Available online at www.sciencedirect.com
SCIENCE DIRECT®
Digital Signal Processing 15 (2005) 73–116
Digital Signal Processing
www.elsevier.com/locate/dsp

Multiresolution local polynomial regression: A new approach to pointwise spatial adaptation

Vladimir Katkovnik [a,b]

[a] *Signal Processing Laboratory, Tampere University of Technology, P.O. Box 553, Tampere, Finland*
[b] *Department of Mechatronics, Kwangju Institute of Science and Technology, Kwangju 500-712, South Korea*

Available online 5 October 2004

Abstract

In nonparametric local polynomial regression the adaptive selection of the scale parameter (window size/bandwidth) is a key problem. Recently new efficient algorithms, based on Lepski's approach, have been proposed in mathematical statistics for spatially adaptive varying scale denoising. A common feature of these algorithms is that they form test-estimates $\hat{y}_h$ different by the scale $h \in H$ and special statistical rules are exploited in order to select the estimate with the best pointwise varying scale. In this paper a novel multiresolution (MR) local polynomial regression is proposed. Instead of selection of the estimate with the best scale h a nonlinear estimate is built using all of the test-estimates $\hat{y}_h$. The adaptive estimation consists of two steps. The first step transforms the data into noisy spectrum coefficients (MR analysis). On the second step, this noisy spectrum is filtered by the thresholding procedure and used for estimation (MR synthesis).

Keywords: Adaptive scale; Kernel estimation; Local polynomial regression; Multiresolution analysis; Local nonparametric regression; Denoising; Sliding least square method

1. Introduction

The present work is devoted to studying the problem of adaptive estimation of a multivariable function given by noisy data. The developed multiresolution approach (MR) and algorithms are of a general nature and can be applied to a variety of univariate or mul-

E-mail address: katkov@cs.tut.fi.

1051-2004/$ – see front matter © 2004 Published by Elsevier Inc.
doi:10.1016/j.dsp.2004.06.004

tidimensional problems. However, we prefer to give the presentation in terms of image processing as it allows a convenient and transparent motivation of basic ideas as well as a good illustration of results. Thus, let the function to estimate be a two-dimensional ($2D$) image intensity given by noisy values on a $2D$ regular grid.

The adaptive estimation used in this paper is based on direct nonparametric pointwise estimation of the image intensity without any preliminary edge recovering. A nonparametric regression estimator is derived from the local polynomial approximation (LPA) in a sliding window with a varying size (estimator's scale) selected in a data-driven way.

Among others nonparametric approaches to regression estimation the LPA can be treated as probably one of the most theoretically justified and well studied. It is a powerful nonparametric technique which provides estimates in a pointwise manner based on a mean square polynomial fitting in a sliding window (e.g., [2,7,10,14–16,30]). In terms of image processing the LPA is a flexible tool to design $2D$ kernels (masks) having prescribed reproducing properties with respect to polynomial (smooth) components of the signal. The invariant and varying scale selection for the LPA has been studied thoroughly by many authors. Optimal, in particular, varying data-driven scale methods are of special interest for problems where the pointwise smooth approximation is natural and relevant. Image denoising provides good examples of this sort of problems.

A crucial difference between the nonparametric LPA and the more traditional parametric methods, say the polynomial mean squared estimates, is that the latter are formed as unbiased ones while the nonparametric estimates are biased and the reasonable choice of the biasedness controlled by the scale parameter is of importance. In the nonparametric regression methods adaptive to unknown smoothness the adaptive selection of the scale is a key point.

The problem of optimal scale selection admits an accurate mathematical formulation in terms of the nonparametric approach, where the optimal scale is defined by a compromise between the bias and the variance of estimation (e.g., [7,16,30]).

The idea of the used Lepski's adaptation method is as follows, [27–29,38]. The algorithm searches for a largest local vicinity of the point of estimation where the LPA assumption fits well to the data. The test-estimates $\hat{y}_h(x)$ are calculated for window sizes h from the set H, $h \in H$, and compared. The adaptive window size is defined as the largest of those windows which estimate does not differ significantly from the estimates corresponding to the smaller window sizes. Special statistics are exploited in order to test these hypotheses and select the best scale giving the optimal balance between the random errors and the biasedness of the estimate. The Lepski's approach algorithms have introduced a number of statistical rules which are proved to be efficient in theory as well as in applications. The nonlinearity of the method is incorporated by an adaptive pointwise choice of the scale.

A novel spatial adaptivity introduced in this paper can be viewed as a development of two independent ideas: wavelet multiresolution analysis (e.g., [11,31]) and a pointwise adaptive scale selection based on the Lepski's approach.

Instead of selection of the estimate with the best scale h we build a nonlinear estimate using all of the available test-estimates $\hat{y}_h(x)$, $h \in H$. The adaptive estimation is divided into two successive steps. The first step transforms the data into noisy spectrum coefficients (MR analysis). In the second step, these coefficients are filtered by the thresholding

procedure and used for estimation (MR synthesis). The LPA based filters are exploited for the nonparametric (pointwise) MR spectrum analysis and synthesis.

In this way we introduce the extension of the conventional scale adaptive nonparametric regression concept and yield a wider class of the adaptive scale regression estimators with a potentially better performance.

The contribution of this paper is two-fold. Firstly, we present a summary of the nonparametric LPA methods including the basic ideas, algorithms for function and derivative estimation, the accuracy analysis. This summary is concluded by a brief review of some recent methods for selection of the adaptive varying scales. Secondly, the new MR local polynomial regression is introduced as a valuable alternative to the conventional adaptive scale nonparametric regression.

The rest of the paper is organized as follows. In Section 2 the observation model as well as the estimation problem are discussed. The LPA method, motivation, basic algorithms and the accuracy analysis are presented in Sections 3 and 4. The Lepski's adaptive varying scale algorithms are reviewed in Section 4.3. The nonparametric regression spectrum and MR analysis are introduced in Section 5. The multiresolution filtering based on the MR spectrum thresholding is presented in Section 6. The optimality of the adaptive scale local regression estimation is discussed briefly in Section 7. Implementation of the introduced algorithms as well as their complexity is a subject of Section 8. Similarity and difference of the MR local polynomial regression versus the wavelet MR techniques are discussed in Section 9. In Section 10 simulation results are presented. It is shown that appropriate filtering of the noisy MR spectrum allows to achieve a better performance that it can be done using the more traditional nonparametric approach based on selection of the best varying adaptive scale.

2. Observation model

The following model, commonly used for image denoising, is assumed:

$$z(x) = y(x) + \sigma\varepsilon(x), \tag{1}$$

where an intensity y of the underlying image is defined as a function of two variables, $y \in R$, ε is an additive noise. It is assumed that all functions in (1) are defined on a $2D$ rectangular regular grid $x \in X \subset R^2$ with pixels $x = (x_1, x_2)$:

$$X = \big\{(x_1, x_2)\colon\ x_1 = s_1\Delta,\ x_2 = s_2\Delta\big\}, \quad s_1 = 1, \ldots, n_1,\ s_2 = 1, \ldots, n_2, \tag{2}$$

where s_1, s_2 are integers, Δ denotes the sampling interval, and $n = n_1 n_2$ is a total number of observations. The random noise ε is assumed to be standard Gaussian i.i.d. for different x with $E\{\varepsilon^2\} = 1$. The basic objective is to reconstruct (estimate) $y(x)$ and derivatives of $y(x)$ for any $x \in X$ from noisy observations $\{z(x), x \in X\}$ with the pointwise mean squared error (MSE) risk which is as small as possible.

It is assumed that y is unknown deterministic. For stochastic y it means that the main intention is to obtain the best result for every realization of y even if they are generated by a probabilistic phenomenon.

The discrete observations $\{y(x), x \in X\}$ are obtained as samples of an underlying continuous argument $y(x), x \in R^2$. Different hypotheses on this y can be applied for derivation and analysis of algorithms. Here we follow the nonparametric regression approach assuming that a parametric representation of y as a function of x, say in the form of a series or function with reasonably small number of invariant parameters, does not exist or unknown.

The following piecewise model of y is appropriate for methods considered in this paper. Let a finite support of y can be separated into Q regions A_q, $q = 1, \ldots, Q$, each of them is a connected set with an edge (boundary) G_q. The function y is assumed to be smooth differentiable within each region A_q:

$$y(x) = \sum_{q=1}^{Q} y_q(x) \mathbb{1}[x \in A_q], \tag{3}$$

where $\mathbb{1}[x \in A_q]$ is an indicator of the region A_q, $\mathbb{1}[x \in A_q] = 1$ if $x \in A_q$ and zero otherwise, and y_q is a continuous differentiable function belonging to the class

$$F_{|r|}(\bar{L}_{|r|}) = \Big\{ y\colon \max_{r_1+r_2=|r|} |y^{(r)}(x)| = L_{|r|}(x) \leqslant \bar{L}_{|r|},\ \forall r_1 + r_2 = |r|,\ x \in R^2 \Big\}. \tag{4}$$

Here and in what follows $r = (r_1, r_2)$ is a multi-index, r_1, r_2 nonnegative integer and $|r| = r_1 + r_2$. A derivative corresponding to r is $y^{(r)}(x) = \frac{\partial^{r_1+r_2}}{\partial x_1^{r_1} \partial x_2^{r_2}} y(x)$, $\bar{L}_{|r|}$ is a finite constant in (4).

The piecewise constant model of y

$$y(x) = \sum_{q=1}^{Q} a_q \mathbb{1}[x \in A_q], \quad 0 \leqslant a_q \leqslant 1, \tag{5}$$

is a particular case of (3) with constant values within each region A_q.

In the models (3) and (5) y_q, a_q as well as the regions A_q are unknown. The boundaries G_q define change points of the piecewise smooth y. The estimation of y can be produced in two different ways. One of the possible approaches deals with a two-stage procedure including estimation of the boundaries G_q on the first stage, which defining the regions A_q. The second stage is a parametric or nonparametric fitting y_q on A_q.

Another approach is connected with the concept of spatially adaptive estimation. In this context, the change points or, more generally, cusps in the curves can be viewed as a sort of an inhomogeneous behavior. One may therefore apply the same procedure, for instance nonlinear wavelet, ridgelet, curvelet estimators for all x, and the analysis focuses on the quality estimation when the change-points are incorporated or not incorporated in the model. Under this approach, the main objective is to estimate the function and not location of change-points which are treated as features of the function surface.

In this paper we follow the second approach. The objective is to develop a method which simultaneously adapts to varying smoothness of the estimated function and which is sensitive to discontinuities of the function and its derivatives.

3. Local polynomial approximation

3.1. Foundations

The idea of the LPA is simple and natural. It is assumed that the function y is well approximated by a polynomial in a neighborhood of the point of interest x. We find the coefficients of the polynomial fit by the weighted least square method and use this approximation in order to calculate the estimate for the point of interest x called also "centre" of the LPA. In fact, the local expansion is applied in order to calculate the estimate for this point of interest only. For the next point the calculations are repeated. This pointwise procedure determines a nonparametric character of the LPA estimation.

The linear LPA estimators have a very long prehistory (e.g., [2,7,14–16,30]). They are a very popular tool in statistics and signal processing with application to a wide variety of the fields for smoothing, filtering, differentiation, interpolation and extrapolation.

Note that the LPA has appeared in signal processing in a number of modifications and under different names: sliding (moving) least square, Savitzky–Golay filter, local regression, reproducing kernel method, moment filters, etc. We prefer the term LPA with a reference to publications on nonparametric estimation in mathematical statistics where the advanced development of this technique can be seen.

In this section we summarize the well known concepts of the discrete LPA.

Let $x \in R^2$ be a "centre" (reference point) of the LPA. Then, the estimate for $v \in R^2$ in the neighborhood of the centre x is presented as an expansion:

$$\begin{aligned} &y(x, v) = C^T \phi(x - v), \\ &\phi(x) = \big(\phi_1(x), \phi_2(x), \ldots, \phi_M(x)\big)^T, \\ &C = (C_1, C_2, \ldots, C_M)^T, \end{aligned} \tag{6}$$

where $\phi(x) \in R^M$ is a vector of linear independent $2D$ polynomials of the powers from 0 up to m, $C \in R^M$ is a vector of parameters of this model. In particular, the following polynomials can be exploited

$$\frac{x_1^{k_1} x_2^{k_2}}{k_1! k_2!}, \quad 0 \leqslant k_1 + k_2 \leqslant m,\ k_1, k_2 \geqslant 0. \tag{7}$$

A total number of these $2D$ polynomials is equal to $M = (m+2)(m+1)/2$. For $m \leqslant 3$ we obtain

$$\begin{aligned} &\phi_1 = 1 \quad \text{for } m = 0, \\ &\phi_2 = x_1, \qquad \phi_3 = x_2 \quad \text{for } m = 1, \\ &\phi_4 = x_1^2/2, \qquad \phi_5 = x_2^2/2, \qquad \phi_6 = x_1 x_2 \quad \text{for } m = 2, \\ &\phi_7 = x_1^3/6, \qquad \phi_8 = x_2^3/6, \qquad \phi_9 = x_1^2 x_2/2, \qquad \phi_{10} = x_1 x_2^2/2 \quad \text{for } m = 3, \end{aligned} \tag{8}$$

with $M = 1, 3, 6$, and 10 for $m = 0, 1, 2, 3$, respectively.

The term "centre" does not assume a central position of x in the neighborhood. It only emphasizes that the LPA is exploited in order to obtain the estimate for this particular value of the argument of y.

Table 1

Window name	$w(x)$
Rectangular symmetric	1, $\|x\| \leqslant 1/2$,
Rectangular nonsymmetric	1, $0 \leqslant x \leqslant 1$,
Exponential	$\frac{1}{2}\exp(-\|x\|)$,
Gaussian	$\frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$,
Epanechnikov	$\frac{3}{4}(1-x^2)$, $\|x\| \leqslant 1$,
Bisquare window	$(1-x^2)^2$, $\|x\| \leqslant 1$,
Tricube	$(1-\|x\|^3)^3$, $\|x\| \leqslant 1$,
Triweight window	$(1-x^2)^3$, $\|x\| \leqslant 1$.

The conventional quadratic criterion function is applied in order to find the coefficient C in (6):

$$J_h(x) = \sum_{v \in X} w_h(x-v)\big(z(v) - y(x,v)\big)^2, \tag{9}$$

where $\{z(v),\ v \in X\}$ are discrete observations and the window

$$w_h(x) = w(x/h)/h^2 \tag{10}$$

is used to formulate mathematically the fitting localized in a neighborhood of x, while the scale parameter $h > 0$ determines the size of the neighborhood. The windowing weight w is usually assumed to satisfy the properties:

$$w(x) \geqslant 0, \qquad w(0) = \max_x w(x), \qquad \int_{R^2} w(x)\,dx = 1, \qquad \int_{R^2} w^2(x)\,dx < \infty. \tag{11}$$

The multiplicative window

$$w(x) = w_1(x_1)w_2(x_2), \tag{12}$$

where w_1 and w_2 are functions of scalar ($1D$) arguments, is commonly applied. If the window is rectangular all observations enter in the criteria (9) with equal weights. Nonrectangular windows such as triangular, quadratic, Epanechnikov, and so on (see [7,16,30]) usually prescribe higher weights to observations which are closer to the centre x. Some typical $1D$ window functions used in local regression estimates are shown in Table 1. Let us also mention windows conventional in signal processing and associated with the names: Kaiser, Hamming, Bartlett, Blackman, Chebyshev, etc. Note that the B-splines also can be used as the windows for the LPA.

There is a simple way to generate nontrivial $2D$ windows different from the multiplicative ones (12). Let us replace the argument x_1 in $w_1(x_1)$ by the norm $\|x\|$, where x is a vector and the norm is not exclusively Euclidian. Then after the corresponding normalization we obtain $2D$ window functions satisfying (11).

Let $\hat{y}_h(x)$ be the LPA estimate of $y(x)$, where the subindex h shows a dependence of the estimate on the scale parameter. This LPA estimate of $y(x)$ is defined according to (6) as $\hat{y}_h(x) = y(x,v)|_{v=x} = y(x,x)$, i.e., the expansion (6) is used for calculation of the estimate for $v = x$ only.

Actually, it is one of the key ideas of the pointwise nonparametric estimate design. We introduce the estimate as the expansion in the local neighborhood of the point x, we estimate the coefficients of this expansion, and finally we use this expansion only in order to estimate at this argument value x. Then, it follows from (6) that

$$\hat{y}_h(x) = y(x,v)|_{v=x} = C^T\phi(0) \tag{13}$$

and for the polynomials (7) it yields

$$\hat{y}_h(x) = y(x,v)|_{v=x} = C_1. \tag{14}$$

Let $\hat{y}_h^{(k)}(x)$ be the estimator of a kth derivative $y_h^{(k)}(x)$ of $y(x)$. Here $k=(k_1,k_2)$ is a multi-index with k_1, k_2 nonnegative integers. The LPA model (6) of the power m can be used for estimation of any derivative of the order k, $|k| \leqslant m$. According to the idea of the pointwise estimation we derive these estimates in the form

$$\begin{aligned}\hat{y}_h^{(k)}(x) &= \left.\frac{\partial^{|k|}y(x,v)}{\partial v_1^{k_1}\partial v_2^{k_2}}\right|_{v=x} = (-1)^{|k|}C^T\phi^{(k)}(0),\\ \phi^{(k)}(0) &= \left.\frac{\partial^{|k|}\phi(x)}{\partial x_1^{k_1}\partial x_2^{k_2}}\right|_{x=0}.\end{aligned} \tag{15}$$

This definition of the derivative estimator assumes that differentiation in (15) is done with respect to v as the approximation $y(x,v)$ is a function of v provided that the LPA centre x is fixed. After the differentiation we assume $v=x$.

For the polynomials (8) the derivative estimates (15) are simple:

$$\begin{aligned}&\hat{y}_h^{(1,0)}(x) = -C_2, \qquad \hat{y}_h^{(0,1)}(x) = -C_3, \qquad \hat{y}_h^{(2,0)}(x) = C_4,\\ &\hat{y}_h^{(0,2)}(x) = C_5, \qquad \hat{y}_h^{(1,1)}(x) = C_6, \text{ etc.}\end{aligned} \tag{16}$$

Thus, the coefficients of the LPA model (6) and (7) gives the estimates of the function and of the corresponding derivatives. This link of the coefficients C with the function and derivative estimation is important for understanding of the LPA.

The idea of the local approximation is applicable not only for the polynomials in the form (7) but also for different polynomials as well as for any basis functions ϕ, which are reasonable for the local fit. In this case, the estimates of the function and the derivatives are defined by the general formulas (13) and (15) and each estimate (function and derivative) can depend on all items of the vector C. The correspondence of the function and derivative estimates with the items of the vector C shown in (14) and (16) is valid only for the polynomials (7).

It deserves to be mentioned that (15) is not a unique definition of the derivative estimate. The estimate of the derivative can be defined as the corresponding derivative of the function-estimate $\hat{y}_h$. Then

$$\hat{y}_h^{(k)}(x) = \frac{\partial^{|k|}}{\partial x_1^{k_1}\partial x_2^{k_2}}\hat{y}_h(x). \tag{17}$$

In general, the estimates (15) and (17) can be quite different at least the derivatives of the window function w appear in (17) while they do not appear in (15) [14,16].

3.2. Nonhomogeneous kernel estimates

According to (9) the coefficients C_k, $k = 1, \ldots, M$, have to be found as a solution of the following quadratic optimization problem:

$$\hat{C}(x,h) = \arg \min_{C \in R^M} J_h(x). \tag{18}$$

In this notation the dependence of the solution $\hat{C}(x,h)$ on the scale h and the variable x is emphasized.

The solution is in the form

$$\hat{C}(x,h) = \Phi_h^{-1} \sum_{v \in X} w_h(x-v)\phi(x-v)z(v),$$

$$\hat{C}(x,h) = \left(\hat{C}_1(x,h), \ldots, \hat{C}_M(x,h)\right)^T, \tag{19}$$

$$\Phi_h = \sum_{v \in X} w_h(x-v)\phi(x-v)\phi^T(x-v), \tag{20}$$

provided that $\det \Phi_h \neq 0$. If the matrix Φ_h is singular a pseudoinverse $\Phi_h^{\#}$ can be used for solution of (18).

Substituting $\hat{C}(x,h)$ (19) into (13) and (15) instead of C we have the function and the derivative estimates in the kernel form:

$$\hat{y}_h(x) = \sum_{v \in X} g_h(x,v)z(v),$$

$$g_h(x,v) = w_h(x-v)\phi^T(x-v)\Phi_h^{-1}\phi(0), \tag{21}$$

and

$$\hat{y}_h^{(k)}(x) = \sum_{v \in X} g_h^{(k)}(x,v)z(v),$$

$$g_h^{(k)}(x,v) = (-1)^{|k|} w_h(x-v)\phi^T(x-v)\Phi_h^{-1}\phi^{(k)}(0), \tag{22}$$

where for the polynomials (7)

$$\phi(0) = [1, 0, \ldots, 0, \ldots, 0]^T \tag{23}$$

is a zero vector-column $M \times 1$ with only 1th elements equal to 1, and

$$\phi^{(k)}(0) = [0, \ldots, 0, \underbrace{1}_{k_1,k_2\text{th}}, 0, \ldots, 0]^T \tag{24}$$

is also a zero vector-column $M \times 1$ with the only element equal to 1 corresponding to location of the polynomial $x_1^{k_1}x_2^{k_2}/k_1!k_2!$ in the vector ϕ.

Thus, the LPA estimates are presented in the form of the linear filters (21) and (22), where for the estimates $\hat{y}_h^{(k)}$ and for the kernels $g_h^{(k)}$ the subindex h indicates a dependence on this important scale parameter.

It is assumed in (2) that X is the regular grid. However, the estimates in the form (18)–(22) are quite universal. They can be applied to any data given on regular or irregular grids,

in particular, to data with lost observations and for data interpolation when the centre x of the LPA does not belong to the grid of the observations X.

It is assumed in the above formulas that the summations $\sum_{v\in X}$ is performed within boundaries of the image support X. It means an accurate LPA fitting inside of these boundaries. There is no boundary problems for these estimates typical for convolution estimates considered latter.

The pointwise LPA estimates as it is in (21) and (22) insure the reproducing properties of the estimates with respect to the polynomial components of y, i.e., for any polynomial y_m with the power less or equal to m the estimates of the function and the derivatives are accurate:

$$\sum_{v\in X} g_h(x,v)y_m(v) = y_m(x), \qquad \sum_{v\in X} g_h^{(k)}(x,v)y_m(v) = \frac{\partial^{|k|}y_m(x)}{\partial x^k}, \quad x\in R^2. \tag{25}$$

Concerning the terminology we note that in statistics the weights $g_h(x,v)$ and $g_h^{(k)}(x,v)$ are named "kernels" and the estimates (21) and (22) are "kernel estimates." The term "bandwidth" is used in statistics for the window size (scale) parameter h. In image processing the term "mask" is commonly used for the weights $g_h(x,v)$ and $g_h^{(k)}(x,v)$.

If the window w has a finite support. For example $w(x) = 0$ if $\|x\| = \sqrt{x_1^2 + x_2^2} > 1$. Then $w_h(x) = 0$ for $\|x\| > h$. Thus, the parameter h defines the window size as well as the support of the masks $g_h^{(k)}(x,v)$ and $g_h(x,v)$. The mask with a finite support defines a finite impulse response (FIR) linear $2D$ filter.

3.3. Homogeneous kernel estimates

In this paper we mainly concern in a special case of the LPA estimate when the kernels are homogeneous shift-invariant depending on the difference of the arguments x and v only. Then, the estimates can be presented as convolutions of these kernels and $2D$ image data.

It happens if the grid X in the formulas for g_h and $g_h^{(k)}$ is regular infinite, $X = \{x_1 = s_1\Delta,\ x_2 = s_2\Delta,\ s_1, s_2 \in \mathbb{Z}\}$, where $\mathbb{Z}$ is a set of integers. In this case the matrix Φ_h in (21) and (22) dependents only on the difference $x - v$ and the kernels defined by the formulas

$$g_h(x,v) = g_h(x-v), \qquad g_h^{(k)}(x,v) = g_h^{(k)}(x-v), \quad x, v \in X.$$

Then, the estimates (21) and (22) can be represented as the convolutions:

$$\hat{y}_h(x) = \sum_v g_h(x-v)z(v) = \sum_v g_h(v)z(x-v), \tag{26}$$

$$\hat{y}_h^{(k)}(x) = \sum_v g_h^{(k)}(x-v)z(v) = \sum_v g_h^{(k)}(v)z(x-v), \tag{27}$$

with the shift-invariant kernels

$$g_h(x) = w_h(x)\phi^T(x)\Phi_h^{-1}\phi(0), \tag{28}$$

$$g_h^{(k)}(x) = (-1)^{|k|}w_h(x)\phi^T(x)\Phi_h^{-1}\phi^{(k)}(0), \tag{29}$$

$$\Phi_h = \sum_v w_h(v)\phi(v)\phi^T(v). \tag{30}$$

Hereafter $\sum_v$ means the double sum $\sum_{s_1=-\infty}^{\infty}\sum_{s_2=-\infty}^{\infty}$ over the infinite regular $2D$ grid with $v_1 = s_1\Delta$ and $v_2 = s_2\Delta$.

The convolutions (26) and (27) assume that the observations $z(x)$ defined on the finite grid (2) are completed by zeros (zero padded) outside of this finite grid for the infinite regular grid.

In what follows the conventional compact notation is used for the convolutions (26) and (27)

$$\hat{y}_h(x) = (g_h \circledast z)(x), \tag{31}$$

$$\hat{y}_h^{(k)}(x) = \big(g_h^{(k)} \circledast z\big)(x). \tag{32}$$

The kernels (28) and (29) satisfy to the polynomial vanishing moment conditions following from (25):

$$\sum_v g_h(v)v^r = \delta_{|r|,0}, \quad 0 \leqslant |r| \leqslant m, \tag{33}$$

$$\frac{1}{r!}\sum_v g_h^{(k)}(v)v^r = (-1)^{|k|}\delta_{k_1,r_1}\cdot\delta_{k_2,r_2}, \quad 0 \leqslant |k| \leqslant m,\ 0 \leqslant |r| \leqslant m. \tag{34}$$

The multi-indexes notation means here that $k = (k_1, k_2)$, $r = (r_1, r_2)$, and $v^r = v_1^{r_1}\cdot v_2^{r_2}$, $|r| = r_1 + r_2$, $1/r! = (1/r_1!)(1/r_2!)$.

The vanishing moment conditions define the polynomial smoothness of the kernels. The support of the kernels is identical to the support of the window w_h.

The first condition (33) means that (26) is a smoothing operator of the order m. The second condition means that the kernel $g_h^{(k)}(x)$ defines the differentiating operator (27) of the order m giving the estimate of the derivative $\partial^{k_1+k_2}/\partial x_1^{k_1}\partial x_2^{k_2}$. The both smoothing and differentiating operators give the accurate results for any polynomial y of the power less or equal to the order of the kernels m.

3.4. Integral homogeneous estimates

Provided that the sampling interval Δ and the scale h are small, such that $\Delta, h \to 0$ and $h/\Delta \to \infty$, the discrete convolutions (31) and (32) are transformed to the corresponding integral forms

$$\hat{y}_h(x) = \frac{1}{h^2}\int_{R^2} g\big((x-u)/h\big)y(u)\,du = \int_{R^2} g(u)y(x-hu)\,du, \tag{35}$$

$$\begin{aligned}\hat{y}_h^{(k)}(x) &= \frac{1}{h^{2+|k|}}\int_{R^2} g^{(k)}\big((x-u)/h\big)y(u)\,du \\ &= \frac{1}{h^{|k|}}\int_{R^2} g^{(k)}(u)y(x-hu)\,du, \quad x \in R^2,\end{aligned} \tag{36}$$

with the kernels

$$g(x) = w(x)\phi^T(x)\Phi^{-1}\phi(0), \tag{37}$$

$$g^{(k)}(x) = (-1)^{|k|}w(x)\phi^T(x)\Phi^{-1}\phi^{(k)}(0), \tag{38}$$

$$\Phi = \int_{R^2} w(x)\phi(x)\phi^T(x)\,du, \tag{39}$$

where $u = (u_1, u_2)$, $\int_{R^2}(\cdot)\,du = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(\cdot)\,du_1\,du_2$.

Some extra technical assumptions are required for the existence of the above integrals and justification of the corresponding limit passages from the sums (31) and (32) to the integrals (35) and (36). They are satisfied, in particular, if the window w is a bounded finite support function and $y(x)$ is continuous.

For the integral kernels the vanishing moment conditions (33) and (34) are as follows:

$$\int_{R^2} g(x)x^r\,dx = \delta_{|k|,0}, \quad 0 \leqslant |r| \leqslant m, \tag{40}$$

$$\frac{1}{r!}\int_{R^2} g^{(k)}(x)x^r\,dx = (-1)^{|k|}\delta_{k_1,r_1}\cdot\delta_{k_2,r_2}, \quad 0 \leqslant |k| \leqslant m,\ 0 \leqslant |r| \leqslant m. \tag{41}$$

3.5. Restricted nonlinear LPA estimates

An 8 bit gray-scale image is defined by its intensity function taking $2^8 = 256$ different values. This intensity is nonnegative and takes values from 0 up to 255. After normalization these conditions have a form

$$0 \leqslant y(x) \leqslant 1. \tag{42}$$

Thus, y is an arbitrary nonnegative function normalized to the interval $[0, 1]$. These conditions can be naturally incorporated in the LPA estimate by modifying (18) to the constrained optimization:

$$\hat{C}(x,h) = \arg \min_{\substack{0\leqslant C_1\leqslant 1,\\ \underline{C}\in R^{M-1}}} J_h(x), \tag{43}$$

where $\underline{C} = (C_2, \ldots, C_M)^T$ is the vector C with the excluded first item C_1.

The constrains (42) according to (14) can be imposed only on C_1 while all other items of C continue to be unconstrained. After the estimates of C are found from (43) the intensity and the derivatives are calculated according to the usual formulas (14) and (16).

Concerning the estimate (43) a number of moments can be noted. First, in general, this estimate is nonlinear with respect to the observations. Its calculation is a much more complex problem as compared with the linear estimate. However, if the linear estimate $\hat{C}_1(x,h)$ found from (18) belongs to the interval $[0, 1]$ then the solutions of (18) and (43) are identical. It gives a simple logic to deal with this nonlinear problem. We start from calculation of the linear estimates for all x and test (42). In this way we identify the pixels

violating the conditions (42) and the nonlinear constrained estimates (43) are calculated only for these pixels.

Second, if the linear estimate does not satisfy to (42) the solutions of the unconstrained and constrained optimizations can be different by all elements of the vector C not only by C_1. The constrains on the intensity function can influence the estimates of the derivatives and enable one to yield both the more accurate estimate of the function as well as the derivatives.

The following compromise is used in order to avoid the complexity of (43). The estimates are obtained in two independent steps. The first step is a solution ignoring the constrains on C_1. The second step defines the estimates according to the equations:

$$\hat{y}_h(x) = \big[(g_h \circledast z)(x)\big]_+, \tag{44}$$

$$\hat{y}_h^{(k)}(x) = \big(g_h^{(k)} \circledast z\big)(x), \tag{45}$$

where $[\cdot]_+$ stays for the projection on the segment $[0, 1]$, which means $[x]_+ = x$ for $0 < x \leqslant 1$, $[x]_+ = 0$ for $x \leqslant 0$, and $[x]_+ = 1$ for $x > 1$. Thus, the estimate of C_1 obtained from the unconstrained optimization is replaced by its projection on the interval $[0, 1]$. All others elements of the vector C are assumed to be equal to the corresponding items of the vector $\hat{C}(x, h)$.

The estimate built in this way can be treated as an approximation of the accurate nonlinear constrained solution (43). It is a conventional practice in image processing to take into consideration the nonnegativity and upper bound of the image intensity by the simple projection of the estimate on the interval $[0, 1]$.

In what follows for the sake of simplicity we use the linear estimate given as the convolutions (26) and (27).

4. LPA accuracy

4.1. Asymptotic bias and variance

A value of the scale h is a crucial point in the efficiency of the local estimators (e.g., [7,16,19,30]). When h is relatively small, the LPA gives a good smooth fit of functions but then fewer data are used and the estimates are more variable and sensitive with respect to the noise. The best choice of h involves a tradeoff between the bias and variance of the estimate. In order to clarify and formalize the meaning of this tradeoff we present some accuracy results.

In what follows in this chapter we present the accuracy results in the terms of the kth derivative estimation as the function estimation is a special case with $k = 0$. The estimation error is a difference between the true $y^{(k)}$ and the estimate $\hat{y}_h^{(k)}$:

$$e_{y^{(k)}}(x, h) = y^{(k)}(x) - \hat{y}_h^{(k)}(x).$$

This error is composed from the systematic (bias) and random components corresponding to the deterministic y and the random noise ε. We have, respectively, for the bias

$$m_{y^{(k)}}(x, h) = y^{(k)}(x) - E\big\{\hat{y}_h^{(k)}(x)\big\},$$

and for the variance

$$\sigma^2_{y^{(k)}}(x,h) = E\{[y^{(k)}(x) - E\{\hat{y}^{(k)}_h(x)\}]^2\},$$

where $E\{\cdot\}$ denotes the mathematical expectation calculated over ε.

The asymptotic formulas for $m_{y^{(k)}}(x,h)$ and $\sigma^2_{y^{(k)}}(x,h)$ can be given in the integral form with the analytical dependence on the scale h. This sort of results are basic for theoretical accuracy analysis and scale optimization.

Hypotheses assumed:

(H1) y is continuous and r-differentiable at the point x such that $y \in F_r(\bar{L}_r)$ (4).
(H2) The window w is finite support continuous.

Let us use the notation

$$M = \min\{m+1, |r|\}, \tag{46}$$

where m is the order of the LPA and r is a multiple-index defining the smoothness (derivative order) of y in the class F_r.

Let the kernel LPA estimates be defined by (31) and (32), where $|k| < M$, and the sampling interval and the scale parameter be small such that Δ, $h \to 0$, and $h/\Delta \to \infty$.

Provided that the hypotheses (H1) and (H2) hold and the derivatives $y^{(r)}(x)$ in (4) are continuous Lipshitz functions

$$|y^{(r)}(x) - y^{(r)}(y)| \leqslant L\|x-y\|^{\gamma}, \quad \gamma > 0, \tag{47}$$

the bias and the variance of the estimate $\hat{y}^{(k)}_h(x)$ are defined by the formulas

$$m_{y^{(k)}}(x,h) = (-1)^{M-1} h^{M-|k|} \sum_{|r|=M} y^{(r)}(x) \cdot \frac{1}{r!} \int_{R^2} u^r g^{(k)}(u)\,du + o(h^{M-|k|}), \tag{48}$$

$$\sigma^2_{y^{(k)}}(x,h) = \sigma^2 \frac{\Delta^2}{h^{2+2|k|}} \int_{R^2} \left(g^{(k)}(u)\right)^2 du + o\left(\frac{\Delta^2}{h^{2+2|k|}}\right), \tag{49}$$

where $g^{(k)}$ is given in (38) and a small $o(x)$ means that $o(x)/x \to 0$ as $x \to 0$.

The derivation of these formulas is based on quite routine techniques using, in particular, the multivariable Taylor series for $y(x-hu)$ on h. Hypothesis (H2) and (47) enable the existence of the integral estimates (35) and (36) as well as the corresponding limits and integrals in (48) and (49).

Omitting the small terms $o(\cdot)$, the following inequality can be derived from (48)

$$|m_{y^{(r)}}(x,h)| \leqslant h^{M-|k|} L_M(x) A^{(k)}_e, \quad A^{(k)}_e = \sum_{|r|=M} \frac{1}{r!} \left| \int_{R^2} u^r g^{(k)}(u)\,du \right|, \tag{50}$$

where $\max_{|r|=M} |\partial^r y(x)/\partial x^r| \leqslant L_M(x)$, according to (4), and (49) is rewritten as

$$\sigma^2_{y^{(k)}}(x,h) = \sigma^2 \frac{\Delta^2}{h^{2+2|k|}} B^{(k)}_e, \quad B^{(k)}_e = \int_{R^2} \left(g^{(k)}(u)\right)^2 du. \tag{51}$$

We can see from (50) that the bias error is restricted by $L_M(x)$ what is the absolute value of the maximum Mth order derivatives of $y(x)$. The constant $A_e^{(k)}$ is defined by the moments of the kernel. The scale h is an important parameter for the bias: a small h means a small bias. The upper bound of the bias is restricted by the value of h of the power M for the function estimation and of the power $M-(k_1+k_2)$ for the derivative estimation. For small h it gives $h^{M-k_1-k_2} > h^M$. Thus, the bias of the derivative estimate is asymptotically larger than the bias of the function estimate. Note also that $M = \min\{m+1, |r|\}$ means that while we increase the power m of the LPA the bias becomes smaller. However, this effect is valid until $m+1 < |r|$.

The variance formula (51) shows that the variance for the derivative estimation (of the order $h^{-2(1+k_1+k_2)}$) is larger than the variance for function estimation (of the order h^{-2}). Thus, in terms of the bias error as well as of the level of the random error the derivative estimation is a more complex problem than estimation of the function itself. The higher order derivative (larger k_1 and k_2) automatically means that the estimation errors become larger.

4.2. Scale optimization

The formulas (50) and (51) define the pointwise mean squared risk $r^{(k)}(x,h)$ for the kth derivative estimation as

$$r^{(k)}(x,h) \triangleq E\big(y^{(k)}(x) - \hat{y}_h^{(k)}(x)\big)^2 \leqslant \big(h^{M-|k|} L_M(x) A_e^{(k)}\big)^2 + \sigma^2 \frac{\Delta^2}{h^{2+2|k|}} B_e^{(k)}. \quad (52)$$

The upper bound of $r^{(k)}(x,h)$ is convex on h. Its minimization on h gives the "ideal" value of the scale found from the equation

$$\frac{\partial}{\partial h}\left[\big(h^{M-|k|} L_M(x) A_e^{(k)}\big)^2 + \frac{\Delta^2\sigma^2}{h^{2+|k|}} B_e^{(k)}\right] = 0.$$

Further calculations give for this ideal scale

$$h_k^*(x) = \left(\Delta^2 \frac{\sigma^2 B_e^{(k)}}{(L_M(x) A_e^{(k)})^2} \gamma_k^2\right)^{1/(2M+2)}, \quad \gamma_k^2 = \frac{1+|k|}{M-|k|}, \quad (53)$$

where γ_k is the ratio of the absolute value of the bias to the standard deviation of the derivative estimate at $h = h_k^*(x)$

$$\gamma_k = \big|m_{y^{(k)}}\big(x, h_k^*(x)\big)\big| \big/ \sigma_{y^{(k)}}\big(x, h_k^*(x)\big). \quad (54)$$

This parameter is a constant (invariant on x) depending on $M = \min\{m+1, |r|\}$ and $|k|$ only. Thus, it depends on the LPA power m, the smoothness of the function defined by $|r|$ and the order $|k|$ of the estimated derivative.

It can be verified also that

$$\big|\overline{m}_{y^{(k)}}\big(x, h(x)\big)\big| \begin{cases} < \gamma_k \cdot \sigma_{y^{(k)}}(x,h), & \text{if } h < h_k^*(x), \\ > \gamma_k \cdot \sigma_{y^{(k)}}(x,h), & \text{if } h > h_k^*(x). \end{cases} \quad (55)$$

It shows that $\gamma_k \cdot \sigma_{y^{(k)}}(x,h)$ is a critical value for the ideal bias. For $h < h_k^*(x)$ the bias is smaller and for $h > h_k^*(x)$ the bias is larger than this critical value. The ICI rule for

the adaptive varying scale selection considered in Section 4.3 is based on testing of the hypotheses: $h \lessgtr h_k^*(x)$.

Let the number of observations (image samples) n and the sampling interval Δ satisfy to the equation $n \approx 1/\Delta^2$, which means that the physical size of the image is 1×1. By substituting $h_k^*(x)$ from (53) into (52) we obtain the ideal pointwise mean squared risk. To be interested only in the order of this risk for large n we immediately can see that

$$\sqrt{r^{(k)}\big(x, h_k^*(x)\big)} = O\big(n^{-(M-|k|)/(2M+2)}\big). \tag{56}$$

This risk approaches zero for a large number of observations and proves the mean square convergence of the studied estimates. The convergence rate is defined by the parameter

$$\psi = \frac{M-|k|}{2M+2} = \frac{\min\{m+1,|r|\}-|k|}{2\min\{m+1,|r|\}+2}, \quad |k| < M. \tag{57}$$

It is seen from (57) that larger values of M result in the higher convergence rate which is restricted by the limit value $O(n^{-1/2})$ for large M.

The converge rate (56) is an ideal one and cannot be achieved in practice as it requires a knowledge of the derivatives of the estimated function for the ideal varying scale selection. However, it gives an useful information on the potential accuracy which can be used as a benchmark for evaluation of algorithms.

Assuming $k = 0$ the corresponding results can be obtained from the above formulas for the function estimation. The ideal scale is

$$h_0^*(x) = \left(\Delta^2 \frac{\sigma^2 B_e^{(0)}}{(L_M(x) A_e^{(0)})^2} \gamma^2\right)^{1/(2M+2)}, \quad \gamma^2 = \frac{1}{M}, \tag{58}$$

and the convergence rate following from (56) is defined as

$$\sqrt{r^{(0)}\big(x, h_0^*(x)\big)} = O\big(n^{-M/(2M+2)}\big).$$

The convergence rate of the derivative estimates is slower than that for the function estimation.

4.3. Adaptive scale selection

The problem of the scale (window size) selection is always solved in favor of the larger scale for any parametric estimate as this estimate is unbiased for any scale and the larger scale means a smaller variance. For the nonparametric estimation the situation is more complex. If there is no noise the scale should be selected as small as possible since a smaller scale means a smaller bias. However, if there is a noise the scale should be increased in order to suppress noise effects. The accuracy analysis produced in Section 4.2 confirms and illustrates this claim. It demonstrates that there is the ideal scale which defines the optimal mean squared balance between the deterministic bias errors and the variance of random errors.

In many signal processing applications this nonparametric nature of the local approximation is ignored. It is assumed that the polynomial model is accurate, noise effects are neglected and the window is taken of the minimum size sufficient for the polynomial fit.

For instance, we may mention most conventional differentiators as well as methods based on the fit by the orthogonal polynomials (e.g., [8,12]).

A number of publications concerning the scale selection is very large and growing quickly. While a review of the field is beyond the scope of this paper, we give few references illustrating the basic progress in various directions.

Two approaches have been first exploited for adaptive (data-driven) invariant scale selection of the LPA estimates. The first one is based on estimation of the biasedness $m_y(x,h)$ and the variance $\sigma_y^2(x,h)$ of the estimates with the ideal scale calculation according to theoretical formulas. However, this bias depends on the derivatives of the signal in question. Thus, in order to find the adaptive h, say from the formula (58), one needs estimate these derivatives. This sort of methods, known as "pilot estimates," are quite complex in implementation and have a few design parameters. Nevertheless, successful methods have been developed based on these ideas and reported by several authors [7,30,36,37].

The second alternative approach, also for the adaptive invariant scale selection, does not require estimation of the bias. These group of methods are based on the quality-of-fit statistics such as cross-validation, generalized cross-validation, C_p, Akaike criteria, etc., which are applied for direct optimization of the accuracy (e.g., [7,13,16,30] and references herein).

The linear LPA with the varying scale found by minimization the so-called "pseudo-mean squared error" is considered in [32]. The target point is left out of the averaging in the pseudo-mean squared error what differs this estimate from the standard mean square methods. It is reported that the proposed pseudo-mean squared error works better then the local cross-validation.

A recent break-through in pointwise varying scale estimation adaptive to unknown smoothness of the function is originated from a general scheme of Lepski [28,29,38] already mentioned in the introduction. The LPA estimates are calculated for a grid of scales and compared. The adaptive scale is defined as the largest of those scales in the grid which estimate does not differ significantly from the estimators corresponding to the smaller scale. These type methods first proposed in few cited above papers for $1D$ nonparametric estimation are mainly differ in (1) grid of window sizes (scales), (2) accuracy criteria, (3) statistics used for scale selection.

The intersection of confidence intervals (ICI) rule can be treated as quite a different implementation of the basic Lepski's approach idea [9,17]. It is proved that the LPA equipped with the ICI rule for the scale selection "possesses simultaneously many attractive asymptotic properties, namely, (1) it is nearly ideal within $\ln n$-factor for estimation of a function (or its derivative) at a single point; (2) it is spatial adaptive in the sense that its quality is close to that one which could achieve if smoothness of the underlying function was known in advance; (3) it is optimal in order or nearly-optimal in order within $\log n$-factor for estimating whole function (or its derivatives) over wide range of classes and global loss functions" [9]. These results demonstrate that the "spatial adaptive abilities" of these estimates are the best possible in the terms of the asymptotic analysis. These results have been extended to multivariable functions provided that the scale parameter of the estimate is scalar [33]. Similar asymptotic accuracy results are proved for different versions of Lepski's algorithm.

Experimental study reveals that a nonasymptotic performance of the ICI rule depends essentially on the threshold parameter the confidence intervals used in the ICI algorithm [17]. It is shown that the cross-validation is able to give adaptive values of the threshold parameter improving the estimation accuracy. Various modifications of the ICI rule are appeared to be efficient for different scale adaptive applications: median filtering [20], beamforming [21], time-frequency analysis [4,22]. An application of the ICI rule to image denoising and deconvolution has been reported in [18,19]. This development of the ICI for the $2D$ image intensity function exploits the $2D$ quadrant (symmetric and nonsymmetric) windows with adaptive varying scale parameters.

A special version of the Lepski's spatially adaptive method is proposed and analyzed in [38]. First of all, a set of test-windows is proposed, which enables a fine cover of a neighborhood of the estimation point. Further, the used test-statistics are based on the residual of estimation, while the original Lepski's algorithms use the function estimates only. The accuracy analysis produced in [38] for estimation at far and near change points shows that the estimates are nearly optimal within the usual $\log n$-factor unavoidable for the adaptive estimation convergence rate.

A $2D$ generalization of the algorithm from [38] is proposed for image denoising in [35]. It is assumed that the image intensity is an unknown piece-wise constant function. The estimate is a sample mean calculated in the sliding varying adaptive size window. The main algorithmic novelty concerns the $2D$ window sets and test-statistics design. For the test-statistics the differences between the estimate in the tested window and the estimates in some subwindows of the tested window are calculated. Near optimal estimation accuracy is proved for the pixels far and near an edge.

Basic papers concerning the Lepski's approach are published mainly in mathematical statistics journals and concentrated on theoretical analysis of the accuracy and adaptivity properties for various classes of the functions to be estimated. Some recent results concerning a development of the adaptive scale multivariable estimation can be seen in [26] where the optimal kernels are derived for different classes of functions to estimate.

To complete these introductory notes we wish to mention a new development concerning a generalization of the approach to multivariable kernel estimates, $x \in R^d$, using different scales for these variables [25]. The adaptation becomes anisotropic and assumes selection of the multivariable scale parameter $h \in R^d$. This generalization is of special interest as there are some principal obstacles for this sort of multivariable adaptivity.

4.4. Lepski's approach

Let us start from the idea of the Lepki's approach. Introduce a set of the scales

$$\tilde{H} = \{h_1 < h_2 < \cdots < h_J\}, \tag{59}$$

starting from a small h_1 and increasing to a maximum h_J, and let $\hat{y}_h$ be the estimate of y defined for $h \in \tilde{H}$ with the estimation standard deviation $\sigma_y(x, h)$. Accordingly to the accuracy analysis produced above, for small h the estimate $\hat{y}_h$ has a small bias and a large standard deviation of the random noise. The adaptive scale algorithm compares the estimates with increasing h. The main intention is to find the maximum scale when the estimate's deviation can be explained by the random component of the estimation error

and there is a balance between the biasedness and randomness in the estimate. The accurate meaning of this sort of balance is discussed above.

The Lepski's approach defines the adaptive scales according to the conditions

$$\hat{h}(x) = \max\left\{h \in \tilde{H}\colon \left|\hat{y}_h(x) - \hat{y}_\eta(x)\right| \leqslant T(h,\eta,x) \text{ for all } \eta < h,\ \eta, h \in \tilde{H}\right\}, \tag{60}$$

where $T(h,\eta,x)$ is a given threshold. The procedure (60) is looking for a largest scale h in order to obtain the maximum smoothing effect for the random errors. However, a large h can result in a significant bias error. All estimates $\hat{y}_\eta(x)$ with the scale $\eta < h$ are compared with the estimate $\hat{y}_h(x)$ of the scale h. If the differences $|\hat{y}_h(x) - \hat{y}_\eta(x)|$ can be explained by the random errors the bias is not large and larger h can be selected. The adaptive scale $\hat{h}$ is defined as maximum in $\tilde{H}$ such that all estimates $\hat{y}_\eta(x)$ with $\eta < h$ are not too different from $\hat{y}_h(x)$. The multiple comparison of the estimates with the different scales is used. The parameter $T(h,\eta,x)$ is a key element of the algorithm as it says when the difference between the estimates is large or small. The procedure (60) enables a multiple statistical test on significance of the systematic error in the differences $\hat{y}_h(x) - \hat{y}_\eta(x)$ in comparison with the corresponding random errors.

A variety of the Lepski's algorithms is defined mainly by the different form of the threshold $T(h,\eta,x)$ which usually depends on the variances of the estimates with the scales h and η. A proper selection of this threshold enables nice statistical properties of these adaptive scale estimate [27–29,38].

Let us describe two different algorithms of the Lepski's class in order to illustrate the approach overall and in order to show that these algorithms are indeed simple in implementation. We also use these algorithms for a reference and further presentation of the novel MR local regression.

4.5. Lepski–Spokoiny algorithm [29]

The adaptive scale $h^+(x)$ is defined as follows:

$$\begin{aligned} &i^+ = \max\left\{i\colon \left|\hat{y}_{h_i}(x) - \hat{y}_{h_j}(x)\right| \leqslant \Gamma_1(h_j)\sigma_y(x,h_j),\ j < i,\ 1 \leqslant i \leqslant J\right\}, \\ &h^+(x) = h_{i^+}, \end{aligned} \tag{61}$$

with the adaptive scale estimate

$$\hat{y}^+(x) = \hat{y}_{h^+(x)}(x),\ h^+(x) = h_{i^+}. \tag{62}$$

Here $\hat{y}_{h_i}(x)$ compared with all estimates having $h_j < h_i$ and in this comparison the inequality in (61) is tested. As in (60) the adaptive scale $h^+(x)$ is equal to the maximum $h_{i^+} \in \tilde{H}$ satisfying all of the corresponding inequalities in (61). In this case the threshold $T(h,\eta,x)$ from (60) is used in the form $T(h_i,h_j,x)$, as according to (61) $h_j < h_i$ and $T(h_i,h_j,x) = \Gamma_1(h_j)\sigma_y(x,h_j)$.

The grid $\tilde{H}$ for (61) is defined inductively starting from the largest h_J by

$$h_{J-k} = \frac{h_{J-k+1}}{1 + \alpha(h_{J-k+1})}, \quad k = 1, 2, \ldots, J-1, \tag{63}$$

$$d(h) = \sqrt{\max\left(1, r\lg(h_j/h)\right)}, \qquad \alpha(h) = \frac{1}{\sqrt{d(h)}}. \tag{64}$$

The threshold $\Gamma_1(h_j)$ in (61) depending on h is as follows:

$$\Gamma_1(h) = \big(1 + \alpha(h)\big)d(h). \tag{65}$$

It is proved in [29] for some asymptotic consideration that the algorithm give the adaptive window sizes which minimizes the risk $E\{|\hat{y}(x, h_i) - y(x)|^r\}$, $r \geqslant 1$. The r is a parameter used in (64). Note that the total number of compared scales is of the logarithmic order and depends on the maximum h_J.

4.6. ICI algorithm [9,17,19]

Being from the Lepski's class this algorithm is derived from different speculations and has quite a different recursive structure.

Determine a sequence of the confidence intervals Q_j of the estimates $\hat{y}_{h_j}(x)$

$$Q_j = \big[\hat{y}_{h_j}(x) - \Gamma \cdot \sigma_y(x, h_j),\ \hat{y}_{h_j}(x) + \Gamma \cdot \sigma_y(x, h_j)\big], \tag{66}$$

where Γ is a threshold parameter.

Consider the intersection of the intervals Q_j, $1 \leqslant j \leqslant i$, with increasing i, and let i^+ be the largest of those i for which the intervals Q_j have a point in common. This i^+ defines the adaptive scale and the adaptive LPA estimate as given by (62).

The following algorithm implements the ICI rule. Determine the sequence of the upper and lower bounds of the confidence intervals Q_j as follows:

$$Q_j = [L_j, U_j], \quad U_j = \hat{y}_{h_j}(x) + \Gamma \cdot \sigma_y(x, h_j), \quad L_j = \hat{y}_{h_j}(x) - \Gamma \cdot \sigma_y(x, h_j), \tag{67}$$

and let

$$\begin{aligned} &\bar{L}_{j+i} = \max\{\bar{L}_j, L_{j+1}\}, \qquad \underline{U}_{j+1} = \min\{\underline{U}_j, U_{j+i}\}, \\ &\quad j = 1, 2, \ldots, J,\ \bar{L}_1 = L_1,\ \underline{U}_1 = U_1. \end{aligned} \tag{68}$$

According to these formulas $\bar{L}_{j+1}$ is a nondecreasing sequence and $\underline{U}_{j+1}$ is a nonincreasing sequence. Find the largest j when

$$\bar{L}_j \leqslant \underline{U}_j, \quad j = 1, 2, \ldots, J, \tag{69}$$

is still satisfied. Denote this largest value as i^+. This i^+ is the largest of those j for which the confidence intervals Q_j have a point in common as it is discussed above and the ICI adaptive scale is $h^+ = h_{i^+}$. It is a procedure for a fixed x giving the varying adaptive scale $h^+(x)$. Figure 1 illustrates this algorithm.

In the ICI algorithm the estimates of the different scale are compared by using their confidence intervals. We may conclude that the confidence intervals Q_i and Q_j intersect if and only if

$$\big|\hat{y}_{h_i}(x) - \hat{y}_{h_j}(x)\big| \leqslant \Gamma\big(\sigma_y(x, h_i) + \sigma_y(x, h_j)\big).$$

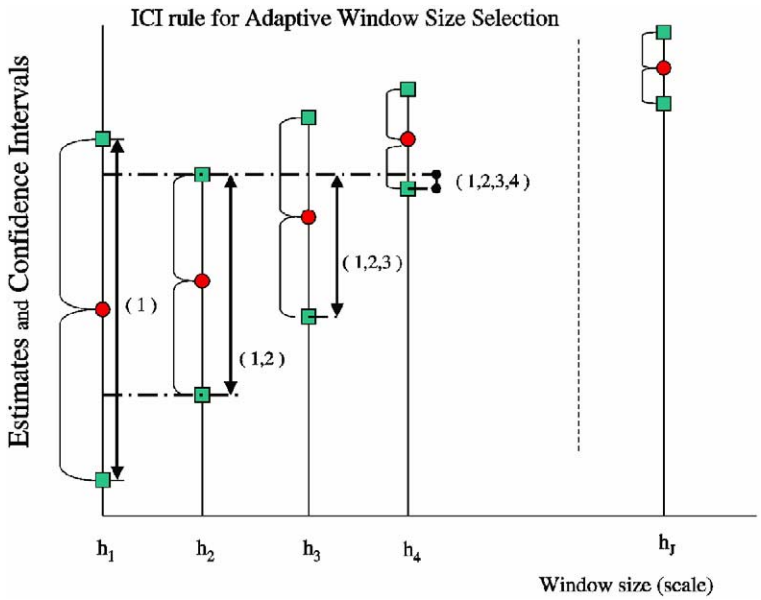

Fig. 1. Intersection of confidence intervals (ICI) rule for the adaptive varying scale selection. The confidence intervals are shown for $h_j \in H$. The vertical lines with arrows show the successive intersections of the confidence intervals $(1, 2)$, $(1, 2, 3)$ and $(1, 2, 3, 4)$. Assuming that the intersection with the forth confidence interval (corresponding $h = h_4$) is the last nonempty one we obtain the adaptive scale as equal to h_4.

Then the ICI rule is reduced to Lepski's scheme (60) with

$$T(h, \eta, x) = \Gamma\big(\sigma_y(x, h) + \sigma_y(x, \eta)\big).$$

The theoretical analysis produced in [9] for $1D$ regression shows that the ICI adaptive scale estimate gives the best possible pointwise and global MSE. A generalization of this result for a multidimensional regression with the scalar h is done in [33].

In the asymptotic analysis most of the scale adaptive nonparametric regression algorithms are equivalent in terms of the convergence rate. However, simulation shows that practically the efficiency of the algorithms can be quite different. It deserves to be mentioned that the similar asymptotic properties concerning the convergence rate as well as the classes of adaptivity are known for the wavelet techniques.

Remind that the ideal scales (53) balance the bias-variance tradeoff. This balance depends on $L_m(x)$, i.e., on the derivatives of the order M, and these derivatives are unknown in advance. The order of these derivative $M = \min(m + 1, |r|)$ depends on the parameter r which also is unknown. The ICI rule gives the adaptive scales close to the ideal ones. The confidence intervals Q_j (67) used in the ICI depend on the estimates and the standard deviations $\sigma_y(x, h_j)$ (48) only and do not use the parameter r as well as the derivatives of y. Thus, the ICI rule produces the estimates which are spatially adaptive to unknown varying smoothness of the estimated signal y.

5. MR nonparametric regression

5.1. Nonparametric regression spectrum

Let us introduce a finite set of scales H

$$H = \{h_0 > h_1 > \cdots > h_J\}, \tag{70}$$

starting from a largest h_0 and decreasing to a smallest h_J. Thus, H is a set of the descending scales, while $\tilde{H}$ in (59) is a set of the ascending scales.

We consider the convolution image intensity estimates as it is defined by (31) and assume that for the smallest scale h_J the LPA kernel $g_h(u)$ is the identical operator

$$\lim_{h\to h_J} g_h(x) = \delta_{x,0}. \tag{71}$$

This assumption is not restrictive and only defines a range of scales starting from a large h_0 and going to sufficiently small h_J. For instance, if the window in (28) is an indicator such that $w(x) = 1$ for $|x_1| < 1$, $|x_2| < 1$ then the LPA with $m = 0$ insures that $g_{h_J}(x) = \delta_{x,0}$ for $h_J = 1$.

If the window w is the $2D$ standard Gaussian density, then h in $w(x/h)/h^2$ (28) is the standard deviation of this distribution and the LPA defines the discrete kernel such that $\lim_{h\to 0} g_h(x) = \delta_{x,0}$. In this case $h_J \to 0$.

Further, we assume that all kernels g_h, $h \in H$, have finite supports embedded into a finite rectangular regular grid U. If the support of the kernel is smaller than U the kernel is completed by zeros in order to have $g_h(x)$, $h \in H$, defined for all $x \in U$.

Let us start from a simple decomposition of the function estimate $\hat{y}_h$ (31) in a sum of differences of various scale estimates:

$$\hat{Y}(x, \boldsymbol{\beta}) = \hat{y}_{h_0}(x) + \sum_{j=1}^{J} \beta_j \cdot \nabla\hat{y}_j(x), \quad \nabla\hat{y}_j(x) = (\nabla g_j \circledast z)(x), \tag{72}$$

$$\nabla g_j(x) = g_{h_j}(x) - g_{h_{j-1}}(x) \quad \text{for } j = 1, 2, \ldots, J, \tag{73}$$

where $\nabla\hat{y}_j(x) = \hat{y}_{h_j}(x) - \hat{y}_{h_{j-1}}(x)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)^T$ is a vector of coefficients.

Consider the items in the right hand-side of (72). The difference $\nabla\hat{y}_j$ is a deviation of the estimate caused by a decrement of h from h_{j-1} to h_j. The largest values $h = h_0$ means a coarser scale and strong smoothing kernel with $\hat{y}_{h_0}(x) = (g_{h_0} \circledast z)(x)$ being the smoothest estimate giving the low frequency picture of y. Smaller h corresponding to a finer scale detecting higher frequency details in the image. In the sum (72) the first term $\hat{y}_{h_0}$ presents the smoothed background of the image while the others add details of different scales. In this way (72) is a decomposition of the image y in the different scale components.

Let ∇g_{h_j} be a spectral (scale) analysis kernel and $\hat{Y}(x, \boldsymbol{\beta})$ be a spectral (scale) decomposition.

The following properties are easy to verify.

(1) According to (33) the analysis kernels, $\nabla g_j(x)$, $j = 1, \ldots, J$, have vanishing moments up to the order m

$$\sum_{x\in U} \nabla g_j(x) x^k = 0, \quad 0 \leqslant |k| \leqslant m, \ j = 1, \ldots, J. \tag{74}$$

Thus, the analysis kernels have the polynomial smoothness defined by the power m of the LPA.

(2) The sum of the analysis kernels ∇g_j assuming $\nabla g_0 = g_{h_0}$ is the identical operator

$$\sum_{j=0}^{J} \nabla g_j(x) = g_{h_J}(x) = \delta_{x,0}. \tag{75}$$

(3) For any $h \in H$ the estimate $\hat{y}_h$ can be represented in the form

$$\hat{y}_h(x) = \hat{Y}\big(x, \boldsymbol{\beta}(h)\big), \tag{76}$$

where the coefficients $\beta_j(h)$ in (72) are defined by the indicator function

$$\beta_j(h) = 1[h_j \geqslant h]. \tag{77}$$

(4) For $\beta_j = 1$ for $1 \leqslant j \leqslant J$, a perfect reconstruction of y has a place

$$\hat{y}_{h_0}(x) + \sum_{j=1}^{J} \nabla \hat{y}_j(x) = y(x). \tag{78}$$

Equation (76) is verified substituting β_j given by (77) in (72). Thus, a varying h in (77) controls a number of spectral items in the expansion (72) and in this way it varies $\hat{Y}(x, \boldsymbol{\beta}(h_J))$ from the perfect reconstruction of the observed $y(x) = \hat{Y}(x, \boldsymbol{\beta}(h_J))$ to the most smoothed estimate $\hat{y}_{h_0}(x) = \hat{Y}(x, \boldsymbol{\beta}(h_0))$.

The problem of the adaptive scale selection for the estimate $\hat{y}_h(x)$ can be formulated as selection of h in $\boldsymbol{\beta}(h)$ (77) for the estimate (72). For h invariant on x (77) gives the same scale selection for all x while h dependent on x results in the pointwise varying scale

$$\beta_j\big(h(x)\big) = 1\big[h_j \geqslant h(x)\big]. \tag{79}$$

5.2. Multiresolution analysis

Developing further the concept of the local regression spectrum we assume that the coefficients β_j in (72) may be not binary. Then, we arrive to the idea of nonparametric estimation of y based on the spectral decomposition (72) with some estimates of the spectral coefficients β_j. In this way we break with the traditional statistical approach to local regression assuming that h in $\hat{y}_h$ is the only scale parameter defining the estimate.

In order to make this approach more constructive we replace the initial spectral analysis kernels ∇g_j by their orthogonal counterparts.

Let the kernels $\nabla g_j(x)$, $j = 0, 1, \ldots, J$, $x \in U$, be a set of $(J+1)$ linear independent functions with the bounded Euclidean norms, $\|\nabla g_j\|^2 = \sum_{x \in U} (\nabla g_j(x))^2 < \infty$.

Then, the standard Gram–Schmidt procedure gives $\omega_{j+1}(x)$ orthogonal with respect to $\omega_k(x)$, $0 \leqslant k \leqslant j$, for $x \in U$ as follows:

$$\begin{aligned} &\omega_{j+1}(x) = \nabla g_{h_{j+1}}(x) - \sum_{k=0}^{j} \frac{\langle \nabla g_{h_{j+1}}, \omega_k \rangle}{\|\omega_k\|^2} \omega_k(x), \\ &\omega_0(x) = g_{h_0}, \quad j = 0, \ldots, J-1, \end{aligned} \tag{80}$$

where the inner product $\langle \cdot, \cdot \rangle$ means $\langle \nabla g_h, \omega_k \rangle = \sum_{x \in U} \nabla g_h(x) \omega_k(x)$ and $\|\omega_k\|^2 = \langle \omega_k, \omega_k \rangle = \sum_{x \in U} \omega_k^2(x)$.

Replace these ω_j by the normalized kernels $\omega_j / \|\omega_j\|$. Then the orthonormal vector-functions $\omega(x) = (\omega_0(x), \omega_1(x), \ldots, \omega_J(x))^T$ and $\nabla g(x) = (\nabla g_{h_0}(x), \nabla g_{h_1}(x), \ldots, \nabla g_{h_J} \times (x))^T$ satisfy to the Gram–Schmidt equation

$$Q\omega(x) = \nabla g(x) \quad \text{for all } x \in U, \tag{81}$$

where $Q = (Q_{j,k})_{j,k=0,1,\ldots,J}$ is a nonsingular lower triangular $(J+1) \times (J+1)$ matrix.

Let α_j be the output of the filter with the kernel ω_j and the input is the accurate signal y. Then

$$\alpha_j(x) = (\omega_j \circledast y)(x) \tag{82}$$

and the following can be verified:

(1) The outputs α_j define a spectral analysis with components varying from a low frequency (coarse scale) base image α_0 to higher frequency (finer scale) image increments α_j. Higher value j corresponds to a higher frequency spectral component of the spectrum.
(2) The spectral kernels have vanishing moments up to the order m

$$(\omega_j \circledast x^k)(0) = 0, \quad 0 < |k| \leqslant m, \ j = 1, \ldots, J. \tag{83}$$

Note that contrary to (74), in general, $(\omega_j \circledast x^k)(0) \neq 0$ for $k = 0$.

Let $l_2(\mathbb{Z}^2)$ be a space of square summable $2D$ functions y defined on the infinite regular grid X, i.e., $y \in l_2(\mathbb{Z}^2)$ if $\sum_{s \in \mathbb{Z}^2} y^2(s\Delta) < \infty$.

Introduce accumulated kernels

$$\Omega_j(x) = \sum_{j=0}^{J} \omega_j(x), \quad j = 0, 1, \ldots, J. \tag{84}$$

Define linear spaces W_j and V_j generated by the kernels ω_j and Ω_j, respectively:

$$W_j = \left\{ \nabla y_j(x) = \sum_{u \in U} \omega_j(u) y(x-u) \colon \ y \in l_2(\mathbb{Z}^2) \right\}, \quad j = 1, \ldots, J, \tag{85}$$

$$V_j = \left\{ y_j(x) = \sum_{u \in U} \Omega_j(u) y(x-u) \colon \ y \in l_2(\mathbb{Z}^2) \right\}, \quad j = 0, 1, \ldots, J. \tag{86}$$

It can be verified that these $\nabla y_j \in l_2(\mathbb{Z}^2)$ and $y_j \in l_2(\mathbb{Z}^2)$ for any $y \in l_2(\mathbb{Z}^2)$. Thus, W_j and V_j are subspaces of $l_2(\mathbb{Z}^2)$ defined as convolutions of $y \in l_2(\mathbb{Z}^2)$ with the kernels ω_j and Ω_j, respectively.

The kernels ω_j are orthogonal. However, it does not mean that the subspaces W_j are orthogonal also. It follows from (84) that the subspaces V_{j+i} can be represented in the form

$$V_{j+1} = V_j + W_{j+1}, \quad j = 0, 1, \ldots, J-1, \tag{87}$$

where the plus "+" stays for the sum of two subspaces. The subspace W_{j+1} is a complement (nonorthogonal in general) of the subspace V_j.

It follows that

$$V_{j+1} = V_0 + \left(\sum_{k=1}^{j+1} W_k \right), \quad j = 0, \ldots, J-1.$$

One requirement on the sequence of subspaces V_j in the MR analysis is completeness:

$$y_j(x) \to y(x) \quad \text{as } j \to J.$$

The sequence V_j defined in (85) is complete because of (71)

$$V_J = l_2(\mathbb{Z}^2). \tag{88}$$

Definition. The sequence of spaces $\{V_j,\ j = 0, 1, \ldots, J\}$ generated by ω_j, $j = 0, 1, \ldots, J$, is a MR analysis of $y \in l_2(\mathbb{Z}^2)$ and ω_j are MR analysis kernels defined on the scale set H.

The following is a *nonparametric local regression decomposition* of y based on the analysis kernels ω_j.

Proposition. *Let $y \in l_2(\mathbb{Z}^2)$ and $q_j = \sum_{l=j}^{J} Q_{l,j}$, where $Q_{l,j}$ are elements of the matrix Q in* (81)*, then*

$$y(x) = \sum_{j=0}^{J} \alpha_j(x) \cdot q_j, \tag{89}$$

$$\alpha_j(x) = (\omega_j \circledast y)(x). \tag{90}$$

Proof of Proposition. Substituting (90) in (89) transforms this representation in the convolution

$$y(x) = (K \circledast y)(x), \tag{91}$$

$$K(x) = \sum_{j=0}^{J} \omega_j(x) q_j. \tag{92}$$

Show that $K(x) = \delta_{x,0}$ and in this way prove that (91) is the identity. Indeed, it follows from (81) that

$$\sum_{l=0}^{J} \nabla g_l(x) = \sum_{l=0}^{J} \sum_{j=0}^{J} Q_{l,j} \omega_j(x) = \sum_{j=0}^{J} \omega_j(x) \sum_{l=0}^{J} Q_{l,j}$$

$$= \sum_{j=0}^{J} \omega_j(x) \sum_{l=j}^{J} Q_{l,j} = \sum_{j=0}^{J} \omega_j(x) q_j = K(x).$$

However, according to (75) $\sum_{l=0}^{J} \nabla g_l = g_{h_J}(x) = \delta_{x,0}$. It shows that $K(x) = \delta_{x,0}$ and completes the proof of the proposition. □

The formulas (90) and (89) yield the accurate MR spectral expansion of y valid for any nonparametric regression $y \in l_2$. Specifically, the equations (90) and (89) define, respectively, the analysis and synthesis steps of this MR analysis. The syntheses formula (89) shows that the projections α_j of y on the MR analysis spaces V_j taken with the weights q_j enables the perfect reconstruction of any $y \in l_2(\mathbb{Z}^2)$.

We wish to mention that the syntheses formula (89) is not unique. In particular, the synthesis of the form

$$y(x) = \sum_{j=0}^{J} \alpha_j(x)\omega_j(0), \quad \alpha_j(x) = (\omega_j \circledast y)(x), \tag{93}$$

is studied in [23,24].

If $g_J(x) \neq \delta_{x,0}$ the formula (89) can be used as an approximate syntheses formula which gives the reconstruction of $y(x)$ within the bias error of the kernel operator g_{h_J}. This bias error is analyzed in Section 4. Note that the following generalization of (89) is valid

$$\hat{y}_{h_i}(x) = \sum_{j=0}^{i} \alpha_j(x) q_j(i), \quad q_j(i) = \sum_{l=j}^{i} Q_{l,j}, \tag{94}$$

where $\hat{y}_{h_i}$ is the LPA estimate of the scale h_i.

5.3. MR for differentiation

For smoothing the accurate identity operator is assumed for the scale J in the perfect reconstruction formula (89). This sort of the accurate differentiation is not possible as there is no a kernel for the discrete convolution (32) which gives the accurate derivative for any y.

The approximate form of (89) is appropriate for differentiation of nonparametric regression functions.

Let us replace g_h in (72) and (73) by the differentiation kernels $g_h^{(k)}$ (29). Then, the Gram–Schmidt formulas (80) defines the orthogonal differentiation analysis kernels $\omega_j^{(k)}$ and the corresponding MR differentiation subspaces.

The analysis and syntheses formulas (89) and (90) are changed to the form:

$$\hat{y}^{(k)}(x) = \sum_{j=0}^{J} \alpha_j^{(k)}(x) q_j^{(k)}, \quad \alpha_j^{(k)}(x) = \left(\omega_j^{(k)} \circledast y\right)(x). \tag{95}$$

Here $q_j^{(k)} = \sum_{l=j}^{J} Q_{l,j}^{(k)}$ with $Q_{l,j}^{(k)}$ being the elements of the matrix $Q^{(k)}$ in the Gram–Schmidt equation

$$Q^{(k)}\omega^{(k)}(x) = \nabla g^{(k)}(x), \quad x \in U, \tag{96}$$

where

$$\omega^{(k)} = \left(\omega_0^{(k)}(x), \omega_1^{(k)}(x), \ldots, \omega_J^{(k)}(x)\right)^T,$$
$$\nabla g^{(k)} = \left(\nabla g_0^{(k)}(x), \nabla g_1^{(k)}(x), \ldots, \nabla g_J^{(k)}(x)\right)^T, \quad \nabla g_0^{(k)}(x) = g_0^{(k)}(x).$$

The $\omega_J^{(k)}$ is the finest scale differentiation analysis operator in the representation (95). The interval of the scales from 0 to J used in (95) and specified in H defines the derivative estimates of different scales to select from or to use jointly in the combined MR estimation.

As $\omega_J^{(k)}$ is not the accurate differentiating operator. The formula (95) defines an approximate reconstruction of the derivative within the accuracy corresponding to the derivative estimate with the kernel $\omega_J^{(k)}$ having the finest scale in the set H, $h = h_J$.

5.4. Examples of kernels

Some examples of the MR analysis smoothing kernels ω_j are shown in Fig. 2. These kernels are obtained according to (80) where g_h are derived from the LPA of the power $m = 2$ for the $2D$ Gaussian window $w = \frac{1}{2\pi}\exp(-\|x\|^2)$ truncated to the squares of the size $h \times h$, where $h \in H$ and $H = \{21, 11, 5, 3, 1\}$. The first MR kernel ω_0 defines a smoothing low-pass filter while others analysis kernels ω_j, $j = 1, \ldots, 4$, define band-pass filters.

The MR differentiating kernels $\omega_j^{(1,0)}(x)$ are shown in Fig. 3. These kernels are obtained from $g_{h_j}^{(1,0)}(x)$ (differentiation on x_1) derived by using the LPA of the power $m = 2$ and the $2D$ Gaussian window $w = \frac{1}{2\pi\sigma^2}\exp(-\|x\|^2/\sigma^2)$, $\sigma = 0.5$. These kernels are truncated to the squares of the size $h \times h$, $h \in H$, with $H = \{21, 15, 11, 5, 3\}$. The first $\omega_0^{(1,0)}$ is the MR

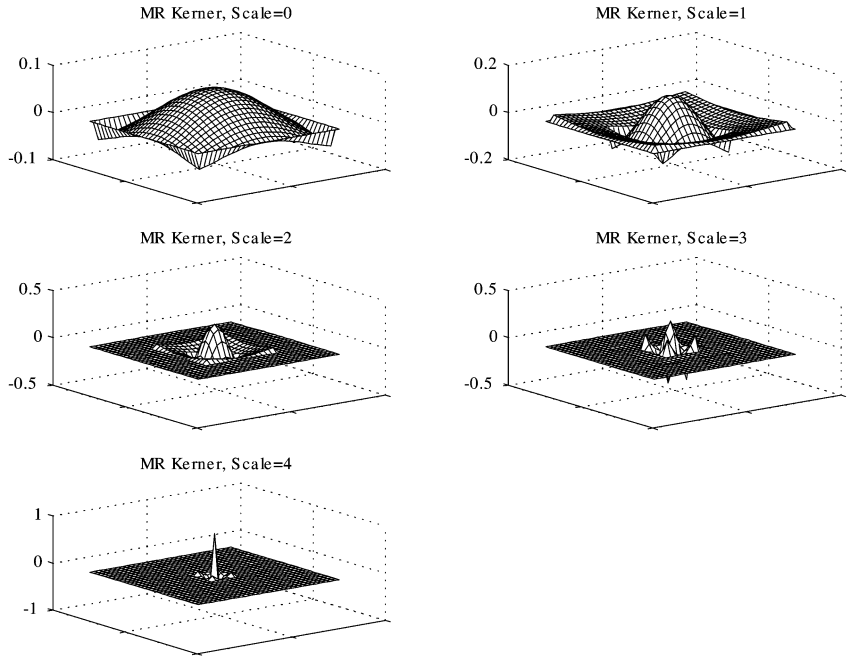

Fig. 2. MR analysis kernels $\omega_j(x)$ obtained using the LPA of the power $m = 2$ and the $2D$ Gaussian window $w(x) = \frac{1}{2\pi}\exp(-\|x\|^2)$.

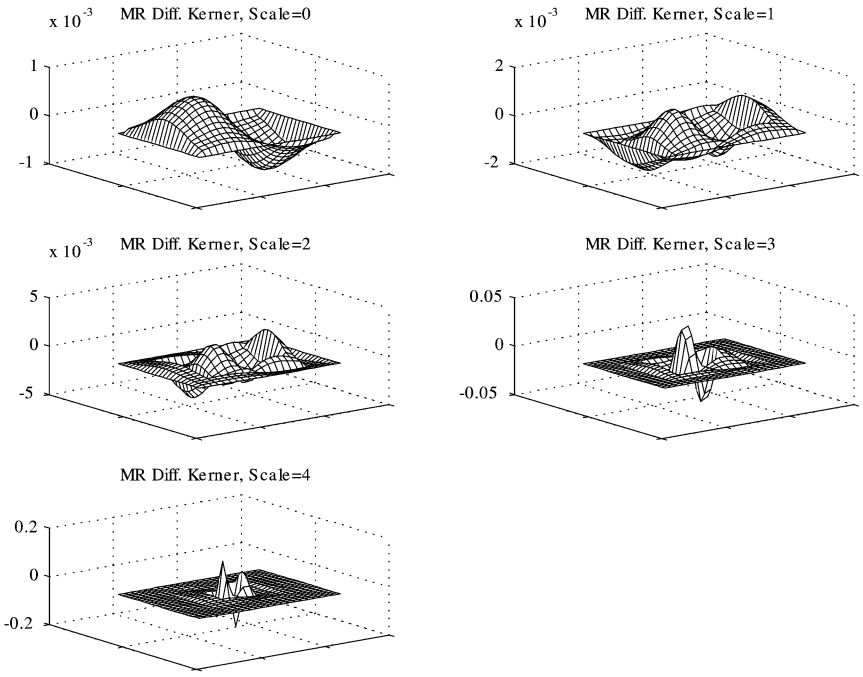

Fig. 3. MR differentiating kernels $\omega_j^{(1,0)}(x)$ obtained using the LPA of the power $m = 2$ and the $2D$ Gaussian window $w(x) = \frac{1}{2\pi\sigma^2}\exp(-\|x\|^2/\sigma^2)$, $\sigma = 0.5$.

differentiation kernel of the largest scale $h_0 = 21$. The higher scale kernels $\omega_j^{(1,0)}$, $j > 0$, become narrower.

A specific feature of the differentiation kernels is that all of them have zero value for the central pixel of the square mask $h \times h$. The smoothing kernels $\omega_j(x)$ have maximum peaks at the central pixel. The support of a square mask differentiating kernel should be larger than 1×1 in order the LPA fitting with $m \geqslant 1$ would be possible. For the minimum scale differentiating kernel we use the mask 3×3, $h = 3$. It is the minimum h in the set H used for differentiating.

6. Filtering by thresholding

A common underlying assumption in multiscale MR curve/surface/signal estimation is that the function to estimate has some redundancy. This is often reflected by the hypothesis that it belongs to a particular functional class. For example, it could be discontinuous but only at a limited number of places, or the function is assumed to have only one mode or to be monotone. Then, the heuristic for the use of, say wavelets, is that the expansion of such a function in a wavelet basis is sparse, i.e., only a few of the wavelet coefficients are large and the rest are small and thus negligible. Hence, in order to estimate the function, one has

to estimate the large wavelet coefficients and discard the rest. This approach has proved useful and successful as shown, in recent years, by various authors (see Refs. [11,31] and references herein). In what follows we apply the thresholding technique to the MR local regression spectral components.

Let the image observation be given by the noisy model (1) and the analysis kernels ω_j be applied to these data. Then the noisy observations

$$\hat{a}_j(x) = (\omega_j \circledast z)(x) \tag{97}$$

of the true spectral coefficients $\alpha_j(x) = (\omega_j \circledast y)(x)$ are linked by the equation

$$\hat{a}_j(x) = \alpha_j(x) + \sigma n_j(x), \quad j = 0, 1, \ldots, J, \tag{98}$$

where $n_j = (\varepsilon \circledast \omega_j)(x)$ are the zero mean Gaussian noise with the standard deviation equal to 1. The orthogonality of the analysis kernels $\omega_j(x)$ means that these noises are uncorrelated for different scales j and a fixed x. The goal is to estimate the unknown vector $\alpha(x) = (\alpha_0(x), \alpha_1(x), \ldots, \alpha_J(x))^T$ from the observations (98). When these estimates of $\alpha_j(x)$ are found the function estimate can be used in the form (89) with the true $\alpha_j(x)$ replaced by the estimates.

This nonparametric estimation via the MR spectral decomposition is divided into two steps. The first step transforms the data into noisy versions of the spectral coefficients $\hat{\alpha}_j(x)$. In the second step, these estimates of the spectral coefficients are filtered using the heuristic, confirmed by simulation, that the spectral MR representation of the signal is sparse and that the noise is evenly spread over the empirical spectral coefficients. Since the spectral MR representation usually is sparse, it is expected that only a small fraction of the spectral coefficients is large and that the rest is small and thus negligible. So if a spectral coefficient is small, it is reasonable to regard it as mostly noise and to set it to zero; if it is large, it is reasonable to keep it. This is known as a hard-thresholding. A soft-thresholding shrinks everything towards zero by a certain amount, thus reducing the variance of the estimation at the cost of a higher bias.

If the noisy $\hat{a}_j(x)$ (98) are substituted in (89) instead of $\alpha_j(x)$ then $\hat{y}(x) = z(x)$, i.e., there is no filtering because the formula (89) gives a perfect reconstruction of any input signal $z(x)$. The mean square error of this straightforward estimate is given by the formula

$$E\left(\left[\sum_{j=0}^{J} q_j\big(\alpha_j(x) - \hat{a}_j(x)\big)\right]^2\right) = \sum_{j=0}^{J} q_j^2 E\big\{\big(\alpha_j(x) - \hat{a}_j(x)\big)^2\big\}, \tag{99}$$

as $\hat{a}_j(x)$ are uncorrelated for different j. The additive structure of (99) with independent contribution by the estimates of the different scales shows that the "diagonal" estimation, i.e., independent estimation of $\alpha_j(x)$ for different j is a reasonable idea.

Assume that this diagonal estimator $\tilde{\alpha}_j(x)$ of $\alpha_j(x)$ has a linear structure [31]

$$\tilde{\alpha}_j(x) = \gamma \hat{a}_j(x), \tag{100}$$

where $0 \leqslant \gamma_j \leqslant 1$ is an attenuation factor of the estimate. Then the estimate (89) has a form

$$\hat{y}(x) = \sum_{j=0}^{J} \tilde{\alpha}_j(x) \cdot q_j = \sum_{j=0}^{J} \gamma_j \hat{a}_j(x) \cdot q_j. \tag{101}$$

Let us start from the "oracle" linear thresholding assuming that $\alpha_j(x)$ is known. It gives the ideal γ_j as $\gamma_j = |\alpha_j(x)|^2/(|\alpha_j(x)|^2 + \sigma^2)$ [31]. The oracle estimate cannot be realized from the data since it depends on the unknown $\alpha_j(x)$. However, this estimate is useful, in particular, as a bench mark for real estimates.

We consider four thresholding algorithms applied to the observations (98) which define nonlinear estimates $\tilde{\alpha}_j(x)$ of $\alpha_j(x)$ in the model (100) by selection of the attenuation factor γ_j as a nonlinear function of $\hat{a}_j(x)$ (e.g., [3,31]):

(a) *Hard-thresholding*

$$\gamma_j(x) = 1\big(\big|\hat{a}_j(x)\big| > t \cdot \sigma\big). \tag{102}$$

Here and in what follows $t > 0$.

(b) *Soft-thresholding*

$$\gamma_j(x) = \big(1 - t \cdot \sigma \big/ \big|\hat{a}_j(x)\big|\big)_+, \tag{103}$$

where $(a)_+ = a$ if $a > 0$ and $(a)_+ = 0$ otherwise.

(c) *Stein's thresholding*

$$\gamma_j(x) = \big(1 - t \cdot \sigma \big/ \big|\hat{a}_j(x)\big|^2\big)_+. \tag{104}$$

(d) *Smoothed Stein's thresholding*

$$\gamma_j(x) = \big(1 - t \cdot \sigma \big/ \big|\hat{b}_j(x)\big|^2\big)_+, \tag{105}$$

where $|\hat{b}_j(x)|^2$ is a mean value of $|\hat{a}_j(x)|^2$ calculated in a square $(M_1 \times M_1)$ neighborhood of the pixel x.

It follows from the MR representation for differentiation (95) that the adaptive scale derivative estimates can be given by the formulas (100)–(105) provided that $\hat{a}_j(x)$ are replaced by the corresponding $\hat{a}_j^{(k)}(x) = (\omega_j^{(k)} \circledast z)(x)$, where $\omega_j^{(k)}$ are defined as in (96).

In this section me exploit a few well known and efficient diagonal thresholding methods while there are many interesting alternatives (e.g., [3,5,6,11,31]). The thresholding overall allows an interesting interpretation in the context of the sequence estimation or model selection framework [1].

7. Optimality of the adaptive scale estimation

The best accuracy which can be achieved using nonparametric regression estimates (31) and (32) equipped with the adaptive varying scale h for y from the class (4) is restricted by the convergence rate

$$\sqrt{r^{(k)}\big(x, h_k^+(x)\big)} = O\left(\left(\frac{\log n}{n}\right)^{(M-|k|)/(2M+2)}\right). \tag{106}$$

It differs by the factor $\log n$ from the formula for the ideal estimator (56) and shows that the adaptive convergence rate is much slower than that is for the ideal estimator. One of the fundamental results of the modern adaptive estimation theory says that this $\log n$-factor is unavoidable in adaptive estimation. Thus, there are no algorithms which could achieve the better accuracy than shown in (106). Moreover, if this convergence rate is proved for some algorithm it means that this algorithm is best possible in terms of the convergence rate.

The theoretical analysis produced for $1D$ regression in [9] and for multidimensional regression in [33] shows that the ICI adaptive algorithms achieve the best convergence rate and in this way the ICI adaptation is asymptotically optimal. Similar results for different classes of function and different accuracy criteria are proved for many versions of Lepski's adaptation algorithms [27,28,35].

The introduced MR spectral decomposition transforms the original nonparametric estimation problem into the sequence estimation framework with the sequence of $\alpha_j(x) = (\omega_j \circledast y)(x)$ defined as the projection of $y(x)$ on the subspaces V_j forming the MR analysis. In other terms, the filtering in the domain of the original argument x is replaced by filtering in the MR spectrum domain.

The sequence estimation framework is quite different from the conventional nonparametric regression methods that mainly exploit the smoothness of the estimated function. The sequence estimation is based on the concept that the sparsity of representation is a more basic notion than the smoothness and that the nonlinear thresholding can be the powerful competitor to traditional linear methods even equipped with the adaptive scale selection algorithms (see [1,3,5,6], and references herein).

A simple example illustrates a source of the possible advantage of processing in the spectral domain. Let us assume that a signal $y(x)$ in the spectrum domain has the only one kth component different from zero

$$y(x) = \alpha_k(x) q_k, \quad \alpha_j(x) = 0, \ j \neq k.$$

Assume that the hard thresholding algorithm identify this nonzero component perfectly, i.e., $\hat{\alpha}_j(x) = 0$ for all $j \neq k$, with the function estimate $\hat{y}(x) = \hat{\alpha}_k(x) q_k$. This estimate is unbiased with the variance

$$E\left\{\left(y(x) - \hat{y}(x)\right)^2\right\} = \sigma^2 q_k^2, \quad q_k = \sum_{l=k}^{J} Q_{l,k}. \tag{107}$$

Further assume that, say ICI algorithm also makes a perfect estimate of the adaptive scale as $i^+ = k$ and gives the estimate as $\hat{y}^+(x) = \hat{y}_{h_k}(x)$. According to (94) this estimate can be presented as

$$\hat{y}_{h_k}(x) = \sum_{j=0}^{k} \alpha_j(x) q_j(k), \quad q_j(k) = \sum_{l=j}^{k} Q_{l,j}.$$

This estimate is also unbiased with the variance

$$E\left\{\left(y(x) - \hat{y}_{h_k}(x)\right)^2\right\} = \sigma^2 \sum_{j=0}^{k} q_j^2(k). \tag{108}$$

Comparing (107) versus (108) assumes for simplicity that Q is the identity matrix, i.e., Δg_{h_j} in (81) are orthonormal. Then $q_k = 1$ and $q_j(k) = 1$ and we obtain for the estimate variances $E\{(y(x) - \hat{y}(x))^2\} = \sigma^2$ versus $E\{(y(x) - \hat{y}_{h_k}(x))^2\} = \sigma^2(k+1)$. Thus, the MR hard thresholding algorithm has a smaller variance values for all scales $k > 0$ with the maximum advantage for the highest frequency scale $k = J$ when the variance of the ICI estimate takes its maximum value $\sigma^2(J+1)$ versus the MR hard thresholding variance $E\{(y(x) - \hat{y}(x))^2\} = \sigma^2$.

This example shows that signals with sparse spectrum representation define a class where the MR adaptive estimation is able to demonstrate a better performance as compare with the nonparametric methods based on the best scale selection.

The general analysis can be produced in order to reveal the ability of the introduced technique. Mainly, this sort of results are of asymptotic nature assuming that the number of observations n, the threshold t and the number J of scales in H are growing. It can be proved that if n, t, and scales h_j are adjusted properly the best possible convergence rate can be achieved. While this sort of analysis is beyond the scope of this paper we wish to note that actually many accuracy results obtained for the wavelet techniques are applicable for the considered estimates at least provided that the usual dyadic scale is assumed for h_j.

8. Algorithm implementation

8.1. Basic MR algorithm

Main steps of the MR algorithm:

(1) Set $H = \{h_0 > h_1 > h_2 > \cdots > h_J\}$, m, t;
(2) For $h = h_j$, $j = 0, \ldots, J$, calculate:
  (a) The kernels $g_{h_j}(x)$ (28),
  (b) The MR kernels $\omega_j(x)$ (81),
  (c) The estimates $\hat{a}_j(x)$ (97);
(3) Apply one of the thresholding rules (102)–(105) to the estimates $\hat{\alpha}_j(x)$;
(4) Calculate the MR adaptive estimate according to the final formulas (101).

Note. The step 2b defines a bank of the linear filters of different scales j. Step 2c serves for calculation of the estimates for all j and x.

The estimate $\hat{\sigma}$ for Step 3 can be obtained from the high scale MR spectrum $\hat{a}_J(x)$ as a robust median estimate $\hat{\sigma} = \text{median}_x(|\hat{a}_J(x)|)/0.6745$.

8.2. Multiple window estimation

A symmetric window w is a good choice in (28) and (29) if y is isotropic in a neighborhood of the estimation point. However, if y is anisotropic, as it happens near discontinuities or image edges a nonsymmetric approximation of y becomes much more reasonable. To deal with the anisotropy of y multiple nonsymmetric window estimates are exploited

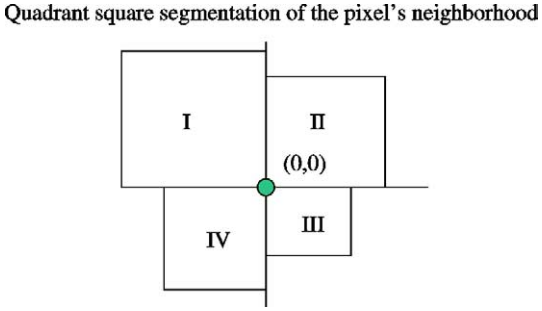

Fig. 4. The quadrant's segmentation of the neighborhood of the LPA centre $(0, 0)$.

[18,19]. It assumes that the neighborhood of the pixel x is separated in a number overlapping or nonoverlapping subareas. Let's K such subareas be introduced. Then, the adaptive scale estimates $\hat{y}_h^{[k]}$, $k = 1, \ldots, K$, are calculated for each of these subareas and fused together in order to yield the final estimate. The four quadrant nonoverlapping segmentation of the pixel neighborhood (see Fig. 4) is a simple and efficient way of fitting y [18,19]. It assumes that the origin of the Euclidian rectangular coordinate $(0, 0)$ is the centre of the LPA estimate for each square quadrant subareas. For each of these quadrants the $\hat{y}_h^{[k]}$ kernel estimates with the adaptive scale are calculated. Thus, for each pixel of the image we are able to obtain four independent estimates based on different observations covered by the corresponding quadrant supports respectively.

There are a number of ways how to fuse the quadrant's estimates into the single final one. In particular, the inverse-variance weighted mean [9,18] or the sample mean can be applied. The last estimate gives

$$\hat{y}(x) = \frac{1}{K} \sum_{k=1}^{K} \hat{y}^{[k]}(x). \tag{109}$$

In our simulation we complete a set of the introduced quadrant's windows by the symmetric window which is centered with respect to the origin point $(0, 0)$. The multiple window estimation significantly improves the performance of the algorithms. As a further development of this idea special directional LPA kernels using narrow beam-wise supports are proposed in [19].

8.3. Algorithm complexity

The calculation of the image estimate $\hat{a}_j(x)$ for given j is the linear convolution requiring $N_{\text{conv}} \sim n \log n$ operations $n = n_1 n_2$. If the sectioning procedure is used for the convolution (e.g., [34]), then $N_{\text{conv}} \sim n \log n_j$, where n_j is a maximum size of the square mask of the kernel ω_j. These calculations are repeated for each of the K subareas (quadrants) of the pixel neighborhood with the fusing the estimates according to (109). The thresholding is produced J times for each of the K subareas. Thus, overall the algorithm complexity is proportional to $J \cdot K \cdot N_{\text{conv}}$, where $K = 5$ for the four quadrant and symmetric window estimate.

9. Parallels with wavelets

Let us provide few facts from the wavelet theory which help to demonstrate similarity and difference with the introduced MR nonparametric local polynomial regression approach. The standard MR continuous wavelet expansion for $1D$ continuous data, $y(x)$, $x \in R$, has a form of the following parametric series (e.g., [11,31]):

$$y_W(x) = \sum_{k\in\mathbb{Z}} \alpha_k \varphi_{0k}(x) + \sum_{k\in\mathbb{Z}} \sum_j \alpha_{jk} \psi_{jk}(x), \tag{110}$$

$$\varphi_{0k}(x) = \varphi(x-k), \qquad \psi_{jk}(x) = 2^{j/2}\psi(2^j x - k), \tag{111}$$

where $\varphi(x)$ and $\psi(x)$ are the scale function (father wavelet) and the wavelet (mother wavelet) respectively, $\mathbb{Z}$ is a set of integers, 2^j stands for jth dyadic scale. The scale h used in this paper is linked with the wavelet dyadic scale by the equation $h = 2^{-j}$. For the orthonormal $\varphi_{0k}(x)$ and $\psi_{jk}(x)$, $x \in R$, the coefficients of the series (110) are calculated as

$$\alpha_k = \langle y, \varphi_{0k}\rangle, \qquad \alpha_{jk} = \langle y, \psi_{jk}\rangle. \tag{112}$$

The inner products used in this section assumes integrals on $x \in R$, for instance $\alpha_{jk} = \langle y, \psi_{jk}\rangle = \int_{-\infty}^{+\infty} y(x)\psi_{jk}(x)\,dx$.

The orthonormality exploited in (112) means that

$$\langle \psi_{il}, \psi_{jk}\rangle = \delta_{ij}\delta_{kl}, \qquad \langle \varphi_{0l}, \psi_{jk}\rangle = 0, \qquad \langle \varphi_{0l}, \varphi_{0k}\rangle = \delta_{kl}, \tag{113}$$

i.e., the intra- and enter-scale orthogonality of the functions (111) is assumed. The intra-scale orthogonality means that the functions of the same scale j but different by the argument shift (variable k in (111)) are orthogonal, i.e., $\langle \psi_{jl}, \psi_{jk}\rangle = \delta_{kl}$, $\langle \varphi_{0l}, \varphi_{0k}\rangle = \delta_{kl}$. The inter-scale orthogonality means the orthogonality of all functions of the different scales, e.g., $\langle \psi_{ik}, \psi_{jk}\rangle = \delta_{ij}$, as well as it is assumed also the orthogonality between the father and mother wavelets $\langle \varphi_{0l}, \psi_{jk}\rangle = 0$. This double intra- and inter-scale orthogonality requirements make a design of the wavelet function quite a complex art and confines the classes of wavelet functions.

The father wavelet φ generates the following linear subspaces in $L_2(R)$:

$$\begin{aligned} V_0 &= \left\{ y(x) = \sum_s \varphi(x-s)c_s\colon \sum_s |c_s|^2 < \infty \right\},\\ V_1 &= \left\{ f(x) = y(2x)\colon\ y \in V_0 \right\},\\ &\cdots\\ V_j &= \left\{ f(x) = y(2^j x)\colon\ y \in V_0 \right\}, \end{aligned} \tag{114}$$

such that

(1) The subspaces V_j are nested, $V_j \subset V_{j+1}$ and can be represented in the form

$$V_{j+1} = V_j \oplus W_{j+1}, \quad j = 0, 1, \ldots, \tag{115}$$

where $\oplus$ stays for the direct sum of the subspaces, W_{j+1} is an orthogonal complement of the subspace V_j and the complement subspaces W_j for every scale j are defined by the wavelets $\psi_{jk}(x)$.

(2) The subspace $\bigcup_{j=0}^{\infty} V_j = V_0 \bigoplus_{j=1}^{\infty} W_j$ is dense in $L_2(R)$.

These orthogonal subspaces define the wavelet MR analysis [11,31]. It means that any $y \in L_2(R)$ can be represented as a series (110)–(112).

Now let us compare the wavelet expansion (110)–(112) with the corresponding nonparametric regression MR expansions (89) and (90):

- The wavelet expansion y_W (110) is a standard orthogonal series with invariant coefficients α_k and β_{jk} (112). As a function of x this series is defined by the wavelets $\varphi_{0k}(x)$, $\psi_{jk}(x)$ (111). It means that (110) is a parametric expansion of y. In contrast to it the expansion (89) is nonparametric as its dependence on x cannot be parameterized and goes through the coefficients ($\alpha_0(x)$ and $\alpha_j(x)$) of the expansion. There are no invariant coefficients in this expansion and the basis functions depending on x, what is typical for the standard series.
- The kernels ω_0 and ω_j, $j = 1, \ldots, J$, in (89) and (90) can be interpreted as the father and mother wavelets, respectively, according to their role in the analysis and the vanishing moment conditions (83). The kernel ω_0 defines a lower frequency background of the signal (father wavelet analysis) while the kernels ω_j, $j = 1, \ldots, J$, define a higher frequency complements to this background (mother wavelet analysis).
- The subspaces W_j and V_j are different for the wavelet and the introduced local regression MR analysis. The principal difference is that in the wavelets W_j are orthogonal complements of V_j while there is no such orthogonality for the kernel MR analysis.
- The dyadic scale in (110)–(112) is a special point defining the wavelet design and fast algorithms. In the local regression MR analysis the scale as defined by the set H is quite arbitrary. The only serious restrictions concern the linear independency of g_h for $h \in H$.
- The wavelet design for multivariable functions is a serious problem mainly solved by using the direct product of the $1D$ wavelets. There is no principal difficulties with design of the LPA kernels for any dimension.

We present here the classical results of the continuous integral wavelet transform with $x \in R$ and $y \in L_2(R)$ versus the discrete MR local regression analysis for the kernel and estimates defined for the discrete $x \in \mathbb{Z}^2$ and $y \in l_2(\mathbb{Z}^2)$. We pragmatically consider only discrete signals as it results in clear numerical algorithms. A generalization of the introduced kernel MR is straightforward for continuous signals belonging to $L_2(R)$ or $L_2(R^2)$.

The discussed parallels concern only the structures and the basic ideas of these two different transforms.

We may conclude that the ideas of the MR used for the wavelets (114) and (115) and for the MR local regression (85)–(87) are quite similar. However, in the considered nonparametric local polynomial version of the MR analysis many strict constrains typical for the wavelet technique may be dropped. The nonparametric local polynomial approach is more flexible and has more freedom for design of the filters (kernels) with concentration on the

signal properties rather than deal with mathematical difficulties essential for the wavelet design.

10. Simulation

As a test signal we use the 256×256 "cameraman" image (8 bit gray-scale) corrupted by an additive zero-mean Gaussian noise. The LPA is used with the uniform square window w, linear polynomials $m = 1$, and a finite set of the scales $H = \{21, 11, 5, 3, 1\}$. The multi-window estimate of y is applied as described in Section 8. We calculate five intermediate estimates obtained for four quadrant and one symmetric windows w, respectively. Each of these five windowing estimates is calculated as spatially adaptive using the developed MR algorithms. The final estimate of y is calculated as the mean (109) of these five adaptive scale estimates.

Figure 5 illustrates how the MR expansion (89) works. It images the items of the MR expansion $\alpha_j(x) \cdot q_j$ for the scales $j = 0, 1, \ldots, 4$. The figures are given for the noise-less cameraman image. The first term $\alpha_0(x) \cdot q_0$ presents a basic smooth lower frequency component of the image. The further terms with $j > 0$ serve as the complements of the basic one providing some finer higher frequency details. The images of the expansion items become sharper for larger values of the scale j. The sum of all five MR expansion items shown in the last image of Fig. 5 gives a perfect reconstruction of the true image.

Histograms of the images from Fig. 5 are shown in Fig. 6. They illustrate the concept of the redundancy of the proposed MR nonparametric regression expansion used in the thresholding filtering. Indeed, the histogram for $j = 0$ covers all segment $[0, 1]$ of possible values of y. The histograms for the complement components of the MR expansion are narrower and more pick-wise. The last scale $j = 4$ has a smallest number of nonzero items which are mainly concentrated in a narrow neighborhood of zero. Actually it means that the space of the MR analysis for $j \geqslant 1$ is sparse: only a few of items of the MR nonparametric expansion are large and the rest are small and thus can be dropped.

In quantitative comparison of the algorithms the following criteria have been used:

(1) Root mean squared error (RMSE): $\mathrm{RMSE} = \sqrt{\frac{1}{\#}\sum_x (y(x) - \hat{y}(x))^2}$;
(2) SNR in dB: $\mathrm{SNR} = 10\log_{10}\sum_x |y(x)|^2 / \sum_x |y(x) - \hat{y}(x)|^2$;
(3) Improvement in SNR (ISNR) in dB: $\mathrm{ISNR} = 20\log_{10}(\hat{\sigma}/\mathrm{RMSE})$;
(4) Peak signal-to-noise ratio (PSNR) in dB: $\mathrm{PSNR} = 20\log_{10}(\max_x |y(x)|/\mathrm{RMSE})$;
(5) Mean absolute error (MAE): $\mathrm{MAE} = \frac{1}{\#}\sum_x |y(x) - \hat{y}(x)|$;
(6) Maximum absolute error: $\mathrm{MAXDIF} = \max_x |y(x) - \hat{y}(x)|$.

These criteria allow to evaluate the performance of the algorithm quantitatively, while PSNR is treated as a criterion linked with a visual image perception. However, it is appeared that these criteria gives quite concordant conclusions while the visual evaluation is an independent performance criterion. In what follows we mainly use only one criteria ISNR.

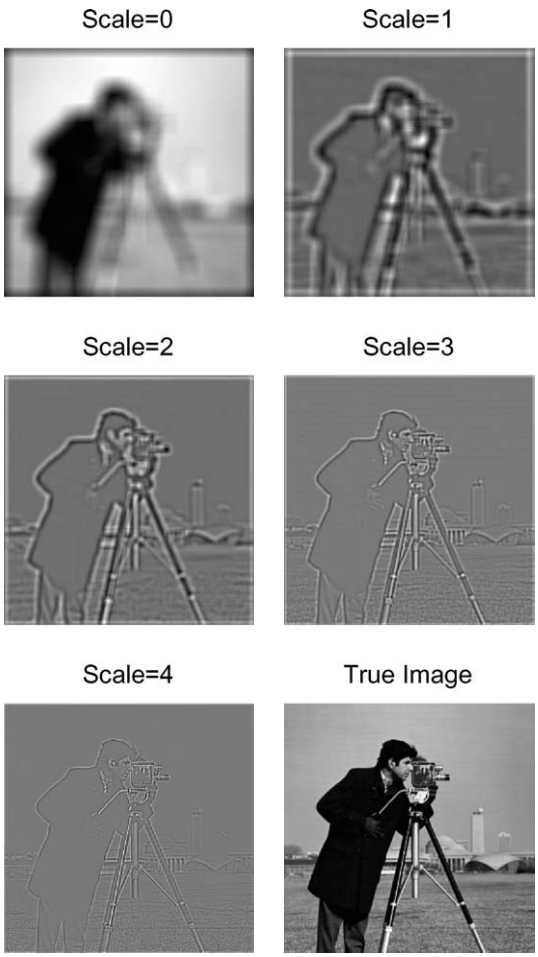

Fig. 5. The spectrum expansion of the noiseless cameraman image. The scale equal to zero corresponds to the first term of this expansion $\hat{\alpha}_0(x)q_0$ and presents a basic lower frequency components of the image. The further items $\hat{\alpha}_j(x)q_j$ with larger scales j serve as the complements of this basic item and provide finer details. The images of these items become sharper for larger values of the scale. The sum of all five MR expansion items shown in the last image is identical to the original cameraman image.

In image denoising we compare the MR algorithms versus the ICI algorithm which demonstrates a performance more less equivalent to the performance of the Lepski–Spokoiny algorithm.

The threshold t is a main design parameter of the thresholding (102)–(105). Multiple simulations and analysis produced for different images show that $t = 1.2$–1.5 is a reasonably good value of this parameter for different scenarios. It is a sort of the rule of thumb for selection of t. For comparison we show also results for the oracle estimator.

Figure 7 shows ISNR as a function of SNR of the observations. The smoothed Stein's algorithm ($M_1 = 5$ in (105)) demonstrates the best performance and outperforms the ICI

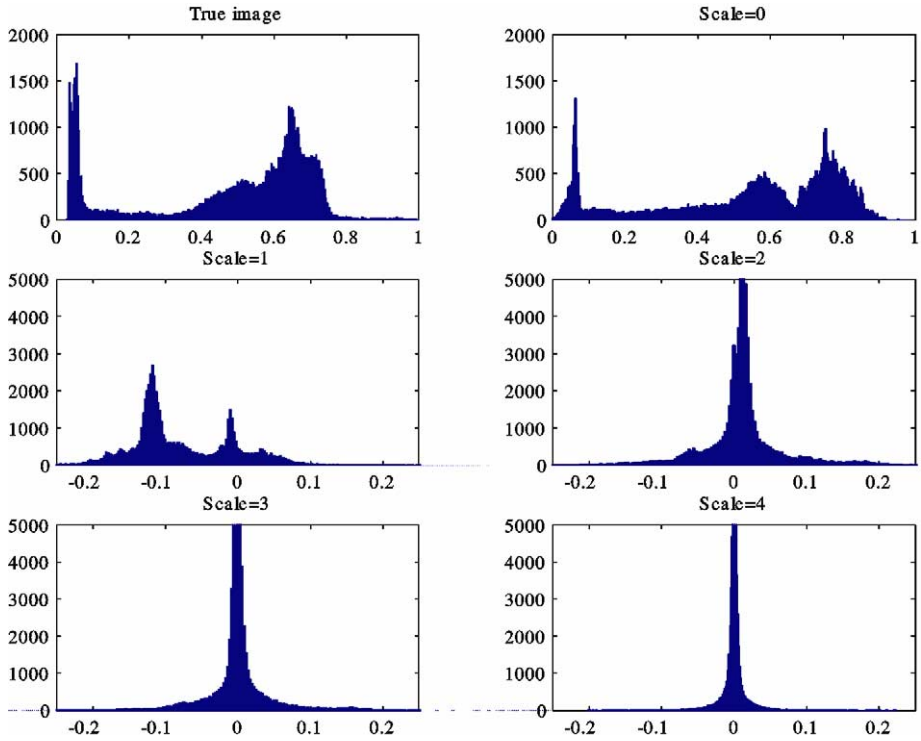

Fig. 6. The histograms of the images shown in Fig. 5. The histogram for the scale equal to 0 is wide covering nearly whole segment $[0, 1]$ even a bit wider that the histogram of the true image. The histograms for the complement components of the MR expansion for the scales $j = 1, 2, 3, 4$ become narrower with smaller frequencies of nonzero items. The last scale 5 have smallest number of nonzero items which are well concentrated in a narrow neighborhood of zero.

algorithm approximately on 1 dB. The soft-thresholding algorithm gives values which are about 1 dB worse than those for the ICI algorithm. The basic Stein's algorithm shows the figures better than the ICI algorithm only for SNR $>$ 15 dB. The oracle estimate naturally demonstrates the best values of ISNR about 2 dB higher than the smoothed Stein's algorithm. We do not show results for the hard-thresholding as they are worse than those for the soft-thresholding.

Examples of the reconstructed images can be seen in Fig. 8, where the noisy image, the smoothed Stein's, ICI and soft-thresholding images are shown. Visual evaluation is in favor of the smoothed Stein's algorithm.

Let us apply the MR algorithms for scale adaptive differentiation. We estimate the first derivatives of the cameraman image intensity on x_1 (horizontal axis) and on x_2 (vertical axis). In all following results we use the simulation scenario and LPA parameters as they are for image denoising. For the sake of simplicity of presentation and discussion for differentiation we apply a single window estimate with the symmetric window function w and the soft-thresholding only. The scales for differentiation are defined by the set $H = \{21, 15, 11, 5, 3\}$ with the minimum admissible scale $h_J = 3$.

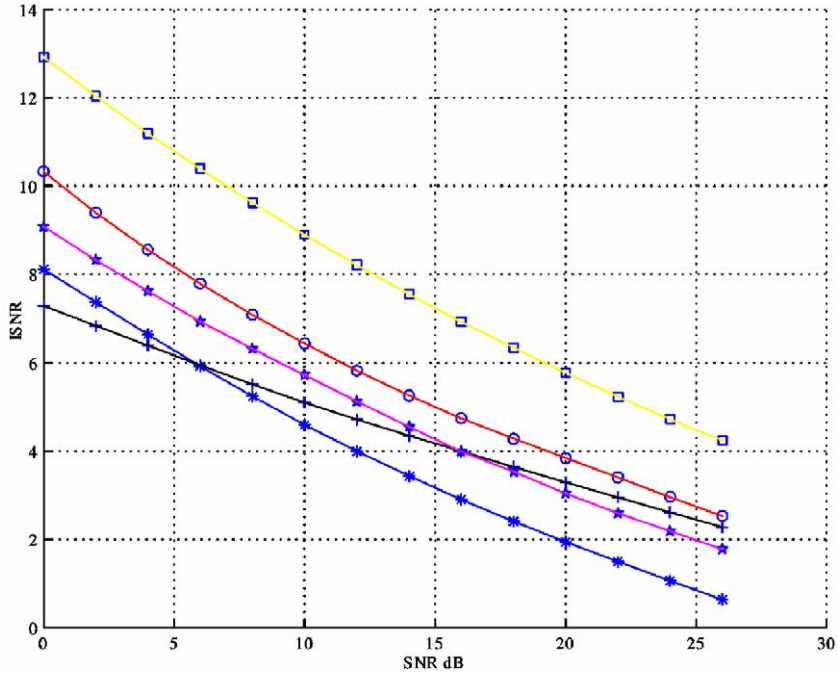

Fig. 7. ISNR as functions of SNR of the noisy observations for the cameraman image. The curves are given for the following four algorithms: soft-thresholding ('star'), Stein's ('plus'), smoothed Stein's ('o'), ICI ('pentagram'). The threshold $t = 1.2$. The oracle estimation results are marked by 'squares.' Overall, the advantage is in favor of the smoothed Stein's algorithm, which outperform the ICI algorithm about on 1 dB.

Figures 9 and 10 illustrate formation of the derivative estimates using the MR spectrum analysis produced by the kernels $\omega_j^{(k)}$. Images in the figures correspond to the items of the sum (95) $\hat{\alpha}_j^{(k)}(x)q_j^{(k)}$ for the scales $j = 0, 1, \ldots, 4$ and the derivatives $k = (1, 0)$ and $k = (0, 1)$. The last sixth image is the derivative estimate $\hat{y}^{(k)}(x) = \sum_{j=0}^{J} \alpha_j^{(k)}(x)q_j^{(k)}$, as it is defined in (95). These MR spectrums are given for the noiseless cameraman image. The influence of the scale of the differentiation operator is clear seen. It varies from smoothed derivative estimate given by the largest scale to the finer contour lines of the smaller scale estimates.

In order to produce a quantitative analysis of differentiation we need to know accurate values of the derivatives. For such image as the cameraman these derivatives are unknown. However, they can be evaluated numerically using the MR analysis. Let us assume that these "accurate" numerical derivatives are defined as the estimates given by the differentiation kernels $g_h^{(1,0)}(x)$ and $g_h^{(0,1)}(x)$ for the noiseless cameraman image provided that the scale h is equal to its minimum value. This minimum admissible scale value is $h = 3$ for the considered LPA with $m = 2$ and squared $2D$ support of the differentiating kernels.

Table 2 provides data illustrating an improvement which can be achieved by using the adaptive varying scale differentiators versus the differentiators with a fixed invariant scale.

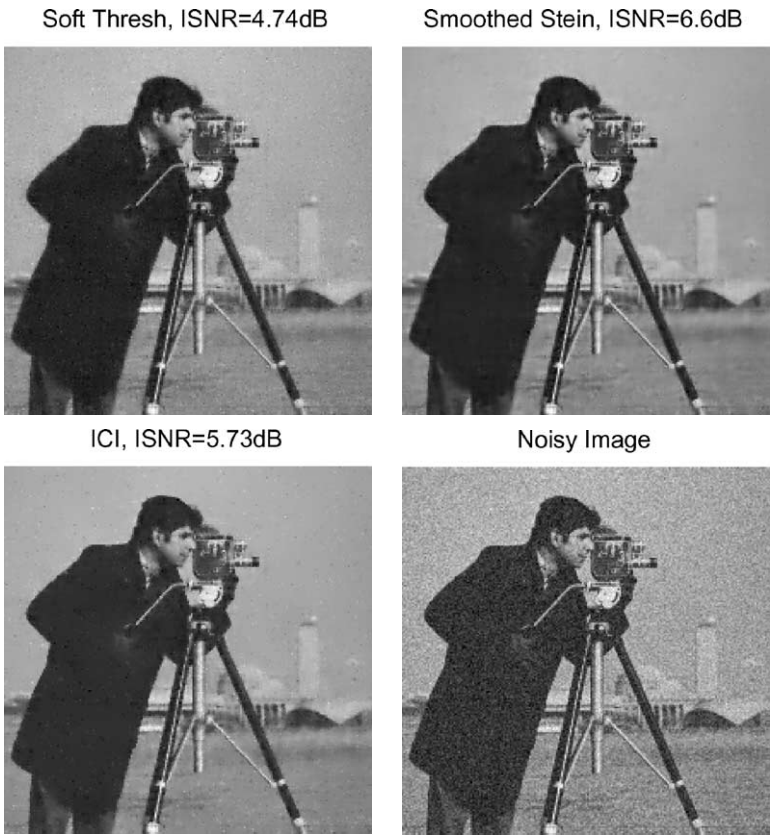

Fig. 8. The images obtained by the soft thresholding, smoothed Stein's and ICI algorithms. SNR = 10 dB. Threshold parameter $t = 1.2$ for the MR algorithms and $\Gamma = 1.5$ for the ICI algorithm. The smoothed Stein's algorithm demonstrates the best performance in terms of the ISNR values as well as visually.

Table 2
Accuracy of the derivative estimation

Invariant scale	RSME	MAE	MAXDIF
$h = 21$	0.0615	0.0275	0.4386
$h = 15$	0.0599	0.0271	0.4590
$h = 11$	0.0576	0.0263	0.4715
$h = 5$	0.0399	0.0216	0.4274
$h = 3$	0.0317	0.0252	0.1486
Adaptive scale	0.0219	0.0145	0.1310

The criteria values are given as a mean of the corresponding values obtained for the derivatives on x_1 and x_2.

The lines 1–5 of the table show the criteria values for the invariant scale estimators and the last line corresponds to the MR soft-thresholding adaptive scale estimator ($t = 1.2$). It can be concluded that the best scale invariant estimator has the scale equal to its minimum

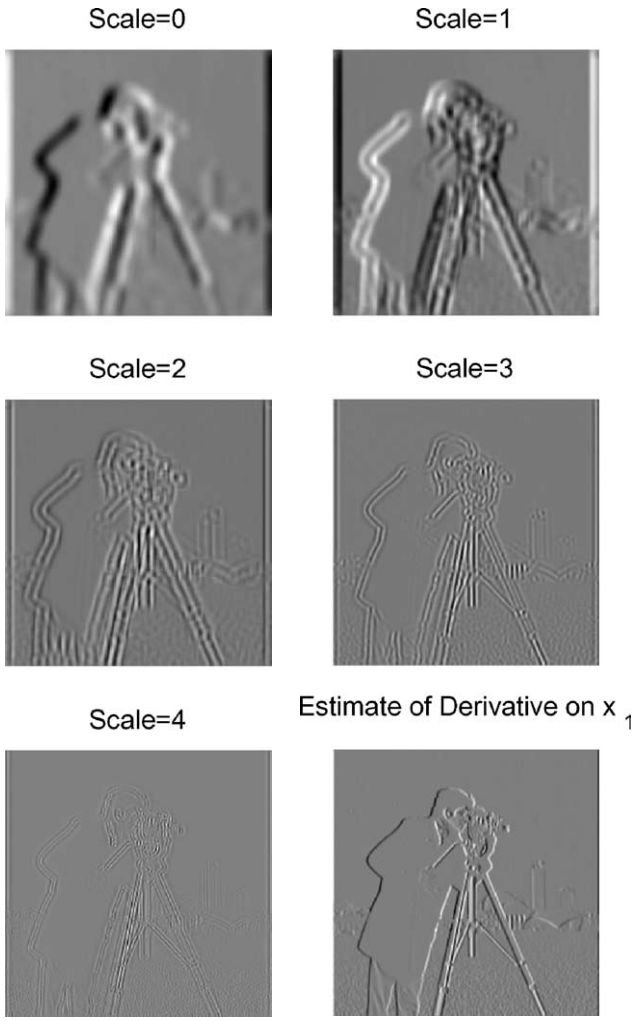

Fig. 9. The spectrum expansion of the derivative $\partial/\partial x_1$. The noiseless cameraman image. The scale equal to zero corresponds to the first term of this expansion $\hat{\alpha}_0^{(1,0)}(x)q_0^{(1,0)}$ and presents a basic lower frequency (smooth) components of the derivative. The further items with larger scales serve as the complements of this basic item and provide finer sharp details. The sum of all five MR expansion items shown in the last image is the MR nonparametric regression estimate of the derivative.

value $h = 3$. It can be noticed also that these accuracy figures are quite sensitive with respect to the scale value.

Further, comparing the criteria values of this best scale invariant estimator versus the corresponding values for the adaptive estimator, we can see quite a significant improvement in values of RMSE and MAE. It consists of about 30% for RMSE and about 40% for MAE. A less improvement can be seen in values of MAXDIF which consists of about 10%. Visual effects of the adaptive scale differentiation are illustrated in Fig. 11.

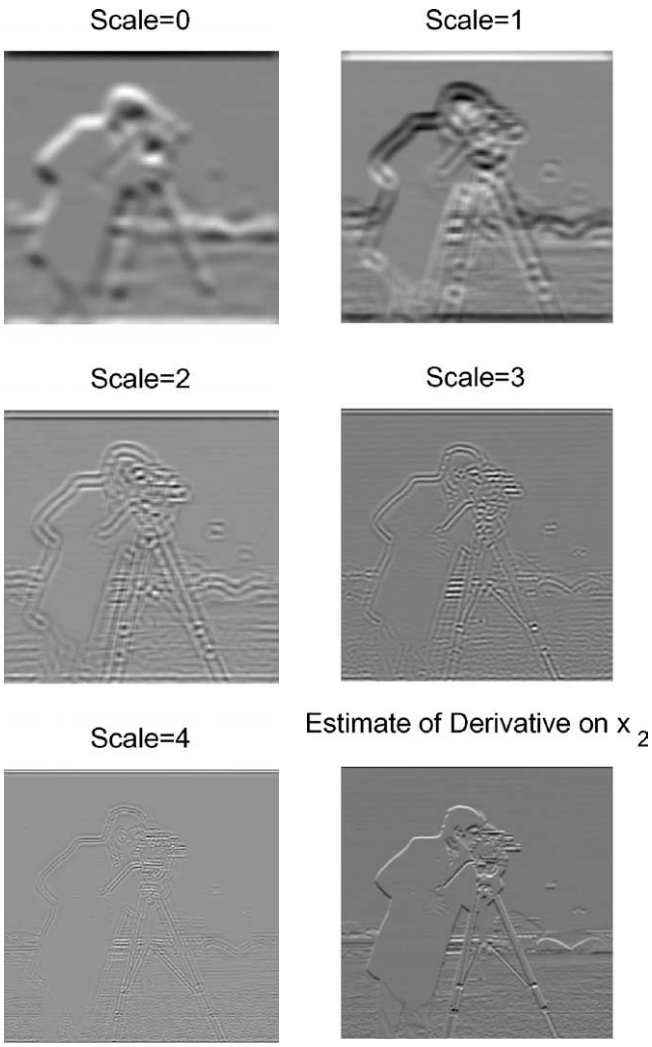

Fig. 10. The spectrum expansion of the derivative $\partial/\partial x_2$. The noiseless cameraman image. The scale equal to zero corresponds to the first term of this expansion $\hat{\alpha}_0^{(0,1)}(x)q_0^{(0,1)}$ and presents a basic lower frequency (smooth) components of the derivative. The further items with larger scales serve as the complements of this basic item and provide fine sharp details. The sum of all five MR expansion items shown in the last image is the MR nonparametric regression estimate of the derivative.

Images presented in Fig. 11 show the sum of absolute values of the estimates of the derivatives $\partial y/\partial x_1$ and $\partial y/\partial x_2$. The left-hand size image is obtained by using the derivative estimators with the best invariant scale $h = 3$. The right-hand side image corresponds to the MR varying adaptive scale soft-thresholding differentiator. Noisy components of the derivative estimates clearly seen in the left-hand side image are well cleared out in the right-hand side image while the fine details of the edges of the cameraman image are

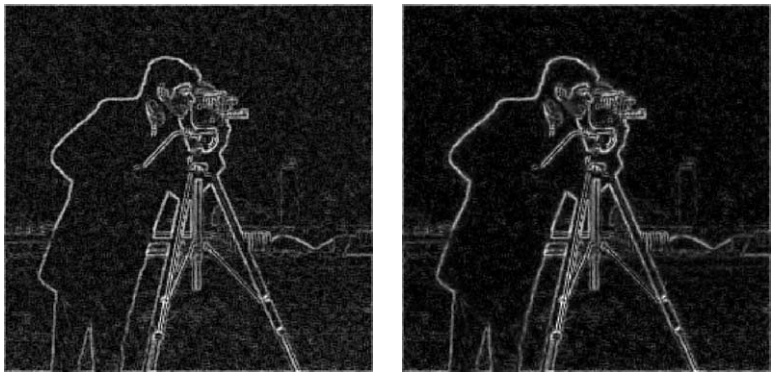

Fig. 11. Images show the sum of absolute values of the estimates of the derivatives $\partial y/\partial x_1$ and $\partial y/\partial x_2$. The left hand size image estimates are obtained using the derivative estimator kernels $g_h^{(1,0)}$ and $g_h^{(0,1)}$ with the best found invariant scale $h = 3$. The right hand side image is obtained by the MR varying adaptive scale soft-thresholding estimators with $t = 1.2$. Noisy components of the derivatives are clearly seen in the left-hand side image are cleared out in the right hand side image while the fine edge details of the cameraman are well preserved.

preserved. Thus, visually and quantitatively, the adaptive scale MR differentiator demonstrates better results as compared with the scale-invariant estimator with the best possible oracle scale selection.

11. Conclusions

A new varying adaptive scale nonparametric local polynomial regression technique is proposed. It is based on the LPA applied for design of the filters joined with the developed spectral MR analysis. The usual scale adaptive local polynomial regression estimates are based on selection of the best scale. The proposed MR analysis assumes multiscale transform of observations, filtering of the obtained local polynomial regression spectrums and fusing these filtered spectrums in the final estimate. This final estimate is composed from the estimates of the different scales but not only single one as it is in the classical adaptive nonparametric local polynomial regression. The MR estimate belongs to a more general class of estimates and is able to provide a better accuracy.

The presentation of the MR analysis is given in terms of image processing. However, the approach is applicable for data of any dimensionality defined on the regular or irregular grids. For the regular grids the MR analysis methods allow fast implementations based on the fast convolution algorithms. For the irregular grids this sort of fast algorithms is not applicable.

The developed MR nonparametric technique is quite universal and can be applied for many different tasks. The introduced spectral expansion allows to involve many traditional techniques of image processing. In particular, one may weigh spectral components or produce their nonlinear transforms in order to obtain desirable image enhancement effects. The MR nonparametric local polynomial technique can be applied for edge detection, image improvements, recognition problems, etc.

Acknowledgments

I express my appreciation to the anonymous referee for fruitful and stimulating comments. This work was supported by BK'21 Project of Kwangju Institute of Science and Technology, South Korea.

References

[1] L. Birgé, P. Massart, Gaussian model selection, J. Eur. Math. Soc. 3 (2001) 203–268.
[2] R. Brown, Smoothing, Forecasting and Prediction of Discrete Time Series, Prentice Hall, Englewood Cliffs, NY, 1963.
[3] L. Cavalier, A.B. Tsybakov, Penalized block-wise Stein's method, monotone oracles and sharp adaptive estimation, Math. Meth. Statist. 10 (3) (2001) 247–282.
[4] I. Djurović, L.J. Stanković, Nonparametric algorithm for local frequency estimation in multidimensional signals, IEEE Trans. Image Process. 13 (4) (2004) 467–474.
[5] D.L. Donoho, I.M. Johnstone, Adapting to unknown smoothness via wavelet shrinkage, J. Amer. Statist. Assoc. 90 (1995) 1200–1224.
[6] D.L. Donoho, I.M. Johnstone, Minimax estimation via wavelet shrinkage, Ann. Statist. 26 (1998) 879–921.
[7] J. Fan, I. Gijbels, Local Polynomial Modelling and Its Application, Chapman & Hall, London, 1996.
[8] L. Ganesan, P. Bhattacharyya, Edge detection in untextured and textured images – a common computational framework, IEEE Trans. Syst. Man Cybern. Part B Cybernet. 27 (5) (1997) 823–834.
[9] A. Goldenshluger, A. Nemirovski, On spatial adaptive estimation of nonparametric regression, Math. Meth. Statist. 6 (1997) 135–170.
[10] W. Hardle, Applied Nonparametric Regression, Cambridge Univ. Press, Cambridge, 1990.
[11] W. Hardle, G. Kerkyacharian, D. Picard, A. Tsybakov, Wavelets, Approximation and Statistical Applications, in: Lecture Notes in Statistics, vol. 129, Springer, 1998.
[12] R.M. Haralick, Digital steps edges from zero crossing of second directional derivatives, IEEE Trans. Pattern Anal. Machine Intell. 6 (1) (1984) 58–68.
[13] C.M. Hurvich, J.S. Simonoff, Smoothing parameter selection in nonparametric regression using an improved AIC criterion, J. Royal Statist. Soc. Ser. B 60 (1998) 271–293.
[14] V. Katkovnik, Homogeneous integral averaging operators obtained by the method of least squares, Autom. Remote Control. 32 (11) (1971) 1767–1775.
[15] V. Katkovnik, Linear and nonlinear methods of nonparametric regression analysis, Soviet J. Autom. Inform. Sci. 5 (1979) 25–34.
[16] V. Katkovnik, Nonparametric Identification and Smoothing of Data (Local Approximation Methods), Nauka, Moscow, 1985, in Russian.
[17] V. Katkovnik, A new method for varying adaptive bandwidth selection, IEEE Trans. Signal Process. 47 (9) (1999) 2567–2571.
[18] V. Katkovnik, K. Egiazarian, J. Astola, Adaptive window size image de-noising based on intersection of confidence intervals (ICI) rule, J. Math. Imaging Vision 16 (3) (2002) 223–235.
[19] V. Katkovnik, K. Egiazarian, J. Astola, Adaptive varying scale methods in image processing, in: Tampere International Center for Signal Processing, TICSP Series, vol. 19, Tampere, TTY, Monistamo, 2003.
[20] V. Katkovnik, K. Egiazarian, J. Astola, Application of the ICI principle to window size adaptive median filtering, Signal Process. 83 (2003) 251–257.
[21] V. Katkovnik, A. Gershman, L.J. Stankovic, Sensor array signal tracking using a data-driven window approach, Signal Process. 80 (12) (2000) 1507–2515.
[22] V. Katkovnik, L.J. Stankovic, Periodogram with varying and data-driven window length, Signal Process. 67 (3) (1998) 345–358.
[23] V. Katkovnik, A multiresolution nonparametric regression and image denoising, in: Proceedings of the ICIP '2003, Barcelona, 2003.
[24] V. Katkovnik, A multiresolution nonparametric regression for spatially adaptive image de-noising, IEEE Signal Process. Lett. 11 (10) (2004) 798–801.

[25] G. Kerkyacharian, O. Lepski, D. Picard, Nonlinear estimation in anisotropic multi-index denoising, Prob. Theor. Relat. Field. 121 (2) (2001) 137–170.
[26] J. Klemela, A.B. Tsybakov, Sharp adaptive estimation of linear functionals, Ann. Statist. 29 (6) (2001) 1567–1600.
[27] O. Lepskii, On a problem of adaptive estimation in Gaussian white noise, Theor. Prob. Appl. 35 (3) (1990) 454–466.
[28] O. Lepski, E. Mammen, V. Spokoiny, Ideal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selection, Ann. Statist. 25 (3) (1997) 929–947.
[29] O. Lepski, V. Spokoiny, Optimal pointwise adaptive methods in nonparametric estimation, Ann. Statist. 25 (6) (1997) 2512–2546.
[30] C. Loader, Local Regression and Likelihood, in: Series Statistics and Computing, Springer, New York, 1999.
[31] S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, 1999.
[32] J.A. McDonald, A.B. Owen, Smoothing with split linear fits, Technometrics 28 (33) (1986) 195–208.
[33] A. Nemirovski, Topics in Non-parametric Statistics, Lecture notes in mathematics, vol. 1738, Springer, New York, 2000, 85–277.
[34] A.V. Oppenheim, R.W. Schafer, Discrete-time Signal Processing, Prentice Hall, Englewood Cliffs, NJ, 1989.
[35] J. Polzehl, V. Spokoiny, Image denoising: pointwise adaptive approach, Ann. Statist. 31 (1) (2003) 30–57.
[36] D. Ruppert, Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation, J. Amer. Statist. Assoc. 92 (439) (1997) 1049–1062.
[37] W.R. Schucany, Adaptive bandwidth choice for kernel regression, J. Amer. Statist. Assoc. 90 (430) (1995) 535–540.
[38] V. Spokoiny, Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice, Ann. Statist. 26 (4) (1998) 1356–1378.