NONPARAMETRIC REGRESSION IN IMAGING: FROM LOCAL KERNEL TO MULTIPLE-MODEL NONLOCAL COLLABORATIVE FILTERING

Vladimir Katkovnik, Alessandro Foi, Karen Egiazarian, and Jaakko Astola

Department of Signal Processing, Tampere University of Technology
P.O. Box 553, 33101, Tampere, Finland
web: www.cs.tut.fi/~lasip email: firstname.lastname@tut.fi

ABSTRACT

We outline the evolution of the nonparametric regression modelling in imaging from the local Nadaraya-Watson estimates to the nonlocal means and further to the latest nonlocal block-matching techniques based on transform-domain filtering. The considered methods are classified mainly according to two leading features: local/nonlocal and pointwise/multipoint. Here nonlocal is an alternative to local, and multipoint is alternative to pointwise. The alternatives, though an obvious simplification, allow to impose a fruitful and transparent classification of the basic ideas in the advanced techniques. Within this framework, we introduce a novel multiple-model interpretation of the basic modelling used in the BM3D algorithm [11], highlighting a source of the outstanding performance of this type of algorithms.

1. INTRODUCTION

Suppose we have independent random observation pairs $\{z_i, x_i\}_{i=1}^n$ given for simplicity in additive form $z_i = y_i + \varepsilon_i$, where $y_i = y(x_i)$ is a signal of interest, $x_i \in \mathbb{R}^d$ denotes a vector of "features" or explanatory variables which determines the signal observations y_i, and $\varepsilon_i = \varepsilon(x_i)$ is an additive noise. The problem is to reconstruct $y(x)$ from $\{z_i\}_{i=1}^n$. In statistics, the function y is treated as a regression of z on x, $y(x) = E\{z|x\}$. In this way, the reconstruction at hand is from the field of the regression techniques. If a parametric model cannot be proposed for y then, strictly speaking, the problem is from a class of the nonparametric ones. Paradoxically, one of the most productive ideas in nonparametric regression is a parametric local modeling. This localization is developed in a variety modifications and can be exploited for the argument feature variables x, in the signal space y, or in the transform/spectrum domains. This parametric modelling "*in small*" makes a big deal of difference versus the parametric modelling "*in large*".

The idea of local smoothing and local approximation is so natural that it is not surprising it has appeared in many branches of science. Citing [39], we can mention early works in statistics using local polynomials by the Italian meteorologist Schiaparelli (1866) and the Danish actuary Gram (1879) (famous for developing the Gram-Schmidt procedure for orthogonalization of vectors). In sixties-seventies of the twentieth century the idea became subject of an intensive theoretical study and applications: in statistics due Nadaraya (1964, [44]), Watson (1964, [59]), Cleveland (1979, [9]) and in engineering due Brown (1963 [5]), Savitzky and Golay (1964, [53]), Katkovnik (1976 [29]).

Being initially developed as local in x, the technique obtained recently a further significant development with localization in the signal y and in the combined x and y domains as nonlocal means algorithm [6]. For imaging, the nonlocal modelling appeared to be extremely successful when exploited in transform domain. This is a promising direction where the current development is focused.

One of the top achievements in the class of nonlocal transform-based methods is represented by the block-matching 3D (BM3D) algorithm recently proposed for image denoising [11]. The quality demonstrated by this algorithm and its modifications are beyond ability of most alternative techniques.

The scope of the paper is twofold. First, we outline the evolution of the nonparametric regression modelling from the local Nadaraya-Watson estimates to nonlocal means and further to the nonlocal block-matching techniques. Second, we propose a novel interpretation of the basic modelling used in the BM3D algorithm. The term "*multiple-model grouping*" is introduced for this modeling. This novel interpretation allows to highlight a source of the outstanding performance of this type of the algorithms.

In what follows, the considered techniques are classified mainly according to two leading features: *local*/*nonlocal* and *pointwise*/*multipoint*. Here nonlocal is an alternative to local, and multipoint is alternative to pointwise.

We call an algorithm *local* if the weights used in the design of the algorithm depend on the distances from the estimation x^0 and observation x_s points in such a way that to distant points correspond small weights, with the size of the estimation support is essentially restricted by these distances. An algorithm is *nonlocal* if these weights and the estimation support are functions of the differences of the corresponding signal (image intensity) values at the estimation point y^0 and observations y_s. In this way, even distant points can be awarded large weights and the support is often composed of disconnected parts of the image domain. Note that the weights used in local algorithms can be dependent also on y_s, but, nevertheless, the weights are overall dominated by the distance $\left|x^0 - x_s\right|$. An important example of this specific type of local filters is the Yaroslavsky filter [63], referred in [6, 7] as a precursor of the nonlocal means filter.

Let us make clear the pointwise/multipoint alternative. We call an estimator *multipoint* if the estimate is calculated for all observation pixels used in estimation. The set of points used in estimation can be an image block or an arbitrarily-shaped region adaptively or non-adaptively selected. In contrast to a *multipoint* estimator, a *pointwise* estimator gives the estimate for a single point only, namely x^0. To be more flexible, we can say that the multipoint estimator gives the estimates for a set of points while the pointwise is restricted to estimation for a single only. The multipoint estimates are typically not the final ones. The final estimates are calculated by aggregating (fusing) a number of multipoint estimates, since typically many such estimates are available for each point (a common of many overlapping neighborhoods). In the pointwise approach the pointwise estimates are calculated directly as the final ones.

We found that the classification of the algorithms according to these two features: local/nonlocal and pointwise/multipoint is fruitful for giving an overview of this quickly developing field. Table 1 illustrates the proposed classification of the algorithms as well as the organization of this paper.

The local algorithms are well documented in numerous papers and books (e.g., [63], [39], [30], [3], [55]). This is why we only slightly touch this direction and mainly are focused on the comparatively novel emerging area of nonlocal modelling and estimation.

In this paper, image denoising is considered as the basic problem convenient for overview of the various ideas, while these types of algorithms are widely used for a plethora of image processing problems including restoration/deblurring, interpolation, reconstruction,

LOCAL	NONLOCAL
POINTWISE	
Section 2 Signal-independent weights (Sections 2.1-2.2): Nadaraya-Watson [9],[5],[53],[44],[59], LPA [19],[29],[39], Lepski's approach [38], LPA-ICI [25],[31],[30],[20], sliding window transform [62],[61]; Signal-dependent weights (Section 2.3): Yaroslavsky filter [63], SUSAN filter [54], Sigma-filter [37], bilateral filter [57],[15], kernel regression [56], AWS [50],[55].	Section 4 Weighted means (Section 4.1): neighborhood filter [6], NL-means algorithm [6], Lebesgue denoising [60], Exemplar-based [33],[35],[34], scale and rotation invariant [40],[65]; Higher-order models (Section 4.2): kernel regression [8].
MULTIPOINT	
Section 3 Sliding-window transform [45],[46],[14],[64],[16],[26],[27]; shape-adaptive transform [22],[21]; learned bases: adaptive PCA [43], K-SVD [18].	Section 5 Single-model groups (Section 5.1): Blockwise NL-means [6]. Multiple-model groups (Section 5.2): BM3D [11], Shape-Adaptive BM3D [12].

Table 1: Organization of the paper and classification of the algorithms.

enhancement, and compression. In our review and classification, we have no pretension of completeness. The methods and algorithms that appear in Table 1, as well as others to which we refer throughout the text, are cited mainly to give few concrete examples of possible implementations of the general schemes discussed in the next four sections.

2. LOCAL POINTWISE MODELLING

2.1 Pointwise weighted means

The weighted local mean as a nonparametric regression estimator of the form

$$\hat{y}_h(x^0) = \sum_s g_{h,x^0}(x_s) z_s, \quad g_{h,x}(x_s) = \frac{w_h(x - x_s)}{\sum_s w_h(x - x_s)}, \qquad (1)$$

has been independently introduced by Nadaraya [44], as a heuristic idea, and by Watson [59], who derived it from the definition of regression as the conditional expectation and using the Parzen estimate of the conditional probability density.

It is convenient to treat this estimator as a zero-order local polynomial approximation and derive it as a minimizer for the windowed (weighted) mean-squares criterion:

$$\hat{y}_h(x^0) = \hat{C}, \hat{C} = \mathrm{argmin}_C \, I_{h,x^0}(C), \qquad (2)$$

$$I_{h,x^0}(C) = \sum_s w_h(x^0 - x_s)[z_s - C]^2. \qquad (3)$$

The window $w_h(x) = w(x/h)$ defines the neighborhood X_h of x^0 used in the estimator. A scalar (for simplicity) parameter $h > 0$ gives the size of this neighborhood as well as the weights for the observations. In particular, for the Gaussian window we have $w(x) = \exp(-||x||^2)$.

2.2 Pointwise polynomial modelling

In the local polynomial approximation (LPA), the observations z_s in the quadratic criterion (3) are fitted by polynomials. The coefficients of these polynomials found by minimization of I_{h,x^0} serve as the pointwise estimates of y and its derivatives at the point x^0 (e.g. [19], [29], [39], [30]). This sort of estimates is a typical example of what we call pointwise local estimates. Of course, for the zero-order polynomial we obtain the Nadaraya-Watson estimates (1).

2.2.1 Adaptivity of pointwise polynomial estimates

The accuracy of the local estimates is quite dependent on the size and shape of the neighborhood used for estimation. Adaptivity of these estimates is a special subject that recently obtained a wide development concerning, in particular, the adaptive selection of the neighborhood size/shape or of the weights. The main idea of the recent methods is to describe a greatest possible local neighborhood of every pixel in which the local parametric assumption is justified by the data [50], [49], [30]. These methods, mainly linked with the Lepski's approach [38], [25], are valid also for the higher-order local modeling. One of the efficient technique is known as the LPA-ICI algorithm [31],[25]. Here ICI stands for the intersection of confidence intervals (ICI) rule, one of the modifications of the Lepski's approach.

A modern overview of the adaptive local image processing is presented in [30] and [20], a general theory of the adaptive image/signal processing developed for quite general statistical models can be seen in [55].

In this line of algorithms using the localized adaptive weights, we wish to emphasize the works by Polzehl and Spokoiny [50], [49], where efficient adaptive algorithms are developed for the class of exponential distributions.

2.3 Signal-dependent windows

There are a variety of works where the local weights $w_h(x^0 - x_s)$ depend also on the observations z_s. A principal difference of these algorithms versus the nonlocal ones is that all the significant weights are localized in the neighborhood of x^0.

In particular, Smith and Brady [54] presented the SUSAN algorithm where the localization is enabled by the weights depending on the distances from x^0 to the observation points x_s:

$$w_h(x^0 - x_s, y^0 - y_s) = e^{-\frac{||x^0 - x_s||^2}{h^2} - \frac{|y^0 - y_s|^2}{\gamma}}, \quad \gamma, h > 0.$$

Similar ideas are exploited in the Sigma-filter by Lee [37] and in the "bilateral filter" by Tomasi and Manduchi [57], [15]. These algorithms are local, mainly motivated by the edge detection problem where the localization is a natural assumption. Further development and interpretation of this sort of local estimator can be seen in [15] and [4]. In the works by Yaroslavsky [63] the localization of the weights is enabled by taking observations from the ball centered at x^0. The accuracy analysis of this algorithms can be seen in [6].

It this context, it is worth mentioning also the kernel estimator by Takeda et al. [56], which is particular higher-order LPA estimator where the weights are defined as in the bilateral filter.

3. LOCAL MULTIPOINT MODELLING

The main progress in the performance of local (as well as of nonlocal) estimation has been achieved in a direction completely different from that pursued in the aforementioned local modelling, where the low-order polynomial approximations is a main tool. In this section, we consider full-rank high-order approximations with a maximum number of basis functions (typically non-polynomials). For the orthogonal basis functions, this modelling is treated as the corresponding transform-domain representation, with filtering produced by shrinkage in the spectrum (transform) domain. The data are typically processed by overlapping subsets, i.e. windows, blocks or generic neighborhoods, and multiple estimates are obtained for each individual point (e.g., [16], [26] and references therein). Estimation is composed from three successive steps: 1) data windowing (blocking); 2) multipoint processing; 3) calculation of the final estimate by aggregating (fusing) the multiple multipoint estimates. It is found, that this sort of redundant approximations with multiple estimates for each pixel essentially improves the performance of the algorithms.

3.1 Sliding-window transform domain

Let the signal be defined on a regular 2-D grid X. Consider a windowing $\mathcal{C} = \{X_r, r = 1, \ldots, N_s\}$ of X with N_s blocks (uniform windows) $X_r \subset X$ of size $n_r \times n_r$ such that $\cup_{r=1}^{N_s} X_r = X$. Mathematically speaking, this windowing is a *covering of* X. Thus, each $x \in X$ belongs to at least one subset X_r. The blocks may be overlapping and therefore some of the elements may belong to more than one block. The noise-free data $y(x)$ and the noisy data $z(x)$ windowed on X_r are arranged in $n_r \times n_r$ blocks denoted as Y_r and Z_r, respectively.

In what follows, we use transforms (orthonormal series) in conjunction with the concept of the redundancy of natural signals. Mainly these are the 2-D discrete Fourier and cosine transforms (DFT and DCT), orthogonal polynomials, and wavelet transforms. The transform, denoted as $\mathcal{T}_r^{2D}$, is applied for each window X_r independently as

$$\theta_r = \mathcal{T}_r^{2D}(Y_r), \quad \left[= D_r Y_r D_r^T \right] \quad r = 1, \ldots, N_s, \qquad (4)$$

where θ_r is the spectrum of Y_r. The equality enclosed in square brackets holds when the transform $\mathcal{T}_r^{2D}$ is realized as a separable composition of 1-D transforms, each computed by matrix multiplication against an $n_r \times n_r$ orthogonal matrix D_r. The inverse $\mathcal{T}_r^{2D\,-1}$ of $\mathcal{T}_r^{2D}$ defines the signal from the spectrum as

$$Y_r = \mathcal{T}_r^{2D\,-1}(\theta_r), \quad \left[= D_r^T \theta_r D_r \right] \quad r = 1, \ldots, N_s.$$

The noisy spectrum of the noisy signal is defined as

$$\tilde{\theta}_r = \mathcal{T}_r^{2D}(Z_r), \quad \left[= D_r Z_r D_r^T \right] \quad r = 1, \ldots, N_s. \qquad (5)$$

The signal y is *sparse* if it can be well approximated by a small number of non-zero elements of the spectrum θ_r. The number of non-zero elements of θ_r, denoted using the standard notation as $||\theta_r||_0$, is interpreted as the complexity of the model in the block.

The blockwise estimates are simpler for calculation than the estimates produced for the whole image because the blocks are much smaller than the whole image. This is a computational motivation for the blocking. Another even more important point is that the blocking imposes a localization of the image on small pieces where simpler models may fit the observations. These shorter models are easy to be compared and selected. Here we can recognize the basic motivation for the zero-order or low-order LPA, which is simple and for small neighborhoods can well fit the data which globally can instead be complex and not allow a simple parametric modelling. By windowing we introduce a small segments exactly with the same reasons in order to use simple parametric models (expansions in the series defining the corresponding transforms) for overall complex data. A principal difference versus the pointwise estimation is that with blocks the concept of the center actually do not have a proper sense and the estimates are thus calculated for all points in the block. Thus, instead of the pointwise estimation we arrive to the blockwise (multipoint) estimation. For the overlapping blocks this leads to the next problem: the multiple estimates for the points and the necessity to aggregate (fuse) these multiple estimates in the final ones.

3.2 Estimation

For the white Gaussian noise, the penalized minus log-likelihood maximization gives the estimates as

$$\hat{\theta}_r = \underset{\vartheta}{\operatorname{argmin}} ||Z_r - \mathcal{T}_r^{2D\,-1}(\vartheta)||_2^2 + \lambda \operatorname{pen}(\vartheta), \qquad (6)$$
$$\hat{Y}_r = \mathcal{T}_r^{2D\,-1}\left(\hat{\theta}_r\right),$$

where $\operatorname{pen}(\vartheta)$ is a penalty term and $\lambda > 0$ is a parameter that controls the trade-off between the penalty and the fidelity term. The penalty $\operatorname{pen}(\vartheta)$ is used for characterizing the model complexity and appears naturally in this modeling, provided that the spectrum θ_r is random with the prior density $p(\theta_r) \propto e^{-\lambda \operatorname{pen}(\theta_r)}$. The estimator (6) can be presented in the following equivalent form

$$\hat{\theta}_r = \underset{\vartheta}{\operatorname{argmin}} ||\tilde{\theta}_r - \vartheta||_2^2 + \lambda \operatorname{pen}(\vartheta), \qquad (7)$$

where the noisy spectrum is calculated as (5).

If the penalty is additive for the items of the spectrum ϑ, $\operatorname{pen}(\vartheta) = \sum_{i,j} \operatorname{pen}(\vartheta_{(i,j)})$, where $\vartheta_{(i,j)}$ is an element of ϑ, then the problem can be solved independently for each element of the matrix $\hat{\theta}_r$ as a scalar optimization problem:

$$\hat{\theta}_{r,(i,j)} = \underset{x}{\operatorname{argmin}} \left(\tilde{\theta}_{r,(i,j)} - x\right)^2 + \lambda \operatorname{pen}(x). \qquad (8)$$

This solution depends on $\tilde{\theta}_{r,(i,j)}$ and λ, and it can be presented in the form

$$\hat{\theta}_{r,(i,j)} = \rho\left(\tilde{\theta}_{r,(i,j)}, \lambda\right), \qquad (9)$$

where ρ is defined by the penalty function in (8).

Hard and soft thresholding are simple and popular techniques [13]:
(1) *Hard thresholding.* The penalty is $||x||_0$, i.e. $||x||_0 = 1$ if $x \neq 0$ and $||x||_0 = 0$ if $x = 0$. It can be shown that

$$\hat{\theta}_{r,(i,j)} = \tilde{\theta}_{r,(i,j)} \cdot 1\left(|\tilde{\theta}_{r,(i,j)}| \geq \lambda\right). \tag{10}$$

In thresholding for the block of the size $n_r \times n_r$ the so-called universal threshold λ is defined depending on n_r as $\lambda = \sigma\sqrt{2\log n_r^2}$.
(2) *Soft thresholding.* The penalty function is $\text{pen}(x) = ||x||_1 = |x|$. The function ρ in (9) is defined as

$$\rho\left(\tilde{\theta}_{r,(i,j)}, \sigma\right) = \tilde{\theta}_{r,(i,j)} \cdot \left(1 - \lambda/|\tilde{\theta}_{r,(i,j)}|\right)_+. \tag{11}$$

3.3 Aggregation

At the points where the blocks overlap, multiple estimates appear. Then, the final estimate is calculated as the average or a weighted average of these multiple estimates:

$$\hat{y} = \frac{\sum_r \mu_r \hat{y}_r}{\sum_r \mu_r \chi(X_r)}, \tag{12}$$

where $\hat{y}_r$ is obtained by returning the blockwise (multipoint) estimates $\hat{Y}_r = \mathcal{T}_r^{2D\,-1}\left(\hat{\theta}_r\right)$ to the respective place X_r (and extending it as zero outside X_r), $\mu_{r(i,j)}$ are the weights used for these estimates, and $\chi(X_r)$ is the characteristic (indicator) function of X_r.

Although in many works equal weights $\mu_r = 1\ \forall r$ are traditionally used (e.g., [10], [28], [45], [46]), it is a well established fact that the efficiency of the aggregated estimates (12) sensibly depends on the choice of the weights.

In particular, using weights μ_r inversely proportional to the variances of the corresponding estimates $\hat{y}_r$ is found to be a very effective choice, leading to a dramatic improvement of the accuracy of estimation [14],[64].

We wish to mention few related works. In [16], Elad considers shrinkage in redundant representations and derives an optimal estimator minimizing a global energy criterion. Guleryuz [26] studies the use of different weights for aggregating blockwise estimates from sliding window transforms. Vice versa, the optimization of the shrinkage function, given fixed simple averaging of the local estimates, is considered by Hel-Or and Shaked [27].

We note also that earlier versions of sliding/running window filters proposed by Yaroslavsky [62, 61] do not belong to the local multipoint filters because only the central pixel is retained from each blockwise estimate. Thus, there are no multiple estimates and no aggregation and these filters are actually pointwise ones with signal-independent weights.

3.4 Shape-adaptive transform domain

A particularly effective sliding window transform domain filter is obtained when the window is made adaptive with respect to the local image content. The adaptation can be in terms of size or, more generally, of shape.

The approach to estimation for a point x^0 can be roughly described as the following four stage procedure:
Stage I (*spatial adaptation*)**:** For every $x \in X$, define a neighborhood $\tilde{U}_x^+$ of x where a simple low-order polynomial model fits the data;
Stage II (*order selection*)**:** apply some localized transform (parametric series model) to the data on the set $\tilde{U}_x^+$, use thresholding operator (model selection procedure) in order to identify the significant (i.e. nonzero) elements of the transform (and thus the order of the parametric model).
Stage III (*multipoint estimation*)**:** Calculate, by inverse-transformation of the significant elements only, the corresponding estimates $\hat{y}_{\tilde{U}_x^+}(v)$ of the signal for all $v \in \tilde{U}_x^+$. These $\hat{y}_{\tilde{U}_x}$ are calculated for all $x \in X$.
Stage IV (*aggregation*)**:** Let $x^0 \in X$ and $I_{x^0} = \left\{x \in X : x^0 \in \tilde{U}_x^+\right\}$ be the set of the centers of the neighborhoods which have x_0 as a common point. The final estimate $\hat{y}(x^0)$ is calculated as an aggregate of $\left\{\hat{y}_{\tilde{U}_x^+}\left(x^0\right)\right\}_{x \in I_{x^0}}$.

This procedure is at the base of the Pointwise Shape-Adaptive DCT algorithm [22],[21], developed for a number of different image filtering problems. The algorithm shows a very good performance, among the best within the class of local estimators.

3.4.1 Learned bases

Another approach to increase the performance of blockwise estimators is to use transforms or redundant bases that have been optimized with respect to the given image or set of images at hand. The Adaptive Principal Components algorithm by Muresan and Parks [43] and, particularly, the K-SVD algorithm by Elad and Aharon [18] are successful examples of this sort of methods.

4. NONLOCAL POINTWISE MODELLING

4.1 Nonlocal pointwise weighted means

Similar to (2), a nonlocal estimator can be derived as a minimizer for

$$I_{h,x^0}(C) = \sum_s w_h(y^0 - y_s)[z_s - C]^2, \quad y^0 = y(x^0), \tag{13}$$

where the weights w_h depend on the distance between the signal values at the observation points y_s and the desirable point $y^0 = y(x^0)$. Minimization of (13) gives the weighted mean estimate in the form (neighborhood filter [6]):

$$\hat{y}_h(x^0) = \sum_s g_{h,s}(y^0) z_s,\ g_{h,s}(x) = \frac{w_h(y^0 - y_s)}{\sum_s w_h(y^0 - y_s)}. \tag{14}$$

This estimator is local in the signal space y similar to (1) while it can be nonlocal in x depending on the type of the function y.

The ideal set of observations for the noiseless data is the set

$$\{x : y(x) = y^0 = y(x^0)\}, \tag{15}$$

where $y(x)$ takes the value y_0.

The estimate (14) is the weighted mean of the observed z_s and the only link with x^0 goes through $y^0 = y(x^0)$. It is a principal difficulty of this estimate, as it requires to know the accurate y^0 and y_s used in (14). In other words, to calculate the estimate we need to know the estimated signal.

There are a number of ways to deal with this problem.

4.1.1 Weights defined by pointwise differences

The simplest and straightforward idea is replace y_s be z_s, then,

$$\begin{aligned} \hat{y}_h(x^0) &= \sum_s g_{h,s}(z^0) z_s, \\ g_{h,s}(z^0) &= \frac{w_h(z^0 - z_s)}{\sum_s w_h(z^0 - z_s)},\ z^0 = z(x^0). \end{aligned} \tag{16}$$

As the observed z_s are used instead of the true values y_s it results in a principal modification of the very meaning of the estimate (14). Indeed, provided a given weight $g_{h,s}$, this estimate is linear with respect to the observations z_s, while when we use $y_s = z_s$ the estimate (16) becomes nonlinear with respect to the observations and the noise in these observations.

4.1.2 Weights defined by neighborhoodwise differences: NL-means algorithm

The weights in the formula (16) are calculated as differences of individual noisy samples z^0 and z_s. In practice, this can yield a quite different outcome from the difference between the true signal samples y^0 and y_s, assumed in (13).

The nonlocal means (NL-means) as they are introduced in [6] are given in different form where these weights calculated over spatial neighborhoods of the points x^0 and x_s. This neighborhoodwise differences can be interpreted as more reliable way to estimate $y^0 - y_s$ from the noise samples alone. Then, the nonlocal mean estimate is calculated in a pointwise manner as the weighted mean with the weights defined by the proximity measure between the image patches used in the estimate. This estimation can be formalized as minimization of the local criterion similar to (13)

$$I_{h,x^0}(C) = \sum_s w_{h,s}(x^0, x_s)[z_s - C]^2, \tag{17}$$

with, say, Gaussian weights (as it in [6])

$$w_{h,s}(x^0, x_s) = e^{-\frac{\sum_{v\in V}\left(z\left(x^0+v\right)-z(x_s+v)\right)^2}{h}} \tag{18}$$

defined by the Euclidean distance between the observations z in V-neighborhoods of the points x^0 and x_s, V being a fixed neighborhood of 0.

The nonlocal means estimate is calculated as

$$\hat{y}_h(x^0) = \sum_s g_{h,s}(x^0) z_s,\ g_{h,s}(x^0) = \frac{w_{h,s}(x^0, x_s)}{\sum_s w_{h,s}(x^0, x_s)}. \tag{19}$$

The detailed review of the nonlocal means estimates with a number of generalizations and developments are presented by Buades, Coll and Morel [6],[7]. From the results in [6], we wish to note the accuracy analysis of the estimator (16) with respect to both signal y and the noise. These asymptotic accuracy results are given for $h \to 0$ and exploited to prove that the nonlocal mean estimates can be asymptotically optimal under a generic statistical image modeling. This sort of estimates has been developed, more less in parallel, in a number of publications with different motivation varying from computer vision ideas to statistical nonparametric regression (see, e.g., [6], [60], [33], [35], [34], [7] and references therein). Extension of the original approach including scale and rotation invariance for the data patches used to define the weights are proposed in [40] and [65].

4.1.3 Recursive reweighting

The next natural idea is to use for the weights $g_{h,s}$ preprocessed observations $\hat{z}_s$, say, prefiltered by a procedure independent of (16):

$$\begin{aligned} \hat{y}_h(x^0) &= \sum_s g_{h,s}(\hat{z}^0) z_s, \\ g_{h,s}(\hat{z}^0) &= \frac{w_h(\hat{z}^0 - \hat{z}_s)}{\sum_s w_h(\hat{z}^0 - \hat{z}_s)}. \end{aligned} \tag{20}$$

For the prefiltering we can exploit the estimate of the same nonlocal average (16) $\hat{z}_s = \hat{y}_h(x^s)$. Then the algorithm becomes recursive with successive of use the estimates for the weight recalculation:

$$\begin{aligned} \hat{y}_h^{(k+1)}(x^0) &= \sum_s g_{h,s}(\hat{y}_h^{(k)}(x^0)) z_s, \qquad x^0 \in X, \\ g_{h,s}(y_h^{(k)}(x^0)) &= \frac{w_h(\hat{y}_h^{(k)}(x^0) - \hat{y}_h^{(k)}(x_s))}{\sum_s w_h(\hat{y}_h^{(k)}(x^0) - \hat{y}_h^{(k)}(x_s))}. \end{aligned} \tag{21}$$

If the algorithm converges, the limit recursive estimate $\hat{y}_h$ is a solution of the set of the nonlinear equations

$$\begin{aligned} \hat{y}_h(x^0) &= \sum_s g_{h,s}(\hat{y}_h(x^0)) z_s, \qquad x^0 \in X, \\ g_{h,s}(\hat{y}_h(x^0)) &= \frac{w_h(\hat{y}_h(x^0) - \hat{y}_h(x_s))}{\sum_s w_h(\hat{y}_h(x^0) - \hat{y}_h(x_s))}. \end{aligned} \tag{22}$$

These estimates can be very different from the estimates (20) which can be treated as a first step of the recursive procedure (21). We do not know results concerning the study of these estimates for the filtering of z which are recursive on $\hat{y}_h^{(k)}$. However, recursive equations of a similar style are considered by the methods referred in Section 4.3.

4.1.4 Weights averaging: Bayesian approach

There is an alternative idea how to deal with the dependence of the weights w_h on the unknown signal y. Let us use the Bayesian rationale and replace the local criterion (13) by an a-posteriori conditional mean calculated provided that the given observations are fixed:

$$\tilde{I}_{h,x^0}(C) = E_y\{I_{h,x^0}(C) | z_s, s = 1, \ldots, N\}. \tag{23}$$

Assume for simplicity that we consider the scalar case, $d = 1$, then y_s are random and independent with the priori p.d.f. $p_0(y_s)$, then the conditional p.d.f. of y_s provided a given z_s is calculated according to the Bayes formula:

$$p(y_s|z_s) = \frac{p(z_s|y_s)p_0(y_s)}{\int p(z_s|y_s)p_0(y_s)dy_s}.$$

For the Gaussian observations model $z_s = \mathcal{N}(y_s, \sigma^2)$ and $p_0(y_s) = const.$, it gives

$$p(y_s|z_s) \propto p(z_s|y_s) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(z_s - y_s)^2}{2\sigma^2}}.$$

Thus, (23) is easily calculated as

$$\begin{aligned} &\tilde{I}_{h,x^0}(C) = \\ &= \sum_s \int\int p(y_0|z_0)p(y_s|z_s)w_h(y^0 - y_s)[z_s - C]^2 dy_s dy_0 = \\ &= \sum_s \tilde{w}_h(z^0 - z_s)[z_s - C]^2. \end{aligned}$$

In particular, for the Gaussian window

$$w_h(y) = \frac{1}{\sqrt{2\pi}h} e^{-\frac{y^2}{2h^2}},$$

tedious calculations show that

$$\tilde{w}_h(z) \propto e^{-\frac{z^2}{2(h^2+2\sigma^2)}},$$

where the proportionality factor depends on h and σ but not on z.

Provided a change of the parameter h in the weight function w_h for $\sqrt{h^2 + 2\sigma^2}$, we have $\tilde{w}_h(z) \propto w_h(z)$, which makes this weight function legitimate for the use with noisy data z_s instead of unknown y_s. The larger value of h, coming from the change of parameter, means a larger window size and stronger smoothing, in some sense equivalent to data prefiltering.

4.2 Nonlocal pointwise higher-order models

Use of the higher-order LPA in the local estimates is well know and well studied area (e.g., [30]). In particular, for the first-order estimate we have the criterion and the estimate in the form

$$I_{h,x^0}(C,C_1)=\sum_s w_h(x^0-x_s)[z_s-C_0-C_1(x^0-x_s)]^2, \quad (24)$$
$$\hat{y}_h(x^0)=\hat{C}_0,\ (\hat{C}_0,\hat{C}_1)=\underset{C_0,C_1}{\operatorname{argmin}} I_{h,x^0}(C_0,C_1),$$

where the weights are defined as in (1). Recall that $\hat{C}_1$ in (24) is an estimate of the derivative $\partial y(x^0)/\partial x$.

Let us try to use this first-order LPA model in the context of the nonlocal mean (13) and combine the weights depending of the distance between the signal values from (13) with the linear on x fit for the observed z_s from (24). Then the nonlocal criterion is of the form

$$\begin{aligned} I_{h,x^0}(C) &= \sum_s w_h(y^0-y_s)[z_s-C_0-C_1(x-x_s)]^2, \quad (25)\\ y^0 &= y(x^0). \end{aligned}$$

Again $\hat{C}_1$ is an estimate of the derivative $\partial y(x^0)/\partial x$. Accordingly to the used windowing the ideal neighborhood X^* is defined as in (15), i.e. it is a set of x where $y(x)=y^0$. However, the derivative $\partial y/\partial x$ can be different for the points in this X^* and then the linear model $C+C_1(x-x_s)$ does not fit $y(x)$ for all $x\in X^*$. Figure 1 illustrates a possible situation, where the set X^* includes all $y(x)=y$ but the derivatives in this points have different signs.

The ideal neighborhood should be different from (15) and include both the signal and derivative values

$$X^*=\left\{x: y(x)=y(x^0),\ \frac{\partial y(x)}{\partial x}=\frac{\partial y(x^0)}{\partial x}\right\}. \quad (26)$$

It follows from this consideration that, for the class of the nonlocal estimators, the windowing function w_h should correspond to the model used in estimation and actually incorporate this model. For the linear model it can be done selecting the window function defining the distance in both the signal and signal derivative values. In particular as follows

$$\begin{aligned} I_{h,x^0}(C) =& \\ =\sum_s w_{h_1}(y^0-y_s)w_{h_2}&\left(\tfrac{\partial y(x^0)}{\partial x}-\tfrac{\partial y(x_s)}{\partial x}\right)\cdot \\ &\cdot[z_s-C-C_1(x-x_s)]^2. \quad (27)\end{aligned}$$

In implementation of this estimation, the unknown y_s and $\partial y(x_s)/\partial x$ could be replaced by the corresponding estimates obtained from LPA or by independent estimates as it is discussed in the previous section.

Figure 1 illustrates the differences between the neighborhoods used for estimation in the case of the local pointwise model (1) and the nonlocal zero and first order models. The area **III** shows the local neighborhood for the local pointwise estimate defined by the window width parameter h. For the nonlocal zero-order modelling (25), the neighborhood is defined as a set of x values where $|y-y^0|\le\Delta$. In the figure this area is a defined as the union of all the subareas **I** and **II**. However, if the first order model is used for the nonlocal modelling according to (26)-(27) at least the sign of the derivative $\partial y/\partial x$ should also be taken in consideration. Thus, if we say that for the desired neighborhoods $\partial y/\partial x>0$ or $\partial y/\partial x<0$, there two different sets defined either as the union of the subareas **I** or as the union of the subareas **II**, respectively. In this sense, the nonlocal zero-order model does not distinguish between the subareas **I** and **II**.

While the low-order polynomial approximations for the local estimates is one of the main streams in the theory and in applications, it has not received sufficient attention in nonlocal setting. The first results in this direction are reported in [8], where the polynomial approximations up to second order are used. However, the polynomial modelling is not included in the window function, where weights depending only on the signal values (and not on the derivatives) are used. We mention also the work [1], where different models of self-similarity in images are studied, with particular emphasis on affine (i.e. first order) similarity between blocks.

While in the above text we considered only polynomial expansions, of course, the higher-order modeling is not restricted to polynomials. The more general case using transforms is illustrated directly in the forthcoming Section 5 for multipoint modeling.

4.3 Variational formulations

A variety of methods for image denoising are derived by considering image processing as a variational problem where the restored image is computed by minimization of some energy functional. Typically, such functionals consist of a fidelity term such as the norm of the difference between the true image and the observed noisy image and a regularization penalty term:

$$J=\lambda||y-z||_2^2+\mathrm{pen}(y). \quad (28)$$

One of the successful filters in this class is the Rudin-Osher-Fatemi (ROF) method [52],[51]. Here, the clear images defined by a variational problem using the total variation penalty. The success of this penalty stems from the fact that it allows discontinuous solutions and hence preserves edges while filtering high-frequency oscillations due to noise. Several other methods are derived from the original ROF model [42],[47],[58]. Overall, these methods can be treated as essentially local methods [36]. The regularization involves only the signal and its derivatives evaluated at the same point, resulting in a Euler-Lagrange equation in differential form.

Recently, a novel class of the variational methods involving nonlocal terms has been proposed (see [36],[24],[23],[40],[41] and references therein) where the corresponding Euler-Lagrange equations takes a differential-integral form. These new methods have been motivated by the concept of the nonlocal means, used to define nonlocal differential operators calculated over some neighborhoods.

First, it is shown that the nonlocal means can be derived by minimizing a special functional. Second, this functional is used as the penalty term in (28), where

$$\mathrm{pen}(y)=\int g\left(\frac{|y(x)-y(v)|^2}{h^2}\right)w(|x-v|)dxdv, \quad (29)$$

$w>0$ is a window function, and g is a differentiable function used for filter design. Minimization of (29) on y gives the equation

$$\begin{aligned} y(x)&=\frac{1}{C(x)}\int g'\left(\tfrac{|y(x)-y(v)|^2}{h^2}\right)y(x')w(|x-v|)dv, \quad (30)\\ C(x)&=\int g'\left(\tfrac{|y(x)-y(v)|^2}{h^2}\right)w(|x-v|)dv. \end{aligned}$$

In particular, for $g=1-\exp(-x)$, it gives $g'\left(\frac{|y(x)-y(v)|^2}{h^2}\right)=\exp\left(-\frac{|y(x)-y(v)|^2}{h^2}\right)$.

The image reconstruction is achieved by a recursive minimization of the criterion (28) using the iteration given by (30):

$$\hat{y}^{(k+1)}(x)=\frac{1}{C(x)}\int g'\left(\frac{|\hat{y}^{(k)}(x)-\hat{y}^{(k)}(v)|^2}{h^2}\right)\hat{y}^{(k)}(v)w(|x-v|)du. \quad (31)$$

The first iteration of this algorithm with $\hat{y}^{(0)}=z$ can be interpreted as the nonlocal estimates (16).

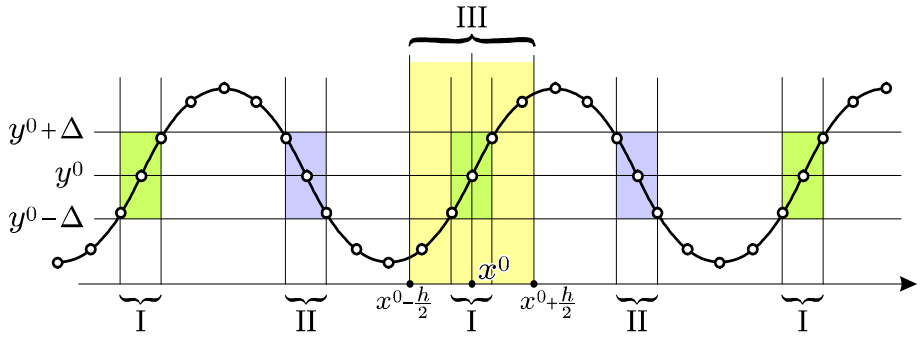

Figure 1: Local versus nonlocal supports for zero- and first-order polynomial fitting: local (1) **III**; nonlocal zero-order model (25) **I**∪**II**; nonlocal first-order model (26)-(27) **II**.

It is interesting to note also that these iterations look similar to the recursive procedure (21). Actually, these iterations deal with the same problem of how to calculate weights that depend on the unknown signal y.

Let us go back to the formulation (28). Using (29)-(30), we arrive to the equation including the observations z

$$y(x)=\frac{1}{C(x)}\left(\lambda z+\int g'\left(\tfrac{|y(x)-y(v)|^2}{h^2}\right)y(v)w(|x-v|)dv\right),$$

To conclude this section, we wish to note that the concepts of locality and nonlocality, as well as the derived algorithms, are different for the nonparametric regression approach, on which we focus, and for the variational formulations, sketched here. Overall, as it is clear from what was discussed, there are very interesting connections and parallels between these different approaches.

5. NONLOCAL MULTIPOINT MODELLING

5.1 Single-model groups

As in Section 3.1, we consider the blocks Y_j obtained by windowing. Furthermore, we assume that there is a similarity between some of these blocks. Following the pointwise nonlocal mean (13), we can introduce a nonlocal multipoint estimator by the criterion

$$I_{Y_r}(\vartheta)=\sum_j w(||Y_j-Y_r||_2^2)||Z_j-\mathcal{T}^{2\mathrm{D}\,-1}(\vartheta)||_2^2+\lambda\,\mathrm{pen}(\vartheta). \tag{32}$$

Here w is a weight function defining a correspondence of the block Y_j to the so-called reference-block Y_r, $||Z_j-\mathcal{T}^{2\mathrm{D}\,-1}(\vartheta)||_2^2$ is a measure of discrepancy between the observed Z_j and the model $\mathcal{T}^{2\mathrm{D}\,-1}(\vartheta)$, the penalty term $\lambda\,\mathrm{pen}(\vartheta)$ controls the complexity of the model or smoothness of the estimate. The model is expressed by the $\mathcal{T}^{2\mathrm{D}}$-spectrum ϑ.

More specifically, due to the orthonormality of $\mathcal{T}^{2\mathrm{D}}$, (32) can be presented in the spectral variables only:

$$I_{Y_r}(\vartheta)=\sum_j w(||\theta_j-\theta_r||_2^2)||\tilde{\theta}_j-\vartheta||_2^2+\lambda\,\mathrm{pen}(\vartheta). \tag{33}$$

5.1.1 Groups

For simplicity, the weights in (32) can be replaced by indicators

$$w(||Y_j-Y_r||_2^2)=\mathbf{1}(||Y_j-Y_r||_2^2<\Delta), \tag{34}$$

where $\Delta>0$ is a similarity threshold. This means that $w(||Y_j-Y_r||_2^2)=1$ if $||Y_j-Y_r||_2^2\le\Delta$ and 0 otherwise. By denoting as K_r^Δ the set of indexes j for which these weights are nonzero, (33) can be given in the form

$$I_{Y_r}(\vartheta)=\sum_{j\in K_r^\Delta}||\tilde{\theta}_j-\vartheta||_2^2+\lambda\,\mathrm{pen}(\vartheta). \tag{35}$$

The set of blocks selected according to (34) is called the *group corresponding to the reference block* Y_r. The ideal set of observations corresponding to the reference block in (32) and (33) is

$$K_r^*=K_r^0=\{x:Y_j=Y_r,\,j=1,\dots,N\}, \tag{36}$$

i.e. the selected blocks Y_j are identical to the reference block Y_r. The inequality in the rule (34) relaxes this strict requirement for the blocks similar enough to the reference block.

The aim of grouping is a joint processing of the windowed data in the group. The criterion (35) can be rewritten as

$$I_{Y_r}(\vartheta)=\#(K_r^\Delta)||\bar{\theta}_r-\vartheta||_2^2+\lambda\,\mathrm{pen}(\vartheta)+\mathrm{const.},$$

where

$$\bar{\theta}_r=\frac{1}{\#(K_r^\Delta)}\sum_{j\in K_r^\Delta}\tilde{\theta}_j, \tag{37}$$

and $\#(K_r^\Delta)$ is the cardinality of the set K_r^Δ of the blocks included in the estimate for the reference block Y_r.

Then

$$\hat{\theta}_r=\operatorname*{argmin}_\vartheta I_{Y_r}(\vartheta)= \tag{38}$$
$$\operatorname*{argmin}_\vartheta||\bar{\theta}_r-\vartheta||_2^2+\frac{\lambda}{\#(K_r^\Delta)}\mathrm{pen}(\vartheta)=\rho\left(\bar{\theta}_r,\frac{\lambda}{\#(K_r^\Delta)}\right).$$

It means that for the penalty additive with respect to the elements of the spectrum, $\hat{\theta}_r$ is obtained by thresholding the sample mean estimate (37). Once the spectrum elements of the reference block are found, the signal multipoint estimate for the reference block is calculated according to the formula

$$\hat{Y}_r=\mathcal{T}^{2\mathrm{D}\,-1}\left(\hat{\theta}_r\right). \tag{39}$$

The final estimate of the signal is obtained according to the aggregation formula (12).

A variety of quite different versions of the considered approach can be developed. First, various estimates of unknown Y_j and Y_r can be used in the block's grouping rule; second, different metrics for comparison of this estimates and the weights $w(||Y_j-Y_r||_2^2)$ in the estimates. Finally, various forms of shrinkage can be applied to the block-wise estimates $\tilde{\theta}_r$ before and after averaging in (37). We wish to note that already in [6], a blockwise version of the nonlocal means is suggested, where similar blocks are averaged together based on their similarity but without a penalty which could regularize the model for the block estimate.

In general, the estimator described above corresponds to what we may call a *single (parametric) model* approach, because for each

group of blocks a unique parametric model [in the form $\mathcal{T}^{2D\,-1}(\vartheta)$ in (32)] is used, where the parameter θ is taking values that will fit for all grouped blocks. It results in a specific use of this blockwise estimates in the group where they are combined as a sample mean or as a weighted mean estimates similar to (37).

As it is already mentioned in the previous subsection, the weighted means in the form (12) allows significantly improve over the multipoint estimate (39), in particular using the inverse variances of the estimates as the weights. However, this weighting does not follow from the used problem formulation and can be treated as a heuristic modification of the algorithm obtained by accurate optimization of the penalized energy criterion.

5.2 Multiple-model groups: collaborative filtering

In this section we formally derive the block-matching 3D (BM3D) algorithm proposed in [11] considering a penalized energy criterion where separate models are used for the data in each block. In this way, we obtain a *multiple-model group*. In our modelling, we use the same $\mathcal{T}^{2D}$-basis functions for all blocks and say that the models are different if their $\mathcal{T}^{2D}$-spectra are allowed to be different.

In (32), for each block, the observed Z_j are fitted by $\mathcal{T}^{2D\,-1}(\vartheta)$, where ϑ is the same for all j. This is a single-model group. Let us assume that, in this fitting, ϑ can take different values ϑ_j for different Z_j in the same group. Then, the quadratic part of the criterion (33) is changed and we obtain the multiple-model criterion:

$$I_{Y_r}(\{\vartheta_j\}_j) = \\ = \left(\sum_j w(||\theta_j-\theta_r||_2^2)||Z_j-\mathcal{T}^{2D\,-1}(\vartheta_j)||_2^2\right)+\lambda\,\mathrm{pen}\left(\{\vartheta_j\}_j\right). \tag{40}$$

In the transform domain it gives

$$I_{Y_r}(\{\vartheta_j\}_j)=\left(\sum_j w(||\theta_j-\theta_r||_2^2)||\tilde{\theta}_j-\vartheta_j||_2^2\right)+\lambda\,\mathrm{pen}\left(\{\vartheta_j\}_j\right), \tag{41}$$

for $\tilde{\theta}_r=\mathcal{T}^{2D}(Z_r)$.

Here, if the penalty term is additive with respect to j, the minimization of I_{Y_r} is trivialized and the very meaning of group is lost, because the solution is obtained by minimizing independently for each j. As a matter of fact, once a multiple-model group is assembled, it is the penalty term that should establish the interaction between different members of the group. A practical way to establish such interaction is the following.

5.2.1 Collaborative filtering

For transparency, let us simplify again the weights w to an indicator of the form (34). In this way, (41) becomes

$$I_{Y_r}(\{\vartheta_j\}_j)=\left(\sum_{j\in K_r^\Delta}||\tilde{\theta}_j-\vartheta_j||_2^2\right)+\lambda\,\mathrm{pen}\left(\{\vartheta_j\}_{j\in K_r^\Delta}\right).$$

Let us consider $\tilde{\Theta}_r=\{\tilde{\theta}_j\}_{j\in K_r^\Delta}$ be the set of $\mathcal{T}^{2D}$-spectra in the group, which we can treat as 3-D array, where j is the index used for the third dimension. Apply a 1-D orthonormal transform $\mathcal{T}^{1D}$ with respect to j. In this way we arrive to a groupwise 3-D spectrum of the group as $\tilde{\Omega}_r=\mathcal{T}^{1D}(\tilde{\Theta}_r)$. Consistent with this representation, we replace the penalty pen$(\{\vartheta_j\}_j)$ with an equivalent penalty pen(Ω), where $\Omega=\mathcal{T}^{1D}(\{\vartheta_j\}_{j\in K_r^\Delta})$ is the corresponding 3-D spectrum obtained by applying the 1-D transform $\mathcal{T}^{1D}$ on the collection of 2-D spectra $\{\vartheta_j\}_{j\in K_r^\Delta}$. We denote the 3-D transform obtained by the composition of $\mathcal{T}^{1D}$ and $\mathcal{T}^{2D}$ as $\mathcal{T}^{3D}$.

We use this 3-D spectrum representation as a special model of data collected in this group, with the penalty pen(Ω) defining the complexity of the data in the group:

$$I_{Y_r}(\Omega)=||\tilde{\Omega}_r-\Omega||_2^2+\lambda\,\mathrm{pen}(\Omega).$$

Then, the estimation of the true Ω_r is defined as

$$\hat{\Omega}_r=\underset{\Omega}{\mathrm{argmin}}\left(||\tilde{\Omega}_r-\Omega||_2^2+\lambda\,\mathrm{pen}(\Omega)\right), \tag{42}$$

$$\hat{\Theta}_r=\{\hat{\theta}_{r,j}\}_{j\in K_r^\Delta}=\mathcal{T}^{1D\,-1}\left(\hat{\Omega}_r\right),$$

$$\hat{Y}_{r,j}=\mathcal{T}^{2D\,-1}\left(\hat{\theta}_{r,j}\right). \tag{43}$$

Again, if the penalty pen(Ω) is additive with respect to the components of Ω, the minimization in (42) is scalar and independent for each element of Ω; thus, it can be solved by thresholding of $\tilde{\Omega}_r$. The consecutive $\mathcal{T}^{1D\,-1}$ and $\mathcal{T}^{2D\,-1}$ inverse transforms return first the estimates $\hat{\Theta}_r=\{\hat{\theta}_{r,j}\}_{j\in K_r^\Delta}$ of $\mathcal{T}^{2D}$-spectra of the blocks in the group, and hence the multipoint estimates $\hat{Y}_{r,j}$ of these blocks. Because these estimates can be different in different groups, we use the double indexes for the signal estimates $\hat{Y}_{r,j}$, where j stays for the index of the block and r for the group where these estimates are obtained.

It gives a clear idea of the principal specific features of the multiple modeling used in this section versus the single-model group modelling.

First, the filtering (thresholding) gives individual estimates for each block in the group. Note that in the single-model group of the previous section a unique estimate is calculated and used for the reference-block only.

Second, an essential difference exists in how the data in the group are processed. The sample mean or weighted mean estimate (37) means that the data in the group are treated as quite relevant (reliable) to the signal estimated for this reference block. Contrary to it, the multiple-model approach produces individual estimates for all participants of the group (collaborative filtering), where the joint spectrum in $\hat{\Omega}_r$ is exploited in order to improve the estimates for each of the blocks in the group. Thus, we obtain a more flexible technique where say an erroneously included block is not able to damage seriously the estimates of other blocks and itself could not be damaged by data from other blocks.

Figure 2 illustrates the collaborative filtering procedure in the particular case of hard-thresholding of the 3-D spectrum: after shrinkage there remain only few nonzero coefficients in the 3-D spectrum. This sparsity is due both to decorrelation within each grouped block operated by the $\mathcal{T}^{2D}$ (intra-block decorrelation) and to decorrelation across the corresponding spectral components of the block operated by the $\mathcal{T}^{1D}$ (inter-block decorrelation). After applying the $\mathcal{T}^{2D\,-1}$ inverse transform, we obtain a number of intermediate block estimates (the red stack at the top-right of the figure). Each of these is obviously $\mathcal{T}^{2D}$-sparse. The blockwise estimates (the purple stack at the bottom-right of the figure) are obtained by applying the $\mathcal{T}^{1D\,-1}$ inverse transform on the intermediate block estimates. As a matter of fact, each one of the blockwise estimates is calculated as a linear combination of the intermediate estimates, where the coefficients of the linear combination are simply the coefficients of one of the $\mathcal{T}^{1D}$ basis elements. Note that the blockwise estimates are not necessarily $\mathcal{T}^{2D}$-sparse than the intermediate estimates, as it is illustrated in the figure. In a very broad sense, our results support the idea that in multipoint image estimation a weighted average of a few sparse estimates is better than single sparse estimate alone [17]. In the single-model group we have a penalty that enforces sparsity on a single estimate, whereas in the multiple-model group the sparsity is enforced for the group as a whole but not on the individual blockwise estimates, which are instead a linear combination of intermediate blockwise estimates that are sparse.

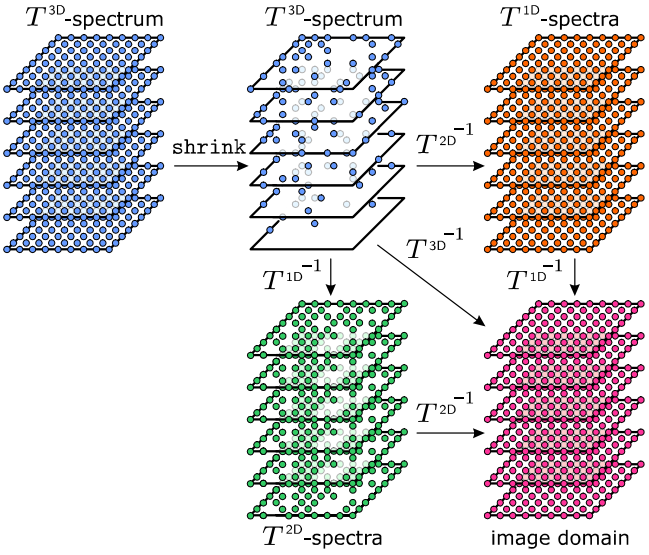

Figure 2: Illustration of the collaborative hard-thresholding.

5.2.2 Aggregation

As a result of the multiple estimation, we obtain multiple estimates for each x in X and the final signal estimate is calculated by fusing these blockwise estimates in a single final one. The main formula used for this aggregation is the weighted mean (12).

6. CONCLUSION

In this paper we reviewed recent developments in the field of nonparametric regression modeling and image processing.

In these conclusive comments, we would like to discuss some theoretical aspects of these developments and, in particular, what problems, of principal importance in our opinion, are not solved.

The considered methods were classified mainly according to two leading features: local/nonlocal and pointwise/multipoint. This discussion is organized according to this classification.

(1) **Local pointwise estimators.**

These estimators are supported by strong theoretical results covering nearly all aspects of estimation: estimation accuracy, adaptation with varying spatially adaptive neighborhoods, etc.

Unsolved problem: simultaneous selection of adaptive neighborhood and order of the local parametric model.

It is a generalized model selection problem where the model support is treated as an element of the model selection. Note that this setting is very different from the classical model selection problem where the model support is assumed to be fixed.

(2) **Local multipoint estimators**.

(2a) These estimators deal with multiple preliminary estimators and the final estimate is calculated by aggregation (fusing) of the preliminary estimates. The existing aggregation methods assume that the models of the preliminary estimates are given.

Unsolved problem: simultaneous optimization of aggregation and models for the preliminary estimates.

(2b) These two step procedures actually heuristic or semiheuristic as the aggregation turned out as the only method to exploit the produced redundant estimates.

Unsolved problem: development of the statistical observation model leading to the two step procedure with the preliminary and aggregation step as a result of some standard statistical estimation technique (ML, EM, etc.).

(3) **Nonlocal pointwise and multipoint estimators.**

(3a) The signal dependent weights define the support of the estimator, i.e. the points included in estimate, and the weights of the included observations. In many developments, in particular in our BM3D algorithm, the use of the indicator window defines the nonlocal support while the details of the comparative weighting of this windows is ignored. Under this simplification, all basic aspects of the algorithms are similar to the ones of standard transform-domain filtering.

The situation becomes much different when we take into consideration the window weights varying according to unknown signal values. Using estimates for these unknown values results in the estimates which are principally different from the usual local ones. The limit estimate is a solution of the nonlinear equation (22):

$$\hat{y}_h(x^0) = \sum_s g_{h,s}(\hat{y}_h(x^0))z_s\,.$$

It is difficult to say what sort of estimate we obtain even for the noiseless signal. For the local estimates with the signal-independent kernel $g_{h,s}$ we know eigenfunctions of this kernel (polynomial for the LPA) and we know the smoothing properties of this filter. For the case of signal-dependent kernel the smoothing properties of this filter are actually unknown. The works by Buades et al. [6] and Kindermann et al. [36] are only very first steps in the direction of studying this sort of nonlinear operators.

Unsolved problem: smoothing properties of the nonlocal pointwise and multipoint estimator with respect to noiseless and noisy signals.

(3b) This point similar to (2b) but for the nonlocal estimators.

Unsolved problem: development of the statistical observation model leading to the windowing, grouping, blockwise estimation and aggregation as a result of some standard statistical estimation technique (ML, EM, etc.).

The model (criteria (40)-(42)) proposed in this paper gives only the blockwise/groupwise estimates while the windowing and the aggregation are treated as separate steps. Use of the mix-distribution for observation modelling in the work [32] was one of the first attempts to move in this direction.

7. ACKNOWLEDGMENTS

This work was supported by the Academy of Finland (application no. 213462, Finnish Programme for Centres of Excellence in Research 2006-2011, and application no. 118312, Finland Distinguished Professor Programme 2007-2010).

REFERENCES

[1] S. Alexander, S. Kovacic, and E. Vrscay, "A Simple Model for Image Self-Similarity and the Possible Use of Mutual Information," *Proc. 15th Eur. Signal Process. Conf., EUSIPCO 2007*, Poznan, Poland, 2007.

[2] L. Alvarez, P.-L. Lions, and J.-M. Morel, "Image selective smoothing and edge detection by nonlinear diffusion. II," *SIAM J. Numer. Anal.*, 29, pp. 845-866, 1992.

[3] J. Astola and L. Yaroslavsky (eds.) *Advances in Signal Transforms: Theory and Applications*, EURASIP Book Series on Signal Processing and Communications, vol. 7, Hindawi Publishing Corporation, 2007.

[4] D. Barash, "A fundamental relationship between bilateral filtering, adaptive smoothing, and the nonlinear diffusion equation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 6, pp. 844-847, June 2002.

[5] R. Brown, *Smoothing, forecasting and prediction of discrete time series*. Prentice-Hall, Englewood Cliffs, NY, USA, 1963.

[6] A. Buades, B. Coll, and J. M. Morel, "A review of image denoising algorithms, with a new one," *SIAM Multiscale Modeling and Simulation*, vol. 4, pp. 490-530, 2005.

[7] A. Buades, B. Coll, and J. M. Morel, "Nonlocal image and movie denoising," *Int. J. Computer Vision*, July 2007.

[8] P. Chatterjee and P. Milanfar, "A Generalization of Non-Local Means via Kernel Regression," *Proc. SPIE Conf. on Computational Imaging*, San Jose, January 2008.

[9] W.S. Cleveland and S.J. Devlin, "Locally weighted regression: an approach to regression analysis by local fitting," *Journal of American Statistical Association*, vol. 83, pp. 596-610, 1988.

[10] R. Coifman and D. Donoho, "Translation-invariant denoising," in *Wavelets and Statistics* (A. Antoniadis and G. Oppenheim (Eds.)), Lecture Notes in Statistics, Springer-Verlag, 125-150, 1995.

[11] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080-2095, August 2007.

[12] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "A Nonlocal and Shape-Adaptive Transform-Domain Collaborative Filtering," in *Proc. 2008 Int. Workshop on Local and Non-Local Approximation in Image Processing, LNLA 2008*, Lausanne, Switzerland, August 2008 (this volume).

[13] D. Donoho and I. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of American Statistical Association*, vol. 90, no. 432, pp. 1200-1224, 1995.

[14] K. Egiazarian, V. Katkovnik, and J. Astola, "Local transform-based image de-noising with adaptive window size selection," *Proc. SPIE Image and Signal Processing for Remote Sensing VI*, vol. 4170, 4170-4, Jan. 2001.

[15] M. Elad, "On the origin of the bilateral filter and ways to improve it," *IEEE Trans on Image Processing*, vol. 10, no. 10, pp. 1141-1151, 2002.

[16] M. Elad, "Why shrinkage is still relevant for redundant representations?" *IEEE Trans. Inf. Theory*, 52, no. 12, pp. 5559-5569, 2006.

[17] M. Elad, "A Weighted Average of Sparse Several Representations is Better than the Sparsest One Alone," presented at *SIAM Imaging Science 2008*, San Diego, CA, USA, July 2008.

[18] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Processing*, vol. 15, no. 12, 2006, pp. 3736-3745.

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

[20] A. Foi, *Anisotropic nonparametric image processing: theory, algorithms and applications,* Ph.D. Thesis, Dip. di Matematica, Politecnico di Milano, ERLTDD-D01290, April 2005.

[21] A. Foi, *Pointwise Shape-Adaptive DCT Image Filtering and Signal-Dependent Noise Estimation*, D.Sc.Tech. Thesis, Institute of Signal Processing, Tampere University of Technology, Publication 710, December 2007.

[22] A. Foi, V. Katkovnik, and K. Egiazarian, "Pointwise Shape-Adaptive DCT for High-Quality Denoising and Deblocking of Grayscale and Color Images," *IEEE Trans. Image Process*., vol. 16, no. 5, pp. 1395-1411, May 2007.

[23] G. Gilboa and S. Osher, "Nonlocal linear image regularization and supervised segmentation," *SIAM Multiscale Modeling and Simulation*, Vol. 6, No. 2, pp. 595-630, 2007.

[24] G. Gilboa and S. Osher, "Nonlocal operators with applications to image processing," *UCLA Computational and Applied Mathematics,* Reports cam (07-23), July 2007, online at `http://www.math.ucla.edu/applied/cam`

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

[26] O. Guleryuz, "Weighted averaging for denosing with overcomplete dictionaries," *IEEE Trans. Image Processing*, vol. 16, no. 12, 2007, pp. 3020-3034.

[27] Y. Hel-Or and D. Shaked, "A Discriminative approach for Wavelet Shrinkage Denoising," *IEEE Trans. Image Process.*, vol 17, no. 4, April 2008.

[28] G. Hua and M. T. Orchard, "A new interpretation of translation invariant image denoising," *Proc. Asilomar Conf. Signals Syst. Comput.*, Pacific Grove, CA, 332- 336, 2003.

[29] V. Katkovnik, *Linear estimation and stochastic optimization problems.* Nauka, Moscow, 1976 (in Russian).

[30] V. Katkovnik, K. Egiazarian, J. Astola, *Local Approximation Techniques in Signal and Image Processing*, SPIE PRESS, Bellingham, Washington, 2006.

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

[32] V. Katkovnik, A. Foi, K. Egiazarian, "Mix-distribution modeling for overcomplete denoising," *Proc. 9th workshop on Adaptation and Learning in Control and Signal Processing (ALCOSP'07)*, St. Petersburg, Russia, August, 29-31, 2007.

[33] C. Kervrann and J. Boulanger, "Local adaptivity to variable smoothness for exemplar-based image regularization and representation," *International Journal of Computer Vision*, vol. 79, pp. 45-69, 2008.

[34] C. Kervrann and J. Boulanger, "Local adaptivity to variable smoothness for exemplar-based image denoising and representation," *Research Report INRIA*, RR-5624, July 2005.

[35] C. Kervrann and J. Boulanger, "Unsupervised Patch-Based Image Regularization and Representation," *ECCV 2006*, Part IV, LNCS 3954, pp. 555-567, 2006.

[36] S. Kindermann, S. Osher, and P.W. Jones, "Deblurring and Denoising of Images by Nonlocal Functionals," *SIAM Multiscale Modeling & Simulation*, vol. 4, no. 4, pp. 1091-1115, 2005.

[37] J.S. Lee, "Digital image smoothing and the sigma filter," Computer Vision, Graphics, and Image Processing, vol. 24, pp. 255-269, 1983.

[38] O. Lepski, Mammen, E. 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.

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

[40] Y. Lou, P. Favaro and S. Soatto, "Nonlocal similarity image filtering," *UCLA Computational and Applied Mathematics,* Reports cam (8-26), April 2008, online at `http://www.math.ucla.edu/applied/cam`

[41] Y. Lou, X. Zhang, S. Osher and A. Bertozzi, "Image Recovery via Nonlocal Operators," Reports cam (8-35), May 2008, online at `http://www.math.ucla.edu/applied/cam`

[42] Y. Meyer, *Oscillating Patterns in Image Processing and Nonlinear Evolution Equations*, Univ. Lecture Ser. 22, AMS, Providence, RI, USA, 2001.

[43] D. Muresan and T. Parks, "Adaptive principal components and image denoising," *Proc. 2003 IEEE Int. Conf. Image Process, ICIP 2003*, pp. 101-104, Sept. 2003.

[44] E.A. Nadaraya, "On estimating regression," *Theory Prob. Appl.,* vol. 9, pp. 141-142, 1964.

[45] R. Öktem, L. Yaroslavsky, K. Egiazarian and J. Astola, *Transform based denoising algorithms: comparative study*, Tampere University of Technology, 1999.

[46] H. Öktem, V. Katkovnik, K. Egiazarian, and J. Astola, "Local adaptive transform based image de-noising with varying window size," *Proc. IEEE Int. Conf. Image Process., ICIP 2001,* Thessaloniki, Greece, 273-276, 2001.

[47] S. Osher, A. Sole, and L. Vese, "Image decomposition and restoration using total variation minimization and the H^{-1}

norm," *Multiscale Model. Simul.*, vol. 1, pp. 349-370, 2003.

[48] P. Perona and J. Malik, Scale space and edge detection using anisotropic diffusion, *IEEE Trans. Patt. Anal. Mach. Intell.*, 12, pp. 629-639, 1990.

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

[50] J. Polzehl and V. Spokoiny, "Image denoising: pointwise adaptive approach," *The Annals of Statistics*, vol. 31, no. 1, pp. 30-57, 2003.

[51] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D,* 60 2, pp. 259-268, 1993.

[52] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear algorithms," *Phys. D*, vol. 60, pp. 259-268, 1992.

[53] A. Savitzky and M. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Analytical Chemistry*, vol. 36, pp. 1627-1639, 1964.

[54] S.M. Smith and J.M. Brady, "SUSAN - a new approach to low level image processing," *International Journal of Computer Vision*, vol. 23, no. 1, pp. 45-78, 1997.

[55] V. Spokoiny, *Local parametric methods in nonparametric estimation*, Springer, to appear 2008.

[56] H. Takeda, S. Farsiu, and P. Milanfar, "Higher Order Bilateral Filters and Their Properties," *Proc. of the SPIE Conf. on Computational Imaging*, San Jose, January 2007.

[57] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," *Proc.of the Sixth Int. Conf. on Computer Vision*, pp. 839-846, 1998.

[58] L. Vese and S. Osher, "Modeling textures with total variation minimization and oscillating patterns in image processing," *J. Sci. Comput.*, vol. 19, pp. 553-572, 2003.

[59] G.S. Watson, "Smooth regression analysis," *Sankhya,* Ser. A, vol. 26, pp. 359-372, 1964.

[60] J. Wei, "Lebesgue anisotropic image denoising," *Int. J. Imaging Systems and Technology*, vol. 15, no. 1, pp. 64-73, 2005.

[61] L. Yaroslavsky and M. Eden, *Fundamentals of Digital Optics*, Birkhäuser Boston, Boston, MA, 1996.

[62] L. Yaroslavsky, "Local adaptive image restoration and enhancement with the use of DFT and DCT in a running window," in *Proc. SPIE Wavelet Applications in Signal and Image Process. IV*, vol. 2825, pp. 1-13, 1996.

[63] L. Yaroslavsky, *Digital picture processing—an introduction*. New York: Springer-Verlag, 1985.

[64] L. Yaroslavsky, K. Egiazarian, and J. Astola, "Transform domain image restoration methods: review, comparison and interpretation," *Proc. SPIE,* vol. 4304 - *Nonlinear Image Process. Pattern Anal. XII*, San Jose, CA, pp. 155-169, 2001.

[65] S. Zimmer, S. Didas, and J. Weickert, "A Rotationally Invariant Block Matching Strategy Improving Image Denoising With Non-Local Means," in *Proc. 2008 Int. Workshop on Local and Non-Local Approximation in Image Processing, LNLA 2008*, Lausanne, Switzerland, August 2008 (this volume).