Weierstraß-Institut
für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Preprint ISSN 0946 – 8633

Local likelihood modeling by adaptive weights smoothing

Jörg Polzehl and Vladimir Spokoiny

Weierstrass Institute for Applied Analysis and Stochastics

E-Mail: polzehl@wias-berlin.de, spokoiny@wias-berlin.de

submitted: 25th November 2002

No. 787
Berlin 2002

2000 *Mathematics Subject Classification.* 62G08.

Key words and phrases. adaptive weights; local likelihood, exponential family, density estimation, volatility, tail index, classification.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Mohrenstraße 39
D — 10117 Berlin
Germany

Fax: + 49 30 2044975
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

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

November 25, 2002

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 estimation of the tail index parameter, classification, density and volatility estimation.

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

AMS 2000 Subject Classification: 62G08,Secondary:

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 the 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 in what follows). The main idea of the AWS estimator is to describe in a data-driven iterative way a maximal possible 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 also dimension free and applies in high dimensional situations. However, the assumption of the regression model with additive errors considered in PS2000 restricts the domain of applications of the AWS method. 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. We also apply in a unified way the AWS method to different problems like density or intensity estimation, classification, tail index estimation and volatility modelling.

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 discusses one important feature of the method which we call the “propagation condition”. Section 6 demonstrates how the AWS method can be applied to the problem of density estimation in $I\!R^d$ for $d \le 3$. Section 7 explains how AWS can be applied to volatility estimation of financial assets. Estimation of the tail-index parameter by the AWS method is discussed in Section 8. The classification problem is considered in Section 9. Section 10 briefly discusses the main advantages of the proposed method. Finally, Section 11 presents one theoretical result about local exponential family models which justifies our adaptive method.

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

Example 2.2. [Local Bernoulli (Binary response) model] Let again $Z_i = (X_i, Y_i)$ with $X_i \in I\!R^d$ and Y_i being 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 $\theta = \theta(X_i)$. Such models are commonly used in the queueing theory, in positron emission tomography (PET), it serves also as the approximation of the density model, obtained by a binning procedure.

Example 2.5. [Local constant 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 3.3 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 density functions on $\mathcal{Y}$ where Θ is a subset in a finite-dimensional space $I\!R^m$. 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 with a ball of some small radius h of the point of estimation, see e.g. Tibshirani and Hastie (1987) or Cai, Fan and Li (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 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 estimator 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 modelling

First we discuss some well known methods of estimation under the global parametric assumption.

3.1 Global parametric estimation

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-estimator $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_n)$ 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 estimator, 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) estimator 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.

3.2 Local estimation

A global parametric assumption can be too restrictive. A 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. This approach includes as particular cases change-point models with piecewise constant function $\theta(x)$ and varying coefficients models with a smooth function $\theta(x)$, see Hastie and Tibshirani (1993), Fan and Zhang (1999), Caroll, Ruppert and Welsh (1998), Cai, Fan and Yao (2000).

The assumption of a local parametric structure leads to the local estimator 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.

3.2.1 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 $u \in I\!R_+$, 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-estimator

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

If the kernel K is supported on $[0,1]$, that is, $K(u) = 0$ for $u \geq 1$, then 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.

3.2.2 Localization by a window

The above discussed localization by 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) or multivariate functions with anisotropic smoothness properties (see Kerkyacharian, Lepski and Picard, 2001), image denoising (see Polzehl and Spokoiny, 2003) require to consider anisotropic local models. The underlying structural assumption can now 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. The test of homogeneity from Polzehl and Spokoiny (2003) is based on the comparison of the estimate $\widehat{\theta}_U$ with similar estimates $\widehat{\theta}_V$ for smaller regions V: the hypothesis of homogeneity is rejected if there is a subregion $V \subset U$ such that the estimates $\widehat{\theta}_U$ and $\widehat{\theta}_V$ differ significantly. If $\widehat{U}$ is the largest non-rejected region (window), then the adaptive estimate of $f(x)$ is $\widehat{f}_{\widehat{U}}$. As shown in Spokoiny (1998) and Polzehl and Spokoiny (2003), the estimate $\widehat{f}_{\widehat{U}}$ possesses nice theoretical properties and demonstrates a reasonable numerical performance. Nonetheless, the approach is computationally very intensive. Additionally, since the selection of the window is carried out independently for every point X_i, this may lead to a high variability of the adaptive estimator.

3.2.3 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 $w_i = w_i(x) \le 1$. Such a local model leads to the local M-estimator 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, such a method for constructing 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 modelling is that a closed form expression for the local estimator $\widehat{\theta}(x)$ is available as well as its confidence regions. Here we consider a general parametric structure and we restrict ourselves to the local maximum likelihood estimator $\widehat{\theta}(x)$ defined by

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

where θ' is an arbitrary point in Θ. This allows to utilize the so called Wilks phenomenon for assessing confidence regions for this estimator, see Fan, Zhang and Zhang (2001). We also denote by W the diagonal matrix with the diagonal entries w_i and use the notation

$$L(W,\theta,\theta') = \sum_{m=1}^{n} w_m \log \frac{p(Y_m,\theta)}{p(Y_m,\theta')}.$$

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 does not depend on θ.

Now we consider the examples introduced in Section 2 and present the structure of local estimators. For all examples, we adopt a local parametric model at the reference point x described by the localization weights w_i and study the local maximum likelihood estimator.

Example 3.1. [Local constant Gaussian regression] The model is described by the equation $Y_i = \theta + \varepsilon_i$ where the ε_i are i.i.d. $\mathcal{N}(0,\sigma^2)$. Here the MLE $\widehat{\theta} = \widehat{\theta}(x)$ coincides with the weighted least squares estimator and is defined as

$$\widehat{\theta} = \operatorname*{arginf}_\theta (2\sigma^2)^{-1} \sum_{i=1}^{n} w_i (Y_i - \theta)^2 = \frac{\sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i}$$

leading by simple algebra to the log-likelihood

$$L(W,\widehat{\theta},\theta) = (2\sigma^2)^{-1} \sum_{i=1}^{n} w_i \left[(Y_i - \theta)^2 - (Y_i - \widehat{\theta})^2 \right] = \frac{N}{2\sigma^2} (\widehat{\theta} - \theta)^2$$

with $N = \sum_{i=1}^n w_i$.

Example 3.2. [Local Bernoulli model] The original model is locally approximated by the Bernoulli model with the parameter $\theta \in [0,1]$: $\boldsymbol{P}(Y_i = 1) = \theta$, $\boldsymbol{P}(Y_i = 0) = 1 - \theta$. The density $p(y,\theta)$ can be written as $p(y,\theta) = \theta^y (1-\theta)^{1-y}$. The local MLE $\widehat{\theta}$ is

$$\widehat{\theta} = \operatorname*{argsup}_{\theta \in [0,1]} \sum_{i=1}^{n} w_i \left(Y_i \log \theta + (1 - Y_i) \log(1-\theta) \right) = \frac{\sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i}$$

leading to

$$\begin{aligned} L(W,\widehat{\theta},\theta) &= \sum_{i=1}^{n} w_i \left(Y_i \log p(Y_i,\theta) - \log p(Y_i,\widehat{\theta}) \right) \\ &= N \left(\widehat{\theta} \log \frac{\widehat{\theta}}{\theta} + (1-\widehat{\theta}) \log \frac{1-\widehat{\theta}}{1-\theta} \right) = NQ(\widehat{\theta},\theta) \end{aligned}$$

where $N = \sum_{i=1}^{n} w_i$ and $Q(\theta,\theta') = \theta \log(\theta/\theta') + (1-\theta)\log(1-\theta)/(1-\theta')$ is the Kullback-Leibler distance between two Bernoulli distributions with the parameters θ, θ'.

Example 3.3. [Local exponential model] The original model is locally approximated by the exponential model with the parameter $\theta \in I\!R_+ = (0,\infty)$: $\boldsymbol{P}(Y_i \ge t) = e^{-t/\theta}$ having the density $p(y,\theta) = \theta^{-1} e^{-y/\theta}$. The local MLE $\widehat{\theta}$ is

$$\widehat{\theta} = \operatorname*{argsup}_{\theta \in I\!R_+} \sum_{i=1}^{n} w_i \left(-\log\theta - Y_i/\theta\right) = \frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i}$$

leading to

$$L(W,\widehat{\theta},\theta) = \sum_{i=1}^{n} w_i \left[-Y_i\left(\widehat{\theta}^{-1} - \theta^{-1}\right) - \log\frac{\widehat{\theta}}{\theta}\right] = -N\left[\log\frac{\widehat{\theta}}{\theta} - \frac{\widehat{\theta}}{\theta} + 1\right] = NQ(\widehat{\theta},\theta)$$

where, similarly to the previous example, $N = \sum_{i=1}^{n} w_i$ and $Q(\theta,\theta') = \theta/\theta' - 1 - \log(\theta/\theta')$ is the Kullback-Leibler distance between two exponential distributions with the parameters θ, θ'.

Example 3.4. [Local Poisson model] The original model is locally approximated by the Poisson model with the parameter θ, that is, the distribution of the observation Y_i is Poisson with the density $p(y,\theta) = \theta^y e^{-\theta}/y!$. The local MLE $\widehat{\theta}$ is

$$\widehat{\theta} = \operatorname*{argsup}_{\theta \in I\!R_+} \sum_{i=1}^{n} w_i \left(Y_i \log\theta - \theta - \log Y_i!\right) = \frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i}$$

leading to

$$\begin{aligned} L(W,\widehat{\theta},\theta) &= \sum_{i=1}^{n} w_i \left[Y_i \log(\widehat{\theta}/\theta) - (\widehat{\theta} - \theta)\right] \\ &= N\left[\widehat{\theta}\log(\widehat{\theta}/\theta) - \widehat{\theta} + \theta\right] = NQ(\widehat{\theta},\theta) \end{aligned}$$

where, again, $N = \sum_{i=1}^{n} w_i$ and $Q(\theta,\theta') = \theta\log(\theta/\theta') - (\theta - \theta')$ is the Kullback-Leibler distance between two Poisson distributions with parameters θ and θ'.

Example 3.5. [Local constant volatility model] The original model is locally approximated by the model $Y_i = \sigma\varepsilon_i$ with $\sigma \in I\!R_+ = (0,\infty)$ and standard normal ε_i's. Here it is convenient to parametrize by $\theta = \sigma^2$, so that $p(y,\theta) = (2\pi\theta)^{-1/2} e^{-y^2/(2\theta)}$. The local MLE $\widehat{\theta}$ is

$$\widehat{\theta} = \operatorname*{argsup}_{\theta \in I\!R_+} \frac{1}{2} \sum_{i=1}^{n} w_i \left(-\log\theta - Y_i^2/\theta\right) = \frac{\sum_{i=1}^{n} w_i Y_i^2}{\sum_{i=1}^{n} w_i}$$

leading to

$$\begin{aligned} L(W,\widehat{\theta},\theta) &= \frac{1}{2}\sum_{i=1}^{n} w_i \left[-Y_i^2\left(\widehat{\theta}^{-1}-\theta^{-1}\right)-\log(\widehat{\theta}/\theta)\right] \\ &= -\frac{N}{2}\left[\log(\widehat{\theta}/\theta)-\widehat{\theta}/\theta+1\right] = NQ(\widehat{\theta},\theta) \end{aligned}$$

where, again, $N = \sum_{i=1}^{n} w_i$ and $Q(\theta,\theta') = \frac{1}{2}\left(\theta/\theta' - 1 - \log(\theta/\theta')\right)$ is the Kullback-Leibler distance between two normal distributions $\mathcal{N}(0,\theta)$ and $\mathcal{N}(0,\theta')$.

3.3 Local exponential family

All the above examples 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 $Q(\theta,\theta')$ for two measures P_θ and $P_{\theta'}$ from this family satisfies

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

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

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

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

$$\widehat{\theta} = S/N.$$

This implies

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

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.

3.4 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. We suppose that the structural assumption is fulfilled for each of these two, that is, the underlying parameter function $\theta(\cdot)$ is nearly constant within every local 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 estimator $\widehat{\theta}_i$ for the local model corresponding to a diagonal matrix $W = \operatorname{diag}\{w_1, \ldots, w_n\}$, 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 $N_i = \sum_j w_{ij}$ can be interpreted as the sample size for the local model centered at X_i and described by the weights W_i.

To compare two local models centered at X_i and X_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 is obtained by the observations Y_k

with $w_{ik} = 1$ and the other one by 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

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

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

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

For instance, in the local normal case,

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

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

In the procedure described in the next section, we apply such a measure to decide about the weight w_{ij} with which the observation Y_j at X_j enters in the local model at X_i. To prevent from situations where "bad" points X_j corresponding to the local models with a small "sample size" N_j are included into a "big" local model at X_i with a large "sample size" N_i, we slightly extend our approach. 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

$$T_{ij}(\alpha) \le T_{ij} = L(W_i, \widehat{\theta}_i, \widehat{\theta}_j) = \sup_{\theta} L\left(W_i, \theta, \widehat{\theta}_j\right). \tag{3.4}$$

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 the point X_i and the other model at the point X_j. Note that this expression is essentially asymmetric, that is, $T_{ij} \neq T_{ji}$. A "symmetrized" version is given by $T_{ij}^s = (T_{ij} + T_{ji})/2$.

We illustrate this definition by examples from Section 2. Using the general representation of $L(W, \widehat{\theta}, \theta')$ from Section 3.3 for all the considered examples, we end up with the expressions

$$N_i = \sum_{l=1}^{n} w_{il}, \qquad S_i = \sum_{l=1}^{n} w_{il} U(Y_l), \qquad \widehat{\theta}_i = S_i/N_i$$

where $U(Y_l) = Y_l$ for the for local Bernoulli, Poisson and exponential modeling and $U(Y_l) = Y_l^2$ for the local volatility modeling, and

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

In the special case of local constant Gaussian regression, we obtain

$$T_{ij} = \frac{N_i}{2\sigma^2} (\widehat{\theta}_i - \widehat{\theta}_j)^2.$$

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 the 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 comparison of 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 3.4.

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

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

  Alternatively, the "symmetrized" statistical penalty is computed as,

$$\boldsymbol{s}_{ij}^{(k)} = \lambda^{-1} \Bigl(L\bigl(W_i^{(k-1)}, \widehat{\theta}_i^{(k-1)}, \widehat{\theta}_j^{(k-1)}\bigr) + L\bigl(W_j^{(k-1)}, \widehat{\theta}_j^{(k-1)}, \widehat{\theta}_i^{(k-1)}\bigr) \Bigr) /2.$$

  Compute

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

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

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

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

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

Here we specify the procedure for the case when $\{P_\theta\}$ is an exponential family, see Section 3.3. 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

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

Therefore, computing this penalty requires only a finite number of operations. Only the estimators $\widehat{\theta}_i^{(k-1)}$ and the values $N_i^{(k-1)}$ and $S_i^{(k-1)}$ have to be stored as the results of the preceding step of the algorithm.

Step 2 of the procedure

The local MLE $\widehat{\theta}_i$ can be represented in the form $\widehat{\theta} = \operatorname{argsup}_\theta L(W_i, \theta, \theta') = S_i / N_i$ 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 estimator $\widehat{\theta}_i^{(k)}$ 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

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

Initialization

The initial estimators $\widehat{\theta}_i^{(0)}$ coincide with the global parametric MLE for all i and they are defined as $\widehat{\theta}_i^{(0)} = S_i^{(0)} / N_i^{(0)}$ with

$$S_i^{(0)} = \sum_{i=1}^n U_j, \qquad N_i^{(0)} = n.$$

Numerical complexity

One can easily estimate the numerical complexity of procedure. 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 $\log_a(h^*/h^{(1)})$ and the complexity of the whole procedure is of order $n\left(M^{(1)} + \ldots + M^{(k^*)}\right)$. Since the value $M^{(k)}$ grows exponentially in our set-up, the whole complexity is of order $nM^{(k^*)}$.

4.4 Choice of parameters

The parameters of the generalized AWS method are selected essentially in the same way as it is suggested in PS2000 for the local constant regression modeling. We briefly discuss each of the parameters.

Kernels K_s and K_l

The kernels K_s and K_l must fulfill $K_s(0) = K_l(0) = 1$ and decrease in the argument u on the positive semiaxis. We recommend to take $K_s(u) = e^{-u}$. The kernel K_l can be taken exponential, however, it is recommended to utilize a compactly supported kernel to reduce the computational effort of the method. PS2000 applied a uniform kernel, here we apply the triangle kernel $K(u) = (1-u)_+$.

"Memory" parameter η

The value $\eta \in (0,1)$ can be viewed as the memory parameter of the algorithm. An increase of η results in a higher stability of the method w.r.t. to iteration, 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 $\eta = 1/2$.

Starting bandwidth $h^{(1)}$, parameter a and maximal bandwidth h^*

The starting bandwidth $h^{(1)}$ should be taken possibly small. In the most of example we select $h^{(1)}$ such that every starting local neighborhood $U_i^{(0)}$ contains only the design point X_i.

The parameter a controls the growth rate of the local neighborhoods for every point X_i. It should be selected to provide that the mean number of points inside every 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, one can use this parameter 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 the cross-validation technique can be applied.

Symmetric and asymmetric versions

In the most of our 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 unstable performance of the method in the homogeneous situation. Large values of λ may result in loss of adaptivity of the method (less sensitivity to the structural changes). A reasonable way to define the parameter λ for specific applications is based on the condition of free extension, which we also call the "propagation condition". This condition means that in the homogeneous situation, when 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, at the end of iteration process all the weights w_{ij} will then 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 the prescribed probability of getting the global model at the end of iteration process for the homogeneous (parametric) model $\theta(x) = \theta$ using Monte-Carlo simulations. The theoretical justification is given by Theorem 5.1 in the next section, that claims that the choice $\lambda = C \log n$ with a sufficiently large $C \leq 4$ yields the "propagation" condition whatever the parameter θ 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 1.

5 "Propagation" condition

The aim of this section is to show that the choice of a sufficiently large value of the parameter λ indeed implies free extension of every local model in a homogeneous situa-

Table 1: 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

tion. We consider only the case of a univariate exponential family $\{P_\theta, \theta \in \Theta \subset I\!R\}$ that corresponds to all our examples. A homogeneous situation corresponds to a global parametric model with observations $Y_1, \ldots, Y_n$ following a distribution P_θ from the given exponential family.

The main difficulty in the proof of the "propagation condition" is that the weights $w_{ij}^{(k)}$ which we use for describing the local models at the kth iteration are random and computed from the same data $Y_1, \ldots, Y_n$ that we use for estimating the local parameters at kth iteration. This makes the precise analysis of the "propagation condition" very complicated.

To simplify the discussion, we focus on one step of the algorithm assuming that the weights $w_{ij} = w_{ij}^{(k)}$ are deterministic or independent of the data $Y_1, \ldots, Y_n$. The latter situation arises if one splits the original sample into a few subsamples and utilizes different subsamples for different iterations. Below in this section, we give some hints how the "propagation condition" can be proved in full generality by induction arguments. We also present the results for the penalty term based on the classical likelihood ratio test statistic T_{ij}^o, see (3.3). The penalty term T_{ij} used in the procedure can be studied similarly but the analysis becomes more involved.

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" $N_i = \sum_j w_{ij}$ is relatively small, even if the global sample size n is large. Theorem 11.1 from the Appendix states that in the homogeneous situation, for every local model centered at X_i and described by the weights W_i holds

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

This immediately yields for the statistical penalty T^o_{ij}

$$\boldsymbol{P}\left(T^o_{ij} > 2\lambda\right) \le \boldsymbol{P}\left(L(W_i, \widehat{\theta}_i, \theta) > \lambda\right) + \boldsymbol{P}\left(L(W_j, \widehat{\theta}_j, \theta) > \lambda\right) \le 4e^{-\lambda}.$$

This leads to the following results.

Theorem 5.1. *Suppose that* $\theta(X_i) \equiv \theta$ *and that the weights* w_{ij} *are deterministic. Then for some absolute constant* $C \le 4$ *holds*

$$\boldsymbol{P}\left(\max_{i,j=1,\ldots,n} T^o_{ij} > C\log n\right) \le 4/n.$$

Indeed, Theorem 11.1 yields

$$\boldsymbol{P}\left(\max_{i,j=1,\ldots,n} T^o_{ij} > 4\log n\right) \le \sum_{i=1}^{n} \boldsymbol{P}\left(L(W_i, \widehat{\theta}_i, \theta) > 2\log n\right) \le 4ne^{-2\log n} = 4/n.$$

An important feature of this results is that it is nonasymptotic and uniform on the parameter θ.

Remark 5.1. The result is stated for the case of deterministic coefficients w_i. Therefore, it formally applies only to the initial step estimators $\widehat{\theta}_i^{(0)}$ and it ensures for λ sufficiently large that all the computed statistical penalties $s_{ij}^{(1)}$ at the next iteration will be close to zero. However, it implies that the next step estimator $\widehat{\theta}^{(k)}$, $k \ge 1$, will be very close to the usual kernel estimators based on the location penalty only. This gives some hints how the "free extension" principle can be proved by induction arguments.

Remark 5.2. It is also worth noting that our way of computing the statistical penalty $\boldsymbol{s}_{ij}^{(k)}$ does not take into account that two "local" models centered at points X_i and X_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 estimators $\widehat{\theta}_i^{(k)}$ 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 estimators $\widehat{\theta}_i^{(k)}$ and $\widehat{\theta}_j^{(k)}$ becomes more and more close to each other. In the extreme case at the end of iteration process both local models coincide with the global one and therefore the local estimators coincide as well. This yields that the statistical penalty is in a homogeneous situation in fact very small, much smaller than the threshold λ. In particular, artifacts like random segmentation of small regions in the homogeneous situation may appear only at the beginning of the iteration process. The use of the asymmetric procedure usually leads to outsmoothing of such random segments, while the symmetric version may keep them until the final iteration.

Theorem 5.1 gives some upper bound for the value λ that provides the "propagation condition". However, this bound is rather conservative leading to a too large value of λ. As mentioned in Section 4.4, for practical applications the λ-value can be selected by Monte-Carlo simulations.

6 Application to nonparametric density estimation

Here we discuss how the AWS procedure can be applied to the problem of 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 $n = M^d$ 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 $Y_1 + \ldots + Y_n = L$ and 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 $\widehat{f}(X_i) = \frac{\widehat{\theta}_i}{L}\,\delta^{-d}$ or as $\widehat{f}(X_i) = \frac{\widehat{\theta}_i}{\sum_{j=1}^n \widehat{\theta}_j}\,\delta^{-d}$.

Note that the usual histogram requires that the bin value δ satisfies the condition $L \min_i p_i \to \infty$. This condition means that the average number of observations within every bin tends to infinity. Our approach admits small values of Lp_i, that means, that an essential part of bins do not contain any observations and the corresponding Y_i are zeros. If $Y_j = 0$ for all X_j near the point X_i, then $\widehat{f}(X_i)$ is simply estimated by zero.

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 6.1. 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 of the estimation results by the AWS is plotted in the left of Figure 1. One can see almost perfect restoration of the unknown density having the piecewise constant structure. Similar behavior was observed for the local constant regression models, see PS2000.

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

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

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.

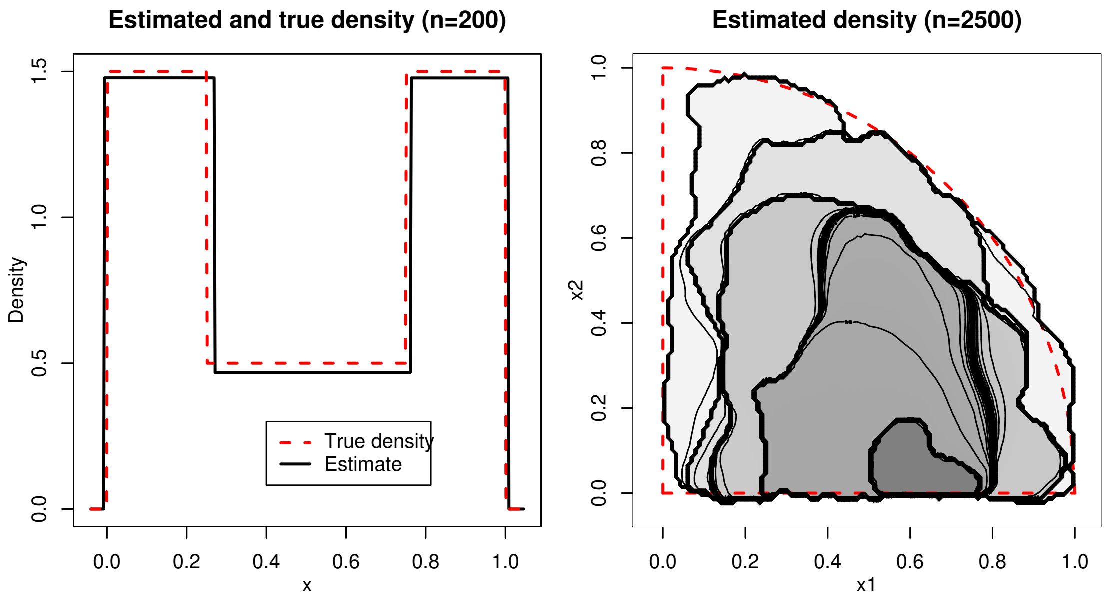

Figure 1: Density estimation: univariate example (left) and bivariate example (right). Solid lines correspond to estimates and dashed lines to the true densities.

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 $x_2 = 0$ and 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 image estimation, see PS2000 and Polzehl and Spokoiny (2003).

7 Application to volatility estimation

Let $S_1, \ldots, S_T$ be an observed stock price (exchange rate, option price etc.) process. The 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:

$$R_t = \sigma_t \varepsilon_t \tag{7.1}$$

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 modeling from Example 2.5 applies to the time dependent volatility from (7.1) 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 required estimate $\widehat{\theta}_t = \widehat{\sigma}_t^2$ of the parameter $\theta_t = \sigma_t^2$ is obtained by running the corresponding AWS procedure on the data $R_1, \ldots, R_T$.

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

Example 7.1. First we produce an artificial series of returns R_t of length $T = 500$ following the model

$$R_t \quad = \quad \sigma_t \varepsilon_t \qquad \text{with} \qquad \sigma_t = 1 + I_{\{t \geq 125\}} - 1.5 I_{\{t \geq 250\}} + .5 I_{\{t \geq 375\}}$$

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.

Example 7.2. In the second example we analyze the exchange rate between the US \$ and the German DM for the period from August 1, 1987 to February 18, 2002. The data are (C) 2001 by Prof. Werner Antweiler, University of British Columbia, 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 behaviour of the asymmetric and symmetric version can be observed, with the symmetric version singleing out several small time intervals with unusually low or high volatility.

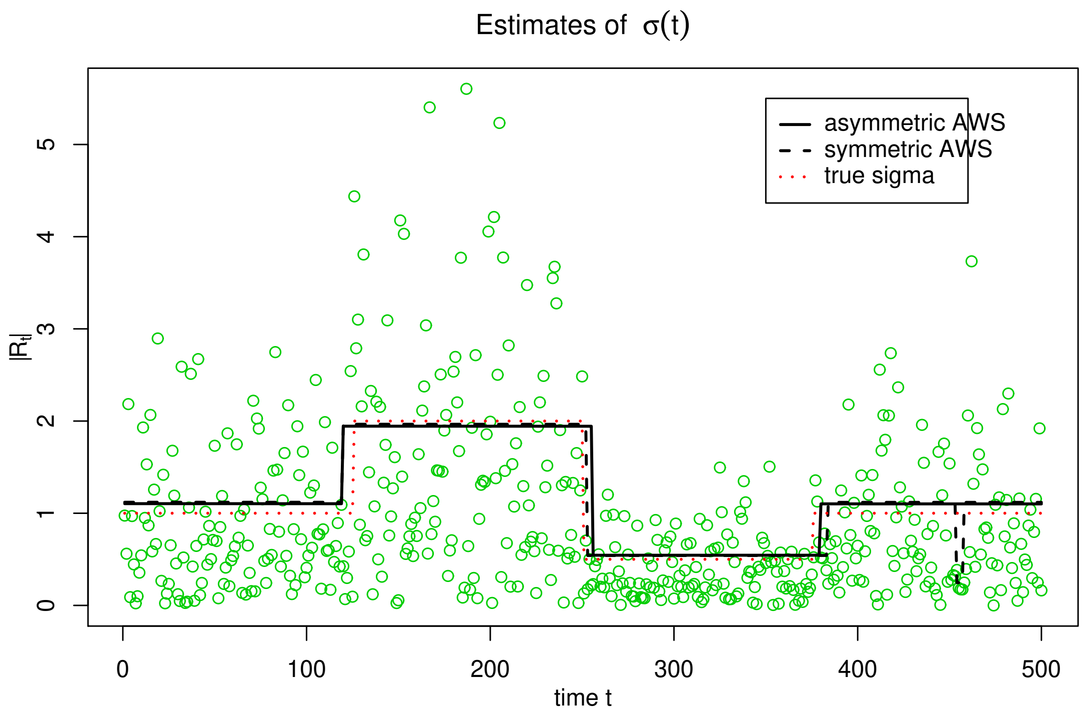

Figure 2: Volatility estimation: Artificial data set with true volatility function and estimates obtained by the asymmetric and symmetric version of AWS.

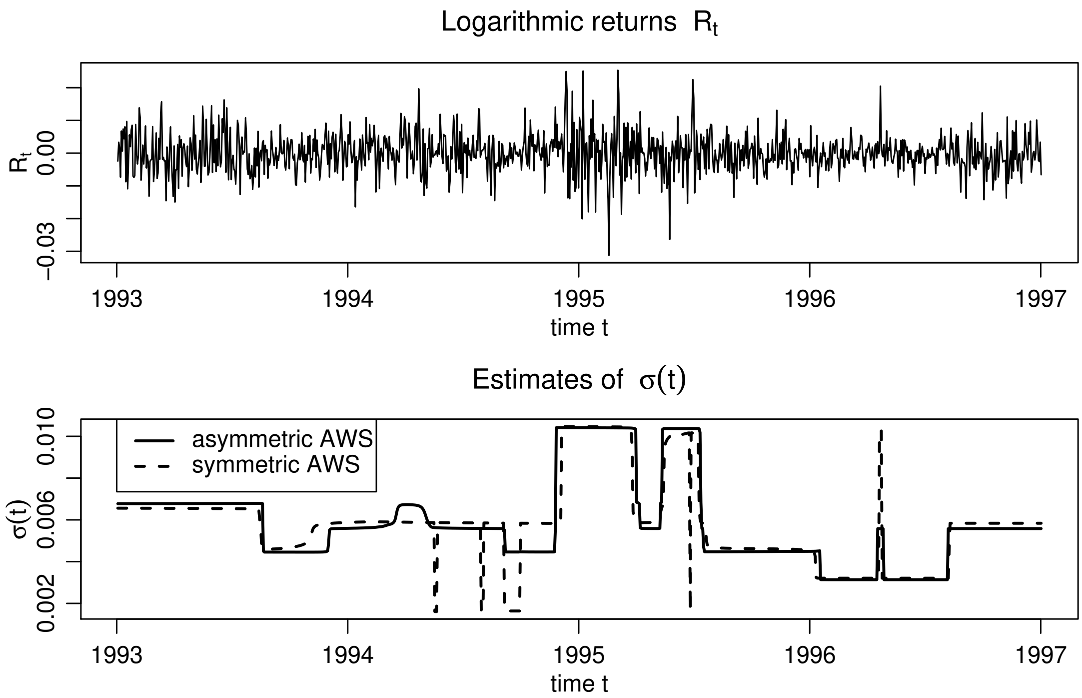

Figure 3: Volatility estimation: Returns for exchange rate between the US \$ and the German DM and estimates obtained by the asymmetric and symmetric version of AWS.

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 $L(x)$ is a slowly varying function and α is the parameter of interest which is usually referred to as the *tail index*. The popular Hill estimator, 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 estimator. 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 estimator 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 estimator, 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 the 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

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

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

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

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 $X_{n,1} > \ldots > X_{n,n}$ be the order statistics pertaining to $X_1, \ldots, X_n$. The goal is to find a natural number k such that on the set $\{X_{n,1}, \ldots, X_{n,k}\}$ 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 $\{X_{n,1}, \ldots, X_{n,k}\}$. 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 $X_1, \ldots, X_n$ be i.i.d. r.v.'s with common strictly increasing d.f. F and $X_{n,1} > \ldots > X_{n,n}$ be the order statistics pertaining to $X_1, \ldots, X_n$. Then the r.v.'s*

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

are i.i.d. standard exponential.

Proof. See for instance Reiss (1989) or Example 4.1.5 in Embrechts, Klüppelberg and Mikosch (1997). □

Let $Y_i = i \log \frac{X_{n,i}}{X_{n,i+1}}$, $i = 1, \ldots, n-1$. Then $Y_i = \alpha_i \xi_i, \quad i = 1, \ldots, n-1$, 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 $i = 1, \ldots, n-1$. These simple considerations reduce the original model to the following inhomogeneous exponential model

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

Table 2: 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

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 estimator of the tail index parameter.

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

Example 8.1. Tail indices are estimated for four distributions, using the Pareto-distribution with tail index $\beta = 1$, 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 centered at the point $X_{n,1}$.

We now reconsider the data used in Example 7.2.

Example 8.2. 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 7.2 equal to 0.1646 and 0.1558, respectively.

Under the hypothesis of a time homogeneous volatility in model (7.1) 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 normality of standardized returns.

9 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

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

This rule can be implemented only if the underlying density functions f_0 and f_1 are known. Since such information is typically lacking 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 with the densities replaced by their estimates.

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 $\rho(x) = \mathbf{1}\left(p(x) \ge \pi_0/(\pi_0 + \pi_1)\right)$. Since the function $p(x)$ is usually unknown, one applies this rule with p replaced by its estimate $\widehat{p}$.

The problem of classification can be easily solved if both densities f_0 and f_1 or their ratio p belong to some parametric family, for instance, if both densities are normal with unknown parameters. The latter assumption leads to linear or quadratic discrimination rules.

Nonparametric methods of estimating the function p are based on local averaging. Two typical examples are given by the k-nearest neighbors estimator and the kernel estimator. 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 estimator 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 estimator 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 both of them 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 9.1. 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. 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.

10 Some important properties of AWS

Here we list some important features of the methods which follow from the construction and are justified by our numerical results. Precise formulations and proofs of these properties are very difficult because of the complex iterative nature of the algorithm and they have to be done elsewhere.

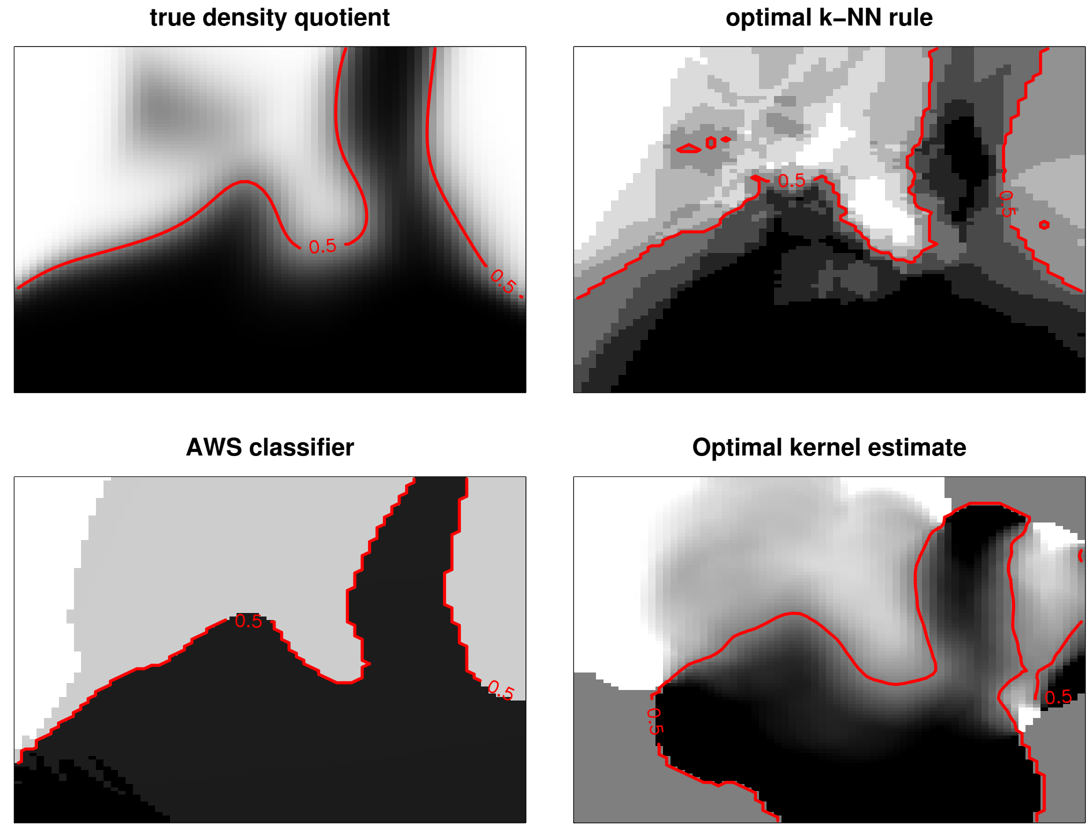

Figure 4: Classification rules obtained by the optimal Bayes decision, the best k-nearest neighbor rule, adaptive weights smoothing (AWS) and the best rule based on kernel estimation.

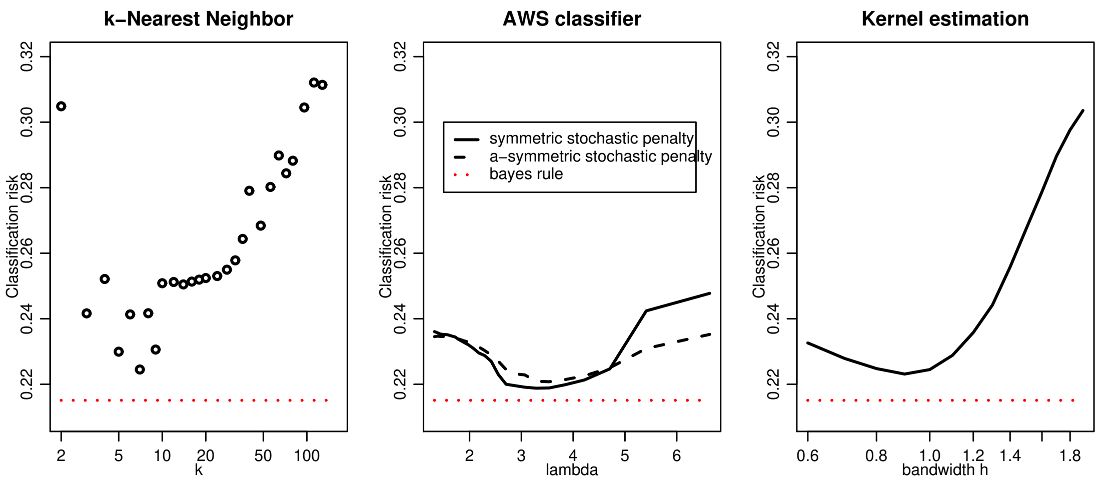

Figure 5: Dependence of the classification error on the main smoothing parameter rules defined by k-nearest neighbor, AWS with symmetric and a-symmetric stochastic penalty, and kernel estimation.

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 application examples considered in the paper, we applied essentially the same procedure. Sometimes, a preliminary model (data) transformation is required, as in density or tail-index estimation.

AWS is fully adaptive to the variable model structure.

This means that the procedure is able to recover the unknown and variable function structure without requiring any specific prior information like degree of smoothness of the underlying function $\theta(x)$, or number and intensity of jumps etc.

Behaviour inside large homogeneous regions

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, due to the "propagation condition", the final estimate at every point coincides with high probability with the fully parametric global estimate.

Estimation near edges and discontinuities

Being stable within homogeneous regions, the procedure is very sensitive to discontinuities. For instance, in the univariate case with a piecewise smooth function $\theta(x)$, the procedure will assign near vanishing weights for every two points from different regions provided that the contrast between these two regions is large enough. This feature of the method can be used for further image segmentation or change-point (edge) analysis.

AWS is dimension free

The dimensionality of the regressors X_i plays absolutely no role for the procedure. This feature of the method 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 restriction were imposed on it. The random design (like for density of tail-index problem) is treated similarly to the case of a deterministic design (image denoising). Design regularity is not assumed in the method. 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 estimator does not indicate the usual Gibbs effect (high variability) near discontinuities like most of the other nonparametric methods.

11 Appendix

This section presents a general deviation result for a local exponential family model. This result is important for justifying the "propagation" condition from Section 5.

We consider an exponential family $(P_\theta\,, \theta \in \Theta)$, 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 $E_\theta Y = \int yp(y,\theta)dP(y) = \theta$ for all $\theta \in \Theta$. Here we suppose that the general definition (see Section 3.3) is applied with $U(y) = y$ to simplify our notation. We also restrict ourselves to the one parameter family, that is, Θ is a subset of the real line. A multivariate exponential family can be considered in a similar way, but the conditions become more involved.

The functions $B(\theta)$ and $C(\theta)$ satisfy the differential equation $B'(\theta) = \theta C'(\theta)$. Moreover, $C'(\theta)$ coincides with the Fisher information of the family (P_θ) at θ .

Let the observations $Y_1, \ldots, Y_n$ be drawn from the distribution P_θ , and let a local model be described by the weights $w_i \in (0,1)$ for $i = 1, \ldots, n$. The corresponding local MLE can be written as $\widehat{\theta} = \sum_{i=1}^n w_i Y_i$. We use the representation $\widehat{\theta} = S/N$ with

$$S = \sum_i w_i Y_i\,, \qquad N = \sum_i w_i$$

see again Section 3.3 for more details.

Theorem 11.1. *Let $\{P_\theta\}$ be an exponential family such that the Fisher information $I(\theta) = C'(\theta)$ is positive on Θ . Then for every $\lambda > 0$ and every $\theta_0 \in \Theta$*

$$\boldsymbol{P}_{\theta_0}\left(L(W, \widehat{\theta}, \theta_0) > \lambda\right) \le 2e^{-\lambda}.$$

Proof. Define $\upsilon = C(\theta)$ and $D(\upsilon) = B(\theta) = B(C^{-1}(\upsilon))$. Since $C'(\theta) > 0$, the new parameter υ is uniquely defined. By simple analysis $D'(\upsilon) = \theta = C^{-1}(\upsilon)$ and $D''(\upsilon) = 1/C'(\theta) = 1/I(\theta) = 1/I(C^{-1}(\upsilon))$. We also set $\upsilon_0 = C(\theta_0)$ and write $\boldsymbol{P}_0$ in place of P_{θ_0}. With this notation

$$L(W,\theta,\theta_0) = L(W,\upsilon,\upsilon_0) = S(\upsilon-\upsilon_0) - N\Bigl(D(\upsilon) - D(\upsilon_0)\Bigr).$$

The MLE $\widehat{\upsilon}$ of the parameter υ is defined by maximizing $L(W,\upsilon,\upsilon_0)$, that is, $\widehat{\upsilon} = \operatorname{argsup}_\upsilon L(W,\upsilon,\upsilon_0)$.

Lemma 11.1. *For given λ and υ_0, there exist two values $\upsilon^* > \upsilon$ and $\upsilon^{**} < \upsilon$ such that*

$$\{L(W,\widehat{\upsilon},\upsilon_0) > \lambda\} \subseteq \{L(W,\upsilon^*,\upsilon_0) > \lambda\} \cup \{L(W,\upsilon^{**},\upsilon_0) > \lambda\}.$$

Proof. It holds

$$\begin{aligned}\{L(W,\widehat{\upsilon},\upsilon_0) > \lambda\} &= \left\{\sup_\upsilon \Bigl[S(\upsilon-\upsilon_0) - N\Bigl(D(\upsilon)-D(\upsilon_0)\Bigr)\Bigr] > \lambda\right\}\\ &= \left\{S > \inf_{\upsilon>\upsilon_0} \frac{\lambda + N\Bigl(D(\upsilon)-D(\upsilon_0)\Bigr)}{\upsilon-\upsilon_0}\right\} \cup \left\{-S > \inf_{\upsilon<\upsilon_0} \frac{\lambda + N\Bigl(D(\upsilon)-D(\upsilon_0)\Bigr)}{\upsilon_0-\upsilon}\right\}.\end{aligned}$$

The function $f(u) = (\lambda + N[D(\upsilon_0+u) - D(\upsilon_0)])/u$ attains its minimum at some point u^* satisfying the equation

$$\lambda + N\Bigl(D(\upsilon_0+u^*) - D(\upsilon_0)\Bigr) - Nu^*D'(\upsilon_0+u^*) = 0$$

and therefore

$$\begin{aligned}\left\{S > \inf_{\upsilon>\upsilon_0} \frac{\lambda + N\Bigl(D(\upsilon)-D(\upsilon_0)\Bigr)}{\upsilon-\upsilon_0}\right\} &= \left\{S > \frac{\lambda + N\Bigl(D(\upsilon^*)-D(\upsilon_0)\Bigr)}{\upsilon-\upsilon_0}\right\}\\ &\subseteq \{L(W,\upsilon^*,\upsilon_0) > \lambda\}\end{aligned}$$

with $\upsilon^* = \upsilon_0 + u^*$. Similarly

$$\left\{-S > \inf_{\upsilon<\upsilon_0} \frac{\lambda + N\Bigl(D(\upsilon)-D(\upsilon_0)\Bigr)}{\upsilon_0-\upsilon}\right\} \subseteq \{L(W,\upsilon^{**},\upsilon_0) > \lambda\}$$

for some $\upsilon^{**} < \upsilon_0$. □

Now we bound the probability $\boldsymbol{P}_0\left(L(W,\upsilon,\upsilon_0)>\lambda\right)$ for every υ. Note that the equality $\theta_0 = D'(\upsilon_0)$ implies for $u=\upsilon-\upsilon_0$

$$L(W,\upsilon,\upsilon_0) = u(S-N\theta_0) - N\left[D(\upsilon_0+u) - D(\upsilon_0) - uD'(\upsilon_0)\right] = (S-N\theta_0)u - NQ(u)$$

with $Q(u) = D(\upsilon_0+u) - D(\upsilon_0) - uD'(\upsilon_0)$. The function Q satisfies $Q'(u) = D'(\upsilon_0+u) - D'(\upsilon_0)$ and $Q''(u) = D''(\upsilon_0+u) = 1/I(C^{-1}(\upsilon_0+u)) > 0$ and thus, it is convex.

We now apply the Chebyshev exponential inequality: for every positive μ

$$\begin{aligned} r(u,\lambda) &:= \log \boldsymbol{P}_0\left(L(W,\upsilon,\upsilon_0)>\lambda\right) \\ &\le -\mu\lambda - \mu NQ(u) + \log \boldsymbol{E}_0 \exp\left(u\mu(S-N\theta_0)\right). \end{aligned}$$

The independence of the Y_i 's implies

$$\log \boldsymbol{E}_0 \exp\left(u\mu(S-N\theta_0)\right) = \log \boldsymbol{E}_0 \exp\left(\sum_{i=1}^{n} u\mu w_i(Y_i-\theta_0)\right) = \sum_{i=1}^{n} \log \boldsymbol{E}_0 e^{u\mu w_i(Y_i-\theta_0)}.$$

Next, for every constant $a>0$, the equalities $\theta_0 = D'(\upsilon_0)$ and $\log \int e^{\upsilon y - D(\upsilon)} P(dy) = 0$ yield

$$\begin{aligned} \log \boldsymbol{E}_0 e^{a(Y-\theta_0)} &= -a\theta_0 + \log \int e^{(a+\upsilon_0)y - D(\upsilon_0)} P(dy) \\ &= -aD'(\upsilon_0) + D(\upsilon_0+a) - D(\upsilon_0) = Q(a). \end{aligned}$$

Therefore

$$r(u,\lambda) \le -\mu\lambda - \mu NQ(u) + \sum_{i=1}^{n} Q(u\mu w_i).$$

Since Q is convex and satisfies $Q(0)=0$, it holds for every $w\in[0,1]$ and every a that $Q(wa) \le wQ(a)$. This and the above inequality applied with $\mu=1$ imply

$$r(u,\lambda) \le -\lambda - NQ(u) + \sum_{i=1}^{n} w_i Q(u) = -\lambda$$

and the result of the theorem follows. $\square$

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

[4] 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.

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

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

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

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

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

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

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

[12] Grama, I. and Spokoiny, V. (2002). Tail index estimation by local exponential modelling. Manuscript in preparation.

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

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

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

[16] Kerkyacharian, G., Lepski, O., and Picard, D. (2001). Nonlinear estimation in anisotropic multi-index denoising. *Probab. Theory Relat. Fields* **121** no.2, 137–170.

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

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

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

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

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

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

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

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

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

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

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

[28] 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.

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

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