Spatially Adaptive Estimation via Fitted Local Likelihood Techniques

Vladimir Katkovnik and Vladimir Spokoiny

***Abstract*—This paper offers a new technique for spatially adaptive estimation. The local likelihood is exploited for nonparametric modeling 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 are used for selection of an adaptive size and shape of this neighborhood. The algorithm is developed for a quite 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 image denoising and compare the new method versus the intersection of confidence interval (ICI) technique that also exploits a selection of an adaptive neighborhood for estimation.**

***Index Terms*—Adaptive non-Gaussian image denoising, adaptive nonparametric regression, anisotropic imaging, fitted local likelihood (FLL), non-Gaussian denoising, Poissonian denoising, varying threshold parameters.**

I. Introduction

THE nonparametric regression originated in mathematical statistics offers an original approach to signal processing problems (e.g., [1] and [2]). It basically results in linear filtering with the linear filters designed using some moving window local approximations. In many applications like speech recognition or image denoising, nonlinear or locally adaptive methods have been shown to be more efficient than the linear ones. The typical examples are given by non-linear wavelet thresholding, [3], and pointwise adaptive kernel smoothing, [4], [5]. The first local pointwise (varying window size) adaptive nonparametric regression statistical procedure was suggested by Lepski [6]; see also [4], [5], and[7]. This approach has received further development as the "intersection of confidence interval" (ICI) rule in application to various signal and image processing problems [8]–[12]. 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 general methodology for design of estimates with adaptive neighborhood. However, 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 general 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. Second, 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 used as statistics for selection of the adaptive estimation neighborhoods. 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 6 shows the accuracy of the adaptive estimate.

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

Further, the paper is organized as follows. The nonparametric observation modeling and local likelihood estimates are discussed in Section II. The local scale adaptive algorithm and the threshold selection are presented in Section III. The anisotropic implementation of the approach for imaging is

Manuscript received July 17, 2006; revised June 3, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Ilya Pollak. This work was supported by the Academy of Finland, project No. 213462 Finnish Centre of Excellence Program (2006–2011).

V. Katkovnik is with the Signal Processing Institute, University of Technology of Tampere, Tampere, Finland (e-mail: katkov@cs.tut.fi).

V. Spokoiny is with the Weierstrass Institute for Applied Analysis and Stochastics, D-10117 Berlin, Germany (e-mail: spokoiny@wias-berlin.de).

Digital Object Identifier 10.1109/TSP.2007.907873

1053-587X/$25.00 © 2008 IEEE

presented in Section IV. The simulation experiments are discussed in Section V. The theory of the approach is a subject of Section VI.

II. Observations and Nonparametric Modeling

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

A. 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 $X_i, \theta_i = \theta(X_i)$. In many cases, the *natural* parametrization is chosen which provides the relation $\theta_i = E\{Y_i\}$. The estimation problem is to reconstruct $\theta(x)$ from the data $\{Z_i\}_{i=1,\dots,n}$.

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 $Y_i = \theta(X_i) + \varepsilon_i$ with a regression function θ 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 $\theta(X_i)$. The probability that $Y_i = k$ given $X_i = x$ is defined by the formula $P(Y_i = k | X_i = x) = \theta^k(x)\exp(-\theta(x))/k!$. This model occurs in digital camera imaging, queueing theory, positron emission tomography, etc.
3) *Bernoulli (binary response) model*. Let Y_i be independent Bernoulli random variables with parameters θ that depends on the d-dimensional vector of "features" $X_i \in \mathbb{R}^d$. This means that $P(Y_i = 1 | X_i = x) = \theta(x)$. 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 $p(\cdot, \theta(x))$ for some unknown function $\theta(x)$ on $\mathcal{X}$ with values in Θ. The considered model can be written as $Y_i \sim P_{\theta(X_i)}$. This means that the distribution of every "observation" Y_i is described by the density $p(Y_i, \theta(X_i))$. 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 $L = \sum_{i=1}^n \log p(Y_i, \theta(X_i))$. 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 $\tilde{\theta} = \arg\max_{\{\theta_i\}} \sum_{i=1}^n \log p(Y_i, \theta_i) = Y$, 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 $\theta(x)$. The last idea is the most popular and it is exploited in a number of quite different forms.

B. Local Constant Likelihood Modeling

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 $\tilde{\theta} = \arg\max_\theta \sum_{i=1}^n \log p(Y_i, \theta)$.

In the nonparametric framework with varying $\theta(x)$, 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 $\theta(x)$, the observations Y_i are used with these weights. This leads to the local likelihood $L_h(\theta) = \sum_{i=1}^n w_{i,h} \log p(Y_i, \theta)$ and the local maximum-likelihood estimate (MLE) defined as

$$\tilde{\theta}_h(x) = \arg\max_\theta L_h(\theta). \tag{1}$$

The weight $w_{i,h}(x)$ in $L_h(\theta)$ 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, bandwidth*) parameter which can be a vector, see Section IV 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 $w(x) = w_1(\|x\|)$. In general, 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 $X_i = x$.

C. Local MLE for the Exponential Family Model

The examples of random observations considered in Section II-A are particular cases of the exponential family of distributions defined in the form $p(y, \theta) = p(y)\exp(yC(\theta) - B(\theta)), \theta \in \Theta, y \in \mathbb{R}$. Here $C(\theta)$ and $B(\theta)$ are some given non-negative functions of θ (see Table I) and $p(y)$ is some non-negative function of y.

For this exponential family the local MLE admits a close form representation as the weighted mean of the observed Y_i. For a given set of weights $\{w_{1,h}, \dots, w_{n,h}\}$ 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 $\{w_{i,h}\}$.

Theorem 1: The local likelihood estimate $\tilde{\theta}_h$ can be represented in the form

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

Moreover, for any θ the difference $L_h(\tilde{\theta}_h, \theta) := L_h(\tilde{\theta}_h) - L_h(\theta)$ reads as

$$L_h(\tilde{\theta}_h, \theta) = N_h \mathcal{K}(\tilde{\theta}_h, \theta) \tag{3}$$

where $\mathcal{K}(\theta, \theta') := E_\theta \log p(y, \theta)/p(y, \theta')$ is the Kullback–Leibler divergence between two measures P_θ and $P_{\theta'}$.

Proof: The definition of $L_h(\theta)$ implies the following representation for $L_h(\theta)$:

$$L_h(\theta) = S_h C(\theta) - N_h B(\theta) + R_h \tag{4}$$

where $R_h = \sum_{i=1}^{n} w_{i,h} \log p(Y_i)$. Differentiating the normalizing condition $\int p(y) \exp\{yC(\theta) - B(\theta)\} dy = 1$ w.r.t. θ together with the condition $E_\theta\{y\} = \theta$ yields the identity $\theta \partial_\theta C(\theta) = \partial_\theta B(\theta)$ for every θ. The estimate $\tilde{\theta}_h$ maximizes $S_h C(\theta) - N_h B(\theta)$ and hence fulfills the equation $S_h \partial_\theta C(\theta) - N_h \partial_\theta B(\theta) = 0$. Substituting in this equation $\partial_\theta B(\theta) = \theta \partial_\theta C(\theta)$ at $\theta = \tilde{\theta}_h$ leads to (2).

Simple algebra yields $\mathcal{K}(\theta, \theta') = \theta\{C(\theta) - C(\theta')\} - B(\theta) + B(\theta')$ and (3) follows by direct substitution of $\tilde{\theta}_h$ in (4). See [13] for more details. ∎

The value $L_h(\tilde{\theta}_h, \theta') = \max_\theta L_h(\theta, \theta')$ is called the *fitted log-likelihood* and it plays an important role in our adaptive procedure.

An important advantage of the maximum-likelihood approach is that it allows to infer on the value of the unknown parameter on the base of the fitted likelihood in the pure parametric situation with $\theta(X_i) \equiv \theta^*$. The basic fact is given by the following rather tight deviation bound for $L_h(\tilde{\theta}_h, \theta)$.

Theorem 2 (Polzehl and Spokoiny [13]): Let $\{w_{i,h}\}$ be a localizing scheme such that $\max_i w_{i,h} \le 1$. If $f(X_i) \equiv \theta^*$ for all X_i with $w_{i,h} > 0$ then for any $\mathfrak{z} > 0$

$$\mathbf{P}_{\theta^*}(L_h(\tilde{\theta}_h, \theta^*) > \mathfrak{z}) = \mathbf{P}_{\theta^*}(N_h \mathcal{K}(\tilde{\theta}_h, \theta^*) > \mathfrak{z}) \le 2e^{-\mathfrak{z}}.$$

This bound is particularly useful for obtaining the risk bounds, testing the hypotheses and building confidence sets in the parametric case. It is worth noting that this result is non-asymptotic and valid for an arbitrary local sample size N_h. This is especially important for our adaptive procedure which starts with the very small vicinity of the point of interest x.

Theorem 3: Under the conditions of Theorem 2, if $\mathfrak{z}_\alpha$ satisfies $2e^{-\mathfrak{z}_\alpha} \le \alpha$, then

$$\mathcal{E}_h(\mathfrak{z}_\alpha) = \{\theta : N_h \mathcal{K}(\tilde{\theta}_h, \theta) \le \mathfrak{z}_\alpha\} \tag{5}$$

is an α-confidence set for the parameter θ^*.

Moreover, for any $r > 0$

$$\boldsymbol{E}_{\theta^*} |L_h(\tilde{\theta}_h, \theta^*)|^r \equiv \boldsymbol{E}_{\theta^*} |N_h \mathcal{K}(\tilde{\theta}_h, \theta^*)|^r \le \mathfrak{r}_r$$

TABLE I
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	$\theta \log \frac{\theta}{\theta'} + (1-\theta) \log \frac{1-\theta}{1-\theta'}$	$\log \frac{\theta}{1-\theta}$	$\log \frac{1}{1-\theta}$
Poisson	$\theta \log \frac{\theta}{\theta'} - (\theta - \theta')$	$\log \theta$	θ

with

$$\mathfrak{r}_r = 2r \int_{\mathfrak{z} \ge 0} \mathfrak{z}^{r-1} e^{-\mathfrak{z}} d\mathfrak{z} = 2r\Gamma(r).$$

Proof: The first statement follows immediately from Theorem 2. Similarly

$$\begin{aligned} \boldsymbol{E}_{\theta^*} |L_h(\tilde{\theta}_h, \theta^*)|^r &\le - \int_{\mathfrak{z} \ge 0} \mathfrak{z}^r d\boldsymbol{P}_{\theta^*}(L_h(\tilde{\theta}_h, \theta^*) > \mathfrak{z}) \\ &\le r \int_{\mathfrak{z} \ge 0} \mathfrak{z}^{r-1} \boldsymbol{P}_{\theta^*}(L_h(\tilde{\theta}_h, \theta^*) > \mathfrak{z}) d\mathfrak{z} \le 2r \int_{\mathfrak{z} \ge 0} \mathfrak{z}^{r-1} e^{-\mathfrak{z}} d\mathfrak{z} \end{aligned}$$

and the second assertion follows. ∎

Theorem 3 particularly claims that the estimation loss measured by $\mathcal{K}(\tilde{\theta}_h, \theta)$ is with high probability bounded by $\mathfrak{z}_\alpha/N_h$ provided that $\mathfrak{z}_\alpha$ is sufficiently large. In the regular situation, the Kullback–Leibler divergence $\mathcal{K}$ fulfills

$$\mathcal{K}(\theta, \theta^*) \approx I_{\theta^*} |\theta - \theta^*|^2 \tag{6}$$

for any point θ in a neighborhood of θ^*, where I_{θ^*} is the Fisher information at θ^*; see, e.g., [14] or [15]. Therefore, the result of Theorem 2 guarantees that $|\tilde{\theta}_h - \theta^*| \le C N_h^{-1/2}$ with a high probability.

Table I provides $\mathcal{K}(\theta_h, \theta')$ for special cases of the exponential distribution considered above.

III. 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 $\tilde{\theta}_h = S_h/N_h$ for $h \in \mathcal{H}$ be the corresponding set of estimates. For conciseness we use the notation $\tilde{\theta}_k = \tilde{\theta}_{h_k}, S_k = S_{h_k}$ and $N_k = N_{h_k}$. We also write $L_k(\theta, \theta')$ instead of $L_{h_k}(\theta, \theta')$ for the log-likelihood ratio 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.

A. FLL Scale Selection Procedure

The presented procedure aims at selecting one estimate $\tilde{\theta}_k$ 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 $\varkappa$ can be presented in the form [16]

$$\varkappa = \max\{k : L_l(\tilde{\theta}_l, \tilde{\theta}_m) \le \mathfrak{z}_l \; l < m \le k\}. \tag{7}$$

With this choice, the resulting adaptive estimate at the point x is $\hat{\theta} = \tilde{\theta}_\varkappa$ and the adaptive scale $\hat{h} = h_\varkappa$. By (3), $L_k(\tilde{\theta}_k, \theta) =$

$N_k\mathcal{K}(\tilde{\theta}_k,\theta)$ for every θ. So the procedure can be rewritten as $\varkappa=\max\{k:\ N_l\mathcal{K}(\tilde{\theta}_l,\tilde{\theta}_m)\le \mathfrak{z}_l,\ l<m\le k\}$.

The procedure (7) can be interpreted as follows. The first estimate $\tilde{\theta}_1$ is always accepted and (7) starts from $k=2$. The estimate $\tilde{\theta}_2$ is checked whether it belongs to the confidence set $\mathcal{E}_{h_1}(\mathfrak{z}_1)$ of the previous step estimate $\tilde{\theta}_1$, see (5) in Theorem 3. If not, the estimate $\tilde{\theta}_2$ is rejected and the procedure terminates selecting $\tilde{\theta}_1$. The estimate $\tilde{\theta}_2$ belongs to the confidence set $\mathcal{E}_{h_1}(\mathfrak{z}_1)$ if the inequality $T_{12}=L_1(\tilde{\theta}_1,\tilde{\theta}_2)\le \mathfrak{z}_1$ is fulfilled then $\tilde{\theta}_2$ is accepted and the procedure considers the next step estimate $\tilde{\theta}_3$. At every step k, the current estimate $\tilde{\theta}_k$ is compared with all the previous estimates $\tilde{\theta}_1,\dots,\tilde{\theta}_{k-1}$ by checking according to (5) the inequalities $T_{lk}=L_l(\tilde{\theta}_l,\tilde{\theta}_k)\le \mathfrak{z}_l$. 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 also be viewed as a multiple testing procedure. The expressions $T_{lk}=L_l(\tilde{\theta}_l,\tilde{\theta}_k)$ is understood as test statistics for testing the hypothesis $H_{lk}:\boldsymbol{E}\tilde{\theta}_l=\boldsymbol{E}\tilde{\theta}_k$, and $\mathfrak{z}_l$ is the corresponding critical value. At the step k the procedure tests the composite hypothesis $\boldsymbol{E}\tilde{\theta}_1=\ldots=\boldsymbol{E}\tilde{\theta}_k$. The choice of the $\mathfrak{z}_k$'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. We also define the random moment $\varkappa(k)$ meaning the largest index accepted after first k steps and the corresponding adaptive estimate

$$\varkappa(k)=\min\{\varkappa,k\},\qquad \hat{\theta}_k=\tilde{\theta}_{\varkappa(k)}. \tag{8}$$

In our simulation study, the proposed procedure is compared with the other proposal, namely, with the *intersection of confidence intervals* (ICI) method from [5], where the ICI was shown to be quite competitive with many smoothing procedures; see also [12]. For the sake of completeness, we present here a brief description of the ICI method.

B. ICI Algorithm

We define a sequence of the confidence intervals of the estimates

$$Q_l=[\tilde{\theta}_l-\mathfrak{z}\cdot\sigma_{\tilde{\theta}_l},\tilde{\theta}_l+\mathfrak{z}\cdot\sigma_{\tilde{\theta}_l}] \tag{9}$$

where $\sigma_{\tilde{\theta}_l}$ is standard deviation of the estimates $\tilde{\theta}_l$ and $\mathfrak{z}$ is the threshold parameter of the confidence interval.

For this sequence with some probability p we may conclude that if $\theta\in Q_l$ holds for $h=h_l, 1\le l\le k$, all of the intervals $Q_l, 1\le l\le k$, have a point in common, namely θ.

Consider the intersection of the intervals $Q_l, 1\le l\le k$, with increasing k, and let $\varkappa$ be the largest of those k for which the intervals $Q_l, 1\le l\le k$, have a point in common. This $\varkappa$ defines the adaptive estimate and the adaptive scale as follows:

$$\hat{\theta}=\tilde{\theta}_{\varkappa},\qquad \hat{h}=h_{\varkappa}. \tag{10}$$

The ICI rule can be presented in the sequential form (7) provided that the inequality $L_l(\tilde{\theta}_l,\tilde{\theta}_k)\le \mathfrak{z}_l$ is replaced by $|\tilde{\theta}_l-\tilde{\theta}_k|\le(\sigma_{\tilde{\theta}_l}+\sigma_{\tilde{\theta}_k})\mathfrak{z}$, where $\sigma_{\tilde{\theta}_l}$ and $\sigma_{\tilde{\theta}_k}$ are standard deviations of the estimates $\tilde{\theta}_l$ and $\tilde{\theta}_k$ and $\mathfrak{z}$ is a parameter similar to $\mathfrak{z}_l$ in (7).

The ICI algorithm differs from the novel FLL by the used statistics, $T_{lk}=|\tilde{\theta}_l-\tilde{\theta}_k|/(\sigma_{\tilde{\theta}_l}+\sigma_{\tilde{\theta}_k})$ and the invariant threshold parameter $\mathfrak{z}$. In order to compare the estimates in the ICI algorithm one has to know or to estimate their variances which in general, in particular for Poisson models, depend on unknown estimated signal θ and as a result the algorithm requires recursive calculations (see [11], [12], and [17]). The proposed FLL procedure (7) does not need variance estimates and recursive calculations.

C. Choice of the Parameters $\mathfrak{z}_k$ for the FLL Method

The critical values $\mathfrak{z}_1,\dots,\mathfrak{z}_{K-1}$ 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 hypothesis means $\theta(X_i)\equiv\theta^*$ for some fixed θ^* and all i. In this case it is natural to expect that the estimate $\hat{\theta}_k$ coming out of the first k steps of the procedure is close to the nonadaptive counterpart $\tilde{\theta}_k$. This particularly means that the probability of rejecting one of the estimates $\tilde{\theta}_2,\dots,\tilde{\theta}_k$ under the null hypothesis should be very small.

Now we give a precise description of the used technique. The risk of estimation for an estimate $\hat{\theta}$ of θ^* can be measured by $\boldsymbol{E}|\mathcal{K}(\hat{\theta},\theta^*)|^r$ for some $r>0$, see Theorem 3. Under the null hypothesis $\theta(X_i)\equiv\theta^*$, every k and every $r>0$ it holds by this theorem that

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

We require that the parameters $\mathfrak{z}_1,\dots,\mathfrak{z}_{K-1}$ of the procedure are selected in such a way that

$$\boldsymbol{E}_{\theta^*}|L_k(\tilde{\theta}_k,\hat{\theta}_k)|^r=\boldsymbol{E}_{\theta^*}|N_k\mathcal{K}(\tilde{\theta}_k,\hat{\theta}_k)|^r\le\alpha\mathfrak{r}_r,\quad k=2,\dots,K. \tag{11}$$

Here, α is the preselected constant having the meaning of the confidence level of procedure. (11) gives us $K-1$ conditions to fix $K-1$ critical values.

The condition (11) 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 $\hat{\theta}_k$ coincides with the nonadaptive counterpart $\tilde{\theta}_k$ in the typical situation. These two estimates can be different only in the case of a "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 $|L_k(\tilde{\theta}_k,\hat{\theta}_k)|^r=|N_k\mathcal{K}(\tilde{\theta}_k,\hat{\theta}_k)|^r$, and it approaches the probability of "false alarm" $\tilde{\theta}_k\ne\hat{\theta}_k$ as r goes to 0. The condition (11) states an upper bound for the risk associated with "false alarms."

Our definition in (11) still involves two parameters α and r. 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. This situation is similar to the hypothesis testing problem where there is no any universal choice

of the testing level. Note that in view of (6), $r = 1/2$ corresponds to the absolute deviation losses while $r = 1$ leads to quadratic type losses. 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.

(11) only gives $K-1$ conditions for choosing $K-1$ critical values but does not explain how these values can be computed. Below we suggest two methods of evaluating the parameters $\mathfrak{z}_k$ which both are based on Monte Carlo simulations from the homogeneous model $\theta(\cdot) \equiv \theta^*$.

1) Sequential Choice of $\mathfrak{z}_k$: First, we only consider the first critical value $\mathfrak{z}_1$ and set the others equal to infinity: $\mathfrak{z}_2 = \ldots = \mathfrak{z}_K = \infty$. This effectively means that every new estimate $\tilde{\theta}_k$ is only compared with $\tilde{\theta}_1$ by checking that $\tilde{\theta}_k \in \mathcal{E}_{h_1}(\mathfrak{z}_1)$. We denote by $\hat{\theta}_k(\mathfrak{z}_1)$ the adaptive estimate which comes out of the proposed procedure with such set of critical values after the first k steps. Now the value $\mathfrak{z}_1$ is defined as the minimal one which provides the condition

$$E_{\theta^*}|L_k(\tilde{\theta}_k, \hat{\theta}_k(\mathfrak{z}_1))|^r = E_{\theta^*}|N_k\mathcal{K}(\tilde{\theta}_k, \hat{\theta}_k(\mathfrak{z}_1))|^r \le \alpha r_r, \quad k = 2, \ldots, K. \tag{12}$$

Now suppose that $\mathfrak{z}_1, \ldots, \mathfrak{z}_{j-1}$ have been already fixed for some $j > 1$ and we want to select $\mathfrak{z}_j$. We proceed in a similar way by fixing already selected value $\mathfrak{z}_1, \ldots, \mathfrak{z}_{j-1}$ and some $\mathfrak{z}_j$ and setting $\mathfrak{z}_{j+1} = \ldots = \mathfrak{z}_{K-1} = \infty$. The adaptive estimate produced by the procedure with such parameter set after k steps is denoted by $\hat{\theta}_k(\mathfrak{z}_1, \ldots, \mathfrak{z}_j), k \ge j$. Now the value $\mathfrak{z}_j$ is defined as the minimal one which provides the condition

$$\begin{aligned} &\boldsymbol{E}_{\theta^*}|L_k(\tilde{\theta}_k, \hat{\theta}_k(\mathfrak{z}_1, \ldots, \mathfrak{z}_j))|^r = \\ &\boldsymbol{E}_{\theta^*}|N_k\mathcal{K}(\tilde{\theta}_k, \hat{\theta}_k(\mathfrak{z}_1, \ldots, \mathfrak{z}_j))|^r \le \alpha \mathfrak{r}_r \qquad k = j+1, \ldots, K. \end{aligned} \tag{13}$$

Continue this calculations for all $\mathfrak{z}_j, j = 1, \ldots, K-1$. It is proved in [16] that such defined $\mathfrak{z}_j$ fulfill (11). It is also obvious that the choice of the critical values $\mathfrak{z}_j$ is determined by the joint distribution of the estimates $\tilde{\theta}_k$ under the null hypothesis H_0 : $\theta(X_1) = \ldots = \theta(X_K) = \theta^*$. The expectations in (12) and (13) are calculated for the random events subject to the distribution with this fixed value θ^* for the estimated values.

2) Simplified Choice of $\mathfrak{z}_k$: Here, we present a simplified procedure which is rather simple for implementation. It is based on following Theorem 4 (Section VI) where it is shown that provided some assumptions there are three constants a_0, a_1, and a_2 depending on r and α such that the choice $\mathfrak{z}_k = a_0 + a_1 \log \alpha^{-1} + a_2 r \log(N_K/N_k)$ ensures (11) for all $k \le K-1$. It suggests to select $\mathfrak{z}_k$ linearly decreasing with k in the form

$$\mathfrak{z}_k = \mathfrak{z}_1 + s(K-k). \tag{14}$$

Then, we only need to fix two parameters, e.g., the first value $\mathfrak{z}_1$ and the slop s. We first identify the first value $\mathfrak{z}_1$ using the condition (12). The other values $\mathfrak{z}_k$ are found in the form $\mathfrak{z}_k = \mathfrak{z}_1 - s(k-1)$ to provide (11).

3) 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 VI. 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.

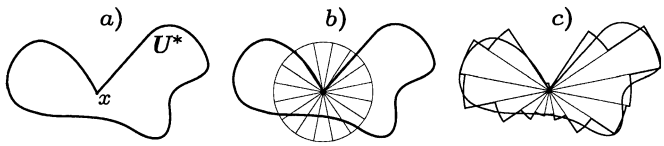


Fig. 1. Nneighborhood of the estimation point x: (a) the best estimation set U^*, (b) the unit ball segmentation, and (c) sectorial approximation of U^*.

The FLL algorithm is implemented in the following two steps:

1) the estimates $\tilde{\theta}_h = \tilde{\theta}_h(x)$ are calculated for all $h \in \mathcal{H}$ by (2) and all x;
2) the adaptive scales $\varkappa$ from (7) and the corresponding adaptive estimates $\hat{\theta} = \tilde{\theta}_\varkappa$ are calculated for all x.

The varying thresholds $\mathfrak{z}_k$ are calculated by the external procedure defined by (13) or (14).

IV. Application to High-Resolution Image Denoising

In many cases the image intensity is an 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 [18] and recent new ridgelet and curvelet transforms [19].

In this paper we exploit star-shaped size/shape adaptive neighborhoods built for each estimation point. Fig. 1 illustrates this concept. A hypothetical ideal neighborhood U^* [Fig. 1(a)] is a largest star-shaped neighborhood of the estimation point x where the constant fits well to the data. A sectorial segmentation of the unit ball with the center at the point x shown in Fig. 1(b) is used for approximation of U^* . This approximation [Fig. 1(c)] is achieved by using varying lengths h_γ of the sectors, where γ is a direction of the sector, and a finite set Γ of different directions γ. Varying size sectors of the length h_γ enable one to get a good approximation for any neighborhood of the point x provided that it is a star-shaped body.

This star-shaped approximation defines both the size and the shape of the estimation neighborhood but requires $|\Gamma|$ parameters h_γ to be defined. Introduce $\eta = (h_\gamma)_{\gamma \in \Gamma}$ as the $|\Gamma|$-dimensional "*window-size*" meaning the set of window (sector) sizes h_γ. Also define $I_\gamma(x)$ as the support of the infinite sector for the direction γ. Note that any two different sectors overlap only in

the central point x. For each η we define the estimate $\tilde{\theta}_\eta$ of the form

$$\tilde{\theta}_\eta = \sum_\gamma \sum_{i\in I_\gamma} w_{i,h_\gamma} Y_i \Big/ \sum_\gamma \sum_{i\in I_\gamma} w_{i,h_\gamma} = S_\eta / N_\eta. \quad (15)$$

As the sectors I_γ overlap at $X_i = x$, this point gets more weights than the others.

The problem of adaptive estimation can be formulated as the choice of the vector η for a given point x. By Theorem 3, the accuracy of the estimate $\tilde{\theta}_\eta$ is measures by the quantity N_η. A natural generalization of the proposal (7) to the vector scale is to select the "largest" (in values of N_η) estimate $\tilde{\theta}_\eta$ which is consistent with all estimates $\tilde{\theta}_{\eta'}$ with "smaller" scales. This approach to adaptation is possible but it is a difficult task encountering some algorithmic and principal problems.

To be practical we use a procedure with independent data-driven selection of the h_γ's for each directions γ using the FLL technique of Section III.

For each direction γ, according to (2), the corresponding directional estimate $\tilde{\theta}_{h,\gamma} = \tilde{\theta}_{h,\gamma}(x)$ is calculated as

$$\tilde{\theta}_{h,\gamma} = \sum_{i\in I_\gamma} w_{i,h} Y_i \Big/ \sum_{i\in I_\gamma} w_{i,h}. \quad (16)$$

With a given set $\mathcal{H}$ of the scales $h_1, \ldots, h_K$ we calculate the corresponding set of the estimates $\{\tilde{\theta}_{h,\gamma}(x), h \in \mathcal{H}\}$ and come back to the problem of selecting for every direction γ one of these estimates in a data driven way. The FLL adaptive procedure from Section III leads to the adaptive scale $\hat{h}_\gamma(x)$ which describes the set of homogeneity in direction γ with the center at x. In total, we have $|\Gamma|$ such sets for different directions γ.

Define $\hat{\eta} = (\hat{h}_\gamma)_{\gamma\in\Gamma}$ and the final adaptive estimate $\hat{\theta} = \tilde{\theta}_{\hat{\eta}}$ due to (15):

$$\hat{\theta} = \sum_\gamma \sum_{i\in I_\gamma} w_{i,\hat{h}_\gamma} Y_i \Big/ \sum_\gamma \sum_{i\in I_\gamma} w_{i,\hat{h}_\gamma}. \quad (17)$$

Selecting a scale $\hat{h}_\gamma$ for every direction γ can be viewed by itself as a multiple testing procedure. Now we perform independently $|\Gamma|$ such procedures which requires an additional correction of the parameters (critical values) $\mathfrak{z}_k$ to account for the direction choice. In the spirit of the proposal (11), we select the critical values $\mathfrak{z}_k$ to provide that

$$\boldsymbol{E}_{\theta^*} |N_{\eta_k} \mathcal{K}(\tilde{\theta}_{\eta_k}, \tilde{\theta}_{\hat{\eta}_k})|^r \le \alpha \mathfrak{r}_r, k = 2, \ldots, K. \quad (18)$$

Here, for every k the restricted set of bandwidths $\{h_1, \ldots, h_k\}$ is considered and η_k is the nonadaptive "oracle" vector scale with the components $h_\gamma \equiv h_k$ for all γ, while $\hat{\eta}_k = (h_{\varkappa_\gamma(k)})_{\gamma\in\Gamma}$ is obtained by the adaptive choice for every direction γ of the index $\varkappa_\gamma(k)$ after the first k steps of the scale selection algorithm using the critical values $\mathfrak{z}_1, \ldots, \mathfrak{z}_{k-1}$ [see (8)].

The sequential or simplified choice of $\mathfrak{z}_k$ can be used exactly as in the case of the scalar scale case. The step of computing the values $\mathfrak{z}_k$ has to be done only once. With the computed parameters $\mathfrak{z}_k$, the total complexity of this procedure is linearly proportional to the number $|\Gamma|$ of different sectors.

The FLL algorithm is implemented in the following three-step procedure:

1) the directional estimates $\tilde{\theta}_{h,\gamma}$ (16) are calculated for all $h \in \mathcal{H}$ and all $\gamma \in \Gamma$;
2) the adaptive scales $\hat{h}_\gamma$ are calculated using (7) for all directions $\gamma \in \Gamma$;
3) the final estimate $\hat{\theta}$ is calculated according to the formula (17).

These steps are performed for all x. The varying thresholds $\mathfrak{z}_k$ are calculated by the external procedure to fulfill (18).

V. Experimental Study

A. Preliminary

The described adaptive star-shaped neighborhood estimates are originated in the works [11], [20], where it is successfully exploited with the ICI adaptive scale selection for different image processing problems.

Multiple studies show that the narrow line-wise window functions w are preferable for high-resolution image denoising. It is demonstrated in [12, ch. 8] that the ICI algorithm is very sensitive with respect to singularities or rapid image intensity variations and combined with the narrow windows gives the estimates which are able to preserve tiny images details while the algorithms with wider sectorial supports usually result in over-smoothed estimates.

The estimates (16) and (17) depends on the product $w_{i,h} I_\gamma$ of the window function $w_{i,h}$ and the sector indicator I_γ. Denote these products by $w_{i,h,\gamma} = w_{i,h} I_\gamma$. Then

$$\tilde{\theta}_{h,\gamma} = \sum_i w_{i,h,\gamma} Y_i \Big/ \sum_i w_{i,h,\gamma} \quad (19)$$

$$\hat{\theta} = \sum_\gamma \sum_i w_{i,\hat{h}_\gamma,\gamma} Y_i \Big/ \sum_\gamma \sum_i w_{i,\hat{h}_\gamma,\gamma}. \quad (20)$$

In what follows $w_{i,h,\gamma}$ is a binary function with values 0, 1. The scale (window size) parameter h takes integer values in the set $\mathcal{H} = \{\lfloor 1.5^k \rfloor, k = 1, \ldots, 7\} = \{1, 2, 3, 5, 7, 11, 17\}$. The length of the window $w_{i,h,\gamma}$ is equal to h. For the horizontal direction, $\gamma = 0$, and $1 \le h \le 5$ the window weights $w_{i,h,0} = 1$ along the horizontal axis only. For $h > 5$ the area where $w_{i,h,0} = 1$ becomes wider as it is illustrated in Fig. 2. For all h the windows $w_{i,h,\gamma}$ are quite narrow. The rotation of these horizontal windows defines the directional windows for eight directions $\gamma_i = (i-1)\pi/4, i = 1, \ldots, 8$.

The ideal neighborhood partition shown in Fig. 1 is possible only for continuous variable functions and in general not possible for functions defined on the regular grid, in particular, if we wish to use these sectors for multidirectional estimation. The windows shown in Fig. 2 are selected in such way that there are some gaps between the sectors. We found from our experiments these windows give better results than, say, the narrow line-wise kernels (of the width equal to one; see [12, ch. 8]) or the windows with wider overlapping sectors.

It is follows also from our experimental study that for the considered set of images using the window weights w different from the binary ones does not improve the numerical results.

$h=1$	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$h=2$	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$h=3$	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
$h=5$	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
$h=7$	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
$h=11$	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
	1	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
$h=17$	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1

Fig. 2. Binary (0,1) windows of the horizontal directional estimates. The length of the window is equal to h. The width becomes larger than 1 starting from $h = 7$.

The directional adaptive estimates $\hat{\theta}_\gamma = \tilde{\theta}_{\hat{h}_\gamma,\gamma}$ are calculated independently for all $\gamma \in \Gamma$. As a result we obtain a number of different estimates of the same signal. Combining these multiple estimates in the final unique estimate is known as a fusing problem. In terms of this problem the formula (20) for the final adaptive estimate has an interesting interpretation at least for the binary zero-one weights $w_{i,h,\gamma}$. Indeed, by Theorem 3 the variability of the estimate $\tilde{\theta}_{h,\gamma}$ is inverse proportional to the sum of weights $N_{h,\gamma} = \sum_i w_{i,h,\gamma}$. Moreover, in the case of binary weights $w_{i,h,\gamma}$ and the homogeneous noise, the variance of $\tilde{\theta}_{h,\gamma}$ is also inverse proportional to $N_{h,\gamma}$. Therefore, a natural way to combine (fuse) the multiple directional estimates $\tilde{\theta}_{\hat{h}_\gamma,\gamma}$ into the unique final one is the weighted mean using inverse variance multipliers $N_{\hat{h}_\gamma,\gamma}$ as weights [12, ch. 6]:

$$\hat{\theta} = \sum_\gamma \tilde{\theta}_{\hat{h}_\gamma,\gamma} N_{\hat{h}_\gamma,\gamma} \Big/ \sum_\gamma N_{\hat{h}_\gamma,\gamma}. \tag{21}$$

It is a simple exercise to verify that the estimates (20) and (21) are identical. Thus, instead of calculation of (20), we can work in (21) with the directional estimates only. The FLL is used for calculation of the directional adaptive scales and the adaptive directional estimates. Then, the final estimate (20) can be calculated as the weighted mean (21) of these adaptive directional estimates. This fact highlights the meaning of the estimate (20) as well as of the more general proposal (15).

An alternative way of fusing of the adaptive directional estimates $\hat{\theta}_\gamma$ is to define $\hat{\theta}$ as the mean over the region $\hat{U}$ obtained by union of the adaptively selected directional sectors $\hat{U}_\gamma$ of length $\hat{h}_\gamma$:

$$\hat{\theta} = \sum_{X_i \in \hat{U}} Y_i \Big/ \hat{N} \tag{22}$$

where $\hat{N}$ means the number of points in $\hat{U}$. The only difference with the estimate (21) is that the central point x is taken with the weight one as all the other points while (21) put much more weight to the point x. The experimental results indicate that for imaging applications, in particular, for texture images, the estimate (21) is visually and numerically preferable. In what follows we show the results obtained by (21).

We mention one more fact used in the algorithm implementation. A pointwise nature of the procedure leads to certain variability of the selected parameter (window size) as a function of the location x, especially for a large noise level; see, e.g., [21]. In order to reduce the stochastic variability of the estimates the FLL algorithm is completed with a special smoothing of the adaptively selected $\hat{h}_\gamma(x)$ as functions of x. For this prefiltering, we apply weighted median filters. The obtained filtered scales $\hat{\eta}(x) = \{\hat{h}_\gamma(x)\}_{\gamma\in\Gamma}$ are used for building the adaptive estimates.

We demonstrate the performance of the developed algorithm for Poissonian, Gaussian and binary Bernoulli image observations. The image $\theta(x)$ and the observations are defined on the finite discrete rectangular grid 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 $\theta(x)$ from the observations $Y(x), x \in X$. For the imaging quality evaluation we use the peak-signal-to-noise-ratio (PSNR) calculated in decibels as PSNR $= 20\log_{10}(\max_x |\theta(x)|/\text{RMSE})$, where the signal peak is $\max_x |\theta(x)|$ and the root mean-squared error (RMSE), RMSE $= \{(n_1 n_2)^{-1} \sum_x [\theta(x) - \hat{\theta}(x)]^2\}^{1/2}$, is used for calculation of the noise level in the image reconstruction.

In 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 the binary test-images: *Cheese* (128×128). For all images, we use the eight sectorial estimators with the window functions shown in Fig. 2.

A special study has been produced for testing the procedures presented for $\mathfrak{z}_k$ selection. The expectations in the corresponding formulas are calculated by the Monte Carlo method. In these calculations the work of the adaptive FLL algorithm is imitated including the prefiltering of the adaptive scales mentioned above. Selection of $\mathfrak{z}_k$ depends on the parameters r and α. As already noticed larger α and smaller r result in smaller critical values $\mathfrak{z}_k$. Smaller $\mathfrak{z}_k$ means decreasing of smoothing properties of the adaptive FLL algorithm.

Our default choice is $r = 1$ and $\alpha = 1$. This recommendation works surprisingly well giving the sets of $\mathfrak{z}_k$ universally good for different images and different distributions. In what follows, we use the sets $\mathfrak{z}_k$ obtained by the simplified threshold parameter choice (Section III-C).

The MATLAB codes of the algorithms used in the following experiments can be found on the website www.cs.tut.fi/~lasip.

B. Poissonian Observations

To achieve different level of randomness [different signal-to-noise ratio (SNR)] in the Poissonian observations we multiply the true signal θ by a scaling factor with the observations defined according to the formula $\tilde{z} \sim \mathcal{P}(\theta \cdot \chi)$, where $\chi > 0$ is a scaling factor. Further, we assume the observations in the form

TABLE II
"LENA" IMAGE: CRITERIA VALUES FOR THE EIGHT DIRECTIONAL AND TWO FINAL ESTIMATES

	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8	Final (21)	Final (22)
PSNR	22.54	22.73	23.67	22.3	22.45	22.62	23.66	22.36	**28.34**	28.10

TABLE III
"CHEESE" IMAGE: CRITERIA VALUES FOR THE EIGHT DIRECTIONAL AND FINAL ESTIMATES

	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8	Final (21)	Final (22)
PSNR	26.25	25.09	26.08	25.26	26.04	24.81	26.16	25.18	**34.3**	34.2

$z = \tilde{z}/\chi$ in order to have the results comparable for different χ as $E\{z\} = E\{\tilde{z}\}/\chi = \theta$ 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 : $\mathrm{var}\{z\} = \mathrm{var}\{\tilde{z}\}/\chi^2 = \theta/\chi$. The SNR is calculated as $E\{z\}/\sqrt{\mathrm{var}\{z\}} = \sqrt{\theta\chi}$. Thus, for larger and smaller χ, we have respectively larger and smaller SNR.

This scaled modeling of Poisson data is exploited in a number of publications [22]–[25] where the advanced performance of the wavelet based denoising algorithms is demonstrated. It is shown in [17] that the ICI-based adaptive algorithm gives quite competitive results and at least numerically demonstrates a better performance then the algorithms in the cited papers. We consider this ICI adaptive algorithm as a main competitor to the proposed FLL technique.

In the scale selection the FLL technique is applied to the Poissonian variables, i.e., to $\tilde{z}$ with the adaptive directional estimates denoted as $\tilde{z}_\gamma$. Then the directional FLL estimates of θ are calculated as $\hat{\theta}_\gamma = \tilde{z}_\gamma/\chi$. The threshold set, calculated according to the simplified choice (14) with $r = 1$ and $\alpha = 1$, is as follows: $\mathfrak{z} = \{1.6\ 1.40\ 1.14\ 0.91\ 0.68\ 0.45\}$.

For the *Lena* image, Table II illustrates numerically the effects of fusing of the directional estimates in the final one. The criterion values for the final estimates compared with the eight directional sectorial ones show a strong improvement in the final estimate. In particular, we have for PSNR the values about 22.5 dB for the sectorial estimates while for the fused estimate (21) PSNR $\simeq 28.34$.

Even a more impressive difference between the directional and final estimates can be seen in Table III given for the *Cheese* image. In Tables II and III, we show the final results obtained by both fusing formulas (21) and (22). The fusing by (21) shows better results. This fact is observed in nearly all our experiments. In what follows, we show the results obtained by using the formula (21).

Some results for the *Lena* image are demonstrated in Fig. 3. The central panel shows the true image and the eight surrounding panels show the FLL adaptive scales $\hat{h}_{\gamma_j}(x)$ for the corresponding eight directions $\gamma_j = (j-1)\pi/4, j = 1, \ldots, 8$. We can see the adaptive scales for directional estimates looking at the horizontal and vertical directions, i.e., to *East*, *North*, *West*, and *South*, as well as to four diagonal directions *North-East*, *North-West*, *South-West*, and *South-East*.

White and black correspond to large and small scale values, respectively. The adaptive scales delineate the image intensity very well as it could be done provided that the intensity function is known in advance. This delineation is obviously directional as the contours of the image are shadowed from the corresponding directions. The eight narrowed windows allow to build the estimates highly sensitive with respect to image details and essentially improve the quality of denoising.

Fig. 3. FLL adaptive directional window sizes $\hat{h}_{\gamma_j}(x), \gamma_j = (j-1)\pi/4, j = 1, \ldots, 8$, for the *Lena* image. The true image is shown in the central panel.

Fig. 4 demonstrates the obtained estimates. The central panel shows the final estimate calculated from the sectorial ones according to the formula (21). The surrounding panels show the sectorial directional adaptive scale estimates $\hat{\theta}_{\gamma_j}(x), j = 1, \ldots, 8$, corresponding to the adaptive scales given in Fig. 3 for the relevant directions.

The noise effects are clearly seen in the adaptive scales $\hat{h}_{\gamma_j}(x)$ as spread black isolated points. The black means that FLL erroneously takes smaller values of the scale. A directional nature of the adaptive estimates $\hat{\theta}_{\gamma_j}(x)$ is obvious with the corresponding directions seen as a line-wise background of this imaging. The fusing of multidirectional estimates allows to delete and smooth these directional line effects and obtain a good quality final estimate. Overall, the multidirectional estimation allows to reveal and preserve a tiny detail of the image and in the same time efficiently suppress the noise.

TABLE IV
PSNR VALUES OBTAINED BY THE FOLLOWING ALGORITHMS: PROPOSED FLL, LPA-ICI RECURSIVE NON-GAUSSIAN (RNG) AND NON-RECURSIVE LPA-ICI USING THE ANSCOMBE TRANSFORM (AT). POISSONIAN DISTRIBUTION

Images	$\chi = 102$			$\chi = 25.5$			$\chi = 12.75$			$\chi = 6.375$		
Algorithm	*FLL*	*RNG*	*AT*	*FLL*	*RNG*	*AT*	*FLL*	*RNG*	*AT*	*FLL*	*RNG*	*AT*
Cheese	38.00	35.66	**39.85**	**34.30**	29.71	33.0	**30.70**	25.58	28.64	**27.58**	19.34	24.46
C-man	**30.77**	29.45	30.2	**26.84**	26.17	26.3	**25.00**	24.42	24.0	**23.13**	21.46	21.0
Boats	**29.87**	29.44	29.6	**26.67**	26.62	26.1	**25.10**	24.98	24.0	**23.59**	21.46	21.7
Peppers	**31.19**	31.16	30.64	28.19	**28.50**	26.95	**26.30**	26.21	24.59	**24.43**	20.80	21.61
Lena	**31.76**	31.64	31.2	28.45	**28.61**	27.29	**26.55**	26.41	25.00	**24.84**	21.11	22.4

Fig. 4. Central panel shows the aggregated final estimate for the *Lena* image and the surrounding ones show the directional FLL adaptive estimates $\hat{\theta}_{\gamma_j}, \gamma_j = (j-1)\pi/4, j = 1, \ldots, 8$.

The FLL numerical results in Table IV are given for the final estimate (21) obtained from eight-directional estimates. It shows PSNR values calculated for the test-images provided different values of the parameter χ defining varying SNR for Poissonian observations. The largest $\chi = 255/2.5 = 102$ corresponds to the smallest level of the noise while the smallest $\chi = 255/40 = 6.375$ corresponds to the highest noise level in our experiments. Recall that SNR is proportional to $\sqrt{\chi}$.

In each cell (image-χ) of this table, we show results given by three different algorithms, respectively: the proposed FLL algorithm, the LPA-ICI recursive non-Gaussian (RNG) algorithm, and the non-recursive LPA-ICI algorithm using the Anscombe transform (AT).

The LPA-ICI is the algorithm using the local polynomial approximation (LPA) for estimation and the ICI for scale adaptation. The zero- and first-order LPA are used in this algorithm with narrow sectorial windows similar to discussed above. The basic LPA-ICI algorithm is developed for the Gaussian observations. The LPA-ICI recursive algorithm is especially developed for non-Gaussian data with the signal dependent variance.

The nonlinear Anscombe transform of observations Y has a form $Z = 2\sqrt{Y + 3/8}$. For Poissonian Y this random Z has the variance approximately equal to 1. This stabilization of the variance is exploited for denoising of Poissonian observations. In the non-recursive LPA-ICI algorithm, we calculate the Anscombe transform of the initial data, filter them using the basic non-recursive LPA-ICI algorithm and inverse the Anscombe transform.

Details of the basic LPA-ICI and LPA-ICI recursive non-Gaussian (RNG) algorithms can be seen in the [12, ch. 12].

For the considered test images, the LPA-ICI recursive non-Gaussian algorithm mainly gives the best result with about seven iterations. The values shown in Table IV are obtained for this number of iterations.

For nearly all cases, PSNR in Table IV shows the best values for the FLL algorithm. For a small level of the noise ($\chi = 102$), the AT algorithm demonstrates a better performance only for the binary *Cheese* image.

For the middle level of the noise with $\chi = 25.5$, the LPA-ICI recursive algorithm shows slightly better results than the FLL algorithm for two texture images *Peppers* and *Lena*. However, this PSNR advantage is in contradiction with visual evaluation. Fig. 5 shows that the images obtained by LPA-ICI recursive non-Gaussian (RNG) algorithm suffer from multiple spot-like artifacts while the FLL images are free from this sort of degradation.

For the higher noise level with $\chi = 255/20 = 12.75$ and $\chi = 6.375$, the FLL algorithm demonstrates the best PSNR values sometimes with a quite valuable improvement.

Note, that the binary "Cheese" image is modeled with $\theta = [0.2, 1.0]$. We do not use the standard binary (0, 1) values because $\theta = 0$ meaning not only the zero value of the signal but also the zero value of its variance. Thus, the observations corresponding to $\theta = 0$ would be noiseless accurate.

Table V shows the processing time (in seconds, 1.5-GHz Intel Centrino Processor) for images of different size using the compared algorithms. The advantage of the FLL algorithm is clear.

C. Gaussian Observations

We assume that the additive zero-mean Gaussian noise has the standard deviation σ. For the scales $\mathcal{H}$, the threshold set calculated according to the simplified choice with $r = 1$ and $\alpha = 1$ is as follows: $\mathfrak{z} = \{3.0, 2.64, 2.28, 1.92, 1.56, 1.2\}$.

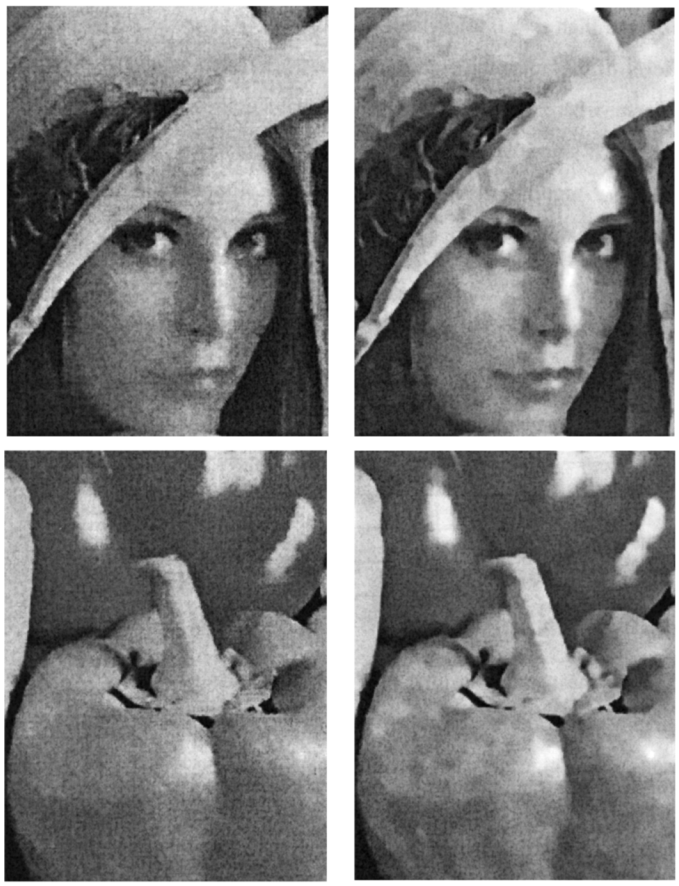


Fig. 5. Fragments of denoised Poissonian images: first line *Lena*, left image FLL with PSNR $= 28.34$, right image LPA-ICI, RNG with PSNR $= 28.61$; second line *Peppers*, left image FLL with PSNR $= 28.06$, right image LPA-ICI, RNG with PSNR $= 28.5$.

TABLE V
PROCESSING TIME IN SECONDS FOR DIFFERENT SIZE IMAGES, POISSONIAN DISTRIBUTION

Image size\ Algorithm	*FLL*	*LPA-ICI, RNG*	*Anscombe*
128 × 128	7	9	33
256 × 256	16	30	40
512 × 512	56	121	76

For comparison, we use the results obtained by the basic non-recursive LPA-ICI denoising algorithm. We select for comparison this algorithm as belonging to the same class of the algorithms mainly different by the statistics used for the adaptive scale selection. Details concerning this algorithm can be seen in [12]. Note that the referred basic LPA-ICI algorithm is a specially designed and optimized for the Gaussian case while the FLL is demonstrated in the form universally tuned for the class of exponential distributions.

In each cell (image-σ) of Table VI, we show results given by two compared algorithms, respectively: the proposed FLL and the basic LPA-ICI algorithms. The PSNR values are shown for small $\sigma = 0.05$, middle $\sigma = 0.1$ and high $\sigma = 0.2$ levels of the noise.

For a small level of the noise, the algorithms demonstrate nearly equivalent performance for all images but *Cheese*, where

TABLE VI
PSNR VALUES OBTAINED BY THE PROPOSED FLL AND BASIC LPA-ICI ALGORITHMS. GAUSSIAN DISTRIBUTION

Images	$\sigma = 0.05$		$\sigma = 0.1$		$\sigma = 0.2$	
Algorithm	*FLL*	*LPA-ICI*	*FLL*	*LPA-ICI*	*FLL*	*LPA-ICI*
Cheese	37.74	**40.97**	**35.80**	35.5	**31.73**	29.07
Cameraman	31.46	**31.54**	**28.08**	27.87	**24.98**	23.95
Boats	**31.34**	31.25	**28.03**	27.88	**24.89**	24.56
Peppers	**32.18**	32.06	**29.30**	28.92	**26.02**	25.23
Lena	**32.92**	32.70	**29.59**	29.21	**26.36**	25.72

TABLE VII
PROCESSING TIME IN SECONDS FOR DIFFERENT SIZE IMAGES, GAUSSIAN DISTRIBUTION

Image size\ Algorithm	*FLL*	*LPA-ICI*
128 × 128	6.2	5.7
256 × 256	12.3	9.2
512 × 512	35.5	26.7

the LPA-ICI algorithm shows the significantly larger PSNR value. For the higher noise the FLL algorithm shows better results than the LPA-ICI algorithm for all cases. This advantage of the FLL algorithm is mainly caused by the varying thresholds $\mathfrak{z}_k$ while these thresholds are the same for all scales in the LPA-ICI algorithm.

Table VII shows the processing time (in seconds, 1.5-GHz Intel Centrino Processor) for images of different size using FLL and ICI algorithms. Compared with Table V, we may note that for the Gaussian case, the FLL algorithm becomes faster. It happens because, for the Poissonian distribution, the Kullback statistics require calculation of logarithm function versus the squared differences for the Gaussian case. Compared with the ICI algorithm, we may conclude that the ICI algorithm is a bit faster than the FLL algorithm.

D. Bernoulli Observations

Bernoulli imaging assumes that the observations 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 of θ is unbiased with the variance equal to $\theta(1-\theta)/n$ and $\mathrm{SNR} = \sqrt{n\theta/(1-\theta)}$, 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 $\tilde{\theta} = 0, 1$ using instead the image $\theta = \tilde{\theta} \cdot \delta + 0.5(1-\delta), 0 < \delta < 1$. For this θ, the Bernoulli random variable takes values 0 and 1 with the probabilities $\theta_0 = 0.5(1-\delta)$ and $\theta_1 = 0.5(1+\delta)$, respectively.

TABLE VIII
PSNR VALUES FOR THE BINARY BERNOULLI IMAGING OBTAINED BY THE PROPOSED FLL ALGORITHM

δ	0.70	0.75	0.80	0.85	0.90	0.95
PSNR, FLL	22.12	23.36	24.49	26.17	28.45	30.80
PSNR, noisy data	7.53	8.5	9.59	10.92	12.71	15.69

Fig. 6. *Cheese* image: binary Bernoulli observations z, estimate errors $|\theta-\hat{\theta}| \times 10$ and estimates $\hat{\theta}$ for $\delta = 0.85$.

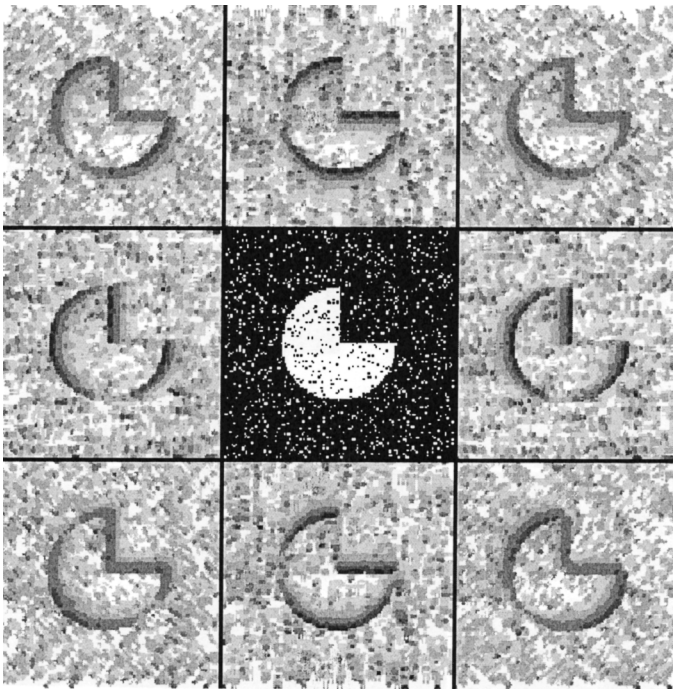


Fig. 7. FLL adaptive directional window sizes $\hat{h}_{\gamma_j}(x), \gamma_j = (j-1)\pi/4, j = 1,\ldots,8$, for the *Cheese* image, $\delta = 0.85$. The noisy image (binary Bernoulli distribution) is shown in the central panel.

The variance of these observations grows rapidly when δ takes smaller values.

The threshold set calculated according to the simplified choice for $r = 1/2$ and $\alpha = 1$ is as follows: $\mathfrak{z} = \{0.7,\ 0.69, 0.67, 0.66, 0.64, 0.63\}$. The modeling results are presented for the binary *Cheese* image ($\tilde{\theta} = 0, 1$) and the varying parameter δ. PSNR values for the FLL filtered and noisy data are shown in Table VIII. The most noisy case corresponds to $\delta = 0.7$ with PSNR $= 7.53$ dB for the noisy data. The lowest level of the noise corresponds to $\delta = 0.95$ with PSNR $= 15.69$ dB for the noisy data. Table VIII confirms a good performance of the algorithm with an improvement in PSNR values about 15 dB. Noisy and denoised images as well as the error of denoising are illustrated in Fig. 6.

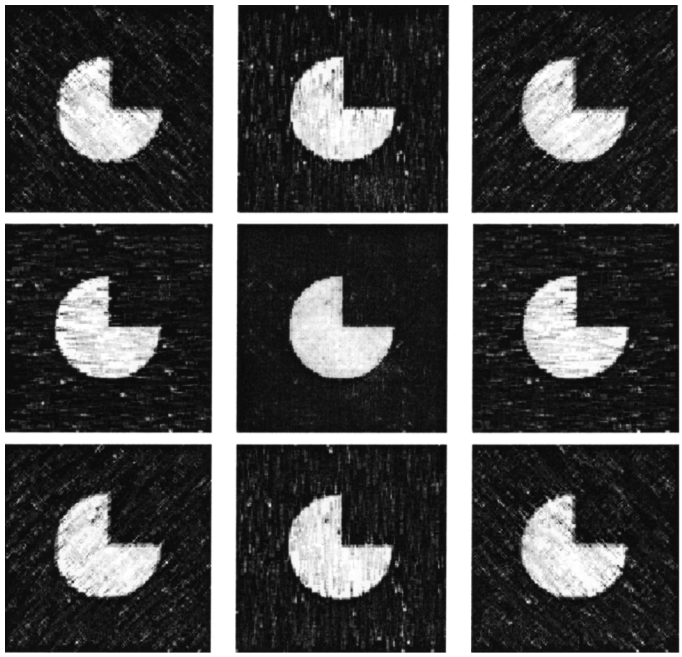


Fig. 8. Central panel shows the aggregated final estimate for the *Cheese* image, and the surrounding ones show the directional FLL adaptive estimates $\hat{\theta}_{\gamma_j}, \gamma_j = (j-1)\pi/4, j = 1,\ldots,8, \delta = 0.85$.

The FLL adaptive scales for different eight directions are shown in Fig. 7. The central panel shows the noisy image and the eight surrounding panels show the FLL adaptive scales $\hat{h}_{\gamma_j}(x)$ for the corresponding eight directions $\gamma_j = (j-1)\pi/4, j = 1,\ldots,8$.

White and black correspond to large and small scale values respectively. The adaptive scales delineate the edges of the binary image quite well. This delineation is obviously directional as the contours of the image are shadowed from the corresponding directions. The eight narrowed kernels allow to build the estimates highly sensitive with respect to image details and essentially improve the quality of denoising.

Fig. 8 demonstrates the obtained estimates. The central panel shows the final fused estimate calculated from the sectorial ones according to the formula (21). The surrounding panels show the sectorial directional adaptive scale estimates $\hat{\theta}_{\gamma_j}(x), j = 1,\ldots,8$, corresponding to the adaptive scales given in Fig. 7 for the relevant directions.

The noise effects are clearly seen in the adaptive scales $\hat{h}_{\gamma_j}(x)$ as spread black isolated points. A directional nature of the adaptive estimates $\hat{\theta}_{\gamma_j}(x)$ is obvious since the corresponding directions are seen as a line-wise background of this imaging. The fusing of multidirectional estimates allows to delete and smooth these directional line effects and obtain a good quality final estimate.

VI. Theoretical Study

This section presents some properties of the adaptive estimates with the scalar scale parameter as they are defined in Section III. In particular, we state the "oracle" estimation quality of the sectorial adaptive scale estimates. The final star-shaped adaptive neighborhood estimate obtained from these adaptive sectorial ones can be considered as a heuristic step of the algorithm design. A full extension of the theory to imaging with star-shaped adaptive neighborhoods is still an open question.

We suppose that the parameters $\mathfrak{z}_k$ of the procedure are selected in such a way that the condition (11) is fulfilled. First, we present some bounds on $\mathfrak{z}_k$ that ensure (11). Next, we study the properties of $\hat{\theta}$ in the parametric and local parametric situation. Finally we extend these results to the general nonparametric situation and prove an "oracle" property of $\hat{\theta}$.

A. 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 $2 \le k \le K$, 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:

(Θ) the set Θ is compact and the Fisher information I_θ is a continuous function of $\theta \in \Theta$.

Our first result claims that under conditions (*MD*) and (Θ), the parameters $\mathfrak{z}_k$ can be chosen in the form $\mathfrak{z}_k = \mathfrak{z}_K + s(K-k)$ to fulfill the "propagation" condition (11). The proof of this and similar statements can be found in [16].

Theorem 4 [16]: Assume (*MD*) and (Θ). 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 (11) for all $k \le K-1$. Particularly, $\boldsymbol{E}_{\theta^*} |N_K \mathcal{K}\big(\tilde{\theta}_K, \hat{\theta}\big)|^r \le \alpha \mathfrak{r}_r$.

This result presents an upper bound for the critical values $\mathfrak{z}_k$. In particular, it claims that these values are at most logarithmic in the sample size. Another important observation is that this upper bound decreases with the index k. The reason can be explained as follows. The choice of the critical values relies only on the behavior of the procedure in the homogeneous situation. The critical values should be large enough to prevent from "false alarms" (rejections of the homogeneity hypothesis). Note, however, that a "false alarm" at an early step of the procedure is more crucial than at the final steps because it leads to the choice of a highly variable estimate $\hat{\theta} = \tilde{\theta}_{\varkappa}$. The criterion (11) automatically accounts for this and the procedure by construction is more conservative at the beginning of the algorithm and less conservative at the end.

B. Risk of Estimation in Nonparametric Situation: "Small Modeling Bias" Condition

Theorem 3 states some results about the accuracy of the local MLE $\tilde{\theta}_k$ in the local homogeneous situation with $\theta(X_i) = \theta$ for all positive weights w_{i,h_k}. In particular, the risk of estimation can be bounded in the form $\boldsymbol{E}_\theta |N_k \mathcal{K}(\tilde{\theta}_k, \theta)|^r \le \mathfrak{r}_r$ for all k.

Here, the bound of Theorem 3 is extended to the nonparametric model $Y_i \sim P_{\theta(X_i)}$ when the function $\theta(\cdot)$ is not any longer constant even in a vicinity of the reference point x, but it can be well approximated by a constant θ for all points X_i from a neighborhood U of x.

Define

$$\Delta_U(\theta) = \sum_{X_i \in U} \mathcal{K}(\theta(X_i), \theta).$$

This quantity $\Delta_U(\theta)$ called the *modeling bias* naturally measures the local distance between the original model given by the regression function $\theta(X_i)$ and the parametric model with $\theta(\cdot) \equiv \theta$ on the set U.

Similarly we define for every scale h_k

$$\Delta_k(\theta) = \sum_{X_i : w_{i,h_k} > 0} \mathcal{K}(\theta(X_i), \theta).$$

We now aim to extend this result to the nonparametric situation under the "small modeling bias" (SMB) condition $\Delta_k(\theta) \le \Delta$ for some $\Delta \ge 0$.

Theorem 5: Let for some $\theta \in \Theta, k^* \le K$, and some $\Delta \ge 0$

$$\max_{k \le k^*} \Delta_k(\theta) \le \Delta. \tag{23}$$

Then, it holds for $r > 0$

$$\begin{aligned} \boldsymbol{E} \log\big(1 + |N_{k^*} \mathcal{K}(\tilde{\theta}_{k^*}, \theta)|^r / \mathfrak{r}_r\big) &\le \Delta + 1 \\ \boldsymbol{E} \log\big(1 + |N_{k^*} \mathcal{K}(\tilde{\theta}_{k^*}, \hat{\theta}_{k^*})|^r / (\alpha \mathfrak{r}_r)\big) &\le \Delta + 1. \end{aligned}$$

Proof: The proof is based on the following general result.

Lemma: Let P and P_0 be two measures such that $\mathcal{K}(P, P_0) \le \Delta < \infty$. Then, for any random variable ζ with $E_0 \zeta < \infty, E \log(1+\zeta) \le \Delta + E_0 \zeta$.

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

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

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

We now apply this lemma with $\zeta = |N_{k^*} \mathcal{K}(\tilde{\theta}_{k^*}, \theta)|^r / \mathfrak{r}_r$ or $\zeta = |N_k \mathcal{K}(\tilde{\theta}_{k^*}, \hat{\theta}_{k^*})|^r / (\alpha \mathfrak{r}_r)$ and utilize that $\boldsymbol{E}_\theta \zeta \le 1$. Clearly in both cases, the estimates $\tilde{\theta}_k$ and $\hat{\theta}_k$ only depend on the observations Y_i with $w_{i,h_k} > 0$. Denote by $\boldsymbol{P}_k$ the joint distribution of such observations for the given function $\theta(\cdot)$ and by $\boldsymbol{P}_{k,\theta}$ the similar distribution in the homogeneous case $\theta(\cdot) \equiv \theta$. Then, with $Z_{k,\theta} = d\boldsymbol{P}_k / d\boldsymbol{P}_{k,\theta}$

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

and the assertion of the lemma follows. ∎

This result particularly means that under the SMB condition (23) with some fixed Δ, the losses $|N_{k^*}\mathcal{K}(\tilde{\theta}_{k^*},\theta)|^r$ are stochastically bounded. Note that this result applies even if Δ is large; however, the bound is only meaningful for small or moderate Δ because it grows exponentially with Δ. It also suggests the following definition of the "oracle" or "ideal" choice k^* of the scale parameter k: it is the largest value for which $\Delta_k(\theta) \le \Delta$ for all $k \le k^*$. Due to Theorems 3 and 5, the "oracle" choice leads to the "oracle" accuracy $1/N_{k^*}$. The next section shows that the adaptive estimate can guarantee essentially the same estimation accuracy.

Note that the given definition of the "oracle" k^* depends upon the value Δ, which measures how far the underlying true model and its the parametric approximation may deviate from each other. This means that the given definition is subjective and there are many "oracle" choices depending on the different Δ -value. However, the procedure does not rely on the "oracle" definition and the theoretical results below apply to any of them.

[13] has shown that the SMB condition is similar to the classical bias-variance tradeoff condition, and Δ can be viewed as a constant that bounds the ratio of the squared bias of the estimate $\tilde{\theta}_k$ and of its variance. This yields that the "oracle" choice of the window is equivalent to the rate optimal scale selection and it leads to the rate optimal estimation quality in the class of smooth functions.

C. "Stability After Propagation" and "Oracle" Result

Our main result claims that the proposed method possesses the "oracle" property: the difference between the "oracle" estimate $\tilde{\theta}_{k^*}$ and the adaptive estimate $\hat{\theta}$ measured by $\mathcal{K}(\tilde{\theta}_{k^*},\hat{\theta})$ is of order of the "oracle" risk $N_{k^*}^{-1}$.

Theorem 6: 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

$$\boldsymbol{E}\log(1+|N_{k^*}\mathcal{K}(\tilde{\theta}_{k^*},\hat{\theta})|^r/\mathfrak{z}_{k^*}^r) \le \Delta + \alpha\mathfrak{r}_r/\mathfrak{z}_{k^*}^r + 1.$$

Proof: The "propagation" result of Theorem 5 applies as long as the SMB condition $\Delta_k(\theta) \le \Delta$ is fulfilled, that is, only to the adaptive estimates $\hat{\theta}_1,\dots,\hat{\theta}_{k^*}$ which come out of the algorithm after the first k^* steps. The "oracle" could tell us to stop exactly after the k^* step. However, our adaptive procedure can continue to work after the step k^* if all the criteria $T_{lk} \le \mathfrak{z}_l, l < k$, are satisfied. To establish the accuracy result for the final estimate $\hat{\theta}$, we have to check that the adaptive estimate $\hat{\theta}_k$ does not vary much at the steps after k^*. The definition of the procedure ensures the following "stability" property:

$$N_{k^*}\mathcal{K}(\tilde{\theta}_{k^*},\hat{\theta})1(\varkappa \ge k^*) \le \mathfrak{z}_{k*} \tag{24}$$

because the estimate $\hat{\theta} = \tilde{\theta}_\varkappa$ is accepted.

The definition of the adaptive estimate $\hat{\theta} = \tilde{\theta}_\varkappa$ and (24) imply

$$|N_{k^*}\mathcal{K}(\tilde{\theta}_{k^*},\hat{\theta})|^r = |N_{k^*}\mathcal{K}(\tilde{\theta}_{k^*},\hat{\theta}_{k^*})|^r + |N_{k^*}\mathcal{K}(\tilde{\theta}_{k^*},\hat{\theta})|^r \times 1(\varkappa > k^*) \le |N_{k^*}\mathcal{K}(\tilde{\theta}_{k^*},\hat{\theta}_{k^*})|^r + \mathfrak{z}_{k^*}^r.$$

By the "propagation" condition (11)

$$\boldsymbol{E}_\theta|N_{k^*}\mathcal{K}(\tilde{\theta}_{k^*},\hat{\theta}_{k^*})|^r \le \alpha\mathfrak{r}_r.$$

Now by Lemma

$$\begin{aligned}&\boldsymbol{E}\log(1+|N_{k^*}\mathcal{K}(\tilde{\theta}_{k^*},\hat{\theta})|^r/\mathfrak{z}_{k^*}^r)\\ &\le \Delta + \boldsymbol{E}_\theta|N_{k^*}\mathcal{K}(\tilde{\theta}_{k^*},\hat{\theta})|^r/\mathfrak{z}_{k^*}^r\\ &\le \Delta + \alpha\mathfrak{r}_r/\mathfrak{z}_{k^*}^r + 1\end{aligned}$$

and the required assertion follows. ∎

The presented result states a kind of "oracle" property for the proposed adaptive estimate $\hat{\theta}$. Indeed, due to this result, the normalized stochastic loss $N_{k^*}\mathcal{K}(\tilde{\theta}_{k^*},\hat{\theta})/\mathfrak{z}_{k^*}$ is bounded in the sense of existence of its log-moment. Theorem 5 states the similar bound for the loss $N_{k^*}\mathcal{K}(\tilde{\theta}_{k^*},\theta)$ of the "oracle" estimate. Therefore, the adaptive estimate provides the same accuracy as the "oracle" one up to a factor $\mathfrak{z}_{k^*}$ which comes from the stability result (24) and can be considered as a kind of "payment for adaptation." Due to Theorem 4, $\mathfrak{z}_{k^*}$ 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 $\{\tilde{\theta}_k\}$ for the choice $k = k^*$ up to a logarithmic factor in the sample size.

Reference [4] 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 $(n^{-1}\log n)^{1/(2+d)}$ under Lipschitz smoothness of the function $\theta(\cdot)$ and the usual design regularity; see [13] for more details. Reference [7] showed that in the problem of pointwise adaptive estimation this rate is optimal and cannot be improved by any estimation method.

VII. Conclusion

A novel technique is developed for spatially adaptive estimation. The fitted local likelihood statistics are used for selecting an adaptive neighborhood. The algorithm is developed for a quite general class of observations subject to the exponential distribution. The estimated signal can be uni- and multivariable. The scale dependent thresholds of the developed statistical tests are an important ingredient of the approach. The developed theory justifies both the adaptive estimation procedure and the varying threshold selection. The main theoretical result formulated in Theorem 6 shows the accuracy of the adaptive estimate.

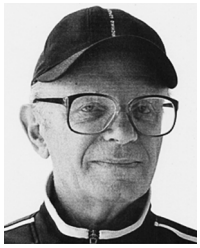
For high-resolution imaging the developed approach is implemented in the form of anisotropic directional estimation with fusing the scale adaptive sectorial estimates. The performance of the algorithm is illustrated for image denoising with data having Poissonian, Gaussian and Bernoulli (binary) random observations. Simulation experiments demonstrate a very good performance of the new algorithm. A demo version of the developed adaptive FLL algorithm is available at the website www.cs.tut.fi/~lasip.

ACKNOWLEDGMENT

The authors would like to thank the associate editor and four anonymous referees for their insightful comments that led to improving the presentation.

REFERENCES

[1] J. Fan J. and I. Gijbels, *Local Polynomial Modelling and its Application*. London, U.K.: Chapman & Hall, 1996.
[2] C. Loader, "Local regression and likelihood," in *Series Statistics and Computing*. New York: Springer-Verlag, 1999.
[3] D. L. Donoho and Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.
[4] O. Lepski, E. Mammen, and V. Spokoiny, "Ideal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selection," *Ann. Stat.*, vol. 25, no. 3, pp. 929–947, 1997.
[5] A. Goldenshluger and A. Nemirovski, "On spatial adaptive estimation of nonparametric regression," *Math. Meth. Stat.*, vol. 6, pp. 135–170, 1997.
[6] O. V. Lepski, "One problem of adaptive estimation in Gaussian white noise," *Theory Probab. Appl.*, vol. 35, no. 3, pp. 459–470, 1990.
[7] O. Lepski and V. Spokoiny, "Optimal pointwise adaptive methods in nonparametric estimation," *Ann. Stat.*, vol. 25, no. 6, pp. 2512–2546, 1997.
[8] V. Katkovnik, "A new method for varying adaptive bandwidth selection," *IEEE Trans. Signal Process.*, vol. 47, no. 9, pp. 2567–2571, Sep. 1999.
[9] V. Katkovnik, K. Egiazarian, and J. Astola, "Adaptive window size image de-noising based on intersection of confidence intervals (ICI) rule," *J. Math. Imag. Vis.*, vol. 16, no. 3, pp. 223–235, 2002.
[10] L. J. Stanković, "Performance analysis of the adaptive algorithm for bias-to-variance trade-off," *IEEE Trans. Signal Process.*, vol. 52, no. 5, pp. 1228–1234, May 2004.
[11] A. Foi, "Anisotropic nonparametric image processing: Theory, algorithms and applications" Ph.D. Thesis, Dip. di Matematica, Politecnico di Milano, Milan, Italy, Apr. 2005, ERLTDD-D01290, [Online]. Available: www.cs.tut.fi/~lasip
[12] V. Katkovnik, K. Egiazarian, and J. Astola, *Local Approximation Techniques in Signal and Image Processing*. Bellingham, WA: SPIE Press, 2006.
[13] J. Polzehl and V. Spokoiny, "Propagation-separation approach for local likelihood estimation," *Probab. Theory Relat. Fields*, vol. 135, no. 3, pp. 335–362, 2005.
[14] I. Ibragimov and R. Khasminskii, *Statistical Estimation*. New York: Springer-Verlag, 1981.
[15] S. Kullback, *Statistics and Information Theory*. New York: Wiley, 1959.
[16] V. Spokoiny, *Local Parametric Methods in Nonparametric Estimation*. New York: Springer, 2008, to be published.
[17] A. Foi, A. R. Bilcu, V. Katkovnik, and K. Egiazarian, "Anisotropic local approximations for pointwise adaptive signal-dependent noise removal," in *Proc. XIII Eur. Signal Process. Conf. (EUSIPCO)*, Antalya, Turkey, Sep. 2005.
[18] W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 9, pp. 891–906, 1991.
[19] J. L. Starck, E. J. Candes, and D. L. Donoho, "The curvelet transform for image denoising," *IEEE Trans. Image Process.*, vol. 11, no. 6, pp. 670–684, 2002.
[20] V. Katkovnik, A. Foi, K. Egiazarian, and J. Astola, "Directional varying scale approximations for anisotropic signal processing," in *Proc. XII Eur. Signal Processing Conf. (EUSIPCO)*, 2004, pp. 101–104.
[21] L. Breiman, "Stacked regression," *Mach. Learn.*, vol. 24, pp. 49–64, 1996.
[22] K. E. Timmermann and R. Nowak, "Multiscale modeling and estimation of Poisson processes with application to photon-limited imaging," *IEEE Trans. Inf. Theory*, vol. 45, no. 3, pp. 846–862, 1999.
[23] R. D. Nowak and R. G. Baraniuk, "Wavelet-domain filtering for photon imaging systems," *IEEE Trans. Image Process.*, vol. 8, no. 5, pp. 666–678, 1999.
[24] R. M. Willett and R. D. Nowak, "Platelets: A multiscale approach for recovering edges and surfaces in photon-limited medical imaging," *IEEE Trans. Med. Imag.*, vol. 22, no. 3, pp. 332–350, 2003.
[25] H. Lu, Y. Kim, and J. M. M. Anderson, "Improved Poisson intensity estimation: Denoising application using Poisson data," *IEEE Trans. Image Process.*, vol. 13, no. 8, pp. 1128–1135, 2004.



Vladimir Katkovnik received the M.Sc., Ph.D., and D.Sc. degrees in technical cybernetics from the Leningrad Polytechnic Institute, Leningrad, Russia, in 1960, 1964, and 1974, respectively.

From 1964 to 1991, he held the positions of Associate Professor and Professor at the Department of Mechanics and Control Processes, Leningrad Polytechnic Institute. From 1991 to 1999, he was a Professor of statistics with the Department of the University of South Africa, Pretoria. From 2001 to 2003, he was a Professor of mechatronics with the Kwangju Institute of Science and Technology, Korea. From 2000 to 2001, and since 2003, he has been a Research Professor with the Institute of Signal Processing, Tampere University of Technology, Tampere, Finland. He has published seven books and more than 200 papers. His research interests include stochastic signal processing, linear and nonlinear filtering, nonparametric estimation, imaging, nonstationary systems, and time-frequency analysis.

Vladimir Spokoiny received the M.Sc. degree in applied mathematics from the Institute of Railway Engineering, Moscow, Russia, and the Ph.D. degree in mathematics from the Lomonosov Slate University, Moscow, Russia.

Since 2000, he has been head of a research group in the Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany, and since January 1, 2002, he has been a Professor of applied statistics with the Department of Mathematics, Humboldt University, Berlin, Germany. His current research directions include adaptive nonparametric smoothing and hypothesis testing, high-dimensional data analysis, statistical methods in finance, image analysis, applications to medicine, classification, and nonlinear time series.