Weierstraß-Institut
für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Preprint ISSN 0946 – 8633

Spatially adaptive estimation via fitted local likelihood techniques

Vladimir Katkovnik, [1] Vladimir Spokoiny [2]

[1] Signal Processing Institute,
University of Technology of Tampere,
P. O. Box 553, Tampere, Finland.
E-Mail: E-mail: katkov@cs.tut.fi

[2] Weierstrass Institute
and Humboldt University Berlin,
Mohrenstr. 39, 10117 Berlin, Germany
E-Mail: spokoiny@wias-berlin.de
URL: www.wias-berlin.de/~spokoiny

No. 1166

Berlin 2006

1991 *Mathematics Subject Classification.* 62G05: secondary: 62G20 .

Key words and phrases. local model selection, fitted likelihood, adaptive estimation .

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Mohrenstraße 39
10117 Berlin
Germany

Fax: + 49 30 2044975
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Abstract

This paper offers a new technique for spatially adaptive estimation. The local likelihood is exploited for nonparametric modelling of observations and estimated signals. The approach is based on the assumption of a local homogeneity of the signal: for every point there exists a neighborhood in which the signal can be well approximated by a constant. The fitted local likelihood statistics is used for selection of an adaptive size of this neighborhood. The algorithm is developed for quite a general class of observations subject to the exponential distribution. The estimated signal can be uni- and multivariable. We demonstrate a good performance of the new algorithm for Poissonian image denoising and compare of the new method versus the intersection of confidence interval (ICI) technique that also exploits a selection of an adaptive neighborhood for estimation.

1 Introduction

The nonparametric local regression originated in mathematical statistics offers an original approach to signal processing problems (e.g. [1], [2]). It basically results in linear filtering with the linear filters designed using some moving window local approximations. The first local pointwise (varying window size) adaptive nonparametric regression statistical procedure was suggested by Lepski [3], [4], [5] and independently by Goldenshluger and Nemirovsky [6]. This approach has received further development as the intersection of confidence interval (ICI) rule in application to various signal and image processing problems [7], [8], [9], [10]. The algorithm searches for a largest local vicinity of the point of estimation where the estimate fits well to the data. The estimates are calculated for a set of window sizes (scales) and compared. The adaptive window size is defined as the largest of those in the grid which estimate does not differ significantly from the estimators corresponding to the smaller window size.

In many applications the noise that corrupts the signal is non-Gaussian and signal dependent. There are a lot of heuristics adaptive-neighborhood approaches to filtering signal and images corrupted by signal-dependent noise. Instead of using fixed-size, fixed-shape neighborhoods, statistics of the noise and the signal are computed within variable-size, variable-shape neighborhoods that are selected for every point of estimation.

The Lepski approach allows a regular and theoretically well justified methodology for design of estimates with adaptive neighborhood. Unfortunately, it is originated from the Gaussian observation model and its modification to the signal dependent noise meets some principal difficulties. Another problem with applications of the Lepski method in practical situations is the choice of tuning parameters, especially of the threshold used for comparing two estimates from different scales. The theory only says that this threshold has to be

large enough (logarithmic in the sample size) and the theory only applies for such thresholds. At the same time, the numerical experiments indicate that a logarithmic threshold recommended by the theory is much too high and leads to a significant oversmoothing of the estimated function. Reasonable numerical results can be obtained by using smaller values of the threshold which shows the gap between the existing statistical theory and the practical applications.

The contribution of this paper is twofold: first, we propose a novel approach to design of the pointwise adaptive estimates especially for non-Gaussian distributions. Secondly, we address in details the question of selecting the parameters of the procedure and prove the theoretical results exactly for the algorithm we apply in numerical finite sample study.

The procedure is given for observations subject to the class of exponential distributions which includes the Poissonian model as an important special case. The fitted local likelihood is developed as statistics for selection of an adaptive size of this neighborhood. The estimated signal can be uni- and multivariable. The varying thresholds of the test-statistics is an important ingredient of approach. Special methods are proposed for selection of these thresholds. The fitted local likelihood approach is founded on theory justifying both the adaptive estimation procedure and the varying threshold selection. The main theoretical result formulated in Theorem 9 shows the accuracy of the adaptive estimate.

The proposed adaptive technique is applied for high-resolution imaging in a special form of anisotropic directional estimates using the size adaptive sectorial windows. The performance of the algorithm is illustrated for image denoising with data having Poissonian, Gaussian and Bernoulli observations. Simulation experiments demonstrate a quite good performance of the new algorithm.

Further, the paper is organized as follows. The nonparametric observation modeling and local likelihood estimates are discussed in Section 2. The local scale adaptive algorithm and the threshold selection are presented in Section 3. The theory of the approach is a subject of Section 4. The anisotropic implementation of the approach for high-resolution imaging is presented in Section 5. The simulation experiments are discussed in Section 6.

2 Observations and nonparametric modeling

This section describes our model and present some basic fact about nonparametric local maximum likelihood estimation.

2.1 Stochastic observations

Suppose we have *independent* random observations $\{Z_i\}_{i=1}^n$ of the form $Z_i = (X_i\,,\ Y_i)$. Here X_i denotes a vector of "features" or explanatory variables which determines the distribution of the " observation" Y_i. The d-dimensional vector $X_i \in \mathbb{R}^d$ can be viewed as a location in time or space and Y_i as the "observation at X_i" . Our model assumes that the values X_i are given and a distribution of each Y_i is determined by a parameter θ_i which may depend on the location $X_i\,,\ \theta_i = f(X_i)$. In many cases the *natural* parametrization is

chosen which provides the relation $\theta_i = E\{Y_i\}$. The estimation problem is to reconstruct $f(x)$ from the observations $\{Z_i\}_{i=1,\dots,n}$ for $x = X_i$.

Let us illustrate this set-up by few special cases.

1. *Gaussian regression.* Let $Z_i = (X_i, Y_i)$ with $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$ obeying the regression equation $Y_i = f(X_i) + \varepsilon_i$ with a regression function f and i.i.d. Gaussian errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. This observation model is standard one for many problems in signal and image processing.

2. *Poisson model.* Suppose that the random Y_i is a nonnegative integer subject to the Poisson distribution with the parameter $f(X_i)$, i.e., $Y_i \sim \mathcal{P}(f(X_i))$. The probability that Y takes the value k provided that $X_i = x$ is defined by the formula $P(Y_i = k|X_i = x) = f^k(x)\exp(-f(x))/k!$. This model occurs in digital camera imaging, queueing theory, positron emission tomography, etc.

3. *Bernoulli (binary response) model.* Let again $Z_i = (X_i, Y_i)$ with $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$ be a Bernoulli random variable with parameter $f(x)$, that is a probability that depends on $X_i = x$ that the random Y_i takes a value equal to one. It means that $P(Y_i = 1|X_i = x) = f(x)$, where $P(Y_i = 1|X_i = x)$ is a conditional probability. Such models arise in many econometric applications, and they are widely used in classification and digital imaging.

Now we describe the general setup. Let $\mathcal{P} = (P_\theta, \theta \in \Theta \subseteq \mathbb{R})$ be a parametric family of distributions dominated by a measure P. By $p(\cdot, \theta)$ we denote the corresponding density. We consider the regression-like model in which every "response" Y_i is, conditionally on $X_i = x$, distributed with the density $p(\cdot, f(x))$ for some unknown function $f(x)$ on $\mathcal{X}$ with values in Θ. The considered model can be written as $Y_i \sim P_{f(X_i)}$. This means that the distribution of every "observation" Y_i is described by the density $p(Y_i, f(X_i))$. In the considered situations with the independent observations Y_i, the joint distribution of the samples $Y_1, \dots, Y_n$ is given by the log-likelihood $L = \sum_{i=1}^n \log p(Y_i, f(X_i))$. In the literature similar regression-like models are also called *varying coefficient* or *nonparametrically driven* models.

Suppose for a moment that given y, the maximum of the density function $p(y, \theta)$ is achieved at $\theta = y$. This is the case for the above examples. Then the unconstrained maximization of the log-likelihood L w.r.t. the collection of parameter values $\theta = (\theta_1, \dots, \theta_n)^\top$ obviously leads to the trivial solution $\widetilde{\theta} = \operatorname{argmax}_{\{\theta_i\}} \sum_{i=1}^n \log p(Y_i, \theta_i) = Y$, where Y means the vector of observations. Thus, there is no smoothing and noise removal in this trivial estimate. It can be introduced assuming the correlation of the observations $\{Z_i\}_{i=1}^n$ or by use some model of the underlying function $f(x)$. The last idea is the most popular and exploited in a number of quite different forms.

2.2 Local likelihood modelling

In the simplest parametric setup, when the parameter θ does not depend on x, i.e., the distribution of every " observation" Y_i is the same, the invariant θ can be estimated well

by the parametric maximum likelihood method $\widetilde{\theta} = \operatorname{argmax}_\theta \sum_{i=1}^n \log p(Y_i, \theta)$.

In the nonparametric framework with varying $f(x)$, one usually applies the local likelihood approach which is based on the assumption that the parameter is nearly constant within some neighborhood of every point x in the "feature" space. This leads to considering a local model concentrated in some neighborhood of the point x.

We use localization by weights as a general method to describe a local model. Let, for a fixed x, nonnegative weights $w_{i,h}(x)$ be assigned to the observations Y_i. The weights $w_{i,h}(x)$ determine a local model corresponding to the point x in the sense that, when estimating the local parameter $f(x)$, the observations Y_i are used with these weights. This leads to the local maximum likelihood estimate

$$\widetilde{\theta}_h(x) = \operatorname*{argmax}_\theta \sum_i w_{i,h}(x) \log p(Y_i, \theta), \tag{1}$$

where the weight $w_{i,h}(x)$ usually depends on the distance between the point of estimation x and the location X_i corresponding to the "observation" Y_i. The index h means a *scale* (window size) parameter which can be a vector, see Section 5 for an example. Usually the weights $w_{i,h}(x)$ are selected in the form $w_{i,h}(x) = w\big(h^{-1}(x - X_i)\big)$, where $w(\cdot)$ is a fixed *window function* in $\mathbb{R}^d$ and h is the scale parameter. This window is often taken either in the product form $w(x) = \prod_{i=1}^n w_i(x_i)$ or in radial form $w(x) = w_1(\|x\|)$. We do not assume any special structure for the window function except that $w(0) = \max_x w(x)$. It means that the maximum weight is given to the observation with $X_i = x$.

2.3 Properties of the local MLE for a varying coefficient exponential family model

The examples of random observations considered in Section 2.1 are particular cases of the exponential family of distributions. This means that all distribution densities in (1) are of the form $p(y, \theta) = p(y) \exp(yC(\theta) - B(\theta)), \theta \in \Theta, y \in \mathcal{Y}$. Here $C(\theta)$ and $B(\theta)$ are some given non-negative functions of θ and $p(y)$ is some non-negative function of y. A *natural* parametrization for this family means the equality $\boldsymbol{E}_\theta Y = \int y p(y, \theta) P(dy) = \theta$ for all $\theta \in \Theta$. This condition is useful because the weighted average of observations is a natural unbiased estimate of θ. This section presents some results for on the properties of such local ML estimates. If $\mathcal{P} = (P_\theta)$ is an exponential family with the natural parametrization, the local log-likelihood and the local maximum likelihood estimates admit a simple closed form representation. For a given set of weights $\{w_{1,h}, \ldots, w_{n,h}\}$ with $w_{i,h} \in [0, 1]$, denote $N_h = \sum_{i=1}^n w_{i,h}$, $S_h = \sum_{i=1}^n w_{i,h} Y_i$. Note that the both sums depend on the location x via the weights $\{w_{i,h}\}$.

Lemma 1 (Polzehl and Spokoiny [11]) *It holds*

$$L_h(\theta) = \sum_{i=1}^n w_{i,h} \log p(Y_i, \theta) = S_h C(\theta) - N_h B(\theta) + R_h$$

where $R_h = \sum_{i=1}^n w_{i,h} \log p(Y_i)$. *Moreover,*

$$\widetilde{\theta}_h = S_h/N_h = \sum_{i=1}^n w_{i,h} Y_i \Big/ \sum_{i=1}^n w_{i,h} \tag{2}$$

and

$$L_h(\widetilde{\theta}_h, \theta) := L_h(\widetilde{\theta}_h) - L_h(\theta) = N_h \mathcal{K}(\widetilde{\theta}_h, \theta)$$

where $\mathcal{K}(\theta, \theta')$ *is the Kullback-Leibler divergence between two distributions with parameter values* θ *and* θ' *:* $\mathcal{K}(\theta, \theta') = E_\theta \log(p(Y, \theta)/p(Y, \theta')) = \int p(y, \theta) \log(p(y, \theta)/p(y, \theta'))dy$.

Here $L_h(\widetilde{\theta}_h, \theta)$ is a "fitted log-likelihood" defined as a difference between the maximized log-likelihood at $\theta = \widetilde{\theta}_h$ and the log-likelihood with an arbitrary θ , $L_h(\widetilde{\theta}_h, \theta) \geq 0$. Table 1 provides $K(\theta, \theta')$, $C(\theta)$, $B(\theta)$ for special cases of the exponential distribution considered above.

Table 1: The Kulback-Leibler divergence for the particular cases of the exponential family.

Model	$\mathcal{K}(\theta, \theta')$	$C(\theta)$	$B(\theta)$
Gaussian	$(\theta - \theta')^2/(2\sigma^2)$	θ/σ^2	$\theta^2/(2\sigma^2)$
Bernoulli	$\theta \log \dfrac{\theta}{\theta'} + (1-\theta) \log \dfrac{1-\theta}{1-\theta'}$	$\log \dfrac{\theta}{1-\theta}$	$\log \dfrac{1}{1-\theta}$
Poisson	$\theta \log \dfrac{\theta}{\theta'} - (\theta - \theta')$	$\log \theta$	θ

Now we present some rather tight exponential inequalities for the fitted log-likelihood $L_h(\widetilde{\theta}, \theta)$ in the parametric situation $\theta_i \equiv \theta^*$ for $i = 1, \ldots, n$ which apply to the arbitrary sample size and arbitrary weighting scheme. These results are essential for explaining our adaptive estimation procedure.

Theorem 2 (Polzehl and Spokoiny [11]) *Let* $\{w_{i,h}\}$ *be a localizing scheme such that* $\max_i w_{i,h} \leq 1$. *If* $f(X_i) \equiv \theta^*$ *for all* X_i *with* $w_{i,h} > 0$ *then for any* $\mathfrak{z} > 0$

$$\boldsymbol{P}_{\theta^*}(L_h(\widetilde{\theta}_h, \theta^*) > \mathfrak{z}) = \boldsymbol{P}_{\theta^*}\left(N_h \mathcal{K}(\widetilde{\theta}_h, \theta^*) > \mathfrak{z}\right) \leq 2e^{-\mathfrak{z}}.$$

In the regular situation, the Kullback-Leibler divergence $\mathcal{K}$ fulfills $\mathcal{K}(\theta, \theta^*) \approx I_{\theta^*}|\theta - \theta^*|^2$ for any point θ in a neighborhood of θ^* , where I_{θ^*} is the Fisher information at θ^* , see e.g. [12] or [13]. Therefore, the result of Theorem 2 guarantees that $|\widetilde{\theta}_h - \theta^*| \leq C N_h^{-1/2}$ with a high probability. Theorem 2 can be used for constructing the confidence intervals for the parameter θ^* .

Theorem 3 *If* $\mathfrak{z}_\alpha$ *satisfies* $2e^{-\mathfrak{z}_\alpha} \leq \alpha$, *then* $\mathcal{E}_h(\mathfrak{z}_\alpha) = \{\theta : N_h \mathcal{K}\big(\widetilde{\theta}_h, \theta\big) \leq \mathfrak{z}_\alpha\}$ *is an* α *-confidence set for the parameter* θ^* .

Theorem 3 claims that the estimation loss measured by $\mathcal{K}(\widetilde{\theta}_h, \theta)$ is with high probability bounded by $\mathfrak{z}_\alpha / N_h$ provided that $\mathfrak{z}_\alpha$ is sufficiently large. Similarly, one can establish a risk bound for a power loss function.

Theorem 4 *Let Y_i be i.i.d. from P_{θ^*}. Then for any $r > 0$*

$$\boldsymbol{E}_{\theta^*} \bigl| L_h(\widetilde{\theta}_h, \theta^*) \bigr|^r \equiv \boldsymbol{E}_{\theta^*} \bigl| N_h \mathcal{K}(\widetilde{\theta}_h, \theta^*) \bigr|^r \le \mathfrak{r}_r \,,$$
$$\mathfrak{r}_r = 2r \int_{\mathfrak{z} \ge 0} \mathfrak{z}^{r-1} e^{-\mathfrak{z}} d\mathfrak{z} = 2r \Gamma(r).$$

Proof. By Theorem 2

$$\boldsymbol{E}_{\theta^*} \bigl| L_h(\widetilde{\theta}_h, \theta^*) \bigr|^r \le - \int_{\mathfrak{z} \ge 0} \mathfrak{z}^r d \boldsymbol{P}_{\theta^*} (L_h(\widetilde{\theta}_h, \theta^*) > \mathfrak{z})$$
$$\le r \int_{\mathfrak{z} \ge 0} \mathfrak{z}^{r-1} \boldsymbol{P}_{\theta^*} (L_h(\widetilde{\theta}_h, \theta^*) > \mathfrak{z}) d\mathfrak{z} \le 2r \int_{\mathfrak{z} \ge 0} \mathfrak{z}^{r-1} e^{-\mathfrak{z}} d\mathfrak{z}$$

and the assertion follows. ■

3 Local scale selection algorithm

Let $\mathcal{H} = \{h_1, \ldots, h_K\}$ be a set of different scales ordered by the smoothing parameter h, and let $\widetilde{\theta}_h = S_h / N_h$ for $h \in \mathcal{H}$ be the corresponding set of estimates. For conciseness we use the notation $\widetilde{\theta}_k = \widetilde{\theta}_{h_k}$, $S_k = S_{h_k}$ and $N_k = N_{h_k}$. We also denote by $L_k(\theta)$ the log-likelihood for the scale h_k, $k = 1, \ldots, K$. We assume that the scale set $\mathcal{H}$ is *ordered* in the sense that the local sample size N_k grows with k.

The presented procedure aims at selecting one estimate $\widetilde{\theta}_k$ out of the given set in a data driven way to provide the best possible quality of estimation. This explains the notion of *local scale selection*. The fitted local likelihood (*FLL*) scale selection rule can be presented in the form [14]:

$$\widehat{k} = \max\{k : T_{lk} \le \mathfrak{z}_l, \; l < k\}, \tag{3}$$
$$T_{lk} = L_l(\widetilde{\theta}_l, \widetilde{\theta}_k) = N_l \mathcal{K}(\widetilde{\theta}_l, \widetilde{\theta}_k).$$

The procedure (3) can be interpreted as follows. The first estimate $\widetilde{\theta}_1$ is always accepted and (3) starts from $k = 2$. For the current estimate $\widetilde{\theta}_2$ is checked whether it belongs to the confidence set $\mathcal{E}_{h_1}(\mathfrak{z}_1)$ of the previous step estimate $\widetilde{\theta}_1$, see Theorem 3. If not, the estimate $\widetilde{\theta}_2$ is rejected and the procedure terminates selecting $\widetilde{\theta}_1$. If the inequality $T_{12} = L_1(\widetilde{\theta}_1, \widetilde{\theta}_2) \le \mathfrak{z}_1$ is fulfilled then $\widetilde{\theta}_2$ is accepted and the procedure considers the next step estimate $\widetilde{\theta}_3$. At every step k, the current estimate $\widetilde{\theta}_k$ is compared with all the previous estimates $\widetilde{\theta}_1, \ldots, \widetilde{\theta}_{k-1}$ by checking the inequalities (3). We proceed this way until the current estimates is rejected or the last estimate in the family for the largest scale is accepted. The adaptive estimate is the latest accepted one.

The proposed method can be viewed as a multiple testing procedure. The expressions $T_{lk} = L_l(\widetilde{\theta}_l, \widetilde{\theta}_k)$ is understood as test statistics for testing the hypothesis $H_{lk} : \boldsymbol{E}\, \widetilde{\theta}_l =$

$\boldsymbol{E}\,\widetilde{\theta}_k$, and $\mathfrak{z}_l$ is the corresponding critical value. At the step k the procedure tests the composite hypothesis $\boldsymbol{E}\,\widetilde{\theta}_1 = \ldots = \boldsymbol{E}\,\widetilde{\theta}_k$. The choice of the $\mathfrak{z}$'s is of special importance for the procedure and it is discussed in the next section.

The random index $\varkappa$ means the largest accepted k. The adaptive estimate $\widehat{\theta}$ is $\widetilde{\theta}_{\varkappa}$, $\widehat{\theta} = \widetilde{\theta}_{\varkappa}$. We also define the random moment $\varkappa_k$ meaning the largest index accepted after first k steps and the corresponding adaptive estimate: $\varkappa_k = \min\{\varkappa, k\}$, $\qquad \widehat{\theta}_k = \widetilde{\theta}_{\varkappa_k}$.

The *ICI* rule mentioned above can be presented in the sequential form (3) provided that the inequality $T_{lk} \le \mathfrak{z}_l$ is replaced by $|\widetilde{\theta}_l - \widetilde{\theta}_k| \le (\sigma_{\widetilde{\theta}_l} + \sigma_{\widetilde{\theta}_k})\mathfrak{z}$ where $\sigma_{\widetilde{\theta}_l}$ and $\sigma_{\widetilde{\theta}_k}$ are standard deviations of the estimates $\widetilde{\theta}_l$ and $\widetilde{\theta}_k$ and $\mathfrak{z}$ is the parameter similar to the varying $\mathfrak{z}_l$ in (3). Thus, to compare the estimates of different scales one has to additionally estimate their variances which in general, in particular for Poisson models, depend on unknown $f(x)$ and requires some recursive calculations, e.g. [10], [15]. Note, that the proposed procedure (3) does not need the estimate variance and the recursive calculations.

3.1 Choice of the parameter $\mathfrak{z}_k$

Following [14], the critical values $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$ are selected by the reasoning similar to the standard approach of hypothesis testing theory: to provide the prescribed performance of the procedure under the simplest (null) hypothesis. In the considered set-up, the null means $f(X_i) \equiv \theta^*$ for some fixed θ^* and all i. In this case it is natural to expect that the estimate $\widehat{\theta}_k$ coming out of the first k steps of the procedure is close to the nonadaptive counterpart $\widetilde{\theta}_k$. This particularly means that the probability of rejecting one of the estimates $\widetilde{\theta}_2, \ldots, \widetilde{\theta}_k$ under the null hypothesis should be very small.

Now we give a precise definition. Similarly to Theorem 4 the risk of estimation for an estimate $\widehat{\theta}$ of θ^* is measured by $\boldsymbol{E}\big|\mathcal{K}(\widehat{\theta}, \theta^*)\big|^r$ for some $r > 0$. Under the null hypothesis $f(X_i) \equiv \theta^*$, every estimate $\widetilde{\theta}_k$ fulfills by Theorem 4 for every $r > 0$

$$\boldsymbol{E}_{\theta^*}\big|L_k(\widetilde{\theta}_k, \theta^*)\big|^r = \boldsymbol{E}_{\theta^*}\big|N_k\mathcal{K}(\widetilde{\theta}_k, \theta^*)\big|^r \le \mathfrak{r}_r$$

for the fixed absolute constant $\mathfrak{r}_r$. We require that the parameters $\mathfrak{z}_1, \ldots, \mathfrak{z}_{K-1}$ of the procedure are selected in such a way that

$$\boldsymbol{E}_{\theta^*}\big|L_k(\widetilde{\theta}_k, \widehat{\theta}_k)\big|^r = \boldsymbol{E}_{\theta^*}\big|N_k\mathcal{K}(\widetilde{\theta}_k, \widehat{\theta}_k)\big|^r \le \alpha\mathfrak{r}_r\,, \qquad k = 2, \ldots, K. \tag{4}$$

Here α is the preselected constant having the meaning of the confidence level of procedure. This gives us $K-1$ conditions to fix $K-1$ critical values.

The condition (4) will be referred to as the *propagation property*. The meaning of "propagation" is that in the homogeneous situation the procedure passes with a high probability at every step from the current scale $k-1$ with the corresponding parameter h_{k-1} to a larger scale k with the parameter h_k. This yields that the adaptive estimate $\widehat{\theta}_k$ coincides with the nonadaptive counterpart $\widetilde{\theta}_k$ in the typical situation. These two estimates can be different only in the "false alarm" when one of the test statistics T_{lm} exceeds the critical value $\mathfrak{z}_l$ for some $l < m \le k$. The loss associated with such "false alarm" is

naturally measured by $\bigl|N_k\mathcal{K}(\widetilde{\theta}_k,\widehat{\theta}_k)\bigr|^r$ and the condition (4) gives the upper bound for the corresponding risk.

Our definition still involves two parameters α and r. It is important to mention that their choice is subjective and there is no way for an automatic local rule. One can apply the cross-validation type technique for a *global* data-driven tuning of these parameters, especially the parameter α. However, it is important to mention that a proper choice of the power r for the loss function as well as the "confidence level" α depends on the particular application and on the additional subjective requirements to the procedure. Taking a large r and small α would result in an increase of the critical values and therefore, improves the performance of the method in the parametric situation at cost of some loss of sensitivity to deviations from the parametric situation. Theorem 5 presents some upper bounds for the critical values $\mathfrak{z}_k$ as functions of α and r in the form $a_0 + a_1 \log \alpha^{-1} + a_2 r(K-k)$ with some coefficients a_0, a_1 and a_2. We see that these bounds linearly depend on r and on $\log \alpha^{-1}$. It is found in our experiments that a relatively small value $r = 1/2$ and $\alpha = 1$ are universally good for most image denoising scenarios. At least the thresholds found for these parameters give a very good initial guess for further threshold optimization. The set of conditions (4) do not directly define the critical values $\mathfrak{z}_k$. We present below two methods for selecting $\mathfrak{z}_k$ to provide these conditions.

3.1.1 The sequential choice

The first one is sequential and it is based on the decomposition

$$\mathcal{K}^r(\widetilde{\theta}_k,\widehat{\theta}_k) = \sum_{l=1}^{k-1} \mathcal{K}^r(\widetilde{\theta}_k,\widetilde{\theta}_l)\mathbf{1}\bigl(\varkappa = l\bigr)$$

for every $k \le K$. The idea is to specify the risk of estimation associated with every particular critical value starting from $\mathfrak{z}_1$. For this we run the procedure with only $\mathfrak{z}_1$ bounded and all the other values $\mathfrak{z}_k = \infty$ for $k \ge 2$. Define for $l > 1$ the events

$$\mathcal{A}_l^{(1)} = \{T_{12} \le \mathfrak{z}_1, \ldots, T_{1l} \le \mathfrak{z}_1\},\ \mathcal{B}_l^{(1)} = \{T_{12} \le \mathfrak{z}_1, \ldots, T_{1,l-1} \le \mathfrak{z}_1, T_{1l} > \mathfrak{z}_1\}.$$

The events $\mathcal{A}_l^{(1)}$ and $\mathcal{B}_l^{(1)}$ mean respectively that the estimate $\widetilde{\theta}_l$ is accepted and rejected when compared with $\widetilde{\theta}_1$. If the rejection happens too often, this is an indication that $\mathfrak{z}_1$ is too small. We therefore select $\mathfrak{z}_1$ as the minimal value providing that

$$\boldsymbol{E}_{\theta^*}\bigl|N_k\mathcal{K}(\widetilde{\theta}_k,\widehat{\theta}_k)\bigr|^r = \sum_{l=2}^{k} \boldsymbol{E}_{\theta^*}\bigl|N_k\mathcal{K}(\widetilde{\theta}_k,\widetilde{\theta}_{l-1})\bigr|^r \mathbf{1}\bigl(\mathcal{B}_l^{(1)}\bigr) \le \alpha\mathfrak{r}_r/(K-1) \tag{5}$$

for all $k = 2, \ldots, K$.

Similarly, we specify $\mathfrak{z}_2$ by considering the situation with the previously fixed $\mathfrak{z}_1$, some finite $\mathfrak{z}_2$ and all the remaining critical values equal to infinity, and so on. For the general definition, suppose that $\mathfrak{z}_1, \ldots, \mathfrak{z}_{j-1}$ have been already fixed for some $j > 1$ and define for any $\mathfrak{z}_j$ and all $k > j$ the events

$$\mathcal{B}_k^{(j)} = \mathcal{A}_k^{(1)} \cap \ldots \cap \mathcal{A}_k^{(j-1)} \cap \{T_{j,j+1} \le \mathfrak{z}_j, \ldots, T_{j,k-1} \le \mathfrak{z}_j, T_{jk} > \mathfrak{z}_j\}.$$

Then similarly to the first step, $\varkappa = k-1$ on $\mathcal{B}_k^{(j)}$ but now the estimate $\widetilde{\theta}_k$ is rejected when compared with $\widetilde{\theta}_j$. The related condition on the risk associated with $\mathfrak{z}_j$ can be written in the form

$$\sum_{l=j+1}^{k} \boldsymbol{E}_{\theta^*} \bigl| N_k \mathcal{K}(\widetilde{\theta}_{l-1}, \widetilde{\theta}_k) \bigr|^r \mathbf{1}\bigl(\mathcal{B}_l^{(j)}\bigr) \leq \alpha \mathfrak{r}_r / (K-1)$$

for all $k = j+1, \ldots, K$.

It is straightforward to check that such defined $\mathfrak{z}_k$ fulfill (4). It is also obvious that the choice of the critical values $\mathfrak{z}_k$ is determined by the joint distribution of the estimates $\widetilde{\theta}_k$ under the null hypothesis $H_0 : f(X_1) = \ldots = f(X_K) = \theta^*$.

3.1.2 Simplified parameter choice

Here we present a simplified procedure which is rather simple for implementation. It suggests to select $\mathfrak{z}_k$ linearly decreasing with k. This simplified selection of $\mathfrak{z}_k$ is based on the upper bound from Theorem 5 that there are constants a_0, a_1, and a_2 such that it holds for every $k \leq K$

$$\mathfrak{z}_k \leq a_0 + a_1 \log \alpha^{-1} + a_2 r \log(N_K / N_k). \tag{6}$$

This result justifies the linear rule

$$\mathfrak{z}_k = \mathfrak{z}_1 - \iota(K-k) \tag{7}$$

in the case when the local sample size measured by the value N_k grows exponentially with k. Then we only need to fix two parameters, e.g. the first value $\mathfrak{z}_1$ and the slop. We first identify the first value $\mathfrak{z}_1$ using the condition (5). The other values $\mathfrak{z}_k$ are found in the form $\mathfrak{z}_k = \mathfrak{z}_1 - \iota(k-1)$ to provide (4).

3.2 Details of implementation

To run the procedure, one has to first fix the set of local weighting schemes $(w_{i,h})$ for every scale parameter $h_1, \ldots, h_K$. The proposed algorithm applies to any such sequence which satisfies the growth condition **(*MD*)** from Section 4. A recommended choice is a geometric progression with the starting value h_1 and the growing factor $a > 1$. This means that $h_k = h_1 a^{k-1}$ for $k = 2, \ldots, K$. The starting bandwidth h_1 is usually the smallest possible value such that the first neighborhood only contains the reference point x. Our numerical results indicate that the procedure is quite stable w.r.t. to the growing factor a, and values in the range $[1.1, 1.5]$ lead to very reasonable estimation quality. The choice of critical values involves two more parameters α and r. Their meaning and impact has been already discussed before.

4 Theoretical study

This section presents some properties of the adaptive estimate $\widehat{\theta}$. We suppose that the parameters $\mathfrak{z}_k$ of the procedure are selected in such a way that the condition (4) is fulfilled. First we present some bounds on $\mathfrak{z}_k$ that ensure (4). Next we study the properties of $\widehat{\theta}$ in the parametric and local parametric situation. Finally we extend these results to the general nonparametric situation and prove an "oracle" property of $\widehat{\theta}$.

4.1 Bounds for the critical values

This section presents some upper and lower bounds for the critical values $\mathfrak{z}_k$. The results are established under the following condition on the local sample sizes N_k.

(*MD*) for some constants $\mathfrak{u}_0, \mathfrak{u}$ with $\mathfrak{u}_0 \le \mathfrak{u} < 1$, the values N_k satisfy for every $2 \le k \le K$ to the following conditions $N_{k-1} \le \mathfrak{u} N_k, \mathfrak{u}_0 N_k \le N_{k-1}$.

In addition, we need the following regularity condition on the parametric set Θ.

($\boldsymbol{\Theta}$) for some constants $\mathfrak{a}$ it holds for every sequence $\theta_0, \theta_1, \ldots, \theta_m \in \Theta$ that

$$\mathcal{K}^{1/2}(\theta_1, \theta_2) \le \mathfrak{a}\bigl\{\mathcal{K}^{1/2}(\theta_1, \theta_0) + \mathcal{K}^{1/2}(\theta_2, \theta_0)\bigr\},$$
$$\mathcal{K}^{1/2}(\theta_0, \theta_m) \le \mathfrak{a}\bigl\{\mathcal{K}^{1/2}(\theta_0, \theta_1) + \ldots + \mathcal{K}^{1/2}(\theta_{m-1}, \theta_m)\bigr\}.$$

[11] showed (Lemma 5.2) that this property is fulfilled under some mild regularity conditions on the parametric family $\mathcal{P}$. Our first result claims that in this situation under condition (*MD*) the parameters $\mathfrak{z}_k$ can be chosen in the form $\mathfrak{z}_k = \mathfrak{z}_K + \iota(K-k)$ to fulfill the "propagation" condition (4). The proof is given in the Appendix.

Theorem 5 *Assume (MD) and (Θ). Let $f(\cdot) \equiv \theta^*$. Then there are three constants a_0, a_1 and a_2 depending on r and $\mathfrak{u}_0$, $\mathfrak{u}$ only such that the choice $\mathfrak{z}_k = a_0 + a_1 \log \alpha^{-1} + a_2 r \log(N_K/N_k)$ ensures (4) for all $k \le K$. Particularly, $\boldsymbol{E}_{\theta^*} \bigl| N_K \mathcal{K}\bigl(\widetilde{\theta}_K, \widehat{\theta}\bigr) \bigr|^r \le \alpha \mathfrak{r}_r$.*

4.2 Risk of estimation in nonparametric situation. "Small modeling bias" condition

This section extends the bound of Theorem 4 to the nonparametric model $Y_i \sim P_{f(X_i)}$ when the function $f(\cdot)$ is not any longer constant even in a vicinity of the reference point x. We, however, suppose that the function $f(\cdot)$ can be well approximated by a constant θ for all points X_i from a neighborhood of x. Let $Z_\theta = d\boldsymbol{P}/d\boldsymbol{P}_\theta$ be the likelihood ratio of the underlying measure $\boldsymbol{P}$ w.r.t. the parametric measure $\boldsymbol{P}_\theta$ corresponding to the constant parameter $f(\cdot) \equiv \theta$. Then $\log Z_\theta = \sum_i \log \frac{p(Y_i, f(X_i))}{p(Y_i, \theta)}$. If we restrict our analysis to a neighborhood U and denote by $\boldsymbol{P}_U$ (respectively $\boldsymbol{P}_{U,\theta}$) the distribution of the observations Y_i for $X_i \in U$, then in a similar way $\log Z_{U,\theta} := \log \frac{d\boldsymbol{P}_{U,\theta}}{d\boldsymbol{P}_U} = \sum_{X_i \in U} \log \frac{p(Y_i, f(X_i))}{p(Y_i, \theta)}$. To

measure the quality of the approximation of the underlying measure $\boldsymbol{P}_U$ by the parametric measure $\boldsymbol{P}_{U,\theta}$, define $\Delta_U(\theta) = \sum_{X_i \in U} \mathcal{K}\bigl(f(X_i), \theta\bigr)$, where $\mathcal{K}\bigl(f(X_i), \theta\bigr)$ means the Kullback-Leibler distance between two parameter values $f(X_i)$ and θ.

Now we define for every scale h_k the neighborhood U_k which includes all the points X_i with $w_{i,h_k} > 0$ and write $\mathcal{F}_k$ instead of $\mathcal{F}_{U_k}$. Define $\Delta_k(\theta) = \sum_{X_i : w_{i,h_k} > 0} \mathcal{K}\bigl(f(X_i), \theta\bigr)$.

By Theorem 4 $\boldsymbol{E}_\theta \bigl| N_k \mathcal{K}(\widetilde{\theta}_k, \theta) \bigr|^r \le \mathfrak{r}_r$ for all k. We now aim to extend this result to the nonparametric situation under the "small modeling bias" condition $\Delta_k(\theta) \le \Delta$ for some $\Delta \ge 0$.

Theorem 6 *Let for some $\theta \in \Theta$ and some $\Delta \ge 0$*

$$\Delta_k(\theta) \le \Delta. \tag{8}$$

Then it holds for $r > 0$

$$\boldsymbol{E} \log\Bigl(1 + \bigl| N_k \mathcal{K}(\widetilde{\theta}_k, \theta) \bigr|^r / \mathfrak{r}_r\Bigr) \le \Delta + 1.$$

This result means that in the nonparametric situation under the condition (8) with some fixed Δ the losses $\bigl| N_k \mathcal{K}(\widetilde{\theta}_k, \theta) \bigr|^r$ are stochastically bounded. Note that this result applies even with if Δ is large, however the bound is proportional to $e^{\Delta+1}$ and grows exponentially with Δ.

4.3 "Stability after propagation" and "oracle" results

The notion of "oracle" result and "oracle" quality becomes more and more popular in statistical literature**.** Our analysis in Section 4.2 suggests the following definition of the "oracle" or "ideal" choice k^* of the scale parameter k: it is the largest value for which the "small modeling bias" condition $\Delta_k(\theta) \le \Delta$ is fulfilled for all $k \le k^*$. The corresponding "oracle" risk $\boldsymbol{E}\, \mathcal{K}^r\bigl(\widetilde{\theta}_{k^*}, \theta\bigr)$ is of order $1/N_{k^*}^r$. We aim to build the estimate $\widehat{\theta}$ which provides the same quality of estimation but does not rely to the "oracle". Our main result, see Theorem 9 below, shows that the proposed method possesses such an "oracle" feature: the difference between the "oracle" estimate $\widetilde{\theta}_{k^*}$ and the adaptive estimate $\widehat{\theta}$ measured by $\mathcal{K}^r\bigl(\widetilde{\theta}_{k^*}, \widehat{\theta}\bigr)$ is of order of the "oracle" risk $N_{k^*}^{-r}$.

The "propagation" result of Theorem 6 applies as long as the "small modeling bias" condition $\Delta_k(\theta) \le \Delta$ is fulfilled. To establish the accuracy result for the final estimate $\widehat{\theta}$, we have to check that the aggregated estimate $\widehat{\theta}_k$ does not vary much at the steps "after propagation" when the divergence $\Delta_k(\theta)$ from the parametric model becomes large.

Theorem 7 *It holds for every $k \le K$*

$$N_k \mathcal{K}\bigl(\widehat{\theta}_k, \widehat{\theta}\bigr) \le \mathfrak{z}_k .$$

Remark 8 *An interesting feature of this result is that it is fulfilled without any condition and with probability one, that is, the control of stability "works" not only with a high probability, it always applies. This property follows directly from the construction of the procedure.*

Proof. The result follows by the definition of $\widehat{\theta} = \widetilde{\theta}_{\varkappa}$ and $\widehat{\theta}_k = \widetilde{\theta}_{\varkappa_k}$ because $\varkappa_k \le \varkappa$ and $\widetilde{\theta}_{\varkappa}$ is accepted. ■

The "stability" result of Theorem 7 and condition (Θ) imply

$$\mathcal{K}^{1/2}\bigl(\widetilde{\theta}_{k^*}, \widehat{\theta}\bigr) \le \mathfrak{a}\mathcal{K}^{1/2}\bigl(\widetilde{\theta}_{k^*}, \widehat{\theta}_{k^*}\bigr) + \mathfrak{a}\mathcal{K}^{1/2}\bigl(\widehat{\theta}_{k^*}, \widehat{\theta}\bigr)$$

for some fixed constant $\mathfrak{a} \ge 1$. Moreover, for any $r > 0$

$$\mathcal{K}^{r/2}\bigl(\widetilde{\theta}_{k^*}, \widehat{\theta}\bigr) \le 2^{(r-1)_+}\mathfrak{a}^r\bigl\{\mathcal{K}^{r/2}\bigl(\widetilde{\theta}_{k^*}, \widehat{\theta}_{k^*}\bigr) + \mathcal{K}^{r/2}\bigl(\widehat{\theta}_{k^*}, \widehat{\theta}\bigr)\bigr\}. \tag{9}$$

Combination of the "propagation" and "stability" statements implies the main result concerning the properties of the adaptive estimate $\widehat{\theta}$. We state the result for $r = 1/2$. An extension to an arbitrary $r > 0$ is obvious.

Theorem 9 *Assume* (MD) *and* (Θ). *Let* θ *and* k^* *be such that* $\max_{k \le k^*} \Delta_k(\theta) \le \Delta$ *for some* $\Delta \ge 0$. *Then*

$$\boldsymbol{E}\log\biggl(1 + \frac{\bigl|N_{k^*}\mathcal{K}\bigl(\widetilde{\theta}_{k^*}, \widehat{\theta}\bigr)\bigr|^{1/2}}{\mathfrak{a}\mathfrak{z}_{k^*}^{1/2}}\biggr) \le \log 2 + \Delta + \frac{\alpha\mathfrak{r}_{1/2}}{\mathfrak{z}_{k^*}^{1/2}}.$$

The presented result states a kind of "oracle" property of the proposed estimate $\widehat{\theta}$. Indeed, due to this result, the normalized stochastic loss $\bigl|N_{k^*}\mathcal{K}\bigl(\widetilde{\theta}_{k^*}, \widehat{\theta}\bigr)\bigr|^{1/2}/\mathfrak{z}_{k^*}^{1/2}$ is bounded in the sense of existence of its log-moment. The "oracle" accuracy from Theorem 6 is stated for the loss $\bigl|N_{k^*}\mathcal{K}\bigl(\widetilde{\theta}_{k^*}, \widehat{\theta}\bigr)\bigr|^{1/2}$. The factor $\mathfrak{z}_{k^*}^{1/2}$ in the risk bound comes from the stability result and it can be considered as a kind of "payment for adaptation". Due to Theorem 5, $\mathfrak{z}_{k^*}$ is bounded from above by $a_0 + a_1\log(\alpha^{-1}) + a_2 r\log(N_K/N_{k^*})$. Therefore, the risk of the aggregated estimate corresponds to the best possible risk among the family $\{\widetilde{\theta}_k\}$ for the choice $k = k^*$ up to a logarithmic factor in the sample size. Lepski, Mammen and Spokoiny [5] established a similar result in the regression setup for the pointwise adaptive Lepski procedure and showed that this result yields the rate of adaptive estimation $\bigl(n^{-1}\log n\bigr)^{1/(2+d)}$ under Lipschitz smoothness of the function $f(\cdot)$ and the usual design regularity, see [11] for more details. It was shown by Lepski [3] that in the problem of pointwise adaptive estimation this rate is optimal and cannot be improved by any estimation method.

5 Application to non-Gaussian image denoising

In many cases the image intensity is a typical anisotropic function demonstrating essentially different nonsymmetric behavior in different directions at each pixel. It follows that

a good local approximation can be achieved only in a non-symmetric neighborhood. To deal with these features oriented/directional estimators are used in many vision and image processing tasks, such as edge detection, texture and motion analysis, etc. To mention a few of this sort of techniques we refer to classical steerable filters [16] and recent new ridgelet and curvelet transforms [17].

In this paper in terms of the considered nonparametric regression approach we exploit starshaped size/shape adaptive neighborhoods built for each estimation point. Figure 1 illustrates this concept and shows sequentially: a local best ideal estimation neighborhood U^* (figure a) , a sectorial segmentation of the unit ball (figure b), and the sectorial approximation of U^* using the scales $h^*_\alpha = h^*(\alpha)$ defining the length of the corresponding sectors (figure b) in the direction α from the finite set of directions $\mathfrak{A}$. Varying size sectors of the length h^*_α enable one to get a good approximation of any neighborhood of the point x provided that it is a starshaped body. This leads to the problem of simultaneous data-driven choice of the set of parameters $h^*_\alpha, \alpha \in \mathfrak{A}$. This is, however, a difficult task encountering some technical and principal points. To be practical we use a procedure with independent selection of the parameters h^*_α for each direction $\alpha \in \mathfrak{A}$. The adaptive procedure applied to the directional estimates $\widetilde{\theta}_{\alpha,h}(x)$ defined as

$$\widetilde{\theta}_{\alpha,h}(x) = \sum_{i \in I_\alpha(x)} w_{i,h}(x) Y_i \Big/ \sum_{i \in I_\alpha(x)} w_{i,h}(x) \tag{10}$$

where $w_{i,h}(x) = w(|X_i - x|/h)$ for some univariate kernel $w(\cdot)$ and $I_\alpha(x)$ is the sectorial set in direction α.

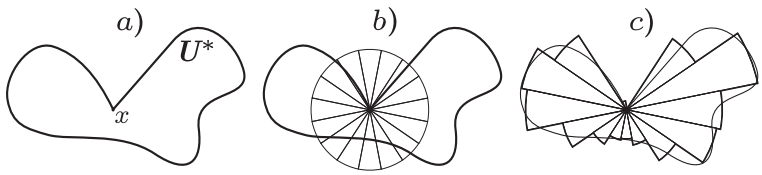

Figure 1: A neighborhood of the estimation point x: a) the best estimation set U^*, b) the unit ball segmentation, c) sectorial approximation of U^*.

With a given set of bandwidths $h_1, \ldots, h_K$ we come back to the problem of selecting for every direction α one of them in a data driven way. The adaptive procedure described in Section 3 leads to the value $\widehat{h}_\alpha(x)$.

When these adaptive scales $\widehat{h}_\alpha(x)$ are found for all $\alpha \in \mathfrak{A}$, the final estimate is calculated as the weighted mean of the observations included in the support of the neighborhoods:

$$\widehat{\theta}(x) = \sum_{\alpha \in \mathfrak{A}}{}' \sum_{i \in I_\alpha(x)} w_{i,\widehat{h}_\alpha(x)}(x) Y_i \Big/ \sum_{\alpha \in \mathfrak{A}}{}' \sum_{i \in I_\alpha(x)} w_{i,\widehat{h}_\alpha(x)}(x). \tag{11}$$

The sets $I_\alpha(x)$ have as a common point (intersection of the sets) at least the origin. The prime $(')$ in the formula (11) means that the estimate is calculated over the union of the directional supports $I_\alpha(x)$. Thus each observation enters in this formula only ones.

In (11) the argument x for $\widehat{h}_\alpha(x)$ indicates that the adaptive scales can be varying for each x. In the estimate (11) the adaptive procedure is used only in order to generate the adaptive neighborhood and the estimate is calculated as the weighted mean of the observations in this neighborhood.

There is another approach to the estimation problem. Let $\widehat{\theta}_\alpha(x)$ be the directional adaptive estimate calculated for the corresponding direction α, that is, $\widehat{\theta}_\alpha(x) = \widetilde{\theta}_{\alpha,\widehat{h}_\alpha(x)}(x)$, see (10). Define also $\widehat{\sigma}^2_\alpha(x) = \sigma^2_{\alpha,\widehat{h}_\alpha(x)}(x)$ where $\sigma^2_{\alpha,h}(x) = \sum_i w^2_{i,h} / (\sum_i w_{i,h})^2$ is the variance of $\widetilde{\theta}_{\alpha,h}$ from (10). Then the final estimate can be yield by fusing of the directional ones as follows

$$\widehat{\theta}(x) = \sum_{\alpha\in\mathfrak{A}} \lambda_\alpha(x)\widehat{\theta}_\alpha(x),\ \lambda_\alpha = \widehat{\sigma}^{-2}_\alpha(x) \Big/ \sum_{\alpha\in\mathfrak{A}} \widehat{\sigma}^{-2}_\alpha(x), \tag{12}$$

The *FLL* adaptive window sizes enable nearly constant value of θ in the starshaped neighborhood. It means that the observations in this neighborhood have equal variances and the variances $\widehat{\sigma}^2_\alpha$ in (12) can be calculated assuming that these variances of the observations are equal to one. The inverse variance weighting in (12) assumes that the directional estimates are unbiased and statistically independent. The estimate (11) is quite different from (12). In particular the origin is used here $T = \#(\mathfrak{A})$ times while it enters in (11) only ones. These estimates are quite competitive. In different cases one or another gives a better result.

The described adaptive starshaped neighborhood estimates are originated in the works [10], [18], where it is successfully exploited with the *ICI* adaptive scale selection for different image processing problems.

Formulas (11)–(12) make clear the algorithm. We introduce the directional estimates $\widehat{\theta}_\alpha(x)$, optimize the scalar scale parameter h_α for each of the directions (sectors) and use these adaptive directional sectors or directional estimates in order to calculate the final fused estimates.

Two points are of the importance here. First, we are able to find good approximations of estimation supports which can be of a complex form. Second, this approximation is composed from the univariate scale optimizations on h, thus the complexity is proportional to the number of sectors.

Multiple studies show that the finite sample performance of estimators based on bandwidth or model selection is often rather unstable, e.g. [19]. It is true for the local pointwise model selection considered in this paper. In spite of nice theoretical properties the *FLL* rule the resulting estimates suffer from a high variability due to a pointwise model choice, especially for a large noise level. In order to reduce the stochastic variability of the estimates the *FLL* algorithm is completed by special filtering of the adaptively selected $\widehat{h}_\alpha$. For this filtering we use a weighted median filters specially designed for each direction of the sectorial starshaped neighborhood. Thus, the adaptive directional estimates are defined as those after this median filtering. In the aggregation formulas (11)–(12) these filtered *FLL* estimates are used.

6 Experimental study

In these simulation experiments we demonstrate the performance of the developed algorithm for Poissonian and Gaussian image observations. It is assumed that the parameter θ is a deterministic unknown image intensity $f(x)$.

The image and the observations are defined on the finite discrete grid $x \in X = \{k_1, k_2 : k_1 = 1, 2, ..., n_1, k_2 = 1, 2, ..., n_2\}$ of the size $n_1 \times n_2$. It is assumed that the observations for each pixel are statistically independent. The problem is to reconstruct the image $f(x)$ from the observations $Y(x)$, $x \in X$. The following standard criteria are used:
(1) Root mean squared error ($RMSE$): $RMSE^2 = (n_1 n_2)^{-1} \sum_{x \in X} (f(x) - \widehat{\theta}(x))^2$;
(2) Signal-to-noise ratio (SNR) in dB: $SNR = 10 \log_{10}(\sum_{x \in X} |f(x)|^2 / \sum_{x \in X} |f(x) - \widehat{\theta}(x)|^2)$;
(3) Improvement in SNR ($ISNR$) in dB: $ISNR = 20 \log_{10}(\widehat{\sigma}_z / RMSE)$, where $\widehat{\sigma}_z$ is an estimate of the observation standard deviation;
(4) Peak signal-to-noise ratio ($PSNR$) in dB: $PSNR = 20 \log_{10}(\max_{x \in X} |f(x)| / RMSE)$.

For our experiments we use the $MATLAB$ texture test-images (8 bit gray-scale): *Boats* (512×512), *Lena* (512×512), *Cameraman* (256×256), *Peppers* (512×512) and two binary test-images: *Testpat1* (256×256) and *Cheese* (128×128). For the texture images we use eight line-wise directional estimators diagonal, vertical and horizontal with windowing function w. The line-wise supports enable high level of directional sensitivity of the adaptive estimators. The sectorial windows (of the angular size $\Delta\alpha \simeq 33.75^0$) work better then the line-wise ones for the images with comparatively large areas of constant or slowly varying intensities, in particular for the binary images considered in our simulation.

For every direction α, we apply the adaptive procedure for the set of window sizes $\mathcal{H}$ with a relatively small number of scales $K = 7$. For uniform linewise and sectorial windows the scale parameter h is integer with the set of values defined as $\mathcal{H} = \{\lfloor 1.5^k \rfloor, k = 1, ..., 7\} = \{1, 2, 3, 5, 7, 11, 17\}$. Then $N_k = h_k$ for all $k \leq K$. The fused estimates are calculated according to the formula (11).

A special study has been produced for testing the procedures presented in Section 3.1 for $\mathfrak{z}_k$ selection. For calculation of the expectations in the corresponding formulas we use Monte-Carlo simulation runs. In implementation of these calculations we accurately imitate the work of the adaptive FLL algorithm and use the adaptive estimates instead of the random event $\mathcal{B}_k^{(j)}$ introduced to check the inequalities $T_{lk} > \mathfrak{z}_l$. The developed algorithms for selection of $\mathfrak{z}_k$ give the results which depend on the parameters r and α, where r is the power of the used criterion functions and α is a parameter, similar to nominal rejection probability in hypothesis testing. These parameters are of purely mathematical origin, our default choice is $r = 1/2$ and $\alpha = 1$. These theoretical recommendations work surprisingly well giving the sets of $\mathfrak{z}_k$ universally good for quite different images and different distributions.

In what follows we use the sets $\mathfrak{z}_k$ obtained by the simplified threshold parameter choice (Section 3.1) with $r = 1/2$ and $\alpha = 1$. Of course, further optimization of $\mathfrak{z}_k$ can be produced for particular images or set of images but in any case what is found for $r = 1/2$

and $\alpha = 1$ can be treated as a good initial guess quite useful for further improvement.

6.1 Poissonian observations

To achieve different level of randomness (i.e. different SNR) in the Poissonian observations we multiply the true signal y by a scaling factor with the observations defined according to the formula $\widetilde{z} \sim \mathcal{P}(y \cdot \chi)$, where $\chi > 0$ is a scaling factor. Further, we assume the observations in the form $z = \widetilde{z}/\chi$ in order to have the results comparable for different χ as $E\{z\} = E\{\widetilde{z}\}/\chi = y$ for all $\chi > 0$. The scaling by χ allows to get the random data z with a different level the random noise and to preserve the mean value : $var\{z\} = var\{\widetilde{z}\}/\chi^2 = y/\chi$. The signal-to-noise ratio is calculated as $E\{z\}/\sqrt{var\{z\}} = \sqrt{y\chi}$. Thus, for larger and smaller χ we have respectively a larger and smaller signal-to-noise ratio.

This scaled modelling of Poisson data is appeared in a number of publications [20], [21], [22], [23] where the advanced performance of the wavelet based denoising algorithms is demonstrated. It is shown in [15] that the *ICI* based adaptive algorithm is quite competitive and at least numerically demonstrates a better performance then the algorithms in the cited papers. Here we compare the proposed *FLL* technique versus the *ICI* adaptive algorithm only.

In the scale selection the *FLL* technique is applied to the Poissonian variables, i.e. to $\widetilde{z}$. However, our linear estimates are calculated for the data $z = \widetilde{z}/\chi$. It means that in the formula for the Kullback divergence θ should be replaced by $\theta\chi$. Then the scale selection rule (3) for the Poissonian data (see the Kullback divergence in Table 1) is modified to the form $\widehat{k} = \max\{m,\ L_m(\widetilde{\theta}^{(m)}, \widetilde{\theta}^{(l)}) \le \mathfrak{z}_l/\chi,\ l < m\}$. In these experiments we use the line-wise and sectorial directional nonsymmetric windows of the scales $\mathcal{H}$. The threshold set calculated according to the simplified choice is as follows $\mathfrak{z} = \{1.2,\ 1.0,\ 0.8,\ 0.6,\ 0.4,\ 0.2\}$.

Table 2: "Cheese" image: criteria values for the eight directional and final estimates.

	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	Fused
$ISNR$, dB	7.62	6.62	7.67	6.84	7.50	6.56	7.76	6.95	16.59
SNR , dB	19.42	18.42	19.47	18.64	19.31	18.35	19.55	18.72	28.22
$PSNR$, dB	27.06	26.02	27.12	26.27	26.93	25.99	27.19	26.38	35.60
$RMSE$	11.32	12.71	11.25	12.39	11.48	12.81	11.15	12.23	4.23

The numerical results in Table 2 are given for the binary Cheeseïmage taking values $\theta = [0.2,\ 1.0]$. The criterion values for the fused (final) estimate compared with the eight directional sectorial ones show a strong improvement in the final estimate. In particular, we have for $ISNR$ the values about $7\ \ dB$ for the sectorial estimates while for the fused estimate $ISNR \simeq 16\ \ dB$. The fusing works very well for all criteria in Table 2. Visually, the improvement effects of the fusing are quite obvious.

Table 3 shows numerical criteria calculated for the test images. Values before and after

Table 3: Accuracy criterion for poissonian *FLL* imaging.

Test Image	*ISNR dB*	*SNR dB*	*PSNR dB*	*RMSE*
Cheese	**16.40**/10.68	**28.04**/22.47	**35.42**/30.1	**4.32**/7.97
Lena	10.65/**11.9**	22.17/**23.58**	27.85/**28.92**	10.32/**9.13**
Cameraman	**9.38**/9.20	**21.17**/21.04	**26.7**5/26.52	**11.71**/12.03
Peppers	10.98/**12.15**	22.58/**23.7**	28.33/**29.5**	9.76/**8.5**
Boats	9.20/**10.02**	20.84/**21.66**	26.19/**27.01**	12.50/**11.38**
Testpat1	9.64/**10.17**	23.31/**23.88**	24.93/**25.53**	14.45/**13.5**

slash correspond to the *FLL* and *LPA–ICI* recursive (after 7 iterations) algorithms, respectively. Numerically the *FLL* algorithm works better for *Cheese* and *Cameraman* while for the other images the *LPA–ICI* algorithm gives better criterion values. However, visual comparison is definitely in favor of the *FLL* algorithm. The recursive *LPA–ICI* estimates typically suffer from multiple spot-like artifacts while the *FLL* estimates are free from this sort of degradation effects.

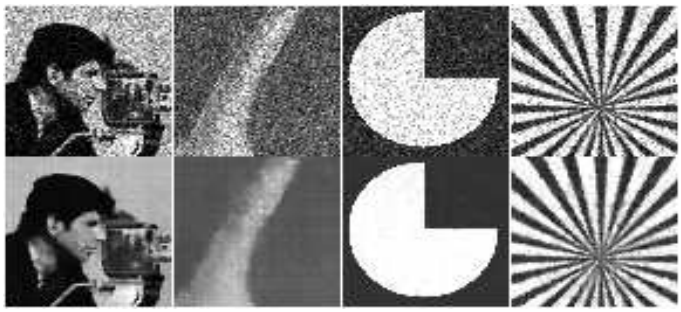

Figure 2: Fragments of noisy and denoised Poissonian images: *Cameraman*, *Peppers*, *Cheese*, *Testpat1*.

Fragments of noisy and denoised (by *FLL* algorithm) images are shown in Figure 2. Overall Table 3 confirms a very good performance of the *FLL* algorithm for Poissonian data.

6.2 Gaussian observations

We assume that the additive zero-mean Gaussian noise has the variance $\sigma^2 = 0.01$. For the scales $\mathcal{H}$ the threshold set calculated according to the simplified choice is as follows

$\mathfrak{z} = \{2.5\,,\ 2.07\,,\ 1.64\,,\ 1.21\,,\ 0.78\,,\ 0.35\}$. Numerically (see Table 4) the performance of the *FLL* algorithm is better (*Cheese*, *Peppers,Testpat1)* or worse (for other images) than that for *LPA–ICI* algorithm. Overall, the compared algorithms are more less equivalent. Note that the referred non-recursive *LPA–ICI* algorithm is a specially designed and optimized for the Gaussian case while the *FLL* is demonstrated in the form universally applicable for the class of exponential distributions and further optimization can improve the performance.

Table 4: Accuracy criterion for Gaussian *FLL* imaging.

Test Image	*ISNR dB*	*SNR dB*	*PSNR dB*	*RMSE*
Cheese	**15.71**/15.26	**28.33**/27.81	**35.71**/35.19	**4.18**/4.43
Lena	9.26/**9.41**	23.59/**24.08**	29.27/**29.42**	8.77/**8.62**
Cameraman	8.00/**8.04**	22.38/**22.53**	27.97/**28.01**	10.18/**10.13**
Peppers	**9.66**/9.46	23.91/**24.72**	**29.67**/29.47	**8.37**/8.57
Boats	7.63/**7.81**	22.30/**22.47**	27.64/**27.82**	10.58/**10.36**
Testpat1	**8.05**/7.60	**26.4**/25.95	**28.02**/27.57	**10.13**/10.66

6.3 Bernoulli observations

Bernoulli imaging assumes that the observations z take random binary values $[0,1]$ subject to the Bernoulli distribution. The image intensity θ is the mean of this random variable to be reconstructed as a function of the argument x. The sample mean estimate is unbiased with the variance equal to $\theta(1-\theta)/n$ and $SNR = \sqrt{n\theta/(1-\theta)}$, where n is a number of the averaged observations. For $\theta = 0$ or $\theta = 1$ the Bernoulli observations are noiseless and give the accurate pattern of the image without any signal processing and averaging. However, for the values θ different from 0 and 1 the observations can be very noisy and difficult for imaging. We illustrate the performance of the FLL algorithm for the piece-wise invariant image intensity. In order to have noisy observations the values of the intensity function should be different from 0 and 1. We control the level of the randomness in the observations by the following transformation of the original $\widetilde{\theta} = 0,1$ using instead the image $\theta = \widetilde{\theta} \cdot \chi + 0.5(1-\chi)$, $0 < \chi < 1$. For this θ the Bernoulli random variable takes values 0 and 1 with the probabilities $\theta_0 = 0.5(1-\chi)$ and $\theta_1 = 0.5(1+\chi)$ respectively. The variance of these observations grows rapidly when χ takes smaller values.

Table 5: Accuracy criterion for binary Bernoulli imaging.

χ	*data SNR dB*	*ISNR dB*	*SNR dB*	*PSNR dB*	*RMSE*
0.8	1.62	16.53	18.14	27.03	11.34
0.85	2.81	16.23	19.04	27.84	10.34
0.9	4.51	15.37	19.88	28.53	9.55

The threshold set calculated according to the simplified choice is as follows $\mathfrak{z} = \{0.7,\ 0.686\,,\ 0.672\,,\ 0.658\,,\ 0.644\,,\ 0.63\}$. The modelling results are presented for the binary *Cheese*

image ($\widetilde{\theta} = 0, 1$) and the varying parameter χ. Results are shown in Table 5. The most noisy case corresponds to $\chi = 0.8$ with $SNR = 1.62$ for these observations. The lowest level of the noise corresponds to $\chi = 0.95$ with $SNR = 7.44$. Numerical criterion values in Table 5 confirms a good performance of the algorithm. Noisy and denoised images as well as the error of denoising are illustrated in Figure 3.

Figure 3: *Cheese* image: binary Bernoulli observations z, estimate errors $|\theta - \widehat{\theta}| \cdot 10$ and estimates $\widehat{\theta}$.

7 Conclusion

A novel technique is developed for spatially adaptive estimation. The fitted local likelihood statistics is used for selection of an adaptive size of this neighborhood. The algorithm is developed for quite a general class of observations subject to the exponential distribution. The estimated signal can be uni- and multivariable. The varying thresholds of the developed statistical test is an important ingredient of the approach. Special techniques are proposed for the pointwise and linear approximation selection of these threshold. The developed theory justifies both the adaptive estimation procedure and the varying threshold selection. The main theoretical result formulated in Theorem 9 shows the accuracy of the adaptive estimate. For high-resolution imaging the developed approach is implemented in the form of anisotropic directional estimation with fusing the scale adaptive sectorial estimates. The performance of the algorithm is illustrated for image denoising with data having Poissonian, Gaussian and Bernoulli (binary) random observations. Simulation experiments demonstrate a very good performance of the new algorithm. A demo version of the developed adaptive *FLL* algorithm and the scale selection procedures are available at the website *www.cs.tut.fi/˜lasip*.

8 Appendix

This section collects the proofs of the main results.

8.1 Proof of Theorem 5

Define for every $k \le K$ the random set $\mathcal{A}_k = \bigcap_{j=1}^{k-1} \Big\{ \max_{j<l\le k} T_{jl} \le \mathfrak{z}_j \Big\}$, where $T_{jl} = N_j \mathcal{K}\big(\widetilde{\theta}_j, \widetilde{\theta}_l\big)$. Note first that $\widehat{\theta}_k = \widetilde{\theta}_k$ on $\mathcal{A}_k$ for all $k \le K$.

Therefore, it remains to bound the risk of $\widehat{\theta}_k$ on the complement $\overline{\mathcal{A}}_k$ of $\mathcal{A}_k$. Define $\mathcal{B}_{k-1} = \mathcal{A}_{k-1} \setminus \mathcal{A}_k$. By definition $\varkappa = \varkappa_k = k-1$ on $\mathcal{B}_{k-1}$ and $\overline{\mathcal{A}}_k = \bigcup_{l<k} \mathcal{B}_l$. First we bound the probability $\boldsymbol{P}_0\big(\mathcal{B}_l\big)$. Assumptions (MD) and (Θ) yield for every $l < k$

$$T_{lk} \le 2\mathfrak{a}^2 N_l \big\{ \mathcal{K}(\widetilde{\theta}_l, \theta) + \mathcal{K}(\widetilde{\theta}_k, \theta) \big\} \le 2\mathfrak{a}^2 \big\{ N_l \mathcal{K}(\widetilde{\theta}_l, \theta) + N_k \mathcal{K}(\widetilde{\theta}_k, \theta) \big\}.$$

Therefore, by Theorem 4, for all $\lambda < 1$ and any θ

$$\boldsymbol{P}_\theta\big(\mathcal{B}_l\big) \le \sum_{j=1}^{l-1} \boldsymbol{P}_\theta\big(T_{jl} > \mathfrak{z}_j\big) \le 2(1-\lambda)^{-1} \sum_{j=1}^{l-1} e^{-\lambda \mathfrak{z}_j / \mathfrak{a}^2}.$$

Similarly for $l < k$

$$\begin{aligned}
&\boldsymbol{E}_\theta \big| N_k \mathcal{K}(\widetilde{\theta}_l, \widetilde{\theta}_k) \big|^r \le \mathfrak{a}^{2r} 2^{(r-1)_+} \big\{ \boldsymbol{E}_\theta \big| N_k \mathcal{K}(\widetilde{\theta}_l, \theta) \big|^r + \boldsymbol{E}_\theta \big| N_k \mathcal{K}(\widetilde{\theta}_k, \theta) \big|^r \big\} \\
&\le \mathfrak{a}^{2r} 2^{(r-1)_+} \Big\{ \frac{N_k^r}{N_l^r} \boldsymbol{E}_\theta \big| N_l \mathcal{K}(\widetilde{\theta}_l, \theta) \big|^r + \boldsymbol{E}_\theta \big| N_k \mathcal{K}(\widetilde{\theta}_k, \theta) \big|^r \Big\} \le \mathfrak{a}^{2r} 2^{r \vee 1} \mathfrak{r}_r N_k^r / N_l^r .
\end{aligned}$$

Now we employ the obvious representation $N_k \mathcal{K}\big(\widetilde{\theta}_k, \widehat{\theta}_k\big) = \sum_{l=1}^{k-1} N_k \mathcal{K}\big(\widetilde{\theta}_k, \widetilde{\theta}_l\big) \mathbf{1}\big(\mathcal{B}_l\big)$. Therefore, for every r and $\lambda < 1$ by the Cauchy-Schwartz inequality

$$\begin{aligned}
&\boldsymbol{E}_\theta \big| N_k \mathcal{K}\big(\widetilde{\theta}_k, \widehat{\theta}_k\big) \big|^r = \sum_{l=1}^{k-1} \boldsymbol{E}_0 \big| N_k \mathcal{K}\big(\widetilde{\theta}_k, \widetilde{\theta}_l\big)^2 \big|^r \mathbf{1}\big(\mathcal{B}_l\big) \\
&\le \sum_{l=1}^{k-1} \boldsymbol{E}_\theta^{1/2} \big| N_k \mathcal{K}\big(\widetilde{\theta}_k, \widetilde{\theta}_l\big) \big|^{2r} \boldsymbol{P}_\theta^{1/2}\big(\mathcal{B}_l\big) \le 2^{r \vee 1} \mathfrak{r}_{2r}^{1/2} (1-\lambda)^{-1/2} \sum_{l=1}^{k-1} \frac{N_k^r}{N_l^r} \Big(\sum_{j=1}^{l-1} e^{-\lambda \mathfrak{z}_j / \mathfrak{a}^2} \Big)^{1/2}.
\end{aligned}$$

It remains to check that the choice $\mathfrak{z}_j = a_0 + a_1 \log \alpha^{-1} + a_2 r \log(N_K / N_j)$ with properly selected a_0, a_1 and a_2 provide under condition (MD) the required bound $\boldsymbol{E}_\theta \big| N_k \mathcal{K}\big(\widetilde{\theta}_k, \widehat{\theta}_k\big) \big|^r \le \alpha \mathfrak{r}_r$ and Theorem 5 follows.

8.2 Proof of Theorem 6

The proof is based on the following general result.

Lemma 10 *Let $\boldsymbol{P}$ and $\boldsymbol{P}_0$ be two measures such that $\mathcal{K}(\boldsymbol{P}, \boldsymbol{P}_0) \le \Delta < \infty$. Then for any random variable ζ with $\boldsymbol{E}_0 \zeta < \infty$, $\boldsymbol{E} \log\big(1 + \zeta\big) \le \Delta + \boldsymbol{E}_0 \zeta$.*

Proof. By simple algebra one can check that for any fixed y the maximum of the function $f(x) = xy - x\log x + x$ is attained at $x = e^y$ leading to the inequality $xy \le x\log x - x + e^y$. Using this inequality and the representation $\boldsymbol{E}\log\bigl(1+\zeta\bigr) = \boldsymbol{E}_0\bigl\{Z\log\bigl(1+\zeta\bigr)\bigr\}$ with $Z = d\boldsymbol{P}/d\boldsymbol{P}_0$ we obtain

$$\begin{aligned}&\boldsymbol{E}\log\bigl(1+\zeta\bigr) = \boldsymbol{E}_0\bigl\{Z\log\bigl(1+\zeta\bigr)\bigr\}\\ &\le \boldsymbol{E}_0\bigl(Z\log Z - Z\bigr) + \boldsymbol{E}_0(1+\zeta) = \boldsymbol{E}_0\bigl(Z\log Z\bigr) + \boldsymbol{E}_0\zeta - \boldsymbol{E}_0 Z + 1.\end{aligned}$$

It remains to note that $\boldsymbol{E}_0 Z = 1$ and $\boldsymbol{E}_0\bigl(Z\log Z\bigr) = \boldsymbol{E}\log Z = \mathcal{K}(\boldsymbol{P},\boldsymbol{P}_0)$. ■

We now apply this lemma with $\zeta = \bigl|N_k\mathcal{K}(\widetilde{\theta}_k,\theta)\bigr|^r/\mathfrak{r}_r$ and utilize that $\boldsymbol{E}_0\zeta \le 1$. This yields

$$\boldsymbol{E}_\theta\bigl(Z_{k,\theta}\log Z_{k,\theta}\bigr) = \boldsymbol{E}\log Z_{k,\theta} = \boldsymbol{E}\sum_{X_i: w_{i,h_k}>0}\log\frac{p(Y_i,f(X_i))}{p(Y_i,\theta)} = \Delta_k(\theta) \le \Delta$$

and the assertion follows.

8.3 Proof of Theorem 9

By (9) and Theorem 7

$$1 + \frac{\bigl|N_{k^*}\mathcal{K}\bigl(\widetilde{\theta}_{k^*},\widehat{\theta}\bigr)\bigr|^{1/2}}{\mathfrak{a}\mathfrak{z}_{k^*}^{1/2}} \le 2 + \frac{\bigl|N_{k^*}\mathcal{K}\bigl(\widetilde{\theta}_{k^*},\widehat{\theta}_{k^*}\bigr)\bigr|^{1/2}}{\mathfrak{z}_{k^*}^{1/2}}.$$

Now by Lemma 10 and Theorem 7

$$\begin{aligned}&\boldsymbol{E}\log\biggl(1 + \frac{\bigl|N_{k^*}\mathcal{K}\bigl(\widetilde{\theta}_{k^*},\widehat{\theta}\bigr)\bigr|^{1/2}}{\mathfrak{a}\mathfrak{z}_{k^*}^{1/2}}\biggr)\\ &\le \log 2 + \Delta + \boldsymbol{E}_\theta\frac{\bigl|N_{k^*}\mathcal{K}\bigl(\widetilde{\theta}_{k^*},\widehat{\theta}_{k^*}\bigr)\bigr|^{1/2}}{2\mathfrak{z}_{k^*}^{1/2}} \le \log 2 + \Delta + \alpha\mathfrak{r}_{1/2}/\mathfrak{z}_{k^*}^{1/2},\end{aligned}$$

and the required assertion follows.

References

[1] J. Fan J. and I. Gijbels, *Local polynomial modelling and its application.* London: Chapman and Hall, 1996.

[2] C. Loader, *Local regression and likelihood,* Series Statistics and Computing, Springer-Verlag New York, 1999.

[3] O.V. Lepski, One problem of adaptive estimation in Gaussian white noise, *Theory Probab. Appl.*, vol. 35, no. 3, pp. 459 - 470, 1990.

[4] O. Lepski and V. Spokoiny, "Optimal pointwise adaptive methods in nonparametric estimation," *The Annals of Statistics,* vol. 25, no. 6, pp. 2512–2546, 1997.

[5] O. Lepski, E. Mammen and V. Spokoiny, "Ideal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selection," *The Annals of Statistic*s, vol. 25, no. 3, 929–947, 1997.

[6] A. Goldenshluger and A. Nemirovski, "On spatial adaptive estimation of nonparametric regression", *Math. Meth. Statistics*, vol.6, pp.135-170, 1997.

[7] V. Katkovnik, "A new method for varying adaptive bandwidth selection," *IEEE Trans. Sig. Proc.*, vol. 47, no. 9, pp. 2567-2571, 1999.

[8] V. Katkovnik, K. Egiazarian and J. Astola, "Adaptive window size image de-noising based on intersection of confidence intervals (ICI) rule," *Journal of Math. Imaging and Vision*, vol. 16, no. 3, pp. 223-235, 2002.

[9] LJ. Stanković, "Performance analysis of the adaptive algorithm for bias-to-variance trade-off," *IEEE Trans. Sig. Proc.*, vol. 52, No. 5, pp. 1228-1234, 2004.

[10] A. Foi, *Anisotropic nonparametric image processing: theory, algorithms and applications,* Ph.D. Thesis, Dip. di Matematica, Politecnico di Milano, ERLTDD-D01290, April 2005. Available: $www.cs.tut.fi/\tilde{}lasip$.

[11] J. Polzehl and V. Spokoiny, "Propagation-separation approach for local likelihood estimation, *Probab. Theory Related Fields*, vol. 135, no. 3, 335–362, 2005.

[12] I. Ibragimov and R. Khasminskii, *Statistical estimation.* Springer-Verlag New York, 1981.

[13] S. Kullback, *Statistics and Information Theory.* Wiley and Sons, New York, 1959.

[14] V. Spokoiny, *Local parametric methods in nonparametric estimation*, Springer, 2006. To appear.

[15] A. Foi, A., R. Bilcu, V. Katkovnik, and K. Egiazarian, "Anisotropic local approximations for pointwise adaptive signal-dependent noise removal", *Proc. XIII European Signal Process. Conf., EUSIPCO 2005*, Antalya, September 2005.

[16] W.T. Freeman and E.H. Adelson, "The design and use of steerable filters," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 9, pp. 891-906, 1991.

[17] J.L. Starck, E.J. Candes, and D.L. Donoho, "The curvelet transform for image denoising," *IEEE Trans. Image Processing*, vol. 11, no. 6, pp. 670-684, 2002.

[18] V. Katkovnik, A. Foi, K. Egiazarian, and J. Astola, "Directional varying scale approximations for anisotropic signal processing", *Proc. of XII European Signal Process. Conf., EUSIPCO 2004*, pp. 101-104, 2004.

[19] L. Breiman, Stacked regression, *Machine Learning*, 24 pp. 49-64, 1996.

[20] K.E. Timmermann and R. Nowak, "Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging," *IEEE Trans. Information Theory,* vol. 45, no. 3, pp. 846-862, 1999.

[21] R.D. Nowak and R. G. Baraniuk, "Wavelet-domain filtering for photon imaging systems," *IEEE Trans. Image Processing,* vol. 8, no. 5, pp. 666-678, 1999.

[22] R. M. Willett and R. D. Nowak, "Platelets: a multiscale approach for recovering edges and surfaces in photon-limited medical imaging", *IEEE Trans Medical Imaging,* vol. 22, no. 3, pp. 332-350, 2003.

[23] H. Lu, Kim Y. and J. M. M. Anderson, "Improved Poisson intensity estimation: denoising application using Poisson data," *IEEE Trans. Image Processing,* vol. 13, no. 8, pp. 1128–1135, 2004.