Weierstraß-Institut
für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.

Preprint ISSN 0946 – 8633

Spatially adaptive regression estimation: Propagation-separation approach

Jörg Polzehl [1] [3] and Vladimir Spokoiny [2] [3]

submitted: 23rd December 2004

[1] Weierstrass Institute
for Applied Analysis and Stochastics,
Mohrenstr. 39, 10117
Berlin, Germany
E-Mail: polzehl@wias-berlin.de

[2] Weierstrass Institute
for Applied Analysis and Stochastics,
Mohrenstr. 39, 10117
Berlin, Germany
E-Mail: spokoiny@wias-berlin.de

No. 998

Berlin 2005

[3] Supported by the DFG Research Center Matheon "Mathematics for key technologies" in Berlin

2000 *Mathematics Subject Classification.* 62G05.

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

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

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

Abstract

Polzehl and Spokoiny (2000) introduced the adaptive weights smoothing (AWS) procedure 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 we present a new method based on the ideas of *propagation* and *separation* which extends the AWS procedure to the case of an arbitrary local linear parametric structure. We also establish some important results about properties of the new 'propagation-separation' procedure including rate optimality in the pointwise and global sense. The performance of the procedure is illustrated by examples for local polynomial regression and by applications to artificial and real images.

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. Finally we present a 'separation' result which indicates that an extension of every local model will be automatically restricted to the region of local homogeneity.

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), Carroll, 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 propagation-separation 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 and the description of the new 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. We also apply the method to the problem of image denoising. Another application of the proposed method to business cycle analysis can be found in Polzehl, Stărică and Spokoiny (2004). 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 described by the equation

$$Y_i = f(X_i) + \varepsilon_i \, . \qquad (2.1)$$

Here $f(x)$ is an unknown regression function and 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. Heteroskedastic 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.

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 $w_i(x) = K_{\mathrm{loc}}(\boldsymbol{l}_i)$ with $\boldsymbol{l}_i = |\rho(x, X_i)/h|^2$ where h is a bandwidth, $\rho(x, X_i)$ is the Euclidean distance between x and the design point X_i and K_{loc} is a *location kernel*. *Localization by a window* simply restricts the model to 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 Propagation-separation using adaptive weights

This section describes a new method of locally adaptive estimation, based on the *propagation-separation* idea. The procedure aims to determine from the data for every point X_i the largest possible local neighborhood in which the model function $f(\cdot)$ can be well approximated by a parametric function $f_{\boldsymbol{\theta}}$ from $\mathcal{F}$. The procedure starts for every point X_i from a very small local neighborhood which is then successively increased. A new point X_j will be included in a neighborhood of X_i only if the hypothesis of local homogeneity $\boldsymbol{\theta}(X_i) = \boldsymbol{\theta}(X_j)$ is not rejected, that means, if there is no significant difference in the values of the estimated parameters obtained at the earlier step of the procedure. The two important properties of the procedure are *propagation* (free extension) of every local neighborhood within the region of local homogeneity and *separation* of every two regions with different parameter values.

The formal description of the method is given in terms of *weights*. For the initial step of the procedure, the estimate $\widehat{\boldsymbol{\theta}}_i^{(0)}$ of $\boldsymbol{\theta}_i = \boldsymbol{\theta}(X_i)$ is computed from a smallest local model defined by a bandwidth $h^{(0)}$, that is,

$$\widehat{\boldsymbol{\theta}}_i^{(0)} = \operatorname*{arginf}_{\boldsymbol{\theta}} \sum_{j=1}^{n} \bigl(Y_j - f_{\boldsymbol{\theta}}(X_j)\bigr)^2 w_{ij}^{(0)}$$

with $w_{ij}^{(0)} = K_{\mathrm{loc}}\bigl(\boldsymbol{l}_{ij}^{(0)}\bigr)$ and $\boldsymbol{l}_{ij}^{(0)} = \bigl|\rho(X_i, X_j)/h^{(0)}\bigr|^2$. In other words, the algorithm starts with the usual local polynomial estimate with bandwidth $h^{(0)}$, which is taken very small. If K_{loc} is supported on $[0,1]$, then for every point X_i the weights $w_{ij}^{(0)}$ vanish outside the ball $U_i^{(0)}$ of radius $h^{(0)}$ with 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. Every point X_j from $U_i^{(k)}$ gets a weight $w_{ij}^{(k)}$ which is defined by testing the hypothesis of homogeneity $\boldsymbol{\theta}(X_i) = \boldsymbol{\theta}(X_j)$ using the estimates $\widehat{\boldsymbol{\theta}}^{(k-1)}(X_i)$ and $\widehat{\boldsymbol{\theta}}^{(k-1)}(X_j)$ obtained in the previous iteration. These weights are then used to compute new improved estimates $\widehat{\boldsymbol{\theta}}^{(k)}(X_i)$ due to (2.2).

The main ingredient of the procedure is the way how the *adaptive weights* $w_{ij}^{(k)}$ are computed. PS2000 suggested to just take the normalized difference of the estimates $\widehat{f}^{(k-1)}(X_i)$ and $\widehat{f}^{(k-1)}(X_j)$ at two different points for checking the hypothesis of homogeneity $f(X_i) = f(X_j)$. Here we extend that approach to the more general local linear parametric assumption. This naturally leads to a test of homogeneity for two local models $W_i^{(k-1)} = \operatorname{diag}\{w_{i1}^{(k-1)}, \ldots, w_{in}^{(k-1)}\}$ and $W_j^{(k-1)} = \operatorname{diag}\{w_{j1}^{(k-1)}, \ldots, w_{jn}^{(k-1)}\}$, to specify the weight $w_{ij}^{(k)}$.

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 populations 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

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

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 two populations likelihood-ratio test statistic is of the form

$$
\begin{aligned}
T_{ij}^{\circ} &= \max_{\boldsymbol{\theta}} L(W_i, \boldsymbol{\theta}, \boldsymbol{\theta}') + \max_{\boldsymbol{\theta}} L(W_j, \boldsymbol{\theta}, \boldsymbol{\theta}') - \max_{\boldsymbol{\theta}} L(W_i + W_j, \boldsymbol{\theta}, \boldsymbol{\theta}') \\
&= L(W_i, \widehat{\boldsymbol{\theta}}_i, \boldsymbol{\theta}') + L(W_j, \widehat{\boldsymbol{\theta}}_j, \boldsymbol{\theta}') - L(W_i + W_j, \widehat{\boldsymbol{\theta}}_{ij}, \boldsymbol{\theta}') \qquad (3.1)
\end{aligned}
$$

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 that is obtained by summing the weights from 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

$$T_{ij} = L(W_i, \widehat{\boldsymbol{\theta}}_i) - L(W_i, \widehat{\boldsymbol{\theta}}_j) = (2\sigma^2)^{-1}(\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_j)^{\top} B_i (\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_j). \tag{3.2}$$

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 for asymmetric situations when the 'size' of the model W_i is much larger than that of W_j. We consider the value $\boldsymbol{s}_{ij} = T_{ij}/\lambda$, with λ being a parameter of the procedure, 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 $\boldsymbol{s}_{ij}$.

3.2 Defining weights

Using the previously described methods, we compute for every pair (i, j) the penalties $\boldsymbol{l}_{ij}^{(k)}$ and $\boldsymbol{s}_{ij}^{(k)}$. 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 $w_{ij}^{(k)}$ as a product

$$w_{ij}^{(k)} = K_{\mathrm{loc}}\bigl(\boldsymbol{l}_{ij}^{(k)}\bigr) K_{\mathrm{st}}\bigl(\boldsymbol{s}_{ij}^{(k)}\bigr), \tag{3.3}$$

where $K_{\mathrm{loc}}, K_{\mathrm{st}}$ are two kernel functions, which are nondecreasing on the positive semiaxis and satisfy the condition $K_{\mathrm{loc}}(0) = K_{\mathrm{st}}(0) = 1$.

3.3 Control of stability using a 'memory' step

The adaptive weights $W_i^{(k)} = \{w_{ij}^{(k)}\}$ defined in (3.3) lead to the local likelihood estimate

$$\widetilde{\boldsymbol{\theta}}_i^{(k)} = \operatorname*{argmax}_{\boldsymbol{\theta}} L(W_i^{(k)}, \boldsymbol{\theta}).$$

If the local parametric assumption continues to hold in $U_i^{(k)}$ then this new estimate improves the previous step estimate $\widehat{\boldsymbol{\theta}}_i^{(k-1)}$ because the effective sample size (sum of weights) increases. At the same time, the adaptive weights procedure attempts to prevent from including the points X_j at a model $W_i^{(k)}$ if the assumption of homogeneity $\theta_i = \theta_j$ is violated. This helps to keep the approximation bias small even when the neighborhoods $U_i^{(k)}$ become large. However, in some situations, for instance, when the parameters change slowly with location, it may happen that the estimation error decreases at the first few steps of the procedure and starts to slowly increase from some iteration due to an increasing error of local parametric approximation. To ensure that the quality of estimation will not be lost during iteration, we introduce a kind of 'memory' in the procedure. This basically means that the new estimate $\widetilde{\boldsymbol{\theta}}_i^{(k)}$ is compared with the previous one $\widehat{\boldsymbol{\theta}}_i^{(k-1)}$. If the difference is significant, the new estimate $\widetilde{\boldsymbol{\theta}}_i^{(k)}$ is forced towards the last estimate $\widehat{\boldsymbol{\theta}}_i^{(k-1)}$. The difference between two estimates is again computed by testing the hypothesis of homogeneity for two local models $W_i^{(k)}$ and $W_i^{(k-1)}$ centered at the same point X_i but defined at two consecutive steps of the procedure. Namely, we utilize the weight $\eta_i = K_{\mathrm{me}}(\boldsymbol{m}_i^{(k)})$ with some kernel function K_{me} and

$$\boldsymbol{m}_i^{(k)} = (2\sigma^2\tau)^{-1}\bigl|\overline{D}_i^{(k)}\bigl(\widetilde{\boldsymbol{\theta}}_i^{(k)} - \widehat{\boldsymbol{\theta}}_i^{(k-1)}\bigr)\bigr|^2$$

where the matrix $\bigl(\overline{D}_i^{(k)}\bigr)^2 = \sum_j \Psi_j\Psi_j^\top K_{\mathrm{loc}}\bigl(\boldsymbol{l}_{ij}^{(k)}\bigr)$ corresponds to the nonadaptive kernel weights $\overline{w}_{ij}^{(k)} = K_{\mathrm{loc}}\bigl(\boldsymbol{l}_{ij}^{(k)}\bigr)$ and the bandwidth $h^{(k)}$, and τ is the parameter of the procedure. The estimate $\widehat{\boldsymbol{\theta}}_i^{(k)}$ is then computed as $\widehat{\boldsymbol{\theta}}_i^{(k)} = \eta_i\widetilde{\boldsymbol{\theta}}_i^{(k)} + (1-\eta_i)\widehat{\boldsymbol{\theta}}_i^{(k-1)}$.

3.4 Formal description of the procedure

Important ingredients of the method are the kernels $K_{\mathrm{loc}}, K_{\mathrm{st}}$ and K_{me}, the parameters λ and τ, 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: Select the parameters λ, τ, a, $h^{(0)}$, $h_{\max}$ and kernels $K_{\mathrm{loc}}, K_{\mathrm{st}}$ and K_{me}. For every i define $w_{ij}^{(0)} = K_{\mathrm{loc}}(\boldsymbol{l}_{ij}^{(0)})$ and $\boldsymbol{l}_{ij}^{(0)} = \bigl|\rho(X_i, X_j)/h^{(0)}\bigr|^2$. Compute

$$B_i^{(0)} = \sum_j \Psi_j\Psi_j^\top w_{ij}^{(0)}, \quad Z_i^{(0)} = \sum_j Y_j\Psi_j^\top w_{ij}^{(0)}, \quad \widehat{\boldsymbol{\theta}}_i^{(0)} = \bigl(B_i^{(0)}\bigr)^{-1} Z_i^{(0)}.$$

Set $k = 1$.

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

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

$$\begin{aligned} \boldsymbol{l}_{ij}^{(k)} &= \bigl|\rho(X_i, X_j)/h^{(k)}\bigr|^2, \\ \boldsymbol{s}_{ij}^{(k)} &= (2\widehat{\sigma}^2\lambda)^{-1}\bigl(\widehat{\boldsymbol{\theta}}_i^{(k-1)} - \widehat{\boldsymbol{\theta}}_j^{(k-1)}\bigr)^\top B_i^{(k-1)}\bigl(\widehat{\boldsymbol{\theta}}_i^{(k-1)} - \widehat{\boldsymbol{\theta}}_j^{(k-1)}\bigr), \end{aligned} \tag{3.4}$$

and obtain the weights $w_{ij}^{(k)}$ and $\overline{w}_{ij}^{(k)}$ as

$$w_{ij}^{(k)} = K_{\mathrm{loc}}\bigl(\boldsymbol{l}_{ij}^{(k)}\bigr) K_{\mathrm{st}}\bigl(\boldsymbol{s}_{ij}^{(k)}\bigr), \qquad \overline{w}_{ij}^{(k)} = K_{\mathrm{loc}}\bigl(\boldsymbol{l}_{ij}^{(k)}\bigr) \tag{3.5}$$

Denote by $W_i^{(k)}$ the diagonal matrix with diagonal elements $w_{ij}^{(k)}$.

- **Compute the new estimate:** Compute

$$\begin{aligned} Z_i^{(k)} &= \Psi W_i^{(k)} Y = \sum_j \Psi_j Y_j w_{ij}^{(k)}, \\ \widetilde{B}_i^{(k)} &= \Psi W_i^{(k)} \Psi^\top = \sum_j \Psi_j \Psi_j^\top w_{ij}^{(k)}, \qquad \overline{B}_i^{(k)} = \sum_j \Psi_j \Psi_j^\top \overline{w}_{ij}^{(k)}, \end{aligned} \tag{3.6}$$

and define the estimate $\widetilde{\boldsymbol{\theta}}_i^{(k)}$ of $\boldsymbol{\theta}_i$ by

$$\widetilde{\boldsymbol{\theta}}_i^{(k)} = \bigl(\widetilde{B}_i^{(k)}\bigr)^{-1} Z_i^{(k)}.$$

- **Control ('memory') step:** Compute $\eta_i^{(k)} = K_{\mathrm{me}}(\boldsymbol{m}_i^{(k)})$ with

$$\boldsymbol{m}_i^{(k)} = (2\sigma^2\tau)^{-1} \bigl| \overline{D}_i^{(k)} \bigl(\widetilde{\boldsymbol{\theta}}_i^{(k)} - \widehat{\boldsymbol{\theta}}_i^{(k-1)}\bigr) \bigr|^2$$

where $\overline{D}_i^{(k)} = \bigl(\overline{B}_i^{(k)}\bigr)^{1/2}$. Define

$$\widehat{\boldsymbol{\theta}}_i^{(k)} = \eta_i^{(k)} \widetilde{\boldsymbol{\theta}}_i^{(k)} + (1-\eta_i^{(k)}) \widehat{\boldsymbol{\theta}}_i^{(k-1)}, \qquad B_i^{(k)} = \eta_i^{(k)} \widetilde{B}_i^{(k)} + (1-\eta_i^{(k)}) B_i^{(k-1)}. \tag{3.7}$$

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 of $\boldsymbol{\theta}_i$ as $\widehat{\boldsymbol{\theta}}_i = \widehat{\boldsymbol{\theta}}_i^{(k^*)}$ with k^* denoting the total number of iterations. The function $f(X_i)$ is estimated as $\widehat{f}_i = \Psi_i^\top \widehat{\boldsymbol{\theta}}_i$.

3.5 Choice of parameters

Here we briefly discuss the impact of every parameter of the procedure and indicate how each of them can be selected.

Kernels K_{st}, K_{loc} and K_{me}: The kernels K_{st}, K_{loc} and K_{me} must be nonnegative and non-increasing on the positive semiaxis. We propose to use $K_{\mathrm{st}}(u) = e^{-u} I_{\{u \le 5\}}$. We recommend to apply a localization kernel K_{loc} supported on $[0,1]$ to reduce the computational effort of the method. As a default we employ the triangle kernel $K_{\mathrm{loc}}(u) = (1-u)_+$. We also set $K_{\mathrm{me}} = K_{\mathrm{loc}}$. Our numerical results indicate that similarly to standard local linear (polynomial) regression the particular choice of kernels K_{loc} and K_{me} does not significantly affect the performance of the method.

Initial bandwidth $h^{(0)}$, parameter a and maximal bandwidth $h_{\max}$: We recommend to select a small $h^{(0)}$ such that every initial local neighborhood $U_i^{(0)}$ contains a sufficient number of design points to assure 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 results in 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}. This ensures that the number of iterations k^* is at most logarithmic in the sample size. Our default choice is $a_{grow} = 1.25$.

The maximal bandwidth $h_{\max}$ can be taken very large. However, if the underlying objective function is very complex, the use of a large final bandwidth $h_{\max}$ may result in oversmoothing and artificial segmentation.

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: The most 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 a homogeneous situation. Large values of λ may result in loss of adaptivity of the method (less sensitivity to structural changes). A reasonable way to define the parameter λ for a specific application is based on the condition of free extension, which we refer to as 'propagation condition'. We discuss this choice in the next section.

Parameter τ: The parameter τ scales the penalty $\boldsymbol{m}_i^{(k)}$ computed for two models $W_i^{(k)}$ and $W_i^{(k-1)}$ centered at the same point for consecutive iterations. The parameter can be chosen by the propagation condition after a value of λ is fixed. In the end of the iteration process the strong overlapping of the models $W_i^{(k)}$ and $W_i^{(k-1)}$ causes a high correlation between the estimates $\widetilde{\boldsymbol{\theta}}_i^{(k)}$ and $\widehat{\boldsymbol{\theta}}_i^{(k-1)}$. This suggests to take a large value of τ in the beginning and decrease it with iterations until a lower bound, say τ_0 is reached. This leads to the following proposal: $\tau = \max\{\tau_1 - \tau_2 \log h^{(k)}, \tau_0\}$ for some τ_0, τ_1 and τ_2. To reduce the numerical effort we also fix $\widehat{\boldsymbol{\theta}}_i^{(k^*)} = \widehat{\boldsymbol{\theta}}_i^{(k-1)}$ if $\eta_i^{(k)} = 0$ occurs.

3.6 Choice of parameters λ and τ by the 'propagation condition'

The 'propagation 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 values 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. The parameter τ can be chosen by the same argument.

The default value for λ is expressed as $\lambda = q_\alpha(\chi^2_p)$, that is the α-quantile of the χ^2 distribution with p degrees of freedom, where α depends on the specified linear parametric family. Defaults for the case of local polynomial regression are given in Section 5.

3.7 Computational complexity of the algorithm

Memory requirements: Note that every estimate is defined as $\widehat{\boldsymbol{\theta}}_i^{(k)} = \bigl(B_i^{(k)}\bigr)^{-1} Z_i^{(k)}$ using the matrix $B_i^{(k)}$ and the vector $Z_i^{(k)}$. Similarly, the new weights $w_{ij}^{(k)}$ are computed on the basis of the same statistics $B_i^{(k-1)}$, $Z_i^{(k-1)}$ from the previous step of the procedure. Therefore, the whole structural information is contained in these two basis elements. During the adaptation step, we compute the weights $w_{ij}^{(k)}$ for every i and all $j \in U_i^{(k)}$ only with the aim to compute the new elements $B_i^{(k)}$, $Z_i^{(k)}$. 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 $\mathcal{O}(n^2)$.

Computational costs: Since the localization kernel K_{loc} is supported on $[0,1]$, for every local model $W_i^{(k)}$, all the weights $w_{ij}^{(k)}$ with 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 $w_{ij}^{(k)}$ 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 $w_{ij}^{(k)}$ 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 $\boldsymbol{m}_i^{(k)}$ and the value $w_{ij}^{(k)}$, for every pair (i,j) with $\rho(X_i, X_j) \le h^{(k)}$. 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. The estimation step involves for every point X_i, computing the $d\times d$-matrix $B_i^{(k)} = \Psi W_i^{(k)} \Psi^\top$ and the vector $Z_i^{(k)} = \Psi W_i^{(k)} Y$ which requires of order M_k operations. Computing $\widetilde{\boldsymbol{\theta}}_i^{(k)} = \bigl(B_i^{(k)}\bigr)^{-1} Z_i^{(k)}$ 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, every $B_i^{(k)}$ coincides with $N_i^{(k)} = \sum_j w_{ij}^{(k)}$. The statistical penalty $\boldsymbol{s}_{ij}^{(k)}$ can be written in the form $\boldsymbol{s}_{ij}^{(k)} = N_i^{(k-1)} \bigl| \widehat{\theta}_i^{(k-1)} - \widehat{\theta}_j^{(k-1)} \bigr|^2 / (2\sigma^2\lambda)$. The weights $w_{ij}^{(k)}$ can be computed as $w_{ij}^{(k)} = K_{\mathrm{loc}}(\boldsymbol{l}_{ij}^{(k)}) K_{\mathrm{st}}(\boldsymbol{s}_{ij}^{(k)})$, this essentially coincides with the proposal from PS2000 if an uniform kernel K_{loc} is applied. The memory penalty reads as $\boldsymbol{m}_i^{(k)} = \overline{N}_i^{(k)} \bigl(\widetilde{\theta}_i^{(k)} - \widehat{\theta}_i^{(k-1)} \bigr)^2 / (2\sigma^2\tau)$.

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 $W_i^{(k)}$ has been computed. We then compute the

p-vector $Z_i^{(k)} = \Psi(X_i) W_i^{(k)} Y$ with entries $Z_{i,m}^{(k)}$ of the form

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

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

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

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

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 as $\widehat{\theta}_{ij,m} = (m!)^{-1} d^m \widehat{f}_j(X_i)/dx^m$.

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). 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 $\{\psi_m(x) = x^m, |m| \le 2\}$. 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

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

Table 1: Default parameters used for the PS procedure

	λ				τ_0			
p	0	1	2	3	0	1	2	3
univariate	$q_{\chi^2;0.966,1}$	$q_{\chi^2;0.65,2}$	$q_{\chi^2;0.92,3}$	$q_{\chi^2;0.92,4}$	3	30	400	4000
bivariate	$q_{\chi^2;0.966,1}$	$q_{\chi^2;0.65,3}$	$q_{\chi^2;0.92,6}$	-	1	4	30	-

Here $\sum_{m':|m'|\leq 2-|m|}$ means the sum over the set of all pairs $m' = (l_1', l_2')$ with $m_1' + m_2' \leq 2 - m_1 - m_2$ 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 in univariate and bivariate regression problems. The aim of this study is to illustrate two important features of the procedure: propagation within large homogeneous regions and sensitivity to 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 packages pspline (J. Ramsay and B. Ripley), waveslim (B. Whitcher) and aws (J. Polzehl, revised version).

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 X_i form an equidistant grid on $(0, 1)$. Errors ε_i are i.i.d. Gaussian.

Local linear ($p = 1$), local quadratic ($p = 2$) and local cubic ($p = 3$) estimates are computed for 1000 simulated data sets using our approach with maximal bandwidth $h_{\max} = 0.3$ and defaults, see Table 1, for the other parameters.

For a 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, but the numerical results (not reported here) were always in favor of smoothing splines, see also PS2000.

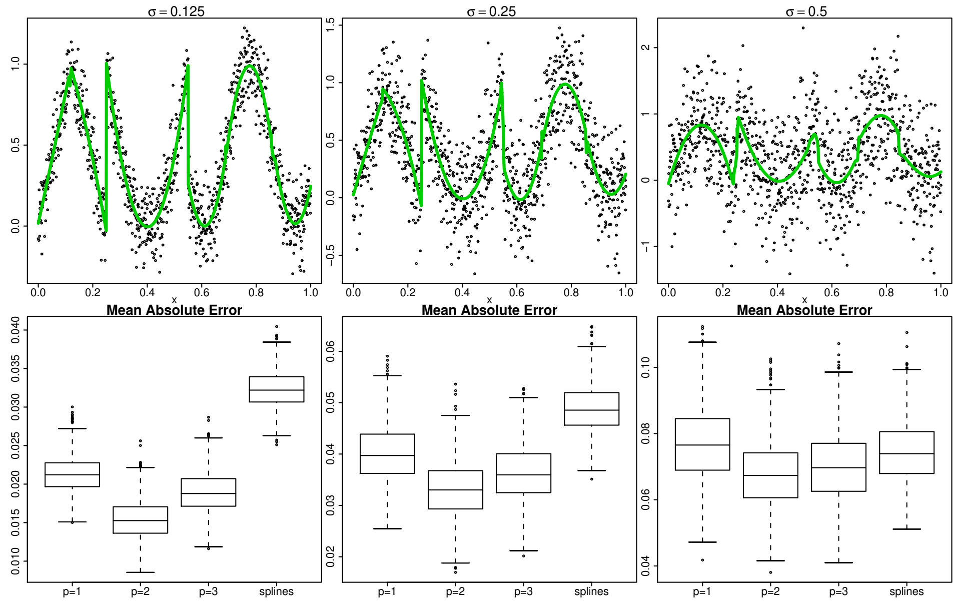

Figure 1: Example 1: Simulated data sets with local quadratic PS estimates. Box-Plots of MAE for local linear, quadratic, cubic PS and penalized cubic smoothing splines.

5.1 Univariate Example 1

Our first example uses the piecewise smooth function

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

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 quadratic PS with default parameters and $h_{\max} = 0.3$. 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, provides pointwise estimates of the MAE in case of $\sigma = 0.125$. The local linear and local quadratic PS estimates are superior to the cubic smoothing spline near the discontinuities and within smooth regions. Advantages are due to the local adaptivity of the PS procedures in contrast to the global nature of the smoothing spline.

5.2 Univariate example 2

The second 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 shows a typical data set for $\sigma = 0.125, 0.25$ and 0.5, respectively, together with the local quadratic PS estimate

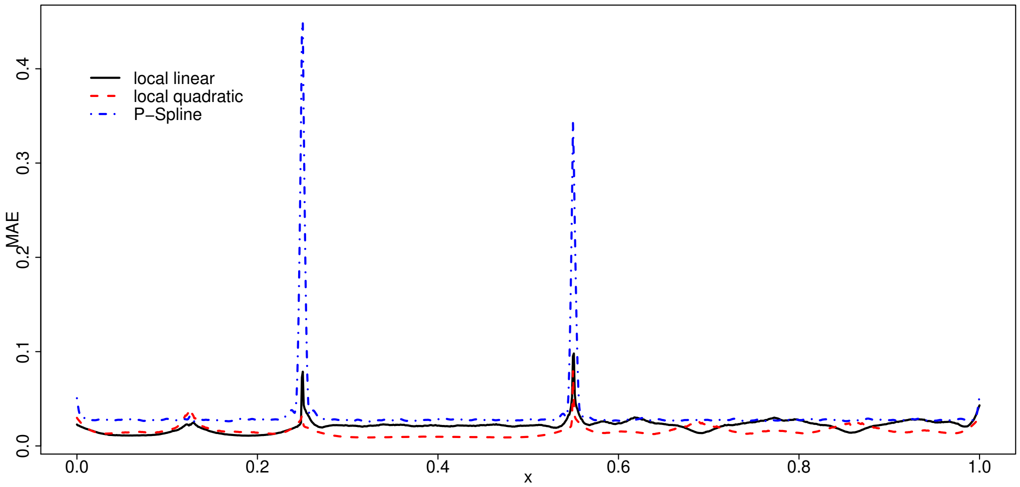

Figure 2: Example 1: Estimated pointwise MAE for local linear and local quadratic PS-estimates and penalized cubic smoothing splines, $\sigma = 0.125$.

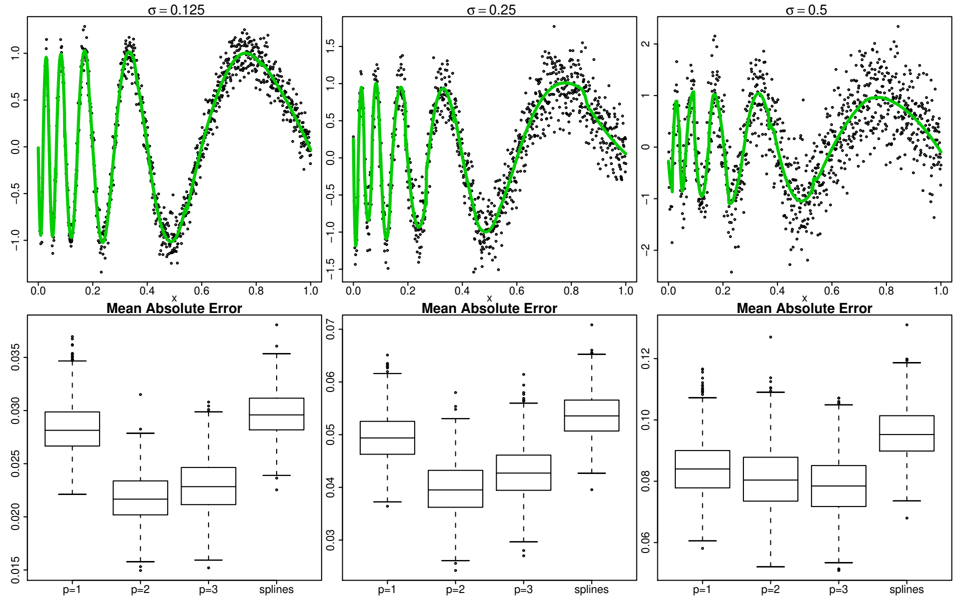

Figure 3: Example 2: Simulated data sets with local quadratic PS estimates. Box-Plots of MAE for local linear, quadratic, cubic PS and penalized cubic smoothing splines.

obtained from this data set using standard parameters and $h_{\max} = 0.3$. The bottom row contains box-plots of MAE obtained for the four procedures in 1000 simulation runs.

5.3 Bivariate Examples

We first use a real image to illustrate the quality of noise reduction achievable by our approach. Figure 4 shows the original image (left), a version of the image with additive

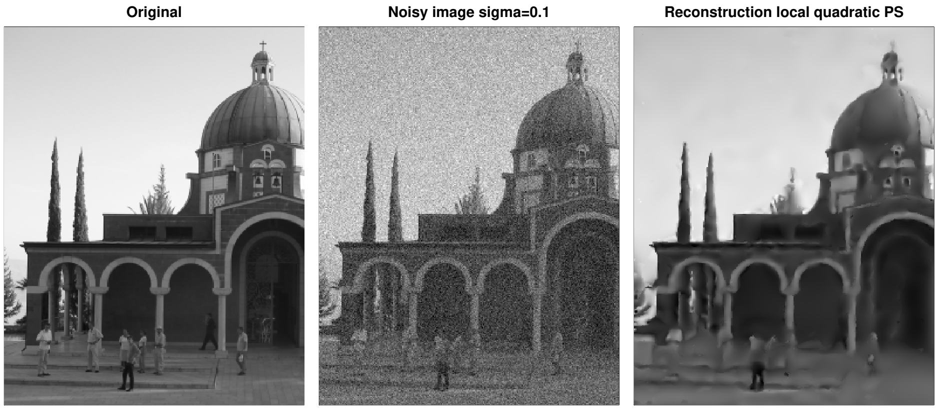

Figure 4: Image 1: Original image (left), Noisy image (center, $\sigma = 0.1$) and local quadratic reconstruction by PS (right, $h_{max} = 25$)

Gaussian noise (center) and the reconstruction of the image obtained by our algorithm (right). The size of the image is 256×330 pixel. Gray values within the original image range from 0.039 to 0.996. Noise standard deviation in the central image is $\sigma = 0.1$. The reconstruction is obtained employing a local quadratic model and using a maximal bandwidth of $h_{\max} = 25$ grid units. All other parameters are set to their defaults. Table 2 (image 1) provides a comparison with some alternative procedures in terms of the MAE of the reconstruction for different noise levels. We present results for our procedure based on a local constant, linear and quadratic assumption, nonadaptive local polynomial regression with degree 0, 1 and 2, 2D discrete wavelet transform (DWT) and 2D maximum overlap discrete wavelet transform (MODWT), see e.g. Gencay, Selcuk and Whitcher (2001). For the latter two we used the *waveslim* package for **R** provided by Brandon Whitcher. Results for local polynomial regression and wavelet procedures are stated for optimized parameters, e.g. bandwidths (in grid units) and basis/depth providing minimal MAE. Parameters used are given in parenthesis.

We use an additional example to demonstrate the potential gain from adaptive local polynomial smoothing. The artificial image is obtained applying the function

$$f(x, y) = 0.5\big[1 + \mathrm{sign}(x^2 - y^2)\big\{\sin(7\phi)\mathbf{1}_{\{r \geq 0.5\}} + \sin(\pi r/2)\mathbf{1}_{\{r<0.5\}}\big\}\big] \tag{5.1}$$

with $r = \sqrt{x^2 + y^2}$ and $\phi = \arcsin(x/r)$ to a grid of size 256×256 on the square $[-1, 1] \times [-1, 1]$. We refer to this image as image 2. Figure 5 shows the original image (upper left), a noisy version with $\sigma = .4$ (upper center), the best nonadaptive local polynomial reconstruction (upper right), the PS reconstructions using a local constant (lower left) and local quadratic (lower center) model together with the best reconstruction using MODWT. Again Table 2 (image 2) provides numerical results in terms of MAE for wide range of noise levels and the alternative procedures with optimized parameters.

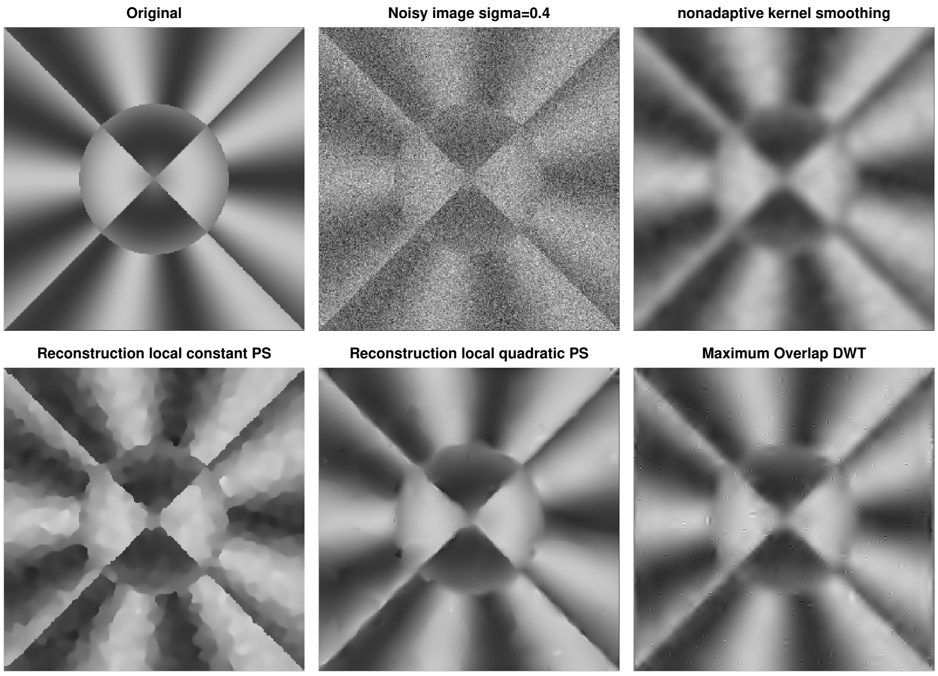

Figure 5: Image 1: Original image (upper left), Noisy image (upper center, $\sigma = 0.4$), best local polynomial (p=0, upper right), local constant PS (lower left), local quadratic PS (lower center) and MODWT reconstruction (lower right)

Table 2: MAE and optimal parameters for reconstructions of image 1 and image 2

Image	σ	PS ($h_{\max}$)			local polynomials (h)			Wavelets (Basis, J)	
No.	(Par.)	$p=0$	$p=1$	$p=2$	$p=0$	$p=1$	$p=2$	DWT	MODWT
1	0.05	0.0152 (8)	0.0139 (20)	0.0133 (25)	0.0218 (2.2)	0.0218 (2.2)	0.0225 (3.4)	0.0230 (Haar, 3)	0.0137 (Haar, 4)
1	0.1	0.0214 (15)	0.0223 (20)	0.0206 (25)	0.0302 (2.7)	0.0303 (2.8)	0.0312 (5.2)	0.0341 (Haar, 3)	0.0220 (Haar, 4)
1	0.2	0.0299 (20)	0.0336 (20)	0.0314 (25)	0.0410 (4.8)	0.0412 (4.4)	0.0423 (8.2)	0.0467 (Fk4, 3)	0.0341 (Haar, 5)
2	0.05	0.0147 (3)	0.0069 (20)	0.0055 (25)	0.0161 (2.8)	0.0161 (2.8)	0.0171 (4.7)	0.0166 (La8, 3)	0.0085 (Mb4, 4)
2	0.1	0.0223 (4.5)	0.0115 (20)	0.0104 (25)	0.0234 (3.9)	0.0234 (4)	0.0246 (7.3)	0.0271 (Mb4, 3)	0.0152 (D4, 4)
2	0.2	0.0323 (6)	0.0208 (20)	0.0185 (25)	0.0335 (5.5)	0.0337 (5.5)	0.0351 (10)	0.0412 (Mb4, 3)	0.0266 (La8, 5)
2	0.4	0.0468 (9)	0.0368 (25)	0.0328 (40)	0.0477 (7.5)	0.0480 (7.6)	0.0487 (14.3)	0.0603 (La8, 3)	0.0439 (La8, 5)
2	0.8	0.0690 (12)	0.0616 (25)	0.0558 (50)	0.0677 (10.4)	0.0683 (10.7)	0.0682 (20.0)	0.0836 (La16, 4)	0.0720 (La8, 6)

Bandwidths are again given in grid units.

The results clearly illustrate the advantages of the PS method compared to local polynomial smoothing if the unknown regression function is piecewise smooth. PS automatically separates regions with different parametric structure and therefore allows for a larger bandwidth within smooth regions, resulting in a larger variance reduction. PS also outperforms wavelet approaches on these examples due to its more flexible handling of boundaries. For image 1 results obtained by local constant and local quadratic PS are comparable with respect to MAE. The local constant approach shows advantages with small detailed structures while the local quadratic PS provides a more acceptable outcome within smooth regions. With the second image best results are obtained for the local quadratic approach, while local constant PS suffers from a segmentation effect caused by its inappropriate structural assumption.

5.4 Summary

The performance of the PS 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.

Local quadratic PS seems to be a reasonable choice for many situations combining good approximating properties with a very good quality of change-point or edge estimation. Local constant PS can be superior in case of very detailed structures or if a local constant assumption is justified.

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

6 Some important properties of the PS estimates

This section discusses some properties of the proposed propagation-separation procedure. In particular we establish the 'propagation' and 'separation' results. 'Propagation' means a free extension of every local model in a homogeneous situation, leading to a nearly parametric estimate at the end of the iteration process. This property and the 'memory' step of the procedure ensure that the resulting estimate is spatially adaptive in the sense of rate optimality over Besov function classes. Finally we show that the procedure separates every two nearly homogeneous regions with significantly different parameter values.

6.1 One step propagation under homogeneity

First we consider the homogeneous case with the constant parameter value $\boldsymbol{\theta}(x) = \boldsymbol{\theta}$ and present some sufficient condition for the 'propagation result'. We proceed by induction. Let the 'propagation' condition be fulfilled for the first k iterations of the algorithm. This means that for every weight $w_{ij}^{(k)}$ its statistical component $K_{\mathrm{st}}(\boldsymbol{s}_{ij}^{(k)})$ is close to one. As a consequence, the k-step estimates $\widehat{\boldsymbol{\theta}}_i^{(k)}$ are close to their non-adaptive counterparts corresponding to the classical local polynomial estimation with the same bandwidth $h^{(k)}$. We now aim to show that the propagation condition continues to hold for the next iteration $k+1$.

Before stating the results we formulate the required assumptions. In our study we restrict ourselves to the case of homogeneous Gaussian errors.

(A1) The errors ε_i are normal with parameters $(0, \sigma^2)$ and the variance σ^2 is known.

This assumption helps to significantly simplify the proofs and to focus on the essential points avoiding technicalities. The procedure does not require a known variance, it is estimated from the data. The theoretical study can be also extended to the case with unknown σ^2, cf. Spokoiny (2002). The case of the non-Gaussian error is more complicated to analyze, however, it also can be considered using the technique from Spokoiny (2001). It is important to mention that the normality of the errors enables us to establish precise nonasymptotic results.

Denote for every i by $U_i^{(k)}$ the ball of radius $h^{(k)}$ with the center at X_i. Let also $\overline{B}_i^{(k)} = \sum_j \Psi_j \Psi_j^\top K_{\mathrm{loc}}(|X_{ij}|^2/|h^{(k)}|^2)$. This matrix arises in the classical local polynomial smoothing with nonadaptive kernel weights corresponding to the bandwidth $h^{(k)}$. Define also $\overline{D}_i^{(k)} = \bigl(\overline{B}_i^{(k)}\bigr)^{1/2}$. We assume that the size of the neighborhoods $U_i^{(k)}$ and the matrices $\overline{D}_i^{(k)}$ grow with k but not too fast. We also assume some local regularity of the design in the neighborhood $U_i^{(k)}$ of every point X_i.

(A2) There exist constants $\nu_1 \le \nu$, $\nu_1, \nu \in (2/3, 1)$ such that for every i

$$\overline{D}_i^{(k-1)} \preceq \nu^{1/2} \overline{D}_i^{(k)}, \qquad \overline{D}_i^{(k)} \preceq \nu_1^{-1/2} \overline{D}_i^{(k-1)}.$$

Here $A \preceq B$ for two symmetric matrices A, B means that $|Av| \le |Bv|$ for every vector v, or equivalently $|v^\top A^2 v| \le |v^\top B^2 v|$.

(A3) There exists a positive constant $\omega^{(k)}$ such that for every i and every $X_j \in U_i^{(k)}$

$$\overline{D}_i^{(k)} \le \omega^{(k)} \overline{D}_j^{(k)}.$$

The conditions A2 and A3 can be easily checked for the equidistant design. They are also fulfilled with a high probability for a random design with a continuous density.

Our theoretical results are stated under one more assumption which helps to gradually simplify the theoretical analysis. The main problem in the theoretical study comes from the iterative nature of the algorithm. At every step we use the same data to compute the estimates $\widehat{\boldsymbol{\theta}}_i^{(k)}$ and the weights $w_{ij}^{(k+1)}$ which will be used to recompute the estimates. As a result, the weights and observations become dependent. To overcome this problem we make the following assumption:

(S0) At step k, the weights $w_{ij}^{(k-1)}$ and $w_{ij}^{(k)}$ are independent of the sample $Y_1, \ldots, Y_n$.

Remark 1. Assumption S0 can be provided using the standard splitting technique, that is, by splitting the original sample into few non overlapping subsamples, cf. Bickel et. al. (1998, pp. 45, 396). However, an application of such a split for practically relevant procedures is questionable. The proposed algorithm utilizes the same sample at every step of the algorithm, and this is not completely unjustified: indeed, it is intuitively clear that the estimates $\widehat{\boldsymbol{\theta}}_i^{(k)}$ obtained by local averaging of the observations are only weakly dependent of the observations Y_j. The same applies to the weights $w_{ij}^{(k)}$ which are defined via the estimates $\widehat{\boldsymbol{\theta}}_i^{(k-1)}$. Our numerical results nicely confirm that the 'propagation' continues to hold even if the same sample is used at every iteration. However, a careful mathematical treatment of this issue might be very complicated.

Under the above conditions and homogeneity of the function $\boldsymbol{\theta}(\cdot)$, we aim to show by induction that the statistical penalties $\boldsymbol{s}_{ij}^{(k)}$ are uniformly bounded by a small constant. This yields that the adaptive weights $w_{ij}^{(k)}$ are close to the nonadaptive kernel weights $\overline{w}_{ij}^{(k)}$ and hence, the estimation results are similar to what we would get for the standard local linear estimation scheme. The results are stated under the additional assumption that the parameters λ, τ of the procedure are taken in the form $\lambda = C_\lambda \log n$ and $\tau = C_\tau \log n$ for some constants C_λ and C_τ depending on the constants from Assumptions A2 and A3.

For the initial estimates $\widehat{\boldsymbol{\theta}}_i^{(k)}$ which are usual local linear estimates with the kernel weights $\overline{w}_{ij}^{(0)}$, Theorem 8.1 implies (see Remark 1) that the values $\bigl|\overline{D}_i^{(0)}\bigl(\widehat{\boldsymbol{\theta}}_i^{(0)} - \boldsymbol{\theta}\bigr)\bigr|$ are with a high probability uniformly bounded by $\sigma\sqrt{C_p \log n}$ with some constant C_p depending on p only. We now assume that after $k-1$ iterations, the following conditions are fulfilled with a high probability for every i

$$D_i^{(k-1)} \succeq \overline{D}_i^{(k-1)}/\sqrt{2}, \quad \bigl|\overline{D}_i^{(k-1)}(\widehat{\boldsymbol{\theta}}_i^{(k-1)} - \boldsymbol{\theta})\bigr| \le \sigma\sqrt{\mu \log n}, \quad \widetilde{D}_i^{(k)} \succeq \nu^{1/2}\overline{D}_i^{(k)}, \tag{6.1}$$

for $\mu = 2C_p$ and ν from Assumption A2. Here $D_i^{(k)} = \bigl(B_i^{(k)}\bigr)^{1/2}$ and similarly $\widetilde{D}_i^{(k)} = \bigl(\widetilde{B}_i^{(k)}\bigr)^{1/2}$, $\overline{D}_i^{(k)} = \bigl(\overline{B}_i^{(k)}\bigr)^{1/2}$, see (3.6) and (3.7). Now we show that the similar result continues to hold for the k th iteration.

Define ρ by $K_{\mathrm{st}}(\rho) = \nu$.

Theorem 6.1. *Suppose that $\boldsymbol{\theta}(\cdot) \equiv \boldsymbol{\theta}$. Let, for the step k of the procedure, Assumptions S0 and A1 through A3 be fulfilled and the parameters λ, τ of the procedure are taken in the form $\lambda = C_\lambda \log n$ and $\tau = C_\tau \log n$ with the constants C_τ and C_λ such that*

$$C_\tau \geq 1.5\mu/(\rho\nu_1), \qquad C_\lambda \geq \mu(1+\omega^{(k)})^2/(2\rho). \tag{6.2}$$

If the condition (6.1) meets, then there exists a random set $\mathcal{A}^{(k)}$ such that $\boldsymbol{P}(\mathcal{A}^{(k)}) \geq 1 - 1/n$, and it holds on $\mathcal{A}^{(k)}$

$$\bigl|\overline{D}_i^{(k)}(\widehat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta})\bigr| \leq \sigma\sqrt{\mu \log n}, \qquad D_i^{(k)} \succeq 2^{-1/2}\overline{D}_i^{(k)}. \tag{6.3}$$

In addition, on $\mathcal{A}^{(k)}$ it holds for every i

$$\min_{X_j \in U_i^{(k)}} K_{\mathrm{st}}(\boldsymbol{s}_{ij}^{(k+1)}) \geq \nu, \quad \widetilde{D}_i^{(k+1)} \succeq \nu^{1/2}\overline{D}_i^{(k+1)}. \tag{6.4}$$

The proof is given in the Appendix. A sequential application of the result of Theorem 6.1 yields the following conclusion for the last step estimate $\widehat{\boldsymbol{\theta}}_i$ under homogeneity:

Corollary 6.2. *Let the conditions of Theorem 6.1 be fulfilled for every iteration k. Then the last step estimate $\widehat{\boldsymbol{\theta}}_i = \widehat{\boldsymbol{\theta}}_i^{(k^*)}$ fulfills*

$$\boldsymbol{P}\Bigl(\max_i \bigl|\overline{D}_i^{(k^*)}(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta})\bigr| > \sigma\sqrt{\mu \log n}\Bigr) \leq k^*/n.$$

6.2 One step propagation under local homogeneity of $\boldsymbol{\theta}(\cdot)$

Here we extend the propagation result to the case when $\boldsymbol{\theta}(\cdot)$ is not constant but can be well approximated by a constant parameter vector in some vicinity of a fixed design point X_i. This would imply a free extension (propagation) of the local model centered at X_i for the first few iterations of the procedure such that the local neighborhoods $U_i^{(k)}$ remain restricted to this region of local homogeneity. Theorem 6.1 claims that under homogeneity the estimate $\widehat{\boldsymbol{\theta}}_i^{(k)}$ of $\boldsymbol{\theta}_i$ satisfies with a high probability the condition $\bigl|\overline{D}_i^{(k)}(\widehat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}_i)\bigr| \leq \sigma\sqrt{\mu \log n}$. We aim to show that if the error of local approximation of the function $\boldsymbol{\theta}(\cdot)$ in the neighborhood $U_i^{(k)}$ of X_i is of the same order, then the result continues to hold.

In the contrary to the previous section where the assertion of Theorem 6.1 applies uniformly to all the points in the design space, we state now a local result in some region $\mathcal{U}^{(k)}$. The reason is that local smoothness properties of $\boldsymbol{\theta}(\cdot)$ and hence the rate of estimation may vary from point to point. The condition we impose on the variability of the function $\boldsymbol{\theta}(\cdot)$ in $\mathcal{U}^{(k)}$ means that $\bigl|\overline{D}_i^{(k)}(\boldsymbol{\theta}_j - \boldsymbol{\theta}_i)\bigr|$ is sufficiently small for all $X_i \in \mathcal{U}^{(k)}$ and $X_j \in U_i^{(k)}$.

(A4) For every $X_i \in \mathcal{U}^{(k)}$ and every $X_j \in U_i^{(k)}$, it holds

$$\sigma^{-1} \bigl| \overline{D}_i^{(k)} (\boldsymbol{\theta}_j - \boldsymbol{\theta}_i) \bigr| \le \delta^{(k)} \sqrt{\log n}.$$

Here $\delta^{(k)}$ is some small constant depending on k and on the region $\mathcal{U}^{(k)}$.

Similarly to the homogeneous case we assume that after $k-1$ iterations, the following conditions are fulfilled with a high probability:

$$D_i^{(k-1)} \succeq 2^{-1/2} \overline{D}_i^{(k-1)}, \quad \bigl| \overline{D}_i^{(k-1)} (\widehat{\boldsymbol{\theta}}_i^{(k-1)} - \boldsymbol{\theta}_i) \bigr| \le \sigma \sqrt{\mu \log n}, \quad \widetilde{D}_i^{(k)} \succeq \nu^{1/2} \overline{D}_i^{(k)}, \tag{6.5}$$

for all $X_i \in \mathcal{U}^{(k)}$. Here ν is from A2 and μ fulfills

$$\sqrt{0.5\mu} \ge \sqrt{C_p} + \delta^{(k)}. \tag{6.6}$$

Theorem 6.3. *Let, for the step k of the procedure, Assumptions S0 and A1 through A4 hold, and let the parameters λ, τ of the procedure fulfill $\lambda = C_\lambda \log n$, $\tau = C_\tau \log n$ with the constants C_τ and C_λ such that*

$$C_\tau \ge 1.5\mu/(\rho\nu_1), \qquad C_\lambda \ge \bigl(\delta^{(k)} + \sqrt{\mu}(1 + \omega^{(k)})\bigr)^2/(2\rho). \tag{6.7}$$

If also μ fulfills (6.6) and (6.5) meets for this μ then there exists a random set $\mathcal{A}^{(k)}$ such that $\boldsymbol{P}(\mathcal{A}^{(k)}) \ge 1 - 1/n$, and it holds on $\mathcal{A}^{(k)}$ for every $X_i \in \mathcal{U}^{(k)}$

$$\bigl| \overline{D}_i^{(k)} (\widehat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}_i) \bigr| \le \sigma \sqrt{\mu \log n}, \qquad D_i^{(k)} \succeq 2^{-1/2} \overline{D}_i^{(k)}. \tag{6.8}$$

Moreover, if X_i is such that $U_i^{(k)} \subset \mathcal{U}^{(k)}$, then on $\mathcal{A}^{(k)}$ it holds

$$\min_{X_j \in U_i^{(k)}} K_{\mathrm{st}}(\boldsymbol{s}_{ij}^{(k+1)}) \ge \nu, \quad \widetilde{D}_i^{(k+1)} \succeq \nu^{1/2} \overline{D}_i^{(k+1)}. \tag{6.9}$$

The proof is given in the Appendix. Here we present one corollary of this result. For a set $\mathcal{U}^{(k)}$ define its $h^{(k)}$-neighborhood $\overline{\mathcal{U}}^{(k)} = \bigcup_{X_i \in \mathcal{U}^{(k)}} U_i^{(k)}$.

Corollary 6.4. *Let, with a fixed k, Assumptions S0 and A1 through A4, (6.6) and (6.7) be fulfilled for every $k' \le k$ with sets $\mathcal{U}^{(k')}$ satisfying $\overline{\mathcal{U}}^{(k'+1)} \subseteq \mathcal{U}^{(k')}$, $k' < k$. Then the k-step estimate $\widehat{\boldsymbol{\theta}}_i^{(k)}$ fulfills*

$$\boldsymbol{P}\Bigl(\max_{X_i \in \mathcal{U}^{(k)}} \bigl| \overline{D}_i^{(k)} (\widehat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}_i) \bigr| > \sigma \sqrt{\mu \log n} \Bigr) \le k/n.$$

Remark 2. The result of Theorem 6.3 and Corollary 6.4 can be reformulated in terms of accuracy of estimation of the function f. Indeed, an estimate $\widehat{\boldsymbol{\theta}}_i^{(k)}$ of $\boldsymbol{\theta}_i = \boldsymbol{\theta}(X_i)$ yields an estimate of the function f at the point X_i in the form $\widehat{f}^{(k)}(X_i) = \Psi_i^\top \widehat{\boldsymbol{\theta}}_i^{(k)}$. In typical situations, the matrix $\overline{B}_i^{(k)} = \sum_j \Psi_j \Psi_j^\top \overline{w}_{ij}^{(k)}$ fulfills the condition $\overline{N}_i^{(k)} \Psi_i^\top \Psi_i \le \varkappa \overline{B}_i^{(k)}$

where $\overline{N}_i^{(k)} = \sum_j \overline{w}_{ij}^{(k)}$ is the 'size' of the local neighborhood $U_i^{(k)}$ and $\varkappa$ is some fixed constant. Therefore

$$\begin{aligned}\overline{N}_i^{(k)} \bigl(\widehat{f}^{(k)}(X_i) - f(X_i)\bigr)^2 &= (\widehat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}_i)^\top \overline{N}_i^{(k)} \Psi_i \Psi_i^\top (\widehat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}_i) \\ &\le \varkappa (\widehat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}_i)^\top \overline{B}_i^{(k)} (\widehat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}_i) = \varkappa \bigl| \overline{D}_i^{(k)} (\widehat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}_i) \bigr|^2\end{aligned}$$

and the result of Corollary 6.4 yields the accuracy of estimation $|\widehat{f}^{(k)}(X_i) - f(X_i)| \le \sigma\bigl(\varkappa\mu \log n / \overline{N}_i^{(k)}\bigr)^{1/2}$ after k steps under propagation. As an interesting special case of Corollary 6.4 consider the situation when the global quality of linear approximation is good in the sense that A4 is fulfilled for all k and $h_{\max}$ is sufficiently large. Then the sizes $\overline{N}_i^{(k^*)}$ of the local neighborhoods at the final step $k = k^*$ are of order of the global sample size n. Therefore, this result claims the root-n consistency of the estimate $\widehat{\boldsymbol{\theta}}_i$.

6.3 Control of stability by the memory step

Due to Theorem 6.3, a small error of the local constant approximation of $\boldsymbol{\theta}(\cdot)$ in a vicinity of a point X_i ensures the propagation condition for the local models $W_i^{(k)}$ and provides with a high probability a certain accuracy of estimation. Now we consider the situation when a local neighborhood $U_i^{(k)}$ extends beyond the region of local homogeneity and A4 is not fulfilled. Of course, the propagation property cannot be stated in this case, and propagation is not desirable when the assumption of local homogeneity is violated. A desirable property of the procedure is that the quality of estimation gained at the 'propagation' phase will not be lost afterwards. This key characteristic is almost a direct consequence of the construction of the 'memory' step. Namely, the following proposition holds.

Proposition 6.5. *For every i and every k, it holds*

$$\bigl|\overline{D}_i^{(k)} \bigl(\widehat{\boldsymbol{\theta}}_i^{(k)} - \widehat{\boldsymbol{\theta}}_i^{(k-1)}\bigr)\bigr| \le \sigma\sqrt{2\tau}. \tag{6.10}$$

Moreover, under A2, it holds for every $k' > k$

$$\bigl|\overline{D}_i^{(k)} \bigl(\widehat{\boldsymbol{\theta}}_i^{(k')} - \widehat{\boldsymbol{\theta}}_i^{(k)}\bigr)\bigr| \le c_1 \sigma\sqrt{2\tau}. \tag{6.11}$$

with $c_1 = \sqrt{\nu}(1 - \sqrt{\nu})^{-1}$.

Remark 3. An interesting feature of this result is that it is fulfilled with probability one, that is, the control of stability 'works' not only with a high probability, it always applies. Assumptions A1 or S0 are not required for this result as well.

Proof. By definition $\widehat{\boldsymbol{\theta}}_i^{(k)} = \eta_i \widetilde{\boldsymbol{\theta}}_i^{(k)} + (1 - \eta_i)\widehat{\boldsymbol{\theta}}_i^{(k-1)}$ with $\eta_i = K_{\mathrm{st}}\bigl(\boldsymbol{m}_i^{(k)}\bigr)$ and $\boldsymbol{m}_i^{(k)} = (2\tau\sigma^2)^{-1}\bigl|\overline{D}_i^{(k)} (\widetilde{\boldsymbol{\theta}}_i^{(k)} - \widehat{\boldsymbol{\theta}}_i^{(k-1)})\bigr|^2$. If $\bigl|\overline{D}_i^{(k)} (\widetilde{\boldsymbol{\theta}}_i^{(k)} - \widehat{\boldsymbol{\theta}}_i^{(k-1)})\bigr| \ge (2\tau\sigma^2)^{1/2}$, then $\eta_i = 0$ and

(6.10) follows automatically. Otherwise

$$\left|\overline{D}_i^{(k)}(\widehat{\boldsymbol{\theta}}_i^{(k)} - \widehat{\boldsymbol{\theta}}_i^{(k-1)})\right| = \eta_i \left|\overline{D}_i^{(k)}(\widetilde{\boldsymbol{\theta}}_i^{(k)} - \widehat{\boldsymbol{\theta}}_i^{(k-1)})\right| \le \sigma\sqrt{2\tau}.$$

Now, Assumption A2 and Proposition 6.5 yield

$$\begin{aligned}
\left|\overline{D}_i^{(k)}(\widehat{\boldsymbol{\theta}}_i^{(k')} - \widehat{\boldsymbol{\theta}}_i^{(k)})\right| &\le \sum_{l=k+1}^{k'} \left|\overline{D}_i^{(k)}(\widehat{\boldsymbol{\theta}}_i^{(l)} - \widehat{\boldsymbol{\theta}}_i^{(l-1)})\right| \le \sum_{l=k+1}^{k'} \nu^{(l-k)/2}\left|\overline{D}_i^{(l)}(\widehat{\boldsymbol{\theta}}_i^{(l)} - \widehat{\boldsymbol{\theta}}_i^{(l-1)})\right| \\
&\le \sigma(1-\sqrt{\nu})^{-1}\sqrt{2\nu\tau}
\end{aligned}$$

which proves (6.11). □

The next theorem states the desirable 'stability' property of the procedure.

Theorem 6.6. *Let A2 hold for all k. If the estimate $\widehat{\boldsymbol{\theta}}_i^{(k)}$ fulfills*

$$\left|\overline{D}_i^{(k)}(\widehat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}_i)\right| \le \sigma\sqrt{\mu\log n} \tag{6.12}$$

for some constant μ, then it holds for the final estimate $\widehat{\boldsymbol{\theta}}_i$

$$\left|\overline{D}_i^{(k)}(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\right| \le c\sigma\sqrt{\log n}$$

with $c = c_1\sqrt{2C_\tau} + \sqrt{\mu}$ and c_1 from Proposition 6.5.

Proof. By Proposition 6.5 $\left|\overline{D}_i^{(k)}(\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_i^{(k)})\right| \le c_1\sigma\sqrt{2\tau} = c_1\sigma\sqrt{2C_\tau\log n}$. Thus

$$\left|\overline{D}_i^{(k)}(\widehat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i)\right| \le \left|\overline{D}_i^{(k)}(\widehat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}_i)\right| + \left|\overline{D}_i^{(k)}(\widehat{\boldsymbol{\theta}}_i - \widehat{\boldsymbol{\theta}}_i^{(k)})\right| \le c_1\sigma\sqrt{2\tau} + \sigma\sqrt{\mu\log n}$$

and the assertion follows. □

6.4 Rate of estimation under smoothness conditions on $f(\cdot)$. Spatial adaptivity

Here we examine the case when $f(\cdot)$ satisfies some smoothness conditions in a neighborhood of a fixed point x. We consider the basis $\{\psi_m\}$ of polynomials of degree less than a given integer number $s \ge 1$ centered at x. In the univariate case $d = 1$ there are exactly $p = s$ basis functions of the form $1, u-x, \ldots, (u-x)^{s-1}$. We also suppose that the design fulfills the property mentioned in Remark 2. We show that under these additional conditions, the results of Theorems 6.3 and 6.6 lead in such a situation to the classical nonparametric rate of estimation of order $(\sigma^2 n^{-1}\log n)^{s/(2s+d)}$.

Let a point $x = X_i$ be fixed. Define $\overline{h}^{(k)} = h^{(1)} + \ldots + h^{(k)}$ for $k \ge 1$ and denote by $\mathcal{B}_i^{(k)}$ the ball with the center at X_i and the radius $\overline{h}^{(k)}$. By definition of $h^{(k)}$, it holds $\overline{h}^{(k)} \le h^{(k)}/(1-a^{-1})$. To ensure the quality of estimation of the function f at the point X_i we assume some smoothness of f and also some design regularity in the neighborhood $\mathcal{B}_i^{(k)}$ for some sufficiently large k.

(A4s) For a fixed k, the function $f(\cdot)$ is $s-1$ times continuously differentiable and the derivative $f^{(s-1)}(u)$ fulfills with some constant L

$$\frac{1}{(s-1)!}\left|f^{(s-1)}(u) - f^{(s-1)}(v)\right| \le Lh^{(k)}, \qquad \forall u, v \in \mathcal{B}_i^{(k)},\ |u-v| \le h^{(k)}.$$

(A5) For a fixed k, it holds for some constants $\nu_2 \le \nu_3$ and $\varkappa$ and all X_j in $\mathcal{B}_i^{(k)}$

$$\overline{N}_i^{(k)} \Psi_i^\top \Psi_i \le \varkappa \overline{B}_i^{(k)}, \qquad \nu_2 \le \frac{\overline{N}_j^{(k)}}{n|h^{(k)}|^d} \le \nu_3.$$

Theorem 6.7. *Define* $\check{h} = \left(L^2\sigma^{-2}n/\log n\right)^{-1/(2s+d)}$ *and fix a constant* $\delta > 0$. *Let* $h^{(k)} = c\check{h}$ *for some iteration number* k *and a sufficiently small constant* c *depending on* a, δ *and* ν_3 *only. Assume that Assumptions A4s and A5 hold for this* k *and, in addition, S0, A1 through A3, (6.6) and (6.7) are satisfied for every* $k' \le k$ *with* $\delta^{(k')} = \delta$. *Then*

$$\boldsymbol{P}\Bigl(\bigl|\widehat{f}_i - f_i\bigr| > C_1 L^{d/(2s+d)} (\sigma^2 n^{-1}\log n)^{s/(2s+d)}\Bigr) \le 4k^*/n \tag{6.13}$$

where C_1 *depends on* c *and the constants in Assumptions A1 through A4s and A5 only.*

The proof is given in the Appendix.

Remark 4. The rate of estimation given in Theorem 6.7 coincides with the optimal rate of estimation for the Sobolev or Hölder smoothness classes up to a log-factor. Moreover, the rate is optimal for the problem of adaptive estimation at a point, cf. Lepski, Mammen and Spokoiny (1997). It was also shown in that paper that this property automatically leads to rate optimality (up to a log-factor) in the Sobolev and Besov function classes $B^s_{p,q}$.

6.5 Separation property

All the results presented earlier discussed the propagation property and its consequences on the quality of estimation. In this section we present one more result which indicates some benefits of using the adaptive weights scheme. Namely we show that the propagation stops when the local parametric approximation does not provide a reasonable accuracy. More precisely, we consider the case when there are two different nearly homogeneous regions, and two points X_{i_1} and X_{i_2}, one from every region, are fixed. We assume that for every of these two points the propagation holds until some step k which leads to the accuracy of estimation $\bigl|\overline{D}_{i_m}^{(k)}(\widehat{\boldsymbol{\theta}}_{i_m}^{(k)} - \boldsymbol{\theta}_{i_m})\bigr| \le \sigma\sqrt{\mu_m \log n}$ for $m = 1, 2$ and some μ_1 and μ_2. We now show that if $\bigl|\overline{D}_i^{(k)}(\boldsymbol{\theta}_{i_1} - \boldsymbol{\theta}_{i_2})\bigr| > C\sigma\sqrt{\log n}$ for a sufficiently large C then the procedure assigns a zero weight $w_{i_1 i_2}^{(k')}$ for all $k' \ge k$.

Theorem 6.8. *Assume A1. Let the statistical kernel* K_{st} *have a compact support on* $[0, A]$ *for some* $A > 0$. *Let, at step* k, *A3 be fulfilled and for two points* X_{i_1} *and* X_{i_2} *hold* $\left|\overline{D}_{i_m}^{(k)}(\widehat{\boldsymbol{\theta}}_{i_m}^{(k)} - \boldsymbol{\theta}_{i_m})\right| \le \sigma\sqrt{\mu_m \log n}$ *with some constants* μ_m *for* $m = 1, 2$. *Let also* $D_{i_1}^{(k)} \succeq b\overline{D}_{i_1}^{(k)}$ *for some* $b > 0$. *If*

$$\left|\overline{D}_{i_1}^{(k)}(\boldsymbol{\theta}_{i_1} - \boldsymbol{\theta}_{i_2})\right| > \sigma\sqrt{\mu_1 \log n} + \sigma\sqrt{\omega^{(k)}\mu_2 \log n} + \sigma\sqrt{Ab^{-1}\lambda}$$

then $w_{i_1 i_2}^{(k+1)} = 0$. *Moreover, there exists a value* Q *depending on* A, b *and the constants from Assumption A2 such that the bounds* $\left|\overline{D}_{i_1}^{(k)}(\boldsymbol{\theta}_{i_1} - \boldsymbol{\theta}_{i_2})\right| > \sigma\sqrt{Q \log n}$ *and* $D_{i_1}^{(k')} \succeq b\overline{D}_{i_1}^{(k)}$ *imply* $w_{i_1 i_2}^{(k')} = 0$ *for every* $k' > k$.

Proof. It suffices to show that $\boldsymbol{s}_{i_1 i_2}^{(k)} = (2\lambda\sigma^2)^{-1}\left|D_i^{(k)}(\widehat{\boldsymbol{\theta}}_{i_1}^{(k)} - \widehat{\boldsymbol{\theta}}_{i_2}^{(k)})\right|^2 > A$. A3 and the inequality $D_{i_1}^{(k)} \succeq b\overline{D}_{i_1}^{(k)}$ yield

$$\begin{aligned}
\left|D_{i_1}^{(k)}(\widehat{\boldsymbol{\theta}}_{i_1}^{(k)} - \widehat{\boldsymbol{\theta}}_{i_2}^{(k)})\right| &\ge \left|D_{i_1}^{(k)}(\boldsymbol{\theta}_{i_1} - \boldsymbol{\theta}_{i_2})\right| - \left|\overline{D}_{i_1}^{(k)}(\widehat{\boldsymbol{\theta}}_{i_1}^{(k)} - \boldsymbol{\theta}_{i_1})\right| - \left|\overline{D}_{i_1}^{(k)}(\widehat{\boldsymbol{\theta}}_{i_2}^{(k)} - \boldsymbol{\theta}_{i_2})\right| \\
&\ge b\left|\overline{D}_{i_1}^{(k)}(\boldsymbol{\theta}_{i_1} - \boldsymbol{\theta}_{i_2})\right| - \sigma\sqrt{\mu_1 \log n} - \sigma\sqrt{\mu_2\omega^{(k)} \log n}
\end{aligned}$$

and the first assertion follows by simple algebra. The second one can be easily shown by involving the result of Proposition 6.5. □

7 Summary and Outlook

The paper presents a new general method of local linear modeling based on the idea of propagation and separation using adaptive weights. The method has a number of remarkable properties. In particular, it applies in a unified way to a broad class of regression models, and the procedure is able to adapt to the unknown and variable structure of the regression function 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 PS 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).

PS applies to models with multidimensional regressors. 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 our method, if the number of local parameters becomes too high (say, more than 6) then the procedure looses sensitivity to structural changes. For such situations, combining the procedure with some dimension reduction methods can be useful.

The proposed 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 the proofs of the main properties claimed in Section 6. First we establish 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 $L(W, \widehat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top B(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})/(2\sigma^2) = (2\sigma^2)^{-1} \bigl| D(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \bigr|^2$ where $B = \Psi W \Psi^\top$ and $D = B^{1/2}$.

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* $z \ge 0$

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

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

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

where $\Sigma = \Psi W^2 \Psi^{\top}$. 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

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

with $R = B^{-1/2}\Sigma B^{-1/2}$. Since $w_i \le 1$, it holds $\Sigma \le B$ and $\|R\| \le 1$, 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. □

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

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

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

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

Now the exponential Tchebychev inequality implies

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

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

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

as required. □

Remark 1. Define z_n by the equality $q_p(z_n) = n^{-2}$, see (8.1). It is easy to see that $p + z_n \le C_p \log n$ where C_p depends on p only. Theorem 8.1 implies for $D = B^{1/2}$

$$\boldsymbol{P}\bigl(2L(W, \widehat{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}) > C_p \log n\bigr) = \boldsymbol{P}\bigl(|D(\widehat{\boldsymbol{\theta}} - \overline{\boldsymbol{\theta}})| > \sigma\sqrt{C_p \log n}\bigr) \le n^{-2}.$$

Proof of Theorem 6.1

Let z_n, C_p be defined by $q_p(z_n) = n^{-2}$ and $p + z_n \le C_p \log n$, see Remark 1. Define

$$\mathcal{A}^{(k)} = \bigl\{\bigl|\widetilde{D}_i^{(k)}(\widetilde{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta})\bigr| \le \sigma\sqrt{C_p \log n},\ \forall i\bigr\}$$

Theorem 8.1 (see Remark 1) yields in the homogeneous situation for every i

$$\boldsymbol{P}\big(\mathcal{A}^{(k)}\big) \geq 1 - \sum_{i=1}^{n} \boldsymbol{P}\Big(\big| \widetilde{D}_i^{(k)} (\widetilde{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}) \big| > \sigma \sqrt{C_p \log n} \Big) \geq 1 - n q_p(z_n) \leq 1 - n^{-1}.$$

We now show that the assertions of the theorem are fulfilled on the set $\mathcal{A}^{(k)}$.

For every i, the memory penalty $\boldsymbol{m}_i^{(k)} = (2\tau\sigma^2)^{-1} \big| \overline{D}_i^{(k)} (\widetilde{\boldsymbol{\theta}}_i^{(k)} - \widehat{\boldsymbol{\theta}}_i^{(k-1)}) \big|^2$ fulfills

$$\begin{aligned}
\sqrt{2\tau\sigma^2 \boldsymbol{m}_i^{(k)}} &\leq \big| \overline{D}_i^{(k)} \big(\widetilde{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}\big) \big| + \big| \overline{D}_i^{(k)} \big(\widehat{\boldsymbol{\theta}}_i^{(k-1)} - \boldsymbol{\theta}\big) \big| \\
&\leq \big| \overline{D}_i^{(k)} \big(\widetilde{D}_i^{(k)}\big)^{-1} \widetilde{D}_i^{(k)} \big(\widetilde{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}\big) \big| + \big| \overline{D}_i^{(k)} \big(\overline{D}_i^{(k-1)}\big)^{-1} \overline{D}_i^{(k-1)} \big(\widehat{\boldsymbol{\theta}}_i^{(k-1)} - \boldsymbol{\theta}\big) \big| \\
&= \big| \overline{D}_i^{(k)} \big(\widetilde{D}_i^{(k)}\big)^{-1} \widetilde{u}_i^{(k)} \big| + \big| \overline{D}_i^{(k)} \big(\overline{D}_i^{(k-1)}\big)^{-1} u_i^{(k-1)} \big|
\end{aligned}$$

where $|\widetilde{u}_i^{(k)}| = \big| \widetilde{D}_i^{(k)} (\widetilde{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}) \big| \leq \sigma \sqrt{0.5\mu \log n}$ on $\mathcal{A}^{(k)}$ and $|u_i^{(k-1)}| = \big| \overline{D}_i^{(k-1)} \big(\widehat{\boldsymbol{\theta}}_i^{(k-1)} - \boldsymbol{\theta}\big) \big| \leq \sigma \sqrt{\mu \log n}$ in view of (6.1). Also by Assumption A2 and (6.1) $\overline{D}_i^{(k-1)} \succeq \nu_1^{1/2} \overline{D}_i^{(k)}$ and $\widetilde{D}_i^{(k)} \succeq \nu^{1/2} \overline{D}_i^{(k)}$. Hence,

$$\sqrt{2\tau \boldsymbol{m}_i^{(k)}} \leq \nu^{-1/2} \sqrt{0.5\mu \log n} + \nu_1^{-1/2} \sqrt{\mu \log n} \leq \sqrt{3\nu_1^{-1} \mu \log n}$$

that yields in view of $\tau = C_\tau \log n \geq 1.5\mu/(\rho\nu_1)$, see (6.2), that $\boldsymbol{m}_i^{(k)} \leq \rho$ and $\eta_i = K_{\mathrm{st}}(\boldsymbol{m}_i^{(k)}) \geq \nu$. It then follows by Assumption A2 and (6.1) for every vector v

$$\begin{aligned}
v^\top B_i^{(k)} v &= \eta_i v^\top \widetilde{B}_i^{(k)} v + (1-\eta_i) v^\top B_i^{(k-1)} v \\
&\geq \eta_i \nu \, v^\top \overline{B}_i^{(k)} v + (1-\eta_i) 0.5 \, v^\top \overline{B}_i^{(k-1)} v \\
&\geq \big(\eta_i \nu + (1-\eta_i)\nu_1/2\big) v^\top \overline{B}_i^{(k)} v \geq 0.5 v^\top \overline{B}_i^{(k)} v
\end{aligned}$$

because of $\eta_i, \nu, \nu_1 \geq 2/3$. Hence, $D_i^{(k)} \succeq 2^{-1/2} \overline{D}_i^{(k)}$.

Further, by definition of $\widehat{\boldsymbol{\theta}}_i^{(k)}$

$$\widehat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta} = \eta_i \big(\widetilde{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}\big) + (1-\eta_i) \big(\widehat{\boldsymbol{\theta}}_i^{(k-1)} - \boldsymbol{\theta}\big).$$

Therefore

$$\begin{aligned}
\big| \overline{D}_i^{(k)} (\widehat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}) \big| &\leq \eta_i \big| \overline{D}_i^{(k)} \big(\widetilde{D}_i^{(k)}\big)^{-1} \widetilde{u}_i^{(k)} \big| + (1-\eta_i) \big| \overline{D}_i^{(k)} \big(\overline{D}_i^{(k-1)}\big)^{-1} u_i^{(k-1)} \big| \\
&\leq \eta_i \sigma \sqrt{0.5\nu^{-1} \mu \log n} + (1-\eta_i) \sigma \sqrt{\nu_1^{-1} \mu \log n} \leq \sigma \sqrt{\mu \log n}
\end{aligned}$$

because of $\eta_i \geq 2/3$, $2/3 \leq \nu_1 \leq \nu \leq 1$. Hence, (6.3) is proved.

By definition $T_{ij}^{(k+1)} = (2\sigma^2)^{-1} \big| D_i^{(k)} (\widehat{\boldsymbol{\theta}}_i^{(k)} - \widehat{\boldsymbol{\theta}}_j^{(k)}) \big|^2$. The definition of $\widetilde{B}_i^{(k)}$ and of $B_i^{(k)}$ clearly implies that $v^\top B_i^{(k)} v \leq v^\top \overline{B}_i^{(k)} v$ for every vector v and therefore, $D_i^{(k)} \preceq \overline{D}_i^{(k)}$. Assumption A3 and (6.3) yield on the set $\mathcal{A}^{(k)}$ for every pair i, j with $X_j \in U_i^{(k)}$

$$\begin{aligned}
\sqrt{2\sigma^2 T_{ij}^{(k+1)}} &\leq \big| \overline{D}_i^{(k)} (\widehat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}) \big| + \big| \overline{D}_i^{(k)} (\widehat{\boldsymbol{\theta}}_j^{(k)} - \boldsymbol{\theta}) \big| \\
&\leq \big| \overline{D}_i^{(k)} (\widehat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}) \big| + \big| \overline{D}_i^{(k)} \big(\overline{D}_j^{(k)}\big)^{-1} \overline{D}_j^{(k)} (\widehat{\boldsymbol{\theta}}_j^{(k)} - \boldsymbol{\theta}) \big| \\
&\leq \sigma \sqrt{\mu \log n} + \omega^{(k)} \sigma \sqrt{\mu \log n} \leq \sigma (1 + \omega^{(k)}) \sqrt{\mu \log n}\,.
\end{aligned}$$

Therefore, on $\mathcal{A}^{(k)}$, it holds for every considered pair i,j

$$\boldsymbol{s}_{ij}^{(k+1)} = \lambda^{-1} T_{ij}^{(k+1)} \le 0.5\mu\bigl(1+\omega^{(k)}\bigr)^2 / C_\lambda \le \rho$$

and $K_{\mathrm{st}}(\boldsymbol{s}_{ij}^{(k+1)}) \ge \nu$. This obviously yields $v^\top \widetilde{B}_i^{(k+1)} v \ge \nu\, v^\top \overline{B}_i^{(k+1)} v$ for any vector v and all i and (6.4) follows.

Proof of Theorem 6.3

The proof follows the line of the proof of Theorem 6.1. We therefore focus only on the specific details. Define

$$\mathcal{A}^{(k)} = \bigl\{ \bigl| \widetilde{D}_i^{(k)} (\widetilde{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{E}\widetilde{\boldsymbol{\theta}}_i^{(k)}) \bigr| \le \sigma \sqrt{C_p \log n},\ \forall i \bigr\}.$$

Here $\boldsymbol{E}\widetilde{\boldsymbol{\theta}}_i^{(k)}$ stands for $\bigl(\widetilde{B}_i^{(k)}\bigr)^{-1} \sum_j w_{ij}^{(k)} \boldsymbol{\theta}_j$ and C_p is defined in Remark 1. Then, similarly to the proof of Theorem 6.1, $\boldsymbol{P}(\mathcal{A}^{(k)}) \ge 1 - n \cdot n^2 = 1 - 1/n$.

Now we check that the assertions of the theorem are satisfied on $\mathcal{A}^{(k)}$. First we bound the estimation error $\widetilde{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}_i$ for $X_i \in \mathcal{U}^{(k)}$. Since $\boldsymbol{E}\widetilde{\boldsymbol{\theta}}_i^{(k)}$ is a convex combination of $\boldsymbol{\theta}_j$ for $X_j \in U_i^{(k)}$, it holds on the set $\mathcal{A}^{(k)}$ by A4

$$\begin{aligned}
\bigl| \widetilde{D}_i^{(k)} \bigl(\widetilde{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}_i\bigr) \bigr| &\le \bigl| \widetilde{D}_i^{(k)} \bigl(\widetilde{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{E}\widetilde{\boldsymbol{\theta}}_i^{(k)}\bigr) \bigr| + \bigl| \overline{D}_i^{(k)} \bigl(\boldsymbol{E}\widetilde{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}_i\bigr) \bigr| \\
&\le \sigma\sqrt{C_p \log n} + \sigma \delta^{(k)} \sqrt{\log n} \le \sigma \sqrt{0.5\mu \log n}
\end{aligned}$$

because of $\sqrt{0.5\mu} \ge \sqrt{C_p} + \delta^{(k)}$, see (6.6). Now (6.8) follows in the same line as (6.3) in the proof of Theorem 6.1.

Now, for every pair i,j with $X_i \in \mathcal{U}^{(k)}$ and $X_j \in U_i^{(k)} \cap \mathcal{U}^{(k)}$, it follows on $\mathcal{A}^{(k)}$ by (6.8) and Assumptions A2, A3, A4

$$\begin{aligned}
\sqrt{2\sigma^2 T_{ij}^{(k+1)}} &\le \bigl| \overline{D}_i^{(k)} (\widehat{\boldsymbol{\theta}}_i^{(k)} - \boldsymbol{\theta}_i) \bigr| + \bigl| \overline{D}_i^{(k)} (\widehat{\boldsymbol{\theta}}_j^{(k)} - \boldsymbol{\theta}_j) \bigr| + \bigl| \overline{D}_i^{(k)} (\boldsymbol{\theta}_j - \boldsymbol{\theta}_i) \bigr| \\
&\le \sigma\sqrt{\mu \log n} + \sigma\omega^{(k)} \sqrt{\mu \log n} + \delta^{(k)} \sigma \sqrt{\log n}.
\end{aligned}$$

Thus, on $\mathcal{A}^{(k)}$ holds $\boldsymbol{s}_{ij}^{(k+1)} = \lambda^{-1} T_{ij}^{(k+1)} \le 0.5\bigl(\sqrt{\mu} + \omega^{(k)}\sqrt{\mu} + \delta^{(k)}\bigr)^2 / C_\lambda \le \rho$, see (6.7), and $K_{\mathrm{st}}(\boldsymbol{s}_{ij}^{(k+1)}) \ge \nu$. The last statement follows similarly to the proof of Theorem 6.1.

Proof of Theorem 6.7

To simplify the proof and avoid tedious tensor notation, we consider the univariate case with $d = 1$ and $s = p$. An extension to the multivariate case is straightforward.

Set $c_0 = \bigl(\delta^2 \alpha^{2s-2} / \nu_3\bigr)^{1/(2s+d)}$ with $\alpha = (1 - 1/a)$ and take k such that $h^{(k)}$ is the largest bandwidth that fulfills $h^{(k)} \le c_0 \check{h}$. Denote $c = h^{(k)} / \check{h}$, $\overline{h}^{(k)} = h^{(1)} + \ldots + h^{(k)}$.

Recall that $\mathcal{B}_i^{(k)}$ is defined as the ball with the center at $x = X_i$ and radius $\overline{h}^{(k)}$. Under condition A4s, the function $f(X_j)$ for $X_j \in \mathcal{B}_i^{(k)}$ can be represented as

$$f(X_j) = f(x) + f'(x)(X_j - x) + \ldots + \frac{f^{(s-2)}(x)}{(s-2)!}(X_j - x)^{s-2} + \frac{f^{(s-1)}(\widetilde{X}_j)}{(s-1)!}(X_j - x)^{s-1}$$

where $\widetilde{X_j}$ is some point between x and X_j. The use of the polynomial basis $\psi_m(u) = (u-x)^m$ for $m = 0, \ldots, s-1$ leads to the local parametrization of the function f given by $f(X_j) = \Psi_j^\top \boldsymbol{\theta}_j$ with $\boldsymbol{\theta}_j = \bigl(f(x), f'(x), \ldots, \frac{f^{(s-2)}(x)}{(s-2)!}, \frac{f^{(s-1)}(\widetilde{X}_j)}{(s-1)!}\bigr)^\top$. For any two points $X_j, X_{j'} \in \mathcal{B}_i^{(k)}$, the corresponding parameter vectors $\boldsymbol{\theta}_j$ and $\boldsymbol{\theta}_{j'}$ differ only in the last coordinate. Moreover, the smoothness condition A4s clearly implies $|\theta_{j,s} - \theta_{j',s}| \le L h^{(k)}$ for any two points $X_j, X_{j'} \in \mathcal{B}_i^{(k)}$ with $|X_j - X_{j'}| \le h^{(k)}$. This yields for every $X_l \in \mathcal{B}_i^{(k)}$ that $|\Psi_l^\top(\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j'})| \le L|\overline{h}^{(k)}|^{s-1} h^{(k)}$ and

$$\begin{aligned}
\bigl|\overline{D}_j^{(k)}(\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j'})\bigr|^2 &= (\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j'})^\top \overline{B}_j^{(k)} (\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j'}) \\
&= \sum_l |\Psi_l^\top(\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j'})|^2 \overline{w}_{il}^{(k)} \le L^2 |\overline{h}^{(k)}|^{2s-2} |h^{(k)}|^2 \overline{N}_j^{(k)}.
\end{aligned}$$

The use of $\overline{h}^{(k)} \le \alpha^{-1} h^{(k)}$, $h^{(k)} = c\check{h}$ with $c \le c_0$ and A5 yields

$$\begin{aligned}
\bigl|\overline{D}_j^{(k)}(\boldsymbol{\theta}_j - \boldsymbol{\theta}_{j'})\bigr|^2 &\le L^2 \nu_3 |\overline{h}^{(k)}|^{2s-2} |h^{(k)}|^{2+d} n \\
&\le L^2 \nu_3 c^{2s+d} \alpha^{2-2s} \check{h}^{2s+d} n = \nu_3 c^{2s+d} \alpha^{2-2s} \sigma^2 \log n \le \delta^2 \sigma^2 \log n
\end{aligned}$$

and A4 holds true for the step k with $\delta^{(k)} = \delta$ and $\mathcal{U}^{(k)} = \{X_i\}$. Obviously A4 also holds for all $k' < k$ with the same δ and $\mathcal{U}^{(k')}$ being the ball centered at X_i of radius $h^{(k'+1)} + \ldots + h^{(k)}$. Corollary 6.4 and Remark 2 ensure with a high probability the following accuracy of estimating the function f by $\widehat{f}^{(k)}$ under A4s and A5:

$$\bigl|\widehat{f}^{(k)}(X_i) - f(X_i)\bigr|^2 \le \frac{\varkappa^2 \sigma^2 \mu \log n}{\overline{N}_i^{(k)}} \le \frac{\varkappa^2 \sigma^2 \mu \log n}{\nu_2 n |h^{(k)}|^d} \le C_1 L^{2d/(2s+d)} \bigl(\sigma^2 n^{-1} \log n\bigr)^{2s/(2s+d)}$$

with some fixed constant C_1 depending on c and the other constants from Assumptions A2–A5. By Theorem 6.6, the same rate of estimation holds for the final estimate $\widehat{f}$.

References

[1] Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1998). *Efficient and adaptive estimation for semiparametric models.* Springer-Verlag New York.

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

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

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

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

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

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

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

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

[10] 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.

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

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

[13] Polzehl, J. and Spokoiny, V. (2002). Local likelihood modeling by adaptive weights smoothing. WIAS-Preprint No. 787, 2002.

[14] Polzehl, J., Stărică, C. and Spokoiny, V. (2004). When did the 2001 recession *really* end? WIAS-Preprint No. 934, 2004.

[15] 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.

[16] Spokoiny, V. (2001). Data driven testing the fit of linear models. *Math. Methods of Statistics*, **10**, no. 4, 465–497, 2001.

[17] Spokoiny, V. (2002). Variance estimation for high-dimensional regression models. *J. of Multivariate Analysis*, **82**(2002), 111–133.