SPATIALLY ADAPTIVE NON-GAUSSIAN IMAGING VIA FITTED LOCAL LIKELIHOOD TECHNIQUE.

Vladimir Katkovnik[1] *and Vladimir Spokoiny*[2]

[1] Signal Processing Institute, University of Technology of Tampere,
P. O. Box 553, Tampere, Finland. E-mail: katkov@cs.tut.fi.
[2] Weierstrass Institute for Applied Analysis and Stochastics,
Mohrenstrasse 39, D - 10117 Berlin, Germany.
E-mail: spokoiny@wias-berlin.de.

ABSTRACT

This paper offers a new technique for spatially adaptive filtering. The fitted local likelihood (FLL) statistics is proposed for selection of an adaptive size estimation neighborhood. The algorithm is developed for quite general observation models subject to the class of the exponential distributions. This algorithm shows a better performance than the intersection of confidence interval (ICI) algorithm, in particular, for Poissonian data. Another principal advantage of the novel technique is that it is non-recursive and does not require knowledge of observation variance.

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] and independently by Goldenshluger and Nemirovsky [5]. This approach has received further development as the intersection of confidence interval (*ICI*) rule in application to various signal and image processing problems [6], [7], [8]. 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 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 and Gaussian observations. Simulation experiments demonstrate a quite good performance of the new algorithm.

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 $\varepsilon_i \sim N(0, \sigma^2)$. 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., $Y_i \sim P(f(X_i))$. 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 $P = (P_\theta, \theta \in \Theta \subseteq \mathbb{R})$ 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 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 $\tilde\theta = \operatorname{argmax}_{\{\theta_i\}} \sum_{i=1}^n \log p(Y_i, \theta_i) = Y$, 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 $\tilde\theta = \operatorname{argmax}_\theta \sum_{i=1}^n \log p(Y_i, \theta)$.

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

$$\tilde\theta_h(x) = \operatorname*{argmax}_\theta \sum_i w_{i,h}(x) \log p(Y_i, \theta), \qquad (1)$$

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 4 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 above are particular cases of the exponential family of distributions. This means that all distribution densities in (1)

are of the form $p(y,\theta) = p(y)\exp(yC(\theta) - B(\theta)), \theta \in \Theta, y \in Y$. 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 $P = (P_\theta)$ 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 [9]) *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,*

$$\tilde\theta_h = S_h/N_h = \sum_{i=1}^n w_{i,h} Y_i \Big/ \sum_{i=1}^n w_{i,h} \qquad (2)$$

and

$$L_h(\tilde\theta_h, \theta) := L_h(\tilde\theta_h) - L_h(\theta) = N_h K(\tilde\theta_h, \theta)$$

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

Here $L_h(\tilde\theta_h, \theta)$ is a "fitted log-likelihood" defined as a difference between the maximized log-likelihood at $\theta = \tilde\theta_h$ and the log-likelihood with an arbitrary θ, $L_h(\tilde\theta_h, \theta) \geq 0$. 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 families.

Model	$K(\theta,\theta')$	$C(\theta)$	$B(\theta)$
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

Now we present some rather tight exponential inequalities for the fitted log-likelihood $L_h(\tilde\theta, \theta)$ 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 [9]) *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* $\mathfrak{z} > 0$

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

In the regular situation, the Kullback-Leibler divergence K fulfills $K(\theta,\theta^*) \approx I_{\theta^*}|\theta - \theta^*|^2$ for any point θ in a neighborhood of θ^*, where I_{θ^*} is the Fisher information at θ^*, see e.g. [10] or [11]. Therefore, the result of Theorem 2 guarantees that $|\tilde\theta_h - \theta^*| \leq C N_h^{-1/2}$ 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* $E_h(\mathfrak{z}_\alpha) = \{\theta : N_h K(\tilde\theta_h, \theta) \leq \mathfrak{z}_\alpha\}$ *is an* α*-confidence set for the parameter* θ^*.

Theorem 3 claims that the estimation loss measured by $K(\tilde\theta_h, \theta)$ 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* $r > 0$

$$\boldsymbol{E}_{\theta^*}\left|L_h(\tilde\theta_h, \theta^*)\right|^r \equiv \boldsymbol{E}_{\theta^*}\left|N_h K(\tilde\theta_h, \theta^*)\right|^r \leq \mathfrak{r}_r\,,$$
$$\mathfrak{r}_r = 2r\int_{\mathfrak{z}\geq 0} \mathfrak{z}^{r-1} e^{-\mathfrak{z}} d\mathfrak{z} = 2r\Gamma(r).$$

Proof. By Theorem 2

$$\boldsymbol{E}_{\theta^*}\left|L_h(\tilde\theta_h, \theta^*)\right|^r \leq -\int_{\mathfrak{z}\geq 0} \mathfrak{z}^r d\boldsymbol{P}_{\theta^*}(L_h(\tilde\theta_h, \theta^*) > \mathfrak{z})$$
$$\leq r\int_{\mathfrak{z}\geq 0} \mathfrak{z}^{r-1} \boldsymbol{P}_{\theta^*}(L_h(\tilde\theta_h, \theta^*) > \mathfrak{z}) d\mathfrak{z} \leq 2r\int_{\mathfrak{z}\geq 0} \mathfrak{z}^{r-1} e^{-\mathfrak{z}} d\mathfrak{z}$$

and the assertion follows. ■

3. LOCAL SCALE SELECTION ALGORITHM

Let $H = \{h_1, \ldots, h_K\}$ be a set of different scales ordered by the smoothing parameter h, and let $\tilde\theta_h = S_h/N_h$ for $h \in H$ 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 denote by $L_k(\theta)$ the log-likelihood for the scale h_k, $k = 1, \ldots, K$. We assume that the scale set H is *ordered* in the sense that the local sample size N_k grows with k.

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 can be presented in the form [12]:

$$\hat k = \max\{k : T_{lk} \leq \mathfrak{z}_l,\ l < k\}, \qquad (3)$$
$$T_{lk} = L_l(\tilde\theta_l, \tilde\theta_k) = N_l K(\tilde\theta_l, \tilde\theta_k).$$

The procedure (3) can be interpreted as follows. The first estimate $\tilde\theta_1$ is always accepted and (3) starts from

$k = 2$. For the current estimate $\tilde{\theta}_2$ is checked whether it belongs to the confidence set $E_{h_1}(\mathfrak{z}_1)$ of the previous step estimate $\tilde{\theta}_1$, see Theorem 3. If not, the estimate $\tilde{\theta}_2$ is rejected and the procedure terminates selecting $\tilde{\theta}_1$. If the inequality $T_{12} = L_1(\tilde{\theta}_1, \tilde{\theta}_2) \leq \mathfrak{z}_1$ 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 $\tilde{\theta}_1, \ldots, \tilde{\theta}_{k-1}$ 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 $T_{lk} = L_l(\tilde{\theta}_l, \tilde{\theta}_k)$ 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 $\mathfrak{z}$'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 $\hat{\theta}$ is $\tilde{\theta}_\varkappa$, $\hat{\theta} = \tilde{\theta}_\varkappa$ .We also define the random moment $\varkappa_k$ meaning the largest index accepted after first k steps and the corresponding adaptive estimate: $\varkappa_k = \min\{\varkappa, k\}, \qquad \hat{\theta}_k = \tilde{\theta}_{\varkappa_k}$.

The *ICI* rule mentioned above can be presented in the sequential form (3) provided that the inequality $T_{lk} \leq \mathfrak{z}_l$ is replaced by $|\tilde{\theta}_l - \tilde{\theta}_k| \leq (\sigma_{\tilde{\theta}_l} + \sigma_{\tilde{\theta}_k})\mathfrak{z}$ 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 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. [8], [13]. 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 [12], 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 $\hat{\theta}_k$ 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 $\hat{\theta}_2, \ldots, \tilde{\theta}_k$ 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 $\tilde{\theta}$ of θ^* is measured by $\boldsymbol{E}\big|K(\hat{\theta}, \theta^*)\big|^r$ for some $r > 0$. Under the null hypothesis $f(X_i) \equiv \theta^*$, every estimate $\tilde{\theta}_k$ fulfills by Theorem 4 for every $r > 0$

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

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 for $k = 2, \ldots, K$

$$\boldsymbol{E}_{\theta^*}\big|L_k(\tilde{\theta}_k, \hat{\theta}_k)\big|^r = \boldsymbol{E}_{\theta^*}\big|N_k K(\tilde{\theta}_k, \hat{\theta}_k)\big|^r \leq \alpha \mathfrak{r}_r\,, \quad (4)$$

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 $\hat{\theta}_k$ coincides with the nonadaptive counterpart $\tilde{\theta}_k$ 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 $l < m \leq k$. The loss associated with such "false alarm" is naturally measured by $\big|N_k K(\tilde{\theta}_k, \hat{\theta}_k)\big|^r$ 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. In this paper 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 that there are constants a_0, a_1, and a_2 such that it holds for every $k \leq K$

$$\mathfrak{z}_k \leq a_0 + a_1 \log \alpha^{-1} + a_2 r \log(N_K / N_k). \quad (5)$$

This result justifies the linear rule

$$\mathfrak{z}_k = \mathfrak{z}_1 - \iota(K - k) \quad (6)$$

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 in a such a way that the condition (4) holds.

4. APPLICATION TO NON-GAUSSIAN IMAGE DENOISING

Points, lines, edges, textures defined by position, orientation and scale even being of small size encode a great proportion of information contained in images. 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 nonsymmetric 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 [14] and recent new ridgelet and curvelet transforms [15].

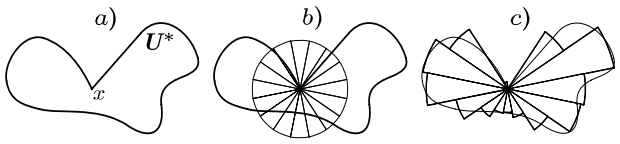

Figure 1. A neighborhood of the estimation point x: a) the best estimation set U^*, b) the unit ball segmentation, c) sectorial approximation of U^*.

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 $\tilde{\theta}_{\alpha,h}(x)$ defined as

$$\tilde{\theta}_{\alpha,h}(x) = \sum_{i \in I_\alpha(x)} w_{i,h}(x) Y_i \Big/ \sum_{i \in I_\alpha(x)} w_{i,h}(x) \qquad (7)$$

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.

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 $\hat{h}_\alpha(x)$.

When these adaptive scales $\hat{h}_\alpha(x)$ 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:

$$\hat{\theta}(x) = \sum_{\alpha \in \mathfrak{A}} {\sum_{i \in I_\alpha(x)}}' w_{i,\hat{h}_\alpha(x)}(x) Y_i \Big/ \sum_{\alpha \in \mathfrak{A}} {\sum_{i \in I_\alpha(x)}}' w_{i,\hat{h}_\alpha(x)}(x). \qquad (8)$$

The sets $I_\alpha(x)$ have as a common point (intersection of the sets) at least the origin. The prime ($'$) in the formula (8) 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 (8) the argument x for $\hat{h}_\alpha(x)$ indicates that the adaptive scales can be varying for each x. In the estimate (8) 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 $\hat{\theta}_\alpha(x)$ be the directional adaptive estimate calculated for the corresponding direction α, that is, $\hat{\theta}_\alpha(x) = \tilde{\theta}_{\alpha,\hat{h}_\alpha(x)}(x)$, see (7). Define also $\hat{\sigma}^2_\alpha(x) = \sigma^2_{\alpha,\hat{h}_\alpha(x)}(x)$ where $\sigma^2_{\alpha,h}(x) = \sum_i w^2_{i,h} / (\sum_i w_{i,h})^2$ is the variance of $\tilde{\theta}_{\alpha,h}$ from (7). Then the final estimate can be yield by fusing of the directional ones as follows

$$\hat{\theta}(x) = \sum_{\alpha \in \mathfrak{A}} \lambda_\alpha(x) \hat{\theta}_\alpha(x), \ \lambda_\alpha = \hat{\sigma}^{-2}_\alpha(x) \Big/ \sum_{\alpha \in \mathfrak{A}} \hat{\sigma}^{-2}_\alpha(x), \qquad (9)$$

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 $\hat{\sigma}^2_\alpha$ in (9) can be calculated assuming that these variances of the observations are equal to one. The inverse variance weighting in (9) assumes that the directional estimates are unbiased and statistically independent. The estimate (8) is quite different from (9). In particular the origin is used here $T = \#(\mathfrak{A})$ times while it enters in (8) 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 [8], [16], where it is successfully exploited with the *ICI* adaptive scale selection for different image processing problems.

Formulas (8)-(9) make clear the algorithm. We introduce the directional estimates $\hat{\theta}_\alpha(x)$, 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. [17]. 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 $\hat{h}_\alpha$. 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 (8)–(9) these filtered *FLL* estimates are used.

5. EXPERIMENTAL STUDY

In these simulation experiments we demonstrate the performance of the developed algorithm for Poissonian and

Gaussian image observations. It is assumes 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 ($RMSE$): $RMSE = \sqrt{\frac{1}{n_1 n_2} \sum_{x \in X} (f(x) - \hat{\theta}(x))^2}$; (2) Signal-to-noise ratio (SNR) in dB: $SNR = 10 \log_{10}(\sum_{x \in X} |f(x)|^2 / \sum_{x \in X} |f(x) - \hat{\theta}(x)|^2)$; (3) Improvement in SNR ($ISNR$) in dB: $ISNR = 20 \log_{10}(\hat{\sigma}_z / RMSE)$, where $\hat{\sigma}_z$ 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: *Testpat* (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 H with a relatively small number of scales $K = 7$. For a linewise uniform window $w(u) = \mathbf{1}(u \leq 1)$ the scale parameter h is integer with the set of values defined as $H = \{\lfloor 1.5^k \rfloor, k = 1, ..., 7\} = \{1, 2, 3, 5, 7, 11, 17\}$. Then $N_k = h_k$ for all $k \leq K$.

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 $B_k^{(j)}$ 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 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.

5.1. 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 χ with the observations defined according to the formula $\tilde{z} \sim P(y \cdot \chi)$, where $\chi > 0$ is a scaling factor. Further, we assume the observations in the form $z = \tilde{z}/\chi$ in order to have the results comparable for different χ as $E\{z\} = E\{\tilde{z}\}/\chi = y$ 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 : $var\{z\} = var\{\tilde{z}\}/\chi^2 = y/\chi$. 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 where the advanced performance of the wavelet based denoising algorithms is demonstrated. It is shown further in [13] that the *ICI* based adaptive algorithms quite competitive and at least numerically demonstrate a very good performance performance. In this paper we compare of the novel proposed *FLL* technique versus these *ICI* adaptive algorithms only.

In the scale selection the *FLL* technique is applied to the Poissonian variables, i.e. to $\tilde{z}$. However, our linear estimates are calculated for the data $z = \tilde{z}/\chi$. 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 for Poissonian distribution in Table 1) is modified to the form $\hat{k} = \max\{m, L_m(\tilde{\theta}^{(m)}, \tilde{\theta}^{(l)}) \leq \mathfrak{z}_l/\chi, l < m\}$.

In these experiments we use the line-wise nonsymmetric windows of the scales H. The linear decreasing threshold set obtained by the simplified choice is as follows $\mathfrak{z} = \{1.2, 1.0, 0.8, 0.6, 0.4, 0.2\}$.

The numerical results in Table 2 are given for the binary "Cheese" image 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 slash correspond to the *FLL* and *LPA–ICI* recursive algorithms (after 7 iterations) respectively. Numerically the *FLL* algorithm works better for *Cheese* and *Cameraman* while for other images the *LPA–ICI* algorithm gives better criterion values. However, visual comparison is always definitely in favor of the *FLL* algorithm as the recursive *LPA–ICI* estimates typically suffer from multiple spot-like artifacts while the *FLL* estimate is free from this sort of degradation effects. 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.

5.2. Gaussian observations

We assume that the additive zero-mean Gaussian noise has the variance $\sigma^2 = 0.01$. For the scales H the linear decreasing thresholds are obtained by the simplified choice $\mathfrak{z} = \{2.5, 2.07, 1.64, 1.21, 0.78, 0.35\}$ with $r = 1/2$ and $\alpha = 1$. Numerically (see Table 4) the performance of the *FLL* algorithm is slightly better (*Cheese*, *Peppers,Testpat)* or slightly worse than that for *LPA–ICI* algorithm. We wish to note that the referred non-recursive *LPA–ICI* algorithm is a specially designed for the Gaussian case while the *FLL* is universally applicable for the class of exponential distributions.

6. 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. 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 $www.cs.tut.fi/\tilde{}lasip$.

7. ACKNOWLEDGMENTS

This work was supported by the Academy of Finland, project No. 213462 (Finnish Centre of Excellence program (2006 - 2011). In part, the work of Dr. Vladimir Katkovnik is supported by *Visiting Fellow* grant from Nokia Foundation.

8. 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, 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.

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

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

[7] 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.

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

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

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

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

[12] V. Spokoiny, *Local parametric methods in nonparametric estimation*, Springer, 2006.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

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

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**
Testpat	9.64/**10.17**	23.31/**23.88**	24.93/**25.53**	14.45/**13.5**

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**
Testpat	**8.05**/7.60	**26.4**/25.95	**28.02**/27.57	**10.13**/10.66

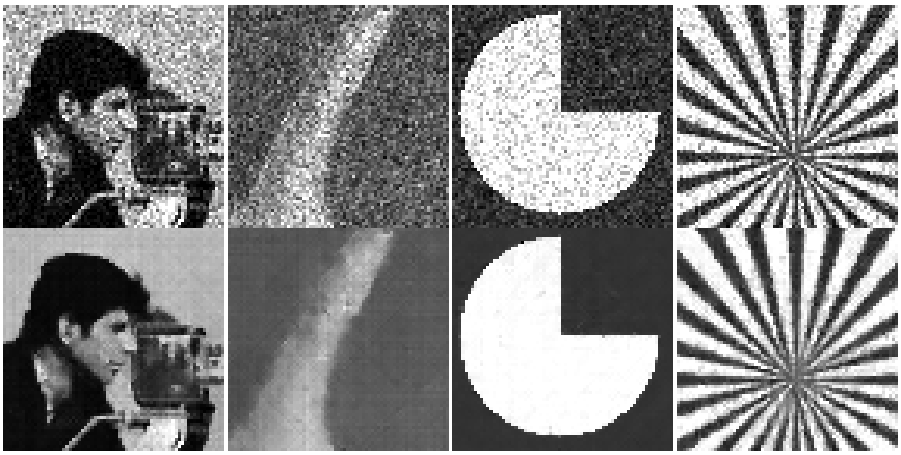

Figure 2. Fragments of noisy and denoised Poissonian images: *Cameraman*, *Peppers*, *Cheese*, *Testpat*.