Local likelihood modeling by adaptive weights smoothing*

Polzehl, Jörg
Weierstrass-Institute,
Mohrenstr. 39, 10117 Berlin, Germany
polzehl@wias-berlin.de

Spokoiny, Vladimir
Weierstrass-Institute,
Mohrenstr. 39, 10117 Berlin, Germany
spokoiny@wias-berlin.de

Abstract

The paper presents a unified approach to local likelihood estimation for a broad class of nonparametric models, including e.g. the regression, density, Poisson and binary response model. The method extends the adaptive weights smoothing (AWS) procedure introduced in Polzehl and Spokoiny (2000) in context of image denoising. Performance of the proposed procedure is illustrated by a number of numerical examples and applications to density or volatility estimation, classification and estimation of the tail index parameter. We also establish a number of important theoretical results on properties of the proposed procedure.

Keywords: adaptive weights, local likelihood, exponential family, density estimation, volatility, classification, tail index

AMS 2000 Subject Classification: 62G05, Secondary: 62G07, 62G08, 62G32, 62H30

*This work was partially supported by the DFG Research Center *Mathematics for key technologies* and the DFG Priority program 1114 *Mathematical methods for time series analysis and digital image processing.*

1 Introduction

Local modeling is one of the most useful nonparametric methods. We refer to the book by Fan and Gijbels (1996) for a rigorous discussion of local linear and local polynomial estimation for regression and some other statistical models and many other references. An extension to the local likelihood approach is discussed in Tibshirani and Hastie (1987), Staniswalis (1989), Loader (1996), among others.

This paper proposes a new approach to local likelihood modeling which is based on the idea of structural adaptation and extends the *Adaptive Weights Smoothing* (AWS) procedure from Polzehl and Spokoiny (2000) (referred to as PS2000). The main idea of AWS is to describe in a data-driven way a maximal local neighborhood of every point in which the local parametric assumption is justified by the data. The method is based on a successive increase of local neighborhoods around every point X_i and a description of the local model within such neighborhoods by assigning weights that depend on the result of the previous step of the procedure. The original AWS procedure was proposed for the regression model in the context of image denoising. The numerical results from PS2000 demonstrate that the AWS method is very efficient in situations where the underlying regression function allows a piecewise constant approximation with large homogeneous regions. The procedure possesses a number of remarkable properties like preservation of edges and contrasts and nearly optimal noise reduction inside large homogeneous regions. It is dimension free and applies in high dimensional situations. However, the assumption of the regression model with additive errors considered in PS2000 restricts its domain of applications. Here we extend the approach from PS2000 to a broad class of nonparametric models including the binary response model, inhomogeneous exponential and Poisson models etc. having local exponential family structure and apply the AWS method in a unified way to different problems like density or intensity estimation, volatility modeling, classification, tail index estimation and establish some remarkable theoretical results on properties of the proposed procedure.

A reference implementation of our algorithms is available as a contributed package (aws) of the R-Project for Statistical Computing from http://www.r-project.org/ .

The paper is organized as follows. Section 2 describes the model and presents the main examples. Local modeling is discussed in Section 3. The local likelihood AWS procedure is introduced in Section 4. Section 5 demonstrates how the AWS method can be used to estimate a density in $\mathbb{R}^d$ for $d \leq 3$. Section 6 explains how AWS can be ap-

plied to volatility estimation of financial assets. The classification problem is considered in Section 7. Section 8 introduces an AWS-estimate of the tail-index parameter. Section 9 discusses main properties of the proposed method, among them the "propagation condition" and the rate of estimation of a smoothly varying parameter. Some technical assertions about the varying coefficient exponential family are collected in the Appendix.

2 Model and problem

This section describes the considered model and states the estimation problem. Suppose we are given random data $Z_1, \ldots, Z_n$ of the form $Z_i = (X_i, Y_i)$. Here every X_i means a vector of "features" or explanatory variables which determines the distribution of the "observation" Y_i. For simplicity we suppose that the X_i's are valued in the finite dimensional Euclidean space $\mathcal{X} = I\!R^d$ and the Y_i's belong to $\mathcal{Y} \subseteq I\!R$. An extension to the case when both the X_i's and Y_i's are valued in some metric spaces is straightforward. The vector X_i can be viewed as a location and Y_i as the "observation at X_i". For ease of exposition, we restrict ourselves to the case of independent Z_i. Our model assumes that the distribution of each Y_i is determined by a finite dimensional parameter θ which may depend on the location X_i, $\theta = \theta(X_i)$. We illustrate this set-up using a few examples.

Example 2.1. (Gaussian regression) Let $Z_i = (X_i, Y_i)$ with $X_i \in I\!R^d$ and $Y_i \in I\!R$ following the regression equation $Y_i = \theta(X_i) + \varepsilon_i$ with a regression function θ and i.i.d. Gaussian errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Example 2.2. (Inhomogeneous Bernoulli (Binary Response) model) Let again $Z_i = (X_i, Y_i)$ with $X_i \in I\!R^d$ and Y_i a Bernoulli r.v. with parameter $\theta(X_i)$, that is, $\boldsymbol{P}(Y_i = 1 \mid X_i = x) = \theta(x)$ and $\boldsymbol{P}(Y_i = 0 \mid X_i = x) = 1 - \theta(x)$. Such models arise in many econometric applications, they are widely used in classification and digital imaging.

Example 2.3. (Inhomogeneous Exponential model) Suppose that every Y_i is exponentially distributed with the parameter $\theta = \theta(X_i)$, that is, $\boldsymbol{P}(Y_i > t \mid X_i = x) = e^{-t/\theta(x)}$. Such models are applied in reliability or survival analysis. They also naturally appear in the tail-index estimation theory.

Example 2.4. (Inhomogeneous Poisson model) Suppose that every Y_i is valued in the set $\mathbb{N}$ of nonnegative integer numbers and $\boldsymbol{P}(Y_i = k \mid X_i) = \theta^k(X_i) e^{-\theta(X_i)}/k!$, that is, Y_i follows a Poisson distribution with parameter $\theta = \theta(X_i)$. This model is commonly used in the queueing theory, it occurs in positron emission tomography, it also serves as an approximation of the density model, obtained by a binning procedure.

Example 2.5. (Inhomogeneous volatility model) The observations Y_t follow a conditional heteroscedastic model $Y_t = \sigma_t \varepsilon_t$ in the discrete time t. The ε_t's are independent standard normal innovations and σ_t is a time dependent parameter (volatility).

All these examples are particular cases of the local exponential family model, see Section 3.2 for more details.

Now we present a formal definition for our model. Let $\mathcal{P} = (P_\theta, \theta \in \Theta)$ be a family of probability measures on $\mathcal{Y}$ where Θ is a subset of the real line $\mathbb{R}^1$. We assume that this family is dominated by a measure P and denote $p(y, \theta) = dP_\theta / dP(y)$. Moreover, we assume that all the densities $p(y, \theta) = dP_\theta / dP(y)$ for $\theta \in \Theta$ are strictly positive, $p(y, \theta) > 0$ for all $y \in \mathcal{Y}$ and $\theta \in \Theta$. We suppose that each Y_i is, conditionally on $X_i = x$, distributed with the density $p(\cdot, \theta(x))$ for some unknown function $\theta(x)$ on $\mathcal{X}$. The aim of the data-analysis is to infer on this function $\theta(x)$.

A standard approach is based on the assumption that the function θ is smooth leading to its local linear (polynomial) approximation within a ball of some small radius h centered in the point of estimation, see e.g. Tibshirani and Hastie (1987), Hastie and Tibshirani (1993), Fan and Zhang (1999), Caroll et.al. (1998), Cai et.al. (2000). This approach has serious problems and has to be substantially extended when functions with discontinuities are considered, see e.g. Müller (1992) or Spokoiny (1998) for a univariate case $\mathcal{X} = \mathbb{R}^1$ and Müller and Song (1994), Qiu (1998), Polzehl and Spokoiny (2003a) for the bivariate case with $d = 2$. Local estimation near discontinuity requires to consider asymmetric neighborhoods of the point of estimation. In the univariate case one can apply one-sided neighborhoods but in the bivariate case the shape of the neighborhood should be quite flexible to provide the optimal rate of estimation, Polzehl and Spokoiny (2003a). Here we apply a completely different approach to estimation of such functions which is based on the idea of *adaptive weights*: the shape of the local model centered at a point x is described by weights which are computed in a data-driven way. This helps to consider in an unified way the models with smoothly varying parameters and "piecewise smooth" models whose parameters may jump with locations. The global parametric model is also naturally incorporated in this framework.

3 Local likelihood modeling

A global parametric structure simply means that the parameter θ does not depend on the location, that is, the distribution of every "observation" Y_i coincides with P_θ for

some $\theta \in \Theta$ and all i. This assumption reduces the original problem to the classical parametric situation and the well developed parametric theory applies here for estimating the underlying parameter θ. In the sequel we consider the parametric maximum likelihood estimate $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ of θ which is defined by maximization of the log-likelihood

$$\widehat{\theta} = \operatorname*{argsup}_{\theta \in \Theta} \sum_{i=1}^{n} \log p(Y_i, \theta).$$

However, a global parametric assumption can be too restrictive. The classical nonparametric approach is based on the idea of localization: for every point x, the parametric assumption is only fulfilled locally in a vicinity of x. This leads to considering a local model concentrated in some neighborhood of the point x.

3.1 Localization

We use *localization by weights* as a general method to describe this local model. Let, for a fixed x, a nonnegative weight $w_i = w_i(x) \leq 1$ be assigned to the observations Y_i at X_i, $i = 1, \ldots, n$. The weights $w_i(x)$ determine a local model corresponding to the point x in the sense that, when estimating the local parameter $\theta(x)$, every observation Y_i is used with the weight $w_i(x)$. This leads to the local (weighted) maximum likelihood estimate

$$\widehat{\theta}(x) = \operatorname*{arginf}_{\theta \in \Theta} \sum_{i=1}^{n} w_i(x) \log p(Y_i, \theta). \tag{3.1}$$

We mention two examples of choosing the weights $w_i(x)$. *Localization by a bandwidth* is defined by weights of the form $w_i(x) = K_{\mathrm{loc}}(\boldsymbol{l}_i)$ with $\boldsymbol{l}_i = |\rho(x, X_i)/h|^2$ where h is a bandwidth, $\rho(x, X_i)$ is the Euclidean distance between x and the design point X_i and K_{loc} is a *location kernel*.

Localization by a window simply restricts the model to a subset (window) $U = U(x)$ of the design space which depends on x, that is, $w_i(x) = \mathbf{1}(X_i \in U(x))$. Observations Y_i with X_i outside the region $U(x)$ are not used when estimating the value $\theta(x)$. This kind of localization arises e.g. in the regression tree approach.

We do not assume any special structure for the weights $w_i(x)$, that is, any configuration of weights is allowed. In what follows we identify the set $W(x) = \{w_1(x), \ldots, w_n(x)\}$ and the local model in x described by these weights and use the notation

$$L(W(x), \theta) = \sum_{i=1}^{n} w_i(x) \log p(Y_i, \theta).$$

Then $\widehat{\theta}(x) = \operatorname{argsup}_\theta L(W(x), \theta)$.

In our procedure we consider a family of local models, one per design point X_i, and denote them as $W_i = W(X_i) = \{w_{i1}, \ldots, w_{in}\}$.

3.2 Local exponential family

The examples introduced in Section 2 can be considered as particular cases of local exponential family distributions. This means that all measures P_θ from this family are dominated by a σ-finite measure P on $\mathcal{Y}$. The density functions $p(y, \theta) = dP_\theta / dP(y)$ are of the form $p(y, \theta) = e^{U(y)C(\theta) - B(\theta)}$. $C(\theta)$ and $B(\theta)$ are some given nonnegative functions. $U(y)$ is a known function of the observation y. The parameter θ is defined by the equations $\int p(y, \theta) P(dy) = 1$ and $\boldsymbol{E}_\theta U(Y) = \int U(y) p(y, \theta) P(dy) = \theta$. The functions $B(\theta)$ and $C(\theta)$ are connected by the differential equation $B'(\theta) = \theta C'(\theta)$. The Kullback-Leibler distance $Q(\theta, \theta') = \boldsymbol{E}_\theta \log\bigl(p(Y, \theta)/p(Y, \theta')\bigr)$ for $\theta, \theta' \in \Theta$ satisfies

$$\begin{aligned} Q(\theta, \theta') &= \bigl(C(\theta) - C(\theta')\bigr) \int U(y) p(y, \theta) P(dy) - \bigl(B(\theta) - B(\theta')\bigr) \\ &= \theta\bigl(C(\theta) - C(\theta')\bigr) - \bigl(B(\theta) - B(\theta')\bigr). \end{aligned}$$

Next, for a given set of weights $W = \{w_1, \ldots, w_n\}$, it holds

$$L(W, \theta) = \sum_{i=1}^n w_i \log p(Y_i, \theta) = C(\theta)\bigr) \sum_{i=1}^n w_i U(Y_i) - B(\theta)\bigr) \sum_{i=1}^n w_i = SC(\theta) - NB(\theta)$$

where $N = \sum_{i=1}^n w_i$ and $S = \sum_{i=1}^n w_i U(Y_i)$. Maximization of this expression w.r.t. θ leads to the estimating equation $NB'(\theta) - SC'(\theta) = 0$. This and the identity $B'(\theta) = \theta C'(\theta)$ yield the local MLE

$$\widehat{\theta} = S/N = \sum_{i=1}^n w_i U(Y_i) \Big/ \sum_{i=1}^n w_i \,.$$

This implies $L(W, \widehat{\theta}) = N\bigl\{\widehat{\theta} C(\widehat{\theta}) - B(\widehat{\theta})\bigr\}$ and $L(W, \widehat{\theta}) - L(W, \theta) = NQ(\widehat{\theta}, \theta)$ for any $\theta \in \Theta$. Table 1 provides the statistics $U(y)$ and the Kullback-Leibler distance $Q(\theta, \theta')$ for the examples from Section 2.

The procedure presented in the next section is effectively based on assigning some measure of inhomogeneity for two different local models. We now discuss how this measure can be naturally defined via likelihood ratio tests of homogeneity for two populations.

3.3 Comparing the parameters of two local models

Consider two local models corresponding to points X_i and X_j and defined by weights $W_i = \{w_{i1}, \ldots, w_{in}\}$ and $W_j = \{w_{j1}, \ldots, w_{jn}\}$. Suppose for a moment that a *structural*

Table 1: $U(y)$ and $Q(\theta, \theta')$ for the examples from Section 2.

Model	$U(y)$	$Q(\theta, \theta')$
Gaussian regression	y	$(\theta - \theta')^2/(2\sigma^2)$
Bernoulli model	y	$\theta \log(\theta/\theta') + (1-\theta)\log\{(1-\theta)/(1-\theta')\}$
Exponential model	y	$\theta/\theta' - 1 - \log(\theta/\theta')$
Poisson model	y	$\theta \log(\theta/\theta') - (\theta - \theta')$
Volatility model	y^2	$0.5(\theta/\theta' - 1 - \log(\theta/\theta'))$

assumption is fulfilled in both local models, that is, the parameter function $\theta(\cdot)$ is nearly constant within the local model in x, i.e. $\theta(X_k) \approx \theta(x)$ if $w_k(x) > 0$. We aim to answer the question whether these two local models can be put into one common parametric model. This can be done by testing the hypothesis that the parameter values $\theta_i = \theta(X_i)$ and $\theta_j = \theta(X_j)$ for the corresponding two local models coincide.

To compare the parameters of two local models W_i and W_j we utilize the likelihood-ratio test statistic. First we consider the situation when both sets W_i and W_j have zero-one entries with positive elements at disjoint positions, that is, the values w_{ik} and w_{jk} and $w_{ik} + w_{jk}$ are either zero or one for all k. This situation corresponds to the two sample problem in which one sample consists of the observations Y_k with $w_{ik} = 1$ and the other one contains the observations Y_k with $w_{jk} = 1$. The classical likelihood-ratio test statistic for the hypothesis $\theta_i = \theta_j$ for this situation is of the form

$$\begin{aligned} T_{ij}^{\circ} &= \max_{\theta} L(W_i, \theta) + \max_{\theta} L(W_j, \theta) - \max_{\theta} L(W_i + W_j, \theta) \\ &= L(W_i, \widehat{\theta}_i) + L(W_j, \widehat{\theta}_j) - L(W_i + W_j, \widehat{\theta}_{ij}) \end{aligned} \tag{3.2}$$

where $\widehat{\theta}_{ij} = \operatorname{argsup}_\theta L(W_i + W_j, \theta)$ is the maximum likelihood estimate from a combined model obtained by adding the weights from both models. The value T_{ij}° characterizes the difference between the two models: if T_{ij}° is larger than some prescribed value λ, then the parameters of these two models are significantly different. A critical level λ can be assigned using the Wilks phenomenon, see Fan et.al. (2001). It means that under the parametric hypothesis $\theta(X_k) \equiv \theta$ the distribution of $2L(W_i, \widehat{\theta}_i) - 2L(W_i, \theta)$ is asymptotically χ^2 with one degree of freedom as the "sample size" $N_i = \sum_j w_{ij}$ grows to infinity. Section 11 (Theorem 11.1) presents a nonasymptotic extension of the Wilks results which applies to a small sample size and arbitrary weights.

Note that the value T_{ij}° is "symmetric" w.r.t. W_i and W_j in the sense that $T_{ij}^{\circ} = T_{ji}^{\circ}$.

In our procedure, described in the next section, we apply a slightly modified asymmetric version of this test statistic, namely

$$T_{ij} = L(W_i, \widehat{\theta}_i) - L(W_i, \widehat{\theta}_j).$$

It has a nice interpretation as a difference between the maximum log-likelihood $L(W_i, \widehat{\theta}_i) = \sup_\theta L(W_i, \theta)$ in model W_i and the "plug-in" log-likelihood $L(W_i, \widehat{\theta}_j)$ in which $\widehat{\theta}_j$ comes from the model W_j. This modification is important because T_{ij} is used for defining the weight w_{ij} with which the observation Y_j at X_j will enter in the local model W_i corresponding to X_i. However, in the "balanced" situation when the "sample sizes" N_i and N_j are of the same order, the values T°_{ij} and T_{ij} have similar properties.

For the local exponential family with a varying parameter we obtain due to Section 3.2

$$T_{ij} = N_i Q(\widehat{\theta}_i, \widehat{\theta}_j). \tag{3.3}$$

This representation is used for the procedure described in the next section.

In our procedure, described below, we consider the value T_{ij} as a "statistical penalty", that is, when computing the new weight w_{ij} at the next iteration step we strongly penalize for a large value of T_{ij}.

We also consider a "symmetrized" version of T_{ij} given by $T^s_{ij} = (T_{ij} + T_{ji})/2$.

4 Adaptive weights smoothing

This section presents the estimation procedure. We start with some heuristic discussion.

The basic assumption of the proposed approach is that for every point X_i, there exists a local model described by weights W_i in which the parametric assumption is nearly fulfilled, that is, the difference $\theta(X_j) - \theta(X_i)$ is insignificant for all points X_j with significantly positive weights w_{ij}. The procedure tries to recover these weights from the data for all local models simultaneously in an iterative way.

We first illustrate this idea for the nonparametric regression with a local constant structural assumption as considered in PS2000. In that case the parameter θ coincides with the function value $f(X_i)$ and the estimate $\widehat{f}(X_i)$ is defined as the mean of the observations Y_j with some weights w_{ij}:

$$\widehat{f}(X_i) = \sum_{j=1}^n w_{ij} Y_j \Big/ \sum_{j=1}^n w_{ij}\,. \tag{4.1}$$

These weights w_{ij} are calculated iteratively, so that the estimate from the previous iteration is used to determine the new weights w_{ij} that in turn lead to the new estimates $\widehat{f}(X_i)$ due to (4.1). At the beginning of the iteration process the weights w_{ij} are taken in the form $w_{ij} = K_{\mathrm{loc}}(\boldsymbol{l}_{ij}^{(0)})$ where $\boldsymbol{l}_{ij}^{(0)} = \left|\rho(X_i, X_j)/h^{(0)}\right|^2$ and K_{loc} is a location kernel. If $K_{\mathrm{loc}} = \mathbf{1}(u \le 1)$ as in PS2000, then for every point X_i the weights w_{ij} vanish outside the ball $U_i^{(0)}$ of radius $h^{(0)}$ with the center at X_i, that is, the local model at X_i is concentrated on $U_i^{(0)}$. Next, at each iteration k, a ball $U_i^{(k)}$ with a larger bandwidth $h^{(k)}$ is considered and every point X_j from $U_i^{(k)}$ gets a weight $w_{ij}^{(k)}$ which is defined by comparing the estimates $\widehat{f}^{(k-1)}(X_i)$ and $\widehat{f}^{(k-1)}(X_j)$ obtained in the previous iteration. One possible interpretation of this procedure is that at each iteration step the *location penalty* $\boldsymbol{l}_{ij}^{(k)}$ is relaxed by increasing $h^{(k)}$ at cost of introducing a data-driven *statistical penalty* which comes from comparison of parameters of different local models.

An extension of this approach to the more general local parametric assumption leads to the comparison of the estimated parameters $\widehat{\theta}_i^{(k-1)}$ and $\widehat{\theta}_j^{(k-1)}$. Obviously the main ingredient of the proposed procedure is the precise way of defining the weights $w_{ij}^{(k)}$.

4.1 Definition of weights

For every pair (i, j), the weight $w_{ij}^{(k)}$ at the kth iteration of the procedure is computed on the base of two quantities: a location penalty $\boldsymbol{l}_{ij}^{(k)} = |\rho(X_i, X_j)/h^{(k)}|^2$ and a statistical penalty $\boldsymbol{s}_{ij}^{(k)} = T_{ij}/\lambda$, see (3.3). It is natural to require that each of these two penalties has an independent influence on the weight $w_{ij}^{(k)}$. This suggests to define the new weight $w_{ij}^{(k)}$ using the product

$$\widetilde{w}_{ij}^{(k)} = K_l\big(\boldsymbol{l}_{ij}^{(k)}\big) K_s\big(\boldsymbol{s}_{ij}^{(k)}\big),$$

where K_{loc} and K_{st} are two kernel functions on the positive semiaxis.

In the algorithm presented below in this section, we use one more (memory) parameter $\eta \in (0, 1)$ which controls the rate of changing the weights for every local model within the iteration process. Namely, we define the new weight $w_{ij}^{(k)}$ as a convex combination $\eta w_{ij}^{(k-1)} + (1-\eta)\widetilde{w}_{ij}^{(k)}$ of the weight $w_{ij}^{(k-1)}$ from the previous iteration step and the just computed value $\widetilde{w}_{ij}^{(k)}$.

Finally, to avoid an identification problem, we initialize the procedure by letting $\widehat{\theta}_i^{(0)} = \widehat{\theta}$ where $\widehat{\theta}$ is the global MLE.

4.2 The procedure

Now we present a formal description of the procedure. The procedure formally applies to any parametric family $\mathcal{P} = \{P_\theta\}$, however, it can be significantly simplified for the case of an exponential family $\mathcal{P}$. We therefore present a general definition and simultaneously give the formulas for the exponential family case. We denote $U_i = U(Y_i)$ with the function U from the definition of the exponential family, see Section 3.2.

Important ingredients of the method are: kernels K_{loc} and K_{st}, parameters λ and η, the initial bandwidth $h^{(1)}$, the factor $a > 1$ and the maximal bandwidth h^*. The choice of the parameters is discussed in Section 4.3. The procedure reads as follows:

1. Initialization: Compute the global MLE $\widehat{\theta}^{(0)}$ of θ:

$$\widehat{\theta}^{(0)} = \operatorname*{argsup}_{\theta\in\Theta} \sum_{i=1}^{n} \log p(Y_i, \theta) = \sum_{i=1}^{n} U_i/n.$$

For every i, set $\widehat{\theta}_i^{(0)} = \widehat{\theta}^{(0)}$, $N_i^{(0)} = n$ and define $W_i^{(0)} = (1, \ldots, 1)$. Set $k = 1$.

2. Iteration: for every $i = 1, \ldots, n$

- **Calculate the adaptive weights:** For every point X_j, compute the penalties

$$\begin{aligned} \boldsymbol{l}_{ij}^{(k)} &= \left|\rho(X_i, X_j)/h^{(k)}\right|^2, \\ \boldsymbol{s}_{ij}^{(k)} &= \lambda^{-1} T_{ij}^{(k)} = \lambda^{-1}\Big\{L\big(W_i^{(k-1)}, \widehat{\theta}_i^{(k-1)}\big) - L\big(W_i^{(k-1)}, \widehat{\theta}_j^{(k-1)}\big)\Big\} \\ &= \lambda^{-1} N_i^{(k-1)} Q\big(\widehat{\theta}_i^{(k-1)}, \widehat{\theta}_j^{(k-1)}\big). \end{aligned} \tag{4.2}$$

  Alternatively, the "symmetrized" statistical penalty $\boldsymbol{s}_{ij}^{(k)} = \lambda^{-1}(T_{ij}^{(k)} + T_{ji}^{(k)})/2$ can be used. Now compute

$$\widetilde{w}_{ij}^{(k)} = K_{\mathrm{loc}}\big(\boldsymbol{l}_{ij}^{(k)}\big) K_{\mathrm{st}}\big(\boldsymbol{s}_{ij}^{(k)}\big)$$

  and define the weight $w_{ij}^{(k)} = \eta w_{ij}^{(k-1)} + (1-\eta)\widetilde{w}_{ij}^{(k)}$.

  Denote $W_i^{(k)} = \{w_{i1}^{(k)}, \ldots, w_{in}^{(k)}\}$, and similarly $\widetilde{W}_i^{(k)} = \{\widetilde{w}_{i1}^{(k)}, \ldots, \widetilde{w}_{in}^{(k)}\}$.

- **Estimation:** Compute the new local MLE estimate $\widehat{\theta}_i^{(k)}$ of θ_i and the value $N_i^{(k)}$:

$$\widehat{\theta}_i^{(k)} = \operatorname*{argsup}_{\theta\in\Theta} L(W_i^{(k)}, \theta) = S_i^{(k)}/N_i^{(k)}$$

  with

$$\begin{aligned} N_i^{(k)} &= \sum_{j=1}^{n} w_{ij}^{(k)} = \eta N_i^{(k-1)} + (1-\eta)\sum_{j=1}^{n} \widetilde{w}_{ij}^{(k)}, \\ S_i^{(k)} &= \sum_{j=1}^{n} w_{ij}^{(k)} U_j = \eta S_i^{(k-1)} + (1-\eta)\sum_{j=1}^{n} \widetilde{w}_{ij}^{(k)} U_j. \end{aligned}$$

3. Stopping: Stop if $ah^{(k)} > h^*$, otherwise increase k by 1, set $h^{(k)} = ah^{(k-1)}$ and continue with step 2.

4.3 Choice of parameters

The parameters of the generalized AWS method are selected similarly to PS2000. We briefly discuss each of the parameters.

Kernels $K_{\rm st}$ and $K_{\rm loc}$: The kernels $K_{\rm st}$ and $K_{\rm loc}$ should be nonnegative with $K_{\rm st}$ decreasing and $K_{\rm loc}$ non-increasing on the positive semiaxis. We recommend to take $K_{\rm st}(u) = e^{-u} I_{\{u \le 6\}}$ and to apply a compactly supported localization kernel $K_{\rm loc}$ to reduce the computational effort of the method. PS2000 used a uniform kernel, here we apply the triangle kernel $K_{\rm loc}(u) = (1-u)_+$. As for kernel or local polynomial smoothing the choice of the kernel has only a minor influence on the final results.

Parameter η: The value $\eta \in (0,1)$ can be used to control the stability of the AWS procedure w.r.t. iterations. An increase of η results in a higher stability, however, it decreases sensitivity to changes of the local structure. A value $\eta > 0$ also guarantees that $Q(\widehat{\theta}_i, \widehat{\theta}_j) < \infty$, so it also serves as regularization parameter. Our default choice is $\eta = 1/2$ although the results change only slightly for η in the range $[0, 1/2]$.

Initial bandwidth $h^{(1)}$, parameter a and maximal bandwidth h^*: The initial bandwidth $h^{(1)}$ should be reasonably small. In most examples we select $h^{(1)} = c/n$ with some c ensuring that every ball $U_i^{(1)}$ with center X_i and radius $h^{(1)}$ contains only the design point X_i. The parameter a controls the growth rate of the local neighborhoods for every point X_i. It should be selected to provide that the mean number of points inside a ball $U_i^{(k)}$ with radius $h^{(k)}$ grows exponentially with k with some factor $a_{grow} > 1$. If X_i are from $\mathbb{R}^d$, then the parameter a can be taken as $a = a_{grow}^{1/d}$. Our default choice is $a_{grow} = 1.25$. Any value in the range $[1.1, 1.3]$ can be taken as well.

The maximal bandwidth h^* can be taken large so that every ball $U_i^{(k)}$ contains the whole sample for the last iteration k and the location penalty nearly vanishes. The parameter h^* can be used to bound the numerical complexity of the procedure, see Section 4.4 below. In some application examples, the use of a very large final bandwidth h^* leads to some oversmoothing of the underlying object. For such situations, a data-driven method of optimal stopping, based, for instance, on cross-validation can be applied.

The geometric grow of the parameter h ensures that the total number of iterations is typically bounded by $C \log n$ for some fixed constant C.

Symmetric and asymmetric versions: In most examples, the results for the symmetric and asymmetric versions of the procedure are very close to each other. The symmetric version is preferable if fine structures in the model should be kept, while the asymmetric version tends to oversmooth such fine structures but performs more stable within large homogeneous regions. Our default choice is the symmetric procedure.

Parameter λ: The most important parameter of the procedure is λ which scales the statistical penalty s_{ij}. Small values of λ lead to overpenalization which may results in a random segmentation of a homogeneous target. Large values of λ may result in loss of adaptivity of the method, i.e. less sensitivity to discontinuities. In some sense this parameter is similar to the wavelet threshold applied in a nonlinear wavelet transform.

A reasonable way to define the parameter λ for specific applications is based on the condition of free extension, which we refer to as "propagation condition". This means that in a homogeneous situation $\theta(X_i) \equiv \theta$, the impact of the statistical penalty in the computed weights w_{ij} is negligible. This would result in a free extension of every local model. If the value h^* is sufficiently large, then at the last iteration all weights w_{ij} will be close to one and every local model will essentially coincide with the global one. Therefore, one can adjust the parameter λ simply selecting by Monte-Carlo simulations the minimal value of λ providing a prescribed probability of getting the weights $w_{ij}^{(k)} \approx 1$ at the end of the iteration process for the parametric model $\theta(x) \equiv \theta$. A theoretical justification is given by Theorem 9.1, that claims that the choice $\lambda = C \log n$ with a sufficiently large C yields the "propagation" condition whatever the parameter θ or the sample size n is. This result suggests to take a slightly increasing λ as the sample size n grows. However, the bound $\lambda \geq C \log n$ is slightly conservative. Our numerical results indicate that an increase of the sample size does not necessarily require to increase λ. Therefore, we utilize as default the constant value $\lambda = t_\alpha(\chi^2_1)$, that is, the α-quantile of the χ^2 distribution with one degree of freedom that relies on the asymptotic distribution of every test statistic T_{ij}. The value α depends on the specified exponential family and the use of an asymmetric or symmetric stochastic penalty. Defaults for α are given in Table 2. They are computed by Monte-Carlo simulation as minimal values providing the "propagation condition".

Remark 4.1. A usual kernel estimate with the kernel K and the bandwidth h can be obtained as the special case of the AWS procedure if λ is taken very large, $K_{\mathrm{loc}} = K$ and $h^* = h$. Moreover, the AWS procedure can be regarded as a sophisticated version

Table 2: Default values for α for different families and for the procedure with symmetric or asymmetric statistical penalty

	Gaussian	Bernoulli	Poisson	Exponential
asymmetric	.966	.953	.958	.914
symmetric	.985	.972	.980	.972

of the kernel estimated with an adaptive asymmetric kernel.

4.4 Numerical complexity of the procedure

The numerical complexity of the procedure is easily analyzed. If the localization kernel K_{loc} is supported on $[0,1]$ and if $M^{(k)}$ denotes the maximal number of points X_j in the neighborhood $U_i^{(k)} = \{x : \rho(x, X_i) \le h^{(k)}\}$ at the kth step of the procedure, then the complexity of this step is of order $nM^{(k)}$. The number of iterations k^* is the largest integer smaller than $\log_a(h^*/h^{(1)})$. Since the value $M^{(k)}$ grows exponentially the whole complexity of the procedure is of order $nM^{(k^*)}$.

5 Application to nonparametric density estimation

Suppose that observations $Z_1, \ldots, Z_L$ are sampled independently from some unknown distribution P on $\mathbb{R}^d$ with density $f(x)$. The problem of adaptive estimation of f can be successfully attacked by the AWS method. Here we consider a small d or moderate, e.g. $d \le 3$. Larger values of d can be handled as well but require a different treatment.

Without loss of generality we suppose that the observations are located in the cube $[0,1]^d$. We do not assume that f is compactly supported or that f is bounded away from zero on $[0,1]^d$. As a first step we apply a *binning* procedure, see e.g. Fan and Marron (1994) or Fan and Gijbels (1996). Let the interval $[0,1]$ be split into M equal disjoint intervals of length $\delta = 1/M$. Then the cube $[0,1]^d$ can be split into $n = M^d$ nonoverlapping small cubes with the side length δ, which we denote by $J_1, \ldots, J_n$. Let X_i be the center point of the cube J_i and let Y_i be the number of observations lying in the ith cube J_i. The pairs (X_i, Y_i) for $i = 1, \ldots, n$ can be viewed as new observations. The joint distribution of $Y_1, \ldots, Y_n$ is described by the multinomial law. This model can be very well approximated by the Poisson model with independent observations Y_i having Poisson distribution with intensity parameter $\theta_i = Lp_i = LP(J_i)$. This is essentially the approach proposed by Lindsay (1974a, 1974b), see also e.g. Efron and Tibshirani (1996).

If the value θ_i has been estimated by $\widehat{\theta}_i$ then the target density f is estimated at

X_i as $\widehat{f}(X_i) = \delta^{-d}\widehat{\theta}_i/L$ or as $\widehat{f}(X_i) = \delta^{-d}\widehat{\theta}_i/\sum_{j=1}^n \widehat{\theta}_j$.

For estimating the values θ_i from the "observations" Y_i we apply the AWS procedure with the local Poisson family from Example 2.4. In addition to the standard parameter set, we need to specify the bin length δ. A reasonable choice is $\delta = c/K$ where K is the smallest integer satisfying $K^d \geq L$ and $c \leq 1$. The procedure applies even if c is small and many bin counts Y_i are zero. Using a small c reduces the discretization error but increases the "sample size" n and therefore, the computational effort by factor c^{-d}.

We use two simulated examples to illustrate the performance of the method. In both examples the symmetrized version of the stochastic penalty was used with defaults for all other parameters for the AWS procedure. For comparison we also computed the kernel density estimates (KDE) with Gaussian kernel and the bandwidth minimizing the Mean Absolute Error (MAE).

Example 5.1. We generate $n = 200$ observations from the univariate distribution with density $f(x) = 1.5 \cdot I_{\{0 \leq x < 0.25\}} + 1.5 \cdot I_{\{0.75 \leq x \leq 1\}} + 0.5 \cdot I_{\{0.25 \leq x < 0.75\}}$.

In the upper left of Figure 1 we provide one typical realization of density estimates using AWS (solid line) and KDE (dashed line). The AWS-estimate was obtained using a regular grid with interval-length $\delta = 0.0025$ and range $(-.1, 1.1)$. The true density (dotted line) is given for comparison. The maximal bandwidth was chosen $h^* = 500\delta = 1.25$. The lower left plot shows the pointwise MAE for both estimates obtained from 500 simulations.

Example 5.2. We generate $n = 2500$ observations from the 2-dimensional density $f(x_1, x_2) = 7.5 \cdot x_1(1 - x_1^2 - x_2^2)_+ I_{\{x_1 \geq 0, x_2 \geq 0\}}$. This density possesses discontinuities along the axis $x_2 = 0$ and discontinuities of the first derivative along the line $x_1 = 0$ and the boundary of the unit disk. The central upper plot of Figure 1 displays 50 contour lines of the estimated density (solid lines) together with the border of the support of the true density (dashed). Results were obtained using a 2-dimensional grid with 120×120 cells on $(-.1, 1.1) \times (-.1, 1.1)$, i.e. with a bin width $\delta = .01$. The maximal bandwidth was set to $h^* = 20\delta = 0.2$.

The external contour can be interpreted as the estimated support of the density. The quality of the estimation of the density support is very good along the line $x_2 = 0$. It is slightly worse along the other axis $x_1 = 0$ where the density goes flatly to zero and along the boundary of the unit circle. This behavior is in agreement with the theoretical results from Korostelev and Tsybakov (1993).

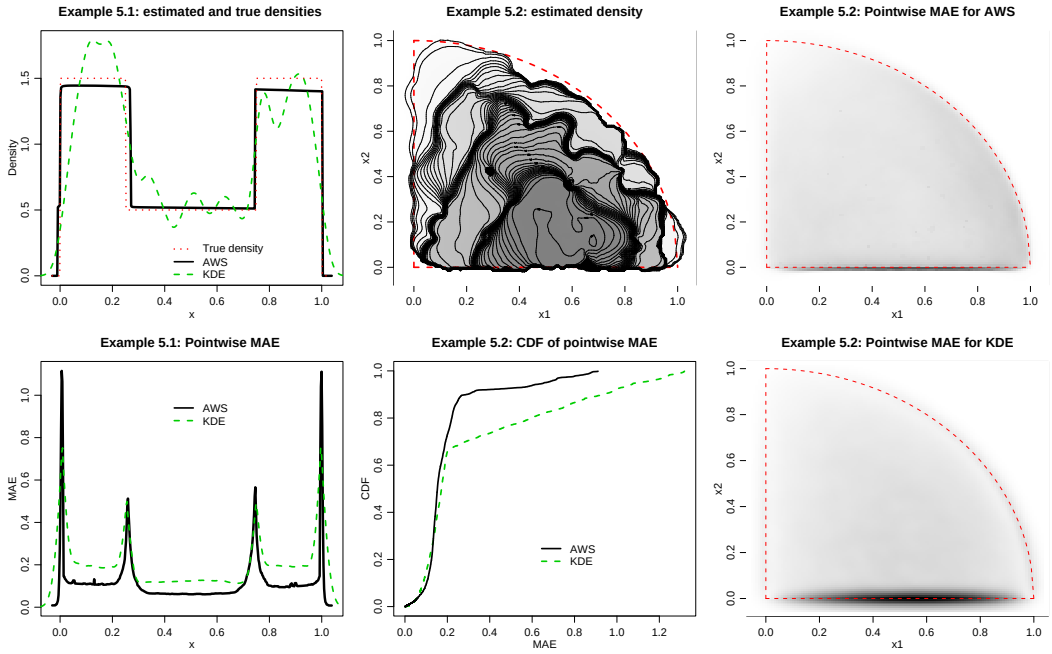

Figure 1: Density estimation: Estimates for typical realizations and simulation result.

Table 3: Simulation results for Examples 5.1 and 5.2

$d=1$: AWS	$d=1$: KDE	$d=2$: AWS	$d=2$: KDE
0.133	0.232	0.131	0.153

The left of Figure 1 provides images of pointwise MAE of for both AWS and KDE obtained from 500 simulations. These two images are summarized in terms of cumulative density functions (CDF) of the MAE-values, sampled on a fine grid, in the central bottom plot. The comparison of the CDF's for AWS and KDE shows a clear advantage of AWS. One can also see from the pointwise plot that the AWS does very well outside of the density support while the KDE oversmooths near boundary. As expected, the KDE performs slightly better in the region of the regularity of the density f.

Table 3 summarizes the results of a simulation of size 500 for both examples comparing the behavior of the AWS and the KDE estimates with respect to MAE.

6 Application to volatility estimation

Let $S_1, \ldots, S_T$ be an observed stock price (exchange rate, option price etc.) process. Log-returns are defined by $R_t = \log(S_t/S_{t-1})$. In many financial market models the log-returns are described by the following *conditional heteroskedasticity* model:

$$R_t = \sigma_t \varepsilon_t \tag{6.1}$$

Table 4: Simulation results based on Example 6.1

	AWS	Regression tree	local linear estimate
MAE	0.0953	0.0967	0.1521

where ε_t are, conditionally on $\mathcal{F}_{t-1} = \sigma(S_1, \ldots, S_{t-1})$, standard normal distributed *innovations* and σ_t is a time dependent predictable *volatility* process, that is $\sigma_t \sim \mathcal{F}_{t-1}$. Aim of the data analysis is to estimate (or forecast) the volatility process σ_t.

The volatility model considered in Example 2.5 is a special case of this model when the volatility process σ_t is deterministic. Note, however, that the local volatility model from Example 2.5 applies to the time dependent volatility from (6.1) in the situation of local time homogeneity, see Mercurio and Spokoiny (2000) for more details. Therefore, we apply the AWS method directly to the time dependent data R_t. The estimate $\widehat{\theta}_t = \widehat{\sigma}_t^2$ of the parameter $\theta_t = \sigma_t^2$ is obtained using AWS on the data $R_1, \ldots, R_T$.

We use two numerical examples to illustrate the behavior of our procedure.

Example 6.1. First we produce an artificial series of returns R_t of length $T = 400$ following the model $R_t = \sigma_t \varepsilon_t$ with $\sigma_t = 1 + I_{\{t \geq 100\}} - 1.5 \cdot I_{\{t \geq 200\}} + 0.5 \cdot I_{\{t \geq 300\}}$.

The left plot in Figure 2 displays for one realization of the process R_t the absolute values $|R_t|$ together with the true volatility σ_t and estimates of the volatility σ_t obtained by the asymmetric version of AWS, with default parameters and maximal bandwidth $h^* = 500$. For a comparison we provide the results for a regression tree and for a local linear smoother, both with smoothing parameters optimized with respect to MAE. Following to Mercurio and Spokoiny (2000), we apply the both methods to the square root of $|R_t|$, leading to a regression like model with an approximately symmetric error distribution. The resulting estimates are then appropriately rescaled and retransformed.

The right plot shows pointwise MAE for all three procedures obtained from 500 simulations. Table 4 provides global simulation results for estimating σ.

AWS demonstrates an almost perfect quality of estimation: the piecewise constant structure of the volatility is reconstructed up to a small error in detecting the location of change-points. The tree based estimate performs similarly, with slightly worse behavior inside the homogeneous regions. Both clearly outperform the local linear estimate.

Example 6.2. In the second example we analyze the exchange rate between the US $ and the German DM for the period from August 1, 1987 to February 18, 2002. The data are (C) 2001 by Prof. Werner Antweiler, University of British Columbia, Van-

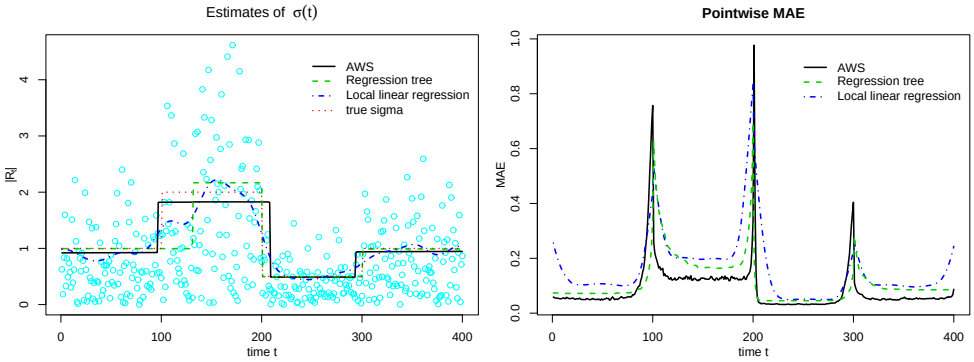

Figure 2: Artificial data set with true volatility function and estimates obtained by AWS and a regression tree (left) and pointwise MAE from 500 simulations.

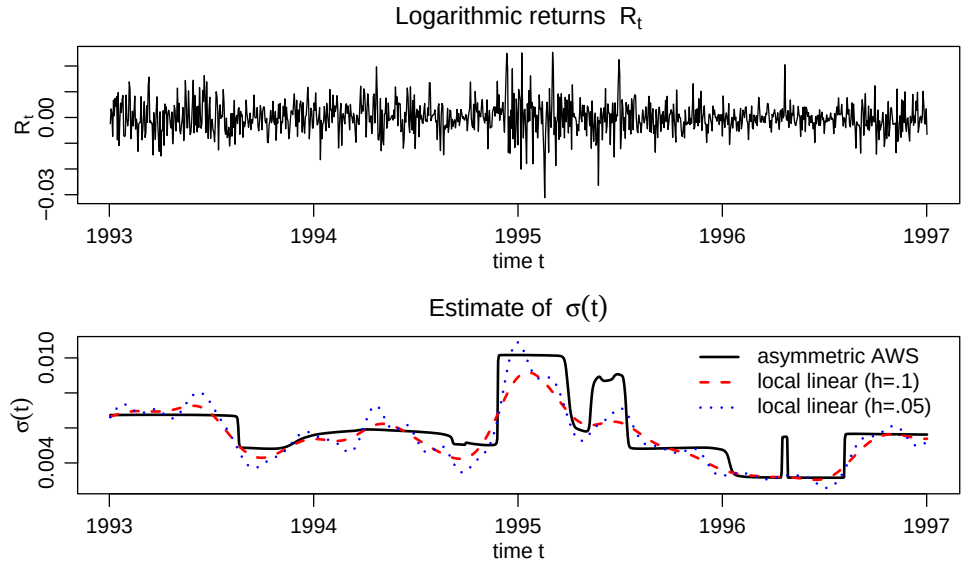

Figure 3: Returns for exchange rate of US $ and German DM and local volatility estimate obtained by AWS and two local linear estimates.

couver BC, Canada, and have been obtained from the Pacific Exchange Rate Service http://pacific.commerce.ubc.ca/xr/data.html. Figure 3 provides the returns $|R_t|$ and estimates of the volatility σ_t obtained by AWS and local linear smoothing of R_t^2 for the time period from January 1993 to December 1997.

Note that the estimates clearly indicate time-inhomogeneity of the volatility.

7 Application to classification

One observes a training sample (X_i, Y_i), $i = 1, \ldots, n$, with X_i valued in a Euclidean space $\mathcal{X} = I\!R^d$ with known class assignment $Y_i \in \{0, 1\}$. Our objective is to construct

a discrimination rule assigning every point $x \in \mathcal{X}$ to one of the two classes.

The classification problem can be naturally treated in the context of a binary response model. It is assumed that each observation Y_i at X_i is a Bernoulli r.v. with parameter $p(X_i)$, that is, $\boldsymbol{P}(Y_i = 0) = 1 - p(X_i)$ and $\boldsymbol{P}(Y_i = 1) = p(X_i)$. The "ideal" discrimination rule is $\rho(x) = \mathbf{1}(p(x) \geq \pi_0)$ where π_0 is the prior probability of the class zero. Since the function $p(x)$ is usually unknown it is replaced by its estimate $\widehat{p}$.

Nonparametric methods of estimating the function p are based on local averaging. Two typical examples are given by the k-nearest neighbors (k-NN) estimate and the kernel estimate. For a given k, define for every point x in $\mathcal{X}$ the subset $\mathcal{D}_k(x)$ of the design $X_1, \ldots, X_n$ containing the k nearest neighbors of x. Then the k-NN estimate of $p(x)$ is defined by averaging the observations Y_i over $\mathcal{D}_k(x)$:

$$\widetilde{p}_k(x) = k^{-1} \sum_{X_i \in \mathcal{D}_k(x)} Y_i .$$

The definition of the kernel estimate of $p(x)$ involves a univariate kernel function $K(t)$ and the bandwidth h:

$$\widetilde{p}_h(x) = \sum_{i=1}^{n} K\left(\frac{\rho^2(x, X_i)}{h^2}\right) Y_i \Big/ \sum_{i=1}^{n} K\left(\frac{\rho^2(x, X_i)}{h^2}\right).$$

Both methods require the choice of a smoothing parameter.

The AWS method can be viewed as a sophisticated extension of both methods using the structural adaption idea. Namely, for estimating the function p at the points $X_1, \ldots, X_n$ we can directly apply the AWS procedure corresponding to the local Bernoulli model from Example 2.2. In order to classify additional observations $X_{n+1}, \ldots, X_{n+m}$ the function p has to be estimated in these points. This can be easily done by applying AWS to the "extended" sample (X_i, Y_i) for $i = 1, \ldots, n+m$, with arbitrary Y_i for $i > n$, and specifying all weights $w_{ij}^{(k)}$ with $j > n$ as zero within the iterative process.

Example 7.1. To illustrate the behavior of AWS in this context we use the data from a simulated two-dimensional discriminant analysis example from Hastie et.al. (2001), page 13. The data and information how they are constructed are available from http://www-stat.stanford.edu/ tibs/ElemStatLearn/. They consist of 200 training observations, 100 from each class. The probability densities for each class are mixtures of Gaussians, see et.al. (2001), page 17, for details.

Figure 4 illustrates the classification rules for the ideal Bayes rule, the k-nearest neighbor rule with optimal $k = 7$, the classification rule obtained by the symmetric

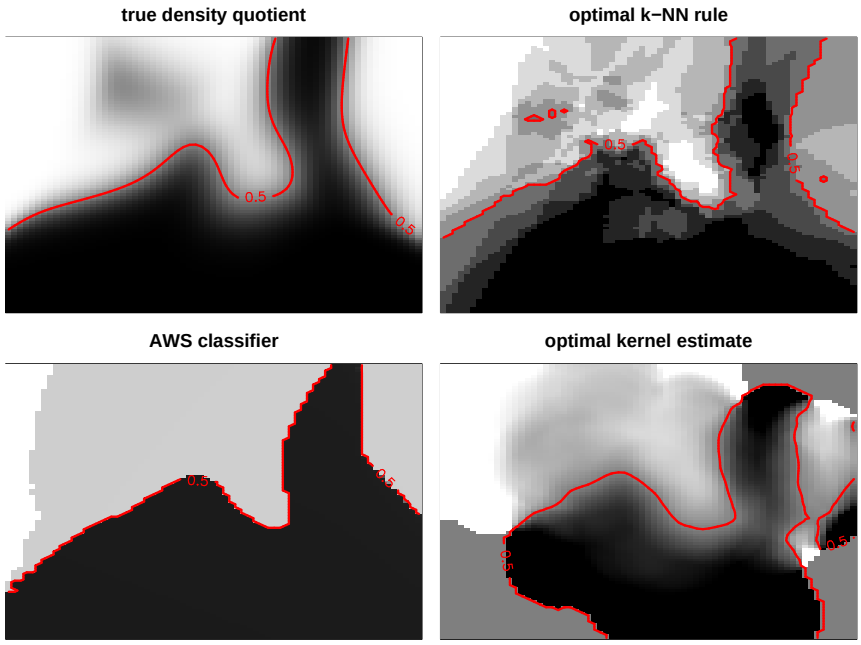

Figure 4: Classification rules obtained by the optimal Bayes decision, the best k-nearest neighbor rule, adaptive weights smoothing (AWS) and the best rule based on kernel estimation.

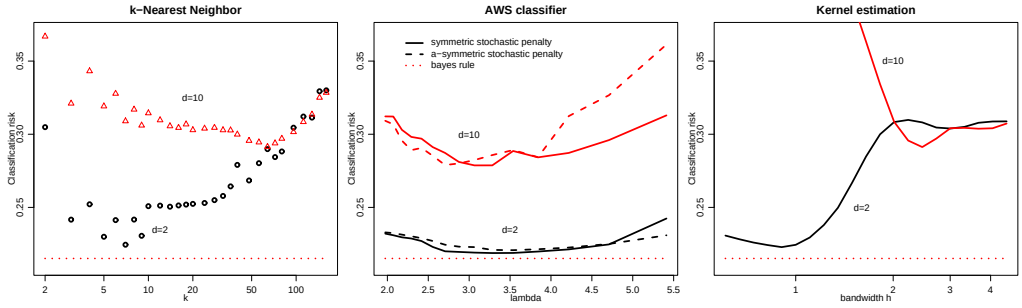

Figure 5: Dependence of the classification error on the main smoothing parameter rules defined by k-nearest neighbor, AWS and kernel estimation.

version of AWS with default value of λ and $h^* = 10$, and the classification rule obtained by the kernel estimate using an Epanechnicov kernel with optimal bandwidth $h = 0.9$. In each case the estimated, or true, function $p(x)$ are provided together with the 0.5-contour line defining the classification rule.

Additionally a 10-dimensional data set has been created adding 8 i.i.d $U(-1, 1)$ nuisance components to each observation. Figure 5 shows graphs of error rates for $d = 2$ and $d = 10$, as functions of the main smoothing parameter for the rules defined by k-nearest neighbor, AWS with symmetric and a-symmetric stochastic penalty, and kernel estimation. Error rates are obtained by classification of 6831 points in predictor space. Numerical integration with respect to class probabilities and Monte Carlo integration are

used in the 2D- and 10D-case, respectively. The ideal Bayes risk is given for a comparison.

Note that the AWS procedure produces the lowest classification errors between the three methods and that the low values are obtained over a wide range of λ-values, with our default setting, $\lambda = 3.9$ and $\lambda = 4.8$ for the a-symmetric and symmetric case, respectively, being conservative for classification. The choice of a smoothing parameter for the other methods is rather critical, with optimal values strongly depending on d and a suboptimal choice leading to significant increases of the error rate.

8 Application to tail index estimation problem

Let $X_1, ..., X_n$ be a sample from a distribution F, and $X_{n,1} \geq ... \geq X_{n,n}$ be their order statistics. The target of the analysis is the tail behavior of this distribution. A popular approach is based on the assumption of a polynomial decay of the value $1 - F(x)$ in the form $1 - F(x) = x^{-1/\alpha} L(x)$ where $L(x)$ is a slowly varying function and α is the parameter of interest which is usually referred to as the *tail index*. The popular Hill estimate, Hill (1975), of α is defined as

$$\widehat{\alpha}_{n,k} = \frac{1}{k} \sum_{i=1}^{k} \log \frac{X_{n,i}}{X_{n,k+1}},$$

where k is the number of upper statistics used in the estimation. There is a vast literature on the asymptotic properties of the Hill estimate. Weak consistency was established by Mason (1982), under the conditions that $k \to \infty$ and $k/n \to 0$ as $n \to \infty$. A strong consistency result can be found in Deheuvels et.al. (1988). However, practical applications of this estimate meet serious problems, see e.g. Embrechts et.al. (1997, p.351). The main difficulty is to chose the parameter k. Another problem is related to the treatment of $L(x)$ which may seriously affect the performance of the estimate, see Embrechts et.al. (1997). Grama and Spokoiny (2002) proposed a new method of adaptive estimation of the parameter α by reducing the original problem to the inhomogeneous exponential model and applying a pointwise adaptive estimation procedure. Here we briefly discuss how the AWS procedure can be used for the same purpose.

Suppose that the distribution $F(x)$ is supported on (a, ∞) where $a > 0$ is a fixed real number. Let the function F be strictly increasing and let it have a continuous density f. Define the function $\alpha(x)$ by the equation

$$\frac{1}{\alpha(x)} = \frac{x f(x)}{1 - F(x)} = -\frac{\frac{d}{dx} \log (1 - F(x))}{\frac{d}{dx} \log x}, \quad x \geq a. \tag{8.1}$$

Since $F(a) = 0$, the d.f. F can be represented as

$$F(x) = 1 - \exp\left(-\int_a^x \frac{dv}{v\alpha(v)}\right), \quad x \geq a. \tag{8.2}$$

Our basic condition is that the function $\alpha(x)$, $x > a$, can be approximated by a constant for large values of x. This is, e.g. the case if there exists an $\alpha > 0$ such that

$$\lim_{x\to\infty} \alpha(x) = \alpha. \tag{8.3}$$

In such a situation, α is precisely the tail index parameter, see the representation theorems in Seneta (1976) or Bingham et.al. (1987).

The problem is to find a number k such that the function $\alpha(x)$, $x \geq a$, can be well approximated on the set $\{X_{n,1}, ..., X_{n,k}\}$ by the value $\alpha(X_{n,1})$ and to estimate this value. The intuitive meaning of this is to find a Pareto approximation for the tail of the d.f. F on the data set $\{X_{n,1}, ..., X_{n,k}\}$. Note that this problem is different from estimating the index of regular variation $\alpha = \alpha(\infty)$. In many examples, the values $\alpha(X_i)$ are essentially different from $\alpha(\infty)$ for all X_i observed for reasonable sample sizes. A typical example is delivered by the so called "Hill horror plot" corresponding to the distribution $F(x) = 1 - x^{-1}\log(x)$.

The function $\alpha(\cdot)$ will be estimated from the approximating exponential model at the points X_i. The construction of the approximating exponential model employs the following lemma, called Renyi representation of order statistics.

Lemma 8.1. *Let $X_1, ..., X_n$ be i.i.d. r.v.'s with common strictly increasing d.f. F and $X_{n,1} > ... > X_{n,n}$ be the order statistics pertaining to $X_1, ..., X_n$. Then the r.v.'s*

$$\xi_i = i \log \frac{1 - F(X_{n,i+1})}{1 - F(X_{n,i})}, \quad i = 1, ..., n-1.$$

are i.i.d. standard exponential.

Proof. See e.g. Reiss (1989) or Example 4.1.5 in Embrechts et.al. (1997). □

Let $Y_i = i \log \frac{X_{n,i}}{X_{n,i+1}}$, $i = 1, ..., n-1$. Then $Y_i = \alpha_i \xi_i$, $i = 1, ..., n-1$, where

$$\alpha_i = -\log \frac{X_{n,i}}{X_{n,i+1}} \Big/ \log \frac{1 - F(X_{n,i})}{1 - F(X_{n,i+1})}.$$

By (8.1) the value α_i can be regarded as an approximation of the value of the function $\alpha(\cdot)$ at the point $X_{n,i+1}$. More precisely, the mean value theorem implies

$$\alpha_i = \alpha\left(X_{n,i+1} + \theta_{n,i+1}\frac{X_{n,i} - X_{n,i+1}}{X_{n,i}}\right),$$

Table 5: MAE of tail-index estimation by AWS for some distributions.

		sample size				
distribution	statistic	100	200	400	800	1600
Pareto	MAE	0.086	0.062	0.046	0.034	0.027
	Bias	0.002	0.001	-0.001	0.002	0.005
	Mean($\alpha(X_{1,n})$)	1.000	1.000	1.000	1.000	1.000
Normal	MAE	0.269	0.197	0.155	0.132	0.110
	Bias	0.268	0.196	0.155	0.132	0.110
	Mean($\alpha(X_{1,n})$)	0.125	0.107	0.095	0.083	0.075
t_2	MAE	0.229	0.177	0.140	0.103	0.082
	Bias	0.221	0.168	0.134	0.097	0.073
	Mean($\alpha(X_{1,n})$)	0.508	0.504	0.502	0.501	0.500
Cauchy	MAE	0.238	0.166	0.129	0.103	0.077
	Bias	0.192	0.126	0.100	0.081	0.057
	Mean($\alpha(X_{1,n})$	1.000	1.000	1.000	1.000	1.000

with some $\theta_{n,i+1} \in [0,1]$, for $i = 1, ..., n-1$. These simple considerations reduce the original model to the following inhomogeneous exponential model

$$Y_i = \alpha_i \xi_i, \quad i = 1, ..., n-1, \tag{8.4}$$

where $\alpha = (\alpha_1, ..., \alpha_{n-1})$ is a unknown parameter vector. It can be estimated by the AWS procedure for the local exponential model, see Example 2.3. The tail index parameter corresponds to the most left piece of local homogeneity of the varying parameter α, or equivalently, to the value α_1. So we use $\widehat{\alpha}_1$ as the estimate of the tail index parameter.

To illustrate the properties of this estimate we present some simulated results and apply the procedure to the exchange rate data.

Example 8.1. Tail indices are estimated for four distributions, using the Pareto-distribution with tail index $\alpha = 1$, the absolute values of standard normal random variables (RV), absolute values of t_2-distributed RV's and absolute values of Cauchy distributed RV's. Table 2 reports the MAE for estimating $\alpha(X_{n,1})$, the estimated bias, i.e. the mean of $\widehat{\alpha}_1 - \alpha(X_{n,1})$, and the mean value of $\alpha(X_{n,1})$, with $\alpha(x)$ defined by (8.1). Results are obtained from 500 simulations. The asymmetric version of the stochastic penalty with default parameters and $h^* = 4n$ is used. The results are very stable and nicely improve with the growing sample size. The bias component in the risk is due to the error of local approximation of the function $\alpha(x)$ near the extreme statistic $X_{n,1}$ by a constant within the local model W_1 centered at the point $X_{n,1}$.

Example 8.2. We reconsider the data used in Example 6.2. The estimated tail index of

the distribution of absolute logarithmic returns $|R_t|$ of the US \$ / DM exchange rate is 0.274. This estimate corresponds to the local model centered at the extreme statistics $|R_{(1)}| = \max_t |R_t|$. Positive weights are supported on the upper 277 values $|R_t|$. This means that α_1 is nothing but the Hill estimate with the adaptive window size 277. The similar tail-index estimates for the standardized absolute logarithmic returns $|R_t|/\widehat{\sigma}_t$ with $\widehat{\sigma}_t$ being the AWS volatility estimate obtained in Example 6.2 equal to 0.1646.

Under the hypothesis of a time homogeneous volatility in model (6.1) the P-value, obtained by Monte-Carlo, of the observed estimate is about 0.001, clearly rejecting this hypothesis for the data at hand. The corresponding P-value of the tail-index estimates for the standardized absolute logarithmic returns is 0.596 not contradicting the hypothesis of homogeneity for the standardized returns.

9 Some important properties of AWS

This section discusses some properties of the proposed AWS procedure. In particular we establish the "propagation condition" which means a free extension of every local model in a homogeneous situation, leading to a nearly parametric estimate at the end of the iteration process. Further we discuss the rate of estimation for a smooth function $\theta(x)$.

9.1 Behavior inside homogeneous regions. Propagation condition

The procedure is designed to provide a free extension of every local model within a large homogeneous region. An extreme case is given by a fully parametric homogeneous model. In that case, a desirable feature of the method is that the final estimate at every point coincides with high probability with the fully parametric global estimate. This property which we call the "propagation" condition is proved here under some simplifying assumptions.

The analysis of the properties of the iterative estimates $\widehat{\theta}_i^{(k)}$ is very difficult. The main reason is that every estimate $\widehat{\theta}_i^{(k)}$ solves the local likelihood problem for the local model defined by the weights $w_{ij}^{(k)}$ which are random and depend on the same observations $Y_1, \ldots, Y_n$. To tackle this problem we make the following assumption:

(A0) for every step k an independent sample $Y_1, \ldots, Y_n$ is available so that the weights $w_{ij}^{(k)}$ are independent of the sample $Y_1, \ldots, Y_n$ for every k.

This assumption can be realized by splitting the original sample into k^* subsamples. Since the number of steps k^* is only of logarithmic order this split can change the

quality of estimation only by a logarithmic factor. Of course, such a split is only a theoretical device, a possibility of using the same sample for all steps of the algorithm still requires further justification.

In our study we restrict ourselves to the case of the varying coefficient exponential family, which is in agreement with all our examples:

(A1) $(P_\theta, \theta \in \Theta \subseteq I\!R)$ is an exponential family with a one-dimensional parameter.

The case of a multi-parameter exponential family can be considered similarly but would be technically much more involved. To simplify the presentation we also assume that

(A2) The statistical penalty $\boldsymbol{s}_{ij}^{(k)}$ is defined via the likelihood ratio test statistic T_{ij}° from (3.2) in Section 3.3.

In our procedure the statistic T_{ij} from (3.3) or its symmetrized version is applied. However, the essential difference between T_{ij} and T_{ij}° may occur only in the situations when the local models W_i and W_j are strongly unbalanced, which do not meet in the specific cases considered in our theoretical study.

First we consider a homogeneous situation which corresponds to a global parametric model with observations $Y_1, \ldots, Y_n$ following a distribution P_θ from the given exponential family. The underlying idea is to apply a nonasymptotic version of the Wilks theorem that claims the asymptotic χ^2-distribution of the test statistic $2L(W, \widehat{\theta}) - 2L(W, \theta)$ under P_θ in the homogeneous situation. The reason for using precise nonasymptotic results is that at the beginning of the iteration process every local "sample size" $N_i = \sum_{j=1}^n w_{ij}$ is small, even if the global sample size n is large. Corollary 11.1 from the Appendix applied with $z = \rho\lambda$ yields in the homogeneous situation for every local model W_i

$$\boldsymbol{P}\left(L(W_i, \widehat{\theta}_i) - L(W_i, \theta) > \rho\lambda\right) \le 2e^{-\rho\lambda}$$

for every $\rho \in (0, 1)$. This immediately implies for the statistical penalty T_{ij}°

$$\begin{aligned} \boldsymbol{P}\left(T_{ij}^{\circ} > 2\rho\lambda\right) &\le \boldsymbol{P}\left(L(W_i, \widehat{\theta}_i) - L(W_i, \theta) > \rho\lambda\right) + \boldsymbol{P}\left(L(W_j, \widehat{\theta}_j) - L(W_j, \theta) > \rho\lambda\right) \\ &\le 4e^{-\rho\lambda} \end{aligned} \tag{9.1}$$

leading to the following results.

Theorem 9.1. *Let (A0), (A1) and (A2) be fulfilled. Suppose that* $\theta(X_i) \equiv \theta$. *If* $\lambda \ge C \log n$ *with a constant* C *depending on the kernel* K_{st} *only, then for each iteration* k

$$\boldsymbol{P}\left(\min_{i,j=1,\ldots,n} K_{\mathrm{st}}(\boldsymbol{s}_{ij}^{(k)}) > 1/2\right) \ge 1 - 4/n.$$

Proof. Define ρ by $K_{\rm st}(\rho) = 1/2$. The bound (9.1) implies for every iteration k

$$\boldsymbol{P}\left(\min_{i,j=1,\ldots,n} K_{\rm st}(\boldsymbol{s}_{ij}^{(k)}) > 1/2\right) = \boldsymbol{P}\left(\max_{i,j=1,\ldots,n} T_{ij}^{(k)} \le \rho\lambda\right) \ge 1 - \sum_{i,j=1}^{n} 2e^{-\rho\lambda} \ge 1 - 1/n$$

provided that $\lambda \ge 3\rho^{-1}\log n$. This yields the assertion. □

This result means that the statistical penalty entering in the weights $w_{ij}^{(k)}$ at every iteration k does not restrict a free extension of every local model.

Corollary 9.1. *Let the assumptions (A0), (A1) and (A2) be fulfilled and $\theta(X_i) \equiv \theta$. If $\lambda \ge C\log n$ and if h^* is sufficiently large then the last step estimate $\widehat{\theta}_i = \widehat{\theta}_i^{(k^*)}$ fulfills for every $z \ge 0$*

$$\boldsymbol{P}\left(nQ(\widehat{\theta}_i, \theta) > 2z\right) \le 4/n + 2e^{-z}.$$

Proof. If h^* is sufficiently large then, at the final iteration $k = k_n$, we have $K_{\rm loc}(\boldsymbol{l}_{ij}^{(k)}) \approx 1$ for every pair (i,j). Theorem 9.1 guarantees $K_{\rm st}(\boldsymbol{s}_{ij}^{(k)}) \ge 1/2$, hence $w_{ij}^{(k)} \ge 1/2$ for all (i,j). This yields $N_i^{(k)} \ge n/2$ and the result follows from Theorem 11.1. □

Due to this result the quantity $nQ(\widehat{\theta}_i, \theta)$ is bounded with a high probability. Since $Q(\theta', \theta) \approx I_\theta|\theta' - \theta|^2/2$, this result claims the root-n consistency of the estimate $\widehat{\theta}_i$. In fact, one can show an even stronger assertion: with a high probability it holds $\widehat{\theta}_i \approx \widehat{\theta}$ where $\widehat{\theta}$ is the global (parametric) MLE of θ from the whole sample $Y_1, \ldots, Y_n$. The explanation is as follows. Our way of computing the statistical penalty $\boldsymbol{s}_{ij}^{(k)}$ does not take into account that two "local" models W_i and W_j have nonzero intersection. This means that there are some points X_l such that the weights $w_{il}^{(k)}$ and $w_{jl}^{(k)}$ are simultaneously positive and hence, the estimates $\widehat{\theta}_i^{(k)}$ and $\widehat{\theta}_j^{(k)}$ are dependent and positively correlated. In the homogeneous situation, for every two fixed points, this dependence grows with iteration, so that the estimates $\widehat{\theta}_i^{(k)}$ and $\widehat{\theta}_j^{(k)}$ become more and more close to each other. In the extreme case at the end of iteration process both local models become very close to each other and the statistical penalties vanish at the end of iteration process.

The propagation condition can be easily extended to the case of a large homogeneous region G in $\mathcal{X}$. Define for every $x \in G$ the distance from x to the boundary of G, i.e. $\rho_G(x) = \min\{\rho(x, X_j) : X_j \notin G\}$. At every step k we consider only internal points $X_i \in G$ which is separated from the boundary with the distance $2h^{(k)}$:

$$\mathcal{G}^{(k)} = \{X_i \in G : \rho_G(X_i) \ge 2h^{(k)}\}.$$

The next result claims the propagation condition (free extension) for all such points.

Theorem 9.2. *Let the assumptions (A0), (A1) and (A2) be fulfilled. Suppose that* $\theta(X_i) \equiv \theta$ *for all* X_i *from some region* G *in* $\mathcal{X}$. *If* $\lambda \geq C \log n$ *with some constant* C *depending on the kernel* K_{st} *only, then for every iteration* k

$$\boldsymbol{P}\left(\min_{(i,j):\, X_i \in \mathcal{G}^{(k)},\, \rho(X_i,X_j) \leq h^{(k)}} K_{\mathrm{st}}(\boldsymbol{s}_{ij}^{(k)}) > 1/2 \right) \geq 1 - 4/n.$$

Proof. It suffices to note that if $X_i \in \mathcal{G}^{(k)}$ then the local model $W_i^{(k)}$ as well as all the models $W_j^{(k)}$ for all X_j with $\rho(X_i, X_j) \leq h^{(k)}$ are homogeneous. Hence, the result follows again by Theorem 11.1. □

9.2 Rate of estimation for a smooth function $\theta(\cdot)$. Spatial adaptivity

Here we consider the case when $\theta(\cdot)$ is a Lipschitz function in some neighborhood of a point $x \in \mathcal{X}$. We first show that this condition ensures a free extension of all the local models within this neighborhood until some critical bandwidth h of order $n^{-1/(2+d)}$ corresponding to the classical nonparametric estimation. This implies the usual nonparametric rate of estimation $n^{-1/(2+d)}$ of the function $\theta(x)$ (corresponding to the smoothness degree one) if the AWS procedure is performed with a control step, see Remark 9.1.

Let a design point $x = X_i$ for some i be fixed, and let h be some bandwidth used in the iteration procedure. We define $U_h(x) = \{x' : |x' - x| \leq h\}$. We consider the following conditions which are specified for the fixed point x and the bandwidth h:

(A3) The function $\theta(\cdot)$ fulfills $|\theta(X_i) - \theta(X_j)| \leq L|X_i - X_j|$ for all $X_j \in U_h(X_i)$.

(A4) There are two positive constants $I_* \leq I^*$ such that $I_* \leq I_{\theta(x')} \leq I^*$ for all $x' \in U_{2h}(x)$, where $I_\theta = C'(\theta)$ is the Fisher information of the family (P_θ) at θ.

(A5) The design points $X_1, \ldots, X_n$ are elements of the Euclidean space $\mathbb{R}^d$ and for some positive constants $C_{X1} \leq C_{X2}$ holds

$$C_{X1} \leq \frac{1}{nh^d} \sum_{j=1}^{n} K_{\mathrm{loc}}(|X_i - X_j|^2/h^2) \leq C_{X2}.$$

(A6) The kernel K_{loc} is compactly supported on $[0, 1]$.

The smoothness condition (A3) allows to approximate the function $\theta(x)$ by a constant within each local model with the precision Lh. For every $X_i \in U_h(x)$ and every k, define $\overline{\theta}_i^{(k)} = \sum_{j=1}^n w_{ij}^{(k)} \theta_j / \sum_{j=1}^n w_{ij}^{(k)}$. Then

$$\left|\overline{\theta}_i^{(k)} - \theta_i\right| \leq Lh. \tag{9.2}$$

The next result claims the propagation condition (free extension) for the local models $W_i^{(k)}$ until $h^{(k)} \le h$.

Theorem 9.3. *Let the assumptions (A0), through (A6) be fulfilled. If $\lambda \ge C \log n$ with some constant C depending on the kernel K_{st} only, and if the bandwidth h fulfills*

$$2C_{X2}\, I^* L^2\, n h^{d+2} \le \rho\lambda/6 \tag{9.3}$$

where ρ is defined by $K(\rho) = 1/2$, then for every iteration k with $h^{(k)} \le h$

$$\boldsymbol{P}\left(\min_{j:\, X_j \in U_h(X_i)} K_{\mathrm{st}}(\boldsymbol{s}_{ij}^{(k)}) > 1/2 \right) \ge 1 - 4/n, \tag{9.4}$$

the estimate $\widehat{\theta}_i^{(k)}$ for the reference point $x = X_i$ fulfills

$$\boldsymbol{P}\left(N_i^{(k)} Q\big(\widehat{\theta}_i^{(k)}, \overline{\theta}_i^{(k)}\big) > \lambda \right) \le 2/n \tag{9.5}$$

and it holds with a probability at least $1 - 4/n$

$$\left| \widehat{\theta}_i^{(k)} - \theta_i \right| \le Lh + 2\sqrt{\lambda/(I_*\, C_{X1} n h^d)}. \tag{9.6}$$

Remark 9.1. The proof is given in the Appendix. The result (9.6) indicates that the first k iterations of the procedure (until $h^{(k)} \le h$) lead to a reasonable quality of estimation of the function $\theta(\cdot)$. However, the procedure has to prevent from losing the obtained quality of estimation during further iterations. This is precisely what the additional *control step* of the original AWS procedure from PS2000, in which the new estimate $\widehat{\theta}_i^{(k)}$ is compared with all the previous estimates $\widehat{\theta}_i^{(k')}$ for $k' < k$, does. If the difference $\widehat{\theta}_i^{(k)} - \widehat{\theta}_i^{(k')}$ became significant, the new estimate was not accepted and the previous step estimate was used. This control step is a very useful device for proving some theoretical properties of the procedure, because it ensures that the gained quality of estimation will not be lost in further iterations. In the case of local exponential family, this control step will accept the estimate $\widehat{\theta}_i^{(k)}$ only if

$$N_i^{(k')} Q(\widehat{\theta}_i^{(k')}, \widehat{\theta}_i^{(k)}) \le \tau, \qquad k' = 1, \ldots, k-1, \tag{9.7}$$

that is, when the differences between the new estimate $\widehat{\theta}_i^{(k)}$ and all the previous ones at the same point X_i are not significant. Our experience with the procedure only shows an effect of the control step in very particular situations, leaving its use questionable. The use of the "memory" parameter η can be regarded as a soft version of the control step.

Although the procedure applies a soft form of the control step, we only show how the *hard* control step can be used for proving the rate result.

Theorem 9.4. *Let the conditions of Theorem 9.3 be fulfilled and let the procedure involve the control step from (9.7) with* $\tau \geq \lambda$. *Then the last step estimate* $\widehat{\theta}_i$ *fulfills* $N_i^{(k)} Q\bigl(\widehat{\theta}_i^{(k)}, \widehat{\theta}_i\bigr) \leq \tau$ *and hence, it holds with a probability at least* $1 - 4/n$

$$\left|\widehat{\theta}_i - \theta_i\right| \leq Lh + 2\sqrt{\lambda/(I_*\, C_{X1} n h^d)} + 2\sqrt{\tau/(I_*\, C_{X1} n h^d)}. \tag{9.8}$$

Proof. This result is a direct corollary of Theorem 9.3 and (9.7). □

Optimization of the bandwidth h under condition (A3) leads to the choice $h \approx \{4\lambda/(I_* n L^2)\}^{-1/(d+2)}$ and to an accuracy of estimation of order $\{\lambda/(I_* n)\}^{1/(d+2)} L^{2/(d+2)}$ which is optimal, up to a logarithmic factor, for the problem of estimation of a Lipschitz function at a point. This means that our procedure is pointwise adaptive in the sense that it automatically adapts to an unknown local smoothness degree measured by the Lipschitz constant L. As shown in Lepski et.al. (1997) this property automatically leads to rate optimality in the Sobolev and Besov function classes $B^1_{p,q}$.

10 Conclusion and outlook

This paper presents a new method of adaptive nonparametric estimation based on the *adaptive weights* idea. An important feature of the AWS procedure is that it applies to a broad class of nonparametric models in a unified way. In many cases its adjustment to the particular situation is trivial. For all the examples in this paper, we essentially applied the same procedure. Sometimes, a preliminary model (data) transformation is required, as in tail index or density estimation.

The procedure can be applied to smooth functions and functions with discontinuities, it adapts automatically to the unknown structure of the model.

The procedure allows for arbitrary dimensionality of $\mathcal{X}$. This makes it feasible to apply the procedure to e.g. image denoising or estimation of a multivariate density and to use it in case of a multidimensional explanatory vector X_i.

The AWS procedure is computationally straightforward and the numerical complexity can be easily controlled by restricting the largest bandwidth h^*, see Section 4.4.

Applications of the AWS procedure are however restricted to models which allow a good piecewise constant approximation. If the underlying model function θ is smooth the local constant approximation may lead to a substantial bias. This problem is well recognized in nonparametric statistics, see e.g. Fan and Gijbels (1996) and local linear (polynomial) smoothing is preferable for such cases. An extension of the AWS approach

to the local linear case is possible but it requires a separate treatment. In particular, the procedure has to be significantly modified because the local likelihood equation cannot be explicitly solved in this case. Local linear (polynomial) regression by AWS is discussed in Polzehl and Spokoiny (2003b). AWS for a local linear exponential family is discussed in context of generalized linear models in Grama, Polzehl and Spokoiny (2003).

11 Appendix

We present some general results for the local exponential family model. The considered exponential family $(P_\theta, \theta \in \Theta \subseteq I\!R)$ is described by the functions $C(\theta)$ and $B(\theta)$, with $p(y,\theta) = dP_\theta/dP(y) = \exp(C(\theta)y - B(\theta))$ and $E_\theta Y = \int y p(y,\theta) dP(y) = \theta$ for all $\theta \in \Theta$, see Section 3.2. We suppose $U(y) = y$ to simplify our notation.

We assume the observation Y_i to be P_{θ_i}-distributed with θ_i depending on location X_i. Let also a local model W be described by the weights $w_i \in [0,1]$ for $i = 1, \ldots, n$. The corresponding local MLE is given as $\widehat{\theta} = \sum_{i=1}^n w_i Y_i \big/ \sum_{i=1}^n w_i$. We use the representation $\widehat{\theta} = S/N$ with $S = \sum_{i=1}^n w_i Y_i$, $N = \sum_{i=1}^n w_i$ and denote $\overline{\theta} = N^{-1} \sum_{i=1}^n w_i \theta_i$.

Our first result can be regarded as a nonasymptotic local version of the Wilks theorem. Namely, we show that the expression $L(W, \widehat{\theta}, \overline{\theta})$ is uniformly bounded with a high probability. It is convenient to introduce the parameter $\upsilon = C(\theta)$ and define $\overline{\upsilon} = C(\overline{\theta})$ and $D(\upsilon) = B(\theta) = B(C^{-1}(\upsilon))$. Since $C'(\theta) > 0$, the new parameter υ is uniquely defined. By simple analysis $D'(\upsilon) = \theta = C^{-1}(\upsilon)$ and $D''(\upsilon) = 1/C'(\theta) = 1/I(\theta) = 1/I(C^{-1}(\upsilon))$. Moreover, $Q(\upsilon_1, \upsilon_2) = D(\upsilon_2) - D(\upsilon_1) - (\upsilon_2 - \upsilon_1) D'(\upsilon_1)$ is the Kullback-Leibler distance between two parametric distributions corresponding to the parameters υ_1 and υ_2. In what follows we use the notation $q(u|\upsilon) = Q(\upsilon, \upsilon + u) = D(\upsilon + u) - D(\upsilon) - uD'(\upsilon)$ and $L(W, \theta, \theta') = L(W, \theta) - L(W, \theta')$ for any pair θ, θ'.

Theorem 11.1. *Let the Fisher information $I(\theta) = C'(\theta)$ be positive on Θ. For a given $z \geq 0$, let $\mathcal{U}(W, z)$ be the set of solutions u of equation $q(u|\overline{\upsilon}) = \int_0^u x D''(\overline{\upsilon} + x) dx = z/N$. If there is some $\alpha > 0$ such that for all $\mu \in (0,1]$ and all $u \in \mathcal{U}(W, z)$*

$$q(\pm w_\ell \mu u | \upsilon_\ell) \leq (1 + \alpha) w_\ell \mu^2 q(u|\overline{\upsilon}), \qquad \ell = 1, \ldots, n, \tag{11.1}$$

then

$$\boldsymbol{P}\Big(L(W, \widehat{\theta}, \overline{\theta}) > z\Big) \leq 2e^{-z/(1+\alpha)}.$$

Remark 11.1. The condition (11.1) can be easily checked in many particular situations. We give two typical examples. The first one corresponds to the homogeneous case when

all υ_i coincide with their mean $\overline{\upsilon}$. Then (11.1) is fulfilled automatically with $\alpha = 0$. Indeed the function $q(\cdot|\upsilon)$ satisfies $q'(u|\upsilon) = D'(\upsilon + u) - D'(\upsilon)$ and $q''(u|\upsilon) = D''(\upsilon + u) = 1/I(C^{-1}(\upsilon + u)) > 0$ and thus, it is convex. Since also $q(0|\upsilon) = 0$, it holds $q(wa|\upsilon) \le wq(a|\upsilon)$ for every $w \in [0,1]$ and every a implying (11.1) with $\alpha = 0$ and arbitrary u. Since this special case is used in the proof of the "propagation condition" in Section 9, we present it as a separate statement.

Corollary 11.1. *If $\theta_i \equiv \theta$, then $\boldsymbol{P}\Big(L(W,\widehat{\theta},\theta) > z\Big) \le 2e^{-z}$ for every $z > 0$.*

In the general inhomogeneous situation, the Taylor expansion yields that $q(wu|\upsilon) = D(\upsilon + wu) - D(\upsilon) - wuD'(\upsilon) = 1/2\, w^2u^2D''(\upsilon + \tau wu)$ for some $\tau \in [0,1]$. If the Fisher information $I(\theta)$ is bounded from zero and infinity, that is, $I_* \le I(\theta) \le I^*$ for all $\theta \in \Theta$, then $1/I^* \le D''(\upsilon) \le 1/I_*$ for all υ and one easily gets for every $u \in \mathcal{U}(W,z)$ that $u^2 \le 2zI^*/N$. Therefore, the condition (11.1) is certainly fulfilled with a small $\alpha \ge 0$ for the case when all υ_i are close to the mean $\overline{\upsilon}$ or when the weights w_i are very small for all ℓ with a large value $|\upsilon_\ell - \overline{\upsilon}|$.

Proof of Theorem 11.1. The log-likelihood ratio can be rewritten for the new parameter υ as

$$L(W,\theta,\overline{\theta}) = L(W,\upsilon,\overline{\upsilon}) = (\upsilon - \overline{\upsilon})S - N\Big(D(\upsilon) - D(\overline{\upsilon})\Big).$$

The MLE $\widehat{\upsilon}$ of the parameter υ is defined by maximizing $L(W,\upsilon,\overline{\upsilon})$, that is, $\widehat{\upsilon} = \operatorname{argsup}_\upsilon L(W,\upsilon,\overline{\upsilon})$.

Lemma 11.1. *For given z, there exist two values $\upsilon^* > \overline{\upsilon}$ and $\upsilon_* < \overline{\upsilon}$ such that*

$$\{L(W,\widehat{\upsilon},\overline{\upsilon}) > z\} \subseteq \{L(W,\upsilon^*,\overline{\upsilon}) > z\} \cup \{L(W,\upsilon_*,\overline{\upsilon}) > z\}.$$

Proof. It holds

$$\begin{aligned}\{L(W,\widehat{\upsilon},\overline{\upsilon}) > z\} &= \left\{\sup_\upsilon \Big[S(\upsilon - \overline{\upsilon}) - N\Big(D(\upsilon) - D(\overline{\upsilon})\Big)\Big] > z\right\} \\ &\subseteq \left\{S > \inf_{\upsilon > \overline{\upsilon}} \frac{z + N\Big(D(\upsilon) - D(\overline{\upsilon})\Big)}{\upsilon - \overline{\upsilon}}\right\} \cup \left\{-S > \inf_{\upsilon < \overline{\upsilon}} \frac{z + N\big(D(\upsilon) - D(\overline{\upsilon})\big)}{\overline{\upsilon} - \upsilon}\right\}.\end{aligned}$$

The function $f(u) = \big[z + N\big(D(\overline{\upsilon} + u) - D(\overline{\upsilon})\big)\big]/u$ attains its minimum at some point u satisfying the equation

$$z + N\big(D(\overline{\upsilon} + u) - D(\overline{\upsilon})\big) - NuD'(\overline{\upsilon} + u) = 0$$

or, equivalently,

$$\int_0^u xD''(\overline{\upsilon} + x)dx = z/N.$$

Therefore

$$\left\{ S > \inf_{\upsilon > \overline{\upsilon}} \frac{z + N\big(D(\upsilon) - D(\overline{\upsilon})\big)}{\upsilon - \overline{\upsilon}} \right\} = \left\{ S > \frac{z + N\big(D(\upsilon^*) - D(\overline{\upsilon})\big)}{\upsilon - \overline{\upsilon}} \right\} \subseteq \{L(W, \upsilon^*, \overline{\upsilon}) > z\}$$

with $\upsilon^* = \overline{\upsilon} + u$. Similarly

$$\left\{ -S > \inf_{\upsilon < \overline{\upsilon}} \frac{z + N\big(D(\upsilon) - D(\overline{\upsilon})\big)}{\overline{\upsilon} - \upsilon} \right\} \subseteq \{L(W, \upsilon_*, \overline{\upsilon}) > z\}$$

for some $\upsilon_* < \overline{\upsilon}$. □

Now we bound the probability $\boldsymbol{P}\left(L(W, \upsilon, \overline{\upsilon}) > z\right)$ for every υ. Note that the equality $\overline{\theta} = D'(\overline{\upsilon})$ implies for $u = \upsilon - \overline{\upsilon}$

$$\begin{aligned} L(W, \upsilon, \overline{\upsilon}) &= u(S - N\overline{\theta}) - N\left[D(\overline{\upsilon} + u) - D(\overline{\upsilon}) - uD'(\overline{\upsilon})\right] \\ &= u(S - N\overline{\theta}) - Nq(u|\overline{\upsilon}). \end{aligned}$$

Now the result of the theorem is a direct corollary of the following general assertion.

Lemma 11.2. *For every u and every z*

$$\begin{aligned} r(u, z) &:= \log \boldsymbol{P}\left(L(W, \overline{\upsilon} + u, \overline{\upsilon}) > z\right) \\ &\leq -\mu z - \mu N q(u|\overline{\upsilon}) + \sum_{\ell=1}^{n} q(u\mu w_\ell | \upsilon_\ell). \end{aligned}$$

Also

$$\begin{aligned} r_1(u, z) &:= \log \boldsymbol{P}\left(L(W, \overline{\upsilon} + u, \overline{\upsilon}) < -z - 2Nq(u|\overline{\upsilon})\right) \\ &\leq -\mu z - \mu N q(u|\overline{\upsilon}) + \sum_{\ell=1}^{n} q(-u\mu w_\ell | \upsilon_\ell). \end{aligned}$$

Moreover, if u fulfills (11.1) then

$$r(u, z) \leq -z/(1 + \alpha), \qquad r_1(u, z) \leq -z/(1 + \alpha).$$

Proof. We apply the Chebyshev exponential inequality: for every positive μ

$$r(u, z) \leq -\mu z - \mu N q(u|\overline{\upsilon}) + \log \boldsymbol{E} \exp\left(u\mu(S - N\overline{\theta})\right).$$

The independence of the Y_ℓ's implies

$$\log \boldsymbol{E} \exp\left(u\mu(S - N\overline{\theta})\right) = \log \boldsymbol{E} \exp\left(\sum_{\ell=1}^{n} u\mu w_\ell (Y_\ell - \theta_\ell)\right) = \sum_{\ell=1}^{n} \log \boldsymbol{E} e^{u\mu w_\ell (Y_\ell - \theta_\ell)}.$$

For every constant a and every $\ell \le n$, the equalities $\log \int e^{\upsilon_\ell y - D(\upsilon_\ell)} P(dy) = 0$ and $\theta_\ell = D'(\upsilon_\ell)$ yield

$$\begin{aligned} \log \boldsymbol{E} e^{a(Y_\ell - \theta_\ell)} &= -a\theta_\ell + \log \int e^{(a+\upsilon_\ell)y - D(\upsilon_\ell)} P(dy) \\ &= -aD'(\upsilon_\ell) + D(\upsilon_\ell + a) - D(\upsilon_\ell) = q(a|\upsilon_\ell). \end{aligned}$$

Therefore

$$r(u,z) \le -\mu z - \mu N q(u|\overline{\upsilon}) + \sum_{i=1}^{n} q(u\mu w_\ell | \upsilon_\ell).$$

This inequality applied with $\mu = (1+\alpha)^{-1}$ and (11.1) imply

$$r(u,z) \le -\mu z - \mu N q(u|\overline{\upsilon}) + (1+\alpha)\mu^2 \sum_{i=1}^{n} w_\ell q(u|\overline{\upsilon}) \le -z/(1+\alpha).$$

Similarly

$$\begin{aligned} r_1(u,z) &= \boldsymbol{P}\left(-u(S - N\overline{\theta}) + Nq(u|\overline{\upsilon}) > z + 2Nq(u|\overline{\upsilon})\right) \\ &\le -\mu z - \mu N q(u|\overline{\upsilon}) + \sum_{i=1}^{n} q(-u\mu w_\ell | \upsilon_\ell). \end{aligned}$$

and the lemma follows. □

Next we consider the likelihood ratio test statistic T°_{ij} defined in Section 3.3 for two local models W_i and W_j. We show that if the difference between two local models defined in terms of the Kullback-Leibler distance, is sufficiently small, then with a large probability, T°_{ij} is smaller then $\rho\lambda$ for some $\rho \le 1$.

Define $\overline{\theta}_i = \sum_{\ell=1}^n w_{i\ell}\theta_\ell / \sum_{\ell=1}^n w_{i\ell}$ and similarly $\overline{\theta}_j$. Define also the mixed model $W_{ij} = (W_i + W_j)/2$ and $\overline{\theta}_{ij} = (N_i\overline{\theta}_i + N_j\overline{\theta}_j)/(N_i + N_j)$.

Theorem 11.2. *Let $\rho \in (0,1]$ and $z = \rho\lambda/6$. Let the condition (11.1) be fulfilled for the local model W_i with $u = |C(\overline{\theta}_i) - C(\overline{\theta}_{ij})|$ and with $u \in \mathcal{U}(W_i, z)$, and for the local model W_j with $u = |C(\overline{\theta}_j) - C(\overline{\theta}_{ij})|$ and with $u \in \mathcal{U}(W_j, z)$. Then the condition*

$$N_i Q(\overline{\theta}_i, \overline{\theta}_{ij}) + N_j Q(\overline{\theta}_j, \overline{\theta}_{ij}) \le \rho\lambda/6 \tag{11.2}$$

implies

$$\boldsymbol{P}\left(T^\circ_{ij} > \rho\lambda\right) \le 4e^{-\frac{\rho\lambda}{6(1+\alpha)}}.$$

Proof. It holds

$$T^\circ_{ij} = L(W_i, \widehat{\theta}_i) + L(W_j, \widehat{\theta}_j) - L(W_i + W_j, \widehat{\theta}_{ij})$$

where $\widehat{\theta}_i = S_i/N_i$ and $\widehat{\theta}_{ij} = (S_i + S_j)/(N_i + N_j)$. We apply this formula with $\theta' = \overline{\theta}_{ij}$. Since $L(W_i + W_j, \widehat{\theta}_{ij}, \overline{\theta}_{ij}) \geq L(W_i + W_j, \overline{\theta}_{ij}, \overline{\theta}_{ij}) = 0$, it holds

$$T^{\circ}_{ij} \leq L(W_i, \widehat{\theta}_i, \overline{\theta}_{ij}) + L(W_j, \widehat{\theta}_j, \overline{\theta}_{ij}).$$

Clearly

$$L(W_i, \widehat{\theta}_i, \overline{\theta}_{ij}) = L(W_i, \widehat{\theta}_i, \overline{\theta}_i) - L(W_i, \overline{\theta}_{ij}, \overline{\theta}_i)$$

Theorem 11.1 implies for every $z \geq 0$ that

$$\boldsymbol{P}\left(L(W_i, \widehat{\theta}_i, \overline{\theta}_i) > z\right) \leq 2e^{-z/(1+\alpha)}.$$

Next, Lemma 11.2 applied with $u = \overline{v}_i - \overline{v}_{ij} = C(\overline{\theta}_i) - C(\overline{\theta}_{ij})$ implies

$$\log \boldsymbol{P}\left(-L(W_i, \overline{\theta}_{ij}, \overline{\theta}_i) > z + 2N_i Q(\overline{\theta}_i, \overline{\theta}_{ij})\right) \leq -z/(1+\alpha).$$

Similar assertion hold for the model W_j. Therefore

$$\boldsymbol{P}\left(T^{\circ}_{ij} > 4z + 2N_i Q(\overline{\theta}_i, \overline{\theta}_{ij}) + 2N_j Q(\overline{\theta}_j, \overline{\theta}_{ij})\right) \leq 4e^{-z/(1+\alpha)}.$$

This inequality with $z = \rho\lambda/6$ and (11.2) imply the assertion. □

We now present some sufficient conditions for separability of two local models. Namely, we aim to establish conditions that ensure $T^{\circ}_{ij} \geq A\lambda$ where A is the length of the support of the kernel K_{st}. With this conditions, it holds $K_{\mathrm{st}}(T_{ij}/\lambda) = 0$ and hence the new computed weight w_{ij} will be equal to zero.

Theorem 11.3. *Let $\rho \in (0, 1]$ and let the condition (11.1) be fulfilled for the local model W_i with $u = |C(\overline{\theta}_i) - C(\overline{\theta}_{ij})|$, for the model W_j with $u = |C(\overline{\theta}_j) - C(\overline{\theta}_{ij})|$ and for the mixed model W_{ij} with $u \in \mathcal{U}(W_{ij}, \rho\lambda)$. Then the conditions*

$$N_i Q(\overline{\theta}_i, \overline{\theta}_{ij}) \geq (6\rho + A)\lambda, \qquad N_j Q(\overline{\theta}_j, \overline{\theta}_{ij}) \geq (6\rho + A)\lambda, \tag{11.3}$$

imply

$$\boldsymbol{P}\left(T^{\circ}_{ij} < A\lambda\right) \leq 4e^{-\rho\lambda/(1+\alpha)}.$$

Proof. Similarly to the proof of Theorem 11.2 we use the representation

$$T^{\circ}_{ij} = L(W_i, \widehat{\theta}_i, \overline{\theta}_{ij}) + L(W_j, \widehat{\theta}_j, \overline{\theta}_{ij}) - L(W_i + W_j, \widehat{\theta}_{ij}, \overline{\theta}_{ij})$$

Theorem 11.1 applied to the local model W_{ij} implies

$$\boldsymbol{P}\left(L(W_{ij}, \widehat{\theta}_{ij}, \overline{\theta}_{ij}) > \rho\lambda\right) \leq 2e^{-\rho\lambda/(1+\alpha)}.$$

Since $\widehat{\theta}_i$ maximizes $L(W_i, \theta, \overline{\theta}_{ij})$ and similarly for $\widehat{\theta}_j$, it follows

$$\boldsymbol{P}\left(T_{ij}^{\circ} < L(W_i, \overline{\theta}_i, \overline{\theta}_{ij}) + L(W_j, \overline{\theta}_j, \overline{\theta}_{ij}) - 2\rho\lambda\right) \le 2e^{-\rho\lambda/(1+\alpha)}. \tag{11.4}$$

Lemma 11.2 applied to $L(W_i, \overline{\theta}_i, \overline{\theta}_{ij}) = -L(W_i, \overline{\theta}_{ij}, \overline{\theta}_i)$ with $z = -(\rho + A/2)\lambda$ and $\mu = (5\rho + A/2)/\{2(6\rho + A)(1+\alpha)\}$ and the conditions (11.1) and (11.3) imply

$$\begin{aligned}
&\log \boldsymbol{P}\left(L(W_i, \overline{\theta}_i, \overline{\theta}_{ij}) < (\rho + A/2)\lambda\right) \\
&\quad = \log \boldsymbol{P}\left(L(W_i, \overline{\theta}_{ij}, \overline{\theta}_i) > -(\rho + A/2)\lambda\right) \\
&\quad \le (\rho + A/2)\lambda\mu - N_i Q(\overline{\theta}_i, \overline{\theta}_{ij})\mu + (1+\alpha)N_i Q(\overline{\theta}_i, \overline{\theta}_{ij})\mu^2 \\
&\quad \le (\rho + A/2)\lambda\mu - (6\rho + A)\lambda\mu + (1+\alpha)(6\rho + A)\lambda\mu^2 \\
&\quad = -\frac{(5\rho + A/2)^2\lambda}{4(6\rho + A)(1+\alpha)} \le -\frac{\rho\lambda}{(1+\alpha)}.
\end{aligned}$$

This and a similar inequality for $L(W_j, \overline{\theta}_j, \overline{\theta}_{ij})$ yield the theorem in view of (11.4). □

Proof of Theorem 9.3

The propagation condition (9.4) follows similarly to the proof of Theorem 9.2. The only difference is that in the local Lipschitz case we apply Theorem 11.2 instead of Corollary 11.1. Let k be such that $h^{(k)} \le h$ and $X_j \in U_h(X_i)$. We apply Theorem 11.2 to the local models $W_i^{(k)}$ and $W_j^{(k)}$. For this we have to check the condition (11.2). Assumption (A5) clearly implies $N_i^{(k)} \le C_{X2} n h^d$ and similarly for $N_j^{(k)}$. Assumption (A3) yields $|\overline{\theta}_i^{(k)} - \overline{\theta}_j^{(k)}| \le 2Lh$. Define $\overline{\theta}_{ij}^{(k)} = \bigl(N_i^{(k)}\overline{\theta}_i^{(k)} + N_j^{(k)}\overline{\theta}_j^{(k)}\bigr)/\bigl(N_i^{(k)} + N_j^{(k)}\bigr)$. Now the inequality $Q(\theta, \theta') \le I^*|\theta - \theta'|^2/2$ and condition (9.3) imply

$$\begin{aligned}
N_i^{(k)} Q(\overline{\theta}_i^{(k)}, \overline{\theta}_{ij}^{(k)}) + N_j^{(k)} Q(\overline{\theta}_j^{(k)}, \overline{\theta}_{ij}^{(k)}) &\le C_{X2} n h^d I^* \left(\bigl|\overline{\theta}_i^{(k)} - \overline{\theta}_{ij}^{(k)}\bigr|^2 + \bigl|\overline{\theta}_j^{(k)} - \overline{\theta}_{ij}^{(k)}\bigr|^2\right)/2 \\
&\le C_{X2} n h^d I^* \bigl|\overline{\theta}_i^{(k)} - \overline{\theta}_j^{(k)}\bigr|^2/2 \\
&\le 2C_{X2} I^* L^2 n h^{d+2} \le \rho\lambda/6.
\end{aligned}$$

Theorem 11.2 now applies with some $\alpha \ge 0$, see Remark 11.1, yielding

$$\boldsymbol{P}\left(\boldsymbol{s}_{ij}^{(k)} < 1/2\right) \le e^{-\frac{\rho\lambda}{6(1+\alpha)}} \le n^{-2}$$

provided that $\lambda = C \log n$ with C fulfilling $C\rho \ge 12(1+\alpha)$, and (9.4) follows. The assertion (9.5) is a corollary of Theorem 11.1.

Let now $h^{(k)} = h$. By (9.4) all the weights $w_{ij}^{(k)}$ for the local model $W_i^{(k)}$ satisfy the condition $w_{ij}^{(k)} \ge 0.5 K_{\mathrm{loc}}(\boldsymbol{l}_{ij}^{(s)})$ with a high probability. This and Assumption (A5) yield $N_i^{(k)} \ge 0.5 C_{X1} n h^d$. Since also $Q\bigl(\widehat{\theta}_i^{(k)}, \overline{\theta}_i^{(k)}\bigr) \ge I_*\bigl|\widehat{\theta}_i^{(k)} - \overline{\theta}_i^{(k)}\bigr|/2$, the last assertion of the theorem follows by (9.2).

References

[1] Bingham, N. H., Goldie, C. M. and Teugels, J. L. (1987) *Regular variation.* Cambridge University Press, Cambridge.

[2] Cai, Z. Fan,J. and Li, R. (2000). Efficient estimation and inference for varying coefficients models. *J. Amer. Statist. Ass.*, **95** 888–902.

[3] Cai, Z. Fan, J. and Yao, Q. (2000). Functional-coefficient regression models for nonlinear time series *J. Amer. Statist. Ass.*, **95** 941–956.

[4] Caroll, R.J., Ruppert, D, and Welsh, A.H. (1998). Nonparametric estimation via local estimating equation. *J. Amer. Statist. Ass.*, **93** 214–227.

[5] Deheuvels, P., Häusler, E. and Mason, D.M. (1988). Almost sure convergence of the Hill estimator. *Math. Proc. Cambridge Philos. Soc.*, **104** 371–381.

[6] Efron, B., Tibshirani, R. (1996). Using specially designed exponential families for density estimation. *Ann. Statist.*, **24**, 2431–2461.

[7] Embrechts, P., Klüppelberg, K., and Mikosch, T. (1997). *Modelling extremal events.* Springer.

[8] Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications.* Chapman & Hall, London.

[9] Fan, J., Marron, J.S. (1994). Fast implementations of nonparametric curve estimators. *J. Comp. Graph. Statist.* **3** 35–56.

[10] Fan, J., Zhang, C. and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.* **29**, 153–193.

[11] Fan, J., Zhang, W. (1999). Statistical estimation in varying coefficient models. *Ann. Statist.* **27** 1491–1518.

[12] Grama, I. and Spokoiny, V. (2003). Tail index estimation by local exponential modelling. WIAS-Preprint 819.

[13] Grama, I., Polzehl, J. and Spokoiny, V. (2003). Adaptive estimation for varying coefficient generalized linear models. Manuscript in preparation.

[14] Hastie, T.J. and Tibshirani, R.J. (1993). Varying-coefficient models (with discussion). *J. Royal Statist. Soc. Ser. B*, **55** 757–796.

[15] Hastie, T.J., Tibshirani, R.J. and Friedman, J. (2001). *The Elements of Statistical Learning.* Springer.

[16] Hill, B. M., (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.* **3** 1163–1174.

[17] Koo, J.-Y. and Kooperberg, C. (2000). Logspline density estimation for binned data. *Statistics & Probability Letters* **46**, no. 2, 133–147.

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

[19] 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.

[20] Lindsay, J. (1974a). Comparison of probabbility distributions. *J. Royal Statist. Soc. Ser. B* **36**, 38–47.

[21] Lindsay, J. (1974b). Construction and comparison of statistical models. *J. Royal Statist. Soc. Ser. B* **36**, 418–425.

[22] Loader, C. R. (1996). *Local likelihood density estimation.* Academic Press.

[23] Mason, D. (1982). Laws of large numbers for sums of extreme values. *Ann. Probab.*, **10** 754–764.

[24] Mercurio, D. and Spokoiny, V. (2000) Statistical inference for time-inhomogeneous volatility models. WIAS-Preprint No. 583.

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

[26] Müller, H.G. and Song, K.S. (1994). Maximum estimation of multidimensional boundaries. *J. Multivariate Anal.* **50**, no.2, 265–281.

[27] Polzehl, J. and Spokoiny, V. (2000). Adaptive weights smoothing with applications to image segmentation. *J. of Royal Stat. Soc.*, **62**, Series **B**, 335–354.

[28] Polzehl, J. and Spokoiny, V. (2003a). Image denoising: pointwise adaptive approach. *Annals of Statistics*, **62**, in print.

[29] Polzehl, J. and Spokoiny, V. (2003b). Varying coefficient regression modeling by adaptive weights smoothing. WIAS-Preprint 818.

[30] Qiu, P. (1998). Discontinuous regression surface fitting. *Annals of Statistics* **26** no. 6, 2218–2245.

[31] Reiss, R.-D. (1989). *Approximate distributions of order statistics: with applications to nonparameteric statistics*. Springer.

[32] Seneta, E. (1976). *Regularly varying Functions*. Lectute Notes in Mathematics, Vol. 508. Springer.

[33] Spokoiny, V. (1998). Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. *Ann. Statist.*, **26** (1998) no. 4, 1356–1378.

[34] Staniswalis, J.C. (1989). The kernel estimate of a regression function in likelihood-based models. *Journal of the American Statistical Association*, **84** 276–283.

[35] Tibshirani, J.R., and Hastie, T.J. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, **82** 559–567.