Complexity Penalized Support Estimation

Jussi Klemelä*
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294, 69120 Heidelberg, Germany
Email: klemela@statlab.uni-heidelberg.de
Fax +49 6221 5331

Abstract

We consider the estimation of the support of a probability density function with iid observations. The estimator to be considered is a minimizer of a complexity penalized excess mass criterion. We present a fast algorithm for the construction of the estimator. The estimator is able to estimate supports which consists of disconnected regions. We will prove that the estimator achieves minimax rates of convergence up to a logarithmic factor simultaneously over a scale of Hölder smoothness classes for the boundary of the support. The proof assumes a sharp boundary for the support.

Mathematics Subject Classifications (AMS 2000): 62G07
Key Words: adaptive estimation, data dependent partitions, quality control, multivariate data, tree structured estimators.
Short title: Support Estimation

1 Introduction

We will present a method for the estimation of a support of a multivariate probability density function. The method works also for the estimation of the support of an intensity function of a Poisson process. The estimator is spatially flexible, allowing us to estimate supports which consist of disconnected components.

*Writing of this article was financed by Deutsche Forschungsgemeinschaft under project MA1026/6-2.

The estimation of density support may be applied to the detection of abnormal behaviour of the system, plant, or machine. We may apply our estimator to define a nonparametric multivariate method for statistical quality control, which could extend the Shewart methodology based on tolerance regions, see Derman and Ross (1997). Support estimation may also be applied to measure performance of an enterprise in terms of technical efficiency measured by distance from the observed productivity to the boundary, see Deprins, Simar and Tulkens (1984). We may apply our estimator to the estimation of the support of a Poisson intensity. This may be applied for example to estimate the boundary of a forest, when the location of individual trees is distributed according to a planar Poisson process with unknown intensity function.

The previous methods for the support estimation may be classified at least to three categories:

1. piecewise polynomial estimators,

2. estimators which are a union of balls centered at observations,

3. estimators which are based on the convex hull of sample points.

Piecewise polynomial estimators are defined for boundary fragments by partitioning the fragment to intervals and by estimating the boundary on each interval by a polynomial. For star shaped sets one may use piecewise polynomial approximation on sectors. A piecewise constant estimator was proposed by Geffroy (1964). Korostelev and Tsybakov (1993*a*) study piecewise polynomial estimator of maximum likelihood type. They derive minimax rates of convergence when the support has a sharp boundary. Härdle, Park and Tsybakov (1995) consider support estimation with a piecewise polynomial estimator when the boundary of the support is not sharp.

The estimator which is a union of balls centered at observations amounts to estimating the support of the density by the support of a kernel estimate whose kernel has a ball shaped support. These types of estimators were considered by Devroye and Wise (1980), Cuevas and Fraiman (1997), Walther (1997), Baíllo, Cuevas and Justel (2000).

When the support is a convex set, it makes sense to estimate it by a convex hull of sample points. This type of estimator was studied by Rényi and Sulanke (1963), Rényi and Sulanke (1964), Chevalier (1976). Ripley and Rasson (1977) defines a blown-up version of the convex hull in order to eliminate bias. A review is given by Schneider (1988). Korostelev and Tsybakov (1994) and Mammen and Tsybakov (1995) derive the minimax rates of convergence for the estimation of a convex set. Korostelev and

Tsybakov (1994) establish 96% efficiency of a certain blown-up version of the convex hull estimator. Korostelev, Simar and Tsybakov (1995) consider sharp asymptotics for the case when the support is a monotone boundary fragment. Gijbels, Mammen, Park and Simar (1999) consider estimation of a support of a distribution when the support is a convex set or bounded by a monotone function. Their problem arises in an econometric problem where the frontier functions of production sets shall be estimated.

Korostelev and Tsybakov (1993*b*) contains results on estimators belonging to all three categories. Hall, Nussbaum and Stern (1997) consider a different type of estimator which is based on order statistics. Mammen and Tsybakov (1995) study density support problem under a general setting of entropy conditions. Their set up includes regions with boundaries that fullfill smoothness conditions (Dudley classes) and convex sets. Polonik (1995) derives rates of convergence for support estimation based on excess mass estimates.

We will define a new type of estimator which does not belong to any of the previous groups. The closest relative is the group of piecewise polynomial estimators, since the simplest form of our estimator may be seen as a histogram type estimator with a data-dependent partition. Our method is related to the classification and regression trees as defined by Breiman, Friedman, Olshen and Stone (1984), and to dyadic CART as defined by Donoho (1997). This type of method was first applied to boundary estimation in Donoho (1999), who studied the estimation of the boundary of two dimensional regression function with regular and fixed design.

The methods of category 1 in the above classification suppose that we know the number and rough location of disconnected components of the support. The methods of category 3 presuppose that the support is a convex set. Our method is in this respect more flexible. The methods of category 2 are vulnerable to the curse of dimensionality, since they are kernel type methods based on local averaging. Our estimator is based on economical splitting of the sample space, making it possible to efficiently estimate high dimensional supports.

In this article we propose to estimate the support by minimizing a complexity penalized *excess mass functional*. Excess mass functional is defined as $-P_n(A)+\lambda \operatorname{mes}(A)$ where $P_n(A)$ is the empirical probability, $\operatorname{mes}(A) = \int_A dx$, $A \subset \mathbf{R}^d$, and $\lambda > 0$. Excess mass functional was proposed to be applied in level set estimation by Hartigan (1987), Müller and Sawitzki (1991), Polonik (1995), Tsybakov (1997). Excess mass functional is useful also for the support estimation when we choose λ to be small. The corresponding estimator is robust to outliers and we have feasible algorithms for solving the mini-

mization problem. Indeed, we may apply a dynamic programming algorithm which solves the minimization problem for spatially localized subsets of the support and then builds the global solution from the previously solved local problems. When the boundary of the support is sharp, that is, the density has a jump on the boundary, then by choosing λ to be smaller than the jump, the level set at level λ is equal to the support of the density.

We will prove that the proposed method has nearly minimax rates of convergence simultaneously over a scale of Hölder smoothness classes for the boundary. We will consider cases when the Hölder smoothness index s is in interval $(0, 2]$. The cases $s \in (0, 1]$ and $s \in (1, 2]$ require different estimators. We will prove the results using the oracle inequality approach. We have followed the approach of Donoho and Johnstone (1994) in that we choose both the basis and a model under that basis instead of choosing only the best model in a single basis. The method of using exponentially growing collection of bases has been applied for example in Donoho (1997) for fixed design regression, in Donoho (1999) for fixed design boundary estimation, in Barron, Birgé and Massart (1999) for various density, regression, and boundary estimation problems, and in Klemelä (2001) for multivariate density estimation.

In the statements of theorems we will make certain assumptions concerning the underlying distribution. This does not mean that the estimator would not behave favorably also in cases where these assumptions are not satisfied. We will define estimators without model assumptions, unlike in some cases where the support has been assumed to be star shaped or convex.

In Section 2 we define two estimators. First one is optimal for Hölder smoothness index $s \in (0, 1]$, $d \geq 2$ and second one for Hölder smoothness index $s \in (1, 2]$ for $d = 2$. In Section 2.3 we present algorithms for the construction of the estimates. In Section 3 we formulate theorems on the rate of convergence of the estimators. In Section 4 we give three simulation examples. The proofs are given in Section 5.

Simulations which were made for this article may be reproduced with a R-package which is downloadable from http://www.denstruct.org.

We will denote $\mathrm{mes}(A) = \int_A dx$. With I_A we denote the indicator of set $A \subset \mathbf{R}^d$: $I_A(x) = 1$ when $x \in A$ and $I_A(x) = 0$ otherwise. Euclidean distance in $\mathbf{R}^d$ is denoted by $\| \cdot \|$. We apply the same notation for the Euclidean distance in $\mathbf{R}^{d-1}$. The relation $a_n \sim b_n$ means $\lim_{n\to\infty}(a_n/b_n) = 1$. Generic positive constants will be denoted by $C, C_1, C_2, \ldots$ Denote $\Pi_{i=1}^d [a_i, b_i] + \eta = \Pi_{i=1}^d [a_i - \eta, b_i + \eta]$ for $\eta \geq 0$.

2 Definition of the estimators

We will consider two types of estimators: (1) an estimator which is a union of rectangles and (2) an estimator which is a union of rectangles and parts of rectangles, resulting from a skew split. Estimators are minimizers of a complexity penalized excess mass criterion among sets which can be represented as a union of sets in a certain partition.

Let $X_1, \ldots, X_n \in \mathbf{R}^d$ be random vectors with density function whose support we want to estimate.

2.1 Block estimator

Let us first consider an estimator which is a union of rectangles. We start with defining the set of partitions with the help of which we define the class of sets on which we search the minimizer. We will consider partitions which are a result of a series of dyadic splits, when by a dyadic split of a rectangle we mean a split along some coordinate axis which divides the rectangle to two equal parts.

Set of partitions. We will denote by $\mathbb{P}_n(R)$ the *set of dyadic partitions of* R, where $R \subset \mathbf{R}^d$ is a rectangle. This set consists of partitions of R that result from of a series of dyadic splits. We will give a recursive definition below.

Definition 1 *We say that* $\mathbb{P}_n(R)$ *is the* set of dyadic partitions of $R = \Pi_{i=1}^d [a_i, b_i]$, *with fineness parameter* $a > 0$, *if*

1. $\{R\} \in \mathbb{P}_n(R)$,
2. *if* $\mathcal{P} \in \mathbb{P}_n(R)$ *and* $P = \Pi_{i=1}^d [c_i, d_i] \in \mathcal{P}$, *and*

$$d_i - c_i > (b_i - a_i) 2^{-J_n}$$

for some $i = 1, \ldots, d$, *where*

$$J_n = \lceil a(d-1)^{-1} \log_2 n \rceil, \tag{1}$$

then $(\mathcal{P} \setminus \{P\}) \cup \{P_1, P_2\} \in \mathbb{P}_n(R)$ *where* P_1, P_2 *are the results of the dyadic split of* P *in the* i*:th direction.*

The definition implies a bound for the maximal fineness of partitions in set $\mathbb{P}_n(R)$: at most J_n splits will be made to any direction and the rectangles in the finest partition have volumes greater or equal to $2^{-dJ_n} \mathrm{mes}(R)$.

For the choice of the rectangle R we apply two methods.

1. With a priori considerations one finds a rectangle which contains the support. We make this assumption to analyze rates of convergence of the estimator.

2. Denote by R^- the smallest rectangle containing observations whose sides are parallel to the coordinate axes, and choose $R = R^- + \eta$ where $\eta \geq 0$ and we apply notation $\Pi_{i=1}^d [a_i, b_i] + \eta = \Pi_{i=1}^d [a_i - \eta, b_i + \eta]$. We choose R in this way in simulation examples.

We will denote later $\mathbb{P}_n = \mathbb{P}_n(R)$.

Collection of sets. As the available class of sets from which we search a minimizer we consider

$$\mathbb{A}_n = \mathbb{A}_n(R) = \{A(\mathcal{P}, W) : \mathcal{P} \in \mathbb{P}_n(R), \;\; W \in \mathcal{W}(\mathcal{P})\} \tag{2}$$

where $\mathcal{W}(\mathcal{P})$ is the set of 0-1-markers associated with partition $\mathcal{P}$,

$$\mathcal{W}(\mathcal{P}) = \{0,1\}^{\mathcal{P}} = \{(w_P)_{P \in \mathcal{P}} : w_P \in \{0,1\}\} \tag{3}$$

and $A(\mathcal{P}, W)$ is the set which is the union of those sets in partition $\mathcal{P}$ which are marked with 1 by set of markers $W = (w_P)_{P \in \mathcal{P}}$,

$$A(\mathcal{P}, W) = \bigcup \{P \in \mathcal{P} : w_P = 1\}. \tag{4}$$

Complexity penalized excess mass criterion. Let the excess mass functional be

$$\gamma_n^e(A) = -\frac{1}{n} \sum_{i=1}^n I_A(X_i) + \lambda \, \mathrm{mes}(A)$$

where $I_A(x) = 1$ when $x \in A$ and $I_A(x) = 0$ otherwise, and $\lambda > 0$. We will define the complexity of a set $A \in \mathbb{A}_n$ to be the number of sets in the corresponding partition. Let

$$D(W) = \#\{w_P = 1 : w_P \in W\} \tag{5}$$

where $W \in \mathcal{W}(\mathcal{P})$ with $\mathcal{P} \in \mathbb{P}_n$. Let the complexity penalized excess mass criterion be

$$\mathcal{E}_n(\mathcal{P}, W, \alpha) = \gamma_n^e(A(\mathcal{P}, W)) + \alpha \, D(W) \tag{6}$$

where $\mathcal{P} \in \mathbb{P}_n$, $W \in \mathcal{W}(\mathcal{P})$, $A(\mathcal{P}, W)$ is defined in (4), and $\alpha > 0$.

The estimator. We define the block estimator with the excess mass criterion by

$$\hat{A}_n^e = A(\hat{\mathcal{P}}_n^e, \hat{W}_n^e) \tag{7}$$

where

$$(\hat{\mathcal{P}}_n^e, \hat{W}_n^e) = \text{argmin}_{\mathcal{P}\in\mathbb{P}_n, W\in\mathcal{W}(\mathcal{P})}\mathcal{E}_n(\mathcal{P}, W, \alpha), \tag{8}$$

$\alpha > 0$ is the smoothing parameter, and $\mathcal{E}_n$ is defined in (6). In addition to the smoothness parameter α, this estimator depends on the "fineness" parameter a and "level set" parameter λ. Theorem 2 gives conditions for the choice of these parameters.

2.2 Half block estimator

Let us consider an estimator which has the form of a union of rectangles and halves of rectangles resulting from a skew split. We will call this estimator half block estimator. We will use a "library" of sets resembling the one defined in Donoho (1999) with the help of wedgelets. In this section we will restrict ourselves to the case $d = 2$.

The definition of the half block estimator differs from the definition of the block estimator only in that we consider a different set of partitions which will define the class of sets from which we search a minimizer.

We will consider partitions which are a result of a series of dyadic splits, with possibly a split not parallel to the coordinate axes at the final stage. We will give a recursive definition below.

Definition 2 *We say that* $\mathbb{P}_n^D(R)$ *is the* set of dyadic partitions of $R = \Pi_{i=1}^d[a_i, b_i]$ which contains skew splits*, with fineness parameters* $a > 0$ *and* $b > 0$*, when* $d = 2$*, if*

1. $\{R\} \in \mathbb{P}_n^D(R)$*,*
2. *if* $\mathcal{P} \in \mathbb{P}_n^D(R)$ *and* $P = \Pi_{i=1}^d[c_i, d_i] \in \mathcal{P}$*, and*

$$d_i - c_i > (b_i - a_i)2^{-\tilde{J}_n}$$

for some $i = 1, \ldots, d$*, where*

$$\tilde{J}_n = \lceil ad^{-1}\log_2 n\rceil, \tag{9}$$

then $(\mathcal{P} \setminus \{P\}) \cup \{P_1, P_2\} \in \mathbb{P}_n^D(R)$ *where* P_1, P_2 *are the results of the dyadic split of* P *in the* i*:th direction,*

3. *if* $\mathcal{P} \in \mathbb{P}_n^D(R)$ *and* $P = \Pi_{i=1}^d [c_i, d_i] \in \mathcal{P}$, *and*

$$d_i - c_i \geq (b_i - a_i) 2^{-\tilde{J}_n}$$

for some $i = 1, \ldots, d$, *then* $(\mathcal{P} \setminus \{P\}) \cup \{P_1, P_2\} \in \mathbb{P}_n^D(R)$ *where* P_1, P_2 *are a result of a skew split of* P *whose endpoints are on the boundary of the rectangle* P, *and the set of possible endpoints forms a grid with step size*

$$\delta_i = (b_i - a_i) 2^{-L_n}, \;\; L_n = \lceil b(d+1)^{-1} \log_2 n \rceil$$

in i:th direction. The grid is such that it contains the four vertices of P *as grid points.*

The definition allows splits not parallel to coordinate axes only for the rectangles: once this kind of split is made, it is not anymore possible to split results of this skew split. To be able to do skew splits we need that $L_n > \tilde{J}_n$.

We define the half block estimator with the excess mass criterion by

$$\tilde{A}_n^e = A(\tilde{\mathcal{P}}_n^e, \tilde{W}_n^e) \tag{10}$$

where $A(\mathcal{P}, W)$ is defined in (4),

$$(\tilde{\mathcal{P}}_n^e, \tilde{W}_n^e) = \text{argmin}_{\mathcal{P} \in \mathbb{P}_n^D, W \in \mathcal{W}(\mathcal{P})} \mathcal{E}_n(\mathcal{P}, W, \alpha),$$

$\mathcal{W}(\mathcal{P})$ is as defined in (3), $\mathcal{E}_n$ is defined in (6), and $\alpha > 0$ is the smoothing parameter. In addition to the smoothness parameter α, the estimator depends on the "fineness" parameters a and b, and "level set" parameter λ. Theorem 3 gives a result on the rate of convergence of this estimator.

Remark 1. The half block estimator is related to the wedgelet estimator as defined in Donoho (1999), who considers the estimation of the boundary of a regression function when the design is fixed and regularly spaced.

The wedgelet estimator has the binwidth $n^{-1/2}$ in the finest rectangular partition. This corresponds to the choice $a = 1$. The wedgelet estimator allows "subpixel" splits of the rectangles, and these splits have discretization step $n^{-2/3}$. This corresponds to the choice $b = 2$.

The partition in the definition of the wedgelet estimator is sligthly more restrictive than the partition of the half block estimator. The partition of the wedgelet estimator is defined by the condition that every rectangle will be splitted by a "quad-split": a split which will result in 4 rectangles. The partition of the half block estimator grows with dyadic splits. This will add flexibility and computational complexity, see Section 2.3.

2.3 Solving the minimization problem.

Let us discuss algorithms for solving the minimization problem in the definition of estimators $\hat{A}_n^e$ and $\tilde{A}_n^e$ which were given in (7) and (10).

One may solve the minimization problem by first building a large multitree whose terminal nodes represent bins of the rectangle containing the support. A path leading to a bin will represent a possible way of choosing splits. Thus to each bin of the initial rectangle R corresponds many terminal nodes of the tree. The minimization problem is solved by pruning the tree.

Growing the tree. Construct a multitree with a single root node and at most $2d$ children for every node. The root node will correspond to the initial rectangle R containing the support. We have d ways of choosing the splitting direction and each binary split will result in two bins. Thus $2d$ children will represent the rectangles resulting from binary splits in d directions.

For the case of block estimator at most J_n splits will be made for each direction, thus the tree will have dJ_n levels where J_n is defined in (1). The half block estimator will have $d\tilde{J}_n$ levels where $\tilde{J}_n$ is defined in (9).

We will record the number of observations in each bin. When some bin is empty we will not split it anymore. The resulting tree will have at most

$$\sum_{i=0}^{dJ_n}(2d)^i = O\left((2d)^{dJ_n}\right) = O\left(n^{ad\log_2(2d)/(d-1)}\right).$$

nodes for the case of block estimator and $O\left((2d)^{d\tilde{J}_n}\right)$ nodes for the case of half block estimator. In the case of half block estimator we have to record also the frequencies at the results of a skew split. Note that in the case of the wedgelet estimator defined in Donoho (1999) the tree would have

$$\sum_{i=0}^{\tilde{J}_n}(2^d)^i = O\left(2^{d\tilde{J}_n}\right)$$

nodes.

Pruning the tree. To prune the tree we start from the next to the highest level, and travel to the root node one level at a time. For each node we find out whether the split to some of the d directions helps (whether it results to a smaller complexity penalized excess mass criterion). If the split does not help, we will cut the tree below the node.

We will formulate a lemma which formalizes the idea that we may solve the global minimization problem (8) by first solving localized subproblems,

and building the global solution from the previously solved local problems. This lemma is given for the block estimator.

Lemma 1 *Let R be the initial rectangle of the estimator and let $R_0 \subset R$ be a rectangle. Let $\mathbb{P}_n(R_0)$ be defined in Definition 1. Define the set which solves the minimization problem when we localize to the rectangle R_0:*

$$\hat{A}_n^e(R_0) = A(\hat{\mathcal{P}}_n^e(R_0), \hat{W}_n^e(R_0))$$

where

$$(\hat{\mathcal{P}}_n^e(R_0), \hat{W}_n^e(R_0)) = \text{argmin}_{\mathcal{P}\in\mathbb{P}_n(R_0), W\in\mathcal{W}(\mathcal{P})}\mathcal{E}_n(\mathcal{P}, W, \alpha).$$

Let $R_0 \subset R$ be now fixed and denote with $R_{1,i}$ and $R_{2,i}$ the left and the right rectangle resulting from dyadic split of R_0 in i:th direction, $i = 1, \ldots, d$. Let

$$\begin{aligned} M \;=\; & \min\{\mathcal{E}_n(\{R_0\}, \alpha), \\ & \mathcal{E}_n\left(\hat{\mathcal{P}}_n^e(R_{1,i}), \hat{W}_n^e(R_{1,i}), \alpha\right) + \mathcal{E}_n\left(\hat{\mathcal{P}}_n^e(R_{2,1}), \hat{W}_n^e(R_{2,i}), \alpha\right), \\ & \mathcal{E}_n\left(\hat{\mathcal{P}}_n^e(R_{1,i}), \hat{W}_n^e(R_{1,i}), \alpha\right), \\ & \mathcal{E}_n\left(\hat{\mathcal{P}}_n^e(R_{2,i}), \hat{W}_n^e(R_{2,i}), \alpha\right) : i = 1, \ldots, d\Big\}. \end{aligned}$$

Then,

$$\hat{A}_n^e(R_0) = \begin{cases} R_0, \;\; \textit{when } M = \mathcal{E}_n(\{R_0\}, \alpha) \\ \hat{A}_n^e(R_{1,i}) \cup \hat{A}_n^e(R_{2,i}), \;\; \textit{when } M = \mathcal{E}_n\left(\hat{\mathcal{P}}_n^e(R_{1,i}), \hat{W}_n^e(R_{1,i}), \alpha\right) \\ \qquad\qquad + \mathcal{E}_n\left(\hat{\mathcal{P}}_n^e(R_{2,i}), \hat{W}_n^e(R_{2,i}), \alpha\right) \\ \hat{A}_n^e(R_{1,i}), \;\; \textit{when } M = \mathcal{E}_n\left(\hat{\mathcal{P}}_n^e(R_{1,i}), \hat{W}_n^e(R_{1,i}), \alpha\right) \\ \hat{A}_n^e(R_{2,i}), \;\; \textit{when } M = \mathcal{E}_n\left(\hat{\mathcal{P}}_n^e(R_{2,i}), \hat{W}_n^e(R_{2,i}), \alpha\right). \end{cases}$$

Proof. Let the collection of sets $\mathbb{A}_n(R_0)$ from which we search a minimizer be defined in (2). We may express $\mathbb{A}_n(R_0)$ recursively:

$$\begin{aligned} \mathbb{A}_n(R_0) \;=\; & \{R_0\} \cup \{A_1 \cup A_2 : A_k \in \mathbb{A}_n(R_{k,i}), \;\; k = 1, 2, \;\; i = 1, \ldots, d\} \\ & \cup \bigcup_{k=1}^{2}\bigcup_{i=1}^{d}\mathbb{A}_n(R_{k,i}). \end{aligned}$$

On the other hand, when $\mathcal{P}_k \in \mathbb{P}_n(R_{k,i})$, $W_k \in \mathcal{W}(\mathcal{P}_k)$, $k = 1, 2$, $i = 1, \ldots, d$, then

$$\mathcal{E}_n(\mathcal{P}_1 \cup \mathcal{P}_2, W_1 \cup W_2, \alpha) = \mathcal{E}_n(\mathcal{P}_1, W_1, \alpha) + \mathcal{E}_n(\mathcal{P}_2, W_2, \alpha).$$

Indeed, this follows directly from definition (6) since $\mathcal{P}_1$ and $\mathcal{P}_2$ are partitions of disjoint rectangles. We have proved the lemma. □

In particular, when we choose $R_0 = R$ in Lemma 1, then $\hat{A}_n^e(R) = \hat{A}_n^e$ is the global solution defined in (7).

We give in the following the pseudo code for the pruning algorithm in the case of the block estimator.

- Input for the algorithm is the smoothing parameter $\alpha > 0$ and a multitree, whose nodes represent certain bins. We will denote by $\text{left}_i(m)$ and $\text{right}_i(m)$ the pointers to the left and right childs of node m, when the split is in the i:th direction, $i = 1, \ldots, d$. Assume that for each node m we have calculated $\text{emComp}(m) = -\text{freq}(m)/n + \lambda\,\text{mes}(m) + \alpha$ where $\text{freq}(m)$ is the number of observations in the set corresponding to m.

- Output of the algorithm is a binary tree. This binary tree is pruned from the original multitree. We represent this subtree by giving for each node pointers "left" and "right", which point to the left and right child of the node.

- An internal data structure of the algorithm is the decoration S which gives for every node of the tree the minimal excess mass complexity for the collection of sets localized to the rectangle associated with this node.

1. **set** maxdep = dJ_n (maxdep is the maximum level of the multitree)
2. **go** through levels starting from the next to the highest level: for dep=(maxdep-1) to 1
   (a) **go** through the nodes m at level dep
   (b) **if** m is leaf node **then** $S(m) = \text{emComp}(m)$
   (c) **else**
      i. let M=min$\{E_i, E_i^{left}, E_i^{right} : i = 1, \ldots, d\}$ where we denote

$$E_i = S(\text{left}_i(m)) + S(\text{right}_i(m))$$
$$E_i^{left} = S(\text{left}_i(m))$$
$$E_i^{right} = S(\text{right}_i(m))$$

      ii. **if** $\text{emComp}(m) < \text{M}$ **then** make m terminal node:
         A. $S(m) = \text{emComp}(m)$

B. left(m)=NIL, right(m)=NIL

iii. **else if** M=E_i **then** node m will be splitted to i:th direction and it has two children:

A. $S(m) = E_i$

B. left(m) = left$_i$(m), right(m) = right$_i$(m)

iv. **else if** M=E_i^{right} **then** node m will be splitted to i:th direction and it has only the right child:

A. $S(m) = E_i^{right}$

B. left(m)=NIL, right(m) = right$_i$(m)

v. **else if** M=E_i^{left} **then** node m will be splitted to i:th direction and it has only the left child:

A. $S(m) = E_i^{left}$

B. left(m)=left$_i$(m), right(m)=NIL

(d) **end if**

(e) **end go**

3. **end go**

In the case of the half block estimator one has to make more comparisons at each node to find out whether some of the skew splits will be better than the splits along the coordinate axis.

3 Rates of convergence of the estimators

We consider estimation of the support of a uniform density $f : \mathbf{R}^d \to \mathbf{R}$,

$$f = I_A/\mathrm{mes}(A)$$

where $d \geq 2$, $A \subset [0,1]^d$, $I_A(x) = 1$ when $x \in A$ and $I_A(x) = 0$ otherwise, and $\mathrm{mes}(A) = \int_A dx$.

We will denote by f the true underlying density and for $B \subset \mathbf{R}^d$ we will denote $g_B = I_B/\mathrm{mes}(B)$. We will denote by $S(g)$ the support of function g so that for example $f = g_{S(f)}$.

Boundary fragments have been a prototype model for studying set estimation. We will assume the boundary fragment model in analyzing the behaviour of the block estimator defined in Section 2.1. To analyze half block estimator defined in Section 2.2 we assume that the support of the density is star shaped. In the case of half block estimator we have assumed also that $d = 2$.

3.1 Block estimator

To prove a result for the rates of convergence of the block estimator, we define a scale of Hölder smoothness classes for smoothness index $0 < s \leq 1$ for the boundary fragment model.

Let $\mathcal{H}_s$ be the Hölder class of functions of smoothness $0 < s \leq 1$ and radius $L > 0$ on $[0,1]^{d-1}$. That is,

$$|h(t) - h(u)| \leq L\|t - u\|^s$$

for all $t, u \in [0,1]^{d-1}$ and $h \in \mathcal{H}_s$. We assume also that for $h \in \mathcal{H}_s$,

$$\gamma \leq h(t) \leq 1$$

for a fixed $\gamma > 0$. Denote by A_h the boundary fragment whose boundary is given by h,

$$A_h = \left\{x = (x_1, \ldots, x_d) \in [0,1]^d : 0 \leq x_d \leq h(x_1, \ldots, x_{d-1})\right\}.$$

A class of uniform densities whose support is a smooth boundary fragment is defined by

$$\mathcal{F}_s = \{g_{A_h} : \ \ h \in \mathcal{H}_s\} \tag{11}$$

where $g_{A_h} = I_{A_h}/\mathrm{mes}(A_h)$.

Consider the loss function

$$d_1(\hat{A}, S(f)) = \mathrm{mes}(\hat{A}\Delta S(f))$$

where $S(f)$ is the support of the true density f and Δ denotes symmetric difference: $A\Delta B = (A \setminus B) \cup (B \setminus A)$. Notation for the loss function reflects the fact that in terms of the boundary functions the loss is equal to the L_1 error. Let

$$r = \frac{s}{s + d - 1}$$

be the exponent of the minimax rate of convergence.

Theorem 2 *Let estimator $\hat{A}_n^e$ be defined in (7) based on iid observations $X_1, \ldots, X_n$. Choose the fineness parameter $a \geq 1$, parameter of the excess mass functional $0 < \lambda < 1$, and initial rectangle $R = [0,1]^d$. Consider class $\mathcal{F}_s$ defined in (11) where $0 < s \leq 1$. Let*

$$\alpha = C_\alpha \frac{\log_e n}{n} \tag{12}$$

where $0 < C_\alpha < \infty$. When C_α is sufficiently large, then

$$\limsup_{n\to\infty} (n/\log_e(n))^r \sup_{f \in \mathcal{F}_s} E_f d_1(\hat{A}_n^e, S(f)) < \infty.$$

A proof of Theorem 2 is given in Section 5. For the choice of α, see equation (28).

Remark 2. A proof that rate n^r is the minimax rate of convergence for Hölder boundary fragments is given in Korostelev and Tsybakov (1993*b*), Section 7.3.

Remark 3. Estimator $\hat{A}_n^e$ does not depend on the smoothness parameter s. Thus Theorem 2 shows that the estimator is adaptive in the sense that it achieves nearly minimax rates simultaneously over a scale of smoothness classes.

Remark 4. To achieve optimal balance between bias and variance we need blocks with width $n^{-1/(s+d-1)}$. On the other hand, to achieve minimax rate the finest partition should have blockwidth smaller than the minimax rate of convergence: $n^{-s/(s+d-1)}$. When $0 < s \leq 1$, then $n^{-1/(s+d-1)}$ satisfies

$$n^{-1/(d-1)} < n^{-1/(s+d-1)}$$

and minimax rate satisfies

$$n^{-1/d} \leq n^{-s/(s+d-1)}.$$

We want to achieve minimax rates simultaneously over scale $s \in (0, 1]$ and thus the finest partition should have blockwidth

$$\min\{n^{-1/(d-1)}, n^{-1/d}\} = n^{-1/(d-1)}. \tag{13}$$

That is why we choose in Theorem 2 the finest binwidth to be $n^{-a/(d-1)}$ where $a \geq 1$.

Remark 5. We have considered iid observations with n as the sample size. When considering regression function estimation with regular fixed design, then the corresponding step of the regular grid is $n^{-1/d}$.

When $0 < s \leq 1$, then by (13), one needs the binwidth of the finest partition to be smaller or equal to $n^{-1/(d-1)}$. Thus, since $n^{-1/(d-1)} < n^{-1/d}$, with fixed regular design we are not able to estimate the support with the rate $n^{-s/(s+d-1)}$. This was pointed out by Korostelev and Tsybakov (1993*b*).

3.2 Half block estimator

To prove a result for the rates of convergence of the half block estimator, we define a scale of Hölder smoothness classes with smoothness index $1 < s \leq 2$ for sets with star shaped boundaries.

Let $\mathcal{H}_s$ be the Hölder class of functions of smoothness $1 < s \leq 2$ and radius $L > 0$ on $[0, 2\pi)$. That is,

$$|h'(t) - h'(u)| \leq L|t - u|^{s-1}$$

for all $t, u \in [0, 2\pi)$ and $h \in \mathcal{H}_s$. We assume that for $h \in \mathcal{H}_s$,

$$\gamma \leq h(\phi) \leq 1/2$$

for $\gamma = 0.1$. Denote by $A_{h,\mu}$ the star shaped set centered at $\mu \in \mathbf{R}^2$ whose boundary is given by h:

$$A_{h,\mu} = \{x = \mu + (r\cos\phi, r\sin\phi) : 0 \leq r \leq h(\phi), \;\; \phi \in [0, 2\pi)\}.$$

A class of uniform densities whose support is a star shaped set is defined by

$$\tilde{\mathcal{F}}_s = \left\{ g_{A_{h,\mu}} : A_{h,\mu} \subset [0,1]^2, \;\; h \in \mathcal{H}_s, \;\; \mu \in \mathbf{R}^2 \right\} \tag{14}$$

where $g_A = I_A/\text{mes}(A)$.

Theorem 3 *Let estimator $\tilde{A}_n^e$ be defined in (10) based on iid observations $X_1, \ldots, X_n$. Choose the fineness parameters $a \geq 1$ and $b \geq 2$, parameter of the excess mass functional $0 < \lambda < 1$, and initial rectangle $R = [0,1]^2$. Consider class $\tilde{\mathcal{F}}_s$ defined in (14) where $1 < s \leq 2$. Let*

$$\alpha = C_\alpha \frac{\log_e n}{n} \tag{15}$$

where $0 < C_\alpha < \infty$. When C_α is sufficiently large, then

$$\limsup_{n\to\infty} (n/\log_e(n))^r \sup_{f \in \tilde{\mathcal{F}}_s} E_f d_1(\tilde{A}_n^e, S(f)) < \infty$$

where $r = s/(s + d - 1)$, $d = 2$.

A proof of Theorem 3 is given in Section 5. For the choice of α, see equation (33).

Remark 6. A proof that rate n^r is the minimax rate of convergence is given in Korostelev and Tsybakov (1993*b*), Section 7.3, for the boundary fragments. A consequence of this is that the same rate is minimax for the star shaped sets.

Remark 7. When $1 < s \leq 2$ (and $d = 2$), then blocksize $n^{-1/(s+d-1)}$ for the optimal bias-variance balancing satisfies

$$n^{-1/d} < n^{-1/(s+d-1)}.$$

Thus we choose in Theorem 3 the finest blocksize to be $n^{-a/d}$ where $a \geq 1$. For $1 < s \leq 2$ the minimax rate $n^{-s/(s+d-1)}$ satisfies

$$n^{-2/(d+1)} \leq n^{-s/(s+d-1)}.$$

We have that

$$\min\{n^{-1/d}, n^{-2/(d+1)}\} = n^{-2/(d+1)}.$$

That is why we choose in Theorem 3 the finest stepsize of skew splits to be $n^{-b/(d+1)}$ where $b \geq 2$.

Note the difference from the case $0 < s \leq 1$, where the minimum blocksize from the bias-variance balancing was smaller than the minimum blocksize from the rate of convergence. See equation (13).

Remark 8. Previously Barron et al. (1999) have proved a similar type of result. Instead of excess mass functional they propose to apply a different contrast function. Their estimator is of piecewise polynomial type and is not able to adapt to the case when the support of the density has a number of disconnected components.

4 Simulation examples

We give simulation examples for the block estimator. In simulation examples we consider examples which do not satisfy the conditions of Theorem 2. The definition of the estimator does not depend on these conditions and we may conjecture that the estimator is usable in a wide range of different situations.

The simulation examples are mixures of standard two dimensional Gaussian densities whose support is in fact the whole $\mathbf{R}^2$.

We chose the initial rectangle R for the simulation examples by first choosing R^- to be the smallest rectangle containing observations whose sides are parallel to the coordinate axes, and then taking $R = R^- + \eta$ where $\eta = 0.1$. The finest partition of R was chosen to contain 64^2 bins. The parameter λ of the excess mass criterion was chosen $\lambda = 0.1$ in all examples.

The first example is the standard Gaussian density in $\mathbf{R}^2$ centered at $(0, 0)$. We generated a sample of 100 observations from this density. Figure 1 shows three block estimates with excess mass criterion. In Figure 1 a) we

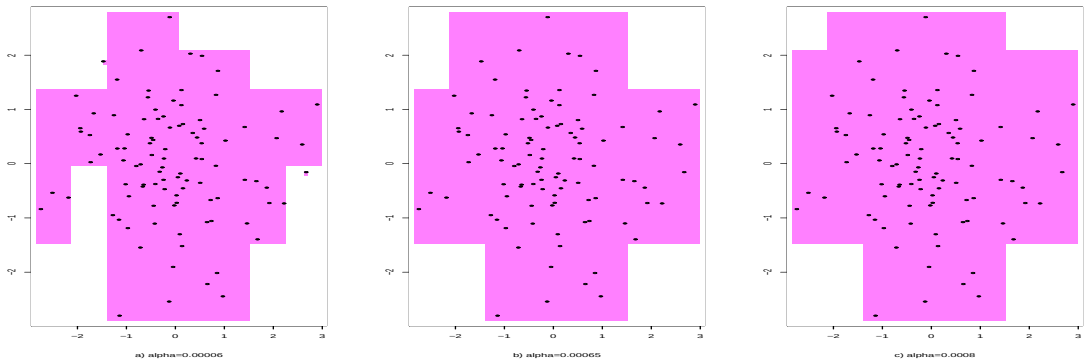

Figure 1: Estimates for a Gaussian density.

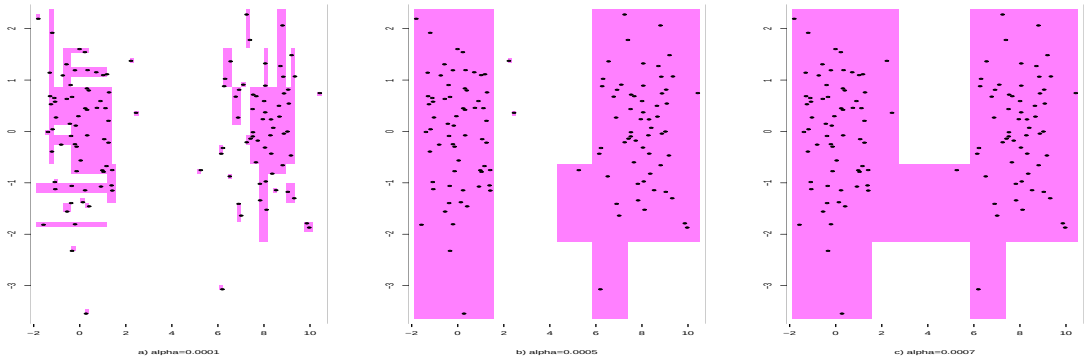

Figure 2: Estimates for a mixture of 2 Gaussian components.

took $\alpha = 0.0006$, in b) we took $\alpha = 0.00065$, and in c) we took $\alpha = 0.0008$. The choice of the smoothing parameter as $\alpha = 0.00065$ gives the best result.

The second example is an equal mixture of two standard Gaussians in $\mathbf{R}^2$. Means of the components of the mixture are $(0, 0)$ and $(8, 0)$. We generated a sample of 125 observations from this density. Figure 2 shows three block estimates with excess mass criterion. In Figure 2 a) we took $\alpha = 0.0001$, in b) we took $\alpha = 0.0005$, and in c) we took $\alpha = 0.0007$. The choice of the smoothing parameter as $\alpha = 0.0005$ gives the best result.

The third example is an equal mixture of three standard Gaussians in $\mathbf{R}^2$. Means of the components of the mixture lie in vertices of a triangle with sidelength $D = 8$, that is, the means are

$$(0, 0), \qquad (D, 0) = (8, 0), \qquad (D/2, D\sqrt{3}/2) \approx (4, 6.9).$$

We generated a sample of 150 observations from this density. Figure 3 shows three block estimates with excess mass criterion. In Figure 3 a) we took

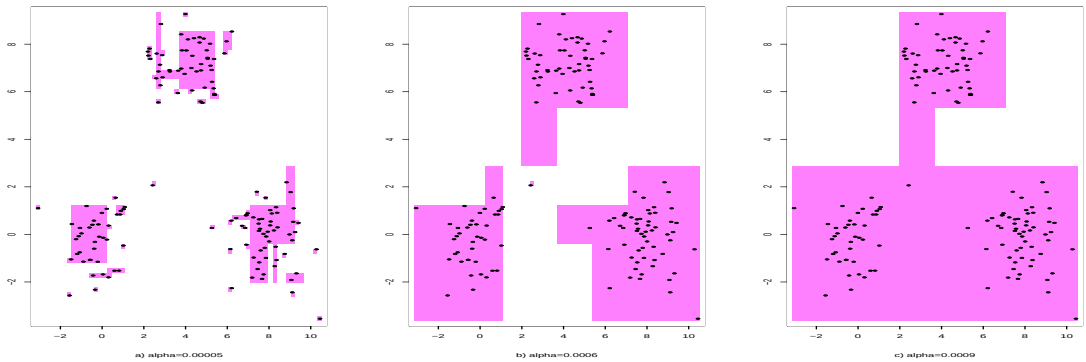

Figure 3: Estimates for a mixture of 3 Gaussian components.

$\alpha = 0.00005$, in b) we took $\alpha = 0.0006$, and in c) we took $\alpha = 0.0009$. The choice of the smoothing parameter as $\alpha = 0.0006$ gives the best result.

5 Proofs

We will give proofs for Theorems 2 and 3. The proofs are organized by giving in Section 5.1 oracle inequalities, giving in Section 5.2 bounds for the theoretical error-complexity, and finishing proofs in Section 5.3. The proof of oracle inequalities is almost the same for both block estimator and half block estimator but the approximation theoretic considerations in Section 5.2 are different for the two estimators.

5.1 Oracle inequality

For $\mathcal{P} \in \mathbb{P}_n$ or $\mathcal{P} \in \mathbb{P}_n^D$ and $W \in \mathcal{W}(\mathcal{P})$, let $K(\mathcal{P}, W, \alpha)$ be the theoretical error-complexity,

$$K(\mathcal{P}, W, \alpha) = d_1(A(\mathcal{P}, W), S(f)) + \alpha\, D(W) \tag{16}$$

where $S(f)$ is the support of the true density f. Let $A^0(f)$ and $A^{0,D}(f)$ be the best approximations to $S(f)$ in terms of theoretical error-complexity, when we search over sets used in the definition of the block estimator and half block estimator:

$$A^0(f) = A(\mathcal{P}^0, W^0) \tag{17}$$

where

$$(\mathcal{P}^0, W^0) = \text{argmin}_{\mathcal{P}\in\mathbb{P}_n, W\in\mathcal{W}(\mathcal{P})} K(\mathcal{P}, W, \alpha)$$

and

$$A^{0,D}(f) = A(\mathcal{P}^{0,D}, W^{0,D}) \tag{18}$$

where

$$(\mathcal{P}^{0,D}, W^{0,D}) = \text{argmin}_{\mathcal{P}\in\mathbb{P}_n^D, W\in\mathcal{W}(\mathcal{P})} K(\mathcal{P}, W, \alpha).$$

We have an upper bound for the theoretical error complexity of complexity penalized excess mass estimators $\hat{A}_n^e$ and $\tilde{A}_n^e$. This upper bound consists of theoretical error-complexity of best approximation with an additional stochastic term.

Lemma 4 *Let $\hat{A}_n^e$ and $\tilde{A}_n^e$ be defined in (7) and (10). Let $0 < \lambda < 1$ be the parameter of the excess mass functional. We have for $f \in \mathcal{F}_s$, when $0 < s \leq 1$,*

$$K(\hat{\mathcal{P}}_n^e, \hat{W}_n^e, \alpha) \leq C_b \left(K(\mathcal{P}^0, W^0, \alpha) + \nu_n(\hat{A}_n^e) - \nu_n(A^0(f)) \right)$$

and for $f \in \mathcal{F}_s$, when $1 < s \leq 2$,

$$K(\tilde{\mathcal{P}}_n^e, \tilde{W}_n^e, \alpha) \leq C_h \left(K(\mathcal{P}^{0,D}, W^{0,D}, \alpha) + \nu_n(\tilde{A}_n^e) - \nu_n(A^{0,D}(f)) \right)$$

for positive constants C_b, C_h, where

$$\nu_n(A) = \frac{1}{n} \sum_{i=1}^{n} I_A(X_i) - P_f(A) \tag{19}$$

for $A \subset \mathbf{R}^d$.

Proof. The proof is same for $\hat{A}_n^e$ and $\tilde{A}_n^e$. We will write the proof for $\hat{A}_n^e$. We have by the definition of $\hat{A}_n^e$ that

$$K_n(\hat{\mathcal{P}}_n^e, \hat{W}_n^e, \alpha) \leq K_n(\mathcal{P}^0, W^0, \alpha). \tag{20}$$

Also, excess mass functional may be written as

$$\gamma_n^e(A) = \lambda \,\text{mes}(A) - \nu_n(A) - P_f(A). \tag{21}$$

Denote by $S_\lambda(f)$ the level set of density f at level λ:

$$S_\lambda(f) = \{x \in \mathbf{R}^d : f(x) \geq \lambda\}.$$

Then, for $A \subset \mathbf{R}^d$,

$$\begin{aligned} &\lambda \,\text{mes}(A) - P_f(A) \\ &\quad = \lambda \,\text{mes}(S_\lambda(f)) - P_f(S_\lambda(f)) + \int |f(x) - \lambda| I_{A \Delta S_\lambda(f)}(x) dx. \end{aligned} \tag{22}$$

From (20) and (21) we have

$$\begin{aligned}&\lambda \operatorname{mes}(\hat{A}_n^e) - \nu_n(\hat{A}_n^e) - P_f(\hat{A}_n^e) + \alpha D(\hat{A}_n^e) \\ &\quad \le \lambda \operatorname{mes}(A^0(f)) - \nu_n(A^0(f)) - P_f(A^0(f)) + \alpha D(A^0(f)).\end{aligned}$$

Combining this with (22) implies

$$\begin{aligned}&\int |f(x) - \lambda| I_{\hat{A}_n^e \Delta S_\lambda(f)}(x) dx + \alpha D(\hat{W}_n^e) \qquad (23)\\ &\quad \le \int |f(x) - \lambda| I_{A^0(f) \Delta S_\lambda(f)}(x) dx + \alpha D(W^0) + \nu_n(\hat{A}_n^e) - \nu_n(A_0(f)).\end{aligned}$$

The minimal jump size of the densities over considered classes at the boundary of the support is 1. Thus

$$\int |f(x) - \lambda| I_{\hat{A}_n^e \Delta S_\lambda(f)}(x) dx \ge \min\{\lambda, 1-\lambda\}\, d_1\left(\hat{A}_n^e, S_\lambda(f)\right).$$

All densities in considered classes are bounded by $M = \gamma^{1-d}$. Thus

$$\int |f(x) - \lambda| I_{A^0(f) \Delta S_\lambda(f)}(x) dx \le \max\{\lambda, M\}\, d_1\left(A^0(f), S_\lambda(f)\right).$$

These two inequalities and (23) imply the lemma, because for $0 < \lambda < 1$, the level set $S_\lambda(f)$ is equal to the support of the density: $S_\lambda(f) = S(f)$. □

We will need an upper bound for the cardinality of the class of all sets in all partitions.

Lemma 5 *Set of partitions $\mathbb{P}_n$ for the block estimator is defined in Definition 1. We have that*

$$\#\left(\bigcup_{\mathcal{P}\in\mathbb{P}_n} \mathcal{P}\right) \le N \stackrel{def}{=} \frac{(2d)^{dJ_n+1} - 1}{2d-1} = O\left(n^{ad\log_2(2d)/(d-1)}\right).$$

Set of partitions $\mathbb{P}_n^D$ for the half block estimator is defined in Definition 2. We have that

$$\#\left(\bigcup_{\mathcal{P}\in\mathbb{P}_n^D} \mathcal{P}\right) \le \tilde{N} \stackrel{def}{=} 4^2 \cdot 2^{2L_n+1} \frac{(d/2)^{d\tilde{J}_n+1} - 1}{(d/2)-1} = O\left(n^{2b/(d+1)} n^{a\log_2(d/2)}\right).$$

Proof. For the case of block estimator cardinality is bounded by the number of nodes in a tree with dJ_n levels, with one root node, and $2d$ children for every node. To the root node corresponds initial rectangle R and every

rectangle may be splitted to two children in d directions, which results to $2d$ children. Thus we have bound

$$\sum_{i=0}^{dJ_n}(2d)^i = \frac{(2d)^{dJ_n+1}-1}{2d-1} = O\left(2^{\log_2(2d)dJ_n}\right).$$

For the case of the half block estimator each rectangle may be splitted with a skew split whose endpoints lie in a grid with cardinality $4\cdot 2^{-i}/\delta$, where $4\cdot 2^{-i}$ is the length of boundaries of rectangles in i:th level, and $\delta = 2^{-L_n}$ is the stepsize of the grid. Thus we have at most $2(4\cdot 2^{-i}/\delta)^2$ children resulting from a skew split. Thus the total number of sets is bounded by

$$\sum_{i=0}^{d\tilde{J}_n} 2\cdot 4^2(2^{-i}/\delta)^2(2d)^i = 4^2\cdot 2^{2L_n+1}\,\frac{(d/2)^{d\tilde{J}_n+1}-1}{(d/2)-1}.$$

□

Now we may prove that the risk of estimators may be bounded by the theoretical error-complexity. We will start with the block estimator.

Lemma 6 *Consider estimator $\hat{A}_n^e$ defined in (7). Let α be defined in (28). We have that*

$$E_f d_1(\hat{A}_n^e, S(f)) \leq C\left[K(\mathcal{P}^0, W^0, \alpha) + n^{-1}\right]$$

for a positive constant C.

Proof. We have

$$E_f d_1(\hat{A}_n^e, S(f)) \leq C_b K(\mathcal{P}^0, W^0, \alpha) + E_f V$$

where

$$V = \max\left\{d_1(\hat{A}_n^e, S(f)) - C_b K(\mathcal{P}^0, W^0, \alpha), 0\right\}$$

and C_b is from Lemma 4. It remains to prove that

$$E_f V = O(n^{-1}). \tag{24}$$

Denote

$$B_n = \left(\sup_{\mathcal{P}\in\mathbb{P}_n}\sup_{W\in\mathcal{W}(\mathcal{P})} w(A(\mathcal{P},W))^{-1}\left|\nu_n(A(\mathcal{P},W)) - \nu_n(A^0(f))\right| \leq \xi\right)$$

where $\xi = \sqrt{8}$, ν_n is defined in (19), and we define with an abuse of notation

$$w(A) = w(A(\mathcal{P},W)) = n^{-1}\left(x + LD(W)\right)$$

with $x > 0$ and

$$L = \log_e(N) \tag{25}$$

where N is defined in Lemma 5. First we prove that on B_n, $V \leq \xi C_b n^{-1} x$, that is

$$(V > \xi C_b n^{-1} x) \subset B_n^c. \tag{26}$$

Secondly we prove that

$$P(B_n^c) \leq C \exp\{-x\}. \tag{27}$$

We have that

$$EV = \xi C_b n^{-1} \int_0^\infty P(V > \xi C_b n^{-1} x)\, dx$$

and thus (24) follows from (26) and (27).

Proof of (26). By the definition of B_n we have that on B_n, $\nu_n(\hat{A}_n^e) - \nu_n(A^0(f)) \leq \xi w(\hat{A}_n^e)$. Thus, by Lemma 4, on B_n,

$$\begin{aligned} K(\hat{\mathcal{P}}_n^e, \hat{W}_n^e, \alpha) &\leq C_b \left[K(\mathcal{P}^0, W^0, \alpha) + \xi w(\hat{A}_n^e) \right] \\ &= C_b \left[K(\mathcal{P}^0, W^0, \alpha) + \xi n^{-1}(x + LD(\hat{W}_n^e)) \right]. \end{aligned}$$

We choose

$$\alpha = \xi C_b n^{-1} L \tag{28}$$

where L is defined in (25), $\xi = \sqrt{8}$, and C_b comes from Lemma 4. Thus, on B_n,

$$d_1(\hat{A}_n^e, S(f)) \leq C_b \left[K(\mathcal{P}^0, W^0, \alpha) + \xi n^{-1} x \right].$$

We have proved (26).

Proof of (27). Define with an abuse of notation

$$\eta_A = \frac{1}{w(A)} \left(I_A - I_{A^0(f)} \right)$$

where $A = A(\mathcal{P}, W)$, $A^0(f) = A(\mathcal{P}^0, W^0)$, $\mathcal{P} \in \mathbb{P}_n$ and $W \in \mathcal{W}(\mathcal{P})$. We have that $-w(A)^{-1} \leq \eta_A(X_i) \leq w(A)^{-1}$. Thus by Hoeffding's inequality, see for example Pollard (1984), page 191,

$$P_f \left(\left| \frac{1}{n} \sum_{i=1}^n \eta_A(X_i) - E_f \eta_A(X_1) \right| > \xi \right) \leq \exp\left\{ - \frac{n\xi^2 w(A)}{8} \right\}. \tag{29}$$

Since $\xi = \sqrt{8}$, we have

$$\frac{n\xi^2 w(A)}{8} = x + LD(W). \tag{30}$$

Now, for $A = A(\mathcal{P}, W)$,

$$\nu_n(A) - \nu_n(A^0(f)) = w(A)\left(\frac{1}{n}\sum_{i=1}^{n}\eta_A(X_i) - E_f\eta_A(X_1)\right).$$

Then, by (29) and (30),

$$P(B_n^c) \leq \sum_{\mathcal{P}\in\mathbb{P}_n}\sum_{W\in\mathcal{W}(\mathcal{P})}\exp\left\{-[x + LD(W)]\right\}. \tag{31}$$

Denote

$$\Psi(k) = \{(\mathcal{P}, W) : \mathcal{P} \in \mathbb{P}_n,\, W \in \mathcal{W}(\mathcal{P}),\, D(W) = k\}$$

so that $\#\Psi(k)$ is equal to the number of ways we may choose k sets from the set of all sets in all partitions. Now, defining N as in Lemma 5, by Stirling's formula,

$$\#\,\Psi(k) \leq \left(\begin{array}{c} N \\ k \end{array}\right) \leq \frac{N^k}{k!} \leq \left(\frac{eN}{k}\right)^k.$$

Thus, continuing from (31),

$$\begin{aligned} P(B_n^c) &\leq \sum_{k=1}^{\infty}\sum_{(\mathcal{P},W)\in\Psi(k)}\exp\left\{-(x + Lk)\right\} \\ &\leq \sum_{k=1}^{\infty}\left(\frac{eN}{k}\right)^k\exp\left\{-(x + Lk)\right\} \\ &\leq C\exp\left\{-x\right\}, \end{aligned} \tag{32}$$

by the choice of L in (25). We have proved (27) and thus the lemma. □

We may prove a similar lemma for the half block estimator.

Lemma 7 *Consider estimator $\tilde{A}_n^e$ defined in (10). Let α be defined in (33). We have that*

$$E_f d_1(\tilde{A}_n^e, S(f)) \leq C\left[K(\mathcal{P}^{0,D}, W^{0,D}, \alpha) + n^{-1}\right]$$

for a positive constant C.

Proof. The proof is similar to the proof of Lemma 6. The difference is that we set

$$\alpha = \xi C_h n^{-1} L \tag{33}$$

where $\xi = \sqrt{8}$, C_h comes form Lemma 4,

$$L = \log_e(\tilde{N}),$$

and $\tilde{N}$ is defined in Lemma 5. □

5.2 A bound for the theoretical error-complexity

So far the proofs have been similar both for the boundary fragment model and for the star shaped sets. In this section we give a separate treatment for the two cases.

Let $A^0(f)$ be defined in (17). We will give a bound for the error-complexity of $A^0(f)$.

Lemma 8 *Let $\mathcal{F}_s$ be defined in (11) for $0 < s \leq 1$ and let K be defined in (16). We have that*

$$\sup_{f \in \mathcal{F}_s} K(\mathcal{P}^0, W^0, \alpha) \leq C \left(\frac{\log_e n}{n} \right)^{s/(s+d-1)}$$

for a positive constant C, when α is defined in (12).

Proof. Let $f \in \mathcal{F}_s$ and let $h : [0,1]^{d-1} \to [0,1]$ be the function defining the boundary of the support of f. That is, $f = g_{A_h} = I_{A_h}/\text{mes}(A_h)$. Let us choose N_0 so that

$$2^{N_0} \sim (n/\log_e(n))^{1/(s+d-1)}.$$

Let $\mathcal{Q}$ be a partition of $[0,1]^{d-1}$ to rectangles whose sidelength is 2^{-N_0} (and volume is $2^{-(d-1)N_0}$). We may construct a function $h_0 : [0,1]^{d-1} \to \mathbf{R}$, which is piecewise constant on partition $\mathcal{Q}$, that is,

$$h_0 = \sum_{P \in \mathcal{Q}} a_P I_P$$

where $a_P \in [\gamma, 1]$, and with the property

$$\int_{[0,1]^{d-1}} |h - h_0| = O\left(2^{-sN_0}\right).$$

This construction may be done with a piecewise constant interpolation of h. On the other hand $S(f) = A_h$ and thus

$$d_1(S(f), A_{h_0}) = d_1(A_h, A_{h_0}) = \int_{[0,1]^{d-1}} |h - h_0| .$$

Now choose N_1 so that $2^{N_1} \sim n^{1/d}$. Make a grid $0 = q_1 < \cdots < q_M = 1$ where the distance between gridpoints is 2^{-N_1}. Define function $\widetilde{h}_0$ which approximates h_0:

$$\widetilde{h}_0 = \sum_{P \in \mathcal{Q}} \widetilde{a}_P I_P$$

where $\widetilde{a}_P$ are the gridpoints closest to a_P:

$$|\widetilde{a}_P - a_P| = \min \{|b - a_P| : b \in \{q_1, \ldots, q_M\}\} .$$

We have that

$$\int_{[0,1]^{d-1}} \left| h_0 - \widetilde{h}_0 \right| = O\left(2^{-N_1}\right) .$$

Then

$$d_1(S(f), A_{\widetilde{h}_0}) = O\left(2^{-sN_0} + 2^{-N_1}\right) .$$

We have that $2^{-N_1} = O(2^{-sN_0})$ for all $0 < s \leq 1$. We have proved that

$$d_1(S(f), A_{\widetilde{h}_0}) = O\left(2^{-sN_0}\right) . \tag{34}$$

Now, because fineness parameter $a \geq 1$, then $A_{\widetilde{h}_0} \in \mathbb{A}_n$ where $\mathbb{A}_n$ is defined in (2). See the discussion leading to equation (13). Let $D(A_{\widetilde{h}_0})$ be the complexity of set $A_{\widetilde{h}_0}$ (with an abuse of notation). By construction, $D(A_{\widetilde{h}_0}) = 2^{(d-1)N_0}$. Thus

$$\alpha D(A_{\widetilde{h}_0}) = O\left(\left(\frac{\log_e n}{n}\right)^{s/(s+d-1)}\right) . \tag{35}$$

Equations (34) and (35) imply the lemma since the bounds are uniform with respect to $f \in \mathcal{F}_s$. □

Consider secondly the case of star shaped sets.

Lemma 9 *Let $\mathcal{F}_s$ be defined in (14) for $1 < s \leq 2$, and let $d = 2$. We have that*

$$\sup_{f \in \mathcal{F}_s} K(\mathcal{P}^{0,D}, W^{0,D}, \alpha) \leq C \left(\frac{\log_e n}{n}\right)^{s/(s+d-1)}$$

for a positive constant C, when α is defined in (15).

Proof. We may apply Lemma 8.5 (Edgel approximation), Lemma 8.6 (Edgelet approximation), and Lemma 8.7 (Counting ancestors) from Donoho (1999) to prove the required bound. Indeed, by Remark 1 the set of partitions $\mathcal{P}_n^D$ is larger than the corresponding set of Donoho (1999). Thus we have at least the same approximation properties. □

5.3 Finishing the proofs

Proof of Theorem 2 follows from Lemma 6 and Lemma 8. Proof of Theorem 3 follows from Lemma 7 and Lemma 9.

Acknowledgements

I wish to thank two referees for the comments that helped to improve the presentation.

References

Baíllo, A., Cuevas, A. and Justel, A. (2000), 'Set estimation and nonparametric detection', *Canadian J. Statist.* **28**, 765–782.

Barron, A., Birgé, L. and Massart, P. (1999), 'Risk bounds for model selection via penalization', *Probab. Theory Relat. Fields* **113**, 301–413.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. J. (1984), *Classification and Regression Trees*, Chapman & Hall, New York.

Chevalier, J. (1976), 'Estimation du support et du contenu du support d'une loi de probabilité', *Ann. Inst. H. Poincaré Sec. B* **12**, 339–364.

Cuevas, A. and Fraiman, R. (1997), 'A plug-in approach to support estimation', *Ann. Statist.* **25**, 2300–2312.

Deprins, D., Simar, L. and Tulkens, H. (1984), Measuring labor efficiency in post offices, *in* M. Marchand, P. Pestieau and H. Tulkens, eds, 'The performance of public enterprises: Concepts and Measurements', Amsterdam, North-Holland, pp. 243–267.

Derman, C. and Ross, S. M. (1997), *Statistical Aspects of Quality Control*, Academic Press, San Diego.

Devroye, L. and Wise, G. L. (1980), 'Detection of abnormal behavior via nonparametric estimation of the support', *SIAM J. Appl. Math.* **38**, 480–488.

Donoho, D. L. (1997), 'Cart and best-ortho-basis: A connection.', *Ann. Statist.* **25**, 1870–1911.

Donoho, D. L. (1999), 'Wedgelets: Nearly minimax estimation of edges', *Ann. Statist.* **27**, 859–897.

Donoho, D. L. and Johnstone, I. M. (1994), 'Ideal denoising in an orthonormal basis chosen from a library of bases', *C. R. Acad. Sci. Paris Sér. I Math.* **319**, 1317–1322.

Geffroy, M. (1964), 'Sur un problème d'estimation géométrique', *Publ. Inst. Statist. Univ. Paris* **13**, 191–120.

Gijbels, I., Mammen, E., Park, B. and Simar, L. (1999), 'On estimation of monotone and concave frontier functions', *J. Amer. Statist. Assoc.* **94**, 220–228.

Hall, P., Nussbaum, M. and Stern, S. E. (1997), 'On the estimation of a support curve of indeterminate sharpness', *J. Multivariate Anal.* **62**, 204–232.

Härdle, W., Park, B. U. and Tsybakov, A. B. (1995), 'Estimation of non-sharp support boundaries', *J. Multivariate Anal.* **55**, 205–218.

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

Klemelä, J. (2001), 'Multivariate histograms with data-dependent partitions'. Submitted for publication.

Korostelev, A. P., Simar, L. and Tsybakov, A. B. (1995), 'Efficient estimation of monotone boundaries', *Ann. Statist.* **23**, 476–489.

Korostelev, A. P. and Tsybakov, A. B. (1993*a*), 'Estimation of the density support and its functionals', *Probl. Inf. Transm.* **29**, 1–15.

Korostelev, A. P. and Tsybakov, A. B. (1993*b*), *Minimax Theory of Image Reconstruction, Lecture Notes in Statistics, 82*, Springer.

Korostelev, A. P. and Tsybakov, A. B. (1994), 'Asymptotic efficiency in estimation of a convex set', *Probl. Inf. Transm.* **30**, 317–327.

Mammen, E. and Tsybakov, A. B. (1995), 'Asymptotical minimax recovery of sets with smooth boundaries', *Ann. Statist.* **23**, 502–524.

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

Pollard, D. (1984), *Convergence of Stochastic Processes*, Springer.

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

Rényi, A. and Sulanke, R. (1963), 'Über die konvexe Hülle von n zufällig gewählten Punkten', *Z. Wahrsch. Verw. Gebiete* **2**, 75–84.

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

Ripley, B. D. and Rasson, J. P. (1977), 'Finding the edge of a Poisson forest', *J. Appl. Probab.* **14**, 483–491.

Schneider, R. (1988), 'Random approximation of convex sets', *Journal of Microscopy* **151**(Pt. 3), 211–227.

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

Walther, G. (1997), 'Granulometric smoothing', *Ann. Statist.* **25**, 2273–2299.

Jussi Klemelä
Institut für Angewandte Mathematik
Universität Heidelberg
Im Neuenheimer Feld 294
69120 Heidelberg, Germany
EMAIL: klemela@statlab.uni-heidelberg.de