Adaptive Weights Smoothing with Applications to Image Segmentation

Jörg Polzehl and Vladimir G. Spokoiny

Weierstrass Institute for Applied Analysis and Stochastics,
Mohrenstr. 39, 10117 Berlin

Keywords: locally constant approximation,
adaptive weights,
edge restoration,
image segmentation,
magnetic resonance imaging

AMS-Classification: Primary 62G05, Secondary 62G25, 62P10
Short Title: Adaptive Weights Smoothing

Abstract

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 Imaging.

1 Introduction

In regression problems arising from many scientific disciplines discontinuities of the regression function play an important role. This is especially the case when analyzing data in form of two or three or even higher dimensional images. Such images meet in several fields, e.g. from satellite imaging, radar, 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) or Speckman (1994) for procedures based on change point detection. 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 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.

Another approach to image estimation was proposed recently in Polzehl and Spokoiny (1998). The method can be viewed as a multidimensional analog of the procedure from Spokoiny (1997) 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 is fully adaptive, that is, no prior information is required.

The proposed procedure is iterative. Let us observe Y_i at a design point X_i, $i = 1, \ldots, n$. We start from a simple pilot estimate $\widetilde{f}_0(X_i)$ of the unknown function f which we use for a preliminary image segmentation. Namely, on the base of this estimate, for each design point X_j from a neighborhood of X_i, we calculate a corresponding weight $w(X_i, X_j)$ which is about zero, if the difference $\widetilde{f}_0(X_i) - \widetilde{f}_0(X_j)$ is essentially larger than its standard deviation, and which is about one otherwise. Then we calculate the next step estimate $\widetilde{f}_1(X_i)$ by averaging the observations Y_j with the weights $w(X_i, X_j)$. We iterate in this way increasing each time the size of the considered neighborhood around each point. In the sequel the proposed method is referred to as adaptive weights smoothing (AWS).

The further discussion and the precise description of the procedure are placed in Section 2. Some possible generalizations are discussed in Section 6. In Section 3 we study some theoretic properties of our AWS method. Section 4 introduces several alternatives for univariate and bivariate situations. These procedures are compared with our estimator by simulation. Finally Section 5 describes an application of AWS to Magnetic Resonance Imaging (MRI).

1.1 Model

The model can be described as

$$Y_i = f(X_i) + \varepsilon_i \qquad X_i \in I\!R^p \qquad \boldsymbol{E}\varepsilon_i = 0 \qquad \boldsymbol{D}\varepsilon_i = \sigma_i^2. \tag{1.1}$$

Here $X_1, \ldots, X_n$ are design points which are usually assumed to be equispaced in the unit cube $[0,1]^d$. At each point X_i we observe the regression function $f(X_i)$ with some additive error ε_i. We suppose independent heterogeneous noise, i.e. ε_i are independent zero mean random variables with unknown distribution which possibly depends on location. 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

$$f(x) = \sum_{m=1}^{M} a_m \mathbf{1}(x \in A_m) \tag{1.2}$$

where $a_1, \ldots, a_M$ are some numbers. 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. We only assume that these regions are large compared with the distance between two neighbor design points.

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 (1.2) can be reformulated 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 an obvious estimate

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

where N_{A_m} denotes the number of design points in A_m. If X_i is a point in A_m, then we simply set $\widehat{f}(X_i) = \widehat{a}_m$. 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 the partition $A_1, \ldots, A_M$. Namely, for each pair of points X_i and X_j, we may decide on the basis of values $\widehat{f}_0(X_i)$ and $\widehat{f}_0(X_j)$ whether they are in the same region. 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 introduce for every design point X_i a set $\widehat{A}(X_i)$ with

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

where $\widehat{\sigma}_0(X_i)$ is the standard deviation of $\widehat{f}_0(X_i)$ and λ is some number. This set estimates the region A_m containing X_i. Using these estimated regions, we may define the 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

$$w_1(X_i, X_j) = \mathbf{1}\left(|\widehat{f}_0(X_i) - \widehat{f}_0(X_j)| \le \lambda \widehat{\sigma}_0(X_i) \right). \tag{2.1}$$

Then 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 essential 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. This means that we calculate the initial pilot estimate $\widehat{f}_0(X_i)$ by averaging of observations over a small neighborhood $U_0(X_i)$ of the point X_i (in many situations it can be the observation Y_i itself). Then we recalculate this estimate by averaging over a larger neighborhood $U_1(X_i)$ but now we use only data points where there are no essential differences between values of the initial estimates. We continue in this way, increasing each time the considered neighborhood $U_k(X_i)$, that is, for each $k \ge 1$,

$$\widehat{f}_k(X_i) = \frac{\sum\limits_{X_j \in U_k(X_i)} w_k(X_i, X_j) Y_j}{\sum\limits_{X_j \in U_k(X_i)} w_k(X_i, X_j)} \tag{2.2}$$

where the weights $w_k(X_i, X_j)$ are computed by comparison of the preceding estimates $\widehat{f}_{k-1}(X_i)$ and $\widehat{f}_{k-1}(X_j)$. Secondly we use continuous weights $w_k(X_i, X_j)$ instead of zero-one weights in (2.1).

Finally, to stabilize the procedure, we add also a control step. The initial estimate $\widehat{f}_0(X_i)$ is very rough in the sense that it is the average over a small neighborhood of X_i. But, if the noise level is not too large, it still allows to make some image classification near each edge. By each iteration we increase the size of the considered neighborhood around each point which allows to reduce the stochastic error drastically. At the same time, we have to prevent a situation when the original correct classification will be lost as a result of iteration. For this we apply the control step which means roughly a comparison of the new estimate with the estimates from preceding iterations. It is worth to mention that this control step alone can be used to construct an adaptive estimate, see Lepski, Mammen and Spokoiny (1997) or Lepski and Spokoiny (1997).

Now we present a formal description of the method starting with a description of the input parameters of the algorithm.

2.2 Parameters of the procedure

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

A sequence of neighborhoods: For each design point x, we assume to be given a sequence of neighborhoods $U_k(x), \quad k = 0, 1, \ldots, k^*$ with $U_k(x) \subset U_{k+1}(x)$ containing x. One possible choice of these neighborhoods $U_k(x)$ would be $U_k(x) = \{X_i : |X_i - x| < d_k\}$ 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.

Further we denote by $N_k(x)$ the number of design points X_i in $U_k(x)$,

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

Estimates of noise variances: We consider the case of a heteroscedastic noise, that is, each error ε_i is a random variable with zero mean and a variance σ_i^2. The variances σ_i^2 are unknown but we suppose to be given consistent estimates $\widehat{\sigma}_i^2$ of σ_i^2.

Parameters: 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 integer value k^* stands for the maximal number of iterations.

A kernel: We fix a univariate kernel K satisfying usual conditions: it is a symmetric, compactly supported smooth function with the maximum at zero and it is nonincreasing on the positive semiaxis.

2.3 The procedure

We begin by an initialization.

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

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

and set $k = 1$.

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

$$w_k(X_i, X_j) = K\left(\frac{\widehat{f}_{k-1}(X_i) - \widehat{f}_{k-1}(X_j)}{\lambda \widehat{s}_{k-1}(X_i)}\right) \tag{2.3}$$

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

$$\widehat{f}_k(X_i) = \frac{\sum\limits_{X_j \in U_k(X_i)} w_k(X_i, X_j) Y_j}{\sum\limits_{X_j \in U_k(X_i)} w_k(X_i, X_j)}, \tag{2.4}$$

$$\widehat{s}_k^2(X_i) = \frac{\sum\limits_{X_j \in U_k(X_i)} w_k^2(X_i, X_j) \widehat{\sigma}_j^2}{(\sum\limits_{X_j \in U_k(X_i)} w_k(X_i, X_j))^2} \tag{2.5}$$

for all X_i.

Control: Denote by $\mathcal{K}$ the geometric set of indices $\mathcal{K} = \{0, 1, 2, 4, \ldots, 2^{\ell}, \ldots\}$. 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 where $k' \in \mathcal{K}$ and $k' < k$. If there is at least one index $k' < k$ from $\mathcal{K}$ 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.

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

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 a homogeneous Gaussian noise with the known variance σ^2. We also consider the uniform kernel $K(x) = \mathbf{1}(|x| \le 1)$.

3.1 Estimation inside a homogeneous region

We consider 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

$$N_1 + \ldots + N_{k^*} \le Cn \tag{3.1}$$

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 $X_j \in U$*

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

$$\le e^{-\lambda^2/4} n^2 C/2 + e^{-\eta^2/2} n k^* \log_2(2k^*)$$

where C is from (3.1).

Proof. 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

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

By (1.1),

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

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 $N_0(X_i, X_j) \le 2N_0$. Therefore,

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

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

$$\begin{aligned}
&\boldsymbol{P}\left(\{|\widehat{f}_0(X_i) - \widehat{f}_0(X_j)| > \lambda\sigma N_0^{-1/2} \text{ for some } i \ne j\}\right) \\
&\le \sum_{i=1}^{n} \sum_{U_1(X_i)} \boldsymbol{P}\left(|\widehat{f}_0(X_i) - \widehat{f}_0(X_j)| > \lambda\sigma N_0^{-1/2}\right) \\
&\le 0.5\, nN_1 e^{-\lambda^2/4}.
\end{aligned}$$

We see that all the weights $w_1(X_i, X_j) = 1$ with a probability greater than $1 - 0.5\, nN_1 e^{-\lambda^2/4}$. 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 $1 - \gamma_k - 0.5nN_{k+1}e^{-\lambda^2/4}$. It remains only to check that the control step does not reject the estimate $\widehat{f}_{k+1}(X_i)$. Let $k' = 0$ or $k' = 2^\ell$ and $k' \le k$. Then obviously

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

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 $s^2_{k'}(X_i) = \sigma^2 N_{k'}^{-1}$

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

where ξ denotes a standard normal r.v. The total number of such control tests is not greater than $n\log_2(2k)$ and the probability that at least one such event occurs at the $(k+1)$-th iteration can be bounded by $e^{-\eta^2/2}n\log_2(2k)$. Therefore, if

$$\gamma_{k+1} = \gamma_k + 0.5\,nN_{k+1}e^{-\lambda^2/4} + e^{-\eta^2/2}n\log_2(2k),$$

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*}

$$\begin{aligned}\gamma_{k^*} &\le 0.5\,ne^{-\lambda^2/4}\sum_{k=1}^{k^*} N_k + e^{-\eta^2/2}n\sum_{k=1}^{k*}\log_2(2k) \\ &\le C\,n^2e^{-\lambda^2/4}/2 + e^{-\eta^2/2}nk^*\log_2(2k^*)\end{aligned}$$

as required. $\square$

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. All results allow 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 which belong to this region together with the corresponding neighborhood $U_0(X_i)$,

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

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 $X_{i_1} \in A^\circ_1$ and $X_{i_2} \in A^\circ_2$ and for all $k \ge 1$.

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*

$$w_k(X_i, X_j) = 0, \qquad \forall X_i \in A^\circ_1,\ X_j \in A^\circ_2,\ \textit{and}\ \forall k \le k^*,$$

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 (3.1).*

Proof. Let us fix one such pair. 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

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

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

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

Summing up over all $k \leq k^*$ and using (3.1), we obtain the required assertion. $\square$

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

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

then this above probability can be bounded by $n^2 e^{-\eta^2}$ and this value is again very small for n sufficiently large.

The results of Propositions 3.1 and 3.2 lead to the following conclusion: inside each large region the probability of a misclassification is small and misclassification errors occur typically only near the edge of each region. If we have only a few such regions with regular edges and if image contrasts are sufficiently large then we would expect a good quality of image recovering. On the other hand, if the image is very complex, then the number of points near edges might be comparable with the total number of design points. In such situations an application of the proposed procedure usually leads to oversmoothing and a combination of different regions.

3.3 Selection of parameters

There are several parameters involved in the algorithm. Our practical recommendations to their choice can be summarized as follows:

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

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

For $1 \leq \alpha/\sigma \leq 2$, we take $N_0 = 5$ and for $\alpha/\sigma < 1$, a choice with $N_0 \geq 9$ can be desirable.

Sequence of neighborhoods U_k: The sequence should satisfy the conditions $X_i \in U_1(X_i)$ 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 grows exponentially, for instance $N_k(X_i) \asymp 2^k$.

$\boldsymbol{\lambda}$: This parameter is very important and its choice 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. Note that the bounds for a wrong classification given by Propositions 3.1 and 3.2 are very rough and conservative. Suitable parameters range is between $\lambda = 2.5$ for a low signal to noise ratio and $\lambda = 3$ for a high signal to noise ratio.

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

4 Simulations

4.1 A segmentation experiment

We first illustrate the behavior of the algorithm near the edge by a simple experiment.

We simulate data using the regression model (1.1) with equidistant design $X_i = i/99, i = 0, 99$ and regression function $f(x) = \mathbf{1}(x > .5)$. Figure 1 shows plots of the

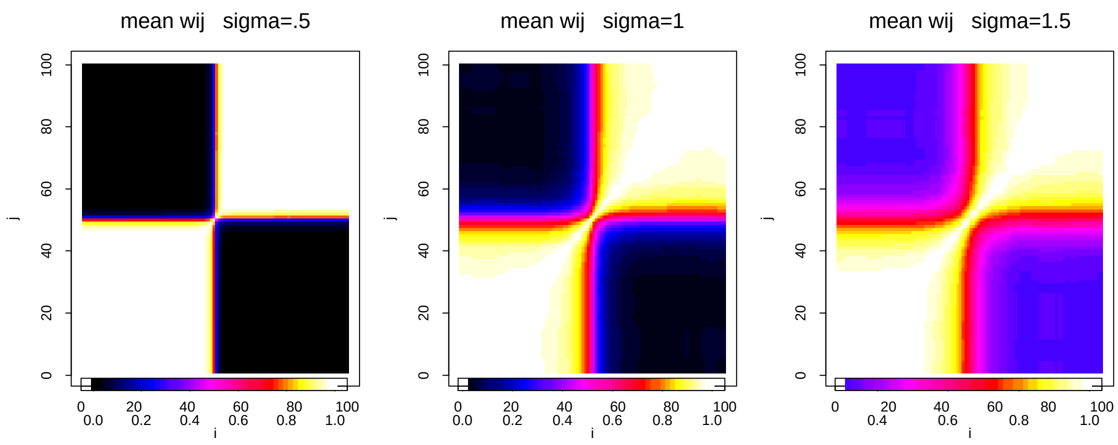

Figure 1: Mean weights w_{ij} from 500 simulations in experiment 1.

mean limiting weights $w_{k^*}(X_i, X_j)$ obtained in 500 simulations for different values of σ corresponding to signal to noise ratios of 2, 1 and 2/3. The left image ($\sigma = .5$) shows a very stable segmentation of the design and detection of the jump in the regression function. For $\sigma = 1$ the jump is detected in about 97% of the simulations. We observe an increase in the variability of its estimated location. Decreasing the signal to noise ratio to 2/3 leads to a situation where both the probability of not detecting the jump in the regression function and the variability of the location of an estimated jump increases. The discontinuity is still detected in about 68% of the 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 signal 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. The following list is far from being complete. We especially omitted Markov Random Field methods, see Geman (1990), and Wavelets, see e.g. Donoho et. al. (1995) or Wang (1995), which would both be suitable alternatives. The main reason is that we feel not to have enough experience with both classes of procedures to select appropriate parameters and to provide a fair comparison.

4.2 Alternative procedures

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

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

Nonlinear Gauss filtering: The Nonlinear Gauss Filter was proposed as an alternative to the Sigma Filter of Lee (1983) by Godtliebsen, Spjotvoll and Marron (1997). It replaces the discontinuous (uniform) weight function of the sigma filter by an Gaussian weight scheme. The filter is defined as

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

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

Modal regression: Modal regression is introduced in Scott (92) as an robust alternative to nonparametric regression procedures estimating a conditional mean. The modal regression curve is defined as

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

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 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 1.

Most of these procedures depend on one or two smoothing parameters. These parameters are chosen minimizing estimates of the mean squared error (MSE), values being determined by separate simulations.

4.3 A univariate simulation example

In our first univariate example we use a piecewise constant regression function with varying size of the homogeneous region, see Figure 2.

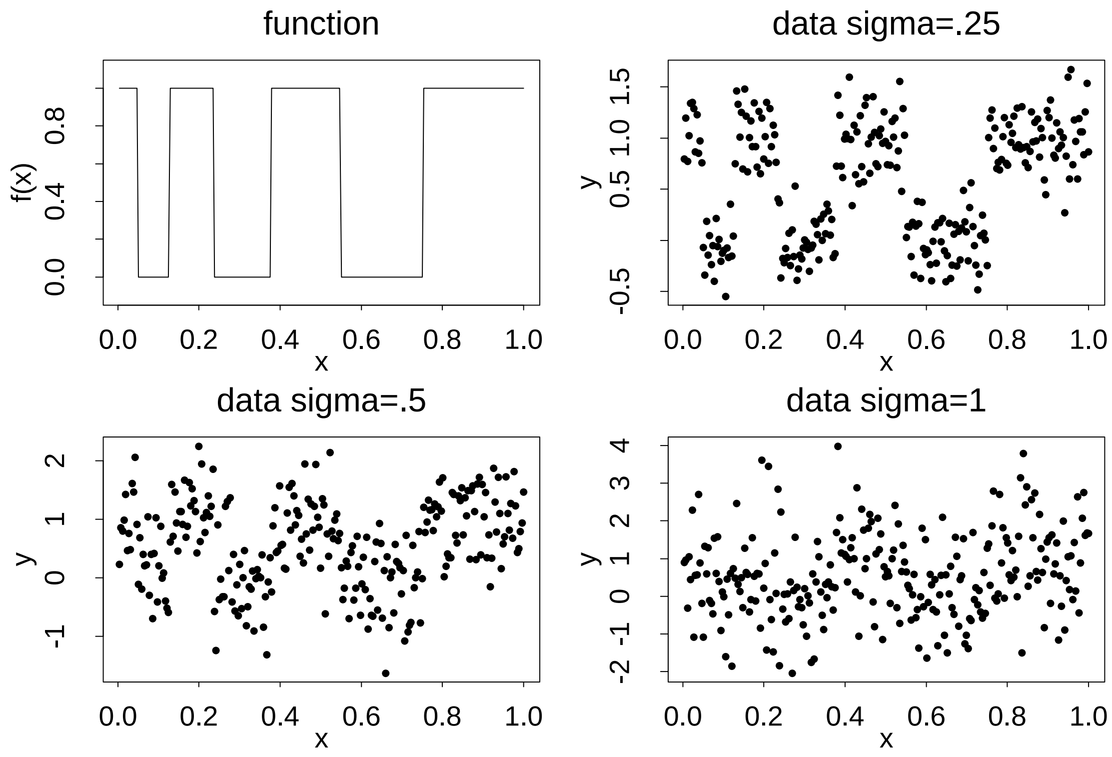

Figure 2: Function used in the univariate simulation experiment and generated observations for standard deviations $\sigma = .25$, .5 and 1.

Figure 3 presents the estimates obtained by AWS for the data displayed in Figure 2. The right plot shows an oversmoothing effect that occurs for some of the simulated datasets in case of $\sigma = 1$.

We run about 1000 simulations with sample size 256 and error standard deviation of $\sigma = .25$, .5 and 1, respectively. Parameters in the AWS estimate are fixed as $\lambda = 3$, $U_0(X^0)$ containing the point itself and U_{k^*} containing the whole sample.

Table 1 displays results of the simulations in terms of estimated MSE and number of points with estimates differing from the true value by more than $\alpha/4 = .25$ (wrong classifications).

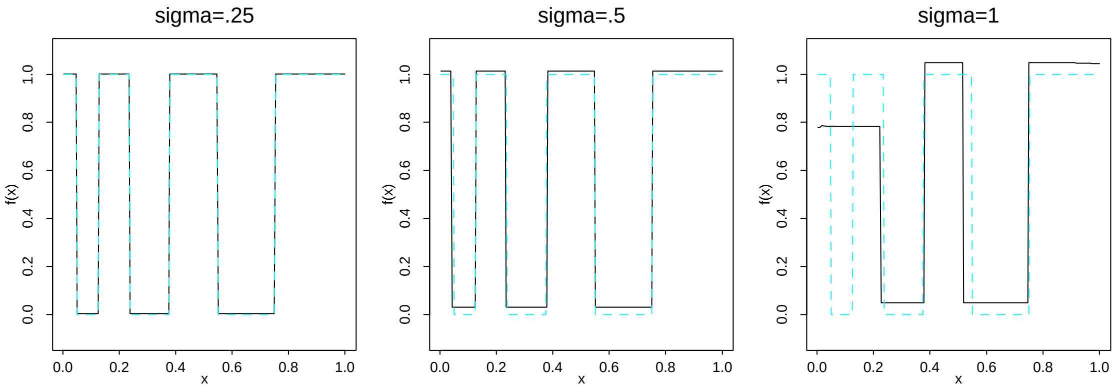

Figure 3: Estimates obtained by AWS (solid lines) and true function (dashed lines) for standard deviations $\sigma = .25$, $.5$ and 1.

	Mean squared error			Wrong classifications		
	$\sigma = .25$	$\sigma = .5$	$\sigma = 1$	$\sigma = .25$	$\sigma = .5$	$\sigma = 1$
Adaptive weights	0.0025	0.0231	0.1225	0.59	6.8	62.2
Gauss filtering	0.0191	0.0392	0.081	17.6	46.6	94.3
Nonlinear gauss	0.0133	0.0397	0.0886	9.82	48.4	100.8
Mode regression	0.0081	0.0403	0.08419	3.71	45.8	96.8
Change point	0.0138	0.043	0.09516	12.0	50.2	99.3
CART	0.0063	0.0314	0.1489	2.06	16.9	106.

Table 1: Estimated mean squared error (MSE) and mean number of wrong classifications in the univariate simulation experiment

Figure 4 and 5 display pointwise estimates of bias and the proportion of wrong classifications for the six procedures under consideration for standard deviation $\sigma = .5$. AWS performs uniformly better with respect to both MSE and mean number of wrong classifications 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 increased squared error and most of the wrong classifications for small x.

Note also the much better behavior of our procedure near the edges with respect to both bias and number of wrong classifications.

4.4 Two bivariate examples

We use two artificial images to demonstrate the power of our procedures in more complicated situations, see Figure 6. The left image possesses two different image contrasts, $\alpha = .5$ and $\alpha = 1$, 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 right and lower left of the image.

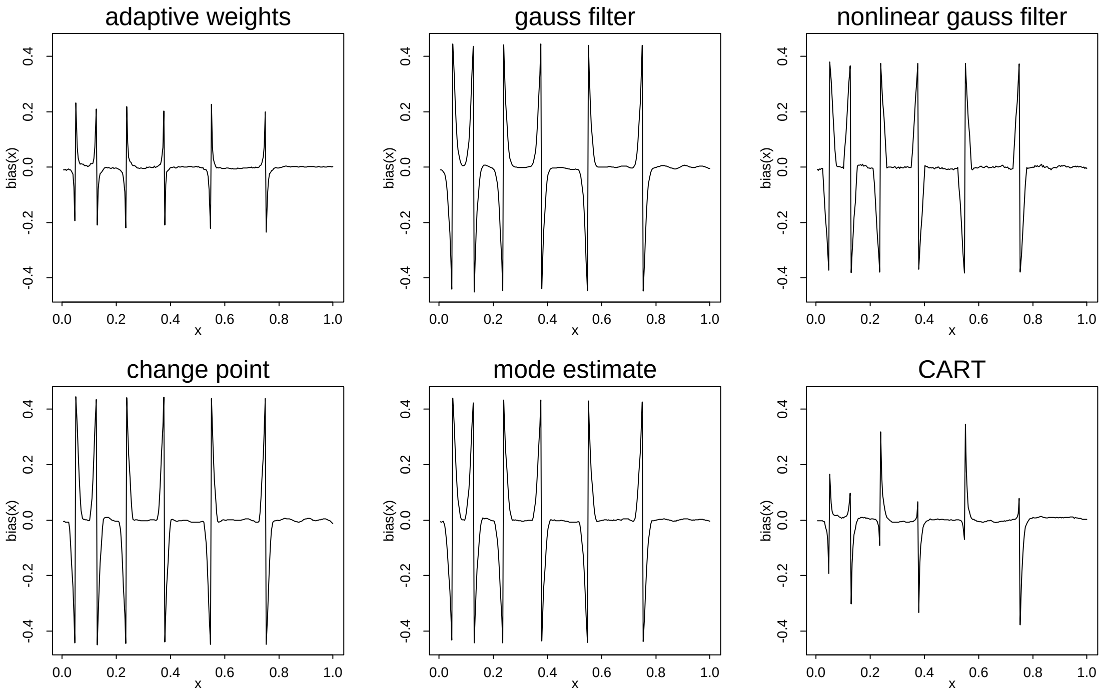

Figure 4: Pointwise bias estimated from 1000 simulations for $\sigma = .5$ in the univariate simulation experiment.

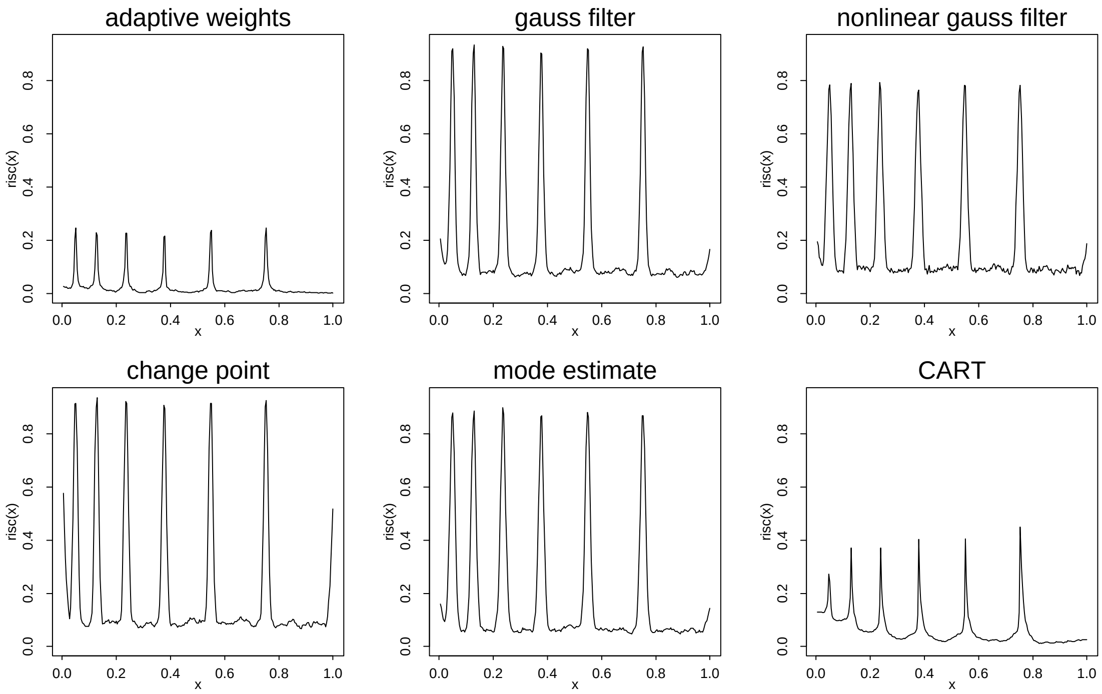

Figure 5: Proportion of wrong classifications estimated from 1000 simulations for $\sigma = .5$ in the univariate simulation experiment.

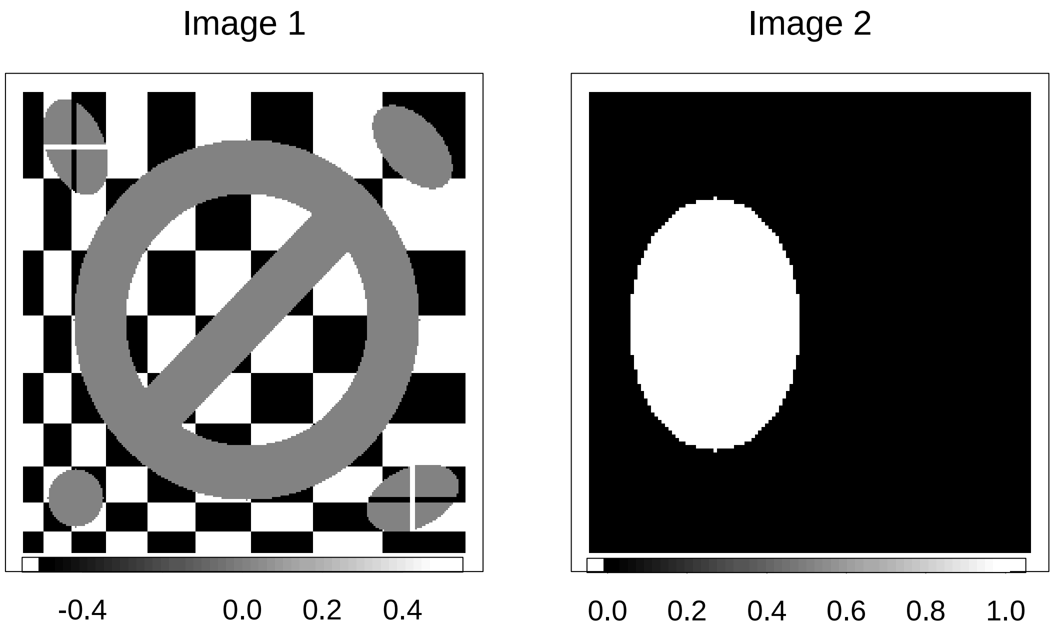

Figure 6: Two artificial test images

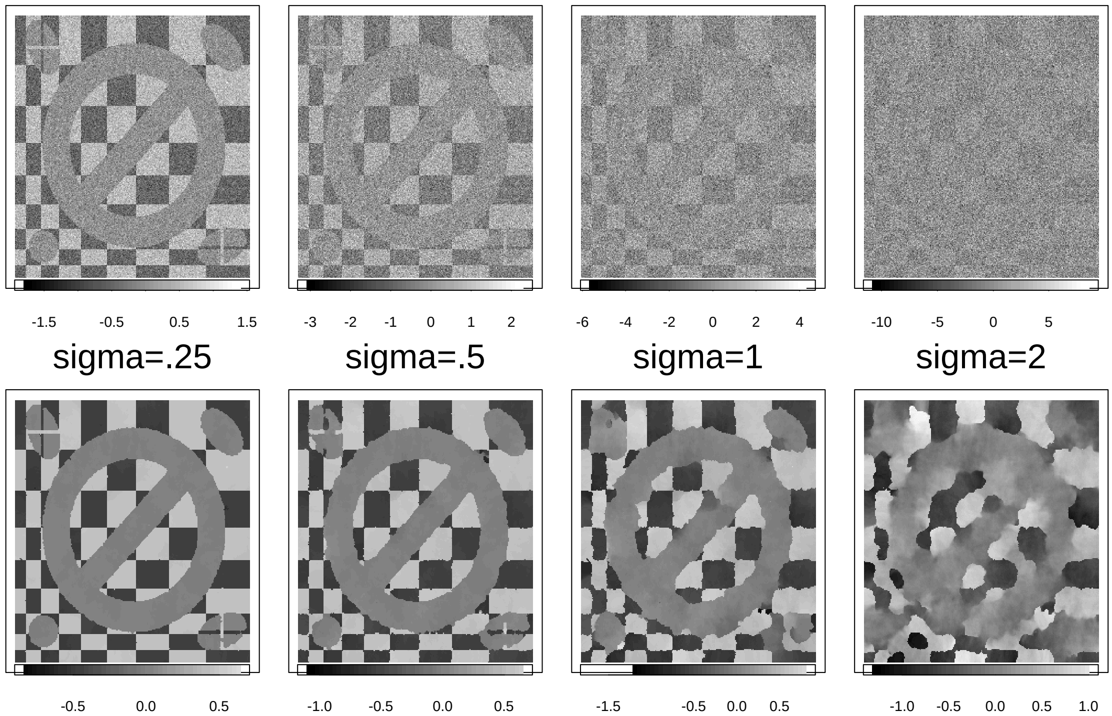

Figure 7: Image 1 plus noise (upper row) and AWS-reconstructions (lower row) for different values of σ

In the upper row of Figure 7 we display this image (image 1) distorted by additive Gaussian noise of standard deviation $\sigma = .25$, $.5$, 1 and 2, respectively. The lower row

	$\sigma = .25$	$\sigma = .5$	$\sigma = 1$	$\sigma = 2$
MSE (AWS estimate)	.0022	.0105	.0327	.0709
wrong classifications (original image)	40311	52556	59073	62265
wrong classifications (AWS estimate)	490	1968	9495	27622

Table 2: Image 1. MSE and number of points where the observed value (second row) and the estimate (third row) differs by more than $\alpha/4$ from the true value

	$\sigma = .5$	$\sigma = 1$	$\sigma = 2$	$\sigma = 4$	$\sigma = 6$	$\sigma = 8$
MSE (AWS estimate)	.0031	.0078	.0193	.0544	.0688	.0775
wrong classifications (image)	10000	13122	14761	15594	15850	15964
wrong classifications (AWS estimate)	49	130	307	1832	2665	5175

Table 3: Image 2. MSE and number of points where the observed value (second row) and the estimate (third row) differs by more than $\alpha/4$ from the true value

contains the reconstruction of the noisy images by AWS using parameters $\lambda = 3$ and initial neighborhoods U_0 of size $N_0 = 1$, 5, 5 and 13 for the four images, respectively.

Table 2 summarizes some statistics characterizing the quality of both the original image and its AWS estimate. We report the mean squared error (MSE) of the estimate and the number of points where the estimate differs from the true value by more than .125, i.e. minimal contrast divided by 4. The first characterizes the mean quality of the estimate while the second indicates quality of segmentation. Note the excellent behavior for $\sigma = .25$. With standard deviation increasing we first loose the most detailed structure ($\sigma = .5$), observe some loss in edge accuracy for the lower contrast level ($\sigma = 1$) and significant blurring of the edges for the highest noise level ($\sigma = 2$). Note that we still recover the main structure that is hardly visible in the noisy original.

The right image in Figure 6 contains $n = 128 \times 128$ pixel. It's structure is much simpler being composed of only two regions with value 0 and 1. We use this image to illustrate the properties of our estimate in case of large homogeneous regions and low signal to noise level. Figure 8 shows the estimates obtained from images with added noise of standard deviation $\sigma = .5$, 1, 2, 4, 6 and 8. Table 3 reports mean squared error and the number of points with estimate differing by more than $\alpha/4 = .25$ from the true value.

For signal to noise ratio $\alpha/\sigma \geq 1$ the segmentation is almost perfect. For $\alpha/\sigma \geq .25$ we still get a clear segmentation but loose the accuracy of the boundary estimate. In case of extremely small α/σ we still detect the presence of a structure, although the procedure performs unstable resulting in a severe blurring.

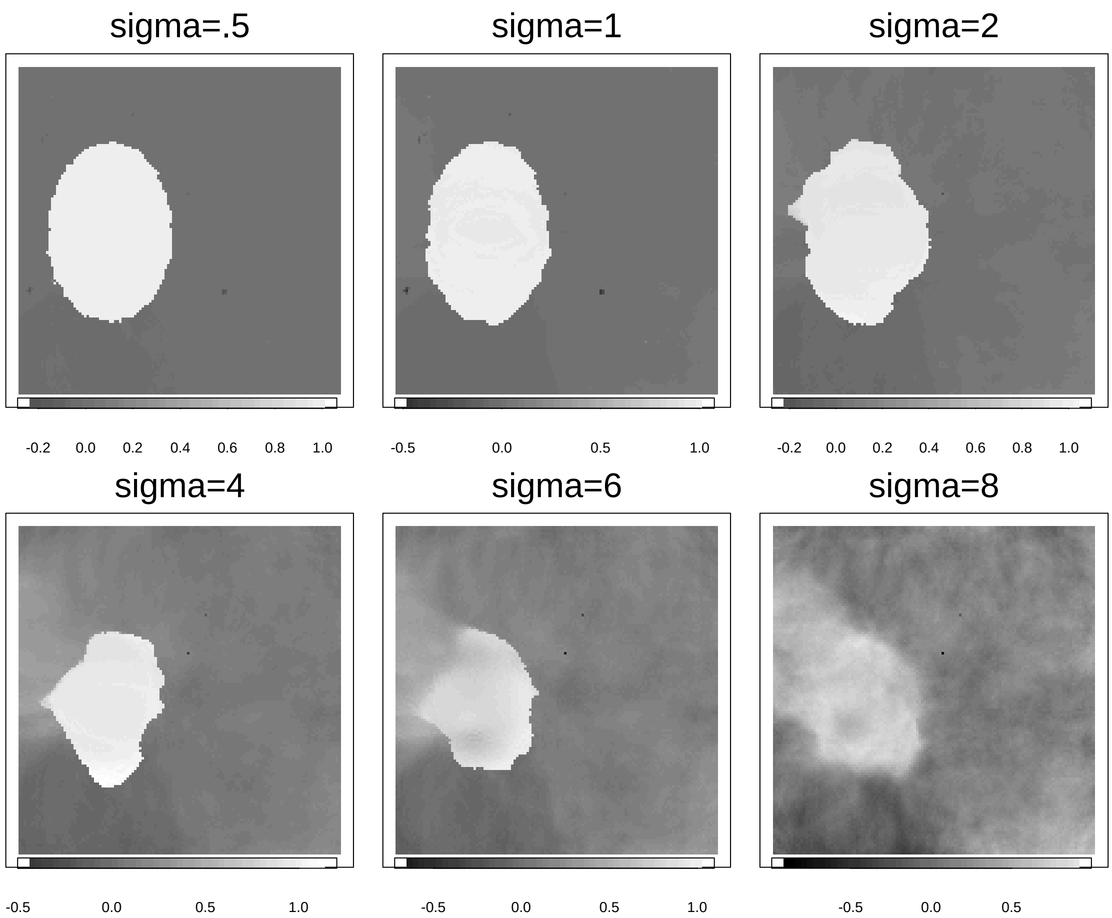

Figure 8: AWS reconstruction of image 2 for different values of σ

4.5 A bivariate simulation

Our second comparative study is based on the test image 1, see Figure 6.

We perform 100 simulations with error standard deviation of $\sigma = .25$, $.5$ and 1, respectively. Parameters of our procedure are again $\lambda = 3$, $U_0(X^0)$ being the point X^0 itself and U_{k^*} now containing the 709 closest design points. Smoothing parameters for the alternative procedures are again chosen approximately MSE-optimal. Table 4 provides the simulation results using the same criteria as in the univariate case.

The images in Figure 9 show the proportion of wrong classified points for each pixel in our test image. The columns contain results for our adaptive procedure, the gaussian filter, the nonlinear gauss filter and mode regression (from left to right), the rows correspond to the different noise levels $\sigma = .25$, $.5$ and 1 (from top to bottom). White regions correspond to a proportion of zero and black to a proportion of one. For the lowest noise level we observe that even the detailed structures in the upper right and lower left of the image are recovered by our method. The nonlinear gauss filter and mode estimation both work reasonable in this situation, providing an improvement to the gauss filter with

	Mean squared error			Wrong classifications		
	$\sigma = .25$	$\sigma = .5$	$\sigma = 1$	$\sigma = .25$	$\sigma = .5$	$\sigma = 1$
Adaptive weights	0.0021	0.0111	0.0334	475	2215	9277
Gauss filtering	0.0138	0.0243	0.0396	13889	20497	29642
Nonlinear gauss	0.0096	0.0262	0.0454	9871	21906	32173
Mode regression	0.0068	0.0254	0.0426	5119	18994	31400

Table 4: Estimated mean squared error and mean number of wrong classifications in the bivariate simulation experiment (image 1)

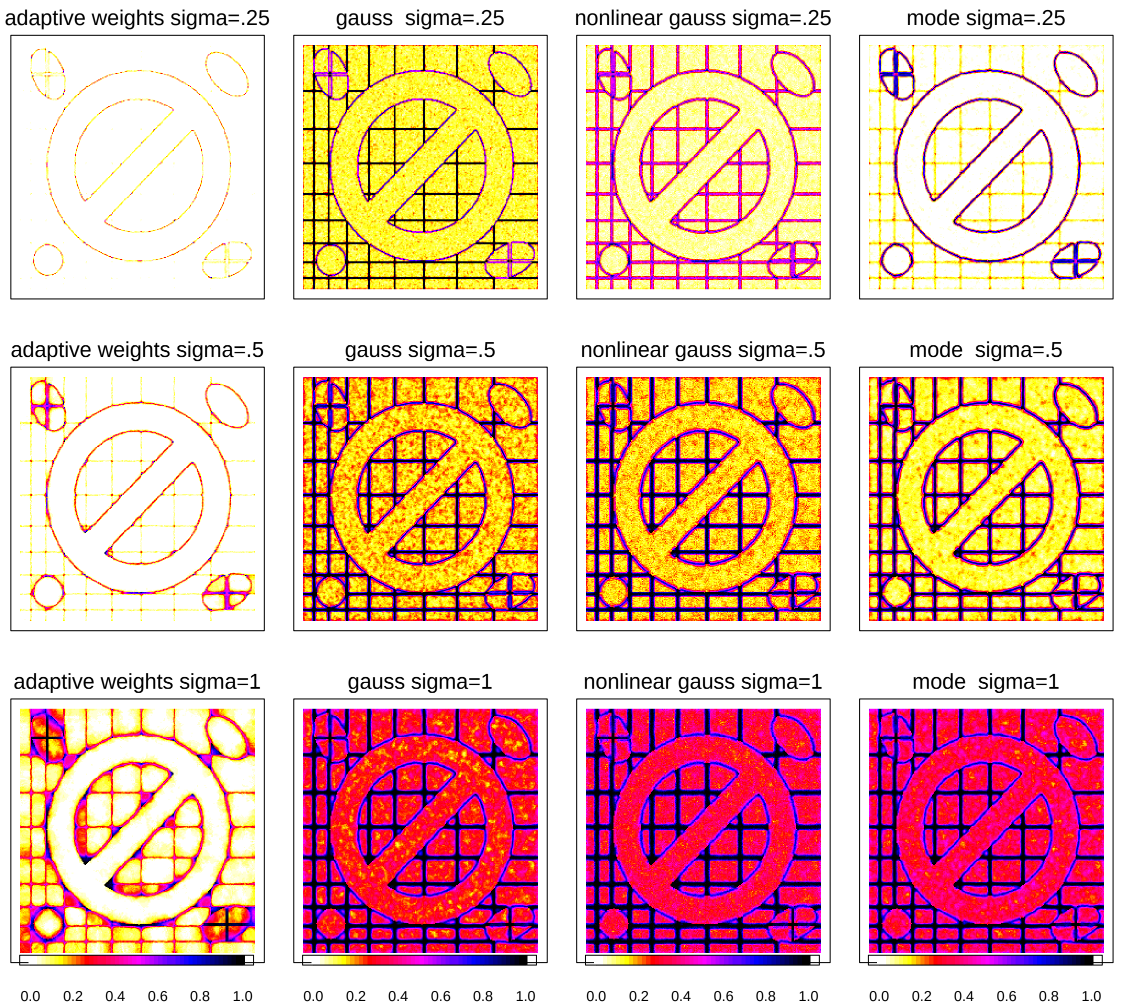

Figure 9: Proportion of wrong classifications estimated from 100 simulations for $\sigma = .25$, $.5$ and 1 in the bivariate simulation experiment.

respect to both criteria. Increasing σ we see a clear advantage of our procedure, the other three possessing almost identical behavior.

5 An Application to Magnetic Resonance Imaging

Our attention has been drawn to the analysis of Magnetic Resonance images (MRI) by colleagues from the Max-Plank-Institute of Cognitive Neuroscience at Leipzig. Magnetic Resonance imaging is a quite new technique of noninvasive analysis providing a delineation of a physical object. The signal, or true image, contained in the 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 Magnetic Resonance imaging see for instance Sebastiani and Barone (1991) and Lange (1996).

We here concentrate on the task of possessing two and three dimensional high resolution MRI although we have also successfully applied modifications of our ideas to problems in functional and dynamic MRI.

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). There is also a systematic error caused by truncating the Fourier series (Gibbs phenomenon), see again e.g. Sebastiani and Barone (1995).

We demonstrate the capabilities of AWS using a two dimensional high resolution MRI recorded at the Max-Plank-Institute of Cognitive Neuroscience at Leipzig. The image contains $n = 512 \times 512$ pixel. The original image is shown in the upper left plot of Figure 10. The upper right plot shows the estimated image obtained with parameters $\lambda = 3$, variance evaluated from the background, $U_0(x^0) = \{x^0\}$ and $U_{k^*}(x^0)$ containing up to 1005 pixel. The lower row of Figure 10 shows a residual image (left) and an image assigning the number of observations used in the estimate at a certain location to the corresponding pixel. The residual image shows almost no structure indicating that AWS preserves the signal. The right plot gives an illustration of the adaptivity of the estimate, large neighborhoods are used in homogeneous regions while only few points are involved in the averaging where the image is highly structured.

Our second 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). Note that

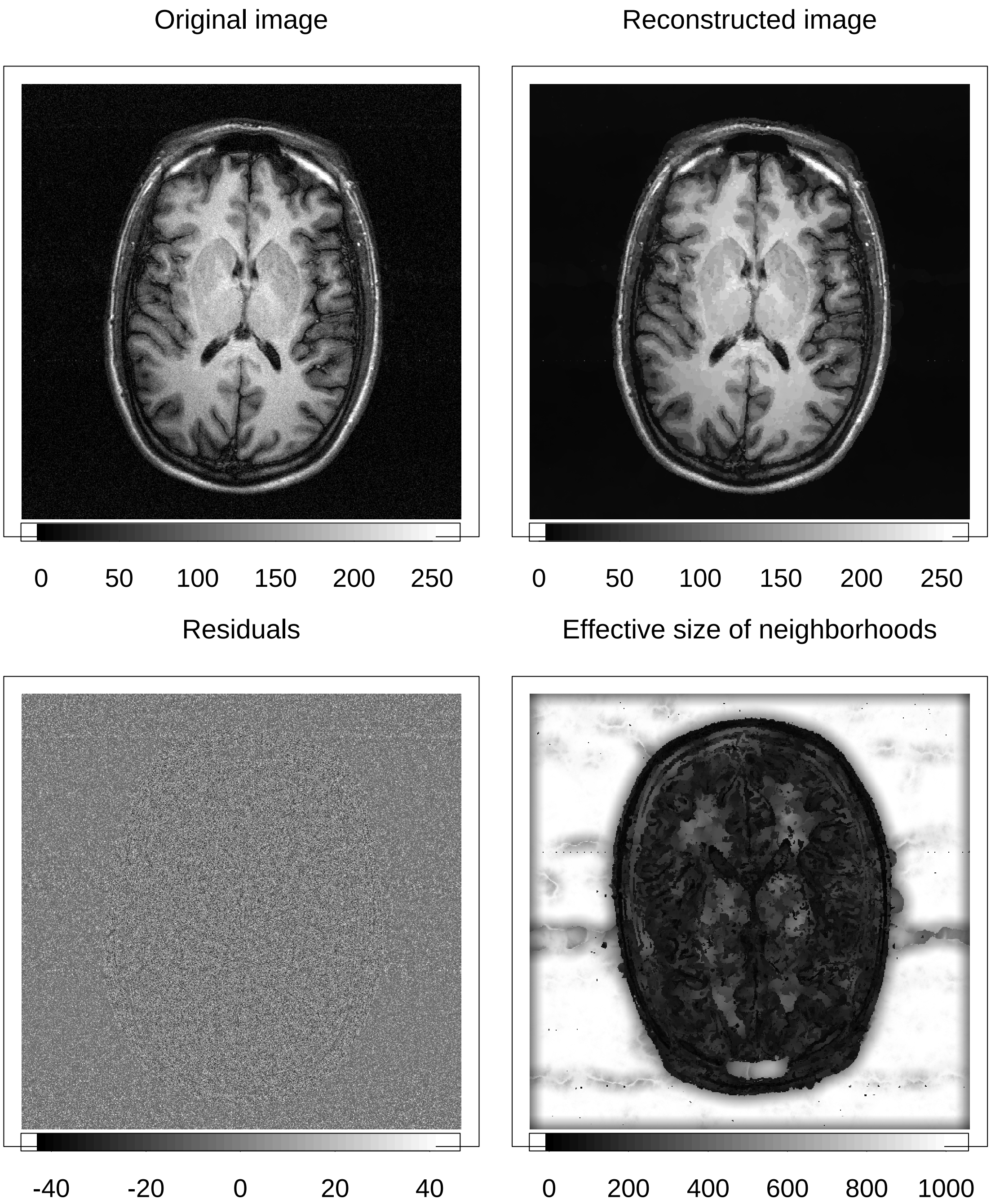

Figure 10: Original and AWS estimate of a Magnetic Resonance Image (upper row); residual image and number of points with positive weights for each pixel in the AWS estimate (lower row)

the original image shown in the upper left plot of Figure 11 is quite noisy. One well established method to reduce the noise level is to average several MR-Images recorded from the same slice of the brain. The result of averaging eight images is shown in the

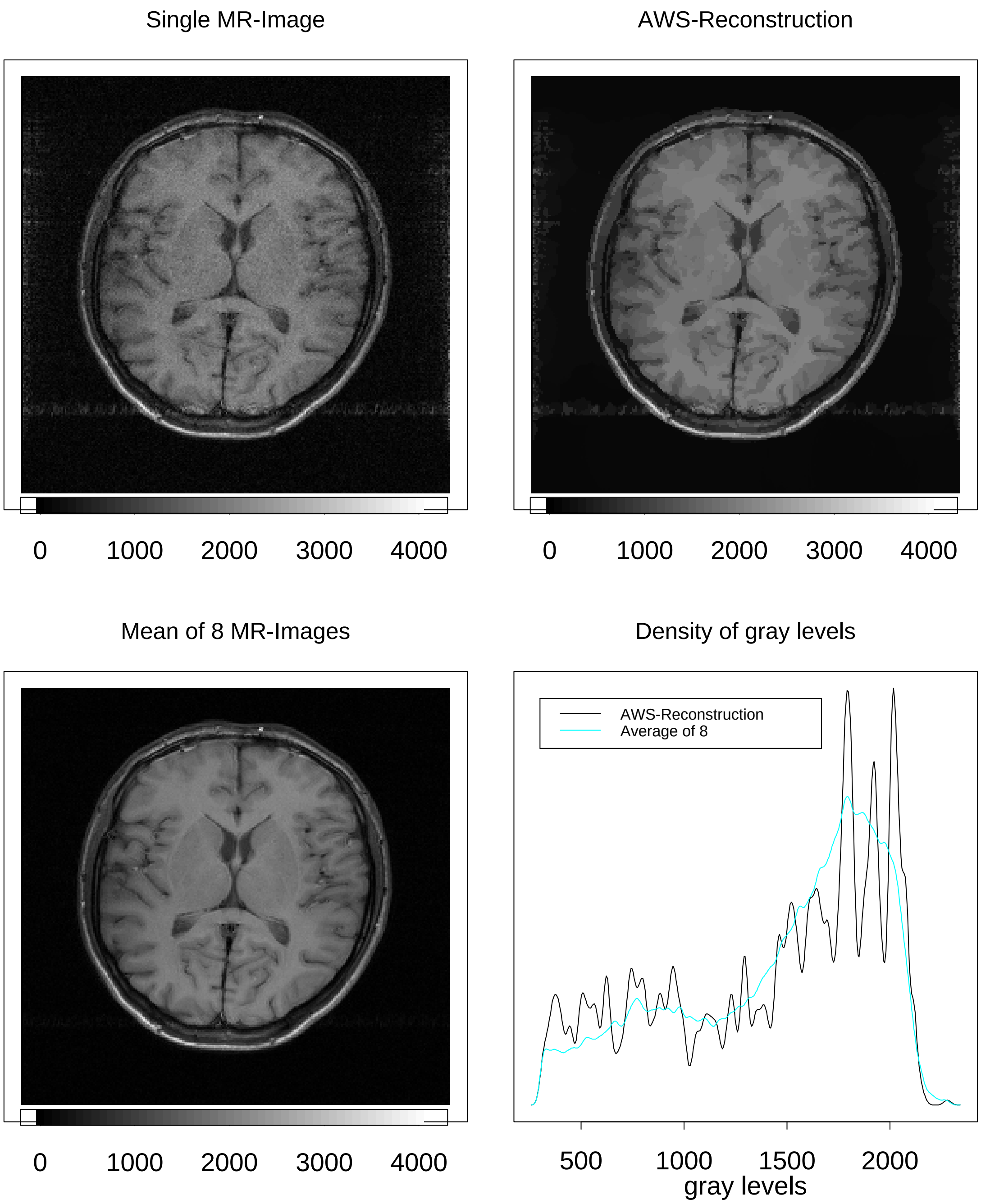

Figure 11: Original and AWS estimate of a Magnetic Resonance Image (upper row); average of eight images of the same slice and estimated densities of gray levels (lower row)

lower left plot. The upper right plot gives the estimate obtained by AWS with parameters $\lambda = 3$, $U_0(x_i) = \{x_i\}$ and $U_{k^*}(x_i)$ containing up to 797 points. Keeping in mind that a MR image exhibits the weighted spin density of nuclei from a rather small number of tissues we would expect this to be reflected in a density estimate of gray levels obtained from the reconstructions. The lower right plot of Figure 11 shows estimates of the density of gray levels obtained from both the average of eight images and the AWS estimate, omitting the gray levels corresponding to the background. Note that the density estimate based on the AWS reconstruction shows significant multimodality while the estimate based on the average of eight is rather smooth.

6 Further developments

Here we briefly discuss some possibilities to extend the method and the domain of applications.

First we mention that in the univariate case, the locally constant approximation does a good job only for functions of a special type like a telegraph signal, see Figure 2. It is well known that, for a general smooth curve, a locally linear approximation leads to an essentially better quality. The presented procedure can be generalized to the case of locally linear approximation but it becomes more complicated. Indeed, for each couple of points (X_i, X_j) we have to construct two locally linear estimates and then to compare them. One way of comparison is to extrapolate one linear fit at X_i to X_j and to compare with the linear fit at X_j. We do not present the results of simulations here but it is worth to mention that the procedure based on the locally linear approximation is more flexible but less stable than one based on the locally constant approximation, and the quality of the initial estimate $\widehat{f}_0$ is even more crucial in such a situation. A locally linear approximation applies for the multivariate case with $d \geq 2$ as well. However, for most applications to the problem of image segmentation, the locally constant approximation seems to be quite reasonable.

Another possible modification of the method is connected to the problem of the quality of the initial estimate $\widehat{f}_0$. We know from Proposition 3.2 that this quality determines the quality of image recognition and that it depends strongly on the size N_0 of the original neighborhoods $U_0(X_i)$. The choice of a large N_0 leads to oversmoothing near edges because this neighborhood is symmetric around X_i. On the other hand, for a small N_0 and small signal-to-noise ration, we have a high probability of a wrong classification. A natural idea would be to use a more sophisticated initial estimator which allows for a good quality of edge recognition. One possibility is to take a median estimator instead of mean over $U_0(X_i)$, like in Tsybakov (1989). Another choice might be to take for the initialization the procedure proposed in Polzehl and Spokoiny (1998). This procedure selects adaptively the geometry of the window $U_0(X_i)$ providing a near-optimal quality of edge recognition.

In this paper we discuss only the problem of function or image denoising and segmentation. In practical applications this is typically the first step of the analysis. At this point we briefly mention two other important classes of problems arising in magnetic resonance imaging: analysis of dynamic and functional MR images. For both cases, a time-series of images is observed, usually of a low resolution and with a lot of distortions. This means that at each design point X_i we are given a time series (curve) $Y_i(t)$, where t is a time parameter. Analysis of functional MRI is focusing on the detection of activations induced by a stimulus signal. Interest is in both an accurate delineation of the activated regions as well as an characterization of the induced signal. Such type of problems meets for instance human brain mapping. Dynamic MRI is actively used in medical applications, for instance, for detection of damaged regions in the brain. The dynamic characteristics are different for damaged and non-damaged regions and the problem can be viewed as curve classification. We can suggest two possible approaches to these problems based on the previously described procedure. One way is to treat the time parameter as the third coordinate and then estimate adaptively the image as a function of space and time. An even more convenient approach is to estimate separately each image in the series but to use adaptive weights which are evaluated on the base of the distance between two estimated curves. Our experience with simulated examples and real data gives an advance to the second approach which demonstrates a quite good performance in all situations.

For our results in Section 3 we assumed homogeneous Gaussian noise but similar results can be stated for heterogeneous and non-Gaussian noise as well. A generalization to a dependent noise is also possible under some mixing conditions.

An interesting question is a possibility to apply the method and the results for other statistical models like distribution density model or nonparametric autoregression etc. We do not discuss this question rigorously here.

References

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

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

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

[4] Chu, C. K., Glad I. K., Godtliebsen, F. and Marron, J. S. (1998). Edge preserving smoothers for image processing (with discussion). *Journal of the American Statistical Association* **93**, to appear.

[5] Donoho, D. L., Johnstone I. M., Kerkyacharian, G. and Picard, D. (1994). Wavelet shrinkage: asymptopia ?. *Journal of the Royal Statistical Society* , Ser B, **57**, 301–369.

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

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

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

[9] Hall, P. (1991a). On the distribution of suprema. *Probab. Theory Rel. Fields* **89**, 447–455.

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

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

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

[13] Lepski, O., Mammen, E. and Spokoiny, V. (1997). Ideal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selection. *Annals of Statistics*, **25**, no. 3, 929–947.

[14] Lepski, O. and Spokoiny, V. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Annals of Statistics*, **25**, no. 6, 2512–2546.

[15] Müller, H.-G. (1992). Change points in nonparametric regression analysis. *Annals of Statistics* **20**, 737–761.

[16] Polzehl, J. and Spokoiny, V. G. (1998). Image denoising: pointwise adaptive approach, submitted.

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

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

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

[20] Speckman, P. L. (1994), Fitting curves with features: Semiparametric change point methods, in *Comput. Sci. Statist: Computat. Intens. Statist. Meth: Proc. 26th Symp. Interface*; Sall, John; Lehman, Ann (Ed); Interface Foundat. North Amer. (Fairfax Station, VA), **26**, 257– 264.

[21] Spokoiny, V. (1996). Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. Preprint **291**, Weierstrass Institute, Berlin. *Ann. Statist.*, tentatively accepted.

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

[23] Venables, W. N. and Ripley, B. D. (1992), *Modern Applied Statistics with S-Plus*, Springer, New York.

[24] Wang, A. (1995). Jump and sharp cusp detection by wavelets. *Biometrika*, **82**, 385–397.