VARYING COEFFICIENT REGRESSION MODELING BY ADAPTIVE WEIGHTS SMOOTHING

JÖRG POLZEHL AND VLADIMIR SPOKOINY

Abstract

The adaptive weights smoothing (AWS) procedure was introduced in Polzehl and Spokoiny (2000) in the context of image denoising. The procedure has some remarkable properties like preservation of edges and contrast, and (in some sense) optimal reduction of noise. The procedure is fully adaptive and dimension free. Simulations with artificial images show that AWS is superior to classical smoothing techniques especially when the underlying image function is discontinuous and can be well approximated by a piecewise constant function. However, the latter assumption can be rather restrictive for a number of potential applications. Here the AWS method is generalized to the case of an arbitrary local linear parametric structure. We also establish some important results about properties of the AWS procedure including the so called "propagation condition" and spatial adaptivity. The performance of the procedure is illustrated by examples for local polynomial regression in univariate and bivariate situations.

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

1991 *Mathematics Subject Classification.* 62G05.

Key words and phrases. adaptive weights; local structure; local polynomial regression.

1. Introduction

Polzehl and Spokoiny (2000), referred to as PS2000 in what follows, offered a new method of nonparametric estimation, *Adaptive Weights Smoothing (AWS)*, in the context of image denoising. The main idea of the procedure is to describe the largest local vicinity of every design point X_i in which the underlying model function can be well approximated by a constant in a data-driven and iterative way. The procedure possesses remarkable properties. It is fully adaptive in the sense that no prior information about the structure of the model is required. It is design adaptive and does not suffer from the Gibbs effect (high variability and increased bias near edges and boundaries). A very important feature of the method is that it is dimension free and computationally straightforward. Our numerical results demonstrate that the new method is, compared to other nonparametric procedures, very efficient in situations when the underlying model allows a piecewise constant approximation within large homogeneous regions. Unfortunately, the iterative nature of the procedure makes a rigorous theoretical analysis of the new method very complicated. PS2000 did not provide any theoretical results about the accuracy of estimation delivered by this method. Another weak point of the procedure from PS2000 is that it applies the simplest method of local smoothing based on local constant approximation. This approach seems reasonable e.g. in image analysis or for statistical inference in magnet resonance imaging, as shown in Polzehl and Spokoiny (2001), referred to as PS2001. Other applications to density, volatility, tail index estimation can be found in Polzehl and Spokoiny (2002). However, in many situations the assumption of a local constant structure can be too restrictive. A striking example is estimation of a smooth or piecewise smooth regression function where a piecewise constant approximation is typically too rough. Local linear (polynomial) smoothing delivers much better results in such cases, see Fan and Gijbels (1996) or our examples in Section 5.

In the present paper we propose an extension of the AWS procedure to the case of varying coefficient regression models and simultaneously present a detailed theoretical study of the new method. We particularly prove an important feature of the procedure, the “propagation condition”, which means a free extension of every local model in a nearly homogeneous situation. We then show that this condition automatically leads to a nearly optimal accuracy of estimation for a smooth regression function.

Varying coefficient regression models generalize classical nonparametric regression and gained much attention within the last years, see e.g. Hastie and Tibshirani (1993), Fan and Zhang (1999), Caroll, Ruppert and Welsh (1998), Cai, Fan and Yao (2000) and references therein. The traditional approach uses an approximation of the varying coefficient by a local linear model in the varying parameter. The model is estimated for every localization point independently by local least squares or local maximal likelihood. Accuracy of estimation is typically studied asymptotically as the localization parameter (bandwidth) tends to zero. Such an approach has serious drawbacks of being unable to incorporate special important cases like a global parametric model, a change-point model or more generally, models with inhomogeneous variability w.r.t. the varying parameter. We propose a completely different approach based on the adaptive weights idea that allows to treat all mentioned special cases in a unified way and to get a nearly optimal accuracy of estimation in every such situation. It is however worth mentioning that the classical local polynomial smoothing appears as a very special case of our procedure when we "turn off" our adaptation step.

The next section discusses the notions of global and local modeling. The basic idea of the generalized AWS and the description of the procedure are given in Section 3. The important special case of a local polynomial regression is discussed in Section 4. The performance of the method is studied for some simulated examples of univariate and bivariate regression in Section 5. Section 6 discusses theoretical properties of the procedure. Proofs and some technical results are provided in the Appendix. A reference implementation of the proposed procedures is available as a contributed package of **R** from *URL: http://cran.r-project.org/*.

2. Local modeling by weights

Suppose that data Y_i are observed at design points X_i from the Euclidean space $I\!R^d$, $i = 1, \ldots, n$. In this paper we restrict ourselves to the regression setup with fixed design. The target of statistical analysis is the mean regression function $f(x) = E(Y|X = x)$. We use a representation

$$Y_i = f(X_i) + \varepsilon_i \tag{2.1}$$

where ε_i can be interpreted as additive random noise with zero mean. The distribution of the ε_i 's is typically unknown. Often noise homogeneity can be assumed, that is, all the ε_i 's are independent and satisfy $\boldsymbol{E}\varepsilon_i = 0$ and $\boldsymbol{E}\varepsilon_i^2 = \sigma^2$ for some $\sigma > 0$. For exposition simplicity we restrict ourselves to this homoscedastic situation. Heteroscedastic noise can be considered as well, see PS2001 for some examples. We assume that an estimate $\widehat{\sigma}^2$ of σ^2 is available, see again PS2000 or PS2001 for specific examples.

A pure nonparametric estimate of the target function $f(x)$ usually performs very poorly, especially in case of a multivariate design. The reason is that the underlying target function $f(x)$ often is too complex to be estimated with a reasonable quality without further specifications of its structure.

The approach proposed in PS2000 and PS2001 can be called *structural adaptation*. We assume that the underlying model has a relatively simple structure in some vicinity of every point X_i. The procedure attempts to recover this local structure using a pilot estimate of the model function. Then it utilizes this estimated local structural information to obtain a new improved estimate of the model function. These two steps are iterated extending each time the degree of locality for every local model.

The original AWS method from PS2000 is based on the simplest local structural assumption: the function f is nearly constant within some neighborhood $U(X_i)$ of the point X_i. Here we extend this to the more general situation of a local linear structure.

2.1. **Global linear modeling.** Suppose we are given a set of functions $\psi_1(x), \ldots, \psi_p(x)$ on $I\!R^d$. We consider a linear parametric family $\mathcal{F} = \{f_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ where Θ is a subset of a p-dimensional Euclidean space and, for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$,

$$f_{\boldsymbol{\theta}}(x) = \theta_1 \psi_1(x) + \ldots + \theta_p \psi_p(x).$$

A global parametric structure for the model (2.1) would mean that the underlying function f belongs to $\mathcal{F}$. The simplest example is a one-parameter family given by $f_\theta(x) \equiv \theta$, corresponding to a constant approximation of the function f. Under the global parametric assumption $f \in \mathcal{F}$, the parameter $\boldsymbol{\theta}$ can be easily estimated from the sample $Y_1, \ldots, Y_n$. A natural estimate of $\boldsymbol{\theta}$ is given by ordinary least squares:

$$\widehat{\boldsymbol{\theta}} = \operatorname*{arginf}_{\boldsymbol{\theta}} \sum_{i=1}^{n} (Y_i - f_{\boldsymbol{\theta}}(X_i))^2 .$$

For an explicit representation of this estimate vector notation is useful. Define vectors Ψ_i in $I\!R^p$ with entries $\psi_m(X_i)$, $m = 1, \ldots, p$, and the $p \times n$-matrix Ψ whose columns are Ψ_i. Let also Y stand for the vector of observations: $Y = (Y_1, \ldots, Y_n)^\top \in I\!R^n$. Then

$$\widehat{\boldsymbol{\theta}} = \left(\sum_{i=1}^{n} \Psi_i \Psi_i^\top \right)^{-1} \sum_{i=1}^{n} \Psi_i Y_i = \left(\Psi \Psi^\top \right)^{-1} \Psi Y$$

provided that the $p \times p$ matrix $\Psi\Psi^\top$ is nondegenerated.

2.2. **Local linear modeling.** The global parametric assumption can be too restrictive and does not allow to model complex statistical objects. A standard approach in nonparametric inference is to apply the parametric (linear) structural assumption locally. The most general way to describe a local model centered at a given point is *localization by weights.* Let, for a fixed x, a nonnegative weight $w_i \le 1$ be assigned to the observation Y_i at X_i. When estimating the local parameter $\boldsymbol{\theta}$ at x we utilize every observation Y_i with the weight $w_i = w_i(x)$. This leads to a local (weighted) least squares estimate

$$\widehat{\boldsymbol{\theta}}(x) = \operatorname*{arginf}_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^{n} w_i (Y_i - f_{\boldsymbol{\theta}}(X_i))^2 = \left(\Psi W \Psi^\top \right)^{-1} \Psi W Y \tag{2.2}$$

with $W = \operatorname{diag}\{w_1, \ldots, w_n\}$.

We mention two examples of choosing the weights w_i. *Localization by a bandwidth* is defined by the weights of the form $w_i(x) = K_l(\boldsymbol{l}_i)$ with $\boldsymbol{l}_i = |\rho(x, X_i)/h|^2$ where h is a bandwidth, $\rho(x, X_i)$ is the Euclidean distance between x and the design point X_i and K_l is a *location kernel. Localization by a window* simply restricts the model to some subset (window) U of the design space, that is, $w_i = \mathbf{1}(X_i \in U)$ and all data points Y_i with X_i outside the region U are not taken into account when estimating $\boldsymbol{\theta}(x)$.

Here we do not assume any special structure for the weights w_i, that is, any configuration of the weights is allowed. In what follows we identify the diagonal weight matrix $W = \operatorname{diag}\{w_1, \ldots, w_n\}$ and the local model defined by these weights.

3. Adaptive weights smoothing

This section describes a new method of locally adaptive estimation, *Adaptive Weights Smoothing*. The idea of the procedure is to determine adaptively for every point X_i the largest possible neighborhood in which the model function $f(x)$ can be well approximated by a parametric function $f_{\boldsymbol{\theta}}$ from $\mathcal{F}$. The local model at X_i is given by weights w_{ij} assigned to every observation Y_j. The procedure is iterative. At every iteration step, the procedure tries to extend the local model at each design point. We first illustrate the idea for the local constant structural assumption as considered in PS2000. Here the estimate $\widehat{\boldsymbol{\theta}}_i = \widehat{f}(X_i)$ is defined as the mean of the observations Y_j with weights w_{ij}:

$$\widehat{f}(X_i) = \sum_{j=1}^{n} w_{ij} Y_j \Big/ \sum_{j=1}^{n} w_{ij}. \tag{3.1}$$

The weights w_{ij} are calculated iteratively. For the initial step, the estimate $\widehat{f}^{(0)}(X_i)$ is computed from a smallest local model defined by a bandwidth $h^{(0)}$, that is,

$$\widehat{f}^{(0)}(X_i) = \widehat{\boldsymbol{\theta}}_i^{(0)} = \sum_{j=1}^{n} K_l\bigl(\boldsymbol{l}_{ij}^{(0)}\bigr) Y_j \Big/ \sum_{j=1}^{n} K_l\bigl(\boldsymbol{l}_{ij}^{(0)}\bigr)$$

with $\boldsymbol{l}_{ij}^{(0)} = \bigl|\rho(X_i, X_j)/h^{(0)}\bigr|^2$. In other words, the algorithm starts with the usual kernel estimate with the bandwidth h^0, which is taken very small. If $K_l = \mathbf{1}(u \le 1)$ as in PS2000, then for every point X_i the weights w_{ij} vanish outside the ball $U_i^{(0)}$ of radius $h^{(0)}$ with the center at X_i, that is, the local model at X_i is concentrated on $U_i^{(0)}$. Next, at each iteration k, a ball $U_i^{(k)}$ with a larger bandwidth $h^{(k)}$ is considered and every point X_j from $U_i^{(k)}$ gets a weight $w_{ij}^{(k)}$ which is defined by comparing the estimates $\widehat{f}^{(k-1)}(X_i)$ and $\widehat{f}^{(k-1)}(X_j)$ obtained in the previous iteration. These weights are then used to compute new improved estimates $\widehat{f}^{(k)}(X_i)$ by use of (3.1).

One possible interpretation of this procedure is that at each iteration step the location penalty $\boldsymbol{l}_{ij}^{(k)} = \bigl|\rho(X_i, X_j)/h^{(k)}\bigr|^2$ is relaxed by increasing $h^{(k)}$ at cost of introducing a data-driven statistical penalty which comes from comparison of different local models.

Note that under the local constant assumption $f(x) = \theta$, the value θ uniquely determines the model function and the comparison of the values $\widehat{f}^{(k-1)}(X_i)$ and $\widehat{f}^{(k-1)}(X_j)$ is equivalent to a comparison of two model functions. The extension of this approach to the more general local parametric assumption leads to a check of homogeneity for two local models $W_i^{(k-1)} = \operatorname{diag}\bigl\{w_{i1}^{(k-1)}, \ldots, w_{in}^{(k-1)}\bigr\}$ and $W_j^{(k-1)} = \operatorname{diag}\bigl\{w_{j1}^{(k-1)}, \ldots, w_{jn}^{(k-1)}\bigr\}$, to specify the weight $w_{ij}^{(k)}$. Now we discuss how a statistical penalty (distance) for two local models can be computed.

3.1. **Measuring the statistical 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 both, that is, the underlying regression function f can be well approximated by some $f_{\boldsymbol{\theta}} \in \mathcal{F}$ within every local model. However, the value of the parameter $\boldsymbol{\theta}$ determining the approximating function $f_{\boldsymbol{\theta}}$ may be different for the two local models. We aim to develop a rule to judge from the data, whether the local model corresponding to the point X_j and described by W_j is not significantly different (in the value of the parameter $\boldsymbol{\theta}$) from the model at X_i described by W_i. More precisely, we want to quantify the difference between the parameters of these two local models in order to assign a weight w_{ij} with which the observation Y_j will enter into the local model at X_i in the next iteration of the algorithm. A natural way is to consider the data from two local models as two different populations and to apply the two population likelihood ratio test for testing the hypothesis $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j$. Suppose that the errors ε_i are normally distributed with parameters $(0, \sigma^2)$. The log-likelihood $L(W_i, \boldsymbol{\theta}, \boldsymbol{\theta}')$ for the local regression model at X_i with the weights W_i is, for any pair $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, defined by

$$\begin{aligned} L(W_i, \boldsymbol{\theta}, \boldsymbol{\theta}') &= \frac{1}{2\sigma^2} \sum_{l=1}^{n} w_{il} \Big[(Y_l - \Psi_l^\top \boldsymbol{\theta}')^2 - (Y_l - \Psi_l^\top \boldsymbol{\theta})^2 \Big] \\ &= \frac{1}{2\sigma^2} \sum_{l=1}^{n} w_{il} \Big[2(Y_l - \Psi_l^\top \boldsymbol{\theta}')\Psi_l^\top (\boldsymbol{\theta} - \boldsymbol{\theta}') - (\boldsymbol{\theta} - \boldsymbol{\theta}')^\top \Psi_l \Psi_l^\top (\boldsymbol{\theta} - \boldsymbol{\theta}') \Big] \end{aligned}$$

yielding

$$L(W_i, \widehat{\boldsymbol{\theta}}_i, \boldsymbol{\theta}') = (2\sigma^2)^{-1} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}')^\top B_i \, (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}'),$$

with $B_i = \Psi W_i \Psi^\top$. The classical likelihood-ratio test statistic is of the form

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

where $\widehat{\boldsymbol{\theta}}_i = \operatorname{argmax}_{\boldsymbol{\theta}} L(W_i, \boldsymbol{\theta}, \boldsymbol{\theta}')$ is the maximum likelihood estimate (MLE) corresponding to the local model described by the weight matrix W_i and similarly for $\widehat{\boldsymbol{\theta}}_j$. Also $\widehat{\boldsymbol{\theta}}_{ij} = \operatorname{argmax}_{\boldsymbol{\theta}} L(W_i + W_j, \boldsymbol{\theta}, \boldsymbol{\theta}')$ is the local MLE corresponding to the combined model which is obtained by summing the weights from the both models.

The simple algebra yields

$$T_{ij}^{\circ} = (2\sigma^2)^{-1} (\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_j)^\top B_i (B_i + B_j)^{-1} B_j (\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_j).$$

Note that the value T_{ij}° is “symmetric” w.r.t. W_i and W_j in the sense that $T_{ij}^{\circ} = T_{ji}^{\circ}$. In our procedure, described in the next section, we apply a slightly modified asymmetric version of this test statistic, namely

$$T_{ij} = L(W_i, \widehat{\boldsymbol{\theta}}_i) - L(W_i, \widehat{\boldsymbol{\theta}}_j) = (2\sigma^2)^{-1}(\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_j)^\top B_i(\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_j). \tag{3.3}$$

It has a nice interpretation as a difference between the maximum log-likelihood $L(W_i, \widehat{\boldsymbol{\theta}}_i) = \sup_{\boldsymbol{\theta}} L(W_i, \boldsymbol{\theta}, \boldsymbol{\theta}')$ in model W_i and the "plug-in" log-likelihood $L(W_i, \widehat{\boldsymbol{\theta}}_j, \boldsymbol{\theta}')$ in which $\widehat{\boldsymbol{\theta}}_j$ comes from the model W_j. This modification is important because T_{ij} is used for defining the weight w_{ij} with which the observation Y_j at X_j will enter in the local model W_i corresponding to X_i. However, in the "balanced" situation when the "sample sizes" N_i and N_j are of the same order, the values T°_{ij} and T_{ij} have similar properties.

We consider the value T_{ij} as a 'statistical penalty', that is, when computing the new weight w_{ij} at the next iteration step we strongly penalize for a large value of T_{ij}.

3.2. **Penalization for extending a local model.** An important feature of the original AWS method from PS2000 is its stability against iteration. It turns out that the generalization of the local constant procedure to the local linear case requires to introduce an additional penalty to prevent from leverage problems. To clarify the idea, suppose for the moment that for every iteration step k, each local model $W_i^{(k)}$ is restricted to the ball $U_i^{(k-1)}$ of the radius $h^{(k-1)}$ centered at X_i. Suppose also that the first $k-1$ iterations of the algorithm have been carried over. As a result, we obtain for every point X_i a local model described by the weights $w_{ij}^{(k-1)}$ for each $X_j \in U_i^{(k-1)}$. At the next iteration the procedure tries to extend every local model by increasing the bandwidth $h^{(k)}$ and assigning the weights $w_{ij} = w_{ij}^{(k)}$ for every point X_j from the larger neighborhood $U_i^{(k)}$ of X_i with the radius $h^{(k)}$. If $X_j \in U_i^{(k)} \backslash U_i^{(k-1)}$, then giving X_j a significantly positive weight w_{ij} can be interpreted as including the point X_j into the local model centered at X_i. In some cases, including even one point X_j with a relatively large value $\rho(X_i, X_j)$ into the local model at X_i may significantly change the estimate $\widehat{\boldsymbol{\theta}}_i$. Such leverage problem does not arise in the local constant modeling but it becomes crucial for local linear (polynomial) regression. To prevent from this danger, we introduce a special penalty for including an influence point.

To measure the influence of the observation Y_j at X_j in the local model described by the weight matrix W_i, one can consider the extended model obtained by adding a single observation at the point X_j and look at the relative difference between the original and the extended model. This leads to the value

$$\begin{aligned} \gamma_{ij} &= \operatorname{tr}\left\{ \left(\Psi \overline{W}_i \Psi^\top\right)^{-1} \left(\Psi \overline{W}_i \Psi^\top + \Psi_j \Psi_j^\top\right) \right\} - p \\ &= \Psi_j^\top \left(\Psi \overline{W}_i \Psi^\top\right)^{-1} \Psi_j = (\operatorname{tr} W_i)\, \Psi_j^\top \left(\Psi W_i \Psi^\top\right)^{-1} \Psi_j. \end{aligned}$$

Here $\Psi_j \in I\!R^p$ is the jth column of Ψ and, for a diagonal matrix W, we denote $\overline{W} = (\operatorname{tr} W)^{-1} W$. A large value of γ_{ij} means that X_j is a leverage point. To make the procedure more stable w.r.t. such influential points, we additionally penalize for including points with a large value γ_{ij}, i.e. assign small weights even when the difference $\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_j$ is statistically insignificant and the statistical penalty $\boldsymbol{s}_{ij}$ is small.

For adjusting the penalty term one can use the 'propagation' principle which means a free extension of the model in the homogeneous situation when the coefficients of the linear model do not vary with location. In that situation, neither the statistical penalty nor the penalty for extending the model would significantly affect the estimate leading after the first $k-1$ iterations to the classical location weights $w_{ij,\mathrm{ho}}^{(k-1)} = K_l\left(\boldsymbol{l}_{ij}^{(k-1)}\right) = K_l\left(\left|\rho(X_i, X_j)/h^{(k-1)}\right|^2\right)$. The influence of the point X_j within the local homogeneous model described by $W_{i,\mathrm{ho}}^{(k-1)}$ is given by

$$\gamma_{ij,\mathrm{ho}} \;=\; \gamma_j\left(W_{i,\mathrm{ho}}^{(k-1)}\right) = \left(\mathrm{tr} W_{i,\mathrm{ho}}^{(k-1)}\right) \Psi_j^\top \left(\Psi W_{i,\mathrm{ho}}^{(k-1)} \Psi^\top\right)^{-1} \Psi_j$$

where $W_{i,\mathrm{ho}}^{(k-1)} = \mathrm{diag}\{w_{i1,\mathrm{ho}}^{(k-1)}, \ldots, w_{in,\mathrm{ho}}^{(k-1)}\}$. This value $\gamma_{ij,\mathrm{ho}}$ can be used for adjusting the penalty for extending the model. Namely, we assign to every observation Y_j at X_j the penalty

$$\boldsymbol{e}_{ij}^{(k)} = \tau^{-1}\left(\gamma_{ij}/\gamma_{ij,\mathrm{ho}} - 1\right)_+$$

where a_+ means $\max\{0, a\}$ and τ is a numerical tuning parameter.

3.3. **Defining weights.** Using the previously described methods, we compute for every pair (i,j) the penalties $\boldsymbol{l}_{ij}^{(k)}$, $\boldsymbol{s}_{ij}^{(k)}$ and $\boldsymbol{e}_{ij}^{(k)}$. It is natural to require that the influence of every such factor is independent of the other factors. This suggests to define the new weight $w_{ij}^{(k)}$ using the product

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

where K_l, K_s and K_e are three kernel functions, which are nondecreasing on the positive semiaxis and satisfy the condition $K_l(0) = K_s(0) = K_e(0) = 1$.

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

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

3.4. **Formal description of the procedure.** Important ingredients of the method are the kernels K_l, K_s and K_e; the parameters λ, τ and η; the initial bandwidth $h^{(0)}$, the factor $a > 1$, the maximal bandwidth $h_{\max}$ and the estimated error variance $\widehat{\sigma}^2$. The choice of these parameters is discussed in detail in Section 3.5.

The generalized procedure reads as follows:

1. Initialization: For every i define the diagonal matrix $W_i^{(0)}$ with diagonal entries $w_{ij}^{(0)} = K_l(\boldsymbol{l}_{ij}^{(0)})$ and $\boldsymbol{l}_{ij}^{(0)} = \left|\rho(X_i, X_j)/h^{(0)}\right|^2$, that is, $W_i^{(0)} = \mathrm{diag}\{w_{i1}^{(0)}, \ldots, w_{in}^{(0)}\}$.

Compute

$$N_i^{(0)} = \mathrm{tr} W_i^{(0)}, \quad B_i^{(0)} = \Psi W_i^{(0)} \Psi^\top, \quad Z_i^{(0)} = \Psi W_i^{(0)} Y \quad \text{and} \quad \widehat{\boldsymbol{\theta}}_i^{(0)} = \left(B_i^{(0)}\right)^{-1} Z_i^{(0)}.$$

Set $k = 1$.

2. Iteration: for every $i = 1, \ldots, n$ define $W_{i,\mathrm{ho}}^{(k-1)} = \mathrm{diag}\{K_l(\boldsymbol{l}_{i1}^{(k-1)}), \ldots, K_l(\boldsymbol{l}_{in}^{(k-1)})\}$,

- **calculate the adaptive weights:** For every point X_j compute

$$\begin{aligned} \gamma_{ij}^{(k)} &= N_i^{(k-1)} \Psi_j^\top \left(B_i^{(k-1)}\right)^{-1} \Psi_j, \\ \gamma_{ij,\mathrm{ho}}^{(k)} &= \mathrm{tr}\left(W_{i,\mathrm{ho}}^{(k-1)}\right) \Psi_j^\top \left(\Psi W_{i,\mathrm{ho}}^{(k-1)} \Psi^\top\right)^{-1} \Psi_j \end{aligned}$$

where Ψ_j is j th column of Ψ. Then calculate the penalties

$$\begin{aligned} \boldsymbol{l}_{ij}^{(k)} &= \left|\rho(X_i, X_j)/h^{(k)}\right|^2, \\ \boldsymbol{s}_{ij}^{(k)} &= (2\widehat{\sigma}^2\lambda)^{-1}\big(\widehat{\boldsymbol{\theta}}_i^{(k-1)} - \widehat{\boldsymbol{\theta}}_j^{(k-1)}\big)^\top B_i^{(k-1)}\big(\widehat{\boldsymbol{\theta}}_i^{(k-1)} - \widehat{\boldsymbol{\theta}}_j^{(k-1)}\big), \qquad (3.4) \\ \boldsymbol{e}_{ij}^{(k)} &= \tau^{-1}\big(\gamma_{ij}^{(k)}/\gamma_{ij,\mathrm{ho}}^{(k)} - 1\big)_+ \end{aligned}$$

and obtain the weight $\widetilde{w}_{ij}^{(k)}$ as

$$\widetilde{w}_{ij}^{(k)} = K_l\big(\boldsymbol{l}_{ij}^{(k)}\big) K_s\big(\boldsymbol{s}_{ij}^{(k)}\big) K_e\big(\boldsymbol{e}_{ij}^{(k)}\big), \qquad (3.5)$$

Denote by $\widetilde{W}_i^{(k)}$ the diagonal matrix whose diagonal elements are $\widetilde{w}_{ij}^{(k)}$.

- **Compute the new estimate:** Compute

$$\begin{aligned} N_i^{(k)} &= \eta N_i^{(k-1)} + (1-\eta)\mathrm{tr}\widetilde{W}_i^{(k)}, \\ Z_i^{(k)} &= \eta Z_i^{(k-1)} + (1-\eta)\Psi \widetilde{W}_i^{(k)} Y, \\ B_i^{(k)} &= \eta B_i^{(k-1)} + (1-\eta)\Psi \widetilde{W}_i^{(k)} \Psi^\top, \end{aligned}$$

and the estimate $\widehat{\boldsymbol{\theta}}_i^{(k)}$ (resp. $\widehat{f}_i^{(k)}$) of $\boldsymbol{\theta}_i$ (resp. of $f_i = f(X_i)$) by

$$\widehat{\boldsymbol{\theta}}_i^{(k)} = \left(B_i^{(k)}\right)^{-1} Z_i^{(k)}, \qquad \widehat{f}_i^{(k)} = \Psi_i^\top \widehat{\boldsymbol{\theta}}_i^{(k)}.$$

3. Stopping: Increase k by 1, set $h^{(k)} = a h^{(k-1)}$. If $h^{(k)} \le h_{\max}$ continue with step 2. Otherwise terminate.

We obtain the final estimates as $\widehat{f}_i = \widehat{f}_i^{(k^*)}$, with k^* denoting the total number of iterations.

3.5. **Choice of parameters.** The parameters of the procedure are selected similarly to PS2000. We briefly discuss each of the parameters.

Kernels K_s, K_l and K_e: The kernels K_s and K_l must be nonnegative and non-increasing on the positive semiaxis. We propose to use $K_s(u) = e^{-u} I_{\{u \le 6\}}$. We recommend to apply a compactly supported localization kernel K_l to reduce the computational effort of the method. PS2000 used a uniform kernel, here the triangle kernel $K_l(u) = (1-u)_+$ is employed. We also set $K_e = K_s$. Our numerical results show that similarly to the standard local linear (polynomial) regression the particular choice of kernels K_s, K_l and K_e does not significantly affect the performance of the method.

Initial bandwidth $h^{(0)}$, parameter a and maximal bandwidth $h_{\max}$: The starting bandwidth $h^{(0)}$ should be small. We recommend to select $h^{(0)}$ such that every initial local neighborhood $U_i^{(0)}$ contains a sufficient number of design points to guarantee identifiability of the local parameter $\boldsymbol{\theta}_i$.

The parameter a controls the growth rate of the local neighborhoods for every point X_i. If X_i are from the unit cube in the space $I\!R^d$ we take the parameter a as $a = a_{grow}^{1/d}$. This allows for an exponential growth, in k, of the mean number of points inside a ball $U_i^{(k)}$ with radius $h^{(k)}$ with the factor a_{grow}. Our default choice is $a_{grow} = 1.25$. This ensures that the number of iterations k^* is at most logarithmic in the sample size.

The maximal bandwidth $h_{\max}$ may be taken very large if the specified local model allows for a good approximation of the regression function in its smooth parts. However, using a too large final bandwidth $h_{\max}$ may lead to oversmoothing and artificial segmentation. A data-driven method of optimal stopping, based, for instance, on cross-validation can be applied for selecting a proper bandwidth $h_{\max}$.

The value of $h_{\max}$ also determines the number of iterations and can therefore be used to control the numerical complexity of the procedure.

Parameter λ: An important parameter of the procedure is λ which scales the statistical penalty $\boldsymbol{s}_{ij}$. Small values of λ lead to overpenalization which may result 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 structural changes). The extreme case is given by $\lambda = \infty$ which leads to nonadaptive local linear procedure with the bandwidth $h_{\max}$.

A reasonable way to define the parameter λ for a specific application is based on the condition of free extension, which we also call the "propagation condition". This condition means that in a homogeneous situation, i.e. 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 under homogeneity.

In a homogenous situation, provided the value $h_{\max}$ is sufficiently large, all weights w_{ij} will be close to one at the end of the iteration process and every local model will essentially coincide with the global one. Therefore, the parameter λ can be adjusted by 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 homogeneous (parametric) model $\boldsymbol{\theta}(x) = \boldsymbol{\theta}$ using Monte-Carlo simulations. The theoretical justification is given by Theorem 6.1 in Section 6.1, that claims that the choice $\lambda = C \log n$ with a sufficiently large C yields the "propagation" condition whatever the parameter $\boldsymbol{\theta}$ is.

Our default value is $\lambda = q_\alpha(\chi^2_p)$, that is the α-quantile of the χ^2 distribution with p degree of freedom, where α depends on the specified linear parametric family. Defaults for the case of local polynomial regression are given in Section 5. An optimal choice of α may also slightly depend on the error distribution.

Parameter τ: The optimal choice of τ depends on the method of smoothing. For the local constant AWS considered in PS2000, there are no influential points (see Section 4.1). For local polynomial smoothing the default choice of τ is given in Section 4.

Parameter η and the control step: A 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 the estimates $\widehat{\boldsymbol{\theta}}_i^{(k)}$ are well defined, that is, all the matrices $B_i^{(k)}$ are positive definite. Our default choice is $\eta = 1/2$.

The original AWS procedure from PS2000 did not involve the "memory" parameter η (it corresponds to $\eta = 0$). Instead it used an additional *control* step in which the new estimate $\widehat{\boldsymbol{\theta}}_i^{(k)}$ was compared with all the previous estimates $\widehat{\boldsymbol{\theta}}_i^{(k')}$ for $k' < k$. If the difference $\widehat{\boldsymbol{\theta}}_i^{(k)} - \widehat{\boldsymbol{\theta}}_i^{(k')}$ 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 6 for more details. In the local linear case this control step would accept the estimate $\widehat{\boldsymbol{\theta}}_i^{(k)}$ only if

$$(2\widehat{\sigma}^2)^{-1}\big(\widehat{\boldsymbol{\theta}}_i^{(k')} - \widehat{\boldsymbol{\theta}}_i^{(k)}\big)^\top B_i^{(k')}\big(\widehat{\boldsymbol{\theta}}_i^{(k')} - \widehat{\boldsymbol{\theta}}_i^{(k)}\big) \le \eta^*, \qquad k' = 1,\ldots,k-1, \tag{3.6}$$

that is, when the new estimate $\widehat{\boldsymbol{\theta}}_i^{(k)}$ lies inside all confidence ellipsoids of previous estimates at the point X_i. However, our numerical results (not reported here) indicate that 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.

3.6. **Computational complexity of the algorithm.** We start with the following two important remarks. First note, that every estimate is defined as $\widehat{\boldsymbol{\theta}}_i^{(k)} = \Big(B_i^{(k)}\Big)^{-1} Z_i^{(k)}$ using the matrix $B_i^{(k)}$ and the vector $Z_i^{(k)}$. Similarly, the new weights $\widetilde{w}_{ij}^{(k)}$ are computed on the basis of the same statistics $B_i^{(k-1)}$, $Z_i^{(k-1)}$ and $N_i^{(k-1)}$ from the previous step of the procedure. Therefore, the whole structural information is contained in these three basis elements. During the adaptation step, we compute for every i the weights $\widetilde{w}_{ij}^{(k)}$ with different j only with the aim to compute the new elements $B_i^{(k)}$, $Z_i^{(k)}$ and $N_i^{(k)}$. This reduces the memory requirements for the algorithm to $\mathcal{O}(np^2)$ or even to $\mathcal{O}(np)$ for local polynomial modeling, see the next section, while keeping all the weights $w_{ij}^{(k)}$ would lead to the memory requirement $\mathcal{O}(n^2)$.

The localization kernel K_l usually has a compact support, say, $[0,1]$. This immediately implies that for every local model at X_i, all the weights $\widetilde{w}_{ij}^{(k)}$ for the points X_j

outside the ball $U_i^{(k)} = \{x : \rho(X_i, x) \le h^{(k)}\}$ vanish. Therefore, it suffices at each step to compute the weights $\widetilde{w}_{ij}^{(k)}$ for pairs X_i, X_j with $\rho(X_i, X_j) \le h^{(k)}$. Denote by M_k the maximal number of design points X_j within a ball of radius $h^{(k)}$ centered at a design point. At the kth step there are at most M_k positive weights $\widetilde{w}_{ij}^{(k)}$ for any X_i.

Therefore, for carrying out the kth adaptation step of the algorithm, we have to compute the penalties $\boldsymbol{l}_{ij}^{(k)}$, $\boldsymbol{s}_{ij}^{(k)}$ and $\boldsymbol{e}_{ij}^{(k)}$ and the value $\widetilde{w}_{ij}^{(k)}$, for every pair (i,j) with $\rho(X_i, X_j) \le h^{(k)}$ due to (3.5). This requires a finite number of operations depending on the number of parameters p only, and the whole kth adaptation step of the algorithm requires of order nM_k operations.

To obtain the estimate we need, for every point X_i, to compute the $d \times d$-matrix $B_i^{(k)} = \eta B_i^{(k-1)} + (1-\eta)\Psi \widetilde{W}_i^{(k)} \Psi^\top$, the vector $Z_i^{(k)} = \eta Z_i^{(k-1)} + (1-\eta)\Psi \widetilde{W}_i^{(k)} Y$ and the value $N_i^{(k)} = \eta N_i^{(k-1)} + (1-\eta)\mathrm{tr}\widetilde{W}_i^{(k)}$. It is clear that the complexity of computing all these values is of order M_k. Computing $\widehat{\boldsymbol{\theta}}_i^{(k)} = \left(B_i^{(k)}\right)^{-1} Z_i^{(k)}$ requires a finite number operations depending on p only. Therefore, the complexity of the whole estimation step is again of order nM_k. Since typically the numbers M_k grow exponentially, the complexity of the whole algorithm is estimated as $n(M_1 + \ldots + M_{k^*}) \asymp nM_{k^*}$ where k^* is the number of iteration steps.

4. Local polynomial regression

We now specify the procedure for adaptive local polynomial estimation of a regression function with univariate and multivariate covariates.

4.1. **Local constant regression.**

The local constant approximation corresponds to the simplest family of basis functions $\{\psi_m\}$ consisting of one constant function $\psi_0 \equiv 1$. The major advantage of this method is that the dimensionality of the regressors plays absolutely no role. In this situation $\Psi = (1, \ldots, 1)$ and, for every diagonal matrix $W = \mathrm{diag}(w_1, \ldots, w_n)$, it holds $\Psi W \Psi^\top = \mathrm{tr} W$ and $\Psi W Y = \sum_{l=1}^n w_l Y_l$. Hence, for the local constant case, the $B_i^{(k)}$'s coincide with the $N_i^{(k)}$'s. The statistical penalty $\boldsymbol{s}_{ij}^{(k)}$ can be written in the form $\boldsymbol{s}_{ij}^{(k)} = (2\sigma^2)^{-1} N_i^{(k-1)} \bigl|\widehat{\boldsymbol{\theta}}_i^{(k-1)} - \widehat{\boldsymbol{\theta}}_j^{(k-1)}\bigr|^2$. Also, for all i and k, it holds $\gamma_{ij}^{(k)} = \mathrm{tr} W_i^{(k-1)} / \mathrm{tr} W_i^{(k-1)} \equiv 1$, and similarly for $\gamma_{ij,\mathrm{ho}}^{(k)}$. Therefore, the penalty e_{ij} is always zero and can be dropped.

The weights $\widetilde{w}_{ij}^{(k)}$ can be computed as $\widetilde{w}_{ij}^{(k)} = K_l(\boldsymbol{l}_{ij}^{(k)}) K_s(\boldsymbol{s}_{ij}^{(k)})$ that essentially coincides with the proposal from PS2000 if the uniform kernel K_l is applied. A small difference remains in the use of the memory parameter η and in a slightly different form of the statistical penalty.

4.2. **Local polynomial univariate regression.**

Local linear (polynomial) smoothing is known to be much more accurate when estimating a smooth function, see e.g. Fan and Gijbels (1996). A generalization of the original AWS to the local linear (polynomial) regression therefore is of special importance.

For local polynomial regression the basis functions could be specified as $\psi_1(x) = 1$, $\psi_2(x) = x$, …, $\psi_p(x) = x^{p-1}$. However, it is well known, that the numerical stability of the procedure will be improved if, for every local model, the basis functions are centered at the reference point X_i, that is, the functions $(x - X_i)^m$ are applied. This is, for fixed i, only a reparametrization, but requires to slightly modify the description of the procedure. Denote by $\Psi(X_i)$ the $p \times n$ matrix with the entries $(X_l - X_i)^m$ for $m = 0, 1, \ldots, p-1$ and $l = 1, \ldots, n$.

The estimation step of the algorithm is performed similarly to the case described in Section 3.4. The only difference is that the family of basis functions (or, equivalently, the matrix Ψ) depends on the central point X_i. Suppose that at the kth step of the procedure, for a point X_i, the matrix $\widetilde{W}_i^{(k)}$ has been computed. We then compute the p-vector $Z_i^{(k)} = \eta Z_i^{(k-1)} + (1-\eta)\Psi(X_i)\widetilde{W}_i^{(k)} Y$ with entries $Z_{i,m}^{(k)}$ of the form

$$Z_{i,m}^{(k)} = \eta Z_{i,m}^{(k-1)} + (1-\eta) \sum_{l=1}^{n} \widetilde{w}_{il}^{(k)} (X_l - X_i)^m Y_l \qquad m = 0, \ldots, p-1,$$

and the matrix $B_i^{(k)} = \eta B_i^{(k-1)} + (1-\eta)\Psi(X_i)\widetilde{W}_i^{(k)}\Psi^\top(X_i)$ whose entries are of the form $B_{i,mm'}^{(k)} = b_{i,m+m'}^{(k)}$ for $m, m' = 1, \ldots, p$ where

$$b_{i,m}^{(k)} = \eta b_{i,m}^{(k-1)} + (1-\eta) \sum_{l=1}^{n} \widetilde{w}_{il}^{(k)} (X_l - X_i)^m \qquad m = 0, \ldots, 2p-2,$$

The estimate $\widehat{\boldsymbol{\theta}}_i^{(k)}$ in the local model at X_i, is obtained as $\widehat{\boldsymbol{\theta}}_i^{(k)} = \left(B_i^{(k)}\right)^{-1} Z_i^{(k)}$.

In the kth adaptation step, we have to compare two estimates corresponding to the local models $W_i^{(k-1)}$ and $W_j^{(k-1)}$. Note however, that this comparison can be done only if the both estimates are computed for the same basis system. Thus, the comparison requires to recompute the estimate for the local model $W_j^{(k-1)}$ w.r.t. the basis centered at the point X_i. Let $\widehat{\boldsymbol{\theta}}_j = (\widehat{\theta}_{j,0}, \ldots, \widehat{\theta}_{j,p-1})^\top$ be the estimate for the local model at X_j. This estimate leads to a local approximation of the unknown regression function by the polynom $\widehat{f}_j(x) = \widehat{\theta}_{j,0} + \widehat{\theta}_{j,1}(x - X_j) + \ldots + \widehat{\theta}_{j,p-1}(x - X_j)^{p-1}$. Now we represent this polynom as a linear combination of the basis functions $(x - X_i)^m$, $m = 0, \ldots, p-1$, that is, we have to find new coefficients $\widehat{\boldsymbol{\theta}}_{ij} = (\widehat{\theta}_{ij,0}, \ldots, \widehat{\theta}_{ij,p-1})^\top$ such that

$$\widehat{f}_j(x) = \widehat{\theta}_{ij,0} + \widehat{\theta}_{ij,1}(x - X_i) + \ldots + \widehat{\theta}_{ij,p-1}(x - X_i)^{p-1}.$$

The coefficients $\widehat{\theta}_{ij,m}$ can be computed from the formula $\widehat{\theta}_{ij,m} = (m!)^{-1} d^m \widehat{f}_j(X_i)/dx^m$.

Suppose that all the estimates $\widehat{\boldsymbol{\theta}}_i^{(k-1)} = (\widehat{\theta}_{i,0}^{(k-1)}, \ldots, \widehat{\theta}_{j,p-1}^{(k-1)})^\top$ have been computed in the previous step. Next, for a fixed i and every j, we compute the estimates $\widehat{\boldsymbol{\theta}}_{ij}^{(k-1)}$ by

$$\widehat{\theta}_{ij,m}^{(k-1)} = \sum_{q=0}^{p-m-1} \binom{q+m}{q} \widehat{\theta}_{j,q+m}^{(k-1)} (X_i - X_j)^q. \qquad m = 0, 1, \ldots, p-1.$$

The estimate $\widehat{\boldsymbol{\theta}}_{ij}^{(k-1)}$ is used in place of $\widehat{\boldsymbol{\theta}}_{j}^{(k-1)}$ for computing the statistical penalty $\boldsymbol{s}_{ij}^{(k)}$ in (3.4). For computing the extension penalty, we apply $\Psi(X_i)$ in place of Ψ and Ψ_j has to be replaced by $\Psi_j(X_i)$ which is the jth column of $\Psi(X_i)$. The remaining steps of the procedure are performed similarly to the basic algorithm.

4.3. **Local linear multiple regression.** Let $X_1, \ldots, X_d$ be points in the d-dimensional Euclidean space $I\!R^d$. Classical linear regression leads to an approximation of the regression function f by a linear combination of the constant function $\psi_0(x) = 1$ and d coordinate functions $\psi_m(x) = x_m$, so that the family $\{\psi_m\}$ consists of $p = d+1$ basis functions. Our procedure attempts to apply this approximation locally for adaptively selected local models. The global linear modeling arises as a special case if the underlying model is entirely linear.

Similarly to the univariate case, we adopt for every design point X_i a local linear model with centered basis functions $\psi_m(x, X_i) = x_m - X_{im}$ for $m = 1, \ldots, d$. The corresponding $p \times n$ matrix $\Psi(X_i)$ has columns $\Psi_l(X_i) = (1, X_{l1} - X_{i1}, \ldots, X_{ld} - X_{id})^\top$ for $l = 1, \ldots, n$. At the estimation step one computes the estimates $\widehat{\boldsymbol{\theta}}_i^{(k)}$ of the parameter $\boldsymbol{\theta} \in I\!R^p$ for every local model, leading to a local linear approximation of the function f by the linear function $\widehat{f}_j(x)$ with

$$\widehat{f}_j(x) = \widehat{\theta}_{j,0} + \sum_{m=1}^{d} \widehat{\theta}_{j,m}(x_m - X_{j,m}).$$

This linear function can be rewritten in the form

$$\widehat{f}_j(x) = \widehat{\theta}_{j,0} + \sum_{m=1}^{d} \widehat{\theta}_{j,m}(X_{i,m} - X_{j,m}) + \sum_{m=1}^{d} \widehat{\theta}_{j,m}(x_m - X_{i,m}).$$

Therefore, only the first coefficient of the vector $\widehat{\boldsymbol{\theta}}_j$ has to be recomputed when the basis system $\Psi(X_i)$ is used in place of $\Psi(X_j)$. This means that at the kth adaptation step, the vector $\widehat{\boldsymbol{\theta}}_j^{(k-1)}$ is replaced by $\widehat{\boldsymbol{\theta}}_{ij}^{(k-1)}$ where $\widehat{\theta}_{ij,m}^{(k-1)} = \widehat{\theta}_{j,m}^{(k-1)}$ for $m = 1, \ldots, d$ and $\widehat{\theta}_{ij,0}^{(k-1)} = \widehat{\theta}_{ij,0}^{(k-1)} + \sum_{m=1}^{d} \widehat{\theta}_{j,m}(X_{i,m} - X_{j,m})$. The rest of the procedure is carried through similarly to the univariate case.

4.4. **Local quadratic bivariate regression.** Finally we shortly discuss the bivariate case with $d = 2$ for local quadratic approximation. The case of a larger d can be handled similarly. The family $\{\psi_m\}$ of basis functions contains one constant function equal to 1, two linear coordinate functions x_1 and x_2 and three quadratic functions x_1^2, x_2^2 and $x_1 x_2$. It is useful to utilize the notation $m = (m_1, m_2)$, $|m| = m_1 + m_2$ and $x^m = x_1^{m_1} x_2^{m_2}$ for $x = (x_1, x_2)^\top \in I\!R^2$ and integers m_1, m_2. The family of basis functions can now be written in the form $\{\psi_m(x), |m| \le 2\}$. For numerical stability the centered functions $\psi_m(x - X_i)$ should be used within each local model.

At the kth estimation step one computes the entries $\widehat{\theta}_{i,m}^{(k)}$, $|m| \le 2$, of the vector $\widehat{\boldsymbol{\theta}}_i^{(k)}$. At the kth adaptation step we additionally need, for every i, to recompute the vectors $\widehat{\boldsymbol{\theta}}_j^{(k-1)}$ for the basis system $\Psi(X_i)$. Similarly to the univariate case, we get

$$\widehat{\theta}_{ij,m}^{(k-1)} = \sum_{m':|m'|\le 2-|m|} \binom{m+m'}{m} \widehat{\theta}_{j,m+m'}^{(k-1)} (X_i - X_j)^{m'}, \qquad |m| \le 2.$$

Here $\sum_{m':|m'|\le 2-|m|}$ means the sum over the set of all pair $m' = (l_1', l_2')$ with $m_1' + m_2' \le 2 - m_1 - m_2$ and $\binom{m}{m'} = \binom{m_1}{m_1'}\binom{m_2}{m_2'}$. Particularly, $\widehat{\theta}_{ij,m}^{(k-1)} = \widehat{\theta}_{j,m}^{(k-1)}$ for all m with $|m| = 2$, and $\widehat{\theta}_{ij,0} = \widehat{f}_j(X_i)$. The rest of the procedure remains as before.

5. Numerical results

We now demonstrate the performance of the method for artificial examples in univariate and bivariate regression. The aim of this study is to illustrate two important features of the procedure: adaptability to large homogeneous regions and sensitivity to sharp changes in the local structure of the model. We also try to give some hints about the choice of the degree of local polynomial approximation.

Estimates are obtained using **R**, a language and environment for statistical computing, and its contributed libraries pspline (J. Ramsay and B. Ripley), waveslim (B. Whitcher) and aws (J. Polzehl).

Our univariate simulations are conducted generating data as (X_i, Y_i) with $Y_i = f(X_i) + \varepsilon_i$. The sample size is $n = 1000$. The design is chosen as an equidistant grid on $(0,1)$. Errors ε_i are i.i.d. Gaussian or t-distributed. The error variance σ^2 is assumed to be unknown.

Local linear ($p = 1$), local quadratic ($p = 2$) and local cubic ($p = 3$) AWS estimates are computed for 1000 simulated data sets using a maximal bandwidth $h_{\max} = 0.25$ and defaults, see Table 1, for the other parameters. The parameter τ was selected to provide the propagation condition in the pure parametric situation.

Table 1. Default parameters used for the AWS procedure

	λ				η	τ		
p	0	1	2	3		1	2	3
univariate	$q_{\chi^2;.966,1}$	$q_{\chi^2;.92,2}$	$q_{\chi^2;.92,3}$	$q_{\chi^2;.92,4}$	.5	4.5	13.5	40
bivariate	$q_{\chi^2;.966,1}$	$q_{\chi^2;.92,3}$	$q_{\chi^2;.92,6}$	-	.5	13.5	150	-

For comparison we use a penalized cubic smoothing spline, with smoothing parameter determined by generalized cross validation. See Heckman and Ramsey (2000) for details. Such a choice was motivated by excellent numerical results delivered by this method for many situations. We also tried other more sophisticated procedures like wavelets,

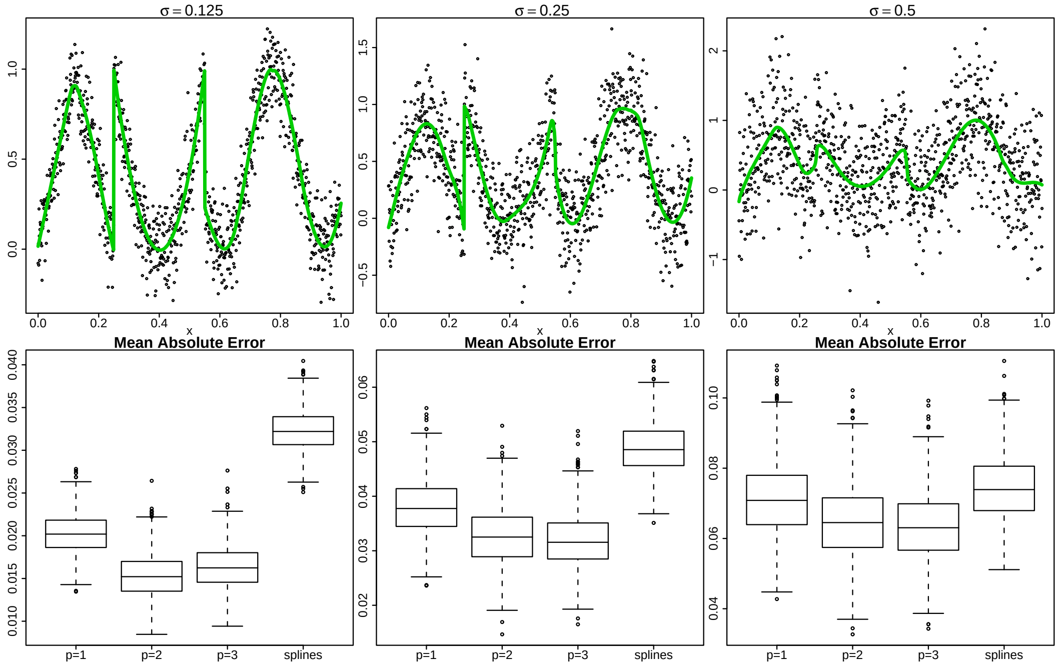

FIGURE 1. Univariate Example 1: Simulated data sets with local cubic AWS estimates ($h_{\max} = 0.6$) and Box-Plots of MAE for local linear, local quadratic and local cubic AWS and penalized cubic smoothing splines.

pointwise adaptive procedures, Markov Random Fields methods but the numerical results (not reported here) were always in favor of smoothing splines, see also PS2000.

5.1. **Univariate Example 1.** Our first example uses the piecewise smooth function

$$f(x) = \begin{cases} 8x & x < 0.125, \\ 2 - 8x & 0.125 \le x < 0.25, \\ 44(x-0.4)^2 & 0.25 \le x < 0.55, \\ 0.5\cos(6\pi(x-0.775) + 0.5 & 0.55 \le x. \end{cases}$$

The upper row of Figure 1 shows plots of the first data set for $\sigma = 0.125, 0.25$ and 0.5, respectively, together with the estimate obtained by local cubic AWS with default parameters and $h_{\max} = 0.6$. The bottom row reports the results in form of box-plots of Mean Absolute Error (MAE) obtained for the four procedures in 1000 simulation runs.

Figure 2, providing pointwise estimates of the Mean Absolute Error for three procedures in case of $\sigma = 0.125$, illustrates the local behavior of the procedures. Especially the local linear AWS is superior to the cubic smoothing spline both near the discontinuities and within smooth regions. Local quadratic AWS seems to be a bit more variable near the first singularity $x = 0.125$, but behaves excellent for the rest of the design. Advantages are due to the local adaptivity of the AWS procedures in contrast to the global nature of the smoothing spline.

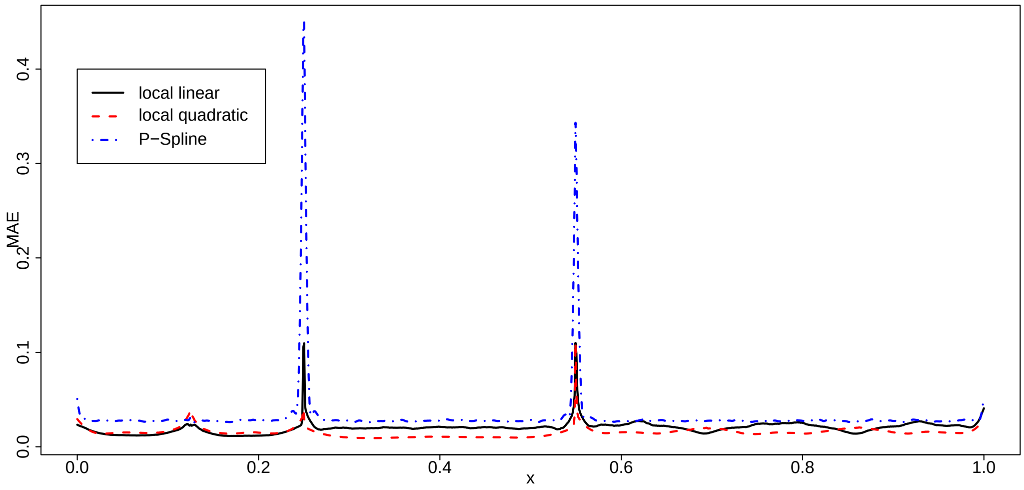

FIGURE 2. Univariate Example 1: Estimated pointwise Mean Absolute Error for local linear and local quadratic AWS and penalized cubic smoothing splines, $\sigma = 0.125$.

5.2. **Univariate example 2.** The second univariate example uses a smooth regression function with varying second derivative:

$$f(x) = \sin(2.4\pi/(x + 0.2)) .$$

The upper row of Figure 3 provides a typical data set for $\sigma = 0.125, 0.25, 0.5$ and 1,

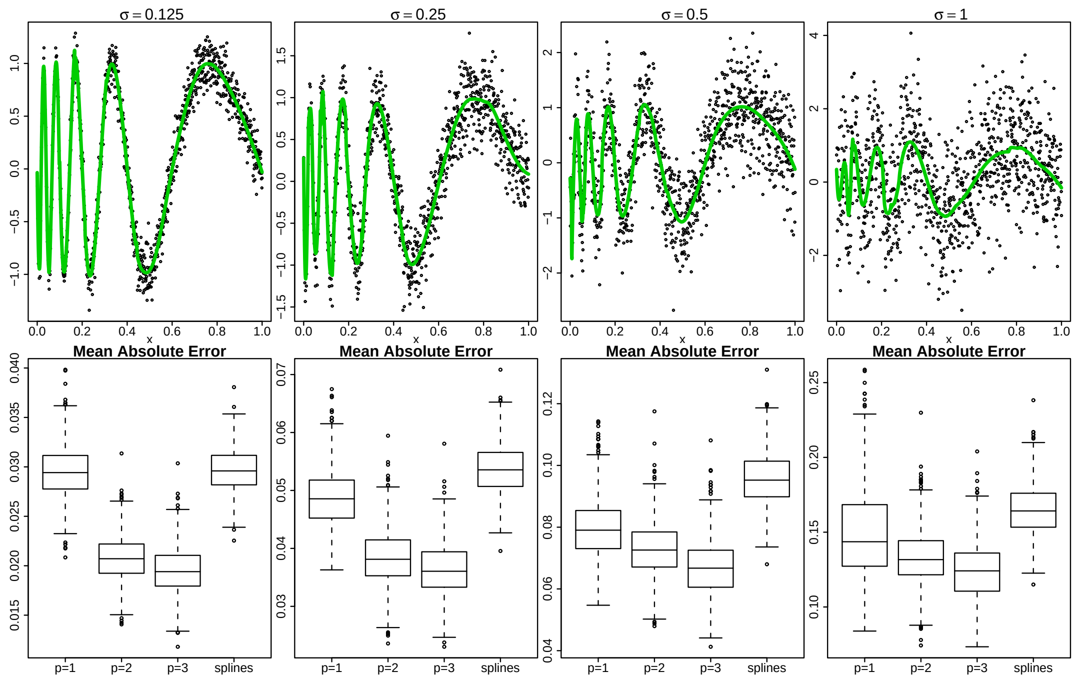

FIGURE 3. Univariate Example 2: Simulated data sets with local cubic AWS estimates ($h_{\max} = 0.3$) and Box-Plots of MAE for local linear, local quadratic and local cubic AWS and penalized cubic smoothing splines.

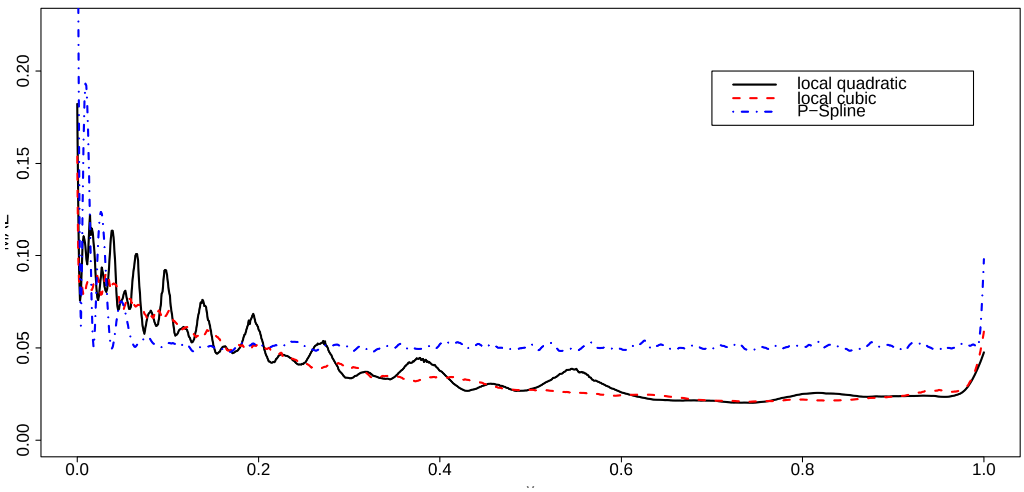

FIGURE 4. Example 2: Estimated pointwise Mean Absolute Error for local quadratic AWS, local cubic AWS and penalized cubic smoothing splines in case of $\sigma = 0.25$.

TABLE 2. MAE for Example 2 with t-distributed errors

Error distribution	linear AWS	quadratic AWS	cubic AWS	splines
t_3	0.0808	0.0730	0.0673	0.0914
t_5	0.0808	0.0728	0.0665	0.0948
t_{10}	0.0810	0.0730	0.0662	0.0957
$\mathcal{N}(0, 0.25)$	0.0813	0.0727	0.0657	0.0955

respectively, together with the local cubic AWS estimate obtained from this data set using standard parameters and $h_{\max} = 0.3$. The bottom row contains box-plots of Mean Absolute Error obtained for the four procedures in 1000 simulation runs.

Figure 4 again gives pointwise estimates of the Mean Absolute Error. Results are shown for local quadratic and cubic AWS and the penalized cubic smoothing splines in case of $\sigma = 0.25$. The AWS procedures perform better in regions where the regression function is highly fluctuating or very smooth while the smoothing spline delivers better results only in a small region of medium fluctuation, i.e. for $x \in (0.05, 0.2)$ where the global smoothing parameter of the spline is nearly optimal in the local sense as well. For small values of x the spline suffers from high bias while for large values variability dominates. AWS delivers a good compromise in all cases.

Table 2 provides simulation results obtained using the same regression function and t-distributed errors, with 3, 5, 10 and ∞ degrees of freedom. Errors are rescaled to have variance $\sigma^2 = 0.25$. A maximal bandwidth of $h_{\max} = 0.3$ is used. Simulation size is 1000. The results show a robust behavior with respect to non-Gaussian errors.

TABLE 3. MAE and optimal parameters of bivariate reconstructions

	AWS ($h_{\max}$)			Local polynomial (h)			Wavelets	
	$p=0$	$p=1$	$p=2$	$p=0$	$p=1$	$p=2$	DWT	MODWT
MAE	0.076	0.042	0.024	0.102	0.098	0.104	0.127	0.063
Parameters	0.075	0.3	0.74	0.053	0.062	0.099	D4, 5	Haar, 5

5.3. **Bivariate Example.** We use the following example to demonstrate the behavior of our procedure in a bivariate design. Data are generated on a equidistant grid of 100×100 points in $[-1, 1]^2$ using the regression function:

$$f(x, y) = (4x^2 + 8y^3)\text{sign}(4x^2 - 4xy - 6y^3)$$

and additive errors with variance $\sigma^2 = 0.25$. The upper left of Figure 5 shows a perspective plot of the data. The three panels in the right column provide plots of the estimated surface obtained by local constant, local linear and local quadratic AWS, with standard values for λ and τ . For a comparison we computed reconstructions of the surface using local polynomial estimates of order 0 , 1 and 2 using optimal global bandwidths and reconstructions using a discrete wavelet transform (DWT) and a maximum overlap discrete wavelet transform (MODWT), see e.g. Gencay, Selcuk and Whitcher (2001), with optimal basis and optimal depth of decomposition. We used R-package *waveslim* of Brandon Whitcher in the last two cases. The reconstructions using the local linear estimator and MODWT are shown in the left column of Figure 5.

Table 3 provides the mean absolute error (MAE) of the reconstructions together with the parameters used, i.e. $h_{\max}$ for AWS, optimal bandwidth h for the local polynomial estimators and basis system and depth of decomposition for wavelets. The results clearly illustrate the advantages of AWS compared to local polynomial smoothing if the unknown regression function is piecewise smooth. AWS automatically detects discontinuities and therefore allows for a larger bandwidth within smooth regions, resulting in a larger variance reduction. AWS also outperforms wavelet approaches on this example due to its more flexible handling of boundaries. Best results are obtained for the local quadratic approach, while local constant AWS suffers from a segmentation effect caused by its inappropriate structural assumption.

5.4. **Summary.** The performance of the AWS method is completely in agreement with what was aimed: it is adaptive to variable smoothness properties of the underlying function and sensitive to discontinuities outperforming the classical smoothing methods. It demonstrates excellent results for a small or moderate noise and it is stable with respect to large noise.

Local quadratic AWS seems to be a reasonable compromise for many situations combining a good approximating properties with a very good quality of change-point or edge

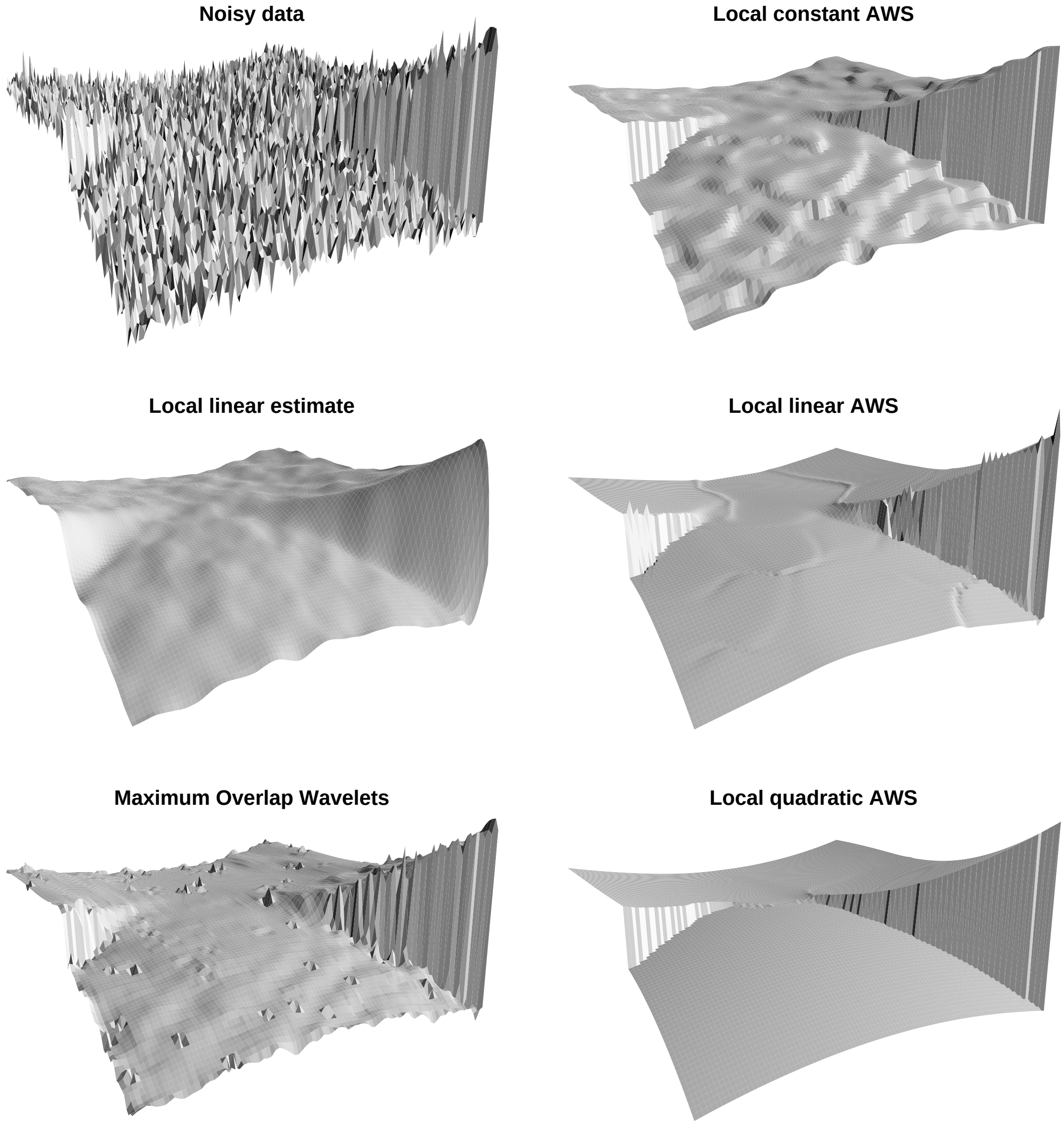

FIGURE 5. Bivariate Example: Perspective plots of data (upper left), local constant (upper right), local linear (lower left) and local quadratic (lower right) reconstruction.

estimation. In situations with large homogeneous regions, local polynomial approximation of a higher order can be slightly preferable. The choice of the polynomial degree can be also done automatically using global cross-validation type criteria.

Our experiments (not reported here) demonstrate that the procedure is rather stable w.r.t. to the choice of the parameters λ , τ , η , $h_{\max}$, that is, a moderate change of these parameters near default values does not significantly affect the quality of estimation. In the most of cases, only a minor improvement can be achieved by tuning these parameters.

6. Some important properties of AWS

This section discusses some properties of the 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 iteration process. Further we discuss the rate of estimation for a smooth function $\boldsymbol{\theta}(x)$.

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

The analysis of the properties of the iterative estimates $\widehat{\boldsymbol{\theta}}_i^{(k)}$ is very difficult. The main reason is that every estimate $\widehat{\boldsymbol{\theta}}_i^{(k)}$ 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 is only of logarithmic order this split can change the quality of estimation only by a logarithmic factor. Of course, this is only a theoretical device, the use of the same sample for all steps of the algorithm still requires a further justification.

In our study we restrict ourselves to the case of the varying coefficient model with homogeneous Gaussian noise:

(A1) The observations $Y_1, \ldots, Y_n$ follow the model $Y_i = f(X_i) + \varepsilon_i$ where $\varepsilon_1, \ldots, \varepsilon_n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$.

The results can be easily extended to the case when the errors ε_i have uniformly bounded exponential moments. To simplify the presentation we also assume that

(A2) The statistical penalty $\boldsymbol{s}_{ij}^{(k)}$ is defined via the likelihood ratio test statistic T_{ij}° from (3.2) in Section 3.1.

In our procedure the statistic T_{ij} from (3.3) is applied. However, an essential difference between T_{ij} and T_{ij}° only occurs in the situations where the local models W_i and W_j are strongly unbalanced, which do not meet in the specific cases considered here.

(A3) The extension penalty $\boldsymbol{e}_{ij}^{(k)}$ is set to zero, that is, $K_e(\boldsymbol{e}_{ij}^{(k)}) = 1$.

Again, this assumption is not restrictive because the extension penalty does not matter as long as the propagation condition is studied.

We first consider the homogeneous situation with $\boldsymbol{\theta}_i = \boldsymbol{\theta}$ which corresponds to a global linear model $f(x) = \theta_1 \psi_1(x) + \ldots + \theta_p \psi_p(x)$.

Theorem 6.1. *Let (A0), (A1), (A2) and (A3) be fulfilled. Suppose that* $\boldsymbol{\theta}(X_i) \equiv \boldsymbol{\theta}$ *, i.e.* $f = \Psi\boldsymbol{\theta}$ *. If* $\lambda \geq C \log n$ *with constant* C *depending on the kernel* K_s *only, then for every iteration* k

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

Proof. Define b by the equation $K_s(b) = 1/2$. Theorem 8.3 from the Appendix yields for every iteration k

$$\boldsymbol{P}\left(\min_{i,j=1,\ldots,n} K_s(\boldsymbol{s}_{ij}^{(k)}) > 1/2\right) = \boldsymbol{P}\left(\max_{i,j=1,\ldots,n} T_{ij}^{(k)} \leq b\lambda\right) \geq 1 - \sum_{i,j=1}^{n} q_p(b\lambda - p)$$

where $q_p(u)$ is defined by $\log q_p(u) = -u/2 + 0.5p\log(1 + u/p)$. It is easy to see that $q_p(u)$ fulfills $\log q_p(u) \leq -2\log n$ for $u \geq C_p \log n$ with some constant C_p depending on p only. This yields the assertion as soon as $b\lambda - p \geq C_p \log n$, or, equivalently, $\lambda \geq (p + C_p \log n)/b$. □

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 any local model.

Corollary 6.2. *Let the assumptions (A0), (A1), (A2) and (A3) be fulfilled and* $\boldsymbol{\theta}(X_i) \equiv \boldsymbol{\theta}$ *. If* $\lambda \geq C \log n$ *and* $h_{\max}$ *is sufficiently large, then the last step estimate* $\widehat{\boldsymbol{\theta}}_i = \widehat{\boldsymbol{\theta}}_i^{(k^*)}$ *fulfills for every* $z \geq 0$

$$\boldsymbol{P}\left((2\sigma^2)^{-1}(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta})^\top \Psi\Psi^\top (\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}) > p + z\right) \leq q_p(z)$$

where $\log q_p(u) = -u/2 + 0.5p\log(1 + u/p)$ *.*

Proof. If $h_{\max}$ is sufficiently large then the location penalty $K_l(\boldsymbol{l}_{ij}^{(k)})$ at the final iteration $k = k^*$ fulfills $K_l(\boldsymbol{l}_{ij}^{(k)}) \approx 1$ for every pair (i,j). By Theorem 6.1 the statistical penalty $K_s(\boldsymbol{s}_{ij}^{(k)}) \geq 1/2$, hence $w_{ij}^{(k)} \geq 1/2$ for all (i,j). This yields $\Psi W_i^{(k)} \Psi^\top \geq 0.5\,\Psi\Psi^\top$ and the result follows from Theorem 8.1 in the Appendix. □

Due to this result the final estimate $\widehat{\boldsymbol{\theta}}_i = \widehat{\boldsymbol{\theta}}_i^{(k^*)}$ delivers the same quality of estimation as the global LSE $\widehat{\boldsymbol{\theta}} = (\Psi\Psi^\top)^{-1}\Psi Y$. In fact, one can show an even stronger assertion: with a high probability it holds $\widehat{\boldsymbol{\theta}}_i \approx \widehat{\boldsymbol{\theta}}$.

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 are 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 6.3. *Let the assumptions (A0), (A1) and (A2) be fulfilled. Suppose that* $\boldsymbol{\theta}(X_i) \equiv \boldsymbol{\theta}$ *for all* X_i *from some region* G *in* $\mathcal{X}$. *If* $\lambda \geq C \log n$ *for some constant* C *depending on the kernel* K_s *only, then for every iteration* k

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

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 8.3. □

6.2. **Accuracy of estimation for a varying coefficient model.** We consider the case of an arbitrary function f which allows a good linear approximation in a 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.

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

- **(A4)** It holds $|\Psi_j^\top [\boldsymbol{\theta}(X_j) - \boldsymbol{\theta}(x)]\,| \leq \delta$ for some $\delta \geq 0$ and all $X_j \in U_h(X_i)$.
- **(A5)** The kernel K_l is compactly supported on $[0,1]$.
- **(A6)** Define $W_i^* = \operatorname{diag}\{w_{i1}^*, \ldots, w_{in}^*\}$ with $w_{ij}^* = K_l(|\rho(X_i,X_j)/h|^2)$, $N_i^* = \operatorname{tr} W_i^*$ and $B_i^* = \Psi W_i^* \Psi^\top$. It holds $N_i^*\,\Psi_i \Psi_i^\top \leq C_B\, B_i^*$.

Condition (A4) means that the value $f(X_j)$ can be approximated by a linear expression $\Psi_j^\top \boldsymbol{\theta}(x)$ with the precision δ for every $X_j \in U_h(x)$. Condition (A6) guarantees a certain design regularity in a neighborhood of the reference point x. The next result claims the propagation condition (free extension) for the local models $W_i^{(k)}$ as long as $h^{(k)} \leq h$ provided that δ is sufficiently small.

Theorem 6.4. *Let the assumptions (A0) through (A6) be fulfilled. Let* $\lambda \geq C \log n$ *for some constant* C *depending on the kernel* K_s *only. If*

$$2\sigma^{-2} p \delta^2 (N_i^* + N_j^*) \leq b\lambda/6, \qquad X_j \in U_h(x), \tag{6.1}$$

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

$$\boldsymbol{P}\left(\min_{j\colon X_j \in U_h(X_i)} K_s(\boldsymbol{s}_{ij}^{(k)}) \geq 1/2\right) \geq 1 - 1/n. \tag{6.2}$$

If $h^{(k)} = h$, *then the estimate* $\widehat{f}_i^{(k)} = \Psi_i \widehat{\theta}_i^{(k)}$ *of* $f_i = f(X_i)$ *fulfills*

$$\boldsymbol{P}\left(\left|\widehat{f}_i^{(k)} - f_i\right| > \sqrt{pC_B}\,\delta + \sigma\sqrt{2C_B\lambda/N_i^*}\right) \leq 2/n. \tag{6.3}$$

The proof is given in the Appendix. The result (6.3) indicates that the first k iterations of the procedure (for $h^{(k)} \leq h$) lead to a reasonable quality of estimation of the function $f(\cdot)$. 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 3.5. The procedure presented here applies this control step in a soft form, however we only show how the *hard* control step (3.6) can be used for proving the rate result.

Theorem 6.5. *Let the conditions of Theorem 6.4 be fulfilled and let the procedure involve the control step from (3.6) with* $\eta^* \ge \lambda$ *. Then the last step estimate* $\widehat{f}_i = \Psi_i^\top \widehat{\boldsymbol{\theta}}_i^{(k^*)}$ *of* $f_i = f(X_i)$ *fulfills with a probability of at least* $1 - 2/n$

$$\left|\widehat{f}_i - f_i\right| \le \sqrt{pC_B}\delta + \sigma\sqrt{2C_B\lambda/N_i^*} + \sigma\sqrt{2C_B\eta^*/N_i^*}.$$

Proof. Let $h = h^{(k)}$ for some k. The control step (3.6) ensures that

$$(2\sigma^2)^{-1}\bigl(\widehat{\boldsymbol{\theta}}_i^{(k)} - \widehat{\boldsymbol{\theta}}_i^{(k^*)}\bigr)^\top B_i^{(k)}\bigl(\widehat{\boldsymbol{\theta}}_i^{(k)} - \widehat{\boldsymbol{\theta}}_i^{(k^*)}\bigr) \le \eta^*.$$

This yields by (A6)

$$\begin{aligned} N_i^{(k)}\bigl|\widehat{f}_i^{(k)} - \widehat{f}_i\bigr|^2 &= N_i^{(k)}\bigl(\widehat{\boldsymbol{\theta}}_i^{(k)} - \widehat{\boldsymbol{\theta}}_i^{(k^*)}\bigr)^\top \Psi_i\Psi_i^\top\bigl(\widehat{\boldsymbol{\theta}}_i^{(k)} - \widehat{\boldsymbol{\theta}}_i^{(k^*)}\bigr) \\ &\le C_B\bigl(\widehat{\boldsymbol{\theta}}_i^{(k)} - \widehat{\boldsymbol{\theta}}_i^{(k^*)}\bigr)^\top B_i^{(k)}\bigl(\widehat{\boldsymbol{\theta}}_i^{(k)} - \widehat{\boldsymbol{\theta}}_i^{(k^*)}\bigr) \le 2\sigma^2 C_B\eta^*. \end{aligned}$$

By Theorem 6.4 $N_i^{(k)} \ge 0.5N_i^*$ with a high probability and the assertion follows directly from (6.3). $\square$

6.3. **Rate of estimation for a smooth function $f(\cdot)$. Spatial adaptivity.** Here we briefly discuss one important special case of the result of Theorem 6.5. Namely, we suppose that $f(\cdot)$ is a smooth function in $I\!R^d$ and consider the polynomial basis $\{\psi_m\}$ of degree less than a given integer number s. In the univariate case $d = 1$ there are exactly $p = s$ basis functions, e.g. $1, u - x, \ldots, (u - x)^{s-1}$. We also suppose that

(A4s) The function $f(\cdot)$ is s times continuously differentiable and $|f^{(s)}(u)| \le Ls!$ for some constant L and all $u \in U_h(x)$.

(A7) For some positive constants $C_{X1} \le C_{X2}$ and for all $h \in [h^{(0)}, h_{\max}]$ holds

$$C_{X1} \le N_h^*/(nh^d) \le C_{X2}.$$

where $N_h^* = \sum_{j=1}^n K_l(|\rho(X_i, X_j)/h|^2)$.

Note that condition (A4s) ensures (A4) with $\delta = Lh^s$. We now apply Theorem 6.5 to this situation with $\eta^* = \lambda$. The result is formulated as a separate statement.

Theorem 6.6. *Suppose that (A0), (A1), (A2), (A3), (A4s), (A5), (A7) are fulfilled and (A6) holds for all* $h \in [h^{(0)}, h_{\max}]$*. If* $\lambda \ge C\log n$ *for some fixed* C*, then*

$$\boldsymbol{P}\left(\bigl|\widehat{f}_i - f_i\bigr| > C_1(\lambda\sigma^2/n)^{s/(d+2s)}L^{d/(d+2s)}\right) \le 2/n$$

where the constant C_1 *depends on* C_{X1}, C_{X2} *and* C_B *only.*

Proof. The bound (6.3) and condition (A7) imply with a high probability

$$\left|\widehat{f}_i - f_i\right| \le \sqrt{pC_B}\delta + 2\sigma\sqrt{2C_B\lambda/N_i^*} \le \sqrt{pC_B}Lh^s + 2\sigma\sqrt{2C_B\lambda/(C_{X1}nh^d)}.$$

Optimizing this expression w.r.t. h leads to $h = C_2\{\lambda\sigma^2/(nL^2)\}^{1/(d+2s)}$. With this choice condition (6.1) is fulfilled in view of (A7) provided that C_2 is not too large. Using such an h results in the accuracy of order $\{\lambda\sigma^2/n\}^{s/(d+2s)}L^{d/(d+2s)}$ as required. □

The accuracy shown in Theorem 6.6 is optimal in rate for the problem of estimation of a smooth function f up to a logarithmic factor λ. Therefore, this result means that our procedure is pointwise adaptive in the sense that it automatically adapts to the unknown local smoothness degree measured by the exponent s and 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 $B^s_{p,q}$.

7. Summary and Outlook

The paper presents a new general method of local linear modeling based on the adaptive weights idea. The method has a number of remarkable properties. In particular, AWS applies in a unified way to a broad class of regression models, and the procedure is able to adapt to the unknown and variable function structure without requiring any specific prior information like the degree of smoothness of the underlying regression function. These features are justified both by our theoretical results and by numerical examples.

Similarly to local polynomial smoothing, the AWS method is design adaptive and has no boundary problem. The produced estimate does not exhibit the usual Gibbs effect (high variability and increased bias near discontinuities).

AWS applies for high dimensional models. However, for local linear or local polynomial modeling, the number of parameters grows dramatically with the dimension d, and the procedure can face the so called “curse of dimensionality” problem: in high dimension, pure nonparametric modeling leads to strong oversmoothing. Specifically for the AWS method, if the number of parameters becomes too high (say, more than 6) then the procedure looses sensitivity to structural changes. For such situation, combining the procedure with some dimension reduction methods can be useful.

The AWS method is computationally straightforward and the numerical complexity can be easily controlled, see Section 3.4.

The presented procedure is however restricted to the case of a local linear model. An extension to generalized linear models with varying coefficients is important for many applications, see Cai, Fan and Li (2000). This will be a subject for further development.

8. Appendix

Here we present some general results on large deviation probabilities for local likelihood ratio test statistics in Gaussian regression.

We consider the varying coefficient regression model $Y_i = f(X_i) + \varepsilon_i$ with homogeneous Gaussian errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. The local model W is described by the weights $w_1, \ldots, w_n$. Local linear modeling assumes the linear structure of the model function f within the local model W: $f(x) = \theta_1 \psi_1(x) + \ldots + \theta_p \psi_p(x)$ for a given system $\{\psi_m(x)\}$. The corresponding local MLE $\widehat{\boldsymbol{\theta}}$ can be represented in the form $\widehat{\boldsymbol{\theta}} = \left(\Psi W \Psi^\top\right)^{-1} \Psi W Y$ with the notation from Section 2.2. The local likelihood ratio test statistic is defined for a given $\boldsymbol{\theta}$ by $L(W, \widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top B (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})/(2\sigma^2)$ where $B = \Psi W \Psi^\top$.

8.1. **Linear parametric case.** Define $\overline{\boldsymbol{\theta}} = B^{-1} \Psi W f$. Then $\Psi \overline{\boldsymbol{\theta}}$ is the best linear approximation of f within the local model W. In the homogeneous case $f = \Psi^\top \boldsymbol{\theta}$, it obviously holds $\overline{\boldsymbol{\theta}} = \boldsymbol{\theta}$. The first result shows that $\widehat{\boldsymbol{\theta}}$ is a good estimate of the vector $\overline{\boldsymbol{\theta}}$. This particularly implies nice properties of the estimate in a homogeneous situation when the local linear assumption is fulfilled and $\boldsymbol{\theta}$ is the true parameter.

Theorem 8.1. *For every* $z \geq 0$

$$\boldsymbol{P}\left(2L(W, \widehat{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}) > p + z\right) \leq q_p(z)$$

where

$$q_p(z) = \exp\left(-0.5z + 0.5p \log(1 + z/p)\right). \tag{8.1}$$

Proof. The model equation $Y = f + \varepsilon$ immediately implies that $\widehat{\boldsymbol{\theta}}_i = B_i^{-1} \Psi W_i Y = \overline{\boldsymbol{\theta}}_i + B_i^{-1} \Psi W_i \varepsilon$. Therefore, $\widehat{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_i = B_i^{-1} \Psi W_i \varepsilon$ does not depend on $\boldsymbol{\theta}$, and we assume without loss of generality that $\boldsymbol{\theta} = 0$, so that the observations Y_i coincide with the noise ε_i. This obviously implies $\boldsymbol{E}\widehat{\boldsymbol{\theta}} = 0$. The covariance matrix V of the estimate $\widehat{\boldsymbol{\theta}}$ can be represented as

$$V = \boldsymbol{E}\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}^\top = \boldsymbol{E} B^{-1} \Psi \varepsilon \varepsilon^\top \Psi^\top B^{-1} = \sigma^2 B^{-1} D B^{-1}$$

where $D = \Psi W^2 \Psi^\top$. Therefore, the estimate $\widehat{\boldsymbol{\theta}}$ can be expressed as $\widehat{\boldsymbol{\theta}} = V^{1/2} \zeta$ where ζ is a standard Gaussian random vector in $\mathbb{R}^p$. This yields

$$L(W, \widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (2\sigma^2)^{-1} \zeta^\top V^{1/2} B V^{1/2} \zeta = 0.5 \zeta^\top R \zeta$$

with $R = B^{-1/2} D B^{-1/2}$. Since $w_i \leq 1$, it holds $D \leq B$ and $\|R\| \leq 1$, that is, the largest eigenvalue of R does not exceed one. Now the desired result follows from the general result for Gaussian quadratic forms in Lemma 8.2. $\square$

Lemma 8.2. *Let a symmetric* $p \times p$*-matrix* R *fulfill* $\|R\| \leq 1$. *Then*

$$\boldsymbol{P}\left(\zeta^\top R \zeta \geq p + z\right) \leq q_p(z).$$

Proof. Let $r_1, \ldots, r_p$ be the eigenvalues of R satisfying $r_m \leq 1$ for all m. It holds for every $\mu < 1$ by simple algebra

$$\log \boldsymbol{E} \exp(\mu \zeta^\top R \zeta / 2) = \log \prod_{m=1}^{p} \frac{1}{\sqrt{1 - \mu r_m}} = -\frac{1}{2} \sum_{m=1}^{p} \log(1 - \mu r_m) \leq -0.5 p \log(1 - \mu).$$

Now the exponential Chebyshev inequality implies

$$\begin{aligned}\log \boldsymbol{P}\Bigl(0.5\zeta^\top R\zeta \ge (p+z)/2\Bigr) &\le -\mu(p+z)/2 + \log \boldsymbol{E}\Bigl(0.5\mu\zeta^\top R\zeta\Bigr)\\ &\le -0.5\mu(p+z) - 0.5p\log(1-\mu).\end{aligned}$$

This expression is maximized by $\mu = z/(p+z)$ leading to

$$\log \boldsymbol{P}\Bigl(\zeta^\top R\zeta \ge p+z\Bigr) \le -0.5z + 0.5p\log(1+z/p)$$

as required. □

Next we consider the likelihood ratio test statistic T°_{ij} defined in Section 3.6 for two local models W_i and W_j.

Theorem 8.3. *Let* $f = \Psi^\top \boldsymbol{\theta}$. *Then for every* $z \ge 0$

$$\boldsymbol{P}\bigl(T^\circ_{ij} > p+z\bigr) \le q_p(z).$$

Proof. We use the expression $2T^\circ_{ij} = \sigma^{-2}(\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_j)^\top B_i(B_i+B_j)^{-1}B_j(\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_j)$. Note that

$$\operatorname{Cov}(\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_j) \le 2\operatorname{Cov}(\widehat{\boldsymbol{\theta}}_i) + 2\operatorname{Cov}(\widehat{\boldsymbol{\theta}}_j) = 2V_i + 2V_j \le 2\sigma^2(B_i^{-1} + B_j^{-1}).$$

Now the result follows from Lemma 8.2 similarly to the proof of Theorem 8.1. □

8.2. **Sufficient conditions for free extension.** We consider the general situation of a varying coefficient model. We show that if the difference between two local models defined in terms of the Kullback-Leibler distance, is sufficiently small, then T°_{ij} is with a large probability smaller then $b\lambda$ for some $b \le 1$.

Theorem 8.4. *Let* $b \in (0,1]$ *be such that* $z = b\lambda/2 - p > 0$. *Then the condition*

$$\varDelta := 0.5\sigma^{-2}(\overline{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_j)^\top B_i(B_i+B_j)^{-1}B_j(\overline{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_j) \le b\lambda/6 \tag{8.2}$$

with $\overline{\boldsymbol{\theta}}_i = B_i^{-1}\Psi W_i f$ *and* $\overline{\boldsymbol{\theta}}_j = B_j^{-1}\Psi W_j f$ *implies*

$$\boldsymbol{P}\bigl(T^\circ_{ij} > b\lambda\bigr) \le q_p(z) + e^{-b\lambda/12}.$$

Proof. We use the decomposition

$$\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_j = \xi_i - \xi_j + \overline{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_j$$

where $\xi_i = B_i^{-1}\Psi W_i\varepsilon$ and similarly for ξ_j. This implies with $B_{ij} = B_i(B_i+B_j)^{-1}B_j$

$$2\sigma^2 T^\circ_{ij} = (\xi_i - \xi_j)^\top B_{ij}(\xi_i - \xi_j) + (\overline{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_j)^\top B_{ij}(\overline{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_j) + 2(\overline{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_j)^\top B_{ij}(\xi_i - \xi_j). \tag{8.3}$$

The result of Theorem 8.3 implies

$$\boldsymbol{P}\Bigl(\sigma^{-2}(\xi_i - \xi_j)^\top B_{ij}(\xi_i - \xi_j) > p+z\Bigr) \le q_p(z).$$

Next, $\zeta_{ij} = \sigma^{-2}(\overline{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_j)^\top B_{ij}(\xi_i - \xi_j)$ is a Gaussian random variable with zero mean satisfying

$$\begin{aligned} \boldsymbol{E}\zeta_{ij}^2 &= \sigma^{-4}(\overline{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_j)^\top B_{ij} \operatorname{Cov}(\xi_i - \xi_j) B_{ij}(\overline{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_j) \\ &\le 2\sigma^{-2}(\overline{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_j)^\top B_{ij}(\overline{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_j) \le 4\Delta. \end{aligned} \tag{8.4}$$

Here we have used that $\operatorname{Cov}(\xi_i - \xi_j) \le 2\sigma^2 B_{ij}$, see the proof of Theorem 8.3. This and condition (8.2) imply

$$\boldsymbol{P}\left(\zeta_{ij} > b\lambda/3\right) \le e^{-b\lambda/12}.$$

Since $p + z = b\lambda$, we finally obtain

$$\begin{aligned} \boldsymbol{P}\left(T_{ij}^\circ > b\lambda\right) &\le \boldsymbol{P}\Big(0.5\sigma^{-2}(\xi_i - \xi_j)^\top B_{ij}(\xi_i - \xi_j) \ge p + z\Big) + \boldsymbol{P}\left(\zeta_{ij} > b\lambda/3\right) \\ &\le q_p(z) + e^{-b\lambda/12} \end{aligned}$$

as required. □

The next assertion delivers some sufficient conditions ensuring (8.2). More precisely, we consider the situation when the function f can be well approximated by a linear function $\Psi^\top\boldsymbol{\theta}$ within both local models W_i and W_j. If $|f(X_l) - \Psi_l^\top\theta| \le \delta$ for some small positive δ and all X_l entering with positive weight in the model W_i, then $(f-\Psi^\top\boldsymbol{\theta})^\top W_i(f-\Psi^\top\boldsymbol{\theta}) = \sum_l w_{il}|f(X_l)-\Psi_l^\top\theta|^2 \le N_i\delta^2$ with $N_i = \sum_l w_{il}$ and similarly for the model W_j.

Lemma 8.5. *The condition*

$$(f - \Psi^\top\boldsymbol{\theta})^\top W_i(f - \Psi^\top\boldsymbol{\theta}) \le \delta^2 N_i$$

implies

$$(\overline{\boldsymbol{\theta}}_i - \boldsymbol{\theta})^\top B_i(\overline{\boldsymbol{\theta}}_i - \boldsymbol{\theta}) \le p\delta^2 N_i\,.$$

If, in addition, $(f - \Psi^\top\boldsymbol{\theta})^\top W_j(f - \Psi^\top\boldsymbol{\theta}) \le N_j\delta^2$, *then*

$$(\overline{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_j)^\top B_{ij}(\overline{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_j) \le 2p\delta^2(N_i + N_j)$$

where $B_{ij} = B_i(B_i + B_j)^{-1}B_j$.

Proof. The use of $B_i = \Psi W_i \Psi^\top$ and $\overline{\boldsymbol{\theta}}_i = B_i^{-1}\Psi W_i f$ gives

$$(\overline{\boldsymbol{\theta}}_i - \boldsymbol{\theta})^\top B_i(\overline{\boldsymbol{\theta}}_i - \boldsymbol{\theta}) = (f - \Psi^\top\boldsymbol{\theta})^\top W_i\Psi^\top B_i^{-1}\Psi W_i(f - \Psi^\top\boldsymbol{\theta})$$

Define $A = W_i^{1/2}\Psi^\top B_i^{-1}\Psi W_i^{1/2}$. Then

$$\operatorname{tr} AA^\top = \operatorname{tr} W_i^{1/2}\Psi^\top B_i^{-1}\Psi W_i\Psi^\top B_i^{-1}\Psi W_i^{1/2} = \operatorname{tr} B_i^{-1}\Psi W_i\Psi^\top = \operatorname{tr} I_p = p.$$

Therefore, by the Cauchy-Schwarz inequality

$$|(\overline{\boldsymbol{\theta}}_i - \boldsymbol{\theta})^\top B_i(\overline{\boldsymbol{\theta}}_i - \boldsymbol{\theta})|^2 \le \|W_i^{1/2}(f - \Psi^\top\boldsymbol{\theta})\|^2 \operatorname{tr} AA^\top \le N_i\delta^2 p$$

and the first assertion follows.

Since $B_{ij} \le B_i$ and similarly $B_{ij} \le B_j$, it holds

$$(\overline{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_j)^\top B_{ij}(\overline{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_j) \le 2(\overline{\boldsymbol{\theta}}_i - \boldsymbol{\theta})^\top B_i(\overline{\boldsymbol{\theta}}_i - \boldsymbol{\theta}) + 2(\overline{\boldsymbol{\theta}}_j - \boldsymbol{\theta})^\top B_j(\overline{\boldsymbol{\theta}}_j - \boldsymbol{\theta}).$$

and the second assertion follows as well. □

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

Theorem 8.6. *The condition*

$$\Delta := 0.5\sigma^{-2}(\overline{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_j)^\top B_i(B_i + B_j)^{-1}B_j(\overline{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_j) > A\lambda \tag{8.5}$$

implies with $b = (\Delta - A\lambda)/\lambda$

$$\boldsymbol{P}\left(T_{ij}^\circ < A\lambda\right) \le e^{-\frac{b^2\lambda}{4(A+b)}}.$$

Proof. As in the proof of Theorem 8.4, decomposition (8.3) and condition (8.4) imply

$$\boldsymbol{P}\left(T_{ij}^\circ < A\lambda\right) \le \boldsymbol{P}\left(\Delta + \zeta_{ij} < A\lambda\right) \le \boldsymbol{P}\left(-\zeta_{ij} > b\lambda\right) \le e^{-b^2\lambda^2/(4\Delta)}.$$

□

Proof of Theorem 6.4. The propagation condition (6.2) follows similarly to the proof of Theorem 6.3. The only difference is that in the local smooth case we apply Theorems 8.4 and 8.5 instead of Theorem 8.3. Let k be such that $h^{(k)} \le h$ and $X_j \in U_h(X_i)$. We apply Theorem 8.4 to the local models $W_i^{(k)}$ and $W_j^{(k)}$. For this we have to check the condition (8.2) using Lemma 8.5. It holds with $\boldsymbol{\theta} = \boldsymbol{\theta}(x)$ by the assumptions (A4) and (A6) that $(f - \Psi^\top\boldsymbol{\theta})^\top W_j^{(k)}(f - \Psi^\top\boldsymbol{\theta}) \le N_j^{(k)}\delta^2 \le N_j^*\delta^2$ for every $X_j \in U_h(X_i)$. Lemma 8.5 yields

$$\big(\overline{\boldsymbol{\theta}}_i^{(k)} - \overline{\boldsymbol{\theta}}_j^{(k)}\big)^\top B_{ij}^{(k)}\big(\overline{\boldsymbol{\theta}}_i^{(k)} - \overline{\boldsymbol{\theta}}_j^{(k)}\big) \le 2p\delta^2(N_i^* + N_j^*)$$

so that the condition (8.2) is fulfilled by (6.1).

Theorem 8.4 now applies yielding

$$\boldsymbol{P}\left(\min_{j=1,\dots,n} \boldsymbol{s}_{ij}^{(k)} < 1/2\right) \le n^{-1}$$

provided that $\lambda = C\log n$ with a sufficiently large C.

The second assertion of the theorem follows from the next lemma.

Lemma 8.7. *Let the assumptions (A4), (A5) and (A6) hold true for some* h *and* $x = X_i$. *Let also the local model* W_i *be such that* $w_{ij} \ge 0.5\overline{w}_{ij} := K_l(\boldsymbol{l}_{ij})$ *for all* j. *If* $\lambda \ge C\log n$ *for some fixed* C, *then*

$$\boldsymbol{P}\left(|\widehat{f}_i - f_i| > \delta\sqrt{pC_B} + \sigma\sqrt{2C_B\lambda/N_i^*}\right) \le 1/n.$$

Proof. Define $W_i^* = \mathrm{diag}\{w_{i1}^*, \ldots, w_{in}^*\}$, $B_i^* = \Psi W_i^* \Psi^\top$ and $N_i^* = \mathrm{tr} W_i^*$. Then the conditions of the lemma yield $N_i \ge 0.5 N_i^*$ and $B_i \ge 0.5 B_i^*$. Next, by Theorem 8.1

$$\boldsymbol{P}\left((\widehat{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_i)^\top B_i (\widehat{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_i) \ge \lambda\sigma^2\right) \le 1/n$$

for $\lambda \le C \log n$ with a sufficiently large C. This implies by (A6) with a high probability

$$(\widehat{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_i)^\top B_i^* (\widehat{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_i) \le 2\lambda\sigma^2.$$

In view of (A6) this gives

$$N_i^* (\widehat{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_i)^\top \Psi_i \Psi_i^\top (\widehat{\boldsymbol{\theta}}_i - \overline{\boldsymbol{\theta}}_i) \le 2 C_B \lambda \sigma^2$$

or equivalently

$$|\widehat{f}_i - \overline{f}_i| \le \sigma \sqrt{2 C_B \lambda / N_i^*}$$

where $\overline{f}_i = \Psi_i^\top \overline{\boldsymbol{\theta}}_i$. Next, Lemma 8.5 and (A4) imply

$$(\overline{\boldsymbol{\theta}}_i - \boldsymbol{\theta})^\top B_i^* (\overline{\boldsymbol{\theta}}_i - \boldsymbol{\theta}) \le p\delta^2 N_i^* .$$

This and (A6) yield using the equality $f_i = \Psi_i^\top \boldsymbol{\theta}$

$$|\overline{f}_i - f_i|^2 = (\overline{\boldsymbol{\theta}}_i - \boldsymbol{\theta})^\top \Psi_i \Psi_i^\top (\overline{\boldsymbol{\theta}}_i - \boldsymbol{\theta}) \le C_B (\overline{\boldsymbol{\theta}}_i - \boldsymbol{\theta})^\top B_i^* (\overline{\boldsymbol{\theta}}_i - \boldsymbol{\theta}) / N_i^* \le p C_B \delta^2$$

and the assertion follows. □

References

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

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

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

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

[6] Gencay, R., F. Selcuk, and B. Whitcher (2001). *An Introduction to Wavelets and Other Filtering Methods in Finance and Economics*, San Diego: Academic Press.

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

[8] Heckman, N. and Ramsay, J.O. (2000). Penalized regression with model-based penalties. *Canadian Journal of Statistics* **28**, 241–258.

[9] 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.

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

[11] Polzehl, J. and Spokoiny, V. (2001). Functional and dynamic magnetic resonance imaging using vector adaptive weights smoothing. *Applied Statistics*, **50**, 485-501.

[12] Polzehl, J. and Spokoiny, V. (2002). Local likelihood modeling by adaptive weights smoothing. Preprint 787. WIAS 2002. `http://www.wias-berlin.de/publications/preprints/787`.

[13] R Development Core Team (2003). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, URL: http://www.R-project.org.

Weierstrass-Institute, Mohrenstr. 39, 10117 Berlin, Germany
E-mail address: `polzehl@@wias-berlin.de`

Weierstrass-Institute, Mohrenstr. 39, 10117 Berlin, Germany
E-mail address: `spokoiny@@wias-berlin.de`