The Annals of Statistics
1998, Vol. 26, No. 5, 1857–1877

THE SILHOUETTE, CONCENTRATION FUNCTIONS AND ML-DENSITY ESTIMATION UNDER ORDER RESTRICTIONS[1]

BY WOLFGANG POLONIK

Universität Heidelberg

Based on empirical Lévy-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 to higher dimensions and arbitrary order restrictions. 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 provides a new point of view to ML-density estimation from which one gains additional insight to this problem, as demonstrated in the present paper.

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 Lévy-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 behavior of the MLE which are based on empirical process theory (see Section 7).

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

Received March 1996; revised April 1998.

[1]Supported by the Deutsche Forschungsgemeinschaft.

AMS 1991 *subject classifications*. Primary 62G07; secondary 62A10, 62G30, 62G20.

Key words and phrases. Empirical processes, excess mass, Grenander density estimator, level set estimation, least concave majorant, minimum volume sets, nonparametric maximum likelihood estimation.

measure space $(\mathscr{X}, \mathscr{A})$. Such order restrictions can be expressed via a class $\mathscr{C}$ of measurable sets [cf. Barlow, Bartholomew, Bremner and Brunk (1972) or Robertson, Wright and Dykstra (1988)]: given a (quasi-) order $\preceq$ (reflexive and transitive), there exists a class of sets $\mathscr{C} = \mathscr{C}_{\preceq}$ (see also Section 3.1) such that f is monotone with respect to $\preceq$ iff f is measurable with respect to $\mathscr{C}_{\preceq}$; this means, iff all *level sets* $\Gamma_f(\lambda) = \Gamma(\lambda) = \{f \geq \lambda\}$, $\lambda \geq 0$, are elements of $\mathscr{C}_{\preceq}$. Hence, order restrictions on f can be reformulated as "$f \in \mathscr{F}_{\mathscr{C}}$" for appropriate classes $\mathscr{C}$, where

$$\mathscr{F}_{\mathscr{C}} = \left\{ f\colon \int f(x)\, d\nu(x) = 1, \Gamma(\lambda) \in \mathscr{C} \text{ for all } \lambda \geq 0 \right\}$$

and where ν is some dominating measure on $(\mathscr{X}, \mathscr{A})$. MLEs under order restrictions based on n i.i.d. observations have been derived and studied, among others, by Grenander (1956), Robertson (1967), Wegmann (1969, 1970) and Sager (1982). It is well known [cf. Robertson (1967), Sager (1982)], that the structure on $\mathscr{X}$ given through $\preceq$ induces a structure on the corresponding class $\mathscr{C}_{\preceq}$: it has to be a σ-lattice. $\mathscr{C}$ is called a σ-lattice if it contains $\mathscr{X}$ and $\varnothing$ and is closed under countable unions and intersections. A simple example is given by $\mathscr{C} = \mathscr{I}_0 = \{[0, x], x \geq 0\}$ 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 σ-lattice which is not a σ-algebra is the class of intervals containing a given point, x_0, say. The corresponding class $\mathscr{F}_{\mathscr{C}}$ is the class of unimodal densities with mode x_0. Discrete analogs are given by the classes $\{\{1, 2, \dots, k\}, k \geq 1\}$ and $\{\{-k, \dots, -1, 0, 1, \dots, k\}, k \geq 0\}$, respectively.

The model $f \in \mathscr{F}_{\mathscr{C}}$ for some class of measurable subsets $\mathscr{C}$ also underlies the construction of the silhouette. However, there the class $\mathscr{C}$ need not correspond to any order; $\mathscr{C}$ can in principle be completely arbitrary. We call a model assumption of the form $f \in \mathscr{F}_{\mathscr{C}}$ *shape restriction* given by $\mathscr{C}$. 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 (see Theorem 2.3), is the MLE in $\mathscr{F}_{\mathscr{C}}$ if the shape restriction actually is an order restriction.

Let us briefly point out the basic distinction between the construction of the silhouette and the classical construction of the MLE. First, note that a MLE $\hat{f}_n$ in $\mathscr{F}_{\mathscr{C}}$ based on an i.i.d. sample of size n has to be of histogram type (see Lemma 5.2); that is, there exists a partition $\{A_1, \dots, A_k\}$ of $\mathbf{R}^d$ such that $\hat{f}_n(x) = \#\{\text{observations} \in A_i\}/n\nu(A_i)$, 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 $\mathscr{C}$. In contrast to that, the silhouette is constructed by putting *estimated level sets* one on 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. Moreover, the construction of the silhouette is of a "global" nature, whereas in contrast, the classical approach can be considered to be of a "local" nature.

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 majorants of certain Lévy-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 (1956)] to higher dimensions and arbitrary order restrictions. The concentration functions under consideration are defined through constrained maximization of certain functionals defined on $\mathscr{C}$. 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 5). Moreover, the concentration functions used for the graphical representation of the MLE also can be used for characterizing the existence of the MLE (see Theorem 5.5).

The present paper is organized as follows. In Section 2 we introduce the silhouette, give some of its properties, and formalize the above-mentioned fact that the silhouette equals the MLE. In Section 3 a list of examples is presented for illustration. In particular, the connection to the Grenander estimator is spelled out. Section 4 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 in Section 5 where also the graphical representation of the MLE is presented. Section 6 discusses algorithmic aspects of this graphical representation. A summary and some final remarks can be found in Section 7. All the proofs are given in Section 8.

2. The silhouette.

If restricted to the continuous case, that is, the dominating measure is Lebesgue measure on $\mathbf{R}^d$, the first part of this section more or less is a short version of Section 2 of Polonik (1995b). 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 $f\colon \mathscr{X} \to \mathbf{R}$, the following key representation holds:

$$f(x) = \int_0^\infty \mathbf{1}_{\Gamma(\lambda)}(x)\, d\lambda \qquad \forall\, x \in \mathscr{X}, \tag{2.1}$$

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 (2.1). As estimators we use so-called empirical generalized λ-clusters. They are defined as follows: let $X_1, X_2, \ldots$ denote i.i.d. observations drawn from a distribution F which has a density f with respect to a measure ν. Let F_n denote the empirical measure based on the first n observations, that is, $nF_n(C) = \#\{\{X_i, \ldots, X_u\} \cap C\}$ and for $\lambda \geq 0$ define the signed measure

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

DEFINITION 2.1. Any set $\Gamma_{n,\mathscr{C}}(\lambda) \in \mathscr{C}$ such that

$$H_{n,\lambda}(\Gamma_{n,\mathscr{C}}(\lambda)) = \sup_{C\in\mathscr{C}} H_{n,\lambda}(C) \tag{2.2}$$

is called an empirical generalized λ-cluster in $\mathscr{C}$.

The sets $\Gamma_{n,\mathscr{C}}(\lambda)$ are called *generalized* since they need not be connected, as one would expect for clusters. Nevertheless, for brevity, we omit the phrase *generalized* and call the sets $\Gamma_{n,\mathscr{C}}(\lambda)$ empirical λ-clusters or sometimes just λ-clusters. Hartigan (1975) used the notion λ-cluster for *connected* components of level sets. Note that the notion "λ-cluster" is in general used for the collection of all λ-clusters, that is, for the collection of λ-clusters at all levels $\lambda \geq 0$. Sometimes, however, we consider a single level λ. We hope this becomes sufficiently clear from the context.

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

$$H_\lambda(\Gamma(\lambda)) = \sup\{H_\lambda(C), C \in \mathscr{A}\}. \tag{2.3}$$

This equation motivates the use of $\Gamma_{n,\mathscr{C}}(\lambda)$ as estimator for the level set $\Gamma(\lambda)$. Note that if ν is a continuous measure then the supremum of $H_{n,\lambda}$ over *all* measureable sets equals 1. Hence, besides the fact that the class $\mathscr{C}$ is used to introduce shape restrictions, it makes sense in general to restrict the supremum to certain subclasses $\mathscr{C}$.

As a function of λ, the maximal value in (2.3), that is, $E(\lambda) = E_F(\lambda) = H_\lambda(\Gamma(\lambda))$, is called *excess mass function*. Note that $E(\lambda)$ is used in majorization ordering. 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 [Marshall and Olkin (1979); see also Hickey (1984), Joe (1993)]. Actually, all of this is formulated in terms of densities. The representation (2.3), however, gives a way to express this in terms of distributions, without using densities explicitly.

The maximal value in (2.2), that is,

$$E_n(\lambda) = H_{n,\lambda}\big(\Gamma_{n,\mathscr{C}}(\lambda)\big)$$

is called *empirical excess mass* at level λ. Hartigan (1987) and Müller and Sawitzki (1987) independently introduced the excess mass approach, which is based on the idea [motivated by (2.3)] that maximizing the signed measure $H_{n,\lambda}$ gives information about the mass concentration of the underlying distribution. The notion *excess mass* was first used by Müller and Sawitzki. For further work on excess mass and on empirical λ-clusters, see Nolan (1991), Müller and Sawitzki (1991) and Polonik (1992, 1995a).

In all of what follows, it is assumed that $\mathscr{C}$ is such that:

(A1) $\varnothing \in \mathscr{C}$.
(A2) *For any $\lambda \geq 0$ there exists an empirical λ-cluster*.
(A3) *Almost surely there exists a set $S \in \mathscr{C}$ with $\nu(S) < \infty$ and $F_n(S) = 1$.*

Let us briefly discuss these assumptions. (A1) assures that the empirical excess mass is nonnegative (as it should be). (A2) for instance holds for finite $\mathscr{X}$ with $\mathscr{A} = \mathscr{P}^{\mathscr{X}}$ and ν the counting measure. In the continuous case, that is, if ν is Lebesgue measure and $\mathscr{A}$ the Borel σ-algebra, empirical λ-clusters exist for standard classes $\mathscr{C}$ such as $\mathscr{I}^d$, $\mathscr{B}^d$, $\mathscr{E}^d$ and $\mathfrak{C}^d$, 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 λ-clusters is that $\mathscr{C}$ is closed under intersections. Of course, this condition is not necessary. (A3) means that a.s. there exist empirical λ-clusters for $\lambda = 0$ with nondegenerate ν measure. Hence, since $\lambda \to \nu(\Gamma_{n,\mathscr{C}}(\lambda))$ is a decreasing function, it follows that all empirical λ-clusters have finite ν-measure.

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

(P) There exist levels $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_{k_n}$, $0 \leq k_n \leq n$, such that $\nu(\Gamma_{n,\mathscr{C}}(\lambda_{k_n})) = 0$ and that the function $\lambda \to \Gamma_{n,\mathscr{C}}(\lambda)$, $\lambda \geq 0$, is constant over the intervals $(\lambda_{j-1}, \lambda_j]$, $j = 1, \dots, k_n$ and obtains different values on different such intervals.

In fact, property (P) is not a necessary assumption for the theory. However, without (P) the silhouette may look quite erratic (this does not happen for σ-lattices $\mathscr{C}$; cf. Lemma 2.2 below). Any choice of empirical λ-clusters satisfying (P) automatically has the property that for any fixed $\mu > 0$, the ν-measure of $\Gamma_{n,\mathscr{C}}(\mu)$ is maximal among all empirical μ-clusters. A way to find the values λ_i of (P) is given by means of the graphical representation of the silhouette (cf. Section 5).

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

$$f_{n,\mathscr{C}}(x) = \int_0^\infty \mathbf{1}_{\Gamma_{n,\mathscr{C}}(\lambda)}(x)\, d\lambda \qquad \forall\, x \in \mathscr{X}. \tag{2.4}$$

The definition of the silhouette depends on the special choice of sets $\Gamma_{n,\mathscr{C}}(\lambda)$. This gives different *versions* of the silhouette. These versions might differ even on sets with positive ν-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) (with $k_n > 0$) the silhouette can be written as

$$f_{n,\mathscr{C}}(x) = \sum_{j=0}^{k_n-1} (\lambda_{j+1} - \lambda_j)\mathbf{1}_{\Gamma_{n,\mathscr{C}}(\lambda_j)}(x). \tag{2.5}$$

Hence, if in addition the sets $\Gamma_{n,\mathscr{C}}(\lambda_j)$, $j = 1, \dots, k_n$, are monotonically decreasing for inclusion, that is, $\Gamma_{n,\mathscr{C}}(\lambda_{j+1}) \subset \Gamma_{n,\mathscr{C}}(\lambda_j)$, then $f_{n,\mathscr{C}}$ can be visualized as putting the slices $\Gamma_{n,\mathscr{C}}(\lambda_j) \times (\lambda_j, \lambda_{j+1}]$ one on top of the other.

The empirical λ-clusters can be chosen to be monotone if $\mathscr{C}$ is a σ-lattice (see Lemma 2.2). Unfortunately, however, monotonicity of empirical λ-clusters does not necessarily hold for non-σ-lattices $\mathscr{C}$ such as $\mathscr{I}^1$ or $\mathfrak{C}^d$. This means that for non-σ-lattices $\mathscr{C}$ the silhouette does not necessarily lie in the model class $\mathscr{F}_{\mathscr{C}}$.

LEMMA 2.2. *If $\mathscr{C}$ is a σ-lattice then*

$$\Gamma_{f_{n,\mathscr{C}}}(\lambda) \in \mathscr{C} \qquad \forall\, \lambda > 0.$$

Moreover, if $k_n \geq 1$,

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

and the empirical λ-clusters can be chosen such that

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

Note that the first assertion of Lemma 2.2 does not say that $f_{n,\mathscr{C}} \in \mathscr{F}_{\mathscr{C}}$, which requires in addition that $\int f_{n,\mathscr{C}} = 1$. In fact, it might happen that $\int f_{n,\mathscr{C}} < 1$ and even $\int f_{n,\mathscr{C}} = 0$. In case of a σ-lattice $\mathscr{C}$ this is connected to the existence of an MLE in $\mathscr{F}_{\mathscr{C}}$ (see Theorem 5.5). Of course $f_{n,\mathscr{C}} \in \mathscr{F}_{\mathscr{C}}$ is a necessary condition for $f_{n,\mathscr{C}}$ to be a maximum likelihood estimator in $\mathscr{F}_{\mathscr{C}}$.

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

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

Now we state one of the main theorems.

THEOREM 2.3. *Let $\mathscr{C}$ be a σ-lattice. If a MLE in $\mathscr{F}_{\mathscr{C}}$ exists, then*

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

3. Examples. To illustrate the implications of the above we now briefly discuss several interesting special situations. In particular, the well-known case of the Grenander estimator is covered. In each case we explicitly mention the corresponding classes $\mathscr{C}$ underlying the construction of the silhouette. For computational aspects we refer to Section 6.

3.1. *Order restrictions.* Here we give some examples of order restrictions and the corresponding classes $\mathscr{C}_{\preceq}$ which lead to known MLEs. It follows from Theorem 2.3 that in all these cases the silhouette corresponding to the class $\mathscr{C} = \mathscr{C}_{\preceq}$ equals the MLE. Recall that for a given (quasi-) order $\preceq$ the class $\mathscr{C}_{\preceq}$ (cf. the Introduction) is the class of so-called *upper sets* for $\preceq$. U is an upper set if and only if $x \in U$, $x \preceq y \Rightarrow y \in U$.

The histogram. Given a partition $\mathscr{P} = \{P_1, \ldots, P_N\}$ of the sample space ($0 < \nu(P_i) < \infty$, $i = 1, \ldots, N$), define $x \preceq y \Leftrightarrow$ "x and y lie in the same set P_i." Hence, one identifies all observations in the individual sets P_i without further restrictions. Here $\mathscr{C}_{\preceq} = \sigma(\mathscr{P})$, the σ-algebra generated by $\mathscr{P}$. The corresponding silhouette equals the histogram based on the partition $\mathscr{P}$. This follows from the well-known fact that the histogram is the MLE in the class of all functions which are constant on each P_i, $i = 1, \ldots, N$.

The Grenander estimator. For the usual order "$\geq$" on the positive real line we have $\mathscr{C}_{\geq} = \mathscr{I}_0 =$ intervals starting at zero. Hence, $\mathscr{F}_{\mathscr{I}_0}$ consists of all probability densities on the positive real line which are nonincreasing. It is well known that the MLE in $\mathscr{F}_{\mathscr{I}_0}$ exists a.s.; it is the Grenander estimator of a monotone density [Grenander (1956)].

Other univariate MLEs. A unimodal MLE on the real line with given mode x_0, say, can (roughly speaking) be represented [see Robertson, Wright and Dykstra (1988)] as the slope of a function $\hat{F}_n$, which is the greatest convex minorant of F_n, on $(-\infty, x_0]$ and the least concave majorant of F_n on $[x_0, \infty)$. (But see Theorem 5.3.) Here the corresponding class of upper sets consists of intervals with midpoint x_0. A bimodal MLE with given or estimated modes $x_1 < x_2$ exists, provided both x_1 and x_2 do not coincide with one of the observations. The class $\mathscr{C}_{\preceq}$ is one of the classes $\mathscr{I}_{x_1, x_2}(a) = \{I_1 \cup I_2; I_1, I_2 \text{ intervals}, x_i \in I_i \text{ and } I_1 \leq a < I_2\}$. For each a, the class $\mathscr{I}_{x_1, x_2}(a)$ is a σ-lattice (cf. Sections 3.2 and 6).

Doubly monotone MLE. A generalization of the Grenander estimator to higher dimensions is the following. A function f: $[0, \infty)^2 \to \mathbf{R}$ is said to be doubly monotone iff it is monotone in both coordinates. The corresponding (partial) order on $\mathbf{R}^2$ is given by $(x_1, y_1) \preceq (x_2, y_2)$ iff $x_1 \geq x_2$ and $y_1 \geq y_2$. Hence, a doubly monotone function in $[0, \infty)^2$ is unimodal with mode $(0, 0)$ and has level sets which are subgraphs of a nonincreasing function from $[0, \infty)$ to $[0, \infty)$. This class of sets in $[0, \infty)^2$ with monotone boundary is the corresponding class of upper sets.

Sager's multivariate unimodal MLEs. For modeling unimodality in $\mathbf{R}^d$, Sager (1982) considered two different classes of sets $\mathscr{C}_{\preceq}$. One is the class of ellipsoids with known (or estimated) location vector μ and scale parameter Σ, a positive definite $d \times d$ matrix. The corresponding order is given through $x \preceq y$ iff $(y - \mu)'\Sigma(y - \mu) \leq (x - \mu)'\Sigma(x - \mu)$. The other σ-lattice considered by Sager, $\mathscr{S}$, say, is defined through the following property: $S \in \mathscr{S}$ iff $x \in S$ implies $[0, x] \in \mathscr{S}$, where $[0, x]$ denotes the d-dimensional interval $[0, x_1] \times [0, x_2] \times \cdots \times [0, x_d]$, for $x = (x_1, \ldots x_d) \geq 0$, and replace $[0, x_i]$ through $[x_i, 0]$ for $x_i < 0$. This class $\mathscr{S}$ corresponds to unimodal densities in higher dimensions with mode 0. For $d = 2$ the corresponding order is the doubly monotone order considered above applied to each quadrant separately. In both cases, estimating the mode results in a MLE conditional on the estimated mode.

Discrete cases. Let $\mathscr{X} = \{x_1, \ldots, x_k\}$, $x_i \in \mathbf{R}$. Without loss of generality assume the x_i's to be ordered, and let $\mathscr{C} = [\{\{x_1, \ldots, x_j\}, j = 1, \ldots, k\}]$. Further, choose ν as the counting measure. The corresponding silhouette is the MLE of the probabilies $p_i = P\{x_i\}$, $i = 1, \ldots, k$ under the restriction that the p_i's are monotone decreasing [for more on this MLE, see Barlow, Bartholomew, Bremner and Brunk (1972)]. Analogously, discrete unimodal situations can be modeled (cf. the Introduction).

3.2. *Shape restrictions.* For the construction of uni- or multimodal, univariate MLEs, one has to assume the mode(s) to be given (see above). Dropping this assumption usually leads to shape restrictions which are no longer order restrictions.

For instance, as a *univariate unimodal* shape restriction, choose $\mathscr{C}$ as the class of *all* intervals. A MLE in the corresponding class $\mathscr{F}_{\mathscr{C}}$ formally does not exist, since the maximum likelihood product is infinite here (see also end of Section 4). In contrast to that, the silhouette in $\mathscr{F}_{\mathscr{C}}$ can be calculated and usually leads to reasonable estimates [cf. Müller and Sawitzki (1991)]. However, it might happen that the silhouette does not lie in $\mathscr{F}_{\mathscr{C}}$; this is to say, it might not be unimodal. Nevertheless, with increasing sample size it converges to the underlying density f [cf. Polonik (1995b)].

Similar remarks apply to *bimodal situations* on the real line without known (or estimated) mode. There an appropriate class is $\mathscr{C} = \mathscr{I} \cup \mathscr{I} = \{C \subset \mathbf{R}\colon C = I_1 \cup I_2; I_1, I_2 \text{ intervals}\}$. But knowing the modes does not necessarily lead to an order restriction, as can be seen from $\mathscr{C} = \mathscr{I}_{x_1, x_2} = \{C \subset \mathbf{R}\colon C = I_1 \cup I_2; I_1, I_2 \text{ intervals}, x_i \in I_i, i = 1, 2\}$.

Multivariate shape restrictions are given through the assumptions of elliptic or convex density contours. In our language this translates to choose $\mathscr{C}$ as the class $\mathscr{E}^d$ or $\mathscr{C}^d$, respectively. Here again the silhouette need not lie in the corresponding model class, but it converges asymptotically to the underlying density if the model is correct. See Sager (1979) for another, but related density estimate with convex level sets.

4. Concentration functions and optimal sets. Above we used the (empirical) excess mass function E_n to define the silhouette. As has already been mentioned, this function can be considered as a concentration function. Now we introduce two additional concentration functions, q_n and $\tilde{F}_n$, which have a close connection to E_n (see below). We shall use them to formulate the graphical representation and an existence theorem for the MLE. They are defined as

$$q_n(\alpha) = \inf_{C \in \mathscr{C}} \{\nu(C)\colon F_n(C) \geq \alpha\}, \qquad \alpha \in [0, 1] \tag{4.1}$$

and

$$\tilde{F}_n(l) = \sup_{C \in \mathscr{C}} \{F_n(C)\colon \nu(C) \leq l\}, \qquad l \geq 0. \tag{4.2}$$

Here q_n is a generalized quantile function in the sense of Einmahl and Mason (1992) [see Polonik (1997) 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 Lévy-type concentration function [see Hengartner and Theodorescu (1973)]. Any set $C_n(\alpha) \in \mathscr{C}$ such that

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

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

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

is called an (*empirical*) *modal set* in $\mathscr{C}$ at level l with respect to ν. Given observations $X_1, \dots, X_n$, the set of all MV sets at level α is denoted by $\mathscr{M}\mathscr{V}_n(\alpha)$, and $\mathscr{M}\mathscr{V}_n = \bigcup_{\alpha\in[0,1]} \mathscr{M}\mathscr{V}_n(\alpha)$ denotes the set of all MV sets. Analogously, let $\mathscr{M}\mathscr{O}_n(l)$ and $\mathscr{M}\mathscr{O}_n$ 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 $\nu = \text{Leb}$, where Leb denotes Lebesgue measure (in $\mathbf{R}^d$). A special case of a MV set is the well-known shorth [cf. Andrews, Bickel, Hampel, Huber, Rodgers and Tukey (1972)] which is the MV set in the class of one-dimensional intervals at the level 1/2. For this class of one-dimensional intervals, q_n has been considered by Grübel (1988). Chernoff (1964) used the midpoint of modal intervals, that is, modal sets in the class of one-dimensional intervals, as estimators of the mode. Similarly, Venter (1967) used MV intervals instead of modal intervals. Note that in the literature the notion "modal set" is also used in a broader sense, so that MV sets are sometimes called modal sets also [e.g., Lientz (1970)].

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

(A4) *Almost surely there exist MV sets and modal sets with finite ν-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, $\mathscr{C}$ is closed under intersections. This closedness of course is not a necessary condition for (A4) to hold, as can be seen from the case $\mathscr{C} = \mathscr{E}^2$.

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 $\mathscr{X} = [0,1]$, $\nu =$ Lebesgue measure and let $\mathscr{C} = \{\varnothing, [0,1/2), [1/2,1], \mathscr{X}\}$. 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 all the sets in $\mathscr{C}$ are MV sets. In general we have the following lemma.

LEMMA 4.1. *Given observations $X_1, \dots, X_n$ the following are equivalent*:

(i) $\exists\, \Gamma \in \mathscr{M}\mathscr{V}_n \cap \mathscr{M}\mathscr{O}_n$ *with* $\nu(\Gamma) = l$, $F_n(\Gamma) = \alpha$;
(ii) $\tilde{F}_n$ *is discontinuous at l and* $\tilde{F}_n(l) = \alpha$;
(iii) q_n *is discontinuous at α and* $q_n(\alpha) = l$.

Note that by definition

$$\Gamma_{n,\mathscr{C}}(\lambda) \in \mathscr{M}\mathscr{V}_n\big(F_n(\Gamma_{n,\mathscr{C}}(\lambda))\big) \cap \mathscr{M}\mathscr{O}_n\big(\nu(\Gamma_{n,\mathscr{C}}(\lambda))\big). \tag{4.3}$$

Therefore, it follows from Theorem 2.3 that for σ-lattices $\mathscr{C}$, every level set of the MLE in $\mathscr{F}_{\mathscr{C}}$ is both (empirical) MV set and modal set. However, not every set which is both MV set and modal set is an empirical λ-cluster (see below). In general the set of all empirical λ-clusters is much smaller than $\mathscr{M}\mathscr{V}_n \cap \mathscr{M}\mathscr{O}_n$. It also follows from (4.3) 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 definitions (4.1) and (4.2), respectively. The level sets of the underlying density f are both (theoretical) MV sets and modal sets, provided all level sets lie in $\mathscr{C}$. MV sets as estimators of level sets are studied in Polonik (1997).

Now we give a 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 1 with at most $n+1$ different values. Hence, $\tilde{F}_n^*$ is a convex function which is piecewise linear, increasing, bounded by 1 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 λ.

LEMMA 4.2. *For each fixed $\lambda \geq 0$, the empirical excess mass $E_{n,\mathscr{C}}(\lambda)$ equals the intercept of the tangent (from above) with slope λ to $\tilde{F}_n^*$. In other words, $E_{n,\mathscr{C}}$ is the Legendre transform of the convex function $-\tilde{F}_n^*$ (restricted to positive λ).*

Closely related to that fact is the graphical representation of the silhouette given in Polonik (1995b): the at most n different positive values $\lambda_1, \dots, \lambda_{k_n}$ of (P) [see (2.5)] 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 λ-clusters $\Gamma_{n,\mathscr{C}}(\lambda_i)$, $i = 1, \dots, k_n$. For σ-lattices $\mathscr{C}$, the values λ_i and the corresponding sets $\Gamma_{n,\mathscr{C}}(\lambda_i)$ are the different levels and level sets, respectively, of the silhouette. The same graphical representation holds for the MLE in $\mathscr{F}_{\mathscr{C}}$, provided $\mathscr{C}$ is a σ-lattice (see Theorem 5.3). This fact then proves Thorem 2.3.

5. A graphical representation of the MLE.

We start this section with two properties of the MLE in $\mathscr{F}_{\mathscr{C}}$. Both will be used to prove the graphical representation of the MLE given below (Theorem 5.3). However, they also have some interest of their own.

LEMMA 5.1. *If f_n^* is a MLE in $\mathscr{F}_{\mathscr{C}}$, then*

$$\frac{1}{n} \sum_{\{i:\ X_i \in C\}} \frac{1}{f_n^*(X_i)} \leq \nu(C) \tag{5.1}$$

for all $C \in \mathscr{C}$ such that $(f_n^* + \varepsilon \mathbf{1}_C)/(1 + \varepsilon \nu(C)) \in \mathscr{F}_{\mathscr{C}}$ for $\varepsilon > 0$ small enough. If $\mathscr{C}$ is a σ-lattice, then (5.1) holds for all $C \in \mathscr{C}$.

It is well known, that a MLE in $\mathscr{F}_{\mathscr{C}}$ does not exist if there exist sets $C \in \mathscr{C}$ with $F_n(C) > 0$ and arbitrary small ν-measure. This can also been seen from Lemma 5.1.

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

LEMMA 5.2. *Suppose that $\mathscr{C}$ is closed under intersections. Given $X_1, \dots, X_n$, let $\mathscr{L}_n = \{L \in \mathscr{C}: L = \bigcap\{C \in \mathscr{C}: X^{\pi} \subset C\}$ for some subset $\pi \subset \{1, \dots, n\}\}$. For any function $f \in \mathscr{F}_{\mathscr{C}}$ with $\prod_{i=1}^{n} f(X_i) > 0$ there exists a function $f^* \in \mathscr{F}_{\mathscr{L}_n}$ with $\prod_{i=1}^{n} f^*(X_i) \geq \prod_{i=1}^{n} f(X_i)$. Hence, if a MLE f_n^* in $\mathscr{F}_{\mathscr{C}}$ exists, then we have*

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

Note that the class $\mathscr{L}_n$ is finite (for a given realization $X_1, \dots, X_n$) and that it contains all MV sets in $\mathscr{C}$ with nonzero ν-measure. We shall see later (Corollary 5.4), that for σ-lattices $\mathscr{C}$ the assertion of Lemma 5.2 holds with the class $\mathscr{L}_n$ replaced by the class of all MV sets (or of all modal sets) which in general is a much smaller class. Lemma 5.2 not only says, that the MLE in $\mathscr{F}_{\mathscr{C}}$ is piecewise constant with at most $(n+1)$ different 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 4 that $\tilde{F}_n^*$ is a piecewise linear, increasing function with at most n changes of slope. These changes of slope occur at levels l where $\tilde{F}_n(l) = \tilde{F}_n^*(l)$. Let $l_1, \dots, l_{k_n}$, $k_n \leq n$ denote those levels in decreasing order and denote by s_i the left-hand derivatives of $\tilde{F}_n^*$ at l_i, $i = 1, \dots, k_n$ (see Figure 1). Note that $s_i < s_{i+1}$, $i = 1, \dots, k_n - 1$. Further denote $\alpha_i = \tilde{F}_n(l_i)$, $i = 1, \dots, k_n$, such that l_i is the ν-measure of the MV set at the level α_i. Given a MLE f_n^* in $\mathscr{F}_{\mathscr{C}}$, let $0 = f_0 < f_1 < \cdots < f_{k_n^*}$, $k_n^* \leq n$ denote the different levels of f_n^* and let $\Gamma_{n,\mathscr{C}}^*(f_i)$ be their corresponding (different) level sets at the levels f_i. With these notations we have the theorem.

THEOREM 5.3. *Suppose that* (A1) *and* (A4) *hold and that a MLE in $\mathscr{F}_{\mathscr{C}}$ exists. If $\mathscr{C}$ is a σ-lattice, then we have for any MLE f_n^* with the above notation that $k_n^* = k_n$ and:*

(i) $f_i = s_i \; \forall\, i = 1, \dots, k_n$;
(ii) $\Gamma_{n,\mathscr{C}}^*(f_i) \in \mathscr{M}\mathscr{V}_n(\alpha_i) \cap \mathscr{M}\mathscr{O}_n(l_i) \; \forall\, i = 1, \dots, k_n$.

Theorem 5.3(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 (see Figure 2). This proves Theorem 2.3, because the silhouette has the same graphical representation [cf. Polonik (1995b)] which, however, not only holds for σ-lattices; see comments after Lemma 4.2.

The following corollary is an easy consequence of Lemma 5.2 and Theorem 5.3.

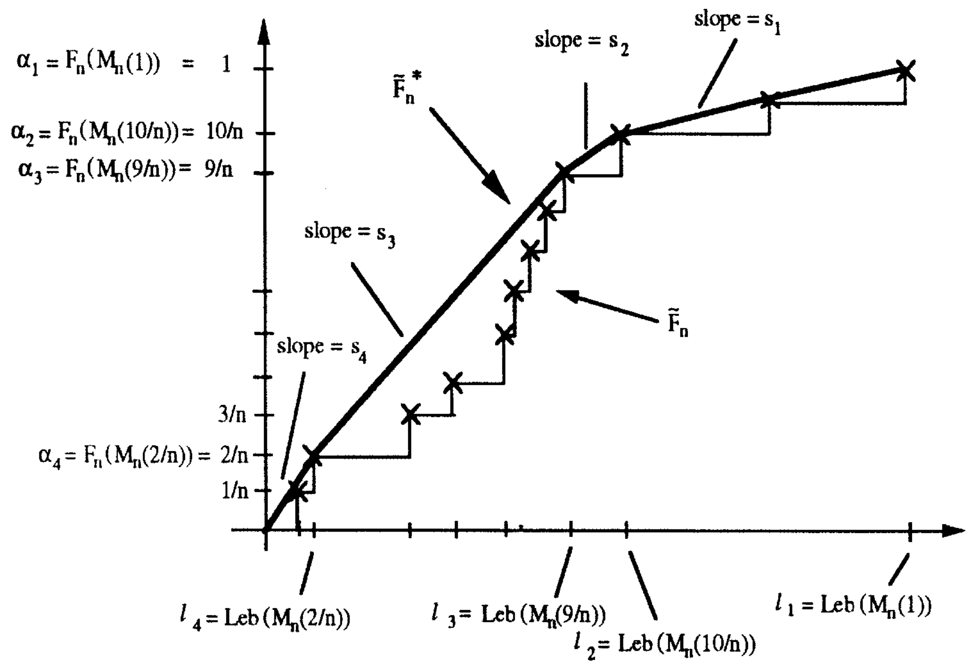

FIG. 1. *The notation used in Theorem* 5.3 *with* $\nu =$ Lebesgue *measure*, *denoted by* Leb. *A possible realization of* $\tilde{F}_n$ *and* $\tilde{F}_n^*$ *for* $n = 12$ *is shown.*

COROLLARY 5.4. *Suppose that the assumptions of Theorem* 5.3 *hold. Let*

$$V_{\mathscr{MV}_n} = \left\{ \frac{F_n(A \setminus B)}{\nu(A \setminus B)} \colon A, B \in \mathscr{MV}_n, B \subset A \right\}$$

and

$$V_{\mathscr{MO}_n} = \left\{ \frac{F_n(A \setminus B)}{\nu(A \setminus B)} \colon A, B \in \mathscr{MO}_n, B \subset A \right\}.$$

Then we have

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

The concentration functions $\tilde{F}_n$ and q_n can also be used to characterize the existence of a MLE in $\mathscr{F}_\mathscr{C}$ which is an assumption in Theorem 5.3.

THEOREM 5.5. *Suppose that* (A1) *and* (A4) *hold and that* $\mathscr{C}$ *is closed under intersection*; *then the following are equivalent*:

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

The well-known fact, that a.s. there exists a MLE in the class of monotone decreasing densities on $[0, \infty)$ follows from Theorem 5.5, since the smallest

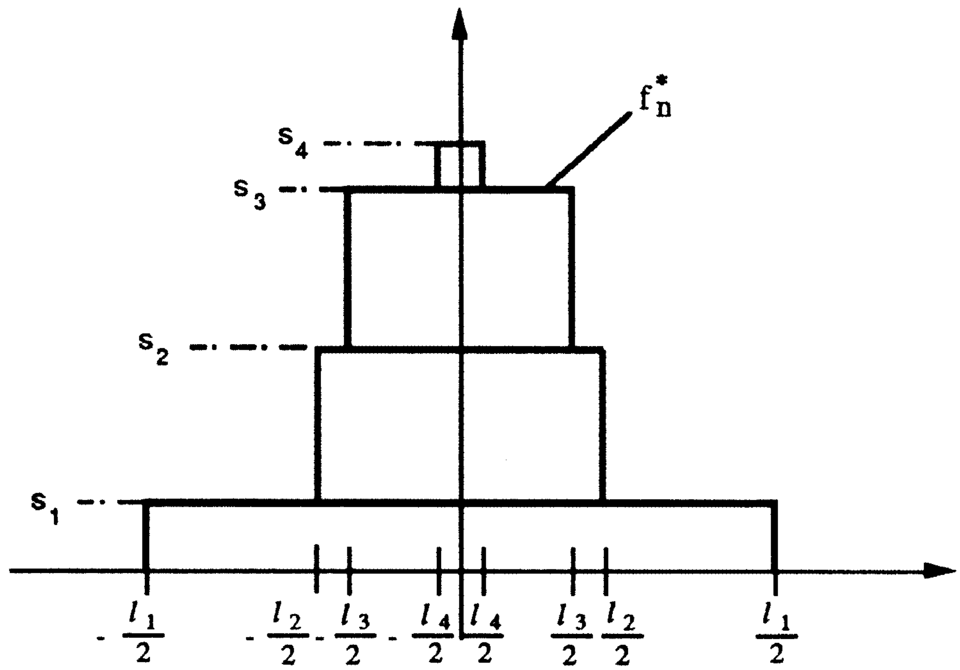

FIG. 2. *The construction of an MLE (corresponding to Figure* 1) *as given in Theorem* 5.3. *The class* $\mathscr{C}$ *is chosen to be the class of all intervals with midpoint zero, which is a* σ*-lattice. The four different level sets are the sets* $M_n(2/n)$, $M_n(9/n)$, $M_n(10/n)$, *and* $M_n(1)$, *respectively, corresponding to Figure* 1.

MV set in $\mathscr{I}_0$ which is $[0, X_{(1)}]$, where $X_{(1)}$ denotes the first order statistic, has a.s. positive Lebesgue measure. Theorem 5.5 also says, that for $\mathscr{C} = \mathscr{I}^1$, or more generally, for the classes $\mathscr{C} = \mathfrak{C}^d$ the MLE in $\mathscr{F}_{\mathscr{C}}$ does not exist. However, if, for example, one removes all sets from $\mathfrak{C}^d$ with Lebesgue measure bigger than a fixed positive ε (with the exception of the empty set), then the MLE exists [see Wegman (1970) or Robertson, Wright and Dykstra (1988)]. Of course there exist other ways to modify the class $\mathscr{C}$ in order to ensure the existence of a MLE. For example, a data dependent approach is given by measuring the significance of a given set through the (empirical) excess mass it carries. More precisely, consider only sets $C \in \mathscr{C}$ with $H_{n,\lambda}(C) > \varepsilon$. Since the value $H_{n,\lambda}(C)$ is interpreted as a measure of mass concentration, it should be easier to choose an ε here. A similar method, also based on excess mass, has been used by Müller and Sawitzki (1991) in the context of the silhouette. These approaches also reduce the well-known problem of *spiking* of the MLE (and of the silhouette) [cf. Wegman (1970)].

6. Algorithmic aspects. Theorem 5.3 immediately provides an algorithm for calculating the MLE under order restrictions: First, calulate all the modal sets or, alternatively, all the minimum volume sets $C_n(i/n)$, $i = 0, \dots, n$ in the corresponding class $\mathscr{C} = \mathscr{C}_{\preceq}$. Then calculate the least concave majorant to the points $(F_n(C_n(i/n)), \nu(C_n(i/n)))$, $i = 0, \dots, n$. The empirical MV sets corrresponding to the extremal points which define a certain (automatically selected) subclass of the MV sets give the different level sets of the MLE. The left-hand slopes of the concave minorant in these extremal points give the corresponding levels. For calculation of the concave majorant, one can use well-known algorithms such as the *pool adjacent violator algorithm* (*PAVA*) [cf. Robertson, Wright and Dykstra (1988)].

All this says is that calculation of the MLE can be reduced to calculation of empirical MV sets in the class $\mathscr{C}_{\preceq}$. This point of view, provides a unified way to look at the problem. However, from an algorithmic point of view, this cannot be studied systematically. In each particular case one has to think about how to calculate the MV sets efficiently. It turns out that in special cases this has already been done implicitly. Hence, as demonstrated below, the MV point of view leads to both new and well-known algorithms or representations, respectively, for MLEs. We refer to Section 3.1 for a brief description of the special cases discussed now.

The Grenander estimator. For $\mathscr{C} = \{[0, x], x \geq 0\}$ we have $C_n(i/n) = [0, X_{(i)}]$, where $X_{(i)}$ denotes the ith order statistic, such that $F_n(C_n(i/n)) = i/n$ and $\nu(C_n(i/n)) = X_{(i)}$. Hence, the above algorithm (or representation) gives us back the well-known representation of the Grenander estimator as the least concave majorant of the empirical distribution function.

Other univariate MLEs. In the unimodal case instead of calculating a convex minorant–concave majorant, one can do the following. Calculate all the empirical MV sets in the class of all intervals which contain the given mode. Plot the empirical mass content of these sets against their length, and calculate the least concave majorant to these points [see Robertson, Wright and Dykstra (1988) for another representation]. Similarly, for the bimodal case calculate the MV sets in the classes $\mathscr{I}_{x_1,x_2}(a)$ for $a \in (x_1, x_2) \cap \{X_1, \dots, X_n\}$, and caculate the corresponding silhouette (or MLE) in each class $\mathscr{I}_{x_1,x_2}(a)$. Finally, select the one with the largest likelihood product.

Elliptic contours. Given the location vector μ and scatter matrix Σ (either known or estimated) the MLE, or conditional MLE, respectively, can be calculated easily: calculate the values $(X_j - \mu)'\Sigma(X_j - \mu)$, $j = 1, \dots, n$, and order them. The ellipsoid corresponding to the ith largest of these values is the set $C_n(i/n)$. The empirical mass content of course is i/n and the corresponding Lebesgue measures can be calculated easily. Now plot i/n versus the Lebesgue measures and calculate the least concave majorant. [This seems to be the algorithm used by Sager (1982).]

Doubly monotone MLE. Here the empirical MV sets are sets which have monotone decreasing piecewise constant boundary with vertices at the observations. Hence, a possible algorithm searches through all these sets to find the MV sets. However, if the number of observations is large, the class to search in becomes very large. Therefore one has to try to develop algorithms using the fact that we are searching for "optimal" sets in the sense of the MV property. This has not been considered in detail yet. Let us mention here that other known algorithms for this problem (as the related minimum lower set algorithm) are also known to be very complex [cf. Robertson, Wright and Dykstra (1988)].

Convex contours. To calculate the silhouette with $\mathscr{C} = \mathfrak{C}^2$, an algorithm of Hartigan (1987) can be used. In contrast to the above, this algorithm calculates the level sets of the silhouette directly by minimizing the (empirical) excess mass function for a given level λ. Hence, in order to calculate the silhouette one has to use this algorithm for all levels $\lambda > 0$. Alternatively, one can use ideas from the Hartigan algorithm to develop an algorithm for calculation of all the convex MV sets. (Such an algorithm is available from the author.)

7. Summary and final remarks.

SUMMARY. The present paper brings together the so-called excess mass approach and ML-density estimation under order restrictions. This connection provides new points of view for ML-density estimation under order restriction. The presented results are of a nonasymptotic nature.

We show that if a MLE under a given order restriction exists, then it equals the silhouette corresponding to the class of upper sets. The silhouette can also be used to characterize the existence of a MLE.

A key result is the graphical representation of the MLE under order restrictions formulated in Theorem 5.3. It generalizes the well-known representation of the Grenander density estimator as the least concave majorant to the empirical distribution function. Algorithmic aspects of this representation are also discussed.

FINAL REMARKS. (a) Results about the asymptotic behavior of the silhouette can be found in Polonik (1995b). In view of Theorem 2.3 of the present paper, these results can also be interpreted as results on the MLE under order restirictions. In other words, the strong tools of empirical process theory can be used to derive results about the MLE under order restriction. This is another aspect to ML-density estimation under order restriction brought in through the connections developed in the present paper.

(b) Note that the dominating measure ν used here need not be Lebesgue or counting measure. This, for example, enables us to do the following: let g denote Lebesgue density of some known measure G. Suppose one is interested in estimating the Lebesgue density f of F under the additional information that $h = f/g$ satisfies some order restrictions. 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 $\hat{h}$ can immediately be translated to results about $\hat{f}$.

(c) The presented MV-approach to ML-density estimation can also be applied to other situations, such as regression problems or reliability theory. This will be studied in a seperate paper. Let us just mention here that in these cases the empirical process has to be replaced by other processes coming out of the particular situation naturally. In the regression case this will be the generalized partial sum process, and in reliability theory an appropriate transformation of the empirical process can be used.

(d) There is another (univariate) shape restriction often considered in the literature: one assumes the underlying target function f to be concave (or convex). This assumption does not directly fit the approach in the present paper. However, basically one might think about letting the derivative f' take over the role of f. This will be studied elsewhere.

8. Proofs.

PROOF OF LEMMA 2.2. Let for any $c > 0$,

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

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

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

it follows that

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

from which the first assertion follows.

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

$$H_{n,\lambda_j}(\Gamma_{n,\mathscr{C}}(\lambda_j) \cup \Gamma_{n,\mathscr{C}}(\lambda_{j+1})) \tag{8.1}$$

$$= H_{n,\lambda_j}(\Gamma_{n,\mathscr{C}}(\lambda_j)) + H_{n,\lambda_j}(\Gamma_{n,\mathscr{C}}(\lambda_{j+1}) \setminus \Gamma_{n,\mathscr{C}}(\lambda_j)), \tag{8.2}$$

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

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

On the other hand we have

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

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

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

PROOF OF LEMMA 4.1. (i) ⇒ (ii): If $\tilde{F}_n$ has no jump at $l = F_n(\Gamma)$ then there exists a set $C \in \mathscr{C}$ with $\nu(C) = l_0 < l$ and $F_n(C) = \alpha = F_n(\Gamma)$. Hence, $q_n(\alpha) \le l_0 < l$ and it follows that $\Gamma \notin \mathscr{MV}_n$.

(ii) ⇒ (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 $C \in \mathscr{C}$ 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) ⇒ (i): Suppose q_n has a jump at α. Then $\exists\, \Gamma$ with $F_n(\Gamma) = \alpha$ and $\nu(\Gamma) = q_n(\alpha)$. If $\Gamma \notin \mathscr{MO}_n(\nu(\Gamma))$ then $\exists\, C \in \mathscr{C}$ with $\nu(C) \le q_n(\alpha)$ and $F_n(C) > \alpha$. This implies that q_n has no jump at α, a contradiction. □

PROOF OF LEMMA 4.2. We have

$$\begin{aligned} E_n(\lambda) &= \sup_{\{C\in\mathscr{C}\}} \{F_n(C) - \lambda\nu(C)\} \\ &= \sup_{l\ge 0} \sup_{\{C\in\mathscr{C}:\ \nu(C)\le l\}} \{F_n(C) - \lambda\nu(C)\} \\ &= \sup_{l\ge 0} \{\tilde{F}_n(l) - \lambda l\}. \end{aligned}$$

The last line is the maximal difference of $\tilde{F}_n$ and a line through the origin with slope λ. 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 (1985) (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 5.1. Let $f^*_{n,\varepsilon,C} = (f^*_n + \varepsilon \mathbf{1}_C)/(1 + \varepsilon\nu(C))$. It then follows that for the ML-estimator f^*_n one has

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

for all $C \in \mathscr{C}$ such that $f^*_{n,\varepsilon,C} \in \mathscr{F}_{\mathscr{C}}$. From this, (5.1) follows by elementary calculations. The fact that (5.1) holds for all $C \in \mathscr{C}$ if $\mathscr{C}$ is a σ-lattice follows directly from the fact that in this case $\mathscr{F}_{\mathscr{C}}$ is a cone [see Robertson, Wright and Dykstra (1988)]. It can also be seen directly by noting that

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

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

PROOF OF LEMMA 5.2. Let $f \in \mathscr{F}_{\mathscr{C}}$ be arbitrary. Denote $f_0 = 0$, $f_j = f(X_j)$, $j = 1, \ldots, n$ and let $\Gamma_j = \{x\colon f(x) \ge f_j\}$ be the level sets of f at the levels f_j.

Without loss of generality assume them to be orderd, $f_0 < f_1 \leq \cdots \leq f_n$. Define

$$g(x) = cf_j \quad \text{for } x \in \Gamma_j \backslash \Gamma_{j+1},\ j = 0, \dots, n,$$

where $\Gamma_{n+1} = \varnothing$ and $c > 0$ is a norming constant to make g integrate to 1. Since $g/c \leq f$, we have $c \geq 1$. Moreover, $g \in \mathscr{F}_{\mathscr{C}}$ 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 an even larger likelihood product and level sets in $\mathscr{L}_n$. Let $\pi_j = \{i\colon X_i \in \Gamma_j\}$, $j = 0, \dots, n$ and define $\tilde{\Gamma}_j = \bigcap\{C \in \mathscr{C}\colon X^{\pi_j} \subset C\}$. Then, since $X^{\pi_{j+1}} \subset X^{\pi_j}$, we have $\tilde{\Gamma}_{j+1} \subset \tilde{\Gamma}_j$. Let $g_j = g(X_j) = cf_j$. Define

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

where as above $\tilde{c}$ is a norming constant. As above, it follows that h has a larger likelihood product than g since the norming constant is bigger than 1. By definition, h has level sets $\tilde{\Gamma}_j \in \mathscr{L}_n$. 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, Devroye (1987)] that for a given partition $A_1, \dots, A_k$ with $\nu(A_j) > 0\, \forall j = 1, \dots, k$ the histogram density, which has constant values $F_n(A_j)/\nu(A_j)$ at A_j has the largest likelihood among all densities which are constant at A_j, $j = 1, \dots, k$. This finishes the proof, since $\nu(\tilde{\Gamma}_j) > 0\, \forall i = 1, \dots, n$. This follows from the assumption that a MLE exists (cf. Theorem 5.5). □

PROOF OF THEOREM 5.3. We first prove Theorem 5.3 under the additional assumptions that (for given observations) the MV sets and modal sets at each level are monotone for inclusion. We refer to this assumption as (M).

Using Lemma 5.1, one easily gets $f_{k_n^*} \geq F_n(C)/\nu(C)$, $\forall C \in \mathscr{C}$. Since by assumption a MLE exists, it follows from Theorem 5.5 and Lemma 5.2 that

$$f_{k_n^*} = \sup_{C \in \mathscr{C}} F_n(C)/\nu(C) = \sup_{L \in \mathscr{L}_n} F_n(L)/\nu(L). \tag{8.3}$$

Clearly, any set maximizing $F_n(C)/\nu(C)$ over all $C \in \mathscr{C}$ has to be in $\mathscr{M}\mathscr{V}_n \cap \mathscr{M}\mathscr{O}_n$. Assume for the moment that this maximizing set is unique. Then it follows from Lemma 5.2 that the maximizing set is $\Gamma^*_{n,\mathscr{C}}(f_{k_n^*})$, the level set of the MLE at the maximal level $f_{k_n^*}$ such that

$$\Gamma^*_{n,\mathscr{C}}(f_{k_n^*}) \in \mathscr{M}\mathscr{V}_n \cap \mathscr{M}\mathscr{O}_n \tag{8.4}$$

and

$$f_{k_n^*} = \frac{F_n(\Gamma^*_{n,\mathscr{C}}(f_{k_n^*}))}{\nu(\Gamma^*_{n,\mathscr{C}}(f_{k_n^*}))}. \tag{8.5}$$

Equation (8.3) 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 restrict ourselves to sets $C \in \mathscr{C}$ with $\Gamma^*_{n,\mathscr{C}}(f_{k_n^*}) \subset C$. Then (5.1) gives

$$\sum_{\{j:\ X_j \in \Gamma^*_{n,\mathscr{C}}(f_{k_n^*})\}} \frac{1}{f_n^*(X_j)} + \sum_{\{j:\ X_j \in C \setminus \Gamma^*_{n,\mathscr{C}}(f_{k_n^*})\}} \frac{1}{f_n^*(X_j)} \le n\nu(C). \tag{8.6}$$

From (8.5) we get that the first term on the left-hand side of (8.6) equals $\nu(\Gamma^*_{n,\mathscr{C}}(f_{k_n^*}))$. Hence it follows that $f_{k_n^*-1}$ has to satisfy

$$f_{k_n^*-1} \ge \frac{F_n(C \setminus \Gamma^*_{n,\mathscr{C}}(f_{k_n^*}))}{\nu(C \setminus \Gamma^*_{n,\mathscr{C}}(f_{k_n^*}))} \qquad \forall\, C \in \mathscr{C} \text{ s.th. } \Gamma^*_{n,\mathscr{C}}(f_{k_n^*}) \subset C. \tag{8.7}$$

As above, it follows that

$$f_{k_n^*-1} = \sup\left\{ \frac{F_n(L \setminus \Gamma^*_{n,\mathscr{C}}(f_{k_n^*}))}{\nu(L \setminus \Gamma^*_{n,\mathscr{C}}(f_{k_n^*}))} : L \in \mathscr{L}_n, \Gamma^*_{n,\mathscr{C}}(f_{k_n^*}) \subset L \right\}. \tag{8.8}$$

Since the numerators and the denominators in (8.8) actually are differences of the corresponding measures of the sets L and $\Gamma^*_{n,\mathscr{C}}(f_{k_n^*})$ it follows by using (M) that the maximizing set in (8.8) lies in $\mathscr{M}\mathscr{V}_n \cap \mathscr{M}\mathscr{O}_n$. If we again assume that it is unique, then we have as above,

$$\Gamma^*_{n,\mathscr{C}}(f_{k_n^*-1}) \in \mathscr{M}\mathscr{V}_n \cap \mathscr{M}\mathscr{O}_n$$

and

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

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

It remains to remove the assumption of uniqueness of the maximizing sets in (8.3), (8.8) and so on and to remove (M). First note that it follows from the graphical representation of the silhouette (cf. the discussion after Lemma 4.2) together with Lemma 2.2 that there exist empirical λ-clusters $\Gamma_{n,\mathscr{C}}(\lambda_j)$, $j = 1, \dots, k_n$, (which by definition all lie in $\mathscr{M}\mathscr{V}_n \cap \mathscr{M}\mathscr{O}_n$) which are monotone for inclusion. These sets correspond to the vertices of $\tilde{F}_n^*$; this means that the points $(\nu(\Gamma_{n,\mathscr{C}}(\lambda_j)), F_n(\Gamma_{n,\mathscr{C}}(\lambda_j)))$ are vertices of the graph of $\tilde{F}_n^*$. From this it follows that the maximal values in (8.5), (8.8) and so on, that is, 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 $\tilde{F}_n^*$, that is, for any maximizing set Γ_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 λ-clusters. The corresponding value of λ equals the maximal value in (8.5), (8.8), and so on. We show below that if Γ_1 and Γ_2 are two empirical λ-clusters to the same value of λ, then also the union $\Gamma_1 \cup \Gamma_2$ is an empirical λ-cluster to this value of λ. From this we can remove the assumption of uniqueness as follows: suppose the maximizing set in (8.5) is not unique, and we did not choose the largest maximizing set, that is, the union of all maximizing sets. Then the next iteration step leads to the same maximal value; that is, we stay at the same level of the MLE, and the (present) level set of the MLE only becomes larger until we finally reach the largest level set. Hence, the uniqueness assumption is not necessary.

It remains to show that if Γ_1 and Γ_2 are two empirical λ_0-clusters, then $\Gamma_1 \cup \Gamma_2$ also is a λ_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 equality it follows that $H_{n,\lambda_0}(\Gamma_2 \setminus \Gamma_1) \le 0$, since by definition Γ_1 maximizes H_{n,λ_0} over all sets in $\mathscr{C}$. Analogously, the second equality gives $H_{n,\lambda_0}(\Gamma_2 \setminus \Gamma_1) \ge 0$ and hence it equals zero. The first equation now gives the assertion. □

PROOF OF THEOREM 5.5. The equivalence of (ii) and (iii) is obvious. (i) ⇒ (ii) follows from Lemma 5.1. This lemma inplies that the ν-measure of sets in $\mathscr{C}$ which have positive F_n-measure has (uniformly) to be bounded away from zero. This implies $\lim_{\alpha\to 0} q_n(\alpha) > 0$ as well as $\lim_{l\to 0} \tilde{F}_n(l) = 0$. Now suppose the latter to be true. This says that all the MV sets at levels $\alpha > 0$ have positive ν-measure. From this and from Lemma 5.2 the assertion follows, since all sets in $\mathscr{L}_n$, defined in Lemma 5.2, have bigger ν-measure than the MV set at level $1/n$. To see that also (iv) is equivalent, just observe that $\int f_{n,\mathscr{C}}(x)\,dx = 1$ iff there exists no set $C \in \mathscr{C}$ with $\nu(C) = 0$ and $F_n(C) > 0$. □

Acknowledgments. This paper was essentially written while I was a research fellow in the Department of Statistics at the University of California, Berkeley. I thank the Department of Statistics for their hospitality and R. Beran and D. Nolan individually.

I am grateful to Lutz Dümbgen for the hint to look at derivatives in directions of indicator functions, which is used in the proof of Lemma 5.1. This lemma eventually turned out to be one of the key results in the proof of the graphical representation of the MLE.

I also thank the Editor, an Associate Editor and the referees for helpful questions, comments and suggestions.

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.

BARLOW, R. E., BARTHOLOMEW, D. J., BREMNER, J. M. and BRUNK, H. D. (1972). *Statistical Inference Under Order Restrictions.* Wiley, London.

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

DEVROYE, L. (1987). A course in density estimation. Birkhäuser, Boston.

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

GRENANDER, U. (1956). On the theory of mortality measuremant, II. *Skand. Aktuarietidskr. J.* **39** 125–153.

GROENEBOOM, P. (1985). Estimating a monotone density. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neymann and Jack Kiefer* (L. Le Cam and R. Olshen, eds.) **2**. Wadsworth, Monterey, CA.

GRÜBEL, R. (1988). The length of the shorth. *Ann. Statist.* **16** 619–628.

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

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

HENGARTNER, W. and THEODORESCU, R. (1973). *Concentration Functions.* Academic Press, New York.

HICKEY, R. J. (1984). Continuous majorisation and randomness. *J. Appl. Probab.* **21** 924–929.

JOE, H. (1993). Generalized majorization orderings. In *Stochastic Inequalities* (M. Shaked and Y. L. Tong, eds.) **22** 145–158. IMS, Hayward, CA.

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

MARSHALL, A. W. and OLKIN, I. (1979). *Inequalities: Theory of Majorization and Its Applications.* Academic Press, New York.

MÜLLER, D. W. and SAWITZKI, G. (1987). Using excess mass estimates to investigate the modality of a distribution. Preprint 398, SFB 123, Univ. Heidelberg.

MÜLLER, D. W. and SAWITZKI, G. (1991). Excess mass estimates and tests of multimodality. *J. Amer. Statist. Assoc.* **86** 738–746.

NOLAN, D. (1991). The excess mass ellipsoid. *J. Multivarite Anal.* **39** 348–371.

POLONIK, W. (1992). The excess mass appraoch to cluster analysis and related estimation procedures. Ph.D. dissertation, Univ. Heidelberg.

POLONIK, W. (1995a). Measuring mass concentrations and estimating density contour clusters: an excess mass approach. *Ann. Statist.* **23** 855–881.

POLONIK, W. (1995b). Density estimation under qualitative assumptions in higher dimensions. *J. Multivariate Anal.* **55** 61–81.

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

ROBERTSON, T. (1967). On estimating a density measurable with respect to a σ-lattice. *Ann. Math. Statist.* **38** 482–493.

ROBERTSON, T., WRIGHT, F. T. and DYKSTRA, R. L. (1988). *Order restricted statistical inference.* Wiley, New York.

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

SAGER, T. W. (1982). Nonparametric maximum likelihood estimation of spatial patterns. *Ann. Statist.* **10** 1125–1136.

VENTER, J. H. (1967). On estimation of the mode. *Ann. Math. Statist.* **38** 1446–1455.

WEGMAN, E. (1969). A note on estimating a unimodal density. *Ann. Math. Statist.* **40** 1661–1667.

WEGMAN, E. (1970). Maximum likelihood estimation of a unimodal density function. *Ann. Math. Statist.* **41** 457–471.

INSTITUT FÜR ANGWANDTE MATHEMATIK
UNIVERSITÄT HEIDELBERG
IM NEUENHEIMER FELD 294
69120 HEIDELBERG
GERMANY
E-MAIL: polonik@statlab.uni-heidelberg.de