one as $n \to \infty$.*

(a) Let $\mathbb{C}$ be a "n-deviation measurable" (v,m)-constructible VC-class and suppose that supp{f} $\in \mathbb{C}$. Then there exists a constant $C = C(v,m)$ such that

$$d_F(supp\{f\},\Gamma_{n,\mathbb{C}}(0))^* \ < \ C\, n^{-1} \log(n).$$

(b) Let the class $\mathbb{C}$ satisfy (3.6) and suppose that supp{f} $\in \mathbb{C}$. Then there exist constants $C(r) = C(r,A)$ such that

$$d_F(supp\{f\},\Gamma_{n,\mathbb{C}}(0))^* \ < \ \begin{cases} C(1)\ n^{-1/2}\log(n), & r = 1 \\ C(r)\ n^{-1/(1+r)}, & r \neq 1 \end{cases}$$

Hence, if $\mathbb{C} = \mathbf{C}^d$, $d \geq 2$, $m, k \in \mathbf{N}$ and supp{f} is compact then we have

$$d_F(supp\{f\},conv_n)^* \quad \leq \quad \begin{cases} C(1)\, n^{-1/2}\log(n), & d = 3 \\ C(\frac{d-1}{2})\ n^{-2/(d+1)}, & d \neq 3 \end{cases}$$

(c) Let U be a uniform distribution on a bounded set S and denote $M = 1/Leb(S)$. If $S \in \mathbb{C}$ then the rates given above also hold for $\sup_{\lambda < M + \delta} d_U(\Gamma(\lambda),\Gamma_{n,\mathbb{C}}(\lambda))^*$, $\delta > 0$ arbitrary.

Remark: For an underlying uniform distribution with a compact convex support in $\mathbf{R}^d$ which has a smooth boundary it is known, that $n^{-2/(d+1)}$ is the exact L_1-rate of the random quantity $d_{Leb}(supp\{f\},conv_n)$. For $d = 2$ this is a well known result of Rényi & Sulanke (1964) (cf. Schneider (1988) for a survey of results in this context). However, also in the case of an unbounded convex support, Proposition 3.8 (b) gives rates of convergence of the convex hull of the sample. We only need to control the metric entropy with bracketing of the corresponding class $\mathbb{C}$. In the example given after Theorem 3.6 we already mentioned, that for $\mathbb{C} = \mathbf{C}^2$ condition (3.6) holds with $r = 1/2$ if a week condition on the tail behavior is satisfied. Hence, in this case Proposition 3.8 (b) gives $d_F(supp\{f\},conv_n) = O_{P*}(n^{-2/3})$.

4. The empirical excess mass, revisited

The consistency results and the rates of convergence for the empirical generalized λ-clusters (derived in the previous chapters) will be used here to study the asymptotic behaviour of the standardized empirical excess mass, which is defined as

$$Z_{n,\mathbb{C}}(\lambda) := n^{1/2} \left(E_{n,\mathbb{C}}(\lambda) - E_{\mathbb{C}}(\lambda) \right).$$

If we (formally) ignore the estimation of $\Gamma_{\mathbb{C}}(\lambda)$ and consider $\widetilde{E}_{n,\mathbb{C}}(\lambda) = H_{n,\lambda}(\Gamma_{\mathbb{C}}(\lambda))$, then the difference $\widetilde{E}_{n,\mathbb{C}}(\lambda) - E_{\mathbb{C}}(\lambda)$ simply equals the difference $(F_n - F)(\Gamma_{\mathbb{C}}(\lambda))$ which is of the order $O_P(n^{-1/2})$. It will turn out, that the random fluctuation which comes in through the estimation of $\Gamma_{\mathbb{C}}(\lambda)$ is asymptotically negligible, so that $n^{1/2}$ is the appropriate normalizing factor. Even in the case where the generalized λ-clusters $\Gamma_{\mathbb{C}}(\lambda)$ are not unique, $Z_{n,\mathbb{C}}(\lambda)$ can be approximated by $F_n - F$ evaluated at the generalized λ-clusters. However, in contrast to the "case of uniqueness", the generalized λ-clusters have to be chosen randomly in $\mathcal{M}_{\mathbb{C}}(\lambda)$.

We say that the set-indexed empirical process ν_n is *stochastically equicontinuous in the limit*, if $\lim_{\delta\to 0}\limsup_{n\to\infty} P^*(\sup_{d_F(C,D) < \delta} | \nu_n(C) - \nu_n(D) | > \eta \} = 0$ for all $\eta > 0$.

Theorem 4.1: *Assume that the following two conditions hold:*

(i) There exists a distribution G which has a strictly positive Lebesgue density such that the space $(\mathbb{C}, d_G)$ is quasicompact.

(ii) ν_n indexed by $\mathbb{C}$ is stochastically equicontinuous in the limit.

Let $\lambda \geq 0$ be fixed. Then there exists a random sequence $\{\Gamma_{\mathbb{C}}(\lambda,n),\ n \in \mathbf{N}\} \subset \mathcal{M}_{\mathbb{C}}(\lambda)$ such that

$$| Z_{n,\mathbb{C}}(\lambda) - n^{1/2} (F_n - F)(\Gamma_{\mathbb{C}}(\lambda,n)) | = o_{P*}(1) \qquad \text{as } n \to \infty.$$

Corollary 4.2: *Assume that conditions (i) and (ii) of Theorem 4.1 hold. Then we have for every $\lambda \geq 0$ such that $\mathcal{M}_{\mathbb{C}}(\lambda)$ is finite that*

$$Z_{n,\mathbb{C}}(\lambda) \;=\; O_{P*}(1) \qquad \text{as } n \to \infty\,.$$

The rate is exact if $F(\Gamma) > 0$ for all $\Gamma \in \mathcal{M}_{\mathbb{C}}(\lambda)$.

If the generalized λ-clusters are uniquely determined (up to F-nullsets), then we can proof stronger results. Let $D(\Lambda)$ denote the space of all real-valued functions on Λ which are continuous from the right and have left limits, equipped with the Skorohod topology.

Theorem 4.3: *Let $\Lambda \subset [0,\infty)$ be compact. Assume that the generalized λ-clusters are unique up to F-nullsets and that the following conditions hold:*

(i) $\sup_{\lambda\in\Lambda} d_F(\Gamma_{\mathbb{C}}(\lambda)),\Gamma_{n,\mathbb{C}}(\lambda) \overset{*}{\to} 0$ *with probability 1 as* n → ∞ *and*
(ii) ν_n *indexed by* $\mathbb{C}$ *is stochastically equicontinuous in the limit,*

then $$\sup_{\lambda\in\Lambda} |\, Z_{n,\mathbb{C}}(\lambda) - B_{n,\mathbb{C}}(\lambda) \,| \;=\; o_{P*}(1) \qquad \text{as } n \to \infty,$$

where $B_{n,\mathbb{C}}(\lambda) = n^{1/2}\,(F_n - F)(\Gamma_{\mathbb{C}}(\lambda))$. Moreover, if in addition

(iii) *F has no flat part in Λ, i.e. (3.3) holds, and*
(iv) $\Gamma(\lambda) \in \mathbb{C}\ \forall\, \lambda \in \Lambda$,

then

$$B_{n,\mathbb{C}}(\lambda) \;\to\; B(a_F(\lambda)) \qquad \text{in distribution as } n \to \infty$$

in $D(\Lambda)$, where B denotes a standard Brownian Bridge and $a_F(\lambda) = F(\Gamma_{\mathbb{C}}(\lambda))$.

Suppose the assumptions of Theorem 4.3 are satisfied with $\Lambda = \Lambda_0 = [0,\lambda_0]$, $\lambda_0 \geq M$ for $\mathbb{C} = \mathbf{E}^2$. Then we have for every $\varepsilon > 0$ that

$$(4.1) \qquad P\,[\, \sup_{\lambda\in\Lambda_0} |\, Z_{n,\mathbb{N}_{m,k}(\mathbf{E}^2)}(\lambda) \,| \leq \varepsilon \,] \;\to\; P\,[\, \sup_{0\leq t\leq 1} |\, B(t) \,| \leq \varepsilon \,].$$

This leads to confidence bands for E(λ). If λ_0 has to be chosen smaller than M, as for example in the case of the uniform distribution, or if λ has to be bounded away from zero, then the right-hand side in (4.1) is asymptotically larger than the left-hand side (cf. Müller & Sawitzki (1987) for the onedimensional case).

5. Tests based on differences of excess masses

The underlying idea for constructing tests based on differences of excess masses has already been explained in the introduction. In general we study the following testing problem: Let $\mathbb{C}$, $\mathbb{D}$ be two classes of measurable subsets of $\mathbf{R}^d$ with $\mathbb{C} \subset \mathbb{D}$ and let Λ be a subset of [0,∞). We consider the hypothesis that the generalized λ-clusters in $\mathbb{D}$ already lie in the smaller class $\mathbb{C}$, i.e. the problem is testing

$$H_0: \ \mathfrak{M}_{\mathbb{D}}(\lambda) \subset \mathbb{C} \qquad \text{for all } \lambda \in \Lambda$$

versus

$$H_1: \ \mathfrak{M}_{\mathbb{D}}(\lambda) \subset \mathbb{D} \setminus \mathbb{C} \qquad \text{for some } \lambda \in \Lambda.$$

Remember that for every fixed $\lambda \geq 0$, $\mathfrak{M}_{\mathbb{D}}(\lambda)$ denotes the set of all generalized λ-clusters in $\mathbb{D}$. Of course we mainly think of cases where the generalized λ-clusters are defined uniquely up to F-nullsets or where the density contour clusters lie in $\mathbb{D}$. Let $\Delta_n(\mathbb{C},\mathbb{D},\lambda) = E_{n,\mathbb{D}}(\lambda) - E_{n,\mathbb{C}}(\lambda)$. As a test statistic for the above testing problem we consider

$$T_n(\mathbb{C},\mathbb{D},\Lambda) \ = \ \sup_{\lambda \in \Lambda} \Delta_n(\mathbb{C},\mathbb{D},\lambda)$$

This test statistic is a generalization of the test statistics proposed by Müller and Sawitzki (1987) and Hartigan (1987), respectively, for testing the hypothesis of multimodality.
$\Delta_n(\mathbb{C},\mathbb{D},\lambda)$ is non-negative for each $\lambda \geq 0$ and large values of this statistic (for some λ) suggest a violation of the hypotheses H_0 (see introduction).
If we consider the univariate case and choose $\mathbb{C} = \mathfrak{I}_1$ and $\mathbb{D} = \mathfrak{I}_2$, then the above testing problem can be regarded as looking for unimodality versus bimodality (cf. introduction). For the analogous problem in two dimensions an appropriate choice is $\mathbb{C} = \mathfrak{C}^2$ and $\mathbb{D} = \mathbb{N}_{3,2}(\mathfrak{C}^2)$ (cf. Fig. 2.2). Tests for the hypothesis of “k modes“, k ≥ 2 against the alternative of “m modes“, k < m, can be constructed analogously. Choosing $\mathbb{C}$ as the class of all balls and $\mathbb{D}$ as the class of all ellipsoids

gives a test which may be interpreted as a test for homoscedasticity.

In the important special case where the (closures of the) density contour clusters are assumed to lie in $\mathbb{D}$, the testing problem reduces to

$$\widetilde{H}_0 : \quad \Gamma(\lambda) \in \mathbb{C} \qquad \text{for all } \lambda \in \Lambda$$

versus

$$\widetilde{H}_1 : \quad \Gamma(\lambda) \in \mathbb{D} \setminus \mathbb{C} \qquad \text{for some } \lambda \in \Lambda,$$

Define $\quad T(\mathbb{C},\mathbb{D},\Lambda) = \sup_{\lambda \in \Lambda} \Delta(\mathbb{C},\mathbb{D},\lambda) = \sup_{\lambda \in \Lambda} (E_{\mathbb{D}}(\lambda) - E_{\mathbb{C}}(\lambda)).$

The following proposition shows that $T_n(\mathbb{C},\mathbb{D},\Lambda)$ converges stochastically to $T(\mathbb{C},\mathbb{D},\Lambda)$. The proof follows immediately by means of the Consistency Lemma 2.2:

Proposition 5.1: *For every choice of* $\mathbb{C}$ *and* $\mathbb{D}$ *we have*

$$(5.1) \qquad \sup_{\lambda \geq 0} | \Delta_n(\mathbb{C},\mathbb{D},\lambda) - \Delta(\mathbb{C},\mathbb{D},\lambda) | \leq || F_n - F ||_{\mathbb{D}} + || F_n - F ||_{\mathbb{C}}.$$

Hence, if $\mathbb{D}$ *is a GC-class for F, then we have for any* $\Lambda \subset [0,\infty)$ *that with probability 1*

$$| T_n(\mathbb{C},\mathbb{D},\Lambda) - T(\mathbb{C},\mathbb{D},\Lambda) |^* \to 0 \qquad \text{as } n \to \infty.$$

If in addition H_0 *holds, then it follows that with probability 1*

$$T_n(\mathbb{C},\mathbb{D},\Lambda)^* \to 0 \qquad \text{as } n \to \infty.$$

If $E_{\mathbb{D}}(\lambda) - E_{\mathbb{C}}(\lambda) > 0$ for some $\lambda \in \Lambda$, then it follows from Proposition 5.1 that the power of a test based on $T_n(\mathbb{C},\mathbb{D},\Lambda)$ converges to 1 as n tends to infinity. This is the case if the generalized λ-clusters are *unique* up to F-nullsets and if $F(\Gamma_{\mathbb{D}}(\lambda) \Delta \Gamma_{\mathbb{C}}(\lambda)) > 0$ for some $\lambda \in \Lambda$. In general the condition $F(\Gamma_{\mathbb{D}}(\lambda) \Delta \Gamma_{\mathbb{C}}(\lambda)) > 0$ does not follow from $\Gamma_{\mathbb{D}}(\lambda) \neq \Gamma_{\mathbb{C}}(\lambda)$, however, in many standard situations this is the case.

Rates of convergence:

The asymptotic distribution of the proposed test statistic is known only for the case of an underlying uniform distribution (cf. Theorem 5.4 below). However, rates of the convergence for the test statistic can be given which give qualitative insight into the behaviour of the test statistic under various testing problems, i.e. under various classes $\mathbb{C}$, $\mathbb{D}$ and sets Λ. In general only upper bounds for the rates of convergence of the test statistics are given. At least in some univariate situations these rates are known to be close (up to a log-term) to the exact rates.

Theorem 5.2: *Let $\mathbb{C}$ be a "n-deviation measurable" (v,m)-constructible VC-class and suppose that (3.5) holds. Then we have under $\widetilde{\mathrm{H}}_0$ that*

$$T_n(\mathbb{C},\mathbb{D},\Lambda) = O_{P*}(n^{-(1+\gamma)/(2+\gamma)} (\log n)^{\gamma/(2+\gamma)}) \qquad \text{as } n \to \infty.$$

Examples: The interesting situation here is the case $\Lambda = [0,\infty)$, because the supremum of the density f clearly is unknown. If F is a smooth univariate unimodal distribution whose density behaves like a parabola near the mode then we have $\gamma = 1/2$ (cf. example after Theorem 3.6). Hence, it follows that

$$T_n(\mathbb{C},\mathbb{D},\Lambda) = O_{P*}(n^{-3/5} (\log n)^{3/5}).$$

This rate has already been derived by Müller & Sawitzki (1991) with the help of the "Hungarian embedding". In higher dimensions, $d \geq 2$, we have in such regular unimodal cases (where the densities behave like a parabola near the mode), that $\gamma = 1$ (see examples given after Theorem 3.6). Hence, we have in this case that

$$T_n(\mathbb{C},\mathbb{D},\Lambda) = O_{P*}(n^{-2/3} (\log n)^{2/3}).$$

Note that this rate is faster than the rate for the onedimensional case. This is caused by the smoothness assumptions, more precisely, by the behaviour of the function Ψ occurring in (3.4). A detailed explanation for this fact is given in Polonik (1992). For more richer classes than the VC-classes the following holds:

Theorem 5.3: *Let $\mathbb{C}$ be such that there exist constants A, r > 0 with*

$$\log N_I(\varepsilon, \mathbb{C}, F) \;\le\; A\, \varepsilon^{-r} \qquad\qquad \forall\, \varepsilon > 0.$$

and suppose that (3.5) holds. Then we have under $\widetilde{\mathrm{H}}_0$ that

$$T_n(\mathbb{C},\mathbb{D},\Lambda) = O_{P*}(\,\alpha_n) \qquad\qquad \text{as } n \to \infty.$$

where

$$\alpha_n \;=\; \begin{cases} n^{-(1+\gamma)/(2+(1+r)\gamma)}, & r < 1 \\ n^{-1/2}\, \log(n), & r = 1 \\ n^{-1/(r+1)}, & r > 1 \end{cases}$$

Examples: We also consider the case $\Lambda = [0,\infty)$ and assume that f has no flat parts, is unimodal and behaves like a parabola near the mode, so that $\gamma = 1$ (cf. examples after Theorem 3.7). Furthermore we assume that the density contour clusters are convex, i.e. we choose $\mathbb{C} = \mathbf{C}^d$, $d \ge 2$, so that $r = (d-1)/2$. Hence, it follows from Theorem 5.3 that

$$T_n(\mathbb{C},\mathbb{D},\Lambda) \;=\; \begin{cases} O_{P*}(n^{-4/7}), & d = 2 \\ O_{P*}(n^{-1/2}\log(n)), & d = 3 \\ O_{P*}(n^{-1/(d+1)}), & d \ge 4. \end{cases}$$

The next theorem shows (together with (5.1)) that for an underlying uniform distribution $n^{-1/2}$ is the exact rate for the proposed test statistic under H_0 if in addition $\mathbb{D}$ is a Donsker class (with the exception of some degenerate cases, as for example $\mathbb{D} = \{\emptyset\}$). Classes $\mathbb{D}$ are called *Donsker classes for F*, if the following two conditions (a) and (b) hold: (a) there exists a $\mathbb{D}$-indexed Brownian Bridge $G_{\mathbb{D}}$ corresponding to F (or in other words, a F-bridge over $\mathbb{D}$, cf. for example Pollard (1984)), and (b) the $\mathbb{D}$-indexed empirical process ν_n converges to $G_{\mathbb{D}}$ in the sense that $\| \nu_n - B_{\mathbb{D}} \|_{\mathbb{D}} \to 0$ in outer probability. Note that the $\mathbb{D}$-indexed empirical process is stochastically equicontinuous in the limit if $\mathbb{D}$ is a Donsker classes .

For a Donsker class $\mathbb{D}$ let

$$Z_{\mathbb{D}}(\lambda) := sup_{D\in\mathbb{D}} \, (G_{\mathbb{D}}(D) - \lambda \, Leb(D)),$$

where $G_{\mathbb{D}}$ denotes a $\mathbb{D}$-indexed Brownian Bridge corresponding to F.

Theorem 5.4: *Let F be a uniform distribution on a bounded set $C_0 \subset \mathbf{R}^d$ and let $\mathbb{D}_0 := \{ D \cap C_0, D \in \mathbb{D} \}$. Suppose that $\mathbb{D}_0$ is a Donsker class for F. Then we have for every interval $\Lambda \subset [0,\infty)$ with $\lambda_0 := 1/Leb(C_0) \in \Lambda$ and every class $\mathbb{C} \subset \mathbb{D}_0$ that*

$$\left| \, n^{1/2} \, T_n(\mathbb{C},\mathbb{D},\Lambda) - sup_{-\infty < \lambda < \infty}(Z_{\mathbb{D}_0}(\lambda) - Z_{\mathbb{C}}(\lambda)) \, \right| \;=\; o_{P*}(1) \qquad as\ n \to \infty.$$

Remarks: (i) The assumption that $\mathbb{D}_0$ forms a Donsker class for F is fulfilled if $\mathbb{D}$ is a Donsker class of *subsets* of C_0. For example choose C_0 as the unit cube in $\mathbf{R}^2$ and $\mathbb{D}$ as the class of all circles, ellipses or closed convex sets in C_0.
(ii) If it is known that $\mathbb{D}$ is not too rich, such that $J = \int_0^1 (\log N_I(\eta^2,\mathbb{D},F))^{1/2} \, d\eta$ is finite then the same holds for $\mathbb{D}_0$. This for example holds for $\mathbb{D} = \mathbb{C}^2$, because it is known (cf. Dudley 1984) that $\log N_I(\eta^2,\mathbb{C}^2,F) \leq A \, \eta^{-1}$, for some constant $A > 0$. The finiteness of J is sufficient for the Donsker property (Dudley 1984). Hence, in this situation the Donsker property of $\mathbb{D}$ carries over to $\mathbb{D}_0$.

For Donsker classes $\mathbb{D}$ Proposition 5.1 shows, that for distributions which have no flat part the generalized λ-clusters are uniformly consistent under $\widetilde{H}_0$ (Theorem 3.5). Hence it follows from Theorem 5.2 that under $\widetilde{H}_0$ the test statistic is asymptotically larger under the uniform distribution than under distributions which have no flat parts. In this situation one could therefore use Monte Carlo simulations under the uniform distribution to determine a critical value for the test, so that the significance of the test could be controlled, at least for large n. In the one dimensional case simulation studies of Müller & Sawitzki (1987) show that this strategy works well for $n \geq 10$. For higher dimensions simulations have not been done yet.

6. Proofs

Proofs of Section 2:

Proof of Proposition 2.1: Since $\varnothing \in \mathbb{C}$ (i) follows directly from the definition of the excess mass. $E_{n,\mathbb{C}}(\lambda)$ is a supremum over affine linear functions of λ, which either are constant or have a negative slope. Hence $E_{n,\mathbb{C}}(\cdot)$ is monotone decreasing and convex in $[0,\infty)$. The assertion (iii) follows from the fact that the affine linear functions $\lambda \to F_n(C) - \lambda \, \mathrm{Leb}(C)$, $C \in \mathbb{C}$, over which the supremum in the definition of $E_{n,\mathbb{C}}$ is extended have at most n+1 different intercepts. .

❏

Proof of the Consistency Lemma 2.2: Using $H_{n,\lambda} = H_\lambda + (F_n - F)$ we get

$$| E_{n,\mathbb{C}}(\lambda) - E_{\mathbb{C}}(\lambda) | = | \sup_{C\in\mathbb{C}} H_{n,\lambda}(C) - \sup_{C\in\mathbb{C}} H_\lambda(C) |$$

$$\leq \sup_{C\in\mathbb{C}} | H_{n,\lambda}(C) - H_\lambda(C) | = \| F_n - F \|_{\mathbb{C}}.$$

❏

Proofs of Section 3:

In order to prove Theorem 3.2 we need two lemmas (Lemma 6.1 and Lemma 6.2) which will be proved first:

Lemma 6.1: (Properties of H_λ)

(a) $\sup_{\lambda\geq 0} | H_\lambda(\Gamma_{n,\mathbb{C}}(\lambda)) - H_\lambda(\Gamma_{\mathbb{C}}(\lambda)) |^* \to 0$ *with probability 1 as* $n \to \infty$.

(b) For every distribution G which has a strictly positive Lebesgue density the function $C \to H_\lambda(C)$, $C \in (\mathbb{C}, d_G)$ *is upper semicontinuous.*

Proof: **(a)** From the definition of $\Gamma_{n,\mathbb{C}}(\lambda)$ it follows $H_{n,\lambda}(\Gamma_{n,\mathbb{C}}(\lambda)) \geq H_{n,\lambda}(\Gamma_{\mathbb{C}}(\lambda))$. Together with $H_{n,\lambda} = H_\lambda + F_n - F$ this leads to

(6.1) $\qquad 0 \leq H_\lambda(\Gamma_{\mathbb{C}}(\lambda)) - H_\lambda(\Gamma_{n,\mathbb{C}}(\lambda)) \leq (F_n - F)(\Gamma_{n,\mathbb{C}}(\lambda)) - (F_n - F)(\Gamma_{\mathbb{C}}(\lambda)),$

and since $\mathbb{C}$ is a GC-class for F (general assumption (A1)) the assertion follows.

(b) First note that F is dominated by G (this follows from (A2)). Therefore it remains to show that $A \to \mathrm{Leb}(A)$ is lower semicontinuous for d_G. In order to see this let $\{K_n\}$ be a sequence of compact sets in $\mathbf{R}^d$ with $K_n \uparrow \mathbf{R}^d$. Then clearly

$$\mathrm{Leb}(A) = \sup_{n \in \mathbf{N}} \mathrm{Leb}(A \cap K_n),$$

and because G has a strictly positive Lebesgue density the functions $A \to \mathrm{Leb}(A \cap K_n)$ are continuous for for d_G. Hence, as a supremum over continuous functions, the function $A \to \mathrm{Leb}(A)$ is lower semicontinuous.

❑

Lemma 6.2: *Let $\Lambda \subset [0,\infty)$. Suppose that conditions (i) and (ii) of Theorem 3.2 are satisfied. Then $\lambda \to \Gamma_{\mathbb{C}}(\lambda)$ is uniformly continuous in Λ for the d_F-pseudometric.*

Proof: Without loss of generality we assume Λ to be compact, because for any $\lambda \geq M$ we have $\mathrm{Leb}(\Gamma_{\mathbb{C}}(\lambda)) = F(\Gamma_{\mathbb{C}}(\lambda)) = 0$. (This follows from the fact that $E_{\mathbb{C}}(\lambda) = 0$ for $\lambda > \max\{f(x)\}$).
Let $\{\lambda_n, n \in \mathbf{N}\}$ be a sequence in Λ with $\lambda_n \to \lambda_0$, $\lambda_0 \in \Lambda$. Because of the compactness of $\mathbb{C}$ we may assume that $\{\Gamma_{\mathbb{C}}(\lambda_n)\}$ converges to a set $D_0 \in \mathbb{C}$ in the d_G-pseudometric.
First assume $\lambda_0 \in \mathrm{int}\,\Lambda$, the interior of Λ. Since $\lambda_n \to \lambda_0$ we have for a given $\varepsilon > 0$ that $\lambda_0 - \varepsilon \leq \lambda_n \leq \lambda_0 + \varepsilon$ for large enough n. Remember that $H_\lambda(\Gamma_{\mathbb{C}}(\lambda)) = E_{\mathbb{C}}(\lambda)$ and that $E_{\mathbb{C}}(\lambda)$ is monotonically decreasing (Proposition 2.1). Therefore we get by using the upper semicontinuity of H_λ (Lemma 6.1) that

$$H_{\lambda_0+\varepsilon}(\Gamma_{\mathbb{C}}(\lambda_0+\varepsilon)) \leq \limsup_n H_{\lambda_n}(\Gamma_{\mathbb{C}}(\lambda_n)) \leq \limsup_n H_{\lambda_0-\varepsilon}(\Gamma_{\mathbb{C}}(\lambda_n)) \leq H_{\lambda_0-\varepsilon}(D_0).$$

Letting $\varepsilon \to 0$ we obtain $H_{\lambda_0}(\Gamma_{\mathbb{C}}(\lambda_0)) \le H_{\lambda_0}(D_0)$ and the assertion follows from the assumed uniqueness of the maximum.

If $\lambda_0 \in \overline{\Lambda} \setminus \Lambda$, where $\overline{\Lambda}$ denotes the closure of Λ then omit the ε on the obvious side in the above inequalities.

❏

Proof of Theorem 3.2 (cf. Müller & Sawitzki (1991b)) **and Theorem 3.3:**

First we prove the special case that Λ consists of a single point λ. In this case the proof of very short and shows the main idea.

We may assume that a given realization of the random sequence $\{\Gamma_{n,\mathbb{C}}(\lambda), n \in \mathbf{N}\}$ converges to a set $D_0 \in \mathbb{C}$ in the d_G-pseudometric. Hence it follows from Lemma 6.1 (a) and (b) that with probability 1

$$H_\lambda(\Gamma_{\mathbb{C}}(\lambda)) = \lim\sup_n H_\lambda(\Gamma_{n,\mathbb{C}}(\lambda))^* \le H_\lambda(D_0)$$

and from the assumed uniqueness of the maximum the assertion follows.

Now we consider the general case where $\Lambda \subset [0,\infty)$ is an arbitrary closed set. We need to show that for every sequence $\{\lambda_n \in \Lambda\}$ we have $d_F(\Gamma_{\mathbb{C}}(\lambda_n),\Gamma_{n,\mathbb{C}}(\lambda_n)) \to 0$ with outer probability 1 as $n \to \infty$. It can be assumed that $\lambda_n \to \lambda_0$, $\lambda_0 \in \Lambda \cup \{\infty\}$.

Since the function $\lambda \to \Gamma_{\mathbb{C}}(\lambda)$ is continuous for the d_F-pseudometric (Lemma 6.2), it is enough to show that with probability 1

$$F(\Gamma_{n,\mathbb{C}}(\lambda_n) \,\Delta\, \Gamma_{\mathbb{C}}(\lambda_0))^* \;\to\; 0 \quad \text{as} \quad n \to \infty.$$

For $\lambda_0 < \infty$ the proof is much the same as the proof of the continuity of $\lambda \to \Gamma_{\mathbb{C}}(\lambda)$. The only difference is, that here in addition the random quantity $H_\lambda(\Gamma_{n,\mathbb{C}}(\lambda))$ comes in. However, $H_\lambda(\Gamma_{n,\mathbb{C}}(\lambda))$ can uniformly be approximated by the non-random quantity $H_\lambda(\Gamma_{\mathbb{C}}(\lambda))$ with outer probability 1 (Lemma 6.1 (a)).

It remains to consider the case $\lambda_0 = \infty$. Assume that $\limsup_n F(D_{n,\mathbb{C}}(\lambda_n)\,) = \limsup_n F(\Gamma_{n,\mathbb{C}}(\lambda_n)) > 0$. Then it follows that $\limsup_n \mathrm{Leb}(\Gamma_{n,\mathbb{C}}(\lambda_n)) > 0$ (because of the bounded Lebesgue density). Hence, for large enough n, i.e. for large enough λ_n, we have $\lambda_n\, \mathrm{Leb}(\Gamma_{n,\mathbb{C}}(\lambda_n)) > 1$. On the other

hand we have $0 \le E_{n,\mathbb{C}}(\lambda_n) = F_n(\Gamma_{n,\mathbb{C}}(\lambda_n)) - \lambda_n \,\mathrm{Leb}(\Gamma_{n,\mathbb{C}}(\lambda_n))$, and hence $\lambda_n \,\mathrm{Leb}(\Gamma_{n,\mathbb{C}}(\lambda_n)) \le 1$. This is a contradiction.
The proof of Theorem 3.3 is the same as the proof of Theorem 3.2 given for the case $\Lambda = \{\lambda\}$.

❑

Proof of Proposition 3.4: First note that

$$H_\lambda(\Gamma(\lambda)) - H_\lambda(C) = \int_{\Gamma(\lambda)} (f(x) - \lambda)\, dx - \int_C (f(x) - \lambda)\, dx$$

$$= \int_{\Gamma(\lambda)\setminus C} (f(x) - \lambda)\, dx - \int_{C \setminus\Gamma(\lambda)} (f(x) - \lambda)\, dx$$

$$= \int_{\Gamma(\lambda)\Delta C} | f(x) - \lambda |\, dx. \qquad (6.2)$$

Inequality (3.2b) follows directly from identity (6.2). To shorten the notation we write $D_{n,\mathbb{C}}(\lambda) = \Gamma_{n,\mathbb{C}}(\lambda) \,\Delta\, \Gamma_{\mathbb{C}}(\lambda)$, so that $F(D_{n,\mathbb{C}}(\lambda)) = d_F(\Gamma_{n,\mathbb{C}}(\lambda), \Gamma_{\mathbb{C}}(\lambda))$. In order to proof (3.2a) we write $F(D_{n,\mathbb{C}}(\lambda))$ as a sum of two terms:

$$F(D_{n,\mathbb{C}}(\lambda)) = F(D_{n,\mathbb{C}}(\lambda) \cap \{x: | f(x) - \lambda | < \eta\}) + F(D_{n,\mathbb{C}}(\lambda) \cap \{x: | f(x) - \lambda | \ge \eta\}).$$

The first term on the right-hand side is dominated by $F\{x: | f(x) - \lambda | < \eta\}$. As for the second term, (6.1) says that

$$H_\lambda(\Gamma(\lambda)) - H_\lambda(\Gamma_{n,\mathbb{C}}(\lambda)) \le (F_n - F)\, (\Gamma_{n,\mathbb{C}}(\lambda)) - (F_n - F)\, (\Gamma(\lambda)).$$

Thus, because of $f \le M$, (3.2a) follows from

$$H_\lambda(\Gamma(\lambda)) - H_\lambda(\Gamma_{n,\mathbb{C}}(\lambda)) = \int_{D_{n,\mathbb{C}}(\lambda)} | f(x) - \lambda |\, dx$$

$$\ge \eta \,\mathrm{Leb}(D_{n,\mathbb{C}}(\lambda) \cap \{x: | f(x) - \lambda | \ge \eta \}).$$

❑

Proofs of Theorem 3.6, Theorem 3.7: Let $\{\delta_n\}$, $\{\eta_n\}$ be sequences of positive real numbers and define

$$B_n = \{ \exists\, C, D \in \mathbb{C} \textit{ such that } d_F(C,D) > \delta_n \textit{ and } d_F(C,D) \le C\, \eta_n^{\gamma} + \eta_n^{-1} M\, [(F_n - F)(D) - (F_n - F)(C)] \}.$$

Then it easily follows from (3.2a) that for all sequences $\{\delta_n\}$, $\{\eta_n\}$ we have

$$P^*[\sup_{\lambda\in\Lambda} d_F(\Gamma_{n,\mathbb{C}}(\lambda),\Gamma(\lambda)) > \delta_n] \le P^*(B_n).$$

Hence, we shall look for the "smallest" sequence $\{\delta_n\}$ such that $P^*(B_n) \to 0$. Now we have

$$P^*(B_n)$$

$$\le P^*\Big(\sup_{d_F(C,D) > \delta_n} | (C\, \eta_n^{\gamma} + \eta_n^{-1} M [(F_n - F)(D) - (F_n - F)(C)]) / d_F(C,D) | > 1\Big)$$

$$\le \quad P^*\Big(\sup_{d_F(C,D) > \delta_n} | [(F_n - F)(D) - (F_n - F)(C)] / M\, \eta_n\, d_F(C,D) | > 1/2 \Big)$$

$$+ \; P^*\Big(\sup_{d_F(C,D) > \delta_n} | C\, \eta_n^{\gamma} / d_F(C,D) | > 1/2 \Big) \quad = \quad I + II.$$

If we choose $\eta_n = \delta_n^{1/\gamma} / 2C$ then II equals zero and it remains to determine $\{\delta_n\}$such that (with this choice of η_n) I tends to zero as $n \to \infty$. Note that

$$\{d_F(C,D) > \delta_n\} \;=\; \bigcup_{k=0}^{k_n} \{2^k \delta_n < d_F(C,D) \le 2^{k+1} \delta_n \},$$

where k_n is chosen as the smallest integer such that $2^{k_n+1} \delta_n \ge 1$. Since the right-hand side is a union of disjoint sets it follows that

$$I \;\le \sum_{k=0}^{k_n} P^*\Big(\sup_{d_F(C,D) \le 2^{k+1}\delta_n} | (F_n - F)(D) - (F_n - F)(C) | > 2^k\, \delta_n^{(1+\gamma)/\gamma} / 4\, M\, C\Big)$$

$$\le \sum_{k=0}^{k_n} P^*\Big(\sup_{A \in (\mathbb{C}\backslash\mathbb{C})_n} | \nu_n(A) | > 2^{k+1} n^{1/2}\, \delta_n^{(1+\gamma)/\gamma} / 16\, M\, C\Big) \;=\; \sum_{k=0}^{k_n} p_{n,k}$$

where we define $(\mathbb{C}\backslash\mathbb{C})_n = \{ C\backslash D,\ C, D \in C,\ F(C\backslash D) < 2^{k+1} \delta_n \}$. Now we seek for conditions to ensure that the last sum converges to zero (as $n \to \infty$). The probabilities $p_{n,k}$ are exactly of the form which is considered in Alexander (1984). For VC-classes $\mathbb{C}$ he derives exponential inequalities of the form

$$P^*\big(\sup_{A \in \mathbb{C}} | \nu_n(A) | > N \big) \le 16 \exp(- (1-\varepsilon)\, \Psi(N,n,\alpha)),$$

where $\alpha \ge \sup_{A \in C} \{ F(A) (1 - F(A) \}$ and Ψ has to lie in a certain class of functions, including $\Psi(N,n,\alpha) = \Psi_2(N,n,\alpha) = N^2 / 2\alpha\ (1 + N/3n^{1/2}\alpha)$, which corresponds to "Bernstein's inequality" (which holds for a single A). Now we fix $\mathbb{C} = (\mathbb{C}\backslash\mathbb{C})_n$, $\varepsilon = 1/2$, $\Psi = \Psi_2$. For any k and n we have $2^{k+1} \delta_n > \sup_{A\in(\mathbb{C}\backslash\mathbb{C})_n} F(A) \ge \sup_{A\in(\mathbb{C}\backslash\mathbb{C})_n} \{ F(A) (1 - F(A) \}$. Therefore we take $\alpha = 2^{k+1} \delta_n$, and by definition of $p_{n,k}$ the quantity N corresponds to $2^{k+1} n^{1/2} \delta_n^{(1+\gamma)/\gamma} / 16\, M\, C$.

Now we split the proof. First we consider the situation of Theorem 3.6. It is easy to see that with $\delta_n = (n / \log n)^{\gamma/(2+\gamma)}$ conditions (2.20), (2.22), (2.23) of Alexander are fulfilled. Hence, Theorem 2.8 of Alexander (1984) gives the following bound for $p_{n,k}$:

$$p_{n,k} \qquad \le 16 \exp\{ - 2^k\, n\, \delta_n^{(2+\gamma)/\gamma} / 64\, M^2 (1 + \delta_n^{1/\gamma} / 24\, M) \}.$$

Hence, for large enough n

$$\sum_{k=0}^{k_n} p_{n,k} \qquad \le 16 \sum_{k=1}^{k_n} \exp\{ - 2^k \log n / 64\, M^2 \}$$

which converges to zero as $n \to \infty$. This proves Theorem 3.6. To finish the proof of Theorem 3.7 we use Corollary 2.4 of Alexander (1984). It is easy to check, that if we choose δ_n as the asserted rates of Theorem 3.7, then condition (2.7) of Alexander is fulfilled. The rest of the proof is the same as above in the situation of Theorem 3.6.

❏

Proof of Proposition 3.8: The idea of the proof is exactly the same as for the proofs of the Theorems 3.6 and 3.7 given above. The only difference is, that here we use inequality (3.2b) instead of (3.2a). Note that for the uniform distribution $\sup_{\lambda<M-\delta} F\{ x: |f(x) - \lambda| < \eta \} = 0$ for $\eta = \eta_0 < \delta$. Hence (3.2a) reduces to (3.2b) with an addition multiplicative constant $M\, \eta_0^{-1}$. Hence we define for a sequence $\{\delta_n\}$ of positive real numbers

$$B_n = \{\, \exists\, C, D \in \mathbb{C} \textit{ such that } d_F(C,D) > \delta_n \textit{ and}$$
$$d_F(C,D) \le M\, \eta_0^{-1}\, [(F_n - F)(D) - (F_n - F)(C)]\}.$$

where $\eta_0 = 1$ for the case of the support estimation (part (a) and (b)). With this definition of B_n the proof works as the proofs of the Theorems 3.6 and 3.7. One has to show, that

$$\sum_{k=0}^{k_n} P^*\big(\sup_{A \in (\mathbb{C}\backslash\mathbb{C})_n} |\, \nu_n(A)\,| \; > \; 2^{k+1}\, n^{1/2}\, \delta_n\, \eta_0 / 8\, M\big) \;\to 0 \quad \text{as } n \to \infty$$

if we choose δ_n as the rates asserted in Proposition 3.8. Here again one can use results of Alexander (1984) in ecaxtly the same way is the proofs of Theorems 3.6 and 3.7.

❑

Proofs of Section 4:

Proof of Theorem 4.1: Remember that for every $\lambda \ge 0$ we have

$$E_{n,\mathbb{C}}(\lambda) = H_{n,\lambda}(\Gamma_{n,\mathbb{C}}(\lambda)) = F_n(\Gamma_{n,\mathbb{C}}(\lambda)) - \lambda\, \mathrm{Leb}(\Gamma_{n,\mathbb{C}}(\lambda)).$$

From Theorem 3.3 we obtain that for every choice of the empirical generalized λ-clusters $\Gamma_{n,\mathbb{C}}(\lambda)$ there exists a sequence $\{\Gamma_{\mathbb{C}}(\lambda,n),\, n \in \mathbf{N}\} \subset M_{\mathbb{C}}(\lambda)$ such that $d_F(\Gamma_{n,\mathbb{C}}(\lambda),\Gamma_{\mathbb{C}}(\lambda,n))^* \to 0$ with probability 1.

Since every set $\Gamma_{\mathbb{C}}(\lambda,n)$ is a generalized λ-cluster it follows with $\widetilde{E}_{n,\mathbb{C}}(\lambda) := H_{n,\lambda}(\Gamma_{\mathbb{C}}(\lambda,n)) = F_n(\Gamma_{\mathbb{C}}(\lambda,n)) - \lambda\, \mathrm{Leb}(\Gamma_{\mathbb{C}}(\lambda,n)$ that

$$n^{1/2}\, (\, \widetilde{E}_{n,\mathbb{C}}(\lambda) - E_{\mathbb{C}}(\lambda)\,) = n^{1/2}\, (F_n - F)(\Gamma_{\mathbb{C}}(\lambda,n)).$$

It remains to show that $|\, n^{1/2}\, (E_{n,\mathbb{C}}(\lambda) - \widetilde{E}_{n,\mathbb{C}}(\lambda))\, | \; = \; o_{P*}(1)$ as $n \to \infty$. We have

$$\begin{aligned} 0 \quad &\le E_{n,\mathbb{C}}(\lambda) - \widetilde{E}_{n,\mathbb{C}}(\lambda) \\ &= H_\lambda(\Gamma_{n,\mathbb{C}}(\lambda)) - H_\lambda(\Gamma_{\mathbb{C}}(\lambda,n)) + (F_n - F)(\Gamma_{n,\mathbb{C}}(\lambda)) \; - (F_n - F)(\Gamma_{\mathbb{C}}(\lambda,n)) \\ &\le \; (F_n - F)(\Gamma_{n.\mathbb{C}}(\lambda)) \; - (F_n - F)(\Gamma_{\mathbb{C}}(\lambda,n)). \end{aligned}$$

Since $d_F(\Gamma_{n,\mathbb{C}}(\lambda),\Gamma_{\mathbb{C}}(\lambda,n))^* \to 0$ with probability 1 the assertion follows from the equicontinuity of the empirical process indexed by $\mathbb{C}$.

❑

Proof of Theorem 4.3: As in the proof of Theorem 4.1 it follows with $\widetilde{E}_{n,\mathbb{C}}(\lambda) = H_{n,\lambda}(\Gamma_{\mathbb{C}}(\lambda))$ that

$$
\begin{aligned}
| Z_{n,\mathbb{C}}(\lambda) - B_{n,\mathbb{C}}(\lambda) | &= | n^{1/2} (E_{n,\mathbb{C}}(\lambda) - \widetilde{E}_{n,\mathbb{C}}(\lambda))) | \\
&\leq (F_n - F)(\Gamma_{n,\mathbb{C}}(\lambda)) - (F_n - F)(\Gamma_{\mathbb{C}}(\lambda)).
\end{aligned}
$$

From the stochastic equicontinuity of ν_n the first assertion follows.

$B_{n,\mathbb{C}}(\cdot)$ is a random element in $D(\Lambda)$ because the sets $\Gamma(\lambda)$ are closed. The convergence of the finite dimensional distributions follows immediately from the multidimensional central limit theorem. The tightness follows from the continuity of $\lambda \to \Gamma(\lambda)$ (which is (3.3); see remark after Theorem 3.5) together with the asymptotic stochastic equicontinuity of ν_n indexed by $\mathbb{C}$.

❑

Proofs of Section 5:

Proof of Theorem 5.2 and Theorem 5.3: First we proof that under $\widetilde{H}_0$ we have

$$
E_{n,\mathbb{D}}(\lambda) - E_{n,\mathbb{C}}(\lambda) \leq (F_n - F)(\Gamma_{n,\mathbb{D}}(\lambda)) + (F_n - F)(\Gamma(\lambda)) \tag{6.3}
$$

Since under H_0 every set $\Gamma_{\mathbb{D}}(\lambda)$ is a generalized λ-cluster for $\mathbb{C}$ and $\mathbb{D}$, we have:

$$
\begin{aligned}
E_{n,\mathbb{D}}(\lambda) - H_\lambda(\Gamma_{\mathbb{D}}(\lambda)) &= H_\lambda(\Gamma_{n,\mathbb{D}}(\lambda)) - H_\lambda(\Gamma(\lambda)) \\
&\quad + (F_n - F)(\Gamma_{n,\mathbb{D}}(\lambda)) \leq (F_n - F)(\Gamma_{n,\mathbb{D}}(\lambda))
\end{aligned}
$$

and

$$
\begin{aligned}
E_{n,\mathbb{C}}(\lambda) - H_\lambda(\Gamma(\lambda)) &= H_{n,\lambda}(\Gamma_{n,\mathbb{C}}(\lambda)) - H_\lambda(\Gamma(\lambda)) \\
&\geq H_{n,\lambda}(\Gamma(\lambda)) - H_\lambda(\Gamma(\lambda)) = (F_n - F)(\Gamma(\lambda)),
\end{aligned}
$$

and (6.3) follows. Because of (6.3) we have for any real numbers β_n and δ_n that

$$\{ n^{1/2} \sup_{\lambda \in \Lambda} \Delta_n(\mathbb{C},\mathbb{D},\lambda) > \beta_n \} \cap \{ \sup_{\lambda\in \Lambda} d_F(\Gamma(\lambda),\Gamma_{n,\mathbb{C}}(\lambda)) \le \delta_n \} \subset \{ \sup_{A \in (\mathbb{C}\backslash\mathbb{C})_n} | \nu_n(A) | > 2 \beta_n \},$$

where $(\mathbb{C}\backslash\mathbb{C})_n = \{ C\backslash D, C, D \in C, F(C\backslash D) < \delta_n \}$. As in the proofs of Theorem 3.6 and Theorem 3.7 the assertions of Theorem 5.2 and Theorem 5.3 now follow by means of Theorem 2.8 and Corollary 2.4, respectively, of Alexander (1984).

❏

Proof of Theorem 5.4: Without loss of generality we assume $\mathrm{Leb}(C_0) = 1$, so that $F(D) = \mathrm{Leb}(D)$ for all $D \in \mathbb{D}_0$. We have for any class $\mathbb{C}$ that

$$n^{1/2} E_{n,\mathbb{C}}(\lambda) = \sup_{C \in \mathbb{C}}(\nu_n(C) - n^{1/2}(\lambda - 1)\mathrm{Leb}(C)).$$

Define $Z_{n,\mathbb{C}}(\lambda) = \sup_{C \in \mathbb{C}}(G_{\mathbb{C}}(C) - n^{1/2}(\lambda - 1)\mathrm{Leb}(C))$. Then we have

$$\sup_{\lambda \in \Lambda} | n^{1/2} (E_{n,\mathbb{D}_0}(\lambda) - Z_{n,\mathbb{D}_0}(\lambda)) | \le \sup_{D \in \mathbb{D}_0} | \nu_n(D) - B_{\mathbb{D}_0}(D) |$$

and it follows

$$\begin{aligned} & | n^{1/2} T_n(\mathbb{C},\mathbb{D}_0,\Lambda) - \sup_{\lambda \in \Lambda} (Z_{n,\mathbb{D}_0}(\lambda) - Z_{n,\mathbb{C}}(\lambda)) | \\ = \quad & | n^{1/2} \sup_{\lambda \in \Lambda} (E_{n,\mathbb{D}_0}(\lambda) - E_{n,\mathbb{C}}(\lambda)) - \sup_{\lambda \in \Lambda} (Z_{n,\mathbb{D}_0}(\lambda) - Z_{n,\mathbb{C}}(\lambda)) | \\ \le \quad & \sup_{D \in \mathbb{D}_0} | \nu_n(D) - B_{\mathbb{D}_0}(D) | + \sup_{C \in \mathbb{C}} | \nu_n(C) - B_{\mathbb{C}}(C) | . \end{aligned}$$

The assertion now follows by means of a continuity argument from the identity

$$\sup_{\lambda \in \Lambda} (Z_{n,\mathbb{D}_0}(\lambda) - Z_{n,\mathbb{C}}(\lambda)) = \sup_{\lambda \in \Lambda_n} (Z_{\mathbb{D}_0}(\lambda) - Z_{\mathbb{C}}(\lambda))$$

where $\Lambda_n = [n^{1/2}(\lambda_0 - 1), n^{1/2}(\lambda_1 - 1)]$ and $\Lambda = [\lambda_0, \lambda_1]$.

❏

Acknowledgements: This paper is a revised version of my Dissertation written in Heidelberg. I would like to thank my supervisor Professor D.W. Müller for drawing my attention to the excess mass approach and for his interest and support during the time of writing my thesis. Furthermore I thank the statistic group of the Universität Heidelberg, in particular W. Ehm and E. Mammen, for valuable discussions, hints and remarks concerning the subject.

References:

Andrews, D.F. Bickel, P.J. Hampel, F.R. Huber, P.J. Rodgers, W.H. and Tukey, J.W.(1972). Robust estimation of location: survey and advances. Princeton Univ. Press, Princeton, N.J.

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

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

Chernoff, H. (1964). Estimation of the mode. *Ann. Inst. Statist. Math.* **16** 31-41.

DeHardt, J. (1971). Generalizations of the Glivenco-Cantelli theorem. *Ann. Math. Statist.* **42** 2050-2055.

Dudley, R.M. (1974). Metric entropy of some classes of sets with differentiable boundaries. *Journal of Approximation Theory* **10** 227-236.

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

Dudley, R.M. (1984). A course on empirical processes, *École d'Été de Probatilités de Saint Flour XII-1982*, *Lecture Notes in Math.* **1097** 1-142. New York: Springer

Dudley, R.M. (1984). An extended wichura theorem, definitions of Donsker classes, and weighted empirical distributions. *Lecture Notes in Math.* **1153** 141-178. New York: Springer

Hartigan, J.A. (1975). Clustering algorithms. New York, London: Wiley

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

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

Müller, D.W. and Sawitzki, G. (1987). Using excess mass estimates to investigate the modality of a distribution. Preprint Nr.398, SFB 123, Universität Heidelberg

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

Müller, D.W. and Sawitzki, G. (1991b). Excess mass estimates and tests of multimodality. Preprint Nr. 632, SFB 123, Universität Heidelberg.

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

Pollard, D. (1984). Convergence of stochastic processes. Springer, New York.

Polonik,W. (1992). The excess mass approach to cluster analysis and related estimation procedures. Dissertation, Universität Heidelberg.

Rao, R.R. (1962). Relation between weak and uniform convergence of measures and applications. *Ann. Math. Statist.* **33** 659-680.

Rényi, A. and Sulanke, R. (1964). Über die konvexe Hülle von n zufällig gewählten Punkten II. *Z. Wahrsch. u. verw. Gebiete* **3** 138-147.

Robertson, T.J. & Cryer, J.D. (1974). An iterative procedure for estimating the mode. *J. Amer. Statist. Assoc.* **69** 1012-1016.

Sager, T.W. (1979). An iterative method for estimating a multivariate mode and isopleth. *J. Amer. Statist. Assoc.* **74** 329-339.

Schneider, R. (1988). Random approximation of convex sets. *Journal of Microscopy* **151** 211-227.