DFG–Schwerpunktprogramm 1114

Mathematical methods for time series analysis and digital image processing

Local likelihood modeling by adaptive weights smoothing

Joerg Polzehl **Vladimir Spokoiny**

Preprint 27

Preprint Series DFG-SPP 1114

Preprint 27 **February 2003**

The consecutive numbering of the publications is determined by their chronological order.

The aim of this preprint series is to make new research rapidly available for scientific discussion. Therefore, the responsebility for the contents is solely due to the autors. The publications will be distributed by the co-ordinator of the DFG-Schwerpunktprogramm 1114 and by the authors.

ISBN 3-88722-564-3

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

February 6, 2003

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.

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

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

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

*This work is partially supported by the Deutsche Forschungsgemeinschaft (DFG) - SPP 1114 *Mathematical methods for time series analysis and digital image processing*

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 the local neighborhoods around every point X_i and a description of the local model within such neighborhoods by assigning weights to every point 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 the algorithms proposed in this paper is available as a contributed package (aws) of the R-Project for Statistical Computing form (http://www.r-project.org/).

The paper is organized as follows. Section 2 describes the considered model and presents the main examples. Different methods of local modeling are discussed in Section 3. The local likelihood AWS procedure is given in Section 4. Section 5 demonstrates how the AWS method can be applied to the problem of density estimation in $I\!R^d$ for $d \leq 3$. Section 6 explains how AWS can be applied to volatility estimation of financial assets. The classification problem is considered in Section 7. Estimation of the tail-index parameter by the AWS method is discussed in Section 8. Section 9 discusses the 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 proposed method starting from a preliminary discussion. Suppose we are given random data $Z_1, \ldots, Z_n$ of the form $Z_i = (X_i, Y_i)$. Here the X_i's are valued in a metric space $\mathcal{X}$ and determine a location. Each Y_i, valued in another metric space $\mathcal{Y}$, is viewed as "observation at X_i'. For ease of exposition, we restrict ourselves to the case of independent Z_i. We also suppose that the distribution of each "observation" Y_i depends on the "location" X_i via a finite dimensional parameter θ which may depend on the location X_i. We illustrate this set-up by means of a few examples.

Example 2.1. [Local constant Gaussian regression] Let $Z_i = (X_i, Y_i)$ with $X_i \in \mathbb{R}^d$ and $Y_i \in \mathbb{R}$ 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. [Local Bernoulli (Binary response) model] Let again $Z_i = (X_i, Y_i)$ with $X_i \in \mathbb{R}^d$ 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 problems and digital imaging.

Example 2.3. [Local 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. [Local 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 the approximation of the density model, obtained by a binning procedure.

Example 2.5. [Local volatility model] The observations Y_t for the discrete time $t = 1, 2, \ldots$ follow the conditional heteroscedastic model $Y_t = \sigma_t \varepsilon_t$ where 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.

Our set-up can be described by the following general *varying coefficient* parametric model. Let $(P_\theta, \theta \in \Theta)$ be a family of probability measures on $\mathcal{Y}$ where Θ is a subset in a finite-dimensional space $\mathbb{R}^m$. We assume that the family is dominated by a measure P and denote $p(y, \theta) = dP_\theta / dP(y)$. Moreover, we assume that all the measures P_θ

are absolutely continuous w.r.t. each other and write $dP_\theta/dP_{\theta'}(y) = p(y,\theta)/p(y,\theta')$ for every pair $\theta, \theta' \in \Theta$.

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, Ruppert and Welsh (1998), Cai, Fan and Yao (2000). Our approach is based on a slightly different assumption of *local homogeneity*: for every point $x \in \mathcal{X}$ there exists a local neighborhood of x in which the parameter θ is nearly constant. This assumption leads to an approximation of the function $\theta(\cdot)$ by a constant within this neighborhood. However, in the contrary to the classical local approach, we allow for an arbitrary shape and size of the local neighborhoods. This helps to consider in an unified way the models with smoothly varying parameters and the "piecewise smooth" models whose parameters may jump with locations. Particular cases of the latter models are "change point" models and non-smooth images. The global parametric model is also naturally incorporated in this framework when the local neighborhood of every point coincides with the whole space.

The procedure we describe below attempts to recover this neighborhood from the data. Afterwards, the value of $\theta(x)$ can be estimated from the observations with X_i lying in this neighborhood by a local maximum likelihood method. In the special case of a global homogeneous model with $\theta(x)$ constant, this would lead to a global parametric estimate of this parameter.

To simplify the exposition, we do not consider the case when the distribution of Y_i depends on some nuisance parameter η. A specific example is given by regression with unknown error distribution. Extensions of the method to such a situation are straightforward.

The next section discusses the notions of *global* and *local* likelihood modeling.

3 Global and 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 M-estimate $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ of θ which is defined by minimization of a sum $\sum_{i=1}^n M(Y_i, \theta)$ with some function $M(\cdot, \theta)$:

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

Particular examples are given by the log-likelihood estimate, with $M(y, \theta) = -\log p(y, \theta)$ being the minus log of the density $p(y, \theta)$ of P_θ, or by the least squares (least absolute deviations) estimate with $M(y, \theta) = |y - f(x, \theta)|^2$ (resp. $|y - f(x, \theta)|$). Here $f(x, \theta)$ is a parametrically specified mean (resp. median) regression function.

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 a local model $\mathcal{L}(Y_i) = P_{\theta(x)}$ described by the parameter $\theta(x)$, for all X_i from the local neighborhood of the point x.

3.1 Localization methods

The assumption of a local parametric structure leads to the local estimate of $\theta(x)$ that is obtained from the observations which belong to this local model. An important question under such an approach is how the local model is defined. Below we discuss three possibilities to localize the considered model.

Localization by a bandwidth: Let $\rho(x, x')$ be the metric in $\mathcal{X}$. Given a *bandwidth* h and a *kernel* function $K(u)$ for $u \in \mathbb{R}_+$, define a local model at x using the *location penalty* $\boldsymbol{l}_{i,h} = h^{-2}\rho^2(x, X_i)$ and assigning a weight $w_{i,h}(x) = K(\boldsymbol{l}_{i,h})$ to every observation X_i. This local model leads to the local M-estimate

$$\widehat{\theta}_h(x) = \operatorname*{arginf}_{\theta \in \Theta} \sum_{i=1}^{n} w_{i,h}(x) M(Y_i, \theta). \tag{3.1}$$

If the kernel K is supported on $[0, 1]$, that is, $K(u) = 0$ for $u \geq 1$, only the points X_i from the ball $U_h(x)$ with the radius h and the center at x get positive weights and enter into the considered local model. The bandwidth h in (3.1) describes the degree of locality of the model see e.g. Tibshirani and Hastie (1987), Cleveland, Grosse and Shyu (1991) or Fan and Gijbels (1996). An optimal or "ideal" choice of the bandwidth h can be defined as the largest h such that the underlying function $\theta(\cdot)$ is well approximated by a constant within the spherical neighborhood of radius h around x.

Localization by a window: Localization by a bandwidth restricts the original global model to the ball with the radius h around the point x. Such a local model is isotropic in the sense, that all the directions in the space $\mathcal{X}$ are equally localized. Some statistical problems like estimation of univariate functions with discontinuites (see Spokoiny, 1998), multivariate functions with anisotropic smoothness properties (see Kerkyacharian, Lepski and Picard, 2001) or image denoising (see Polzehl and Spokoiny, 2003) require anisotropic local models. The underlying structural assumption can be formulated as follows: *for every point x, the function $\theta(\cdot)$ can be well approximated by some θ from Θ within a region $U(x)$ containing x.*

Given a window U containing the point of estimation x, define a local model simply by restricting to observations with $X_i \in U$. This leads to a local M-estimate

$$\widehat{\theta}_U(x) = \operatorname*{arginf}_{\theta\in\Theta} \sum_{i=1}^{n} w_{i,U} M(Y_i, \theta) \tag{3.2}$$

where $w_{i,U} = \mathbf{1}(X_i \in U)$. Statistical inference under such a structural assumption focuses on searching for every point x for the largest neighborhood $U = U(x)$ where the hypothesis of structural homogeneity is not rejected.

Localization by weights: The most general approach is to localize *by weights*. For the reference point x, the corresponding local model is described by assigning to every observation Y_i at X_i some nonnegative weight $w_i = w_i(x) \leq 1$. In what follows we identify the diagonal weight matrix $W = \mathrm{diag}\{w_1, \ldots, w_n\}$ and the local model defined by these weights. Such a local model leads to the local M-estimate of the form

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

For the specific example of a local constant regression, a method for constructing such a local model is discussed in PS2000. Polzehl and Spokoiny (2002) generalises this method to local polynomial regression. The main advantage of local constant and local polynomial regression modeling is that a closed form expression for the local estimate $\widehat{\theta}(x)$ is available as well as its confidence regions.

Here we consider a general parametric structure and restrict ourselves to the local maximum likelihood estimate $\widehat{\theta}(x)$. This allows to utilize the Wilks phenomenon for assessing confidence regions for this estimate, see Fan, Zhang and Zhang (2001). We denote by W the diagonal matrix with the diagonal entries w_i and use the notation

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

where θ' is an arbitrary point in Θ. Then $\widehat{\theta}(x) = \widehat{\theta} = \operatorname{argsup}_\theta L(W, \theta, \theta')$ for any θ'. The Wilks phenomenon means that under the parametric hypothesis for the considered local model (at least for the case when the weights w_{ij} are either zero or one) the distribution of $2L(W, \widehat{\theta}, \theta)$ under the parametric model with the true parameter θ is approximately χ^2 and is asymptotically independent of θ.

3.2 Local exponential family

The examples introduced in Section 2 can be considered in a unified way as particular cases of local exponential family distributions. This means that all the measures P_θ from this family are dominated by a σ-finite measure P on $\mathcal{Y}$ and the density functions $p(y, \theta) = dP_\theta / dP(y)$ of the form $p(y, \theta) = e^{U(y)C(\theta) - B(\theta)}$ where $C(\theta)$ and $B(\theta)$ are some nonnegative functions, $U(y)$ is a known function of the observation y and the parameter θ is defined by the equations $\int p(y, \theta) P(dy) = 1$ and $\boldsymbol{E}_\theta U(Y) = \int U(y) p(y, \theta) P(dy) = \theta$. One can easily check that 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 localizing matrix $W = \operatorname{diag}\{w_1, \ldots, w_n\}$ the local log-likelihood for the corresponding local model is of the form

$$\begin{aligned} L(W, \theta, \theta') &= \sum_{i=1}^n w_i \log \frac{p(Y_i, \theta)}{p(Y_i, \theta')} \\ &= \bigl(C(\theta) - C(\theta')\bigr) \sum_{i=1}^n w_i U(Y_i) - \bigl(B(\theta) - B(\theta')\bigr) \sum_{i=1}^n w_i \\ &= S\bigl(C(\theta) - C(\theta')\bigr) - N\bigl(B(\theta) - B(\theta')\bigr) \end{aligned}$$

where

$$N = \sum_{i=1}^n w_i \,, \qquad S = \sum_{i=1}^n w_i U(Y_i) \,.$$

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

$$L(W, \widehat{\theta}, \theta') = N \left[\widehat{\theta} \left(C(\widehat{\theta}) - C(\theta') \right) - \left(B(\widehat{\theta}) - B(\theta') \right) \right] = NQ(\widehat{\theta}, \theta').$$

Table 1 provides the statistics $U(y)$ and the Kullback-Leibler distance $Q(\theta, \theta')$ for the examples from Section 2.

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	$\frac{1}{2}(\theta/\theta' - 1 - \log(\theta/\theta'))$

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 Measuring the difference between two local models

Consider two local models corresponding to points X_i and X_j and defined by diagonal weight matrices W_i and W_j. Suppose for a moment that the structural assumption is fulfilled for each of these two, that is, the parameter function $\theta(\cdot)$ is nearly constant within each model. We aim to answer the question whether these two local models can be put into one common parametric model. This can be done testing the hypothesis that the values $\theta_i = \theta(X_i)$ and $\theta_j = \theta(X_j)$ describing two local models coincide.

We use the notation from the previous section. The local maximum likelihood estimate $\widehat{\theta}_i$ for the local model corresponding to a diagonal matrix $W = \mathrm{diag}\{w_1, \ldots, w_n\}$, is defined for any θ' by the local optimization problem

$$\widehat{\theta}_i = \operatorname*{argsup}_{\theta \in \Theta} L(W_i, \theta, \theta') = \operatorname*{argsup}_{\theta \in \Theta} \sum_{j=1}^{n} w_{ij} \log \frac{p(Y_j, \theta)}{p(Y_j, \theta')}$$

The value $N_i = \sum_{j=1}^{n} w_{ij}$ can be interpreted as the sample size for the local model W_i. To compare two local models W_i and W_j we utilize the likelihood-ratio test statistic corresponding to the hypothesis that the parameters θ_i and θ_j for two local models coincide. First we consider the situation when both matrices W_i and W_j have

zero-one diagonal 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

$$T_{ij}^{\circ} = \max_{\theta} L(W_i,\theta,\theta') + \max_{\theta} L(W_j,\theta,\theta') - \max_{\theta} L(W_i+W_j,\theta,\theta') \qquad (3.3)$$

where $\widehat{\theta}_{ij} = \operatorname{argsup}_\theta L(W_i+W_j,\theta,\theta')$ is the maximum likelihood estimate corresponding to the combined model which is obtained by summing the weights from both models. The value T_{ij}° characterizes the difference between the two models in the statistical sense: if T_{ij}° is larger than some prescribed value λ, then these two models are significantly different in the value of the underlying parameter θ.

The value T_{ij}° is "symmetric" w.r.t. the local models located at the points X_i and X_j in the sense that $T_{ij}^{\circ} = T_{ji}^{\circ}$. However, in the "unbalanced situation" when the "sample sizes" $N_i = \operatorname{tr} W_i$ and $N_j = \operatorname{tr} W_j$ are essentially different, the contribution of every local model into the value T_{ij}° is also essentially different.

For instance, in the case of local Gaussian regression,

$$T_{ij}^{\circ} = \frac{N_i N_j}{N_i+N_j}\left(\widehat{\theta}_i - \widehat{\theta}_j\right)^2 .$$

In the situation when e.g. N_j/N_i is close to zero, $T_{ij}^{\circ} \approx N_j\left(\widehat{\theta}_i-\widehat{\theta}_j\right)^2 \approx N_j\left(\theta-\widehat{\theta}_j\right)^2$. (Since the "sample size" N_i is large, it is not restrictive to suppose here that $\widehat{\theta}_i$ is a "good" estimate of $\theta=\theta(X_i)$ i.e. $\widehat{\theta}_i-\theta\approx 0$.) This means that the contribution of the model with smaller sample size to the value T_{ij}° strongly dominates.

In our procedure, described in the next section, such a statistic based on the estimates $\widehat{\theta}_i$ and $\widehat{\theta}_j$ is applied to decide about the weight w_{ij} with which the observation Y_j at X_j enters in the local model W_i. If the variability of the estimate $\widehat{\theta}_j$ is much higher than the variability of $\widehat{\theta}_i$ (that is, if $N_i \gg N_j$), then we apply the rule for computing the weight w_{ij} in a more conservative way. Namely, when computing the value T_{ij} which determines the weight w_{ij}, we artificially increase the "sample size" N_j by multiplying the weights for the second model at X_j with some factor α and then optimize the resulting test statistic w.r.t. this parameter. The use of the factor α leads to the test

statistics

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

where

$$\widehat{\theta}_{ij}(\alpha) = \operatorname*{argsup}_{\theta} L(W_i + \alpha W_j, \theta, \theta') = \operatorname*{argsup}_{\theta} \sum_{l=1}^{n} (w_{il} + \alpha w_{jl}) \log \frac{p(Y_l, \theta)}{p(Y_l, \theta')}.$$

The application of $\theta' = \widehat{\theta}_j$ yields

$$T_{ij}(\alpha) = L(W_i, \widehat{\theta}_i, \widehat{\theta}_j) - L(W_i + \alpha W_j, \widehat{\theta}_{ij}(\alpha), \widehat{\theta}_j)$$

implying

$$T_{ij}(\alpha) \leq T_{ij} = L(W_i, \widehat{\theta}_i, \widehat{\theta}_j) = \sup_{\theta} L\left(W_i, \theta, \widehat{\theta}_j\right).$$

Moreover, it is easy to check that $T_{ij} = \lim_{\alpha \to \infty} T_{ij}(\alpha)$. This expression will be used in the procedure to measure the statistical difference between the local model at point X_i and the other model at point X_j. Note that this expression is essentially asymmetric, that is, $T_{ij} \neq T_{ji}$. A "symmetrized" version is given by $T^s_{ij} = (T_{ij} + T_{ji})/2$.

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

$$T_{ij} = N_i Q(\widehat{\theta}_i, \widehat{\theta}_j). \tag{3.4}$$

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

4 Adaptive weights smoothing

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

4.1 Preliminaries

The basic assumption of the proposed approach is that for every point X_i, there exists a vicinity of x in which the underlying model described by the function $\theta(x)$ can be well approximated by a parametric model with the constant parameter θ. The idea of the procedure is to describe simultaneously the local models for all points X_i by assigning for every point X_i a weight w_{ij} to every observation Y_j at another point X_j.

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}:

$$\widehat{f}(X_i) = \sum_{\ell=1}^{n} w_{i\ell} Y_\ell \Big/ \sum_{\ell=1}^{n} w_{i\ell}. \tag{4.1}$$

These weights w_{ij} are calculated iteratively, so that the estimate from the previous iteration is used to determine the new weights w_{ij} which in turn leads to the new estimates $\widehat{f}(X_i)$ due to (4.1). For the initial step, the estimate $\widehat{f}^{(0)}(X_i)$ is calculated using the data from a small neighborhood $U_i^{(0)}$ of the point X_i. At each iteration k a larger neighborhood $U^{(k)}(X_i)$ 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 at the previous iteration. Note that under the local constant assumption $f(x) = \theta$, the value θ uniquely determines the model and to comparing the values $\widehat{f}^{(k-1)}(X_i)$ and $\widehat{f}^{(k-1)}(X_j)$ is equivalent to the comparison of two local constant models.

An extension of this approach to the more general local parametric assumption compares two local models described by the weights $W_i^{(k-1)} = \operatorname{diag}\left\{w_{i1}^{(k-1)}, \ldots, w_{in}^{(k-1)}\right\}$ and $W_j^{(k-1)} = \operatorname{diag}\left\{w_{j1}^{(k-1)}, \ldots, w_{jn}^{(k-1)}\right\}$ when determining the weight $w_{ij}^{(k)}$. This can be done using the proposal from Section 3.3.

In addition we extend the original AWS procedure by introducing a memory parameter η such that the new weight $w_{ij}^{(k)}$ at the step k is defined 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 $\widetilde{w}_{ij}^{(k)}$.

4.2 The procedure

Now we present a formal description. Important ingredients of the method are: kernels K_l and K_s, 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 4.4. The procedure reads as follows:

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

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

For every i, set $\widehat{\theta}_i^{(0)} = \widehat{\theta}^{(0)}$ and define $W_i^{(0)}$ as the unit matrix. Set $k = 1$.

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

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

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

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

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

  and define the weight $w_{ij}^{(k)}$ as

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

  Denote by $W_i^{(k)}$ the diagonal matrix whose diagonal elements are $w_{ij}^{(k)}$, that is, $W_i^{(k)} = \mathrm{diag}\{w_{i1}^{(k)}, \ldots, w_{in}^{(k)}\}$, and similarly $\widetilde{W}_i^{(k)} = \mathrm{diag}\{\widetilde{w}_{i1}^{(k)}, \ldots, \widetilde{w}_{in}^{(k)}\}$.

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

$$\widehat{\theta}_i^{(k)} = \operatorname*{argsup}_{\theta\in\Theta} L(W_i^{(k)}, \theta, \theta') = \operatorname*{argsup}_{\theta\in\Theta} \left[\eta L(W_i^{(k-1)}, \theta, \theta') + (1-\eta) L(\widetilde{W}_i^{(k)}, \theta, \theta')\right].$$

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 The case of a local exponential family

We now discuss some features of the procedure for the case when $\{P_\theta\}$ is an exponential family, see Section 3.2. This holds for all the examples considered in this paper.

Statistical penalty: The statistical penalty $\boldsymbol{s}_{ij}^{(k)}$ from (4.2) can, in this case, be represented in the form

$$\boldsymbol{s}_{ij}^{(k)} = \lambda^{-1} L\big(W_i^{(k-1)}, \widehat{\theta}_i^{(k-1)}, \widehat{\theta}_j^{(k-1)}\big) = \lambda^{-1} N_i^{(k-1)} Q\left(\widehat{\theta}_i^{(k-1)}, \widehat{\theta}_j^{(k-1)}\right).$$

Step 2 of the procedure: The local MLE $\widehat{\theta}_i$ can be represented in the form $\widehat{\theta} = \mathrm{argsup}_\theta L(W_i, \theta, \theta') = S_i/N_i$ where $N_i = \sum_{j=1}^n w_{ij}$ and $S_i = \sum_{j=1}^n w_{ij} U_j$. In our examples, $U_j = Y_j$ for local Gaussian, Bernoulli, Poisson and exponential models and $U_j = Y_j^2$ for the local volatility model. Therefore, in the estimation step, the new estimate $\widehat{\theta}_i^{(k)}$ can be written as

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

with

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

Initialization: The initial estimates $\widehat{\theta}_i^{(0)}$ coincide with the global parametric MLE for all i. They are obtained as $\widehat{\theta}_i^{(0)} = S_i^{(0)}/N_i^{(0)} = \sum_{i=1}^n U_j/n$.

Numerical complexity: The numerical complexity of procedure is easily analyzed. If the localization kernel K_l 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 $nM^{(k^*)}$.

The use of the "memory" parameter η: The original AWS procedure from PS2000 does not involve the parameter η (it corresponds to $\eta = 0$) when computing the new weights $w_{ij}^{(k)}$, or equivalently, the new estimate $\widehat{\theta}_i^{(k)}$. Instead it contained one additional *control* step in which the new estimate $\widehat{\theta}_i^{(k)}$ is compared with all the previous estimates $\widehat{\theta}_i^{(k')}$ for $k' < k$. 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, see Section 9 for more details. In the case of local exponential family, this control step will accept the estimate $\widehat{\theta}_i^{(k)}$ only if

$$N_i^{(k')} Q(\widehat{\theta}_i^{(k')}, \widehat{\theta}_i^{(k)}) \le \tau, \qquad k' = 1, \ldots, k-1, \tag{4.3}$$

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. However, as shown in our numerical results, the usefulness of the control step for practical purpose is questionable. The use of the "memory" parameter η can be regarded as a soft version of the control step.

4.4 Choice of parameters

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

Kernels K_s and K_l: The kernels K_s and K_l must fulfill $K_s(0) = K_l(0) = 1$, with K_s decreasing and K_l non-increasing on the positive semiaxis. We recommend to take $K_s(u) = e^{-u} I_{\{u \le 6\}}$. We also recommend to apply a compactly supported localization kernel K_l to reduce the computational effort of the method. PS2000 applied a uniform kernel, here we apply the triangle kernel $K_l(u) = (1-u)_+$.

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 the sensitivity to changes of the local structure. The use of the memory parameter also guarantees that $Q(\widehat{\theta}_i, \widehat{\theta}_j) < \infty$. Our default choice is $\eta = 1/2$.

Initial bandwidth $h^{(1)}$, parameter a and maximal bandwidth h^*: The initial bandwidth $h^{(1)}$ should be reasonably small. In most examples we select $h^{(1)}$ such that every ball $U_i^{(1)}$ with center X_i and radius $h^{(1)}$ contains only one 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 the factor a_{grow}. If X_i are from the unit cube in the space $\mathbb{R}^d$, then the parameter a can be taken as $a = a_{grow}^{1/d}$. Our default choice is $a_{grow} = 1.25$.

The maximal bandwidth h^* can be taken very large. However, this parameter can be used to bound the numerical complexity of the procedure, see Section 4.3. 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.

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. 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, where the underlying parameters for every two local models coincide, 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 the minimal value of λ still providing a prescribed probability of getting the global model at the end of the iteration process for the parametric model $\theta(x) = \theta$ using Monte-Carlo simulations. A theoretical justification is given by Theorem 9.1 in the next section, 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.

Our default value is $\lambda = t_\alpha(\chi^2_1)$, that is the α-quantile of the χ^2 distribution with 1 degree of freedom, where 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.

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

5 Application to nonparametric density estimation

Suppose that the observations $Z_1, \ldots, Z_L$ were sampled independently from some unknown distribution P on $\mathbb{R}^d$ having a density $f(x)$ w.r.t. the Lebesgue measure. The problem of adaptive estimation of f can be successfully attacked by the AWS method. Here we consider the case with a relatively small d, e.g. $d \leq 3$. The case of a larger d can be considered as well but requires a separate treatment.

Without loss of generality we suppose that the observations are located in the cube $[0,1]^d$. Note that we do not assume that f is compactly supported or that f is bounded away from zero on $[0,1]$. 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 variables $Y_1, \ldots, Y_n$ are not independent because of the obvious equation $Y_1 + \ldots + Y_n = L$. The joint distribution of $Y_1, \ldots, Y_n$ is described by the multinomial law. However, 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 Tibshirany (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 have to specify the choice of the bin length δ. A reasonable choice is given by the rule $\delta = c/K$ where K is the smallest integer satisfying $K^d \geq L$ and $c \leq 1$. The use of a small c helps to reduce the discretization error but increases the "sample size" n and therefore, the computational effort by factor c^{-d}.

We illustrate the performance of the method by means of two simulated examples with piecewise smooth density function. We start with the univariate case.

Example 5.1. We generate $n = 200$ observations from the univariate distribution with density

$$f(x) = \begin{cases} 1.5 & x \in [0, .25) \wedge x \in [.75, 1] \\ .5 & x \in [.25, .75) \\ 0 & \text{otherwise} \end{cases}$$

The density estimate (solid line) provided in the left part of Figure 1 was obtained using an equispaced grid of 440 intervals of length $\delta = 0.0025$ and range $(-.1, 1.1)$. The true density is given for comparison (dotted line). A large value $h^* = 2000\delta = 5$ was used to have a vanishing influence of the location penalty. The symmetrized version of the stochastic penalty was applied. All other parameters equal to their defaults. A typical example is illustrated in the left of Figure 1. Note the almost perfect restoration of the unknown piecewise constant density.

The next example presents a piecewise smooth bivariate density having 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.

Example 5.2. We generate $n = 2500$ observations from the 2-dimensional density

$$f(x_1, x_2) = 7.5x_1(1 - x_1^2 - x_2^2)_+ I_{\{x_1 \geq 0, x_2 \geq 0\}}$$

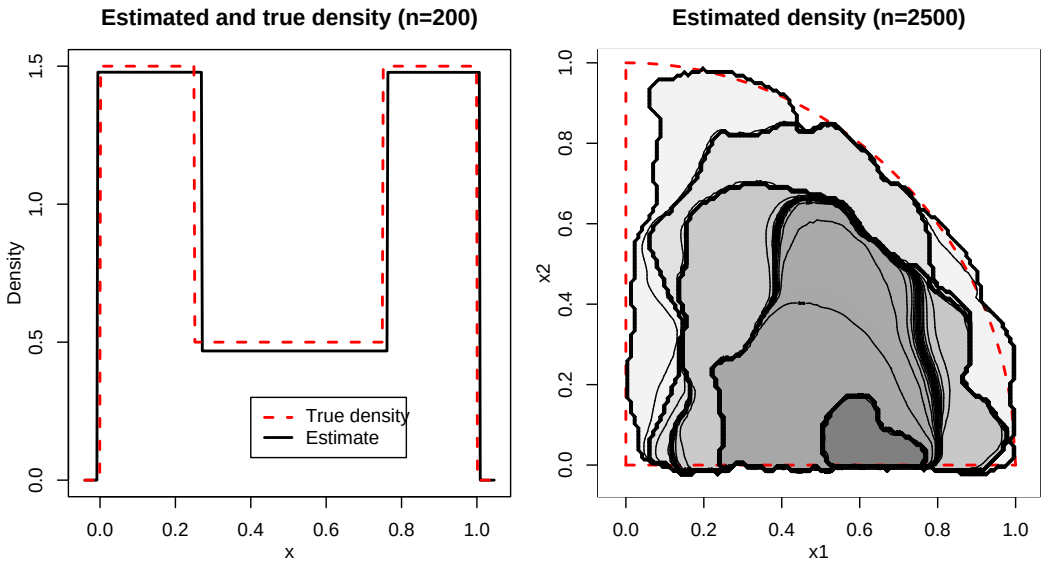

Figure 1: Density estimation: univariate example (left) and bivariate example (right). Solid lines correspond to estimates and dashed lines to the true densities.

The right part 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 symmetrized version of the stochastic penalty was used with $h^* = 400\delta = 4$ and defaults for all other parameters.

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) and is similar to the case of the edge estimation in imaging, see PS2000 and Polzehl and Spokoiny (2003).

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:

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

where ε_t are *innovations* which are conditionally on $\mathcal{F}_{t-1} = \sigma(S_1, \ldots, S_{t-1})$ standard normal distributed and σ_t is the 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 behaviour of our procedure.

Example 6.1. First we produce an artificial series of returns R_t of length $T = 500$ following the model $R_t = \sigma_t \varepsilon_t$ with

$$\sigma_t = 1 + I_{\{t \geq 125\}} - 1.5 I_{\{t \geq 250\}} + 0.5 I_{\{t \geq 375\}} .$$

Figure 2 displays the absolute values $|R_t|$ together with the true volatility σ_t and estimates of the volatility σ_t obtained by the symmetric and asymmetric version of AWS, both with default parameters and maximal bandwidth $h^* = 2000$. Both procedures demonstrate 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 symmetric version sometimes randomly segments small regions, Figure 2 shows a typical example.

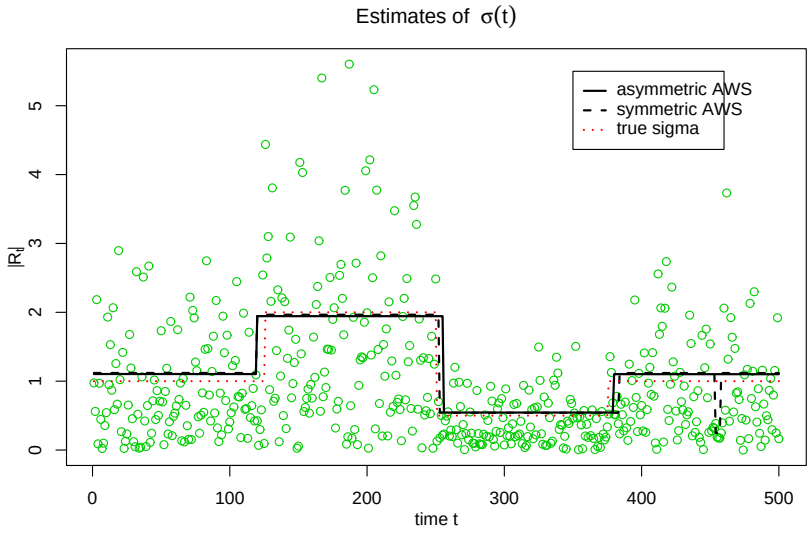

Figure 2: Volatility estimation: Artificial data set with true volatility function and estimates obtained by the asymmetric and symmetric version of AWS.

Example 6.2. In the second example we analyze the exchange rate between the US $

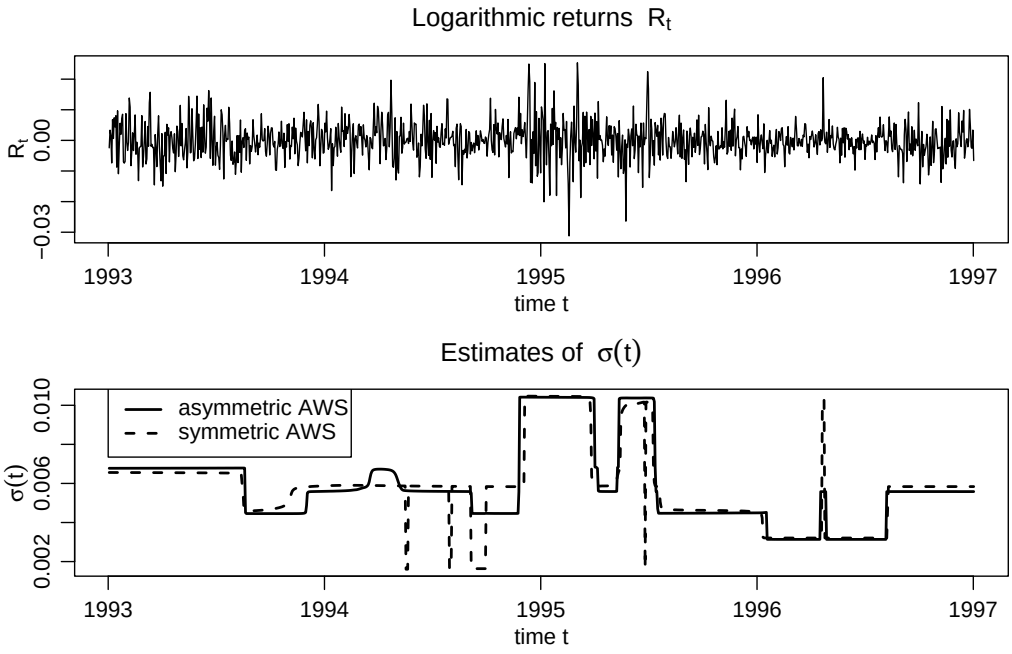

Figure 3: Volatility estimation: Returns for exchange rate between the US $ and the German DM and estimates obtained by the asymmetric and symmetric version of AWS.

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, Vancouver 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 the symmetric and asymmetric version of AWS for the time period from January 1993 to December 1997.

Note that both estimates indicate time-inhomogeneity of the volatility and that most discontinuities occur at the same points in time for both estimates. Again a different behavior of the asymmetric and symmetric version can be observed, with the symmetric version singling out several small time intervals with unusually low or high volatility.

7 Application to classification

We consider the following discrimination problem for two populations. One observes a training sample (X_i, Y_i), $i = 1, \ldots, n$, with X_i valued in a metric space $\mathcal{X}$ with known class assignment $Y_i \in \{0, 1\}$. The goal is to construct a discrimination rule to decide for every point $x \in \mathcal{X}$ whether it belongs to class "zero" or class "one'.

The standard approach in classification is based on the Bayes discrimination rule.

Suppose that for $k = 0, 1$, all the X_i's with $Y_i = k$ (that is, all the points from the k-th population) are randomly sampled from a distribution F_k with the density $f_k(x)$ with respect to some measure μ on $\mathcal{X}$. Let also π_k be the prior probability of the population $k = 0, 1$. Then the Bayes discrimination rule is

$$\rho(x) = \mathbf{1}\left(\pi_1 f_1(x) \geq \pi_0 f_0(x)\right).$$

Since f_0 and f_1 are usually unknown in practical applications, one first constructs estimates of the densities f_0 and f_1 or of the ratio $f_1(x)/f_0(x)$ and then applies the above rule plugging in the estimated quantities.

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 $\boldsymbol{P}(Y_i = 1) = p(X_i)$. Here the parameter $p(X_i)$ equals to the density ratio $f_1(X_i)/(f_0(X_i) + f_1(X_i))$. The "ideal" discrimination rule for this model is $\rho(x) = \mathbf{1}\left(p(x) \geq \pi_0/(\pi_0 + \pi_1)\right)$. Since the function $p(x)$ is usually unknown, one applies this rule with p 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 estimate and the kernel estimate. Given a natural k, define for every point x in $\mathcal{X}$ the subset $\mathcal{D}_k(x)$ of the design $X_1, \ldots, X_n$, including the k closest to x points with respect to the metric $\rho(x, x')$ in $\mathcal{X}$. Then the k-nearest neighbors 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 value k for the first and the bandwidth h for the second method) and meet the "curse of dimensionality" problem: high dimensional data are very sparse which leads to a large estimation bias.

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 practical applications, one has to estimate the function p in some other points $X_{n+1}, \ldots, X_{n+m}$. This extension can be naturally incorporated in the procedure by applying the procedure to the "extended" sample (X_i, Y_i) for $i = 1, \ldots, n+m$, where Y_i are arbitrary for $i > n$. At every step of the procedure, all the weights $w_{ij}^{(k)}$ with $j > n$ are set to zero, because the corresponding "observations" Y_j are not informative.

The kernel estimate is an extreme case of the AWS estimate, it is computed in case of parameters $\lambda = \infty$ and $h^* = h$. The k-nearest neighbors (k-NN) estimate can be obtained by a slightly modified AWS procedure, that uses the nearest neighbor idea for the location penalty.

Example 7.1. To illustrate the behaviour of AWS in this context we use the data from a simulated two-dimensional discriminant analysis example from Hastie, Tibshirani and Friedman (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 Hastie, Tibshirani and Friedman (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 version of AWS with $\lambda = 3.28$, i.e. the 0.93-quantile of χ^2_1, 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.

Figure 5 shows graphs of error rates 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. 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, in particular, for our default setting. The choice of a smoothing parameter for the other methods is rather critical and a suboptimal choice leads to a significant increase of the error rate.

8 Application to tail index estimation problem

Let $X_1, ..., X_n$ be a sample from the distribution F. The target of the analysis is the tail behaviour 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 $L(x)$

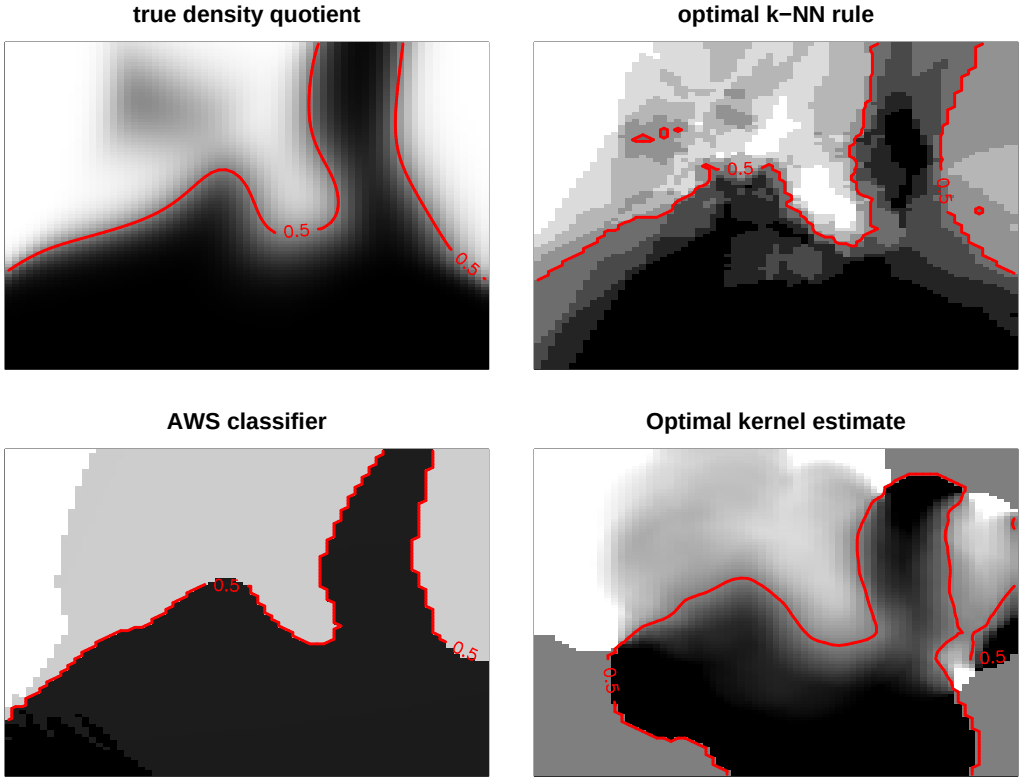

Figure 4: Classification rules obtained by the optimal Bayes decision, the best k-nearest neighbor rule, adaptive weights smoothing (AWS) and the best rule based on kernel estimation.

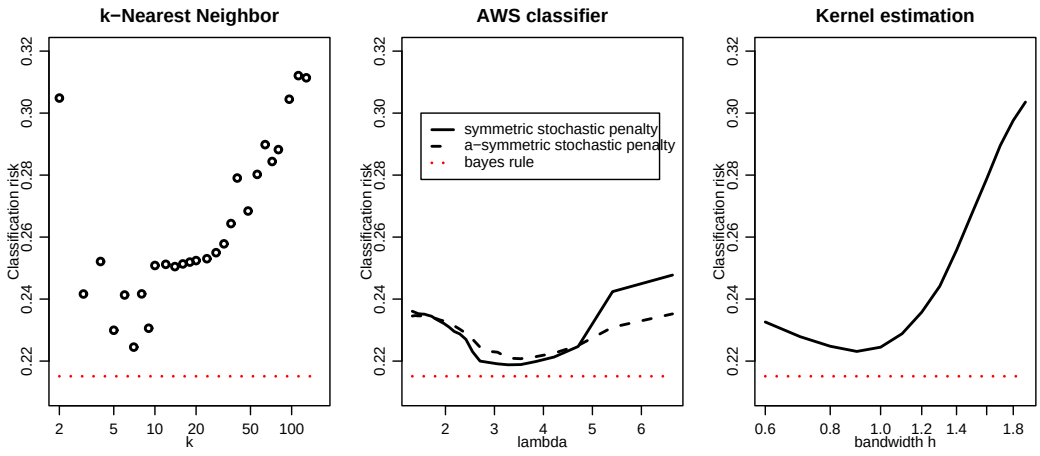

Figure 5: Dependence of the classification error on the main smoothing parameter rules defined by k-nearest neighbor, AWS and kernel estimation.

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 $X_{n,1} \geq ... \geq X_{n,n}$ are the order statistics pertaining to $X_1, ..., X_n$ and 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, Häusler and Mason (1988). However, practical applications of this estimate meet serious problems, see e.g. Embrechts, Klüppelberg and Mikosch (1997, p.351). The main practical difficulty is dealt with the choice of the parameter k. Another problem is related with the treatment of the slowly varying function $L(x)$ which may seriously affect the performance of the estimate, see Embrechts, Klüppelberg and Mikosch (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 $a > 1$ is a fixed real number. Assume that the function F is strictly increasing and has a continuous density f. Define the function $\alpha(x)$ by the equation

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

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

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

The basic condition imposed on the model is that the function $\alpha(x)$, $x > a$, can be approximated by a constant for large values of x. For instance, this is the case when there exists an $\beta > 0$ such that

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

Many regularly varying at infinity d.f.'s F satisfy the assumptions (8.2) and (8.3), see representation theorems in Seneta (1976) or Bingham, Goldie and Teugels (1987).

Our problem can be formulated as follows. Let $X_{n,1} > ... > X_{n,n}$ be the order statistics pertaining to $X_1, ..., X_n$. The goal is to find a natural number k such that on the set $\{X_{n,1}, ..., X_{n,k}\}$ the function $\alpha(x)$, $x \geq a$, can be well approximated 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 that of estimating the index of regular variation β defined by the limit (8.3). Indeed, the value β can be regarded as $\lim_{x\to\infty}\alpha(x)$. However, 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 $\alpha(\cdot)$ at the points X_i will be estimated from the approximating exponential model. Our motivation is somewhat similar to that of Hill (1975). 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 for instance Reiss (1989) or Example 4.1.5 in Embrechts, Klüppelberg and Mikosch (1997). □

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

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

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 vector of unknown parameters. This vector can be estimated by the AWS procedure for the local exponential model, see Example 2.3. The target 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.

Table 3: 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	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	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	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	1.000	1.000	1.000	1.000	1.000

Example 8.1. Tail indices are estimated for four distributions, using the Pareto-distribution with tail index $\beta = 1$, 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. Sample sizes of $n = 100$, $n = 200$, $n = 400$, $n = 800$ and $n = 1600$ are used in each case. Table 2 reports the mean absolute error (MAE) for estimating $\alpha(x_{\max})$, the estimated bias, i.e. the mean of $\widehat{\alpha}_1 - \alpha(x_{\max})$, and the mean value of $\alpha(x_{\max})$, with $\alpha(x)$ defined by (8.1) and $x_{\max}$ the maximal value from the sample. 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. It is worth noting that 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 $|R_{(1)}| = \max_t |R_t|$. The sums of weights for this local model is approximately equal to 277, and the positive weights are effectively 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 asymmetric or symmetric AWS volatility estimate obtained in Example 6.2 equal to 0.1646 and 0.1558, respectively.

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-values of the tail-index estimates for the standardized absolute logarithmic returns are 0.596 and 0.693 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 $\theta(x)$. We start by listing some attractive features of the method which directly follow from the construction and are also justified by our numerical results and applications.

AWS applies in a unified way to a broad class of nonparametric models: The proposed method is very general and its adjustment to the particular situation is trivial in many cases. For all the examples considered in the paper, we applied essentially the same procedure. Sometimes, a preliminary model (data) transformation is required, as in tail index or density estimation.

AWS is dimension free: The dimensionality of the regressors X_i plays absolutely no role for the procedure. This feature is extremely important making it feasible to apply the procedure to e.g. image denoising or inference for high dimensional models.

AWS is computationally straightforward and the numerical complexity can be easily controlled: Indeed, the AWS requires of order nM_{k^*} operations with k^* being the number of iterations and M_k being the corresponding size of the typical neighborhood $U_i^{(k)}$ at the step k. Therefore, the complexity of the method can be controlled simply by restricting k^*, or, equivalently the largest bandwidth h^*, see Section 4.3.

AWS is design adaptive and has no boundary problem: The method proceeds with the given "design" $X_1, \ldots, X_n$, no assumptions or restrictions are imposed on it. A random design, e.g. in density estimation, is treated similarly to the case of a deterministic design, e.g. in image denoising. The local constant modeling applied in the algorithm does not suffer from nonregular design. This feature is important in connection to change point and edge estimation, the produced estimate does not exhibit the typical Gibbs effect (high variability) near discontinuities of most other nonparametric methods.

Now we turn to more involved properties of the method which require a theoretical justification.

9.1 Behaviour 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 it 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 (3.3) in Section 3.3.

In our procedure the statistic T_{ij} from (3.4) 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 $2L(W, \widehat{\theta}, \theta)$ 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 relatively small, even if the global sample size n is large. Corollary 10.1 from the Appendix applied with $z = \rho\lambda$ yields in the homogeneous situation for every local model W_i

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

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

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

leading to the following results.

Theorem 9.1. *Let (A0), (A1) and (A2) be fulfilled. Suppose that* $\theta(X_i) \equiv \theta$. *If* $\lambda \ge C \log n$ *where a constant* C *depends on the kernel* K_s *only, then for every iteration* k

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

Proof. Define ρ by the equation $K_s(\rho) = 1/2$. The bound (9.1) implies for every iteration k

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

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 the location penalty $K_l(\boldsymbol{l}_{ij}^{(k)})$ at the final iteration $k = k_n$ fulfills $K_l(\boldsymbol{l}_{ij}^{(k)}) \approx 1$ for every pair (i,j). By Theorem 9.1 the statistical penalty $K_s(\boldsymbol{s}_{ij}^{(k)}) \ge 1/2$, 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 10.1. □

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 a more strong 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 $2h^{(k)}$:

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

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_s only, then for every iteration k*

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

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 10.1. □

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 would imply 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 the control step, see the discussion at end of Section 4.3.

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 $U_h(x) = \{x' : |x' - x| \le h\}$. We consider the following conditions which are specified for the fixed point x and the bandwidth h:

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

(A4) There are two positive constants $I_* \le I^*$ such that $I_* \le I_{\theta(x')} \le I^*$ for all $x' \in U_{2h}(x)$, where I_θ 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 $C_{X1} \le C_{X2}$ holds

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

(A6) The kernel K_l 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

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

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_s only, and if the bandwidth h fulfills*

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

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

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

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 $1 - 4/n$

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

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 *control step* of the original AWS procedure from PS2000 does, see the discussion at the end of Section 4.3. The procedure presented here applies this control step in a soft form, however 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 (4.3) with* $\tau \ge \lambda$. *Then the last step estimate* $\widehat{\theta}_i$ *fulfills* $N_i^{(k)} Q\big(\widehat{\theta}_i^{(k)}, \widehat{\theta}_i\big) \le \tau$ *and hence, it holds with a probability at least* $1 - 4/n$

$$\left|\widehat{\theta}_i - \theta_i\right| \le Lh + 2\sqrt{\lambda/(I_*\, C_{X1} n h^d)} + 2\sqrt{\tau/(I_*\, C_{X1} n h^d)}. \tag{9.7}$$

Proof. This result is a direct corollary of Theorem 9.3 and (4.3). □

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

10 Appendix

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

We consider the varying coefficient model in which observation Y_i is P_{θ_i}-distributed where θ_i depends on the location X_i. Let also a local model W be described by the weights $w_i \in [0,1]$ for $i = 1, \dots, n$. The corresponding local MLE can be written as $\widehat{\theta} = \sum_{i=1}^n w_i Y_i / \sum_{i=1}^n w_i$. We use the representation $\widehat{\theta} = S/N$ with $S = \sum_{i=1}^n w_i Y_i$, $N = \sum_{i=1}^n w_i$. We also 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 $v = C(\theta)$ and define $\overline{v} = C(\overline{\theta})$ and $D(v) = B(\theta) = B(C^{-1}(v))$. Since $C'(\theta) > 0$, the new parameter v is uniquely defined. By simple analysis $D'(v) = \theta = C^{-1}(v)$ and $D''(v) = 1/C'(\theta) = 1/I(\theta) = 1/I(C^{-1}(v))$. Moreover, $Q(v_1, v_2) = D(v_2) - D(v_1) - (v_2 - v_1)D'(v_1)$ is the Kullback-Leibler distance between two parametric distributions corresponding to the parameters v_1 and v_2. In what follows we also use the notation $q(u|v) = Q(v, v+u) = D(v+u) - D(v) - uD'(v)$.

Theorem 10.1. *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 the equation $q(u|\overline{v}) = \int_0^u xD''(\overline{v}+x)dx = z/N$. If there is some $\alpha > 0$ such that for all $\mu \in (0, 1]$ and all $u \in \mathcal{U}(W, z)$*

$$q(\pm w_\ell \mu u | v_\ell) \leq (1+\alpha) w_\ell \mu^2 q(u|\overline{v}), \qquad \ell = 1, \ldots, n, \tag{10.1}$$

then

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

Remark 10.1. The condition (10.1) can be easily checked in many particular situations. We give two typical examples. The first one corresponds to the homogeneous case when all v_i coincide with their mean $\overline{v}$. Then (10.1) is fulfilled automatically with $\alpha = 0$. Indeed the function $q(\cdot|v)$ satisfies $q'(u|v) = D'(v+u) - D'(v)$ and $q''(u|v) = D''(v+u) = 1/I(C^{-1}(v+u)) > 0$ and thus, it is convex. Since also $q(0|v) = 0$, it holds $q(wa|v) \leq wq(a|v)$ for every $w \in [0, 1]$ and every a implying (10.1) 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 10.1. *If $\theta_i \equiv \theta$, then $\boldsymbol{P}\left(L(W, \widehat{\theta}, \theta) > z\right) \leq 2e^{-z}$ for every $z > 0$.*

In the general inhomogeneous situation, the Taylor expansion yields that $q(wu|v) = D(v+wu) - D(v) - wuD'(v) = 1/2\, w^2 u^2 D''(v + \tau wu)$ for some $\tau \in [0, 1]$. If the Fisher information $I(\theta)$ is bounded from zero and infinity, that is, $I_* \leq I(\theta) \leq I^*$ for all $\theta \in \Theta$, then $1/I^* \leq D''(v) \leq 1/I_*$ for all v and one easily gets for every $u \in \mathcal{U}(W, z)$ that $u^2 \leq 2zI^*/N$. Therefore, the condition (10.1) is certainly fulfilled with a small $\alpha \geq 0$ for the case when all v_i are close to the the mean $\overline{v}$ or when the weights w_i are very small for all ℓ with a large value $|v_\ell - \overline{v}|$.

Proof of Theorem 10.1. The log-likelihood ratio can be rewritten for the new

parameter v as

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

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

Lemma 10.1. *For given z, there exist two values $v^* > \overline{v}$ and $v_* < \overline{v}$ such that*

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

Proof. It holds

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

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

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

or, equivalently,

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

Therefore

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

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

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

for some $v_* < \overline{v}$. □

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

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

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

Lemma 10.2. *For every u and every z*

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

Also

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

Moreover, if u fulfills (10.1) 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

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

The independence of the Y_ℓ's implies

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

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

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

Therefore

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

This inequality applied with $\mu = (1+\alpha)^{-1}$ and (10.1) imply

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

Similarly

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

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 Kulback-Leibler distance, is sufficiently small, then T_{ij}° is with a large probability smaller then $\rho\lambda$ for some $\rho \leq 1$.

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 10.2. *Let $\rho \in (0, 1]$ and $z = \rho\lambda/6$. Let the condition (10.1) 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*

$$N_i Q(\overline{\theta}_i, \overline{\theta}_{ij}) + N_j Q(\overline{\theta}_j, \overline{\theta}_{ij}) \leq \rho\lambda/6 \tag{10.2}$$

imply

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

Proof. It holds

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

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

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

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 10.1 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 10.2 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

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

This inequality with $z = \rho\lambda/6$ and (10.2) imply the assertion. □

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

Theorem 10.3. *Let $\rho \in (0,1]$ and let the condition (10.1) 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*

$$N_i Q(\overline{\theta}_i, \overline{\theta}_{ij}) \geq (6\rho + A)\lambda, \qquad N_j Q(\overline{\theta}_j, \overline{\theta}_{ij}) \geq (6\rho + A)\lambda, \tag{10.3}$$

imply

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

Proof. Similarly to the proof of Theorem 10.2 we use the representation

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

Theorem 10.1 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

$$\boldsymbol{P}\left(T_{ij}^{\circ} < L(W_i, \overline{\theta}_i, \overline{\theta}_{ij}) + L(W_j, \overline{\theta}_j, \overline{\theta}_{ij}) - 2\rho\lambda\right) \leq 2e^{-\rho\lambda/(1+\alpha)}. \tag{10.4}$$

Lemma 10.2 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 (10.1) and (10.3) imply

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

This and a similar inequality for $L(W_j, \overline{\theta}_j, \overline{\theta}_{ij})$ yield the theorem in view of (10.4). □

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 10.2 instead of Corollary 10.1. Let k be such that $h^{(k)} \le h$ and $X_j \in U_h(X_i)$. We apply Theorem 10.2 to the local models $W_i^{(k)}$ and $W_j^{(k)}$. For this we have to check the condition (10.2). 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 $\overline{\theta}_{ij}^{(k)} = \left(N_i^{(k)}\overline{\theta}_i^{(k)} + N_j^{(k)}\overline{\theta}_j^{(k)}\right)/\left(N_i^{(k)} + N_j^{(k)}\right)$. Now the inequality $Q(\theta,\theta') \le I^*|\theta-\theta'|^2/2$ and condition (9.3) imply

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

Theorem 10.2 now applies with some $\alpha \ge 0$, see Remark 10.1, yielding

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

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 10.1.

Let now $h^{(k)} = h$. By (9.4) all the weights $w_{ij}^{(k)}$ for the local model $W_i^{(k)}$ satisfy with a high probability the condition $w_{ij}^{(k)} \ge 0.5 K_l(\boldsymbol{l}_{ij}^{(s)})$. This and Assumption (A5) yield $N_i^{(k)} \ge 0.5 C_{X1} n h^d$. Since also $Q\left(\widehat{\theta}_i^{(k)}, \overline{\theta}_i^{(k)}\right) \ge I_* \left|\widehat{\theta}_i^{(k)} - \overline{\theta}_i^{(k)}\right|/2$, 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] Cleveland, W.S., Grosse, E. and Shyu, W.M. (1991). Local regression model. In *Statistical Models in S* (Chambers, J.M. and Hastie, T.J. eds.) Wadsworth & Brooks, Pacific Grove. 309–376.

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

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

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

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

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

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

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

[13] Grama, I. and Spokoiny, V. (2002). Tail index estimation by local exponential modelling. 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] Kerkyacharian, G., Lepski, O., and Picard, D. (2001). Nonlinear estimation in anisotropic multi-index denoising. *Probab. Theory Relat. Fields* **121** no.2, 137–170.

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

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

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

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

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

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

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

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

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

[27] Polzehl, J. and Spokoiny, V. (2002). Varying coefficient regression modeling by adaptive weights smoothing. Manuscript in preparation.

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

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

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

[31] 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.

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

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

Preprint Series DFG-SPP 1114

http://www.math.uni-bremen.de/zetem/DFG-Schwerpunkt/SP.preprints.html

Reports

1. W. Horbelt, J. Timmer, H.U. Voss, Parameter Estimation in Nonlinear Delayed Feedback Systems from Noisy Data, May 2002 (ISBN: 3-88722-530-9).

2. A. Martin, Propagation of Singularities, July 2002 (ISBN: 3-88722-533-3).

3. T.G. Müller, J. Timmer, Fitting parameters in partial differential equations from partially observed noisy data, August 2002 (ISBN: 3-88722-536-8).

4. G. Steidl, S. Dahlke, G. Teschke, Coorbit Spaces and Banach Frames on Homogeneous Spaces with Applications to the Sphere, August 2002 (ISBN: 3-88722-537-6).

5. J. Timmer, T.G. Müller, I. Swameye, O. Sandra, U. Klingmüller, Modeling the non-linear dynamics of cellular signal transduction, September 2002 (ISBN: 3-88722-539-2).

6. M. Thiel, M.C. Romano, U. Schwarz, J. Kurths, J. Timmer, Surrogate Based Hypothesis Test without Surrogates, September 2002 (ISBN: 3-88722-540-6).

7. K. Keller, H. Lauffer, Symbolic Analysis of High-dimensional Time Series, September 2002 (ISBN: 3-88722-538-4).

8. F. Friedrich, G. Winkler, O. Wittich, V. Liebscher, Elementary Rigorous Introduction to Exact Sampling, October 2002 (ISBN: 3-88722-541-4).

9. S.Albeverio, D.Belomestny, Reconstructing the intensity of non-stationary poisson, November 2002 (ISBN: 3-88722-544-9).

10. O. Treiber, F. Wanninger, H. Führ, W. Panzer, G. Winkler, D. Regulla, An adaptive algorithm for the detection of microcalcifications in simulated low-dose mammography, November 2002 (ISBN: 3-88722-545-7).

11. M. Peifer, J. Timmer, H.U. Voss, Nonparametric Identification of Nonlinear Oscillating Systems, November 2002 (ISBN: 3-88722-546-5).

12. S.M. Prigarin and G. Winkler, Numerical solution of boundary value problems for stochastic differential equations on the basis of the Gibbs sampler, November 2002 (ISBN: 3-88722-549-X).

13. A. Martin, S.M.Prigarin and G. Winkler, Exact numerical algorithms for linear stochastic wave equation and stochastic Klein-Gordon equation, November 2002 (ISBN: 3-88722-547-3).

14. A. Groth, Estimation of periodicity in time series by ordinal analysis with application to speech, November 2002. (ISBN: 3-88722-550-3).

15. H.U. Voss, J. Timmer, J. Kurths, Nonlinear dynamical system identification from uncertain and indirect measurements, December 2002 (ISBN: 3-88722-548-1).

16. U. Clarenz, M. Droske, M. Rumpf, Towards fast non-rigid registration, December 2002. (ISBN: 3-88722-551-1).

17. U. Clarenz, S. Henn, M. Rumpf, K. Witsch, Relations between optimization and gradient flow with applications to image registration, December 2002 (ISBN: 3-88722-552-X).

18. M. Droske, M. Rumpf, A variational approach to non-rigid morphological registration, December 2002 (ISBN: 3-88722-553-8).

19. T. Preusser, M. Rumpf, Extracting motion velocities from 3D image sequences and spatio-temporal smoothing, December 2002 (ISBN: 3-88722-555-4).

20. K. Mikula, T. Preusser, M. Rumpf, Morphological image sequence processing, December 2002 (ISBN: 3-88722-556-2).

21. V. Reitmann, Observation stability for controlled evolutionary variational inequalities, January 2003 (ISBN: 3-88722-557-0).

22. K. Koch, A New Family of Interpolating Scaling Vectors, January 2003 (ISBN: 3-88722-558-9).

23. A. Martin, Small Ball Asymptotics for the Stochastic Wave Equation, January 2003 (ISBN: 3-88722-559-7).

24. P. Maass, T. Koehler, R. Costa, U. Parlitz, J. Kalden, U. Wichard and C. Merkwirth, Mathematical methods for forecasting bank transaction data, January 2003 (ISBN: 3-88722-569-4).

25. D. Belomestny and H. Siegel, Stochastic and self-similar nature of highway traffic data, February 2003 (ISBN: 3-88722-568-6).

26. G. Steidl, J. Weickert, T. Brox, P. Mrazek and M. Welk, On the Equivalence of Soft Wavelet Shrinkage, Total Variation Diffusion, and SIDEs, February 2003 (ISBN: 3-88722-561-9).

27. J. Polzehl and V. Spokoiny, Local likelihood modeling by adaptive weights smoothing, February 2003 (ISBN: 3-88722-564-3).