Statistica Sinica **19** (2009), 159-176

MULTIVARIATE HISTOGRAMS WITH DATA-DEPENDENT PARTITIONS

Jussi Klemelä

University of Oulu

Abstract: We consider estimation of multivariate densities with histograms which are based on data-dependent partitions. We find data-dependent partitions by minimizing a complexity-penalized error criterion. The estimator may also be characterized as a series estimator whose basis is chosen empirically. We show that the estimator achieves minimax rates of convergence up to a logarithmic factor over a scale of smoothness classes containing functions with anisotropic and spatially varying smoothness. The method may also be viewed as based on the presmoothing of data. We show how the optimal amount of presmoothing depends on the spatial inhomogeneity of the density.

Key words and phrases: Adaptive estimation, dyadic CART, multivariate density estimation, presmoothing, tree structured estimators.

1. Introduction

We consider density estimation based on i.i.d. multivariate random vectors taking values in $\mathbf{R}^d$. We estimate densities with histograms, which we define to be rectangularwise constant estimates, and the value of the estimate in each rectangle is taken to be the empirical probability divided by the volume of the rectangle. The main problem is how to choose the partition defining the histogram in an optimal way.

Histograms with equispaced bins are not able to adapt to spatially varying smoothness. This problem appears already in the one-dimensional case. Furthermore, in the multivariate case the density to be estimated may have anisotropic smoothness: the density function may vary more in one direction than in the other directions. We should choose bins to be thinner in the direction where the density varies more.

When we choose the partition in a flexible way, then we are not so vulnerable to the curse of dimensionality. Indeed, in high-dimensional cases accurate estimation may be possible if the "effective dimension" of the density is small. The effective dimension could mean, for example, the number of variables with respect which the density has variability. A density which is almost constant on its support with respect to most variables would have a low effective dimension (this

would be an extreme case of anisotropic smoothness). These types of densities could be estimated well if we had a method of choosing the partition of the histogram economically: we should choose a partition which does not contain splits in those directions where there is no variation (the partition would only delineate the support with respect to those directions where there is no variation).

We define the histogram estimator as a minimizer of a complexity-penalized error criterion. As the error criterion we take the empirical risk with the L_2 contrast function, and the complexity of the histogram is defined to be the number of sets in the partition. The set of candidate partitions is fixed, and defined by the set of dyadic splitting sequences. Thus the estimator is similar to the dyadic regressograms considered in Donoho (1997).

An important property of the estimator is that we can define it in two ways: (1) as a histogram estimator, and (2) as a series estimator associated to a basis of multivariate Haar functions. The characterization of the estimator as a series estimator makes it possible to analyze asymptotic properties of the estimator and the definition of the estimator as a histogram makes it possible to find a fast algorithm for evaluating the estimates. In the histogram characterization the partition is chosen empirically, and in the series estimator characterization the basis is chosen empirically. Instead of thresholding the empirical coefficients in a fixed basis, the method chooses empirically a basis where the thresholding is performed.

We show that the estimator achieves minimax rates up to a logarithmic factor over a scale of anisotropic smoothness classes, for the L_2 loss. We consider histograms with unequal binwidths in every direction and thus we nearly achieve the minimax rates over smoothness classes containing functions with considerable spatially varying smoothness. To apply the estimator we have to choose a bound for the maximal fineness of the partitions we consider. We may increase the flexibility of the estimator by choosing the maximal allowed resolution to be fine. On the other hand this will increase the computational complexity of the estimator. We shall show how the bound for the maximal fineness depends on the spatial inhomogeneity of the density. We show also how the computational complexity depends on this bound for the maximal fineness. The method we propose may be seen as based on presmoothing the data since the estimator uses only the frequencies on the partition defined by the finest resolution level.

We give some references to the previous literature on histograms with irregular data-dependent partitions, and on other spatially flexible estimation methods.

1. *Multivariate regressograms.* Breiman, Friedman, Olshen and Stone (1984) introduced CART (Classification and Regression Trees) as a method for estimating classification and regression functions with piecewise constant estimates. They constructed data-dependent partitions by a two-step procedure. First

they found a set of candidate partitions by minimizing an empirical error criterion in a myopic fashion, and then they chose the final partition by minimizing an error-complexity criterion among the set of candidate partitions.
Donoho (1997) considered 2-dimensional Gaussian regression on a fixed and regularly spaced design. He considers an estimator which is defined as a minimizer of an error-complexity criterion. Unlike in CART, where the set of candidate partitions is constructed empirically, he considered candidate partitions which are obtained by sequential dyadic splitting of the rectangle containing the support of the regression function.

2. *Multivariate histograms.* Density estimation with CART-type methods was considered by Shang (1994), Sutton (1994), Ooi (2002). Hüsemann and Terrell (1991) consider the problem of optimal fixed and variable cell dimensions in bivariate histograms. Lugosi and Nobel (1996) present L_1-consistency results on density estimators based on data dependent partitions. Barron, Birgé and Massart (1999) constructed a multivariate histogram which achieves asymptotic minimax rates over anisotropic Hölder classes for the L_2 loss. Their histograms had different numbers of bins in different directions, but in a single direction bins were equispaced. A modified Akaike criterion for histogram estimation with irregular splits was studied in the multivariate case by Castellan (2000) who gives oracle inequalities for Kullback-Leibler and Hellinger loss.
3. *Other methods.* Multivariate density estimation based on wavelet expansions has been considered in Tribouley (1995). Neumann (2000) constructed an estimator based on wavelet expansions which achieves minimax rates up to a logarithmic factor over a large scale of anisotropic Besov classes in the Gaussian white noise model. Kerkyacharian, Lepski and Picard (2001) consider a kernel-based adaptation scheme to cope with anisotropic smoothness.

In Section 2 we define the estimator in two ways as a histogram, and as a series estimator. We present an algorithm for the computation of an estimate. In Section 3 we give the rates of convergence of the estimator. Section A. illustrates the properties of the estimator with simulation examples. Some of the proofs are in the Appendices.

2. Estimators

2.1. Dyadic histogram

Let $X^1, \dots, X^n \in \mathbf{R}^d$ be i.i.d. random vectors whose density function we want to estimate. A histogram with partition $\mathcal{P}$ is defined by

$$\hat{f}(x, \mathcal{P}) = \sum_{R \in \mathcal{P}} \frac{n_R}{n \operatorname{vol}(R)} I_R(x), \qquad x \in \mathbf{R}^d, \tag{2.1}$$

where $n_R = \#\{X^i \in R\}$ are the frequencies for the sets of the partition. We define a collection of dyadic partition generating trees. The optimal partition will be searched from the collection of partitions generated by these partition generating trees.

Definition 1. A *collection of dyadic partition generating trees* $\mathbb{T}(R_0, J)$, associated with a rectangle $R_0 \subset \mathbf{R}^d$, and with a bound for split numbers $J = (J_1, \ldots, J_d)$, $J_l \in \{0, 1, \ldots\}$, consists of binary trees where each node is associated with a rectangle, and each non-leaf node is associated with a splitting direction in $\{1, \ldots, d\}$.

1. The root node is associated with R_0.
2. Let a non-leaf node be associated with rectangle $R = \Pi_{m=1}^d [c_m, d_m]$ and direction $l \in \{1, \ldots, d\}$. The split point is $s = (d_l - c_l)/2$. Write

$$R_{l,s}^{(0)}(R) = \{x \in R : x_l < s\}, \qquad R_{l,s}^{(1)}(R) = \{x \in R : x_l \geq s\}.$$

   The left child of the node is associated with $R_{l,s}^{(0)}(R)$ and the right child is associated with $R_{l,s}^{(1)}(R)$.
3. In direction l at most J_l splits will be made, $l = 1, \ldots, d$.

We make some remarks concerning the definition.

- In fact, a set of dyadic partition generating trees is completely determined by the initial rectangle and by the splitting directions; since the splits are always made at the midpoints of the sides of the rectangles, the association of the nodes with rectangles is redundant.
- The simplest dyadic partition generating tree is the tree which consists only of the root node, and this tree is the single member of $\mathbb{T}(R_0, 0)$.
- The bound J for the split numbers implies a bound for the depth of the tree: the depth is at most $|J| = \sum_{i=1}^d J_i$. (We define the depth of a tree to be equal to the largest depth among the depths of its nodes, and we stipulate that the depth of the root node is 0, the depth of the children of the root is 1, and so on.)
- Note that a tree generating a dyadic partition may be an unbalanced tree: some terminal nodes may have depth equal to $|J|$ but the depth of some other terminal nodes may be less than $|J|$.

Each tree in the set $\mathbb{T}(R_0, J)$ generates a partition: the partition is the collection of the rectangles associated with the leaf nodes of the tree. This is the content of the definition below.

Definition 2. (*Collection of dyadic partitions.*) The dyadic partition associated to tree $\mathcal{T} \in \mathbb{T}(R_0, J)$, where $\mathbb{T}(R_0, J)$ is defined in Definition 1, is

$$\mathcal{P}(\mathcal{T}) = \{R(t) : t \in \mathit{Ter}(\mathcal{T})\}, \tag{2.2}$$

where $\mathit{Ter}(\mathcal{T})$ is the set of terminal nodes of $\mathcal{T}$, and $R(t)$ is the rectangle associated to node t. *The collection of dyadic partitions* $\mathbb{P} = \mathbb{P}(R_0, J)$, with base rectangle R_0 and with depth bound J, is denoted by

$$\mathbb{P}(R_0, J) = \{\mathcal{P}(\mathcal{T}) : \mathcal{T} \in \mathbb{T}(R_0, J)\}. \tag{2.3}$$

Complexity-penalized error criterion. Define the empirical risk of a density estimator $\hat{f} : \mathbf{R}^d \to \mathbf{R}$ by

$$\gamma_n\left(\hat{f}\right) = \frac{1}{n}\sum_{i=1}^{n} \gamma\left(\hat{f}, X^i\right), \tag{2.4}$$

where $\gamma(g, x)$ is the L_2 contrast function,

$$\gamma(g, x) = -2g(x) + \|g\|_2^2, \qquad g : \mathbf{R}^d \to \mathbf{R}, \;\; x \in \mathbf{R}^d.$$

Minimization of $\|\hat{f} - f\|_2^2$ over estimators $\hat{f}$ is equivalent to the minimization of $\|\hat{f} - f\|_2^2 - \|f\|_2^2$, and minimization of $\gamma_n(\hat{f})$ amounts to the minimization of $\|\hat{f} - f\|_2^2 - \|f\|_2^2$, up to the approximation $\int_{\mathbf{R}^d} \hat{f} f \approx n^{-1}\sum_{i=1}^{n} \hat{f}(X^i)$. Indeed,

$$\begin{aligned} \|\hat{f} - f\|_2^2 - \|f\|_2^2 &= -2\int_{\mathbf{R}^d} f\hat{f} + \|\hat{f}\|_2^2 \\ &\approx -2n^{-1}\sum_{i=1}^{n} \hat{f}(X^i) + \|\hat{f}\|_2^2 \\ &= \gamma_n\left(\hat{f}\right). \end{aligned} \tag{2.5}$$

A histogram $\hat{f}(\cdot, \mathcal{P})$ is uniquely defined through its partition $\mathcal{P}$, and we use the notation

$$\mathit{ERR}_n(\mathcal{P}) = \gamma_n\left(\hat{f}(\cdot, \mathcal{P})\right). \tag{2.6}$$

We have that

$$\mathit{ERR}_n(\mathcal{P}) = -\sum_{R \in \mathcal{P}} \frac{n_R^2}{n^2 \mathrm{vol}(R)} = -\left\|\hat{f}^2(\cdot, \mathcal{P})\right\|_2^2. \tag{2.7}$$

The complexity of a histogram is taken to be the number of sets in the partition of the histogram. Let $0 \le \alpha < \infty$ and define the complexity-penalized error criterion as

$$\mathit{COPERR}_n(\mathcal{P}, \alpha) = \mathit{ERR}_n(\mathcal{P}) + \alpha \cdot \#\mathcal{P}. \tag{2.8}$$

The dyadic histogram is defined as a minimizer of the complexity-penalized empirical risk, when we minimize the complexity-penalized empirical risk over the set of dyadic partitions.

Definition 3. *(Dyadic histogram.)* Define the partition corresponding to parameter α as

$$\hat{\mathcal{P}}_\alpha = \text{argmin}_{\mathcal{P}\in\mathbb{P}(R_0,J)} COPERR_n(\mathcal{P},\alpha), \tag{2.9}$$

where $\mathbb{P}(R_0, J)$ is defined in (2.3). The *dyadic histogram* is

$$\hat{f}_{n,\alpha} = \hat{f}(\,\cdot\,,\hat{\mathcal{P}}_\alpha). \tag{2.10}$$

where $\hat{f}(\,\cdot\,,\mathcal{P})$ is defined in (2.1).

Remark 1. The estimator depends, besides the smoothing parameter α, on the maximal directionwise split numbers J, and on the initial rectangle R_0. Theorem 1 gives conditions for the choice of these parameters. In particular, α and J will depend on the sample size n. In Theorem 1 we take $R_0 = [0,1]^d$, but in practice one would estimate R_0. A reasonable choice is to define R_0 as the smallest rectangle, containing the observations, whose sides are parallel to the coordinate axis.

2.2. Series estimator

We define a series estimator by using a basis of Haar wavelets. We prove that the series estimator is in fact identical with a dyadic histogram. A dyadic histogram is a useful representation of the estimator when we want to find algorithms for the calculation of the estimates. The representation of the estimator as a series estimator is useful when we want to find its asymptotic properties.

Let

$$\tilde{f}(x,W,\Theta,\mathcal{B}) = I_{[0,1]^d}(x) + \sum_{\phi\in\mathcal{B}} w_\phi\theta_\phi\phi(x), \qquad x\in\mathbf{R}^d, \tag{2.11}$$

where $\mathcal{B}$ is an orthonormal system of functions in $L_2([0,1]^d)$, $W = (w_\phi)_{\phi\in\mathcal{B}} \in \{0,1\}^{\mathcal{B}}$, $\Theta = (\theta_\phi)_{\phi\in\mathcal{B}} \in \mathbf{R}^{\mathcal{B}}$. Vector W chooses a subset of $\mathcal{B}$ and vector Θ gives the coefficients of the expansion. We assume that $\int_{[0,1]^d}\phi = 0$ for all $\phi\in\mathcal{B}$ and, since we estimate densities, we may include the indicator $I_{[0,1]^d}$ in all expansions.

Multivariate Haar wavelets. The univariate Haar scaling function is $\eta^{(0)} = I_{[0,1]}$, and the univariate Haar wavelet is $\eta^{(1)} = I_{[1/2,1)} - I_{[0,1/2)}$. Let

$$\eta^{(\iota)}_{j_m,k_m}(t) = 2^{\frac{j_m}{2}}\eta^{(\iota)}(2^{j_m}t - k_m), \qquad t\in[0,1],$$

with $\iota\in\{0,1\}$, $j_m\in\{0,1,\ldots\}$, and $k_m\in\{0,\ldots,2^{j_m}-1\}$. Let

$$\phi^{(l)}_{j,k}(x) = \eta^{(1)}_{j_l,k_l}(x_l)\Pi^d_{m=1,m\neq l}\eta^{(0)}_{j_m,k_m}(x_m), \qquad x=(x_1,\ldots,x_d)\in\mathbf{R}^d, \tag{2.12}$$

where $l \in \{1, \ldots, d\}$, $j = (j_1, \ldots, j_d) \in \{0, 1, \ldots\}^d$, $k = (k_1, \ldots, k_d) \in K_j$, and

$$K_j = \{k = (k_1, \ldots, k_d) : k_l = 0, \ldots, 2^{j_l} - 1, l = 1, \ldots, d\} \tag{2.13}$$

is the set of translation coefficients corresponding to resolution index j. Function $\Pi_{m=1}^d \eta_{j_m,k_m}^{(0)}(x_m)$ is (a constant times) the indicator of a rectangle but we have multiplied by Haar wavelet $\eta_{j_l,k_l}^{(1)}(x_l)$ in (2.12).

Dyadic rectangles. Write the rectangle corresponding to the pair of multi-indeces $(j, k) \in \{0, 1, \ldots\}^d \times K_j$ as

$$R_{jk} = \Pi_{l=1}^d \left[\frac{k_l}{2^{j_l}}, \frac{k_l + 1}{2^{j_l}} \right), \tag{2.14}$$

where K_j is defined in (2.13). We have defined in Definition 1 a collection of dyadic partition generating trees; when the root node is associated with rectangle $[0, 1)^d$, then every node of the tree is associated with a dyadic rectangle. We have a bijective correspondence between dyadic rectangles and pairs of multi-indeces, defined by (2.14). We denote by $\mathcal{I}(t)$ the pair of multi-indeces associated with a node, that is, when a node is associated with rectangle R_{jk}, then $\mathcal{I}(t) = (j, k)$.

Collection of pre-bases.

Definition 4 of a collection of pre-bases is a counterpart of Definition 2. A difference is that now we take the initial rectangle $R_0 = [0, 1)^d$.

In (2.2) we defined the partition associated with a partition generating tree, we define analogously a pre-basis $\mathcal{B}(\mathcal{T})$ associated with a partition generating tree. Collection $\mathcal{B}(\mathcal{T})$ is a finite orthonormal system and $\int_{[0,1]^d} \phi = 0$ for each $\phi \in \mathcal{B}(\mathcal{T})$. We call these collections "pre-bases", since it is possible to extend them to be bases of $L_2([0, 1]^d)$.

Definition 4. *(Collection of pre-bases.)* When $\mathcal{T} \in \mathbb{T}([0, 1]^d, J)$ is a dyadic partition generating tree, and t is a node of $\mathcal{T}$, let $s(t) \in \{1, \ldots, d\}$ be the direction associated with t and let $\mathcal{I}(t)$ be the pair of multi-indeces associated with t. Denote by $NT(\mathcal{T})$ the set of non-terminal nodes of $\mathcal{T}$. The *pre-basis associated to tree* $\mathcal{T}$is

$$\mathcal{B}(\mathcal{T}) = \left\{ \phi_{\mathcal{I}(t)}^{(s(t))} : t \in NT(\mathcal{T}) \right\}, \tag{2.15}$$

where $\phi_{j,k}^{(l)}$ is defined in (2.12). *The collection of pre-bases* $\mathcal{L}(\mathcal{J})$, with depth bound $J = (J_1, \ldots, J_d)$, is

$$\mathcal{L}(J) = \left\{ \mathcal{B}(\mathcal{T}) : \mathcal{T} \in \mathbb{T}([0, 1]^d, J) \right\}. \tag{2.16}$$

Collection of tree weights.

We define a series estimator whose terms are a subset of a pre-basis $\mathcal{B}(\mathcal{T})$. The series estimator is defined with the help of 0-1-weights that fix a subset of the pre-basis. In order for the series estimator to be equivalent to a dyadic histogram we need a restriction on the weights of the series estimator. The pre-basis $\mathcal{B}(\mathcal{T})$ is associated with tree $\mathcal{T}$ and we require that the weights are such that they correspond to a pruning of the associated tree. The *collection of tree-weights* $\mathcal{W}_{tree,J} = \mathcal{W}_{tree,J}(\mathcal{B})$, associated with $\mathcal{B} \in \mathcal{L}(J)$, is the set of vectors $W = (w_\phi)_{\phi\in\mathcal{B}} \in \{0,1\}^{\mathcal{B}}$, which satisfy the condition that a weight can be zero only when all the "ancestor" weights are zero at the coarser resolution levels. Define

$$\mathcal{W}_{tree,J} = \left\{(w_\phi)_{\phi\in\mathcal{B}} \in \{0,1\}^{\mathcal{B}} : \text{ if } w_\phi = 0 \text{ then } w_{\phi'} = 0 \text{ for all } \phi' \subset \phi\right\}. \quad (2.17)$$

Here $\phi' \subset \phi$ means for $\phi = \phi_{\mathcal{I}(t)}^{(s(t))}, \phi' = \phi_{\mathcal{I}(t')}^{(s(t'))} \in \mathcal{B}$, that $R_{\mathcal{I}(t')} \subset R_{\mathcal{I}(t)}$, where R_{jk} is defined in (2.14).

When $\phi' \subset \phi$, we say that ϕ' is a child of ϕ. The tree condition says that if $w_\phi = 0$, then $w_{\phi'} = 0$ for all children ϕ' of ϕ. Choosing a subset of $\mathcal{B}(\mathcal{T})$ with the help of weights $W \in \mathcal{W}_{tree,J}(\mathcal{B}(\mathcal{T}))$ is equivalent to the pruning of tree $\mathcal{T} \in \mathbb{T}([0,1]^d, J)$.

Definition of the series estimator.

Analogously to (2.8), consider a complexity-penalized error criterion

$$\mathcal{E}_n(W, \Theta, \mathcal{B}, \alpha) = \gamma_n\left(\tilde{f}(\cdot, W, \Theta, \mathcal{B})\right) + \alpha \cdot D(W), \quad (2.18)$$

where γ_n is defined in (2.4), and the complexity penalization is taken to be the number of terms in the expansion:

$$D(W) = \#\{w_\phi : w_\phi = 1\} + 1, \quad (2.19)$$

where $W = (w_\phi)_{\phi\in\mathcal{B}} \in \{0,1\}^{\mathcal{B}}$. We have added 1 in the definition of $D(W)$ since the function $I_{[0,1]^d}$ is also in the expansion (2.11). The series estimator $f_{n,\alpha}^*$ is a minimization estimator where we search a best pre-basis $\mathcal{B}_{n,\alpha}^*$ and a best sub-set of $\mathcal{B}_{n,\alpha}^*$ so that the tree condition is satisfied. The coefficients of the expansion are given by the empirical coefficients $\Theta_n(\mathcal{B})$:

$$\Theta_n(\mathcal{B}) = \left(\hat{\theta}_\phi\right)_{\phi\in\mathcal{B}}, \qquad \hat{\theta}_\phi = \frac{1}{n}\sum_{i=1}^n \phi(X^i). \quad (2.20)$$

Definition 5. *(Dyadic series estimator.)* The empirical choice for the basis $\mathcal{B}$ and for the coefficient vector W is given by

$$\left(\mathcal{B}_{n,\alpha}^*, W_{n,\alpha}^*\right) = \text{argmin}_{\mathcal{B}\in\mathcal{L}(J), W\in\mathcal{W}_{tree,J}(\mathcal{B})}\mathcal{E}_n(W, \Theta_n(\mathcal{B}), \mathcal{B}, \alpha). \quad (2.21)$$

The *dyadic series estimator* is

$$f^*_{n,\alpha}(x) = \tilde{f}\left(x, W^*_{n,\alpha}, \Theta_n(\mathcal{B}^*_{n,\alpha}), \mathcal{B}^*_{n,\alpha}\right), \qquad x \in \mathbf{R}^d, \tag{2.22}$$

where $\tilde{f}(\cdot, W, \Theta, \mathcal{B})$ is defined in (2.11).

2.3. Equivalence between estimators

We prove that the a dyadic histogram is equivalent to a series estimator.

Lemma 1. *We have that* $\hat{f}_{n,\alpha} = f^*_{n,\alpha}$, *where* $\hat{f}_{n,\alpha}$ *is defined in* (2.10) *and* $f^*_{n,\alpha}$ *is defined in* (2.22), *when the initial rectangle of the dyadic histogram is* $R_0 = [0,1)^d$.

A proof of Lemma 1 may be found in the technical report. See also Engel (1994).

2.4. Algorithms and computational complexity

Let us discuss algorithms for solving the minimization problem (2.9). The solution is the partition defining the estimator. One may solve the minimization problem by first building a large multitree which contains all paths leading to partitions, and then pruning the tree.

2.4.1. Growing the tree

First we 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_0. We have d ways of choosing the splitting direction and each binary split gives two bins. Thus $2d$ children will represent the rectangles resulting from the binary splits in d directions. At most J_l splits will be made in direction l, thus the depth of the tree will is $|J|_{max} = \max_{l=1,\ldots,d} |J_l|$. We record the number of observations n_R in each bin R, and calculate $-n_R^2/(n^2 \mathrm{vol}(R))$, so that we are able to calculate (2.7) for all partitions. When some bin is empty of observations we do not split it further. The resulting tree has at most

$$\sum_{i=0}^{|J|_{max}} (2d)^i = O\left((2d)^{|J|_{nax}}\right) \tag{2.23}$$

nodes. For the choice $J = J_n$ as in (3.4), there are $O\left(n^{a \log_2(2d)}\right)$ nodes in the tree.

2.4.2. 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 in some of the d directions helps (whether it results in a smaller complexity-penalized error criterion). If the split does not help, we cut the tree below the

node. This is a multivariate version of the Fast algorithm for Dyadic CART given in Donoho (1997). The number of flops required by the algorithm is bounded by the number of nodes of the tree given in (2.23).

We formulate a lemma which states that the minimization problem is solved by this bottom-up algorithm.

Lemma 2. *Let T be the tree grown in Section* 2.4.1. *Let t be some non-terminal node of T and t_{il}, $i = 1, 2$, $l = 1, \ldots, d$, be the children of t. Denote with R_t and R_{il}, respectively, the rectangles associated with these nodes. Denote the partition minimizing the complexity-penalized error criterion, when we localize to rectangle R which is associated with a node of T, by*

$$\widehat{\mathcal{P}}_{n,\alpha}(R) = \operatorname{argmin}_{\mathcal{P}\in\tilde{\mathbb{P}}(R)} COPERR_n\left(\mathcal{P}, \alpha\right),$$

where $\tilde{\mathbb{P}}(R)$ is the set of partitions $\mathbb{P}(R, J')$ and $J' = J - depth(R)$, where $depth(R)$ is the vector of the number of splits which has been made in each direction to reach R. Let

$$\mathcal{M} = \min\left\{COPERR_n\left(\{R_t\}, \alpha\right), \right.$$
$$\left. COPERR_n\left(\widehat{\mathcal{P}}_{n,\alpha}(R_{1l}), \alpha\right) + COPERR_n\left(\widehat{\mathcal{P}}_{n,\alpha}(R_{2l}), \alpha\right), l = 1, \ldots d\right\}.$$

Then,

$$\widehat{\mathcal{P}}_{n,\alpha}(R_t) = \begin{cases} \{R_t\}, \text{ when } \mathcal{M} = COPERR_n\left(\{R_t\}, \alpha\right), \\ \widehat{\mathcal{P}}_{\alpha}(R_{1l}) \cup \widehat{\mathcal{P}}_{\alpha}(R_{2l}), \text{ when} \\ \mathcal{M} = COPERR_n\left(\widehat{\mathcal{P}}_{n,\alpha}(R_{1l}), \alpha\right) + COPERR_n\left(\widehat{\mathcal{P}}_{n,\alpha}(R_{2l}), \alpha\right). \end{cases}$$

Proof. When $\mathcal{P}_{il} \in \tilde{\mathbb{P}}(R_{il})$, $i = 1, 2$, $l = 1, \ldots, d$, then

$$COPERR_n\left(\mathcal{P}_{1l} \cup \mathcal{P}_{2l}, \alpha\right) = COPERR_n\left(\mathcal{P}_{1l}, \alpha\right) + COPERR_n\left(\mathcal{P}_{2l}, \alpha\right). \quad (2.24)$$

Indeed, $\mathcal{P}_{1l}$ and $\mathcal{P}_{2l}$ are partitions of disjoint rectangles and thus (2.24) follows from (2.7) and the fact that $\#(\mathcal{P}_{1l} \cup \mathcal{P}_{2l}) = \#\mathcal{P}_{1l} + \#\mathcal{P}_{2l}$. On the other hand

$$\tilde{\mathbb{P}}(R_t) = \{\{R_t\}\} \cup \left\{\mathcal{P}_{1l} \cup \mathcal{P}_{2l} : \mathcal{P}_{il} \in \tilde{\mathbb{P}}(R_{il}), i = 1, 2, l = 1, \ldots, d\right\}.$$

We have proved the lemma.

Remark 2. In particular, when we choose t in Lemma 2 to be the root of tree T, then $R_t = R_0$ and $\widehat{\mathcal{P}}_{n,\alpha}(R_0) = \widehat{\mathcal{P}}_{n,\alpha}$ is the global solution defined in (2.9).

Remark 3. In higher-dimensional cases one could modify the definition of the dyadic histogram in order to get an estimator with less computational complexity

and less memory requirements. One possibility is to use CART type algorithms, as in Breiman, Friedman, Olshen and Stone (1984). CART has been used in the case of estimation of classification and regression functions, but the algorithm can be applied also in the case of density estimation. The algorithm for the calculation of regression trees would be modified by replacing the sum of the squared prediction errors by the negative log-likelihood or by the L_2 empirical error, and regressograms would be replaced by histograms.

In CART one constructs empirically (using a greedy algorithm) a sequence of nested partitions, and the final partition is found from this sequence by minimizing the complexity-penalized empirical error. In the case of dyadic histograms one is minimizing the complexity-penalized empirical error over a much larger collection of partitions. Thus one could try to speed up the algorithm by finding various ways to reduce the size of the collection of partitions. The challenge would be to try to reduce the size of the collection of partitions, but simultaneously not to lose the good theoretical properties of dyadic histograms.

3. Rates of Convergence

We prove that the estimator achieves optimal rates of convergence up to a logarithmic factor over anisotropic Besov classes $B_{sp}(L)$. The parameter $p = (p_1, \ldots, p_d)$ of the Besov ball may be such that $p_l < 2$ for each $l = 1, \ldots, d$. In order to reach optimal rates of convergence over such function classes containing functions with high spatial variability, it is essential that the bin widths have variable length in any single direction. We denote the intersection of the Besov ball with the set of bounded densities as

$$\mathcal{F} = \mathcal{F}_{sp}(L, B_\infty) = B_{sp}(L) \cap \Big\{ f : \int_{[0,1]^d} f = 1, 0 \le f \le B_\infty \Big\}, \tag{3.1}$$

where $0 < B_\infty < \infty$. We define the anisotropic Besov ball $B_{sp}(L)$, where $s = (s_1, \ldots, s_d) \in (0,1]^d$, $p = (p_1, \ldots, p_d) \in [1, \infty]^d$, $0 < L < \infty$, to be the set of functions $f : [0,1]^d \to \mathbf{R}$ satisfying, for $l = 1, \ldots, d$,

$$\|D_h^l f\|_{L_{p_l}(A_h^l)} \le L h^{s_l}.$$

Here $0 < h < 1$, $D_h^l f(x) = f(x + he_l) - f(x)$, $e_l \in \mathbf{R}^d$ with the l:th coordinate one and the other coordinates zero, and

$$A_h^l = \{(x_1, \ldots, x_d) : 0 \le x_m \le 1, m \ne l, 0 \le x_l < 1 - h\}. \tag{3.2}$$

For more on anisotropic Besov spaces, see Nikol'skii (1975).

The result. The exponent r of the optimal rate of convergence and the anisotropic smoothness index σ are defined by

$$r = \frac{\sigma}{2\sigma+1}, \qquad \sigma = \Big(\sum_{l=1}^{d} s_l^{-1}\Big)^{-1}. \tag{3.3}$$

Besides the smoothing parameter α, the estimator depends on the vector of maximal directionwise split numbers J and we take

$$J_n = (J_{n,1}, \ldots, J_{n,d}), \qquad J_{n,l} = \left\lceil \frac{\sigma}{s_l}\, a \log_2 n \right\rceil, \tag{3.4}$$

where $a \geq 0$ is the fineness parameter. The initial rectangle of the dyadic histogram is $R_0 = [0,1]^d$.

Theorem 1. *Let $X^1, \ldots, X^n$ be i.i.d. observations from the density $f \in \mathcal{F}$. When s_l, p_l, and the fineness parameter a in* (3.4) *are such that*

$$\sigma - (\frac{1}{p_l} - \frac{1}{2})_+ > 0, \qquad l = 1, \ldots, d, \tag{3.5}$$

$$\frac{\sigma}{2\sigma+1}\, \frac{1}{\sigma - (\frac{1}{p_l} - \frac{1}{2})_+} < a < 1, \qquad l = 1, \ldots, d, \tag{3.6}$$

then

$$\limsup_{n\to\infty} \left(\frac{n}{\log_e n}\right)^{2r} \sup_{f\in\mathcal{F}} E_f \int_{[0,1]^d} \Big(f - \hat{f}_{n,\alpha_n}\Big)^2 < \infty,$$

where $\hat{f}_{n,\alpha}$ is defined in (2.10),

$$\alpha_n = C B_\infty \frac{\log_e n}{n}, \tag{3.7}$$

and $C > 0$ is a sufficiently large constant.

A proof of Theorem 1 is given in Section 3.1.

Remark 4. (Adaptiveness of the estimator.) The choice of penalization parameter α in Theorem 1 does not depend on the smoothness parameters $s_1, \ldots, s_d$, nor on $p_1, \ldots, p_d$, or L. Vector J depends on $s_1, \ldots, s_d$, and on the fineness parameter a. The lower bound for a depends on the parameters s_l and p_l, but we may take a arbitrarily close to 1.

Remark 5. (The fineness parameter and restrictions on the smoothness.) Because $s_i \leq 1$ we have $\sigma \leq 1/d$. By (3.5), $\sigma > (1/p_l - 1/2)_+$. Thus, Theorem 1 holds only for σ satisfying

$$\max_{l=1,\ldots,d} \left(\frac{1}{p_l} - \frac{1}{2}\right)_+ < \sigma \leq \frac{1}{d}.$$

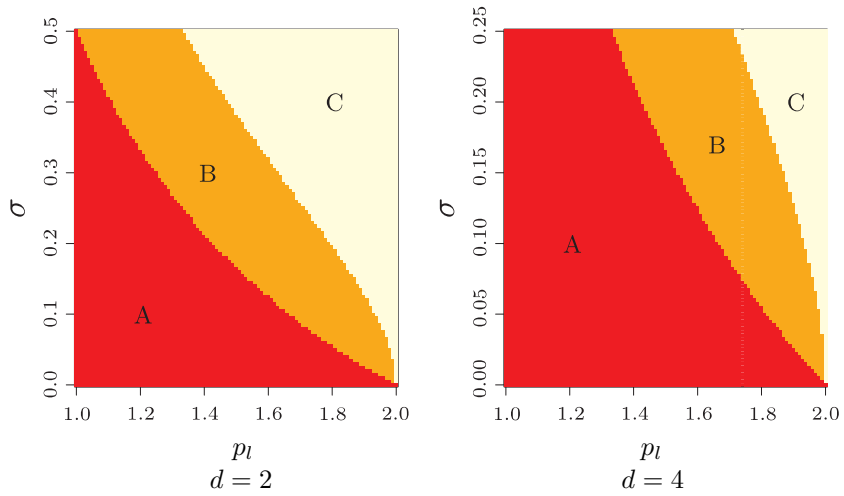

Figure 1. The admissible range C of σ and p_l for a) $d = 2$ and b) $d = 4$.

For large values of d, parameters p_l cannot be much smaller than 2. When $\min_{l=1,\dots,d} p_l \geq 2$, Theorem 1 holds for $0 < \sigma \leq 1/d$. Figure 1 shows the possible values of $(p_l, \sigma) \in [1, 2] \times [0, d^{-1}]$ for a) $d = 2$ and (b) $d = 4$. Region A is the region where condition (3.5) is violated, so that $\sigma \leq (1/p_l - 1/2)_+$. Region B is the region where condition (3.6) may not be satisfied because $c' = [\sigma/(2\sigma + 1)][\sigma - (1/p_l - 1/2)_+]^{-1} > 1$. Region C is the region where condition (3.5) is satisfied and $c' \leq 1$.

3.1. Proof of Theorem 1

Since dyadic histograms are equivalent to dyadic series estimators, as proved in Lemma 1, it is enough prove the theorem for the series estimator. We go through the steps of the proof and after that we give details for the proofs of Step 2 and Step 4. Details for Step 1 are given in Appendix B and Step 3 is proved in the technical report.

Step 1. (*Application of an oracle inequality.*) The first step is to bound the MISE of the estimator by a minimal complexity-penalized approximation error. We have for the series estimator f^*_{n,α_n}, for continuous densities $f : [0, 1]^d \to \mathbf{R}$, that

$$E_f \left\| f^*_{n,\alpha_n} - f \right\|_2^2 \leq C_1 \min_{(W,\Theta,\mathcal{B}) \in \mathbb{K}_0} K(f, W, \Theta, \mathcal{B}, \alpha_n) + C_2 n^{-1}, \tag{3.8}$$

where C_1 and C_2 are positive constants,

$$K(f, W, \Theta, \mathcal{B}, \alpha) = \left\| \tilde{f}(\,\cdot\,, W, \Theta, \mathcal{B}) - f \right\|_2^2 + \alpha \cdot D(W), \tag{3.9}$$

and

$$\mathbb{K}_0 = \left\{ (W, \Theta, \mathcal{B}) \in \mathcal{W}(\mathcal{B}) \times \mathbf{R}^{\mathcal{B}} \times \mathcal{L} : \|\tilde{f}(\cdot, W, \Theta, \mathcal{B})\|_\infty \leq 2B_\infty \right\}, \tag{3.10}$$

where $\mathcal{W}(\mathcal{B}) = \mathcal{W}_{tree,J_n}(\mathcal{B})$, $\mathcal{L} = \mathcal{L}(J_n)$, and $B_\infty > \|f\|_\infty$ is a positive constant. Eq. (3.8) is proved in Appendix B.

Step 2. *(Choosing a basis.)* We bound the approximation error by finding a pre-basis $\mathcal{B}^*_{\alpha_n} \in \mathcal{L}(J_n)$, which is in a sense the best pre-basis for $f \in B_{sp}(L)$. After fixing the pre-basis to be $\mathcal{B}^*_{\alpha_n}$, we choose the vector of coefficients to be the coefficients of f in the pre-basis $\mathcal{B}^*_{\alpha_n}$; $\Theta = \Theta_f(\mathcal{B}^*_{\alpha_n})$, where

$$\Theta_f(\mathcal{B}) = \left(\int_{\mathbf{R}^d} f\phi \right)_{\phi \in \mathcal{B}}. \tag{3.11}$$

We have the upper bound

$$\min_{(W,\Theta,\mathcal{B}) \in \mathbb{K}_0} K(f, W, \Theta, \mathcal{B}, \alpha_n) \leq \min_{W \in \mathcal{W}(\mathcal{B}^*_{\alpha_n})} K(f, W, \Theta_f(\mathcal{B}^*_{\alpha_n}), \mathcal{B}^*_{\alpha_n}, \alpha_n),$$

where $\mathcal{W}(\mathcal{B}^*_{\alpha_n}) = \mathcal{W}_{tree,J_n}(\mathcal{B}^*_{\alpha_n})$.

Step 3.

The minimization is restricted to the tree weights. One may show that this restriction does not greatly increase the complexity-penalized approximation error: we do not get much better approximation by minimizing the weights over a larger collection of weights. When (3.5) holds,

$$\begin{aligned} &\sup_{f \in B_{sp}(L)} \min_{W \in \mathcal{W}_{tree,J_n}(\mathcal{B}^*_{\alpha_n})} K(f, W, \Theta_f(\mathcal{B}^*_{\alpha_n}), \mathcal{B}^*_{\alpha_n}, \alpha_n) \\ &\quad \leq C \sup_{f \in B_{sp}(L)} \min_{W \in \{0,1\}^{\mathcal{B}^*_{\alpha_n}}} K(f, W, \Theta_f(\mathcal{B}^*_{\alpha_n}), \mathcal{B}^*_{\alpha_n}, \alpha_n) \end{aligned} \tag{3.12}$$

for a positive constant C, depending on s, p, L, d. A proof of (3.12) may be found in the technical report. See also Donoho (1997).

Step 4. The last step is to bound the complexity-penalized approximation error in (3.12). We have that

$$\sup_{f \in B_{sp}(L)} \min_{W \in \{0,1\}^{\mathcal{B}^*_{\alpha_n}}} K(f, W, \Theta_f(\mathcal{B}^*_{\alpha_n}), \mathcal{B}^*_{\alpha_n}, \alpha_n) \leq C\alpha_n^{2r}. \tag{3.13}$$

This proves the theorem by the choice of α_n in (3.7).

3.1.1. Step 2

We define a best basis for the approximation of functions in $B_{sp}(L)$. Let $h : \{1, 2, \ldots\} \to \{1, \ldots, d\}$. We apply h as a *direction selection rule*, for choosing trees and multi-indeces:

- Every tree in $\mathbb{T}([0,1]^d, J)$ is uniquely determined by the splitting directions. Let $\mathcal{T}_{h,M}$ be the partition generating tree determined by the following rules.
  1. Split the root node in direction $h(1)$.
  2. The 2^m nodes at depth m are splitted in direction $m+1$, for $m \in \{0, \ldots, M-1\}$.
- For any direction selection rule h, we denote the corresponding sequence of multi-indeces $\mathcal{J} = \mathcal{J}_h : \{0,1,\ldots\} \to \{0,1,\ldots\}^d$, by $\mathcal{J} = (j_1, \ldots, j_d)$, where $j_l(m)$ is the number of times direction l was chosen by h up to step m: $j_l(0) = 0$ and
$$j_l(m) = \#\{m' \le m : h(m') = l\}, \qquad m = 1, 2, \ldots, \tag{3.14}$$
$l = 1, \ldots, d$.

(*Definition of h^*.*) We focuse on a direction selection rule h^* that depends on the vector of smoothness indeces $s = (s_1, \ldots, s_d)$ of the anisotropic Besov space $B_{sp}(L)$. Define the sequence $\mathcal{Z}(m) = \mathcal{Z}_s(m) = (z_1(m), \ldots, z_d(m)) \in [0,\infty)^d$, $m = 0, 1, \ldots$, satisfying $\mathcal{Z}(0) = (0, \ldots, 0)$, and

$$\begin{cases} z_1(m)s_1 = \cdots = z_d(m)s_d \\ z_1(m) + \cdots + z_d(m) = m. \end{cases} \tag{3.15}$$

That is, $\mathcal{Z}(1) \in \{x \in [0,\infty)^d : x_1 s_1 = \cdots = x_d s_d\}$ is such that $\sum_{l=1}^d z_l(1) = 1$, and $\mathcal{Z}(m) = m\mathcal{Z}(1)$ for $m \ge 1$ integer. We take h^* so that $\mathcal{J}^* = \mathcal{J}_{h^*}$ is an approximation to $\mathcal{Z}_s$, taking values on a grid. The direction selection rule h^* is defined by the following rules.

1. Choose $h^*(1) = \operatorname{argmax}_{l \in \{1,\ldots,d\}} z_l(1)$, that is, $h^*(1) = \operatorname{argmin}_{l \in \{1,\ldots,d\}} s_l$.
2. Write $\mathcal{J}^*(m) = (j_1^*(m), \ldots, j_d^*(m))$. Define for $m = 1, 2, \ldots$,
$$h^*(m+1) = \operatorname{argmax}_{l \in \{1,\ldots,d\}} z_l(m) - j_l^*(m).$$
That is, we choose the direction where $\mathcal{J}^*(m)$ is furthest below from $\mathcal{Z}(m)$.

Then $z_l(m)s_l = m\sigma$, for $l = 1, \ldots, d$, and

$$j_l^*(m)s_l \sim m\sigma \tag{3.16}$$

as $m \to \infty$, where σ is defined in (3.3). This means that the proportion in which direction l was chosen, $j_l^*(m)/m$, is approximately equal to σ/s_l.

(Definition of the pre-basis.) We choose

$$M_{\alpha_n}^* = \left[a \log \alpha_n^{-1}\right].$$

We have, see (3.16),

$$j_l^*(M_{\alpha_n}^*) \le J_{n,l}, \qquad l = 1, \ldots, d, \tag{3.17}$$

$$\mathcal{T}_{h^*, M_{\alpha_n}^*} \in \mathbb{T}([0,1]^d, J_n). \tag{3.18}$$

With

$$\mathcal{B}_{\alpha_n}^* = \mathcal{B}(\mathcal{T}_{h^*, M_{\alpha_n}^*}), \tag{3.19}$$

(3.18) implies that $\mathcal{B}_{\alpha_n}^* \in \mathcal{L}(J_n)$. Take

$$\mathcal{B}^* = \left\{ I_{[0,1]^d} \right\} \cup \mathcal{B}\left(\mathcal{T}_{h^*, \infty}\right) \tag{3.20}$$

as a spatially homogeneous anisotropic basis. A proof that $\mathcal{B}^*$ is a basis of $L_2([0,1]^d)$ may be found in the technical report.

3.1.2. Step 4

Largeness of the wavelet coefficients in basis $\mathcal{B}^*$.

We need a bound on the coefficients $\int_{[0,1]^d} f\phi$, $\phi \in \mathcal{B}^*$. To give the bound it is convenient to write the basis in terms of Haar-basis functions. We have that

$$\mathcal{B}^* = \bigcup_{m=0}^{\infty} \left\{ \phi_{\mathcal{J}^*(m),k}^{(h^*(m+1))} : k \in K_{\mathcal{J}^*(m)} \right\},$$

where the $\phi_{j,k}^{(l)}$ are defined in (2.12) and the K_j in (2.13). We denote the coefficients of f by

$$\tau_{mk} = \int_{[0,1]^d} f \phi_{\mathcal{J}^*(m),k}^{(h^*(m+1))}, \tag{3.21}$$

where $m = 0, 1, \ldots$ and $k \in K_{\mathcal{J}^*(m)}$.

Lemma 3. *Let $f \in B_{sp}(L)$, $s = (s_1, \ldots, s_d) \in (0,1]^d$, $p = (p_1, \ldots, p_d) \in [1, \infty]^d$. Then we have, for $m = 0, 1, \ldots$, that*

$$\left(\sum_{k \in K_{\mathcal{J}^*(m)}} |\tau_{mk}|^{\tilde{p}_{l_m^*}} \right)^{\frac{1}{\tilde{p}_{l_m^*}}} \le 2^{\frac{d}{2}} L 2^{-m(\sigma + \frac{1}{2} - \frac{1}{\tilde{p}_{l_m^*}})}, \tag{3.22}$$

where we use the notation $l_m^* = h^*(m+1)$ and

$$\tilde{p}_l = \min\{p_l, 2\}, \qquad l = 1, \ldots, d. \tag{3.23}$$

A proof of Lemma 3 may be found in the technical report.

Lemma 4. *Let* (3.5) *be satisfied, and let* $\tilde{a}$ *satisfy*

$$\tilde{a} > \frac{\sigma}{2\sigma+1}\,\frac{1}{\sigma-(\frac{1}{p_l}-\frac{1}{2})_+}\,, \qquad l=1,\dots,d. \tag{3.24}$$

Let $M = M_\alpha$ *be an integer satisfying*

$$M_\alpha \geq \tilde{a}\log_2 \alpha^{-1} \tag{3.25}$$

and let $\mathcal{B}^*_\alpha = \mathcal{B}(\mathcal{T}_{h^*,M_\alpha})$. *Then*

$$\sup_{f\in B_{sp}(L)}\min_{W\in\{0,1\}^{\mathcal{B}^*_\alpha}} K\left(f,W,\Theta_f(\mathcal{B}^*_\alpha),\mathcal{B}^*_\alpha,\alpha\right) \leq C\alpha^{2r} \tag{3.26}$$

for a positive constant C, *depending on* s,p,L,d, *when* $0<\alpha<1$ *is sufficiently small, where* $r=\sigma/(2\sigma+1)$ *is defined in* (3.3).

Lemma 4 is proved in Appendix C; and (3.13) follows from it.

Acknowledgements

Writing of this article was financed by Deutsche Forschungsgemeinschaft under project MA1026/8-2. I wish to thank the editors and the referees for helpful comments.

References

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

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

Castellan, G. (2000). Sélection d'histogrammes à l'aide d'un critère de type Akaike. *C. R. Acad. Sci., Paris I* **330**, 729-732.

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

Engel, J. (1994). A simple wavelet approach to nonparametric regression from recursive partitioning schemes. *J. Multivariate Anal.* **49**, 242-254.

Hüsemann, J. A. and Terrell, G. R. (1991). 'Optimal parameter choice for error minimization in bivariate histograms'. *J. Multivariate Anal.* **37**, 85-103.

Kerkyacharian, G., Lepski, O. and Picard, D. (2001). Nonlinear estimation in anisotropic multiindex denoising. Probab. Theory Relat. Fields 121, 137-170.

Lugosi, G. and Nobel A. (1996). Consistency of data-driven histogram methods for density estimation and classification. *Ann. Statist.* **24**, 687-706.

Neumann, M. H. (2000). Multivariate wavelet thresholding in anisotropic function spaces. *Statist. Sinica* **10**, 399-431.

Nikol'skii, S. M. (1975). *Approximation of Functions of Several Variables and Imbedding Theorems.* Springer-Verlag, Berlin.

Ooi, H. (2002). Density visualization and mode hunting using trees. J. Com- put. Graph. Statist. 11, 328-347.

Shang, N. (1994). Tree-structured density estimation and dimensionality reduction. In *Proceedings of the* 26*rd Symposium on the Interface*, 172-176.

Sutton, C. D. (1994). Tree structured density estimation. Computing Science and Statistics 26, 167-171.

Tribouley, K. (1995). Practical estimation of multivariate densities using wavelet methods. *Statist. Neerlandica* **49**, 41-62.

Department of Mathematical Sciences, University of Oulu, P.O. Box 3000, FIN-90014, University of Oulu, Finland.

E-mail: jussi.klemela@oulu.fi

(Received November 2006; accepted August 2007)