# The silhouette, concentration functions, and ML-density estimation under order restrictions

by

Wolfgang Polonik

Technical Report No.445
December 1995

Department of Statistics
University of California
Berkeley, Califonia

This work has been supported by the Deutsche Forschungsgemeinschaft.

# The silhouette, concentration functions, and ML-density estimation under order restrictions

Wolfgang Polonik
University of California, Berkeley

**Abstract**

Based on empirical Levy-type concentration functions a new graphical representation of the ML-density estimator under order restrictions is given. This representation generalizes the well-known representation of the Grenander estimator of a monotone density as the slope of the least concave majorant of the empirical distribution function. From the given representation it follows that a density estimator called silhouette which arises naturally out of the excess mass approach is the ML-density estimator under order restrictions. This fact brings in several new aspects to ML-density estimation under order restrictions. Especially, it provides new methods for deriving asymptotic results for ML-density estimators under order restrictions based on empirical process theory.

## 1 Introduction

In the present paper we give the connection of what is called excess mass approach and of ML-density estimation under order restrictions. The link between those two is established by means of certain empirical Levy-type concentration functions. Based on these concentration functions we derive a graphical representation of the ML-density estimator (MLE) under order restrictions. It turns out that this graphical representation is the same as the one of the silhouette, (and hence, that the silhouette is the MLE), where the silhouette is a density estimator which arises naturally out of the excess mass approach (see Section 2). This fact brings in several new aspects to ML-density estimation under order restrictions. A more philosophical aspect, for example, is given by the fact that the original motivation of the excess mass approach is measuring mass concentration which (at least at a first view) is not related to order restrictions or ML-density estimation. Another aspect comes in through the construction of the silhouette (see below). Their construction is completely different from the classical construction of the MLE under order restrictions based on (generalized) isotonic regression. One also obtains new methods to study the asymptotic behaviour of the MLE which are based on empirical process theory (see Section

5).

Estimating a density $f$ under order restrictions means estimating $f$ under the assumption that $f$ is monotone with respect to an order on the underlying measure space $(\mathcal{X}, \mathcal{A})$. Such order restrictions can be expressed via a class $\mathcal{C}$ of measurable sets ([3, 28]): given a (quasi-)order $\preceq$ (reflexive and transitive) there exists a class of sets $\mathcal{C} = \mathcal{C}_{\preceq}$ such that $f$ is monotone with respect to $\preceq$ iff $f$ is measurable with respect to $\mathcal{C}_{\preceq}$, this means, iff *level sets* $\Gamma(\lambda) = \{f \geq \lambda\}$, $\lambda \geq 0$, all are elements of $\mathcal{C}_{\preceq}$. Hence, order restrictions on $f$ can be reformulated as "$f \in \mathcal{F}_{\mathcal{C}}$" for appropriate classes $\mathcal{C}$, where

$$\mathcal{F}_{\mathcal{C}} = \{f : \int f(x)d\nu(x) = 1, \Gamma(\lambda) \in \mathcal{C} \textit{ for all } \lambda \geq 0\}$$

and where $\nu$ is some dominating measure on $(\mathcal{X}, \mathcal{A})$. MLEs under order restrictions based on $n$ i.i.d. observations have been derived and studied, among others, by Grenander [9], Robertson [27], Wegmann [31] and Sager [29]. It is well-known (cf. [27, 29]), that the structure on $\mathcal{X}$ given through $\preceq$ induces a structure on the corresponding class $\mathcal{C}$: it has to be a $\sigma$-lattice. $\mathcal{C}$ is called a $\sigma$-lattice if it contains $\mathcal{X}$ and $\emptyset$ and is closed under countable unions and intersections. A simple example is given by $\mathcal{C} = \mathcal{I}_0 = \{[0, x], x \geq 0\}$ which corresponds to the class of decreasing (left continuous) densities in $[0, \infty)$ with respect to the usual order on the real line. Another example for a $\sigma$-lattice which is not a $\sigma$-algebra is the class of intervals containing a given point, $x_0$, say. The corresponding class of densities is the class of unimodal densities with mode $x_0$. Discrete analogs are given by the classes $\{\{1, 2, \ldots, k\}, k \geq 1\}$ and $\{\{-k, \ldots, -1, 0, 1, \ldots, k\}, k \geq 0\}$, respectively.

The model $f \in \mathcal{F}_{\mathcal{C}}$ for some class of measurable subsets $\mathcal{C}$ also underlies the construction of the silhouette. However, there the classes $\mathcal{C}$ need not correspond to any order. $\mathcal{C}$ can in principle be completely arbitrary. We call a model assumption of the form $f \in \mathcal{F}_{\mathcal{C}}$ *shape restriction* given by $\mathcal{C}$. A standard choice for a shape restriction (which is not an order restriction) is the class of convex sets in $\mathbf{R}^d$. In this terminology the silhouette is a density estimator under shape restrictions which, as shown in this paper, is the MLE in $\mathcal{F}_{\mathcal{C}}$ if the shape restrictions actually are order restrictions.

Let us briefly indicate the principle difference between the construction of the silhouette and the classical construction of the MLE. First note, that a MLE $\hat{f}_n$ in $\mathcal{F}_{\mathcal{C}}$ based on an i.i.d. sample of size $n$ has to be of histogram type (see Lemma 4.2), that is, there exists a partition $\{A_1, \ldots, A_k\}$ of $\mathbf{R}^d$ such that $\hat{f}_n(x) = \#\{\textit{observations} \in A_i\}/n\, \nu(A_i)$, for all $x \in A_i$. Now, constructing the MLE using ideas of isotonic regression means constructing the sets $A_i$ by building them as unions of certain generating sets in $\mathcal{C}$. In contrast to that the silhouette is constructed by putting *estimated level sets* one on the top of each other. The sets $A_i$ then automatically pop up as symmetric differences of successive level sets. Hence, in constructing the silhouette, one does not look at the individual observations $X_i$ and hence on the horizontal "axis", but one builds the estimator in "moving up" the *vertical* axis.

As mentioned earlier, the proof of the fact that the silhouette is the MLE under order restrictions is based on a graphical representation of the MLE. This graphical representation is based on least concave minorants of certain Levy-type concentration functions. It generalizes the well-known representation of the Grenander density estimator of a monotone density on the real line as the slope of the concave majorant of the empirical distribution function (Grenander [9]). The concentration functions under consideration are defined through constrained maximization of certain functionals defined on $\mathcal{C}$. The corresponding maximizing sets (minimum volume sets and modal sets) serve as level set estimators and are used to build the silhouette as described above. The given graphical representation also immediately provides an algorithm for calculating the MLE (see Section 4).

Note that the dominating measure $\nu$ used in our here need not be Lebesgue or counting measure. This for example enables us to do the following: Suppose one wants to estimate the Lebesgue density of $F$ under the additional information that $h = f/g$ satisfies some order restriction where $g$ denotes Lebesgue density of some known measure $G$. Then the MLE of $f$ under this additional information is given by $\hat{f} = g\ \hat{h}$ where $\hat{h}$ is the MLE of $h$ under the corresponding order restriction with $\nu = G$. Hence, the results given in the present paper for $\hat{h}$ (as for example asymptotic rates) immediately can be translated into results about $\hat{f}$ also.

The present paper is organized as follows. In Section 2 we introduce the silhouette and give some of their properties. Section 3 deals with concentration function and the corresponding maximizing sets. Some properties of these objects are given. A characterization of the existence of the MLE under order restrictions in terms of these concentration functions is given is Section 4 where also the graphical representation of the MLE is presented. In Section 5 we derive rates of convergence for the silhouette under metric entropy conditions on $\mathcal{C}$. All the proofs are in Section 6.

# 2 The silhouette

If restricted to the continuous case, this is, the dominating measure is Lebesgue measure in $\mathbf{R}^d$, the first part of this section more or less is a short cut of Section 2 of Polonik [26]. Proofs of several facts given below can be found there. Although they are given there for the continuous case, they apply to the general case considered here also.

For any density $f : \mathcal{X} \to \mathbf{R}$ the following key equality holds:

$$f(x) = \int \mathbf{1}_{\Gamma(\lambda)}(x)\, d\lambda \qquad \forall x \in \mathcal{X}, \tag{1}$$

where $\mathbf{1}_C$ denotes the indicator function of a set $C$. The idea for the construction of the silhouette is to plug in estimators for $\Gamma(\lambda)$ into equation (1). As

estimators we use so-called *empirical generalized $\lambda$-clusters*. They are defined as follows: Let $X_1, X_2, \ldots$ denote i.i.d observations from a distribution $F$ which has a density $f$ with respect to $\nu$. Let $F_n$ denote the empirical measure based on the first $n$ observations, i.e. $n\, F_n(C) = \#\{X_i \in C, i = 1, \ldots, n\}$ and define the signed measure

$$H_{n,\lambda} = F_n - \lambda\nu.$$

**Definition 2.1** *Any set $\Gamma_{n,\mathcal{C}}(\lambda) \in \mathcal{C}$ such that*

$$H_{n,\lambda}(\Gamma_{n,\mathcal{C}}(\lambda)) = \sup_{C \in \mathcal{C}} H_{n,\lambda}(C) \tag{2}$$

*is called an* empirical generalized $\lambda$-cluster in $\mathcal{C}$.

The sets $\Gamma_{n,\mathcal{C}}(\lambda)$ are called *generalized* since they need not be connected, as one would expect for clusters. Nevertheless, for briefity, we omit the word *general* and call the sets $\Gamma_{n,\mathcal{C}}(\lambda)$ empirical $\lambda$-clusters or sometimes just $\lambda$-clusters. Hartigan [12] used the notion $\lambda$-cluster for *connected* components of level sets. Note that the notion $\lambda$-cluster is in general used for the collection of all $\lambda$-clusters, that is, for the collection of $\lambda$-clusters at all levels $\lambda \geq 0$. Sometimes, however, we consider a single level $\lambda$. We hope it becomes sufficiently clear out of the context.

The motivation for defining $\Gamma_{n,\mathcal{C}}(\lambda)$ as above is given by the following equality. Let $H_\lambda = F - \lambda\nu$ then it is easy to see that

$$H_\lambda(\Gamma(\lambda)) = \sup\{H_\lambda(C), C \in \mathcal{A}\}. \tag{3}$$

This equation gives a justification for regarding the sets $\Gamma_{n,\mathcal{C}}(\lambda)$ as estimators for the level sets $\Gamma(\lambda)$ if they lie in $\mathcal{C}$. Note that if $\nu$ is a continuous measure then the supremum of $H_{n,\lambda}$ over all measureable sets equals one. Hence, besides the fact that the class $\mathcal{C}$ is used to introduce shape restrictions it makes sense in general to restrict the supremum to certain subclasses $\mathcal{C}$.

As a function of $\lambda$ the maximal value in (3), i.e. $E(\lambda) = E_F(\lambda) = H_\lambda(\Gamma(\lambda))$, is called *excess mass function*. Note that $E(\lambda)$ is used in majorisation orderings. There two distributions $F$ and $G$ with Lebesgue densities $f$ and $g$, respectively, are ordered by comparing their excess mass functions. If $E_F(\lambda) \leq E_G(\lambda) \forall \lambda \geq 0$ then $G$ is said to majorize $F$ ([15]; see [16] for a brief overview). Actually all this is formulated in terms of densities. The representation (3) however, gives a way to express this in terms of distributions, without using densities explicitely.

The maximal value in (2), i.e.

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

is called *empirical excess mass* at level $\lambda$. Hartigan [13] and Müller and Sawitzki [19] independently introduced the excess mass approach which is based on the idea (motivated by equation (3)) that maximizing the signed measure $H_{n,\lambda}$

gives information about mass concentration of the underlying distribution. The notion *excess mass* has first been used by Müller and Sawitzki. For further work on the excess mass and on empirical $\lambda$-clusters see Nolan [21], Müller and Sawitzki [20], and Polonik [23, 25].

In all of what follows it is assumed that $\mathcal{C}$ is such that

**(A1)** $\emptyset \in \mathcal{C}$

**(A2)** *for any* $\lambda \geq 0$ *there exists an empirical* $\lambda$*-cluster*

**(A3)** *almost surely there exists a set* $S \in \mathcal{C}$ *with* $\nu(S) < \infty$ *and* $F_n(S) = 1$

Empirical $\lambda$-clusters of course exist for finite $\mathcal{X}$ with $\mathcal{A} = \mathcal{P}^{\mathcal{X}}$ and $\nu$ the counting measure. In the continuous case, i.e. if $\nu$ is Lebesgue measure and $\mathcal{A}$ the Borel $\sigma$-algebra, empirical $\lambda$-clusters exist for standard classes $\mathcal{C}$ like $\mathcal{I}^d$, $\mathcal{B}^d$, $\mathcal{E}^d$, and $\mathcal{C}^d$ which denote the classes of all closed intervals, balls, ellipsoids and convex sets in $\mathbf{R}^d$, respectively. A general sufficient condition for the existence of empirical $\lambda$-clusters is that $\mathcal{C}$ is closed under intersections. Of course this condition is not necessary. The assumption $\emptyset \in \mathcal{C}$ assures that the empirical excess mass is nonnegative (as it should be). (**A3**) means that a.s. there exist empirical-$\lambda$ clusters with non-degenerate Lebesgue measure and it follows from (**A3**) together with the fact that the $\nu$-measures of the empirical $\lambda$-clusters are decreasing in $\lambda$ that all empirical $\lambda$-clusters have finite $\nu$-measure.

The sets $\Gamma_{n,\mathcal{C}}(\lambda)$ need not be uniquely determined. It even may happen that there exist empirical $\lambda$-clusters for the same $\lambda$ which carry different empirical mass and hence have also different Lebesgue measure. However, the sets $\Gamma_{n,\mathcal{C}}(\lambda)$ can be chosen such that the following property **(P)** holds:

**(P)** there exist levels $0 = \lambda_0 < \lambda_1 < \ldots < \lambda_{k_n}, k_n \leq n$ such that $\nu(\Gamma_{n,\mathcal{C}}(\lambda_{k_n})) = 0$ and that the function $\lambda \to \Gamma_{n,\mathcal{C}}(\lambda)$, $\lambda \geq 0$, is constant at the intervals $(\lambda_{j-1}, \lambda_j], j = 1, \ldots, k_n$ and has different values on different such intervals.

Property **(P)** actually is not necessary for proving asymptotic results for the silhouette. Without **(P)**, however, the silhouette may look quite erratic (this does not happen for $\sigma$-lattices $\mathcal{C}$, cf. Lemma 2.1 below). Any choice of empirical $\lambda$-clusters satisfying **(P)** automatically have the property that for any fixed $\mu > 0$ the $\nu$-measure of $\Gamma_{n,\mathcal{C}}(\mu)$ is maximal among all empirical $\mu$-clusters. A way to find the values $\lambda_i$ of **(P)** is given by means of the graphical representation of the silhouette (cf. end of Section 3).

For every choice of sets $\Gamma_{n,\mathcal{C}}(\lambda)$ satisfying **(P)** we define (a version of) the silhouette as

$$f_{n,\mathcal{C}}(x) = \int \mathbf{1}_{\Gamma_{n,\mathcal{C}}(\lambda)}(x)\, d\lambda \quad \forall x \in \mathcal{X}. \tag{4}$$

The definition of the silhouette depends on the special choice of sets $\Gamma_{n,\mathcal{C}}(\lambda)$. This gives different *versions* of the silhouette. These versions might differ on

sets with positive $\nu$-measure. However, all the results given below hold for any of these versions. We do not mention this further and only speak of "the" silhouette.

Under **(P)** the silhouette can be written as

$$f_{n,\mathcal{C}}(x) = \sum_{j=0}^{k_n-1} (\lambda_{j+1} - \lambda_j) \mathbf{1}_{\Gamma_{n,\mathcal{C}}(\lambda_j)}(x). \tag{5}$$

Hence, if in addition the sets $\Gamma_{n,\mathcal{C}}(\lambda_j), j = 1, \ldots, k_n$, are monotonically decreasing for inclusion, i.e. $\Gamma_{n,\mathcal{C}}(\lambda_{j+1}) \subset \Gamma_{n,\mathcal{C}}(\lambda_j)$, then $f_{n,\mathcal{C}}$ can be visualized as putting the slices $\Gamma_{n,\mathcal{C}}(\lambda_j) \times (\lambda_j, \lambda_{j+1}]$ one on top of the other. The empirical $\lambda$-clusters can be chosen to be monotone if $\mathcal{C}$ is a $\sigma$-lattice (see Lemma 2.1). Unfortunately, however, the monotonicity of the empirical $\lambda$-clusters does not necessarily hold for non- $\sigma$-lattices $\mathcal{C}$ like $\mathcal{I}^1$ or $\mathcal{C}^d$. This means that for non-$\sigma$-lattices $\mathcal{C}$ the silhouette does not necessarily lie in the model class $\mathcal{F}_{\mathcal{C}}$.

**Lemma 2.1** *If $\mathcal{C}$ is a $\sigma$-lattice then*

$$\Gamma_{f_{n,\mathcal{C}}}(\lambda) \in \mathcal{C} \quad \forall \lambda \geq 0.$$

*Moreover,*

$$\nu(\Gamma_{n,\mathcal{C}}(\lambda_{j+1}) \backslash \Gamma_{n,\mathcal{C}}(\lambda_j)) = 0, \qquad \forall\, j = 0, \ldots, k_n - 1,$$

*and the empirical $\lambda$-clusters can be chosen such that*

$$\Gamma_{n,\mathcal{C}}(\lambda_{j+1}) \subset \Gamma_{n,\mathcal{C}}(\lambda_j), \qquad \forall\, j = 0, \ldots, k_n - 1.$$

Note that the first assertion of Lemma 2.1 does not say that $f_{n,\mathcal{C}} \in \mathcal{F}_{\mathcal{C}}$ which in addition requires $\int f_{n,\mathcal{C}} = 1$. In fact it might happen that $\int f_{n,\mathcal{C}} < 1$ and even $\int f_{n,\mathcal{C}} = 0$. This is closely connected to the existence of an MLE in $\mathcal{F}_{\mathcal{C}}$ (see Theorem [?] below). Note that $f_{n,\mathcal{C}} \in \mathcal{F}_{\mathcal{C}}$ of course is a necessary condition for $f_{n,\mathcal{C}}$ to be a maximum likelihood estimator in $\mathcal{F}_{\mathcal{C}}$.

A density $f \in \mathcal{F}_{\mathcal{C}}$ is called MLE in $\mathcal{F}_{\mathcal{C}}$ iff

$$\prod_{i=1}^{n} f(X_i) = \sup_{g \in \mathcal{F}_{\mathcal{C}}} \prod_{i=1}^{n} g(X_i) < \infty.$$

Now we state one of the main theorems.

**Theorem 2.1** *Let $\mathcal{C}$ be a $\sigma$-lattice. If a MLE in $\mathcal{F}_{\mathcal{C}}$ exists, then*

$$f_{n,\mathcal{C}} \in argmax_{f \in \mathcal{F}_{\mathcal{C}}} \prod_{i=1}^{n} f(X_i).$$

**Remarks:** (*i*) It is well known that the MLE in $\mathcal{F}_{\mathcal{C}}$ exists a.s. for $\mathcal{C} = \mathcal{I}_0$ and, more generally, for each class of intervals containing a given point $x_0$ (the mode) where $\nu$ is Lebesgue measure. The MLE in $\mathcal{I}_0$ is the Grenander estimator of a monotone density. Hence, Theorem 2.1 says that the silhouette corresponding to $\mathcal{I}_0$ is the Grenander estimator. This fact has already been shown in Polonik [26].
(*ii*) Theorem 2.1 also says that the silhouette equals the (multivariate) unimodal MLEs considered in Sager [29] (which exist a.s. (see Theorem 4.2)). For modeling unimodality Sager actually used two different classes of sets both of which are $\sigma$-lattices. One of these classes is given by the class of ellipsoids with known or estimated location and scale parameters and the other is a $\sigma$-lattice $\mathcal{S}$ defined through the following property: $S \in \mathcal{S}$ iff $x \in S$ implies $[0, x] \in \mathcal{S}$, where $[0, x]$ denotes the d-dimensional interval $[0, x_1] \times [0, x_2] \ldots \times [0, x_d]$, and $x = (x_1, \ldots x_d)$. This class $\mathcal{S}$ corresponds to unimodal densities in higher dimensions with mode 0.
(*iii*) In the discrete case there of course also exist well-known MLEs under order restrictions. Consider for example a multinomial distribution on $\mathcal{X} = \{x_1, \ldots, x_k\}, x_i \in \mathbf{R}$ with corresponding probabilies $p_i = P\{x_i\}, i = 1, \ldots, k$. The MLE under the restriction that the $p_i$'s are monotone can for example be found in [3]. Without loss of generality let the $x_i$ be ordered, and let $\mathcal{C} = \{\{x_1, \ldots, x_j\}, j = 1, \ldots, k\}$. The corresponding silhouette is the MLE. This follows from Theorem 2.1.

# 3 Concentration functions

Besides the (empirical) excess mass function which has been used in the previous section to define the silhouette we now consider two more concentration functions, $q_n$ and $\tilde{F}_n$. They will be used to formulate the graphical representation and an existence theorem of the MLE. They are defined as:

$$q_n(\alpha) = \inf_{C \in \mathcal{C}} \{\nu(C) : F_n(C) \geq \alpha\}, \quad \alpha \in [0, 1] \tag{6}$$

and

$$\tilde{F}_n(l) = \sup_{C \in \mathcal{C}} \{F_n(C) : \nu(C) \leq l\}, \quad l \geq 0. \tag{7}$$

$q_n$ is a generalized quantile function in the sense of Einmahl and Mason [8] (see Polonik [24] for weak Bahadur-Kiefer approximations of the normalized $q_n$ and for tests of multimodality based on $q_n$). The function $\tilde{F}_n$ is an empirical Levy-type concentration function (see [14]). It has recently been used in [2] for constructing test for multimodality. Any set $C_n(\alpha) \in \mathcal{C}$ such that

$$q_n(\alpha) = \nu(C_n(\alpha))$$

is called an *(empirical) minimum volume (MV) set* in $\mathcal{C}$ at level $\alpha$ with respect to $\nu$. Any set $M_n(l) \in \mathcal{C}$ such that

$$\tilde{F}_n(l) = F_n(M_n(l))$$

is called *(empirical) modal set* in $\mathcal{C}$ at level $l$ with respect to $\nu$. Given observations $X_1, \ldots, X_n$ the set of all MV-sets at level $\alpha$ is denoted by $\mathcal{MV}_n(\alpha)$, and $\mathcal{MV}_n = \cup_{\alpha \in [0,1]} \mathcal{MV}_n(\alpha)$ denotes the set of all MV-sets. Analogously, let $\mathcal{MO}_n(l)$ and $\mathcal{MO}_n$ denote the sets of modal sets at level $l$ and the set of all modal sets, respectively.

The notion *minimum volume* set of course is motivated by the case $\nu = Leb$, where *Leb* denotes Lebesgue measure (in $\mathbf{R}^d$). A special case of a MV-set is the well-known shorth which is the MV-set in the class of 1-dim. intervals at the level 1/2. For this class of 1-dim. intervals $q_n$ has been considered by Gruebel [11]. Chernoff [4] used the midpoint of modal intervals, i.e. modal sets in the class of 1-dim. intervals, as estimators of the mode. Note that in the literature the notion modal set is also used in a more broader sense, such that for example MV-sets are sometimes called modal sets also (see, for example, Lientz [17]).

We assume that $\mathcal{C}$ is such that

**(A 4)** *almost surely there exist MV-sets and modal sets with finite $\nu$-measure for every $\alpha \in [0,1]$ and $l \geq 0$, respectively.*

**(A4)** can for example be assured if in addition to the assumptions given above $\mathcal{C}$ is closed under intersections. This closedness of course is not a necessary condition for **(A4)** to hold, as can be seen from the case $\mathcal{C} = \mathcal{E}^2$.

If **(A4)** holds, then we have $q_n(\alpha) \leq l \Leftrightarrow \tilde{F}_n(l) \geq \alpha$. However, for given observations, the class of all MV-sets does not coincide with the class of all modal sets, in general. Consider for example the case $\mathcal{X} = [0,1]$, $\nu =$ Lebesgue measure and let $\mathcal{C} = \{\emptyset, [0,1/2), [1/2,1], \mathcal{X}\}$. Let $\alpha_1 = F_n([0,1/2))$, and $\alpha_2 = F_n([1/2,1])$. If $\alpha_1 \neq \alpha_2$ then either $[0,1/2)$ or $[1/2,1]$ is not a modal set, depending on whether $\alpha_1 < \alpha_2$ or $\alpha_1 > \alpha_2$. But in any case all sets in $\mathcal{C}$ are MV-sets. In general we have:

**Lemma 3.1** *Given observations $X_1, \ldots, X_n$ the following are equivalent:*

*(i)* $\exists\, \Gamma \in \mathcal{MV}_n \cap \mathcal{MO}_n$ *with* $\nu(\Gamma) = l, F_n(\Gamma) = \alpha$

*(ii)* $\tilde{F}_n^*$ *is discontinuous at* $l$ *and* $F_n^*(l) = \alpha$

*(iii)* $q_n$ *is discontinuous at* $\alpha$ *and* $q_n(\alpha) = l$.

Note that by definition

$$\Gamma_{n,\mathcal{C}}(\lambda) \ \in \mathcal{MV}_n(F_n(\Gamma_{n,\mathcal{C}}(\lambda)) \bigcap \mathcal{MO}_n(\nu(\Gamma_{n,\mathcal{C}}(\lambda)). \tag{8}$$

Therefore, it follows from Theorem 2.1 that for $\sigma$-lattices $\mathcal{C}$ every level set of the MLE in $\mathcal{F}_{\mathcal{C}}$ is both, (empirical) MV-set and modal set. However, not every set which is both, MV-set and modal set is an empirical $\lambda$-cluster (see below). In general the set of all empirical $\lambda$-clusters is much smaller than $\mathcal{MV}_n \cap \mathcal{MO}_n$. It also follows from (8) that assumption **(A4)** implies **(A2)** and **(A3)**.

*Theoretical* MV-sets and modal sets can be defined analogously to the sets $C_n(\alpha)$ and $M_n(l)$ as maximizers of corresponding theoretical concentration functions. These theoretical concentration functions are defined through replacing the empirical measure by the true measure $F$ in the definitions (6) and (7), respectively. The level sets of the underlying density $f$ are both, (theoretical) MV-sets and modal sets, provided all level sets lie in $\mathcal{C}$. MV-sets as estimators of level sets are studied in Polonik [24].

Now we give the connection of the excess mass functional and $\tilde{F}_n$. To that end define

$$\tilde{F}_n^* = \textit{least concave majorant of } \tilde{F}_n,$$

where the least concave majorant of a function $g$ is defined to be the smallest concave function lying above $g$. Note that $\tilde{F}_n$ is a piecewise constant, increasing function bounded by one with at most $n+1$ different values. Hence, $\tilde{F}_n^*$ is a convex function which is piecewise linear, increasing, bounded by one with at most $n$ changes of slope. Therefore, for every given $\lambda \geq 0$ there exists a tangent (from above) to $\tilde{F}_n^*$ which has slope $\lambda$. The connection of $\tilde{F}_n$ and the empirical excess mass function is given through this tangent:

**Lemma 3.2** *For each fixed $\lambda \geq 0$ the empirical excess mass $E_{n,\mathcal{C}}(\lambda)$ equals the intercept of the tangent (from above) with slope $\lambda$ to $\tilde{F}_n^*$.*

Closely related to that fact is the graphical representation of the silhouette given in Polonik [26]: the at most $n$ different positive values $\lambda_1, \ldots, \lambda_{k_n}$ of **(P)** (see(5)) are given by the different slopes (left-hand derivatives) of $\tilde{F}_n^*$. The corresponding modal sets (which also are MV-sets) at the levels where the slope changes are the empirical $\lambda$-clusters $\Gamma_{n,\mathcal{C}}(\lambda_i), i = 1, \ldots, k_n$. For $\sigma$-lattices $\mathcal{C}$ the values $\lambda_i$ and the corresponding sets $\Gamma_{n,\mathcal{C}}(\lambda_i)$ are the different values and level sets of the silhouette. The same graphical representation holds for the MLE in $\mathcal{F}_{\mathcal{C}}$, provided $\mathcal{C}$ is a $\sigma$-lattice (see Theorem 4.1). This fact then proves Thorem 2.1.

# 4 A graphical representation of the MLE

We start this section with two properties of the MLE in $\mathcal{F}_{\mathcal{C}}$. Both will be used to prove the graphical representation of the MLE given below (Theorem 4.1). However, they also have some interest for their own.

**Lemma 4.1** *If $f_n^*$ is a MLE in $\mathcal{F}_{\mathcal{C}}$ then*

$$\frac{1}{n} \sum_{\{i:X_i \in C\}} \frac{1}{f_n^*(X_i)} \leq \nu(C) \tag{9}$$

*for all $C \in \mathcal{C}$ such that $(f_n^* + \epsilon \mathbf{1}_C)/(1 + \epsilon\nu(C)) \in \mathcal{F}_{\mathcal{C}}$ for $\epsilon > 0$ small enough. If $\mathcal{C}$ is a $\sigma$-lattice, then (9) holds for all $C \in \mathcal{C}$.*

It is well-known, that a MLE in $\mathcal{F}_{\mathcal{C}}$ does not exist if there exist sets $C \in \mathcal{C}$ with $F_n(C) > 0$ and arbitrary small $\nu$-measure. This can also been seen from Lemma 4.1. Moreover, Lemma 4.1 also gives a quantification of this fact. It is an easy consequence of (9) that $\max_x f_n^*(x) \geq 1/(n\epsilon)$, if $\mathcal{C}$ is a $\sigma$-lattice with $\inf\{\nu(C) : C \in \mathcal{C}\} \leq \epsilon$.

Another property of the MLE in $\mathcal{F}_{\mathcal{C}}$ is the following. Let for a subset $\pi \subset \{1, \ldots, n\}$ denote $X^\pi = \{X_i : i \in \pi\}$. Then we have:

**Lemma 4.2** *Suppose that $\mathcal{C}$ is closed under intersections. Given $X_1, \ldots, X_n$ let $\mathcal{L}_n = \{L \in \mathcal{C} : L = \bigcap\{C \in \mathcal{C} : X^\pi \subset C\}$ for some subset $\pi \subset \{1, \ldots, n\}\}$. For any function $f \in \mathcal{F}_{\mathcal{C}}$ with $\prod_{i=1}^n f(X_n) > 0$ there exists a function $f^* \in \mathcal{F}_{\mathcal{L}_n}$ with $\prod_{i=1}^n f^*(X_i) \geq \prod_{i=1}^n f(X_i)$. Hence, if a MLE $f_n^*$ in $\mathcal{F}_{\mathcal{C}}$ exists, then we have*

$$f_n^*(x) \in \{\frac{F_n(A \setminus B)}{\nu(A \setminus B)} : A, B \in \mathcal{L}_n, B \subset A\}.$$

Note that the class $\mathcal{L}_n$ is finite (for a given realization $X_1, \ldots, X_n$) and that it contains all MV-sets in $\mathcal{C}$ with non-zero $\nu$-measure. We shall see later (Corollary 4.1), that for $\sigma$-lattices $\mathcal{C}$ the assertion of Lemma 4.2 holds with the class $\mathcal{L}_n$ replaced by the class of all MV-sets (or of all modal sets) which in general is much smaller. Lemma 4.2 not only says, that the MLE in $\mathcal{F}_{\mathcal{C}}$ is piecewise constant with at most (n+1) distinct levels and that it is of histogram type (which is well-known). It also gives a finite number of levels among which the levels of the MLE can be found and it gives the corresponding class of sets among which the sets can be found where the MLE is constant.

Now we formulate the graphical representation of the MLE which is based on $\tilde{F}_n^*$. It has already been mentioned in Section 3 that $\tilde{F}_n^*$ is a piecewise linear, increasing function with at most $n$ changes of slope. These changes of slope occure at levels $l$ where $\tilde{F}_n(l) = \tilde{F}_n^*(l)$. Let $l_1, \ldots, l_{k_n}$, $k_n \leq n$ denote those levels in decreasing order and denote by $s_i$ the left-hand derivatives of $\tilde{F}_n^*$ at $l_i$, $i = 1, \ldots, k_n$ (see Figure 1, below). Note that $s_i < s_{i+1}$, $i = 1, \ldots, k_n - 1$. Further denote $\alpha_i = \tilde{F}_n(l_i)$, $i = 1, \ldots, k_n$, such that $l_i$ is the $\nu$-measure of the MV-set at the level $\alpha_i$.

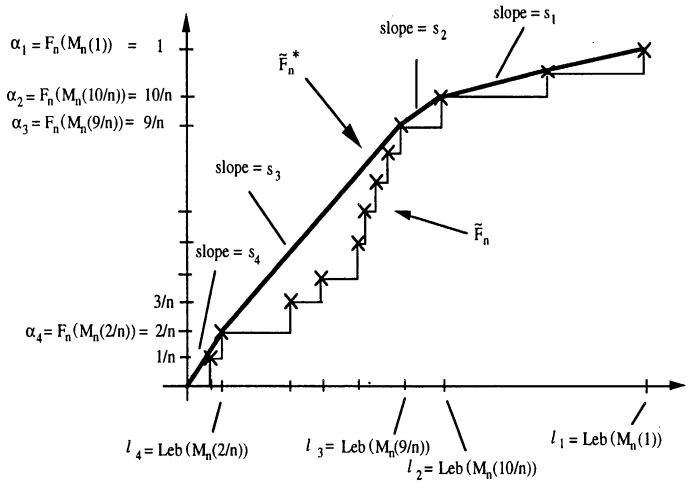


**Fig. 1:** The notation introduced above and used in Theorem 4.1 is illustrated with $\nu$ = Lebesgue measure, denoted by *Leb*. A possible realization of $\tilde{F}_n$ and $\tilde{F}_n^*$ for $n = 12$ is shown. Fig. 2 below gives a corresponding ML-density estimate (or silhouette).

Given an MLE $f_n^*$ in $\mathcal{F}_{\mathcal{C}}$ let $0 = f_0 < f_1 < \ldots < f_{k_n^*}, k_n^* \leq n$ denote the distinct levels of $f_n^*$ and let $\Gamma_{n,\mathcal{C}}^*(f_i)$ be their corresponding (distinct) level sets at the levels $f_i$.

**Theorem 4.1** *Suppose that* **(A1)** *and* **(A4)** *hold and that a MLE in $\mathcal{F}_{\mathcal{C}}$ exists. If $\mathcal{C}$ is a $\sigma$-lattice, then we have for any MLE $f_n^*$ with the above notation that $k_n^* = k_n$ and*

$$\textit{(i)}\ f_i = s_i \qquad \forall i = 1, \ldots, k_n$$

$$\textit{(ii)}\ \Gamma_{n,\mathcal{C}}^*(f_i) \in \mathcal{MV}_n(\alpha_i) \cap \mathcal{MO}_n(l_i) \qquad \forall i = 1, \ldots, k_n.$$

Theorem 4.1 $(i)$ says that the different values of the MLE are given by the slopes of the least concave majorant of $\tilde{F}_n$, and $(ii)$ says that the corresponding level sets are the modal sets at these levels. Since the silhouette has the same graphical representation (which, however, not only holds for $\sigma$-lattices, see comments after Lemma 3.2) this proves Theorem 2.1.

*An algorithm:* Theorem 4.1 immediately provides an algorithm to calculate the MLE: First calulate all the modal sets or, alternatively, all minimum volume sets $C_n(i/n), i = 0, \ldots, n$. Then the concave majorant corresponding to the points $(F_n(C_n(i/n)), \nu(C_n(i/n))), i = 0, \ldots, n$, gives the different levels and the corresponding level sets of the MLE as indicated in Theorem 4.1.

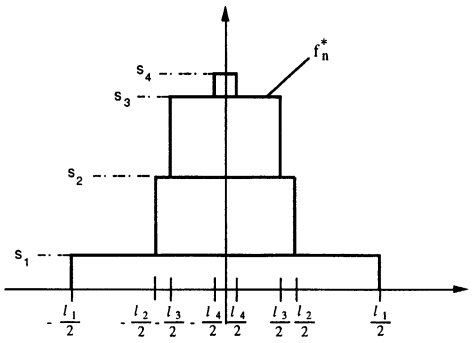


**Fig. 2:** The construction of an MLE (corresponding to Figure 1) as given in Theorem 4.1 is illustrated. The class $\mathcal{C}$ is chosen to be the class of all intervals with midpoint zero which is a $\sigma$-lattice. The four different level sets are the sets $M_n(2/n)$, $M_n(9/n)$, $M_n(10/n)$, and $M_n(1)$, respectively, corresponding to Fig.1.

The following corollary is an easy consequence of Lemma 4.2 and Theorem 4.1:

**Corollary 4.1** *Suppose that the assumptions of Theorem 4.1 hold. Let*

$$V_{\mathcal{MV}_n} = \{\frac{F_n(A \setminus B)}{\nu(A \setminus B)} : A, B \in \mathcal{MV}_n, B \subset A\}$$

*and*

$$V_{\mathcal{MO}_n} = \{\frac{F_n(A \setminus B)}{\nu(A \setminus B)} : A, B \in \mathcal{MO}_n, B \subset A\}.$$

*Then we have*

$$f_n^*(x) \in V_{\mathcal{MV}_n} \cap V_{\mathcal{MO}_n} \qquad \forall x \in \mathcal{X}.$$

The concentration functions $\tilde{F}_n$ and $q_n$ can also be used to characterise the existence of a MLE in $\mathcal{F}_{\mathcal{C}}$ which is an assumption in Theorem 4.1.

**Theorem 4.2** *Under* **(A1)** *and* **(A4)** *we have the following: Suppose that $\mathcal{C}$ is closed under intersection, then the following are equivalent:*

*(i) a MLE in $\mathcal{F}_{\mathcal{C}}$ exists*
*(ii)* $\lim_{\alpha \to 0} q_n(\alpha) > 0$
*(iii)* $\lim_{l \to 0} \tilde{F}_n(l) = 0$
*(iv)* $\int f_{n,\mathcal{C}}(x)\, dx = 1$

We want to stress here that the silhouette might nevertheless be a reasonable density estimator even if a MLE does not exist, as for example in the class $\mathcal{F}_{\mathcal{C}^d}$.

The well known fact, that there exists a.s. a MLE in the class of monotone decreasing densities on $[0, \infty)$ follows from Theorem 4.2, since the smallest MV-set in $\mathcal{I}_0$ which is $[0, X_{(1)}]$ has a.s. positive Lebesgue measure. Here $X_{(1)}$ denotes the first order statistic. Theorem 4.2 also says, that for $\mathcal{C} = \mathcal{I}^1$, or more general, for the classes $\mathcal{C} = \mathcal{C}^d$ the MLE in $\mathcal{F}_{\mathcal{C}}$ does not exist. However, if one for example removes all sets from $\mathcal{C}^d$ with Lebesgue measure bigger than a fixed positive $\epsilon$ (with the exception of the empty set), then the MLE exists (see [31] or [28]). Of course there exist other ways to modify the class $\mathcal{C}$ in order to ensure the existence of a MLE. For example, a datadependent approach is given by measuring the significance of a given set through the (empirical) excess mass it carries. More precisely, only consider sets $C \in \mathcal{C}$ with $H_{n,\lambda}(C) > \epsilon$. Since the value $H_{n,\lambda}(C)$ has some interpretation (cf. [19, 20, 25]), it should be easier to choose an $\epsilon$ in the latter case. A similar approach, also based on the excess mass, has been used by Müller and Sawitzki [20] in the context of the silhouette. The just mentioned approaches also reduce the well-known problem of *spiking* of the MLE (and of the silhouette) (cf. [31]).

# 5 Rates of convergence

In this section we give rates of convergence of the silhouette (and hence for MLEs under order restrictions). We use $L_1(\nu)$-distance, denoted by $\|\cdot\|_1$. We give rates of the silhouette under the only assumption on $f$ that $f \in \mathcal{F}_{\mathcal{C}}$, this means, that the model is correct.

The given rates depend on the richness of the underlying model, that is, the richness of $\mathcal{C}$. This richness is measured by bracketing covering numbers, or the metric entropy (with inclusion, or bracketing) which are the log-covering numbers. The bracketing covering numbers are defined as

$$N_B(\epsilon, \mathcal{C}, F) = \inf \{m \in \mathbf{N} : \exists C_1, \ldots, C_m \text{ measurable, such that for every } C \in \mathcal{C} \\ i, j \in \{1, \ldots, m\} \text{ with } C_i \subset C \subset C_j \text{ and } F(C_j \setminus C_i) < \epsilon\}.$$

The rates of convergence given below also depend on the tail behaviour of $f$ and the behaviour around the mode(s). These behaviours are measured here by the behaviour of $\bar{E}(\lambda) = 1 - E(\lambda)$ as $\lambda \to 0$ and $E(\lambda)$ as $\lambda \to \infty$, respectively. Geometrically $\bar{E}(\lambda)$ equals the "area" under $min\{f, \lambda\}$, i.e. $\bar{E}(\lambda) = \int_{\mathcal{X}} min(f(x), \lambda)\, d\nu(x)$. By using Fubini's theorem this can also be written as

$$\bar{E}(\lambda) = \int_0^{\lambda} \nu(\Gamma(\mu)) d\mu.$$

**Theorem 5.1** *Suppose that $\mathcal{C}$ is closed under intersection and satisfies $q_n(1) > 0$ and $\bar{F}_n(0) = O(n^{-1})$ a.s. If $f \in \mathcal{F}_{\mathcal{C}}$ then we have the following:*

*(a) Let $1 < M_n < n$. Suppose there exists constants $\gamma_1, A_1 > 0$ such that*

$$N_B(\epsilon, \mathcal{C}, F) \leq A_1 \, \epsilon^{-\gamma_1} \qquad \forall \epsilon > 0,$$

*then*

$$\| f_{n,\mathcal{C}} - f \|_1 = O_P( \bar{E} \, (\left( \frac{n}{M_n \log n} \right)^{-1/3} ) \vee E(M_n) ).$$

*(b) Suppose there exist constants $\gamma_2, A_2 > 0$ such that*

$$\log N_B(\epsilon, \mathcal{C}, F) \leq A_2 \, \epsilon^{-\gamma_2} \qquad \forall \epsilon > 0.$$

*(i) Case $\gamma_2 < 1$: Let $M_n = O(n^{1/4})$. Then*

$$\| f_{n,\mathcal{C}} - f \|_1 = O_P( \bar{E} \left( \frac{n}{M_n^{1-\gamma_2}} \right)^{-\frac{1}{\gamma_2+3}} ) \vee E(M_n)).$$

*(ii) Case $\gamma_2 \geq 1$:*

$$\| f_{n,\mathcal{C}} - f \|_1 = O_P(\bar{E}(n^{-\frac{1}{2(\gamma_2+1}}) \vee E(n^{(2\gamma_2-1)/2(\gamma_2+1)})).$$

*If $f$ is bounded, then the assumption that $\mathcal{C}$ is closed under intersection is not necessary.*

**Remarks:** *(i)* If $sup_x f(x) < M$, then we have $E(M_n) = 0$ for $M_n \geq M$. Hence, the above rates reduce to the rates given in Polonik [26], with the exception of case $\gamma_2 = 1$ in part (b). Here we are able to remove an unpleasant log-term which appears in [26].
*(ii)* Part (a) applies to the Grenander estimator of a monotone density on $[0, \infty)$, say, because the corresponding class $\mathcal{C}$ is the class $\mathcal{I}_0 = \{[0, x], x \geq 0\}$ which is a so-called Vapnik-Cervonenkis class (or a class with polynomial discrimination), and hence satisfies the condition on the covering numbers of part (a) for all $F$ (cf. Pollard [22]). Clearly $\bar{E}(\lambda) \leq \nu(S) \, \lambda$, as $\lambda \to 0$ if $f$ has bounded support $S$. Hence, in that case the above theorem gives for bounded $f$ an upper bound for the rates of convergence of the Grenander estimator of $(n/\log n)^{-1/3}$ under no smoothness assumption on the underlying density and without using the monotonicity at all. Only the fact that $\mathcal{I}_0$ is a Vapnik-Cervonenkis class enters the proof. Note that $n^{-1/3}$ is known to be the exact $L_1$-rate of the Grenander estimator if $f$ has bounded second derivative and compact support. This has been shown by Groeneboom [10].
*(iii)* Another example is given by the class $\mathcal{C}^d$, which is no $\sigma$-lattice. For this class we have for $d \geq 2$ that $\gamma_2 = (d-1)/2$. Hence, if $f$ is bounded and has bounded support, such that $\bar{E}(\lambda) = O(\lambda)$ then part (b) of the above theorem gives the rates $O_P(n^{-2/7})$ for $d = 2$ and $O_P(n^{-1/(d+1)})$ for $d \geq 3$. This has

already been mentioned in Polonik [26] (up to an additional log-term for $d = 3$). *(iv)* The assumptions made in Theorem 5.1 that $q_n(1) > 0$ and $\tilde{F}_n(0) = O(n^{-1})$ *a.s.* are fulfilled for all standard classes $\mathcal{C}$ mentioned in the present paper. These assumption (together with **(A3)**) imply that the silhouette integrates to $1 - \tilde{F}_n(0) = 1 - O(n^{-1})$. Actually the assumption that the silhouette integrates to $1 - O(\alpha_n)$ *a.s.*, where $\alpha_n$ is the rate asserted in Theorem 5.1, would be enough. This is equivalent to $q_n(1) > 0$ and $\tilde{F}_n(0) = O(\alpha_n)$ *a.s.*
*(v)* The rates given in Theorem 5.1 are not faster than $O_{P^*}(n^{-1/3}(\log n)^{1/3})$. However, for example for finite $\mathcal{X}$ the MLE under order restriction is known to converge at rate $O_P(n^{-1/2})$ (see for example [28]). We can rederive this rate up to an additional log-term for underlying densities which attain only finitely many different values. Here one needs the additional assumption that $\mathcal{C}$ is closed under finite unions und one has to combine the ideas of the proof of Theorem 3.5 of Polonik [26] with the ideas of the proof of Theorem 5.1 given below.

Finally we give a brief heuristic comparison of the rates given in Theorem 5.1 (a) and (b), case $\gamma_2 < 1$ with rates given in Theorem 2 of Wong and Severini [32] for MLEs in infinite dimensional parameter spaces. The comparison is not so easy, because of the different nature of the assumptions used. Moreover, different metrices are used in both papers. Wong and Severini use a certain $L_2$-type metric, called Fisher metric. For that reasons the given comparison is of very heuristic nature and leaves several open questions which perhaps seem worth for further investigations. We shall not do this here, since the rates of convergence are not our major interest. Nevertheless, the comparison provides interesting connections.

First note that in both cases metric entopy (log-covering numbers) determine the (upper bounds) for rates of convergence. In Theorem 5.1 we use metric entropy with bracketing of the underlying class $\mathcal{C}$ whereas Wong and Severini use metric entropy (without bracketing) of a class of score functions with respect to the sup-norm. In both cases the same upper bounds of the form $A\epsilon^{-r}$, $A, r > 0$, are used for the metric entopy (both use results of Alexander (1984)). The rates stated in Wong and Severini are of the form $n^{-1/(2+r)}$, where $r$ is the exponent in the bound for the log-covering numbers of the class of score functions. In contrast to that, the rates given above for $\gamma_2 < 1$ (and in part (a) (by ignoring the log-term)) are of the form $n^{-1/(3+\gamma_2)}$, where here $\gamma_2$ is the exponent in the upper bound for the metric entropy of $\mathcal{C}$. Now, a rough heuristic upper bound for the covering numbers of the class of *functions* $\mathcal{F}_\mathcal{C}$ is given by $N(\epsilon)^{1/\epsilon}$, where $N(\epsilon)$ denotes the covering number of $\mathcal{C}$. The heuristic holds, if $f$ is bounded an has bounded support otherwise the behaviour around the mode(s) and in the tails enters. The idea is to construct approximating functions as follows: divide the $y$-axis (levels) into a regular grid of distance $\epsilon$. This gives (assuming boundedness) of the order $1/\epsilon$ different levels. Approximate the level sets of a function $f \in \mathcal{F}_\mathcal{C}$ at these levels by the approximating sets corresponding to the covering number $N(\epsilon)$. If in addition $f$ has bounded support this leads

to approximating functions for $f$ at an $L_1(F)$-distance $\epsilon$. An upper bound for the number of these approximating function is of the order $N(\epsilon)^{1/\epsilon}$. Using $N(\epsilon) \leq A_2\epsilon^{-\gamma_2}$ gives an upper bound for the $L_1$-covering number of $\mathcal{F}_{\mathcal{C}}$ of the form $A\epsilon^{-\gamma^*}$ with $\gamma^* = \gamma_2 + 1$. Hence the rates in Theorem 5.1 are of the order $n^{-1/(2+\gamma^*)}$, which is of the same form as the rates in Wong and Severini. Examples in Wong and Severini are given where the covering numbers of the class of score functions can be bounded by the covering numbers of the class $\mathcal{F}_{\mathcal{C}}$. Hence, in the situation just discussed, the rates in Wong and Severini and the rates given here (for $\gamma_2 < 1$) are of the same nature.

# 6 Proofs

**Proof of Lemma 2.1:** Let for any $c > 0$

$$\hat{\Gamma}_n(c) = \{x : f_{n,\mathcal{C}}(x) \geq c\}.$$

Define further $J_c = \{\pi \subset \{0, \ldots, k_n - 1\} : \sum_{j\in\pi}(\lambda_{j+1} - \lambda_j) \geq c\}$. Then, since

$$x \in \hat{\Gamma}_n(c) \quad \Leftrightarrow \quad \exists \pi \in J_c : \; x \in \bigcap_{j\in\pi} \Gamma_{n,\mathcal{C}}(\lambda_j)$$

it follows that

$$\hat{\Gamma}_n(c) = \bigcup_{\pi\in J_c} \left( \bigcap_{j\in\pi} \Gamma_{n,\mathcal{C}}(\lambda_j) \right),$$

from which the first assertion follows.

To see that $\nu(\Gamma_{n,\mathcal{C}}(\lambda_{j+1}) \setminus \Gamma_{n,\mathcal{C}}(\lambda_j)) = 0$ assume that it actually is $> 0$. Since

$$H_{n,\lambda_j}(\Gamma_{n,\mathcal{C}}(\lambda_j) \cup \Gamma_{n,\mathcal{C}}(\lambda_{j+1})) = \tag{10}$$

$$H_{n,\lambda_j}(\Gamma_{n,\mathcal{C}}(\lambda_j)) + H_{n,\lambda_j}(\Gamma_{n,\mathcal{C}}(\lambda_{j+1}) \setminus \Gamma_{n,\mathcal{C}}(\lambda_j)), \tag{11}$$

and $\Gamma_{n,\mathcal{C}}(\lambda_j) \cup \Gamma_{n,\mathcal{C}}(\lambda_{j+1}) \in \mathcal{C}$ it follows by definition of the empirical $\lambda$-clusters as maximizers of the functional $H_{n,\lambda}$ that $H_{n,\lambda_j}(\Gamma_{n,\mathcal{C}}(\lambda_{j+1}) \setminus \Gamma_{n,\mathcal{C}}(\lambda_j)) \leq 0$. Hence, since $\lambda_j < \lambda_{j+1}$ and $\nu(\Gamma_{n,\mathcal{C}}(\lambda_{j+1}) \setminus \Gamma_{n,\mathcal{C}}(\lambda_j)) > 0$ (by assumption) it follows that

$$H_{n,\lambda_{j+1}}(\Gamma_{n,\mathcal{C}}(\lambda_{j+1}) \setminus \Gamma_{n,\mathcal{C}}(\lambda_j)) < 0.$$

On the other hand we have

$$\begin{aligned} &H_{n,\lambda_{j+1}}(\Gamma_{n,\mathcal{C}}(\lambda_j) \cap \Gamma_{n,\mathcal{C}}(\lambda_{j+1})) \\ &\quad = H_{n,\lambda_{j+1}}(\Gamma_{n,\mathcal{C}}(\lambda_{j+1})) - H_{n,\lambda_{j+1}}(\Gamma_{n,\mathcal{C}}(\lambda_{j+1}) \setminus \Gamma_{n,\mathcal{C}}(\lambda_j)) \\ &\quad > H_{n,\lambda_{j+1}}(\Gamma_{n,\mathcal{C}}(\lambda_{j+1})). \end{aligned}$$

Since $\mathcal{C}$ is closed under intersection this gives a contradition by definition of empirical $\lambda$-clusters.

These arguments also show how the empirical $\lambda$-clusters can be chosen in order to be monotone for inclusion. Namely, if actually $\Gamma_{n,\mathcal{C}}(\lambda_{j+1}) \backslash \Gamma_{n,\mathcal{C}}(\lambda_j) \neq \emptyset$ then replace $\Gamma_{n,\mathcal{C}}(\lambda_{j+1})$ by $\Gamma_{n,\mathcal{C}}(\lambda_{j+1}) \cap \Gamma_{n,\mathcal{C}}(\lambda_j)$.

**Proof of Lemma 3.1:** (i) $\Rightarrow$ (ii): Suppose (i) holds. If $\tilde{F}_n$ would have no jump at $l = F_n(\Gamma)$ then there exists a set $C \in \mathcal{C}$ with $\nu(C) = l_0 < l$ and $F_n(C) = \alpha = F_n(\Gamma)$. Hence, $q_n(\alpha) \leq l_0 < l$ and it follows that $\Gamma \notin \mathcal{MV}_n$. This is a contradition.
(ii) $\Rightarrow$ (iii): Suppose $\tilde{F}_n$ has a jump at $l$. If $q_n$ would have no jump at $\alpha = \tilde{F}_n(l)$ then there exists a set $C \in \mathcal{C}$ with $F_n(C) = \alpha$ and $\nu(C) = l_0 < l$. This implies $\tilde{F}_n(l_0) = \alpha$ which is a contradiction to the assumption that $\tilde{F}_n$ has a jump at $l$.
(iii) $\Rightarrow$ (i): Suppose $q_n$ has a jump at $\alpha$. Then $\exists \Gamma$ with $F_n(\Gamma) = \alpha$ and $\nu(\Gamma) = q_n(\alpha)$. If $\Gamma \notin \mathcal{MO}_n(\nu(\Gamma))$ then $\exists C \in \mathcal{C}$ with $\nu(C) < q_n(\alpha)$ and $F_n(\alpha) \geq \alpha$. This implies that $q_n$ has no jump at $\alpha$. Contradiction.

**Proof of Lemma 3.2:** We have

$$\begin{aligned} E_n(\lambda) &= \sup_{\{C \in \mathcal{C}\}} \{F_n(C) - \lambda\, \nu(C)\} \\ &= \sup_{l \geq 0} \sup_{\{C \in \mathcal{C}: \nu(C) \leq l\}} \{F_n(C) - \lambda\, \nu(C)\} \\ &= \sup_{l \geq 0} \{\tilde{F}_n(l) - \lambda\, l\}. \end{aligned}$$

The last line is the maximal difference of $\tilde{F}_n$ and a line through the origin with slope $\lambda$. This supremum is attained at a point where $\tilde{F}_n = \tilde{F}_n^*$ and the maximal value itself of course is the intercept of the tangent at this point. If there exist more than one point where this supremum is attained, then they all lie on the same tangent. This argument has been used in Groeneboom [10] (with $F_n$ instead of $\tilde{F}_n$). He used this argument for proving exact $L_1$-rates of convergence for the Grenander density estimator.

**Proof of Lemma 4.1:** Let $f^*_{n,\epsilon,C} = (f^*_n + \epsilon \mathbf{1}_C)/(1 + \epsilon \nu(C))$. It then follows that for the ML-estimator $f^*_n$ one has

$$\frac{d}{d\epsilon} \left\{ \frac{1}{n} \sum_{j=1}^{n} log\, f^*_{n,\epsilon,C}(X_j) \right\} \Bigg|_{\epsilon=0} \leq 0$$

for all $C \in \mathcal{C}$ such that $f^*_{n,\epsilon,C} \in \mathcal{F}_{\mathcal{C}}$. From this, (9) follows by elementary calculations. The fact that (9) holds for all $C \in \mathcal{C}$ if $\mathcal{C}$ is a $\sigma$-lattice follows directly from the fact that in this case $\mathcal{F}_{\mathcal{C}}$ is a cone (see [28]). It can also be seen easily directly be noting that

$$\{x : f(x) + \epsilon\, \mathbf{1}_C(x) > \lambda\} = \{x : f(x) > \lambda\} \cup \{\{x : f(x) > \lambda - \epsilon\} \cap C\}.$$

However, this essentially is the proof of the fact that $\mathcal{F}_{\mathcal{C}}$ is a cone for $\sigma$-lattices $\mathcal{C}$.

**Proof of Lemma 4.2:** Let $f \in \mathcal{F}_{\mathcal{C}}$ be arbitrary. Denote $f_0 = 0, f_j = f(X_j), j = 1, \ldots, n$ and let $\Gamma_j = \{x : f(x) \geq f_j\}$ be the level sets of $f$ at the levels $f_j$. Without loss of generality assume them to be orderd, $f_0 < f_1 \leq \ldots \leq f_n$. Define

$$g(x) = c\, f_j \;\text{ for }\; x \in \Gamma_j \backslash \Gamma_{j+1}, j = 0, \ldots, n$$

where $\Gamma_{n+1} = \emptyset$ and $c > 0$ is a norming constant to make $g$ integrate to 1. Since $g/c \leq f$ we have $c \geq 1$. Moreover, $g \in \mathcal{F}_{\mathcal{C}}$ and $\prod_{j=1}^{n} g(X_j) = c^n \prod_{j=1}^{n} f(X_j) \geq \prod_{j=1}^{n} f(X_j)$.

Now we construct a density with even larger likelihood product and level sets in $\mathcal{L}_n$. Let $\pi_j = \{i : X_i \in \Gamma_j\}, j = 0, \ldots, n$ and define $\tilde{\Gamma}_j = \bigcap\{C \in \mathcal{C} : X^{\pi_j} \subset C\}$. Then, since $X^{\pi_{j+1}} \subset X^{\pi_j}$ we have $\tilde{\Gamma}_{j+1} \subset \tilde{\Gamma}_j$. Let $g_j = g(X_j) = c\, f_j$. Define

$$h(x) = \tilde{c}\, g_j \;\; for\; x \in \tilde{\Gamma}_j \setminus \tilde{\Gamma}_{j+1},$$

where as above $\tilde{c}$ is a norming constant. As above it follows that $h$ has larger likelihood product than $g$ since the norming constant is bigger than 1. By definition $h$ has level sets $\tilde{\Gamma}_j \in \mathcal{L}_n$. The density $h$ is constant at $\tilde{\Gamma}_j \setminus \tilde{\Gamma}_{j+1}$. These sets define a partition of $\tilde{\Gamma}_0$ and it is not difficult to see (see for example [5]) that for a given partition $A_1, \ldots, A_k$ with $\nu(A_j) > 0\ \forall j = 1, \ldots, k$ the histogramm density, which has constant values $F_n(A_j)/\nu(A_j)$ at $A_j$ has the largest likelihood among all densites which are constant at $A_j, j = 1, \ldots, k$. This finishes the proof, since $\nu(\tilde{\Gamma}_j) > 0\ \forall i = 1, \ldots, n$. This follows from the assumption that a MLE exists (cf. Theorem 4.2).

**Proof of Theorem 4.1**: We first prove Theorem 4.1 under the additional assumptions that (for given observations) the MV-sets and modal sets at each level are monotone for inclusion. We refere to this assumption as **(M)**.

Using Lemma 4.1 one easily gets $f_{k_n^*} \geq F_n(C)/\nu(C),\ \forall C \in \mathcal{C}$. Since by assumption a MLE exists, it follows from Theorem 4.2 and Lemma 4.2 that

$$f_{k_n^*} = sup_{C \in \mathcal{C}} F_n(C)/\nu(C) = sup_{L \in \mathcal{L}_n} F_n(L)/\nu(L). \tag{12}$$

Clearly, any set maximizing $F_n(C)/\nu(C)$ over all $C \in \mathcal{C}$ has to be in $\mathcal{MV}_n \cap \mathcal{MO}_n$. Assume for the moment that this maximizing set is unique. Then it follows from Lemma 4.2 that the maximizing set is $\Gamma^*_{n,\mathcal{C}}(f_{k_n^*})$, the level set of the MLE at the maximal level $f_{k_n^*}$ such that

$$\Gamma^*_{n,\mathcal{C}}(f_{k_n^*}) \in \mathcal{MV}_n \cap \mathcal{MO}_n \tag{13}$$

and

$$f_{k_n^*} = \frac{F_n(\Gamma^*_{n,\mathcal{C}}(f_{k_n^*}))}{\nu(\Gamma^*_{n,\mathcal{C}}(f_{k_n^*}))}. \tag{14}$$

Equation (12) says that $f_{k_n^*}$ equals the steepest slope of $\tilde{F}_n^*$ (note that $\tilde{F}_n^*$ starts at $(0,0)$ since a MLE exists) which is the left-hand derivative of $\tilde{F}_n^*$ at $l_{k_n}$. Hence we have

$$f_{k_n^*} = s_{k_n}.$$

In the next step we can restict ourselves to sets $C \in \mathcal{C}$ with $\Gamma_{n,\mathcal{C}}^*(f_{k_n^*}) \subset C$. Then (9) gives

$$\sum_{X_j \in \Gamma_{n,\mathcal{C}}^*(f_{k_n^*})} \frac{1}{f(X_j)} + \sum_{X_j \in C \setminus \Gamma_{n,\mathcal{C}}^*(f_{k_n^*})} \frac{1}{f(X_j)} \leq \nu(C), \tag{15}$$

From (14) we get that the first term on the left hand side of (15) equals $\nu(\Gamma_{n,\mathcal{C}}^*(f_{k_n^*}))$. Hence it follows that $f_{k_n^*-1}$ has to satisfy

$$f_{k_n^*-1} \geq \frac{F_n(C \setminus \Gamma_{n,\mathcal{C}}^*(f_{k_n^*}))}{\nu(C \setminus \Gamma_{n,\mathcal{C}}^*(f_{k_n^*}))} \qquad \forall \quad C \in \mathcal{C} \text{ s.th. } \Gamma_{n,\mathcal{C}}^*(f_{k_n^*}) \subset C. \tag{16}$$

As above it follows that

$$f_{k_n^*-1} = sup\{\frac{F_n(L \setminus \Gamma_{n,\mathcal{C}}^*(f_{k_n^*}))}{\nu(L \setminus \Gamma_{n,\mathcal{C}}^*(f_{k_n^*}))} : L \in \mathcal{L}_n, \Gamma_{n,\mathcal{C}}^*(f_{k_n^*}) \subset L\}. \tag{17}$$

Since the nominators and the denominators in (17) actually are differences of the corresponding measures of the sets $L$ and $\Gamma_{n,\mathcal{C}}^*(f_{k_n^*})$ it follows by using **(M)** that the maximizing set in (17) lie in $\mathcal{MV}_n \cap \mathcal{MO}_n$. If we again assume that it is unique then we have as above

$$\Gamma_{n,\mathcal{C}}^*(f_{k_n^*-1}) \in \mathcal{MV}_n \cap \mathcal{MO}_n$$

and

$$f_{n,k_n^*-1} = s_{k_n-1}.$$

This argument can be repeated and leads to the desired result.

It remains to remove the assumption of uniqueness of the maximizing sets in (12), (17), e.t.c. and to remove **(M)**. First note that it follows from the graphical representation of the silhouette (cf. discussion after Lemma 3.2) together with Lemma 2.1 that there exist empirical $\lambda$-clusters $\Gamma_{n,\mathcal{C}}(\lambda_j)$, $j = 1, \ldots, k_n$, (which by definition all lie in $\mathcal{MV}_n \cap \mathcal{MO}_n$) which are monotone for inclusion. These sets correspond to the vertices of $\tilde{F}_n^*$, this means, that the points $(\nu(\Gamma_{n,\mathcal{C}}(\lambda_j)), F_n(\Gamma_{n,\mathcal{C}}(\lambda_j)))$ are vertices of the graph of $\tilde{F}_n^*$. From this it follows, that the maximal values in (14), (17), e.t.c., i.e. the different values of the MLE, are the slopes of $\tilde{F}_n^*$ even if **(M)** is not assumed to hold. It also follows, that all the maximizing sets correspond to points on $\tilde{F}_n^*$, i.e. for any maximizing set $\Gamma_n$ the point $(\nu(\Gamma_n), F_n(\Gamma_n))$ lies on the graph of $\tilde{F}_n^*$. In other words, the maximizing sets are empirical $\lambda$-clusters. The corresponding value

of $\lambda$ equals the maximal value in (14), (17), e.t.c. We show below, that if $\Gamma_1$ and $\Gamma_2$ are two empirical $\lambda$-clusters to the same value of $\lambda$, then also the union $\Gamma_1 \cup \Gamma_2$ is a empirical $\lambda$-cluster to this value of $\lambda$. From this we can remove the assumption of uniqueness as follows: Suppose the maximizing set in (14) is not unique, and we did not choose the largest maximizing set, i.e. the union of all maximizing sets. Then the next iteration step leads to the same maximal value, i.e. we stay at the same level of the MLE, and the (present) level set of the MLE only becomes larger, until we finally reached the largest level set. Hence, the uniqueness assumption is not necessary.

It remains to show, that if $\Gamma_1$ and $\Gamma_2$ are two emprical $\lambda_0$-clusters, then $\Gamma_1 \cup \Gamma_2$ also is a $\lambda_0$-cluster. We have

$$H_{n,\lambda_0}(\Gamma_1 \cup \Gamma_2) = H_{n,\lambda_0}(\Gamma_1) + H_{n,\lambda_0}(\Gamma_2 \setminus \Gamma_1)$$

and

$$H_{n,\lambda_0}(\Gamma_1 \cap \Gamma_2) = H_{n,\lambda_0}(\Gamma_2) - H_{n,\lambda_0}(\Gamma_2 \setminus \Gamma_1).$$

From the first equalility it follows that $H_{n,\lambda_0}(\Gamma_2 \setminus \Gamma_1) \leq 0$, since by definition $\Gamma_1$ maximizes $H_{n,\lambda_0}$ over all sets in $\mathcal{C}$. Analogously, the second equality gives $H_{n,\lambda_0}(\Gamma_2 \setminus \Gamma_1) \geq 0$ and hence it equals zero. The first equation now gives the assertion.

**Proof of Theorem 4.2:** The equivalence of $(ii)$ and $(iii)$ is obvious. $(i) \Rightarrow (ii)$ follows from Lemma 4.2. $(ii) \Rightarrow (i)$ also follows from Lemma 4.2: $(ii)$ says that all the MV-sets at levels $\alpha > 0$ have positive $\nu$-measure. Hence $(i)$ follows from Lemma 4.2, since all sets in $\mathcal{L}_n$, defined in Lemma 4.2, have bigger $\nu$-measure than the MV-set at level $1/n$. $(v) \Leftrightarrow (iii)$ follows from the fact that $\int f_{n,\mathcal{C}}(x)\, dx = \tilde{F}_n(1) - \tilde{F}_n(0)$ (cf. Polonik [26]).

**Proof of Theorem 5.1:** The proof is very similar to the proof of Theorem 3.4 of Polonik [26]. We only give the main steps and indicate the argument for removing the log-term.

First note that $|| f_{n,\mathcal{C}} - f ||_1 \leq 2 \int \nu(\Gamma_{n,\mathcal{C}}(\lambda) \setminus \Gamma(\lambda))\, d\lambda + O(n^{-1})$. This follows from Fubini's theorem, Jensen's inequality and the fact that the silhouette integrates to $1 + O(n^{-1})$. (Equality holds, if the sets $\Gamma_{n,\mathcal{C}}(\lambda)$ are level sets of $f_{n,\mathcal{C}}$ as for $\sigma$-lattices $\mathcal{C}$ (see Section 2).) We use estimates for $\nu(\Gamma_{n,\mathcal{C}}(\lambda) \setminus \Gamma(\lambda))$ in order to obtain rates of convergence of the silhouette in $L_1(\nu)$. The truncation argument for unbounded densities given below essentially uses the fact that $F(\Gamma_{n,\mathcal{C}}(\lambda) \setminus \Gamma(\lambda)) \leq \lambda\, \nu(\Gamma_{n,\mathcal{C}}(\lambda) \setminus \Gamma(\lambda))$. Since this in general does not hold for the symmetric difference $\Gamma_{n,\mathcal{C}}(\lambda) \Delta \Gamma(\lambda)$ we don't use this symmetric difference here, as is done in Polonik [26]. There on uses similar inequalities as below, but with $\Gamma_{n,\mathcal{C}}(\lambda) \Delta \Gamma(\lambda)$ instead of $\Gamma_{n,\mathcal{C}}(\lambda) \setminus \Gamma(\lambda)$, which hold without the additional assumption that $\mathcal{C}$ is closed under intersection. Therefore the proof given here also shows, how to remove the additional log-term of Polonik [26] for the case $\gamma_2 = 1$ in (b) by analog argument.

Now, there is the following basic inequality. For any $\alpha > 0$ we have with $\Psi_\alpha(\lambda) = \nu\{x : \lambda - \alpha < f(x) < \lambda\}$ that

$$\nu(\Gamma_{n,\mathcal{C}}(\lambda) \setminus \Gamma(\lambda)) \leq \Psi_\alpha(\lambda) + \frac{1}{\alpha}( (F_n - F)(\Gamma_{n,\mathcal{C}}(\lambda) \setminus \Gamma(\lambda)) ). \tag{18}$$

The proof of (18) is as follows. First note that $H_{n,\lambda}(\Gamma_{n,\mathcal{C}}(\lambda) \setminus C) \geq 0$ for all $C \in \mathcal{C}$, provided $\mathcal{C}$ is closed under intersections. This follows from the last equation given in proof of Theorem 4.1 above, by replacing $\Gamma_2$ by $C$. Hence, we have with $D_n(\lambda) = \Gamma_{n,\mathcal{C}}(\lambda) \setminus \Gamma(\lambda)$ that $0 \leq H_{n,\lambda}(D_n(\lambda)) = H_\lambda(D_n(\lambda)) + (F_n - F)(D_n(\lambda))$, and it follows

$$0 \leq -H_\lambda(D_n(\lambda)) \leq (F_n - F)(D_n(\lambda)). \tag{19}$$

Moreover, we have

$$-H_\lambda(D_n(\lambda)) = \int_{D_n} (\lambda - f(x)\, d\nu(x) \geq \alpha\, \nu(D_n(\lambda) \cap \{x : f(x) \leq \lambda - \alpha\}). \tag{20}$$

By writing $\nu(D_n(\lambda)) = \nu(D_n(\lambda) \cap \{\lambda - \alpha < f < \lambda\}) + \nu(D_n \cap \{f \leq \lambda - \alpha\})$ inequality (18) follows from (19) and (20).

Note that (18) does not help us for "small" $\lambda$ if $\nu(\mathcal{X}) = \infty$. Therefore we use that for any $0 \leq \alpha_n \leq M_n$ we have

$$||f_{n,\mathcal{C}} - f||_1 \leq 2 \int_{\alpha_n}^{M_n} \nu(\Gamma_{n,\mathcal{C}}(\lambda) \setminus \Gamma(\lambda)) d\lambda + (\bar{E}(\alpha_n) + E(M_n)) + O_{P^*}(||F_n - F||_{\mathcal{C}}) \tag{21}$$

where $||\cdot||_{\mathcal{C}}$ denotes sup-norm over $\mathcal{C}$. Here and below $P^*$ denotes outer measure. The proof of (21) is given below. Inequality (18) can be exploited to get that under the assumptions on $\alpha_n$ and $M_n$ formulated in the theorem

$$\sup_{M_n > \lambda > \alpha_n} \frac{\nu(\Gamma_{n,\mathcal{C}}(\lambda) \setminus \Gamma(\lambda))}{max(\Psi_{\alpha_n}(\lambda),\ \alpha_n^K)} = O_{P^*}(1), \tag{22}$$

for some large $K > 0$ determined later. We use this power of $\alpha_n$ in order to avoid considering levels $\lambda$ with $\Psi_{\alpha_n}(\lambda) = 0$ separately. The proof of (22) is also given below. Plugging in (22) into (21) gives the assertion:

$$\begin{aligned} || f_{n,\mathcal{C}} - f ||_1 \ &\leq\ O_{P^*}(1) \textstyle\int_{\alpha_n}^{M_n} max(\Psi_{\alpha_n}(\lambda),\ \alpha_n^K)\, d\lambda + \\ &\qquad \bar{E}(\alpha_n) + E(M_n) \ +\ O_{P^*}(||F_n - F||_{\mathcal{C}}) \\ &\leq\ O_{P^*}(1)[\ \textstyle\int_{\alpha_n}^{M_n} \Psi_{\alpha_n}(\lambda)\, d\lambda \ +\ \alpha_n^K(M - \alpha_n)\ ] + \\ &\qquad 2\bar{E}(\alpha_n) \ +\ E(M_n) \ +\ O_{P^*}(|F_n - F||_{\mathcal{C}}) \\ &\leq\ O_{P^*}(1)[\ \bar{E}(\alpha_n) \ +\ E(M_n) \ +\ \alpha_n^K(M - \alpha_n)\ ] + \end{aligned}$$

$$\bar{E}(\alpha_n) \; + \; E(M_n) \; + \; O_{P^*}(||F_n - F||_{\mathcal{C}})$$

The last inequality follows easily by using the fact that up to at most countably many levels $\lambda$ we have $\Psi_\eta(\lambda) = \nu(\Gamma(\lambda - \eta)) - \nu(\Gamma(\lambda + \eta))$. Now choose $K$ large enough such that $\alpha_n^K(M_n - \alpha_n) = O(n^{-1})$. As for the term $O_{P^*}(||F_n - F||_{\mathcal{C}}$ it follows from Corollary 2.4 of Alexander [1], Correction, that this term is negligible here. The rate of this term is $O_{P^*}(n^{-1/2})$ for classes of sets as in part (a) and for classes of sets with $\gamma_2 < 1$ as in (b). For classes with $\gamma_2 = 1$ the rate is $O_{P^*}(n^{-1/2} \log n)$ and $O_{P^*}(n^{-1/(\gamma_2+1)})$ if $\gamma_2 > 1$.

It remains to proof (21) and (22). As for the latter we get from (18) that for any $c > 0$:

$$\left\{ \sup_{M_n > \lambda > \alpha_n} \frac{\nu(\Gamma_{n,\mathcal{C}}(\lambda) \setminus \Gamma(\lambda))}{max(\Psi_{\alpha_n}(\lambda),\, \alpha_n)} \geq c \right\}$$

$$\subset \left\{ \frac{(F_n - F)(\Gamma_{n,\mathcal{C}}(\lambda) \setminus \Gamma(\lambda))}{\nu(\Gamma_{n,\mathcal{C}}(\lambda) \setminus \Gamma(\lambda))} \geq \alpha_n(1 - \frac{1}{c}) \right.$$

$$\left. \textit{for some } M_n > \lambda > \alpha_n \;\; \textit{s.th.} \;\; \nu(\Gamma_{n,\mathcal{C}}(\lambda) \setminus \Gamma(\lambda)) > \alpha_n \right\},$$

Now we use empirical process theory to find the "smallest" $\alpha_n$ such that the probability of the last event tends to zero. For $m \in \mathbf{R}$ let $\mathcal{G}_{n,m} = \{g = r(C \setminus D) : r \leq m,\; C, D \in \mathcal{C}, C \setminus D \subset \Gamma(M_n)\}$. Then it is enough to find $\alpha_n$, such that for any $\eta > 0$ there exists a $c > 0$ such that for all $n$ large enough

$$P( \sup_{g \in \mathcal{G}_{n,1} : ||g||_1 > \alpha_n} n^{1/2}(F_n - F)(g)/||g||_1 > n^{1/2}\alpha_n(1 - 1/c) < \eta \qquad (23)$$

This follows from

$$\sum_{j=1}^{\infty} P( \sup_{\{g \in \mathcal{G}_{n,1} :\, ||g||_1 \leq 2^j \alpha_n\}} n^{1/2}(F_n - F)(g) \geq 2^{j-1}\, n^{1/2}\alpha_n^2\} < \eta. \qquad (24)$$

Now we are ready to use Theorem 2.3 of Alexander [1]. This theorem gives conditions under which for classes of sets $\mathcal{C}$ with $\infty > M_n \geq \sup_{f \in \mathcal{F}} |||f||_\infty$ and $\sigma_{max}^2 \geq \sup_{C \in \mathcal{C}} F(C)(1 - F(C))$ we have

$$P(\sup_{C \in \mathcal{C}} n^{1/2}(F_n - F)(C) \geq L) \leq 3exp\{-\frac{1 - \delta}{2}\left(\frac{L}{\sigma_{max}}\right)^2\}$$

for some $0 < \delta < 1$. The result of Alexander can easily be generalized to classes of functions like $\mathcal{G}_{n,m}$, by replacing the $L_1$-bracketing entropy of the class of sets $\mathcal{C}$ by the $L_1$-bracketing entropy of $\mathcal{G}_{n,m}$. Note that $\sup_{\{g \in \mathcal{G}_{n,m} : ||g||_1 \leq 2^j \alpha_n\}} var(g)$ $\leq M_n\; m\; 2^j\alpha_n$. If we apply this result (for each $j$) with $L = 2^{j-1}n^{1/2}\alpha_n^2$ and $m = 1$ such that $\sigma_{max}^2 = M_n\; 2^j\alpha_n$ then we directly obtain the assertion of the theorem with the exception of the case $\gamma_2 = 1$ in part (b). In this case a direct

application gives an additional log-term (see below) as in Polonik [26]. This can be avoided by the following trick. Instead of applying Alexander's theorem to the probabilities in (24) (for each $j$), we instead apply it to

$$P(\sup_{\{g \in \mathcal{G}_{n,\alpha_n^\beta}:\, ||g||_1 \leq 2^j \alpha_n^{1+\beta}\}} n^{1/2}(F_n - F)(g) \geq 2^{j-1}\, n^{1/2} \alpha_n^{2+\beta}\}.$$

for some $\beta \in \mathbf{R}$. This means, we multiply "everything" inside the probability by $\alpha_n^\beta$. Of course, this does not change the actual probability. However, it changes the crucial condition (2.8) from Alexander's theorem (and the appropriate generalization) in the case $\gamma_2 = 1$. This crucial condition rewritten for the case of an underlying function class $\mathcal{G}_{n,m}$ is as follows. Let $\tilde{H}(\epsilon) = A(\epsilon/m^{1/2})^{-2\,r}$ and define $t = \tilde{H}^{-1}(\, c(1-c)\, L^2/8\, \sigma_{max}^2\, )$ and $s = (c\, L/16\, n^{1/2})^{1/2}$. Then $\log N_B(\epsilon, \mathcal{G}_{n,m}, F) \leq \tilde{H}(\epsilon)$ and $L$ needs to satisfy

$$L > 2^9 c^{-3/2}\, m^{1/2} \int_{s/4}^{t} \tilde{H}(\epsilon)^{1/2}\, d\epsilon.$$

With $L$ as above and since $m = \alpha_n^\beta$ we get for $\gamma_2 = 1$ (by collecting constants): $n^{-1/2}\alpha_n^{2+\beta} > const.\ \alpha_n^{\beta/2}\alpha_n^{\beta r/4} \log s$ and hence

$$n^{-1/2}\alpha_n^2 > const.\ \log \alpha_n^{2+\beta}.$$

Therefore, if we choose $\beta = -2$ we obtain $\alpha_n > const.\ n^{-1/4}$ without an additional log-term.

Now we proof (21). As mentioned above $||f_{n,\mathcal{C}} - f||_1 \leq 2 \int \nu(\Gamma_{n,\mathcal{C}}(\lambda) \backslash \Gamma(\lambda))\, d\lambda + O(n^{-1})$. Now split the integral into a sum of three integrals:

$$\int \nu(\Gamma_{n,\mathcal{C}}(\lambda) \setminus \Gamma(\lambda))d\lambda = \int_\alpha^M + \int_0^\alpha + \int_M^\infty \nu(\Gamma_{n,\mathcal{C}}(\lambda) \setminus \Gamma(\lambda))d\lambda.$$

Note that $\nu(\Gamma_{n,\mathcal{C}}(\lambda) \setminus \Gamma(\lambda)) \leq \nu(\Gamma_{n,\mathcal{C}}(\lambda))$ and that $\nu(\Gamma_{n,\mathcal{C}}(\lambda))$ is the derivative of $E_{n,\mathcal{C}}$ almost everywhere. This follows directly from the definition of the empirical excess mass (cf. Polonik [25]). Hence, the last two integrals are smaller than or equal to $1 - E_{n,\mathcal{C}}(\alpha)$ and $E_{n,\mathcal{C}}(M)$, respectively. Since $sup_{\lambda \geq 0}|E_{n,\mathcal{C}}(\lambda) - E(\lambda)| \leq ||F_n - F||_{\mathcal{C}}$ (Polonik [25], Lemma 2.2) the assertion follows.

A small step remains open. In order to formulate the rates as in the theorem we need that for each $K > 0$ there exists a $K^* > 0$ such that $\bar{E}(K\lambda) \leq K^* \bar{E}(\lambda)$ for all $\lambda$ small enough. However, this follows easily from the fact that $E$ is continuous. (Note that trivially $E(K\, M) \leq E(M)$ for $M$ large enough, since $E$ is decreasing.)

**Acknowledgement:** I am greatful to Lutz Dümbgen for the hint to look at derivatives in directions of indicator functions which is used in the proof of

Lemma 4.1. This lemma eventually turned out to be one of the key results in the proof of the graphical representation of the MLE. I also want to thank Jianhua Huang for careful reading of the manuscript and for suggestions that lead to an improvement in the presentation of the proof of Theorem 5.1 and to the discovery of a serious error in an earlier version of the manuscript.

# References


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

[2] Barbe, P. and Wei, S. (1994). A test for multimodality. Unpublished maunscript

[3] Barlow, R.E., Bartholomew, D.J., Bremner, J.M., and Brunk, H.D. (1972). *Statistical inference under order restrictions.* Wiley, London

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

[5] Devroye, L. (198?). A course in density estimation. Birkhaeuser

[6] Dharmadhikari, S. and Joag-Dev, K. (1988). *Unimodality, convexity, and applications.* Academic Press, San Diego

[7] Dudley, R.M. (1984). A course in empirical processes. *Ecole d'Ete de Probabilites de Saint Flour XII-1982, Lecture Notes in Math.* **1097** 1-142, Springer, New York

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

[9] Grenander, U. (1956). On the theory of mortality measuremant, Part II. *Skand. Akt.* **39** 125-153

[10] Groeneboom, P (1985). Estimating a monotone density. In *Proceedings of the Berkeley Conference in honor of Jerzy Neymann and Jack Kiefer*, Vol. II, Le Cam, L. and Olshen, R. (eds.), Monterey: Wadsworth

[11] Grübel, R. (1988). The length of the shorth. *Ann. Statist.* **16** 619-628

[12] Hartigan, J.A. (1975). *Clustering algorithms.* Wiley, New York

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

[14] Hengartner, W. and Theodorescu, R. (1973). *Concentration functions.* Academic Press, New York

[15] Hickey, R.J. (1984). Continuous majorisation and randomness. *J. Appl. Prob.* **21** 924-929

[16] Joe, H. (1993). Generalized majorization orderings, in *Stochastic Inequalities*, IMS Lecture Notes - Monograph series, Vol. 22, Shaked, M. and Tong, Y.L., eds.

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

[18] Marshall, A.W. and Olkin, I. (1979). *Inequalities: Theory of majorization and its applications.* Academic Press, New York

[19] Müller, D.W. and Sawitzki, G. (1987). Using excess mass estimates to investigate the modality of a distribution. Preprint No.398, SFB 123, Univ. Heidelberg.

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

[21] Nolan, D. (1991). The excess mass ellipsoid. *J. Multivarite Anal.* **39** 348-371

[22] Pollard, D. (1984) *Convergence of stochastic processes.* Springer, New York

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

[24] Polonik, W. (1994). Minimum volume sets and generalized quantile processes. Beitraege zur Statistik No. 20, Universität Heidelberg

[25] Polonik,W. (1995). Measuring mass concentrations and estimating density contour clusters - an excess mass approach. *Annals of Statistics* **23** 855-881

[26] Polonik, W. (1995). Density estimation under qualitative assumptions in higher dimensions. *J. Multivariate Anal.* **55** 61-81

[27] Robertson, T. (1967). On estimating a density measurable with respect to a $\sigma$-lattice. *Ann. Math. Statist.* **38** 482-493.

[28] Robertson, T., Wright, F.T., and Dykstra, R.L. (1988). *Order restricted statistical inference.* Wiley, New York

[29] Sager, T.W. (1982). Nonparametric maximum likelihood estimation of spatial patterns. *Ann. Statist.* **10** 1125-1136.

[30] Wegman, E. (1969). A note on estimating a unimodal density. *Ann. Math. Statist.* **40** 1661-1667.

[31] Wegman, E. (1970). Maximum likelihood estimation of a unimodal density function. *Ann. Math. Statist.* **41** 457-471

[32] Wong, W. H. and Severini, T. A. (1991). On maximum likelihood estimation in infinite dimensional parameter spaces. *Ann. Statist.* **19** 603-632