Adaptive weights smoothing with applications to image restoration

Jörg Polzehl [1] **and Vladimir G. Spokoiny**

Weierstrass Institute for Applied Analysis and Stochastics, Berlin, Germany

Summary: We propose a new method of nonparametric estimation which is based on locally constant smoothing with an adaptive choice of weights for every pair of data-points. Some theoretical properties of the procedure are investigated. Then we demonstrate the performance of the method on some simulated univariate and bivariate examples and compare it with other nonparametric methods. Finally we discuss applications of this procedure to magnetic resonance and satellite imaging.

Keywords: adaptive smoothing; image restoration; Change point and edge estimation; magnetic resonance imaging, satellite imaging.

1 Introduction

In this paper we introduce a new locally adaptive method for two and three dimensional image processing, i.e. image denoising and image enhancement. This method can be applied if the underlying structure can be well approximated by a local constant function. Such images meet in several fields, e.g. from satellite imaging, x-rays, ultrasound or magnetic resonance imaging. Usually these images will suffer from distortions, leading to the problem of recovering the underlying structure of the image. Often interesting structures correspond to discontinuities in the image, i.e. procedures used in this context should both reduce distortions as well as preserve discontinuities. Classical nonparametric regression procedures are based on smoothness assumptions about the underlying function which are not fulfilled in the neighborhood of discontinuities. This leads to so called oversmoothing of the function in such regions. In univariate situations several proposals exist how to overcome this problem, see e.g. Müller (1992), Speckman (1994), Wu and Chu (1993) for procedures based on change point detection, or Banerjee and Rosenfeld (1993) for maximum a posteriori estimation.

The generalization of this idea to the multidimensional case leads to the edge estimation problem. This problem is studied in details in Korostelev and Tsybakov (1993) where the optimal rate of edge estimation is established for the case of an image with the structure of a

[1] *Address for correspondence:* Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, 10117 Berlin, Germany. E-mail: polzehl@wias-berlin.de

boundary fragment. The reader can find further references there. Unfortunately the proposed procedures are based on some quite restrictive assumptions like the structure of a boundary fragment. Another inconvenience is that the methods and results apply only to the case of a random or jittered design which rarely meets in practice.

A large class of methods commonly used in this context is based on Markov random field models (MRF), see e.g. Besag (1986), Geman (1990) or Winkler (1995). A concurrent approach, used especially in image segmentation, is based on recursive partitioning and merging procedures. This class contains CART, see Breiman et al. (1984), or the region based segmentation method of Bose and O'Sullivan (1997). Wu (1993) discussed similar ideas based on testing homogeneity for subimages.

There exist other methods which estimate the image directly without estimating edges but which still pay special attention to the quality of estimation near edges. We mention modal regression, see e.g. Scott (1992), the nonlinear Gaussian filter, see Godtliebsen et al. (1997), the M-smoother of Chu et al. (1998) and different proposals based on wavelets, see e.g. Nason and Silverman (1994), Engel (1994) or Donoho (1997) and references there. One more approach in this direction was proposed recently in Polzehl and Spokoiny (1998). The method can be viewed as a multidimensional analog of the procedure from Spokoiny (1998) assigned to estimation of a univariate function allowing jumps or jumps of derivatives. The idea is to estimate the regression function separately at each design point using a locally constant (or locally polynomial) modeling with an adaptive choice of a neighborhood (a window) from a large class of neighborhoods in which the applied model fits well the data. An inconvenience of this approach is that the class of considered windows has to be really large to get a reasonable quality of estimation. This makes the procedure difficult to realize and computationally intensive.

In this paper we modify this idea. Namely we do not specify the class of considered windows but we determine in a data-driven way the form of the neighborhood around the point of interest x in which the function f can be well approximated by a constant. A similar idea was discussed in Tsybakov (1989) but the proposed method uses essentially some prior information about the structure of the image and about the image values within each region.

Our method, which in the sequel will be referred to as adaptive weights smoothing (AWS), is fully adaptive, that is, no prior information is required. It is important to remark that the method does not depend on the dimensionality of the image and can be applied to smooth three and even higher dimensional images as well.

The AWS procedure is assigned for image estimation and can in general be applied to an arbitrary image. However, successful applications of the proposed method can be expected in situations when the image contains large homogeneous regions, not necessary connected and

of may be complicated shape.

The further discussion and the precise description of the procedure are placed in Section 2. In Section 3 we study some theoretical properties of our method. Section 4 provides a simulative comparison with several alternative procedures for univariate and bivariate situations. Finally Section 5 describes an application of our method to Magnetic Resonance Imaging and Satellite Imaging.

1.1 Model

The model can be described as

$$Y_i = f(X_i) + \varepsilon_i \quad X_i \in I\!R^d, \quad \boldsymbol{E}\varepsilon_i = 0, \quad \mathrm{Var}\,\varepsilon_i = \sigma^2. \tag{1}$$

Here $X_1, \ldots, X_n$ are design points which are usually assumed to be equispaced in the unit cube $[0,1]^d$. $\boldsymbol{E}$ and Var denote expectation and variance, respectively. At each point X_i we observe the regression function $f(X_i)$ with some additive error ε_i. We suppose the errors ε_i to be independent identically distributed zero mean random variables with unknown distribution.

The regression function f is supposed piecewise constant. This means that the unit cube $[0,1]^d$ can be split into disjoint regions $A_1, \ldots, A_M$ and

$$f(x) = \sum_{m=1}^{M} a_m \mathbf{1}(x \in A_m) \tag{2}$$

where $a_1, \ldots, a_M$ are some numbers, e.g. gray levels in an image, and $\mathbf{1}$ stands for the indicator function. Obviously the function f is constant within each region A_m. The regions A_m, the values a_m and even the total number of regions M are unknown. Clearly such an assumption on the underlying structure is valid for an arbitrary image, since each region A_m may consist of one point. However, an application of the procedure proposed below seems to be reasonable for situations where the underlying image really contains large homogeneous regions. When studying theoretical properties of the method proposed we impose some additional assumptions on the size of these regions.

2 Adaptive weights smoothing

In this section we present our estimation procedure.We start with some heuristic explanation.

2.1 Preliminaries

The problem of estimating the function f of the form (2) can be treated as follows: to recover the values $a_1, \ldots, a_M$ and to decide for each point X_i in which region A_m it is.

To explain the idea of the method, we imagine for a moment that the regions $A_1, \ldots, A_M$ are known and only the values a_m are to be estimated. This leads to obvious estimates

$$\widehat{a}_m = \frac{1}{N_{A_m}} \sum_{X_i \in A_m} Y_i$$

where N_{A_m} denotes the number of design points in A_m, $m = 1, \ldots, M$. Then we simply set $\widehat{f}(X_i)$ equal to the mean $\widehat{a}_m$ of Y_j's over the region A_m containing X_i. Therefore, given a partition $A_1, \ldots, A_M$, we can easily estimate the underlying function f.

Next we consider the inverse situation when the partition $A_1, \ldots, A_M$ is unknown but we are given a pilot estimate $\widehat{f}_0$ of the regression function f. It is natural to use this estimate to recover for every point X_i the corresponding region A_m. Namely, for each pair of points X_i and X_j, we know, if the value $|\widehat{f}_0(X_i) - \widehat{f}_0(X_j)|$ is large compared with its standard deviation then these two points are almost definitely in different regions. We therefore, for every design point X_i, estimate the region A_m containing X_i by

$$\widehat{A}(X_i) = \{X_j : |\widehat{f}_0(X_i) - \widehat{f}_0(X_j)| \le \lambda \widehat{\sigma}_0(X_i)\}$$

where $\widehat{\sigma}_0(X_i)$ is the standard deviation of $\widehat{f}_0(X_i)$ and λ is some number.

Using these estimated regions, we define a new estimate $\widehat{f}_1$ by

$$\widehat{f}_1(X_i) = \frac{\sum\limits_{X_j \in \widehat{A}(X_i)} Y_j}{N_{\widehat{A}(X_i)}} = \frac{\sum\limits_j w_1(X_i, X_j) Y_j}{\sum\limits_j w_1(X_i, X_j)}$$

with

$$w_1(X_i, X_j) = \mathbf{1}\left(|\widehat{f}_0(X_i) - \widehat{f}_0(X_j)| \le \lambda \widehat{\sigma}_0(X_i)\right). \tag{3}$$

We can repeat this calculation using $\widehat{f}_1$ in place of $\widehat{f}_0$ and so on.

Our adaptive procedure mostly realizes this idea with two modifications. First of all, at each iteration k, we restrict the estimated region $\widehat{A}(X_i)$ to some local neighborhood $U_k(X_i)$ of the point X_i such that the size of $U_k(X_i)$ grows with k. Secondly we use continuous weights $w_k(X_i, X_j)$ instead of zero-one weights in (3).

Finally, to stabilize the procedure, we also add a control step, comparing the new estimate with the estimates from preceding iterations.

Now we present a formal explanation of the method starting with a description of the input parameters of the algorithm. Our recommendations for a default choice of these parameters and for a data-driven selection are given in Section 3.3 and 3.4.

2.2 Parameters of the procedure

The most important element of the procedure is an increasing sequence of neighborhoods around each design point.

For each design point x, we assume to be given a sequence of neighborhoods $U_k(x)$, $k = 0, 1, \ldots, \infty$ with $U_k(x) \subset U_{k+1}(x)$ containing x. One reasonable choice of these neighborhoods $U_k(x)$ is $U_k(x) = \{X_i : |X_i - x| \leq d_k\}$ with d_k being a sequence of increasing radii. Another possibility is to define $U_k(x)$ as the set of the N_k nearest neighbors of x, where N_k is an increasing sequence of integers. In the sequel, $N_k(x)$ denotes the number of design points X_i in $U_k(x)$,

$$N_k(x) = \#\{X_i \in U_k(x)\}.$$

Parameter k^* denotes the maximal index of neighborhoods used.

The procedure involves numerical parameters λ and η which are used as critical values for tests entering in the adaptation and the control steps.The role of these parameters and recommendations for their choice are discussed in Section 3.3 and 3.4.

We fix a univariate kernel K satisfying usual conditions: it is a symmetric smooth function with the maximum at zero and nonincreasing on the positive semiaxis. We assume the kernel to be integrable, i.e. $\int_0^\infty K(x)dx < \infty$.

In most applications the noise variance σ^2 is unknown and an estimate $\widehat{\sigma}^2$ can be obtained from the data, see e.g. Gasser et al. (1986) or Wu and Chu (1993) for different proposals. A general form of such estimates is

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{e}_i^2$$

where *pseudo-residuals* $\widehat{e}_i$ are defined on the base of the difference $Y_i - \widehat{f}(X_i)$ with a local regression estimate $\widehat{f}(X_i)$. *Pseudo-residuals* can also be defined using differences of observations. In the univariate case one can use either $\widehat{e}_i = \sqrt{2}(Y_i - Y_{i-1})$ or $\widehat{e}_i = \sqrt{6}\,(-Y_{i-1} + 2Y_i - Y_{i+1})$ and in the two-dimensional case

$$\begin{aligned} \widehat{e}_{i_1,i_2} &= \sqrt{6}\,\{2Y_{i_1,i_2} - Y_{i_1+1,i_2} - Y_{i_1,i_2+1}\} \qquad \text{or} \\ \widehat{e}_{i_1,i_2} &= \sqrt{20}\,\{4Y_{i_1,i_2} - (Y_{i_1+1,i_2} + Y_{i_1-1,i_2} + Y_{i_1,i_2+1} + Y_{i_1,i_2-1})\}\,. \end{aligned}$$

In case of a complicated underlying image structure, an estimate based on the inter-quartile-range $\widehat{\sigma} = (t_{75\%} - t_{25\%})/1.35$, with $t_{25\%}$ and $t_{75\%}$ being the .25- and .75-quantile of the empirical distribution of the *pseudo-residuals*, is preferable.

For some further discussion concerning the variance estimation, see Section 3.4.

2.3 The procedure

We begin with an initialization.

Initialization: For each point X_i, we calculate initial estimates of $f(X_i)$ and $\mathrm{Var}\,\widehat{f}(X_i)$ as

$$\begin{aligned}\widehat{f}_0(X_i) &= \frac{1}{N_0(X_i)} \sum_{X_j \in U_0(X_i)} Y_j \\ \widehat{s}_0^2(X_i) &= \frac{\widehat{\sigma}^2}{N_0(X_i)}\end{aligned}$$

and set $k = 1$. Here $\widehat{\sigma}^2$ is the variance estimate defined previously.

Adaptation: Compute weights $w_k(X_i, X_j)$ as

$$w_k(X_i, X_j) = K\left(\frac{\widehat{f}_{k-1}(X_i) - \widehat{f}_{k-1}(X_j)}{\lambda \widehat{s}_{k-1}(X_i)}\right) \tag{4}$$

for all points X_j in $U_k(X_i)$ and compute new estimates of $f_k(X_i)$ and $\mathrm{Var}\,\widehat{f}_k(X_i)$ as

$$\widehat{f}_k(X_i) = \frac{\sum\limits_{X_j \in U_k(X_i)} w_k(X_i, X_j) Y_j}{\sum\limits_{X_j \in U_k(X_i)} w_k(X_i, X_j)}, \tag{5}$$

$$\widehat{s}_k^2(X_i) = \frac{\widehat{\sigma}^2 \sum\limits_{X_j \in U_k(X_i)} w_k^2(X_i, X_j)}{\left(\sum\limits_{X_j \in U_k(X_i)} w_k(X_i, X_j)\right)^2} \tag{6}$$

for all X_i.

Control: After the estimate $\widehat{f}_k(X_i)$ is computed, we compare it with the previous estimates $\widehat{f}_{k'}(X_i)$ at the same point X_i for all $k' < k$. If there is at least one index $k' < k$ such that

$$\left|\widehat{f}_k(X_i) - \widehat{f}_{k'}(X_i)\right| > \eta \widehat{s}_{k'}(X_i)$$

then we do not accept $\widehat{f}_k(X_i)$ and keep the estimates $\widehat{f}_{k-1}(X_i)$ from the preceding iteration. This means that in such a situation we replace $\widehat{f}_k(X_i)$ and $\widehat{s}_k(X_i)$ by $\widehat{f}_{k-1}(X_i)$ and $\widehat{s}_{k-1}(X_i)$, respectively. It is worth mentioning that this control step alone can be used to construct an adaptive estimate, see Lepski, Mammen and Spokoiny (1997) or Lepski and Spokoiny (1997).

Stopping: Stop if $k = k^*$ or if $\widehat{f}_k(X_i) = \widehat{f}_{k-1}(X_i)$ for all i, otherwise increase k by 1 and continue with the adaptation step.

3 Properties and computational details

Because of the iterative and complex nature of the algorithm theoretical properties are extremely difficult to obtain in a general situation. We consider two specific cases which are of the most interest. The first situation corresponds to estimation inside a large homogeneous region and the second one to estimation near an edge.

For simplicity we assume homogeneous Gaussian noise with known variance σ^2. We also consider the uniform kernel $K(x) = \mathbf{1}(|x| \le 1)$. All properties can be easily extended to the case of a continuous kernel K.

3.1 Estimation inside a homogeneous region

We study an idealized situation where the underlying image function is constant, $f(x) \equiv a$. For simplicity we also assume that each neighborhood $U_k(X_i)$ contains exactly N_k design points where N_k is a prescribed increasing sequence. We aim to show that in this situation our estimate is, with a very large probability, also a constant and the deviations $\widehat{f}(X_i) - a$ are of order $n^{-1/2}$.

In the next statement we need an estimate for the sum $N_1 + \ldots + N_{k^*}$. Since the sequence N_k typically grows exponentially, this sum is of order N_{k^*}. Also we assume that $N_{k^*} = n$, that is, we stop when the largest possible neighborhood is taken. This leads to the bound

$$N_1 + \ldots + N_{k^*} \le Cn \tag{7}$$

with some $C > 0$.

Proposition 3.1 *Let $f(x) \equiv a$ and $\lambda^2 \ge (2+\delta)\log(n)$ with some $\delta > 0$. Then for all $k \le k^*$ and all pairs X_i and $X_j \in U$*

$$\boldsymbol{P}\left(w_k(X_i, X_j) = 0 \text{ for some } k \le k^* \text{ and } i \ne j\right) \le \gamma_{k^*}$$

with

$$\gamma_{k^*} = \exp\{-\lambda^2/4\} n^2 C/2 + \exp\{-\eta^2/2\} n k^*(k^*+1)/2$$

and C is from (7).

We defer the proof of this and the next proposition to the appendix.

The quantity γ_{k^*} is small provided that $\lambda^2 \ge (8+\delta)\log n$ and $\eta^2 \ge (2+\delta)\log n$ with some constant $\delta > 0$. Then with a probability of at least $1 - \gamma_{k^*}$ all estimates $\widehat{f}_{k^*}(X_i)$ coincide with the mean values of all observations Y_j.

3.2 The case of many regions

Now we discuss the situation when there are more than one regions. To simplify the presentation, we suppose that there are only two large regions A_1 and A_2 in the image and hence the function f has only two values a_1 and a_2. The result allows straightforward generalization to the case of many regions.

By $\alpha = |a_1 - a_2|$ we denote the image contrast. We also denote by A_m° the set of points X_i in each region A_m for which the initial neighborhood $U_0(X_i)$ belongs completely to A_m,

$$A_m^\circ = \{X_i : U_0(X_i) \subset A_m\}, \qquad m = 1, 2.$$

We intend to show that if the image contrast is sufficiently large compared to the noise level then we typically get $w_k(X_{i_1}, X_{i_2}) = 0$ for all pairs (X_{i_1}, X_{i_2}) with $X_{i_1} \in A_1^\circ$ and $X_{i_2} \in A_2^\circ$ and for all $k \geq 1$.

Proposition 3.2 *Let* $f(x) = a_1 \mathbf{1}(x \in A_1) + a_2 \mathbf{1}(x \in A_2)$. *Then it holds*

$$w_k(X_i, X_j) = 0, \qquad \forall X_i \in A_1^\circ,\ X_j \in A_2^\circ, \quad \textit{and } \forall k \leq k^*,$$

with a probability greater or equal to

$$1 - 0.5Cn^2 \exp\left\{ - \left(\sigma^{-1} N_0^{1/2} |a_1 - a_2| - 2\eta \right)^2 / 4 \right\}$$

where C *is from (7) and* N_0 *is the number of points in the initial neighborhood.*

We know from Proposition 3.1 that a proper choice of η^2 is $(2 + \delta)\log(n)$. Therefore, if

$$\sigma^{-1} N_0^{1/2} |a_1 - a_2| > 4\eta$$

the probability of $w_k(X_i, X_j) = 0$ for any two points X_i, X_j from the same region can be bounded by $n^2 \exp\{-\eta^2\}$. If n sufficiently large, this probability is again very small.

The results of Propositions 3.1 and 3.2 lead to the following conclusion. Let a point X_i lie inside a large region and let for all $j \in U_{k+1}(X_i)$ the neighborhood $U_k(X_j)$ belong to the same region. Then according to Proposition 3.1 all weights $w_{k+1}(X_i, X_j)$ are 1 and hence the estimate $\widetilde{f}_{k+1}(X_i)$ is very close to the mean of observations over $U_{k+1}(X_i)$. Such an estimate is unbiased and its variance is of order $\sigma^2 / N_{k+1}(X_i)$. Moreover, the control step guarantees that further iterations do not lead to an essential decrease of the accuracy of estimation. Therefore, inside every large region, the estimate should perform quite well.

At the same time, for points near an edge, the probability to assign a weight $w(X_i, X_j)$ of 1 for two points X_i, X_j from different regions could be quite high. This leads to a larger bias in estimating the image function especially when the image contrast is small compared to the standard deviation of the errors, see Proposition 3.2. It can be also shown that the procedure delivers the rate optimal quality of edge recognition in the sense discussed in Polzehl and Spokoiny (1998). This issue is also in agreement with simulation results, see the next sections.

For weights equal 0 or 1 and for a particular iteration k the estimated regions of homogeneity $\widehat{A}(X_i)$ are restricted to a local neighborhood $U_k(X_i)$. Therefore these regions strongly depend

on the point X_i and do not yield a segmentation of the data domain. But, if the noise is small compared to the image contrast, then due to Proposition 3.1 and 3.2, for a point X_i lying in a region A_m, we get with a high probability at k-th iteration $\widehat{A}_k(X_i) = A_m \cap U_k(X_i)$. In such case, for sufficiently large k^*, we have $A_m \cap U_{k^*}(X_i) = A_m$. Hence $\widehat{A}_k(X_i) = A_m$ for all X_i in A_m. For a large noise, these arguments do not apply. Both issues are again in agreement with our simulation results, see Section 4.

3.3 Computational issues

Now we discuss how the parameters of the procedure can be selected and indicate one possible default choice used in our simulations. Although this choice involves some arbitrariness we observe that moderate changes of the parameters lead to essentially similar results. A way for a data-driven parameter choice, used in our examples, is also presented.

Size of U_0: The size $N_0(X_i)$ is important in the context of image recognition and edge estimation, see Proposition 3.2. For the cases with contrast-to-noise ratio $\alpha/\sigma > 2$, the choice $N_0 = 1$ can be advised. Here α is the (minimal) image contrast,

$$\alpha = \min\{|a_m - a_{m'}|,\, m \neq m',\, a_m \neq a_{m'}\}.$$

For smaller contrast-to-noise ratio $N_0 = 5$ or $N_0 = 9$ may be desirable.

Sequence of neighborhoods U_k: The sequence should satisfy the conditions $X_i \in U_0(X_i)$ and $U_{k-1}(X_i) \subset U_k(X_i)$. It can be recommended to select sequences $U_k(X_i)$ in a way that the numbers $N_k(X_i)$ of points in every such neighborhood grow exponentially with k.

In our simulation study and all examples we use neighborhoods $U_k(x) = \{X_i : |X_i - x| \leq d_k\}$ with $d_k \in \{0:8, 2\cdot(5:12), 4\cdot(7:12), 8\cdot(7:12), 16\cdot(7:10), 32\cdot(6:8)\}$, $k^* = 35$ for univariate situations, and $d_k \in \{(0:8)/2, 4.4, 5:10, 2\cdot(6:10)\}$, $k^* = 19$ for images ($(a:b)$ denotes a sequence of integers from a to b). This choice gives $N_{k^*} = 513$ and $N_{k^*} = 1257$ points in the largest neighborhood for univariate and bivariate situations, respectively.

k^*: The value of k^*, and therefore N_{k^*}, is mainly determined by the degree of locality that one wishes to maintain and the computational effort one is able to spend. Increasing k^* allows for additional variance reduction in large homogeneous regions but usually does not change the estimates where local structure is present. A data-driven choice of k^* is discussed in Section 3.4.

K: Our default choice for the kernel is $K(x) = \exp(-x^2)$.

λ: The choice of this parameter mostly determines the properties of the procedure. Increasing the parameter reduces the probability of detection of artificial jumps in a homogeneous situation

(error of first kind) and increases the probability not to detect an existing discontinuity (error of second kind), see Propositions 3.1 and 3.2 . Our default choice, for the above K, is $\lambda = 3$. A data-driven choice of λ is discussed in Section 3.4.

$\boldsymbol{\eta}$: The control step prevents the algorithm from loosing previously detected discontinuities, see Proposition 3.2. Suitable values for η are between 3 and 4. We use $\eta = 4$ in all cases.

Remark: There is no magic behind the recommended choice $\lambda \approx 3$ and $\eta \approx 4$. We illustrate this on the simplest situation corresponding to the first step of the algorithm assuming $U_0(X_i) = \{X_i\}$ for all i. Then the initial estimates $\widetilde{f}_0(X_i)$ coincide with the observations Y_i. Therefore, if points X_i and X_j belong to the same region A_m, then the difference $\widetilde{f}_0(X_j) - \widetilde{f}_0(X_i)$ coincides with the differences $\varepsilon_i - \varepsilon_j$ of the corresponding stochastic errors, see (1). If these errors are normally $\mathcal{N}(0, \sigma^2)$-distributed and independent, then the difference is also normal $\mathcal{N}(0, 2\sigma^2)$. Calculating the weight $w_1(X_i, X_j)$ we compare this difference with $\lambda\sigma_0(X_i) = \lambda\sigma$. The parameter λ is chosen to provide an essentially large probability of the event $\{|\varepsilon_i - \varepsilon_j| \le \lambda\sigma\}$. The value $\lambda = 3$ corresponds to the probability $2\Phi(\sqrt{9/2}) - 1 \approx 0.966$, Φ being the standard normal CDF (note that similar arguments hold for further iterations assuming that the neighborhoods $U_k(X_i)$ and $U_k(X_j)$ are still inside the region A_m). Of course larger values of λ lead to even larger probability of such an event. But, when λ increases, the quality of estimation near an edge decreases. The choice $\lambda = 3$ provides a reasonable compromise for most cases. However, we keep a possibility to tune this parameter in some specific situations depending on what is important in each particular case. In many cases (especially for a large contrast-to-noise ratio) the choice of λ between 2.8 and 4 does not change the result of the procedure significantly. If the noise is comparable with the image contrast this choice becomes more crucial: increasing λ decreases the probability to detect a discontinuity and therefore results in oversmoothing while decreasing λ may lead to a random segmentation of homogeneous regions. A Bootstrap-based choice of the parameters λ and k^* is discussed in Section 3.4.

The iterative algorithm introduced in Section 2 is computationally intensive but still feasible. The number of operations necessary to process an image containing n pixel is of order $O(nN_{k^*})$ if the sequence N_{k^*} is exponentially growing. We illustrate the speed of the algorithm giving the CPU-Time reported for the MR-images analyzed in Section 5.1. We implemented AWS using Fortran for the time critical parts of the algorithm and Splus as an user interface. On a 255 Mhz DEC-Alpha Workstation the CPU-Time (User) taken by our implementation using the default parameter settings is 87 s for an image of 256×256 pixel and 383 s for an image of 512×512 pixel. For many applications this can be reduced significantly by using a

smaller value of k^*.

3.4 Bootstrap-based choice of the parameters of the procedure

The performance of the proposed procedure strongly depends on the choice of the involved parameters, especially on λ and k^*. Our simulated results and applications to real data show that there is no one universal optimal choice for all situations, and the quality of the procedure can be improved by tuning these parameters.

Another important point is connected to the quality of variance estimation, see Section 2.2. It turns out that the proposed variance estimator overestimates the true variance in case of a complicated underlying structure, e.g. in MRI applications. The use of $\widehat{\sigma}^2$ in place of σ^2 is clearly equivalent to replacing λ by $\lambda\widehat{\sigma}/\sigma$ which leads again to the question of an optimal choice of parameter λ for each particular example. Below we discuss one possibility for a data-driven choice of these parameters based on the resampling (bootstrap) idea.

The underlying idea is that the proper choice is connected to the complexity of the image and this complexity is recovered by our estimate with the default choice of these parameters. Then we can resample the data using the estimate as a reference image and select a proper set of parameters for this known reference. Finally we apply this bootstrap-based choice to the original data. This procedure may be iterated by repeating the resampling step with the new estimate. The procedure reads as follows:

Run with default parameter set. AWS is used with the default parameters $\lambda = 3$, $\eta = 4$ and $k^* = 19$. This provides us with an estimate of the image $\widehat{f}(X_i)$ and two sums

$$W_1(X_i) = \sum_j w(X_i, X_j), \quad \text{and} \quad W_2(X_i) = \sum_j w^2(X_i, X_j),$$

based on the weights $w(X_i, X_j)$ used at the last iteration of AWS, all of this for every $i = 1, \ldots, n$. Clearly $W_1(X_i) = W_2(X_i)$ for the case of zero-one weights.

Variance estimation on the base of $\widehat{f}$. Next we recalculate the noise variance using the estimate $\widehat{f}$. The representation $\widehat{f}(X_i) = \frac{1}{W_1(X_i)} \sum_j w(X_i, X_j) Y_j$ with $w(X_i, X_i) = 1$ leads to the variance estimate

$$(\sigma^*)^2 = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \widehat{f}(X_i)\right)^2 \frac{W_1^2(X_i)}{W_2(X_i) + W_1(X_i)^2 - 2W_1(X_i)}.$$

Resampling. We draw new bootstrap samples $Y^*_{i,m}$ using the model $Y^*_{i,m} = \widehat{f}(X_i) + \sigma^* \varepsilon^*_{i,m}$ where $\varepsilon^*_{i,m}$ are independent standard normal errors. Here m denotes the number of the bootstrap sample, $m = 1, \ldots, M$.

Parameter optimization. For every considered set of parameters λ, k^*, and for every bootstrap sample $Y^*_{1,m}, \ldots, Y^*_{n,m}$, we carry over the AWS procedure resulting in the image estimate f^*_m and compute the quality criterion

$$\frac{1}{M}\sum_{m=1}^{M}\sum_{i=1}^{n}\psi\left(\widehat{f}(X_i) - f^*_m(X_i)\right)$$

where ψ is some loss function, e.g. a quantile function, or $\psi(t) = |t|$ or $\psi(t) = t^2$. Parameters are chosen minimizing this criterion w.r.t. λ and k^*.

Since our criterion is defined by summation over all design point, one can expect a degenerated behavior of the optimized quantity even for one bootstrap sample, that is, for $M = 1$.

Final run. Finally apply AWS to the original data using the selected set of parameters.

4 Simulations

In the following subsections we demonstrate the capabilities of our approach using some uni- and bivariate simulations. We illustrate the behavior of our algorithm for different contrast-to-noise ratios ranging from easy to handle situations ($\alpha/\sigma = 4$) to situations where the signal is hardly visible by eye ($\alpha/\sigma = 1$ and larger) and different size of the homogeneous regions.

We compare our AWS procedure with some established alternative approaches. It should be mentioned that the following list is far from being complete.

4.1 Alternative procedures

Gauss filtering: Here we use an Nadaraya-Watson kernel estimate with Gaussian kernel and smoothing parameter h

$$\widehat{f}(X_i) = \frac{\sum_j Y_j \exp\{-(X_j - X_i)^2/(2h^2)\}}{\sum_j \exp\{-(X_j - X_i)^2/(2h^2)\}}.$$

Nonlinear Gauss filtering: The Nonlinear Gauss Filter was proposed by Godtliebsen et al. (1997) as a generalization of the Sigma Filter of Lee (1983). It replaces the discontinuous (uniform) weight function of the sigma filter by an Gaussian weight scheme. The filter is defined as

$$\widehat{f}(X_i) = \frac{\sum_{j\in U(X_i)} Y_j \exp\{-(Y_j - Y_i)^2/(2g^2)\}}{\sum_{j\in U(X_i)} \exp\{-(Y_j - Y_i)^2/(2g^2)\}}$$

where the radius of $U(x)$ and g are smoothing parameters.

Modal regression: Modal regression is introduced in Scott (1992) as a robust alternative to nonparametric regression procedures estimating a conditional mean. The modal regression curve is defined as

$$\widehat{f}(x) = \arg\max_y \widehat{p}(y, x)$$

with $\widehat{p}(y,x)$ being an estimate of the joint density of y and x. Although Scott proposes to use multiple modes simultaneously we concentrate on the mode closest to the observed Y. The estimate depends on two bandwidths in x and y domain.

Change point methods: An alternative in case of well separated jumps can be based on methods of change point estimation. We use the procedures of change point estimation proposed by Müller (1992). The change point estimate is only used in the univariate case.

CART: A suitable procedure for the univariate case can be based on the classification- and regression trees (CART) introduced by Breiman et al. (1984). We use CART as implemented in Splus with the number of splits determined by CV-pruning. CART is only used in the univariate simulations since it is not flexible enough to allow for a reasonable reconstruction of our test image.

Wavelets: For our comparisons we use the Wavelet package *wavethresh* of G. P. Nason, see Nason and Silverman (1994) for a description of the software. We use the Haar basis and the biorthogonal Haar basis for univariate and bivariate situations, respectively. We suppose this choice to be the most adequate for local constant functions from the selection of bases offered. Parameters, i.e. threshold value and levels for thresholding, are selected to provide optimal mean integrated squared error (MISE) for the underlying true structure. We used hard or soft thresholding depending on which method provided better results in terms of MISE. We do not discuss more sophisticated wavelet procedures like the translation invariant wavelet transform, see Coifman and Donoho (1995), or anisotropic wavelet bases, see Daubechies (1992), Chapter 10.1, or Neumann (1998).

Markov Random Fields (MRF): Out of the wide range of Markov Random Fields methods in image analysis we use an Metropolis algorithm, see e.g. Winkler (1995), page 133. We start with initial values $\widehat{f}(X_i) = Y_i$. A new proposal y^* in a randomly chosen point x is generated from a strictly positive probability distribution $G(y^*|x)$ with support on the range of Y. A new proposal y^* in point x is accepted as a new value for $\widehat{f}(x)$ with probability

$$\min(1, \exp\left(H(\widehat{f}(x)|x,Y) - H(y^*|x,Y)/\tau\right)),$$

otherwise the old value is kept. Here

$$H(y^*|X_i,Y) = \frac{(Y_i - \widehat{f}(X_i))^2}{2\sigma^2} \; - \sum_{X_j \in U(X_i)/\{X_i\}} \frac{\beta}{1 + ((\widehat{f}(X_i) - \widehat{f}(X_j))/\delta)^2}$$

is an energy function designed for a continuous state space. The temperature τ is chosen to slowly decrease with the number of iterations. For more discussions see e.g. Winkler (1995) Chapter 10 for the Metropolis algorithm and Chapter 2 for the energy function used.

Except the Gauss filter all of the procedures considered are designed to handle discontinuities. Most of the alternative procedures depend on smoothing parameters. These parameters are chosen to minimize an estimate of MISE in the situation studied.

4.2 A univariate simulation example

In our first univariate example we use a piecewise constant regression function with varying size of the homogeneous region. The left column of Figure 1 presents three data sets generated for different values of σ. The central plots show the true function together with the AWS estimates. The third column provides estimates obtained by the best alternative procedures, wavelets and CART, for a comparison.

********** put Figure 1 around here **********

********** put Table 1 around here **********

We run 1000 simulations with sample size 256 and error standard deviation $\sigma = .25$, .5 and 1, respectively. Table 1 displays results of the simulations in terms of estimated MISE and mean, over x, estimates of $\mathbf{P}(|\widehat{\mathbf{f}}(\mathbf{x}) - \mathbf{f}(\mathbf{x})| > \alpha/\mathbf{4})$, $\alpha = 1$. We call this quantity large deviation probability (LDP). Note that AWS performs best with respect to both MSE and LDP in case of $\sigma \leq .5$. For $\sigma = 1$ our adaptive procedure does not always detect the discontinuities for small x, i.e. where the homogeneous regions are small. This leads to an increased mean squared error and large deviation probability for small x.

4.3 Bivariate simulations

We use an artificial image to demonstrate the power of our procedures in more complicated situations, see Figure 2.

********** put Figure 2 around here **********

The image possesses two different image contrasts, $\alpha = .5$ and 2 $\alpha = 1$, and homogeneous regions of various size and form. The image contains $n = 256 \times 256$ pixel. Note that in the image the size of homogeneous regions increases from the lower left to the upper right. There are very detailed structures in the upper left and lower right of the image.

********** put Figure 3 around here **********

In the left column of Figure 3 we display this image distorted by additive Gaussian noise of standard deviation $\sigma = .25$, .5, and 1, respectively. The second column contains the reconstruction of the noisy images by AWS (with default parameters). For a comparison we provide the results obtained by modal regression, wavelets and Markov random fields. For the alternative procedures estimated gray levels, that are out of scale are projected.

With standard deviation increasing we first loose the most detailed structure ($\sigma = .5$) and

observe some loss in edge accuracy for the lower contrast level ($\sigma = 1$). Note that we still recover the main structure that is hardly visible in the noisy original. The AWS-estimates behave very stable with respect to changes of the parameters, e.g. each $\lambda \in (3, 3.6)$ gives essentially the same quality of reconstruction for all σ considered.

To illustrate the local behavior of our procedure in more detail we conduct a comparative simulation study based on the test image, see Figure 5. We perform 100 simulations with error standard deviation of $\sigma = .25$, .5 and 1, respectively. We use our default parameter settings for AWS and again approximately MISE-optimal parameters for the alternative procedures.

Table 2 provides the simulation results using the same criteria as in the univariate case.

********** put Table 2 around here **********

For the lowest noise level we observe that even the detailed structures in the upper left and lower right of the image are recovered by our method. MRF and modal regression both work reasonable in this situation, providing an improvement to the gauss filter with respect to both criteria. The wavelet estimate suffers especially for detailed structures and where the boundaries are not parallel to the axes. Increasing σ we see a clear advantage of our procedure.

5 Applications

5.1 An Application to Magnetic Resonance Imaging

Magnetic Resonance imaging (MRI) is a new technique of noninvasive analysis providing a delineation of a physical object. The signal, or true image, can be interpreted as a weighted spin density of the system of atomic nuclei the physical object consists of. For an excellent introduction into the mathematics and physics of MRI see for instance Sebastiani and Barone (1991) and Lange (1996).

In Fourier imaging, which is the most common MR imaging technique, a finite number of coefficients from the 2-D Fourier series expansion of the true image are measured. The MR image is then obtained applying the discrete Fourier transform to the raw data, i.e. the MR image can be viewed as a truncated Fourier series of the weighted spin density distorted by noise, see e.g. Barone and Sebastiani (1992).

It is reasonable to characterize the underlying image by a piecewise constant function, with homogeneous regions corresponding to the same type of tissue and therefore having a similar spin density and discontinuities at the interface between adjacent tissues. Random errors can be modeled as additive white Gaussian noise, see e.g. Sebastiani and Barone (1995).

********** put Figure 4 around here **********

Our first example is based on a MR image recorded at the MR center at Trondheim, kindly provided to us by F. Godtliebsen. The same data were analyzed e.g in Barone and

Sebastiani (1992), Chu et al. (1998), and Godtliebsen et al. (1997). Reconstructions of the same image using Markov random field methods can be found in Godtliebsen and Sebastiani (1994). The upper left plot of Figure 4 shows the central part of the image. The upper right plot gives the estimate obtained by AWS. Parameters are selected by the procedure described in Subsection 3.4 using $\psi(t) = |t|$.

To illustrate the quality of reconstruction we present the result of another well established method to reduce the noise level by averaging several MR-Images recorded from the same slice of the brain, see the lower left plot. Images, recorded successively, can not be assumed to have exactly the same location. This leads to some convolution in the averaged image. In the lower right plot we show gray level densities for both the averaged image and the AWS estimate. Since gray levels correspond to certain tissues in the brain, a density with spikes is more what one would expect. By averaging images this property is lost. AWS allows to preserve the structure, although at the given noise level there is no definite decision whether peaks of the gray level density are due to the structure or introduced by the procedure.

5.2 An example from satellite imaging

In our last example, suggested by a referee, we use a log-transformed C-band, HH-polarization, synthetic aperture radar (SAR) image recorded by Dr. E. Attema at the European Space Research and Technology Centre in Noordwijk, Netherlands. The example is also used in Glasbey and Horgan (1995). The data can be obtained from *ftp://peipa.essex.ac.uk/ipa/pix/books/glasbey-horgan/* . The image shows an area near Thetford forest, England.

********** put Figure 5 around here **********

In Figure 5 we show the noisy original, the reconstruction obtained by AWS and a residual image. Parameters for AWS are selected according to Subsection 3.4 using $\psi(t) = |t|$.

6 Conclusions

The simulated results and the examples demonstrate a reasonable performance of the proposed procedure especially in situations where the underlying image is piecewise constant or can be approximated by such images. In such cases the procedure outperforms most other methods. The nice visual quality of restoration for such examples is due to the two most important features of the method: the estimated image is homogeneous within every large homogeneous regions independently of its shape and, simultaneously the procedure provides a reasonable quality of estimation near image edges. The procedure is very stable w.r.t. increasing noise level. All these issues are in agreement with theoretical properties of the procedure which surely should be investigated further. The algorithm can be easely applied to higher dimensional

situations.

Acknowledgements:

We would like to thank Fritjof Kruggel and Fred Godtliebsen for their introduction into MRI and for useful discussions. We also thank two unknown referees for their helpful suggestions.

Appendix: Proofs

Proof of Proposition 3.1. We argue by induction in k. First we consider the weights $w_1(X_i, X_j)$. Since every initial neighborhood $U_0(X_i)$ contains exactly N_0 design points, each estimate $\widehat{f}_0(X_i)$ is normal with the mean a and the variance $s_0^2(X_i) = \sigma^2/N_0$. First we evaluate the probability of the event

$$\{|\widehat{f}_0(X_i) - \widehat{f}_0(X_j)| > \lambda\sigma N_0^{-1/2} \text{ for some } i \neq j\}.$$

By (1),

$$\widehat{f}_0(X_i) - \widehat{f}_0(X_j) = \frac{1}{N_0} \sum_{U_0(X_i)} \varepsilon_\ell - \frac{1}{N_0} \sum_{U_0(X_j)} \varepsilon_\ell$$

and this is a linear combination of Gaussian errors. Therefore this difference itself is a Gaussian zero mean random variable with

$$\boldsymbol{E}|\widehat{f}_0(X_i) - \widehat{f}_0(X_j)|^2 = \sigma^2 N_0(X_i, X_j)/N_0^2$$

where $N_0(X_i, X_j)$ is the number of design points lying either in $U_0(X_i)$ or in $U_0(X_j)$ but not in their intersection,

$$N_0(X_i, X_j) = \#\left\{U_0(X_i) \cup U_0(X_j) \setminus U_0(X_i) \cap U_0(X_j)\right\}.$$

Obviously $N_0(X_i, X_j) \leq 2N_0$. Therefore,

$$\begin{aligned}
&\boldsymbol{P}\left(|\widehat{f}_0(X_i) - \widehat{f}_0(X_j)| > \lambda\sigma N_0^{-1/2}\right) \\
&\quad \leq \exp\left\{-\frac{\lambda^2\sigma^2 N_0^{-1}}{2\sigma^2 N_0(X_i, X_j) N_0^{-2}}\right\} \\
&\quad \leq \exp\left\{-\frac{\lambda^2 N_0}{2N_0(X_i, X_j)}\right\} \\
&\quad \leq \exp\{-\lambda^2/4\}.
\end{aligned}$$

In the adaptation step we compute the weights $w_1(X_i, X_j)$ for all X_i and for every X_j from $U_1(X_i)$. This involves about $nN_1/2$ comparisons for different pairs (X_i, X_j). Therefore

$$\boldsymbol{P}\left(\{|\widehat{f}_0(X_i) - \widehat{f}_0(X_j)| > \lambda\sigma N_0^{-1/2} \text{ for some } i \neq j\}\right)$$

$$\le \sum_{i=1}^{n} \sum_{U_1(X_i)} \boldsymbol{P}\left(|\widehat{f}_0(X_i) - \widehat{f}_0(X_j)| > \lambda\sigma N_0^{-1/2}\right)$$
$$\le 0.5\, nN_1 \exp\{-\lambda^2/4\}.$$

We see that all the weights $w_1(X_i, X_j) = 1$ with a probability greater than $1 - 0.5\, nN_1 \exp\{-\lambda^2/4\}$. Therefore, assuming that an event of type $\{w_1(X_i, X_j) = 0\}$ does not occur, all estimates $\widehat{f}_1(X_i)$ are simply mean values of the observations Y_j over $U_1(X_i)$.

All these arguments apply to the next iteration with $\widehat{f}_1$ in place of $\widehat{f}_0$ and so on.

Now suppose that we have got the equal weights $w_{k'}(X_i, X_j) = 1$ for all $k' \le k$ with a probability of at least $1 - \gamma_k$ with some number γ_k. We intend to estimate the similar probability to the next iteration. First we note that by the previous arguments $w_{k+1}(X_i, X_j) = 1$ for all $i \ne j$ with a probability of at least $1 - \gamma_k - 0.5 nN_{k+1} \exp\{-\lambda^2/4\}$. It remains only to check that the control step does not reject the estimate $\widehat{f}_{k+1}(X_i)$. Let $k' \le k$. Then obviously

$$\widehat{f}_{k+1}(X_i) - \widehat{f}_{k'}(X_i) = N_{k+1}^{-1} \sum_{U_{k+1}(X_i)} Y_j - N_{k'}^{-1} \sum_{U_{k'}(X_i)} Y_j = N_{k+1}^{-1} \sum_{U_{k+1}(X_i)} \varepsilon_j - N_{k'}^{-1} \sum_{U_{k'}(X_i)} \varepsilon_j.$$

Since all errors ε_j are independent $\mathcal{N}(0, \sigma^2)$ r.v.'s, this difference is also a normal zero mean r.v. with the variance

$$\boldsymbol{E}\left(\widehat{f}_{k+1}(X_i) - \widehat{f}_{k'}(X_i)\right)^2 = \sigma^2 (N_{k'}^{-1} - N_{k+1}^{-1}) \le \sigma^2 N_{k'}^{-1}.$$

Therefore, using $s_{k'}^2(X_i) = \sigma^2 N_{k'}^{-1}$

$$\boldsymbol{P}\left(|\widehat{f}_{k+1}(X_i) - \widehat{f}_{k'}(X_i)| > \eta\sigma N_{k'}\right) \le \boldsymbol{P}\left(|\xi| > \eta\right) \le \exp\{-\eta^2/2\}$$

where ξ denotes a standard normal r.v. The total number of such control tests is not greater than nk and the probability that at least one such event occurs at the $(k+1)$-th iteration can be bounded by $\exp\{-\eta^2/2\} nk$. Therefore, if

$$\gamma_{k+1} = \gamma_k + 0.5\, nN_{k+1} \exp\{-\lambda^2/4\} + \exp\{-\eta^2/2\} nk,$$

then, with a probability greater or equal to $1 - \gamma_{k+1}$, we get all $w_{k+1}(X_i, X_j) = 1$.

Summing over all iterations we get the following upper bound for γ_{k*}

$$\begin{aligned} \gamma_{k^*} &\le 0.5\, n \exp\{-\lambda^2/4\} \sum_{k=1}^{k^*} N_k + \exp\{-\eta^2/2\} n \sum_{k=1}^{k*} k \\ &\le C\, n^2 \exp\{-\lambda^2/4\}/2 + \exp\{-\eta^2/2\} nk^*(k^*+1)/2 \end{aligned}$$

as required. □

Proof of Proposition 3.2. Let us fix one pair (X_{i_1}, X_{i_2}) with $X_{i_m} \in A_m^\circ$, $m = 1, 2$. First we note that $\widehat{f}_0(X_{i_m}) \sim \mathcal{N}(a_m, \sigma^2 N_0^{-1})$ and we may represent these estimates in the

form $\widehat{f}_0(X_{i_m}) = a_m + \sigma N_0^{-1/2}\xi_{i_m}$ where $m = 1, 2$ and ξ_{i_1} and ξ_{i_2} are independent standard Gaussian r.v.'s.

Next, in view of the control step, we have for every $k \geq 1$

$$|\widehat{f}_k(X_{i_m}) - \widehat{f}_0(X_{i_m})| \leq \eta\sigma N_0^{-1/2}, \qquad m = 1, 2.$$

Therefore

$$\begin{aligned}
|\widehat{f}_k(X_{i_1}) - \widehat{f}_k(X_{i_2})| &\geq |\widehat{f}_0(X_{i_1}) - \widehat{f}_0(X_{i_2})| \\
&\quad - |\widehat{f}_k(X_{i_1}) - \widehat{f}_0(X_{i_1})| - |\widehat{f}_k(X_{i_2}) - \widehat{f}_0(X_{i_2})| \\
&\geq |a_1 - a_2| - 2\eta\sigma N_0^{-1/2} - \sigma N_0^{-1/2}|\xi_{i_1} - \xi_{i_2}|.
\end{aligned}$$

The difference $\xi_{i_1} - \xi_{i_2}$ is a zero mean Gaussian r.v. with the variance 2 and hence

$$\begin{aligned}
&\boldsymbol{P}\left(|\widehat{f}_k(X_{i_1}) - \widehat{f}_k(X_{i_2})| < \lambda\sigma N_0^{-1/2}\right) \\
&\quad \leq \boldsymbol{P}\left(|\xi_{i_1} - \xi_{i_2}| > \sigma^{-1}N_0^{1/2}|a_1 - a_2| - 2\eta\right) \\
&\quad \leq \exp\left\{-\left(\sigma^{-1}N_0^{1/2}|a_1 - a_2| - 2\eta\right)^2/4\right\}.
\end{aligned}$$

The number of such pairs can be very roughly bounded by $nN_k/2$ and therefore the probability to meet by k-th iteration at least one such event is smaller than

$$0.5nN_k \exp\left\{-\left(\sigma^{-1}N_0^{1/2}|a_1 - a_2| - 2\eta\right)^2/4\right\}.$$

Summing up over all $k \leq k^*$ and using (7), we obtain the required assertion. □

References

[1] Banerjee, S. and Rosenfeld, A. (1993) MAP estimation of piecewise constant digital signals. *CVGIP: Image understanding*, **57**, 63–80.

[2] Besag, J. (1986) On the statistical Analysis of dirty pictures (with discussion). *J. R. Statist. Soc.*, Ser. **B**, **48** , 259–302.

[3] Barone, P. and Sebastiani, G. (1992) A new method of magnetic resonance image reconstruction with short acquisition time and truncation artifact reduction. *IEEE Trans. on Medical Imaging*, **11** , 250–259.

[4] Bose, S. and O'Sullivan, F. (1997) A Region-Based Segmentation Method for Multichannel Image Data. *J. Amer. Statist. Ass.*, **92**, 92–106.

[5] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*. Monterey: Wadsworth and Brooks/Cole.

[6] Chu, C. K., Glad I. K., Godtliebsen, F. and Marron, J. S. (1998) Edge preserving smoothers for image processing (with discussion). *J. Amer. Statist. Ass.*, **93**, 526–556.

[7] Coifman, R. R. and Donoho, D. L. (1995) Translation-invariant de-noising. In *Lecture Notes in Statistics: Wavelets and Statistics*, (ed. A. Antoniadis), 125–150. New York: Springer.

[8] Daubechies, I. (1992) *Ten Lectures on Wavelets*. Philadelphia: SIAM.

[9] Donoho, D. L. (1997) CART and best-ortho-basis: A connection. *Ann. Statist.*, **25**, 1870–1911.

[10] Donoho, D. L., Johnstone I. M., Kerkyacharian, G. and Picard, D. (1995) Wavelet shrinkage: asymptopia ?. *J. Royal Statistical Society*, Ser. **B**, **57**, 301–369.

[11] Engel, J. (1994) A simple wavelet approach to nonparametric regression from recursive partitioning schemes. *J. Multivariate Anal.*, **49**, 242-254.

[12] Gasser, T. and Sroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika*, **73**, 625–633.

[13] Geman, D. (1990) *Random Fields and Inverse Problems in Imaging. Lecture Notes in Mathematics*, New York: Springer.

[14] Glasbey, C. A. and Morgan, G. W. (1995) *Image Analysis for the Biological Sciences*, New York: Wiley.

[15] Godtliebsen, F. and Sebastiani, G. (1994) Statistical methods for noisy images with discontinuities. *J. Applied Statistics*, **21** , 459–477.

[16] Godtliebsen, F., Spjøtvoll, E. and Marron, J. S. (1997) A nonlinear Gaussian Filter applied to images with discontinuities. *J. Nonparametric Statistics*, **8**, 21-43.

[17] Korostelev, A. and Tsybakov, A. (1993) *Minimax Theory of Image Reconstruction*. New York–Heidelberg–Berlin: Springer.

[18] Lange, N. (1996) Tutorial in biostatistics: Statistical approaches to human brain mapping by functional magnetic resonance imaging. *Statistics in Medicine*, **15**, 389–428.

[19] Lee, J.S. (1983) Digital image smoothing and the sigma-filter. *Computer Vision, Graphics and Image Processing*, **24**, 255–269.

[20] Lepski, O., Mammen, E. and Spokoiny, V. (1997) Ideal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selection. *Ann. Statist.*, **25**, 929–947.

[21] Lepski, O. and Spokoiny, V. (1997) Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.*, **25**, 2512–2546.

[22] Müller, H.-G. (1992) Change points in nonparametric regression analysis. *Ann. Statist.*, **20**, 737–761.

[23] Nason, G. P. and Silverman, B. W. (1994) The discrete wavelet transform in S. *Journal of Computational and Graphical Statistics*, **3**, 163-191.

[24] Neumann, M. H. (1998) Multivariate wavelet thresholding in anisotropic function spaces, Manuscript.

[25] Polzehl, J. and Spokoiny, V. G. (1998) Image denoising: pointwise adaptive approach, DP 38/98, SFB 373, Humboldt University Berlin.

[26] Scott, D. W. (1992) *Multivariate Density Estimation*, New York: Wiley.

[27] Sebastiani, G. and Barone, P. (1991) Mathematical principles of basic magnetic resonance imaging in medicine. *Signal Processing*, **25**, 227–250.

[28] Sebastiani, G. and Barone, P. (1995) Truncation artifact reduction in magnetic resonance imaging by markov random field methods. *IEEE Trans. on Medical Imaging*, **24**, 434–441.

[29] Speckman, P. L. (1994) Fitting curves with features: Semiparametric change point methods, In *Computing Science and Statistics* (ed. J. Sall, A. Lehman), pp. 257– 264. Fairfax Station: Interface Foundation of North Amererica, Inc.

[30] Spokoiny, V. (1998) Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. *Ann. Statist.*, **26**, 1356–1378.

[31] Tsybakov, A. (1989) Optimal estimation accuracy of nonsmooth images. *Problem Inf. Trans.*, **25**, 180–191.

[32] Winkler, G. (1995) *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*, Heidelberg: Springer.

[33] Wu, Z. (1993) Homogeneity testing for unlabeled data: A performance evaluation. *CVGIP: Graphical Models and Image Processing*, **55**, 370–380.

[34] Wu, J. S. and Chu, C. K. (1993) Kernel-type estimators of jump points and values of a regression function. *Ann. Statist.*, **21**, 1545–1566.

	MISE			LDP		
	$\sigma = .25$	$\sigma = .5$	$\sigma = 1$	$\sigma = .25$	$\sigma = .5$	$\sigma = 1$
AWS	0.003	0.023	0.123	0.002	0.026	0.243
Gauss filtering	0.019	0.039	0.081	0.069	0.182	0.368
Nonlinear gauss	0.013	0.040	0.089	0.038	0.189	0.394
Modal regression	0.008	0.040	0.084	0.015	0.179	0.378
Change point	0.014	0.043	0.095	0.047	0.196	0.388
CART	0.006	0.031	0.149	0.008	0.066	0.414
Wavelets	0.009	0.038	0.097	0.008	0.146	0.379
MRF	0.010	0.044	0.112	0.021	0.191	0.422

Table 1: Estimated mean integrated squared error (MISE) and large deviation probability (LDP) in the univariate simulation experiment

	MISE			LDP		
	$\sigma = .25$	$\sigma = .5$	$\sigma = 1$	$\sigma = .25$	$\sigma = .5$	$\sigma = 1$
AWS	0.0021	0.0109	0.0328	0.007	0.032	0.119
Gauss filtering	0.0138	0.0243	0.0396	0.212	0.313	0.452
Nonlinear gauss	0.0096	0.0262	0.0454	0.151	0.334	0.491
Modal regression	0.0068	0.0254	0.0426	0.078	0.290	0.479
Wavelets	0.0079	0.0147	0.0468	0.073	0.172	0.437
MRF	0.0050	0.0204	0.0475	0.048	0.248	0.497

Table 2: Estimated mean integrated squared error (MISE) and large deviation probability (LDP) in the bivariate simulation experiment

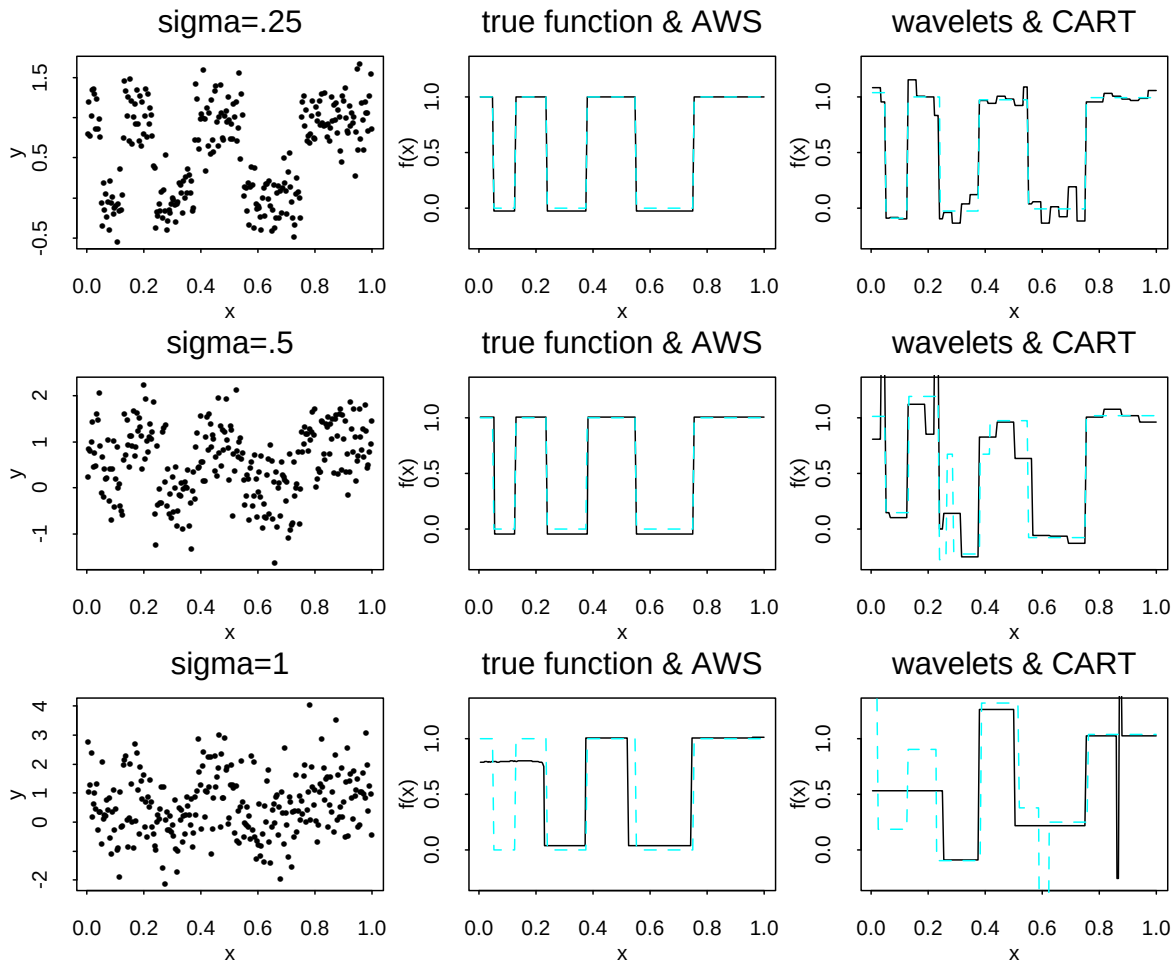

Figure 1: Data generated in the univariate experiment (left column), true function (dashed line), AWS estimate (solid line) (central column), wavelet estimate (best hard thresholding,solid line) and CART estimate (dashed line) (right column) for standard deviations $\sigma = .25$, .5 and 1.

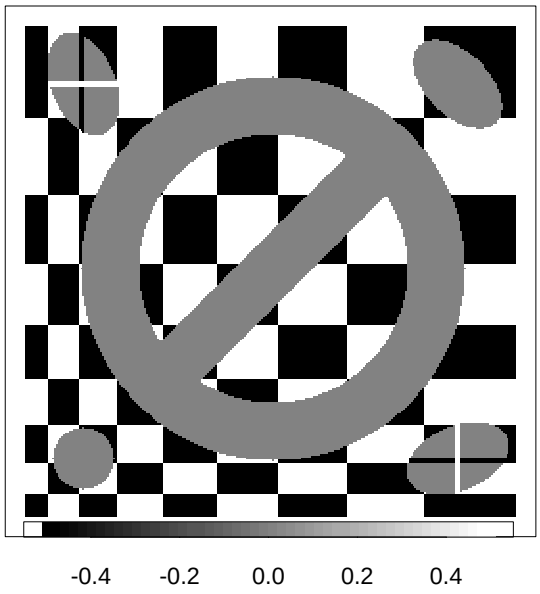

Figure 2: Artificial test image

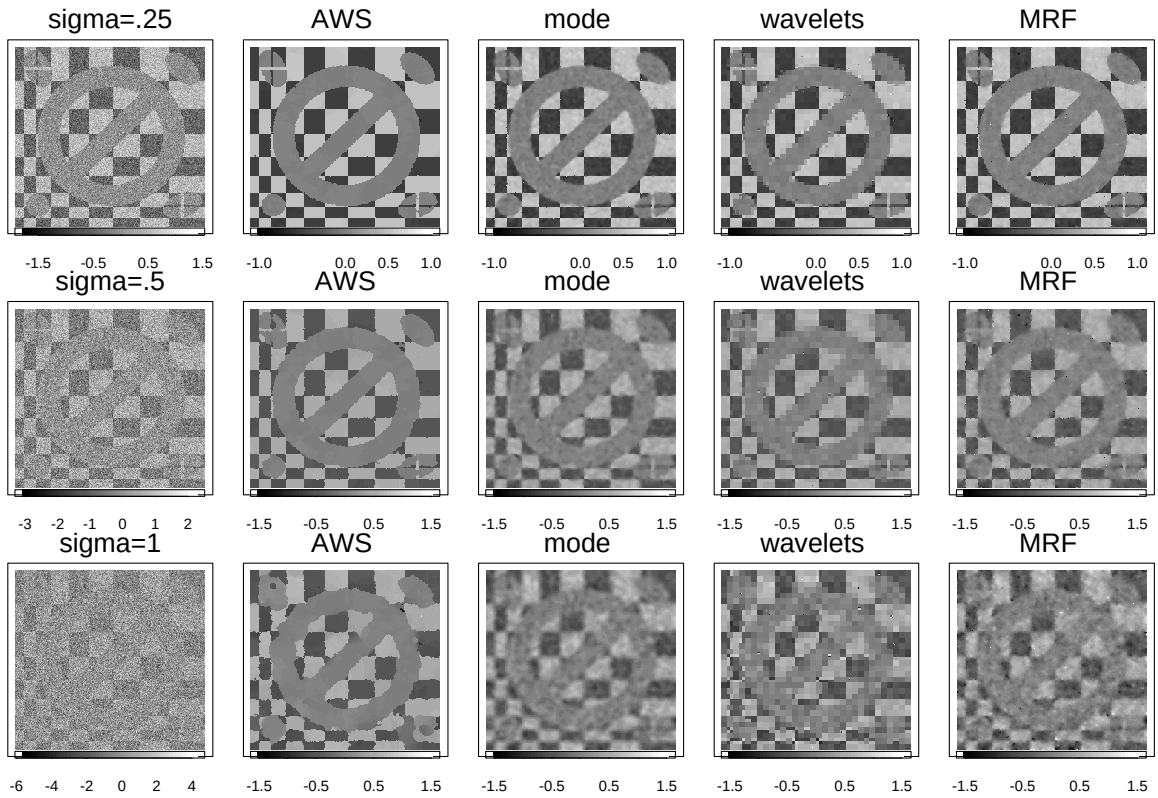

Figure 3: Image plus noise (left), AWS-reconstructions (second), modal regression (third), wavelet (fourth) and MRF estimate (right column) for different values of σ.

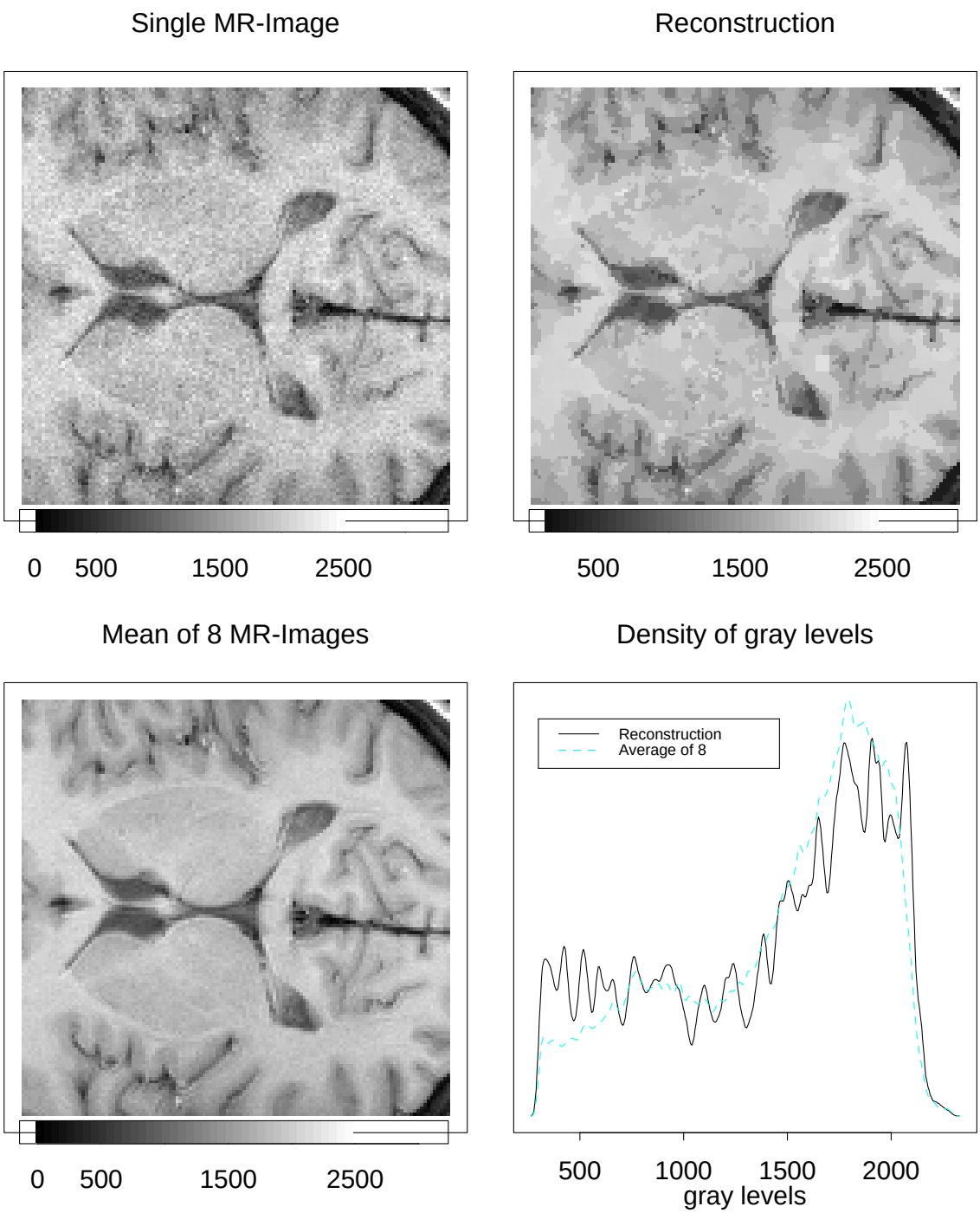

Figure 4: Central part of Original Magnetic Resonance image and AWS estimate (upper row); average of eight images of the same slice and estimated densities of gray levels (lower row)

Figure 5: Synthetic aperture radar (SAR) data: original image (left), AWS Reconstruction (central) and residuals (right).

Figure 1: Data generated in the univariate experiment (left column), true function (dashed line), AWS estimate (solid line) (central column), wavelet estimate (best hard thresholding,solid line) and CART estimate (dashed line) (right column) for standard deviations $\sigma = .25$, .5 and 1.

Figure 2: Artificial test image

Figure 3: Image plus noise (left), AWS-reconstructions (second), modal regression (third), wavelet (fourth) and MRF estimate (right column) for different values of σ.

Figure 4: Central part of Original Magnetic Resonance image and AWS estimate (upper row); average of eight images of the same slice and estimated densities of gray levels (lower row)

Figure 5: Synthetic aperture radar (SAR) data: original image (left), AWS Reconstruction (central) and residuals (right).

Table 1: Estimated mean integrated squared error (MISE) and large deviation probability (LDP) in the univariate simulation experiment

Table 2: Estimated mean integrated squared error (MISE) and large deviation probability (LDP) in the bivariate simulation experiment