Propagation-Separation Approach for Local Likelihood Estimation

Polzehl, Jörg

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

Spokoiny, Vladimir

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

Abstract

The paper presents a unified approach to local likelihood estimation for a broad class of nonparametric models, including e.g. the regression, density, Poisson and binary response model. The method extends the adaptive weights smoothing (AWS) procedure introduced in Polzehl and Spokoiny (2000) in context of image denoising. The main idea of the method is to describe a greatest possible local neighborhood of every design point X_i in which the local parametric assumption is justified by the data. The method is especially powerful for model functions having large homogeneous regions and sharp discontinuities. The performance of the proposed procedure is illustrated by numerical examples for density estimation and classification. We also establish some remarkable theoretical nonasymptotic results on properties of the new algorithm. This includes the "propagation" property which particularly yields the root-n consistency of the resulting estimate in the homogeneous case. We also state an "oracle" result which implies rate optimality of the estimate under usual smoothness conditions and a "separation" result which explains the sensitivity of the method to structural changes.

Keywords: adaptive weights, local likelihood, exponential family, propagation, separation, density estimation, classification

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

1 Introduction

Local modeling is one of the most useful nonparametric methods. We refer to the book by Fan and Gijbels (1996) for a rigorous discussion of local linear and local polynomial estimation for regression and some other statistical models and many other references. An extension to the local likelihood approach is discussed in Tibshirani and Hastie (1987), Staniswalis (1989), Loader (1996), Fan, Farmen and Gijbels (1998) among others.

This paper offers a new approach to local likelihood modeling which is based on the idea of structural adaptation and extends the *Adaptive Weights Smoothing* (AWS) procedure from Polzehl and Spokoiny (2000) (referred to as PS2000). The main idea of AWS is to describe in a data-driven way a maximal local neighborhood of every design point X_i in which the local parametric assumption is justified by the data. This is realized using the "propagation" idea: at the beginning of the procedure a very small neighborhood of every point is taken. Then, during iteration every such neighborhood is extended by including new points at which the local parametric assumption is not violated. The precise description is given in term of weights: for every point X_i, the corresponding neighborhood at step k is described by a collection of weights $w_{ij}^{(k)}$ having values between zero and one. Points X_j with significantly positive weights $w_{ij}^{(k)}$ can be treated as included in the local model at X_i at the step k. The weights $w_{ij}^{(k)}$ are computed from the data using the estimated results at the preceding step $k-1$. Note that our way of using data-driven weights (for describing the largest possible region of local homogeneity) makes our approach essentially different from other nonparametric procedures like *sieve empirical likelihood* (see e.g. Fan and Zhang, 2003) or *boosting* (see e.g. Friedman, 2001) also using data-driven weights.

The proposed approach also differs essentially from other adaptive methods of estimation like global or variable bandwidth selection. In particularly, we allow the shape of data-driven local neighborhoods to be different for different points while a global bandwidth selection assumes the same shape of all local neighborhoods, cf. Fan, Farmen and Gijbels (1998) and references therein. However, our weights $w_{ij}^{(k)}$ describing the shape of the local model at the point X_i at kth iteration step depend on the estimation results at the other points, while the other pointwise adaptive methods proceed independently for every point, c.f. Fan and Gijbels (1995) or Lepski, Mammen and Spokoiny (1997).

The original AWS procedure was proposed for the regression model in the context of image denoising. The numerical results from PS2000 demonstrate that the AWS method is very efficient in situations where the underlying regression function allows a piecewise constant approximation with large homogeneous regions. The procedure possesses a number of remarkable properties like preservation of edges and contrasts and nearly optimal noise reduction inside large homogeneous regions. It is dimension free and applies in high dimensional situations. However, the AWS procedure from PS2000 has some limitations. The assumption of the regression model with additive errors considered in PS2000 restricts the domain of its applications. The assumption of a local constant structure might be restrictive when a smooth function is considered. Finally, the iterative nature of the AWS procedure makes its theoretical analysis quite complicated and PS2000 did not provide any result about the quality of the proposed procedure. The aim of this paper is to extend the *propagation-separation* (PS) approach to a broad class of nonparametric models and to study it from a theoretical viewpoint. We explain how the procedure can be applied in a unified way to different models like binary response, inhomogeneous exponential and Poisson etc. having local exponential family structure. We also apply the PS method to problems like density or Poisson intensity estimation and classification. More applications can be found in Polzehl and Spokoiny (2002) (tail index estimation), Polzehl and Spokoiny (2004b) (local constant and GARCH volatility estimation), Polzehl and Spokoiny (2004a) (analysis of macroeconomic time series). We also establish some remarkable theoretical *nonasymptotic* results on properties of the new PS algorithm. This includes the "propagation" result which yields the root-n consistency of the resulting estimate in a special homogeneous case. We also state an "oracle" result which implies rate optimality of the estimate under usual smoothness conditions. The present paper, similarly to PS2000, focuses on the local constant approximation of the underlying model function. An extension of the method to local polynomial estimation in the regression setup is given in Polzehl and Spokoiny (2004c). An extension of both method and theory to a general local likelihood modeling remains a challenging problem.

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

The paper is organized as follows. Section 2 describes the model, presents the main examples, discusses the problem of local modeling and estimation and gives some important deviation bounds for the likelihood. The local likelihood PS procedure is introduced in Section 3. Section 4.1 demonstrates how the PS method can be used to estimate a density function. The classification problem is considered in Section 4.2. Section 5 discusses main properties of the proposed method, among them the "propagation condition" and the rate of estimation of a smoothly varying parameter.

2 Local likelihood estimation

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

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

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

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

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

All these examples are particular cases of the local exponential family model, see Section 2.1 for more details.

Now we present a formal definition of our model. Let $\mathcal{P} = (P_\theta, \theta \in \Theta)$ be a family of probability measures on $\mathcal{Y}$ where Θ is a subset of the real line $I\!R^1$. We assume that this family is dominated by a measure P and denote $p(y,\theta) = dP_\theta/dP(y)$. We suppose that each Y_i is, conditionally on $X_i = x$, distributed with density $p(\cdot,\theta(x))$. The density is parameterized by some unknown function $\theta(x)$ on $\mathcal{X}$ which we aim to estimate.

A global parametric structure simply means that the parameter θ does not depend on the location, that is, the distribution of every "observation" Y_i coincides with P_θ for some $\theta \in \Theta$ and all i. This assumption reduces the original problem to the classical parametric situation and the well developed parametric theory applies here for estimating the underlying parameter θ. In particular, the maximum likelihood estimate $\widetilde{\theta} = \widetilde{\theta}(Y_1,\ldots,Y_n)$ of θ which is defined by maximization of the log-likelihood $L(\theta) = \sum_{i=1}^n \log p(Y_i,\theta)$ is root-n consistent and asymptotically efficient under rather general conditions. However, a global parametric assumption is typically too restrictive. The classical nonparametric approach is based on the idea of localization: for every point x, the parametric assumption is only fulfilled locally in a vicinity of x. This leads to considering a local model concentrated in some neighborhood of the point x.

The most general way to describe a local model is based on weights. Let, for a fixed x, a nonnegative weight $w_i = w_i(x) \le 1$ be assigned to the observations Y_i at X_i, $i = 1,\ldots,n$. When estimating the local parameter $\theta(x)$, every observation Y_i is used with the weight $w_i(x)$. This leads to the local (weighted) maximum likelihood estimate

$$\widetilde{\theta}(x) = \operatorname*{arginf}_{\theta\in\Theta} \sum_{i=1}^n w_i(x) \log p(Y_i,\theta). \tag{2.1}$$

Note that this definition is a special case of a more general local linear (polynomial) likelihood modeling when the underlying function θ is modelled linearly (polynomially) in x, see e.g. Fan, Farmen and Gijbels (1998). However, our approach focuses on the choice of localizing weights in a data-driven way rather than on the method of local approximation of the function θ.

We mention two examples of choosing the weights $w_i(x)$. *Localization by a bandwidth* is defined by weights of the form $w_i(x) = K_{\mathrm{loc}}(\boldsymbol{l}_i)$ with $\boldsymbol{l}_i = |\rho(x,X_i)/h|^2$ where h is a bandwidth, $\rho(x,X_i)$ is the Euclidean distance between x and the design point X_i and K_{loc} is a *location kernel.* This approach is intrinsically based on the assumption that the function θ is smooth leading to its local linear (polynomial) approximation within a ball of some small radius h centered in the point of estimation, see e.g. Tibshirani and Hastie (1987), Hastie and Tibshirani (1993), Fan, Farmen and Gijbels (1998), Carroll et.al. (1998), Cai et.al. (2000). The method has serious problems and has to be substantially extended when functions with inhomogeneous smoothness and especially with discontinuities are considered.

Localization by a window simply restricts the model to a subset (window) $U = U(x)$ of the design space which depends on x, that is, $w_i(x) = \mathbf{1}(X_i \in U(x))$. Observations Y_i with X_i outside the region $U(x)$ are not used when estimating the value $\theta(x)$. This kind of localization arises e.g. in the regression tree approach, in change point estimation Müller (1992) and Spokoiny (1998), in image denoising, Qiu (1998), Polzehl and Spokoiny (2003) among many others.

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

$$L(W(x), \theta) = w_1(x) \log p(Y_1, \theta) + \ldots + w_n(x) \log p(Y_n, \theta).$$

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

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

2.1 Local likelihood estimation for an exponential family model

The examples introduced in Section 2 can be considered as particular cases of local exponential family distributions. The density functions $p(y, \theta) = dP_\theta / dP(y)$ are of the form $p(y, \theta) = p(y) e^{yC(\theta) - B(\theta)}$. Here $C(\theta)$ and $B(\theta)$ are some given nondecreasing functions on Θ and $p(y)$ is some nonnegative function on $\mathcal{Y}$. The presentation simplifies if we assume $p(y)$ strictly positive. This assumption is not restrictive because the factor $p(y)$ cancels in the likelihood ratio. The parameter θ is defined by the equations $\int p(y, \theta) P(dy) = 1$ and $\boldsymbol{E}_\theta Y = \int y p(y, \theta) P(dy) = \theta$.

In our study we suppose that the considered parametric family $\mathcal{P}$ satisfies the following standard regularity condition:

(A1) $\mathcal{P} = (P_\theta, \theta \in \Theta \subseteq I\!R)$ is an exponential family, the parameter set Θ is compact and the functions $B(\theta)$ and $C(\theta)$ are continuously differentiable on Θ.

Under this condition, the functions $B(\theta)$ and $C(\theta)$ are connected by the differential equation $B'(\theta) = \theta C'(\theta)$. The Kullback-Leibler divergence $\mathcal{K}(\theta, \theta') = \boldsymbol{E}_\theta \log\bigl(p(Y, \theta) / p(Y, \theta')\bigr)$

for $\theta, \theta' \in \Theta$ and the Fisher information $I(\theta) := \boldsymbol{E}_\theta |p'_\theta(Y,\theta)/p_\theta(Y,\theta)|^2$ satisfy

$$\mathcal{K}(\theta,\theta') = \theta\bigl(C(\theta) - C(\theta')\bigr) - \bigl(B(\theta) - B(\theta')\bigr), \qquad I(\theta) = C'(\theta).$$

Moreover, the regularity condition A1 implies for some constant $\varkappa$ that

$$|I(\theta')/I(\theta'')|^{1/2} \le \varkappa, \qquad \theta', \theta'' \in \Theta.$$

Next, for a given set of weights $W = \{w_1, \dots, w_n\}$ with $w_i \in [0,1]$, it holds

$$L(W,\theta) = \sum_{i=1}^n w_i \log p(Y_i,\theta) = SC(\theta) - NB(\theta) + R$$

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

$$\widetilde{\theta} = S/N = \sum_{i=1}^n w_i Y_i \Big/ \sum_{i=1}^n w_i .$$

This also implies for any $\theta \in \Theta$ that $L(W,\widetilde{\theta}) = N\bigl\{\widetilde{\theta} C(\widetilde{\theta}) - B(\widetilde{\theta})\bigr\} + R$ and

$$L(W,\widetilde{\theta},\theta) := L(W,\widetilde{\theta}) - L(W,\theta) = N\mathcal{K}(\widetilde{\theta},\theta).$$

Our procedure and the theoretical study heavily relies on a deviation bound for the "fitted log-likelihood" $L(W,\widetilde{\theta},\theta)$. A general result is given in Section 6. Here we present two special cases that are important for our presentation. Our first result considers the parametric situation and can be regarded as a nonasymptotic local version of the Wilks theorem, cf. Fan, Zhang and Zhang (2001).

Theorem 2.1. *Let* $W = \{w_i\}$ *be a local model such that* $\max_i w_i \le 1$. *If* $\theta(\cdot) \equiv \theta$ *then*

$$\boldsymbol{P}\bigl(L(W,\widetilde{\theta},\theta) > z\bigr) = \boldsymbol{P}\bigl(N\mathcal{K}(\widetilde{\theta},\theta) > z\bigr) \le 2e^{-z}, \qquad \forall z > 0.$$

Remark 2.1. Condition A1 ensures that the Kullback-Leibler divergence $\mathcal{K}$ fulfills $\mathcal{K}(\theta',\theta) \le I|\theta' - \theta|^2$ for any point θ' in a neighborhood of θ, where I is the maximum of the Fisher information over this neighborhood. Therefore, the result of Theorem 2.1 guarantees with a high probability that $|\widetilde{\theta} - \theta| \le CN^{-1/2}$. In other words, the value N^{-1} can be used to measure variability of the estimate $\widetilde{\theta}$. Theorem 2.1 can be used for constructing the confidence interval of the parameter θ. Indeed, under homogeneity, the true parameter value θ lies with a high probability in the region $\{\theta' : N\mathcal{K}\bigl(\widetilde{\theta},\theta'\bigr) \le z\}$ for a sufficiently large z.

Theorem 2.1 can be extended to the case when $\theta_i \approx \theta$ for all X_i with positive weights w_i. In this case the "fitted log-likelihood" $L(W, \widetilde{\theta}, \overline{\theta}) = N\mathcal{K}(\widetilde{\theta}, \overline{\theta})$ with $\overline{\theta} := \boldsymbol{E}\widetilde{\theta} = N^{-1}\sum_{i=1}^{n} w_i \theta(X_i)$ can also be bounded with high probability.

Theorem 2.2. *Let $W = \{w_i\}$ be a local model such that $\max_i w_i \le 1$. If the family $\mathcal{P}$ satisfies A1, then there is an $\alpha \ge 0$ depending on $\varkappa$ only such that for every $z > 0$*

$$\boldsymbol{P}\bigl(L(W, \widetilde{\theta}, \overline{\theta}) > z\bigr) = \boldsymbol{P}\bigl(N\mathcal{K}(\widetilde{\theta}, \overline{\theta}) > z\bigr) \le 2e^{-z/(1+\alpha)}.$$

More details and proofs can be found in Section 6.

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 for every point X_i the largest possible local neighborhood in which the model function $\theta(\cdot)$ can be well approximated using a constant parametric value θ. 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 the neighborhood of X_i only if the hypothesis of local homogeneity $\theta(X_i) = \theta(X_j)$ 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. 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 $\widehat{\theta}_i^{(0)}$ of $\theta_i = \theta(X_i)$ is the local MLE computed from a smallest local model defined by the kernel weights $w_{ij}^{(0)} = K_{\mathrm{loc}}\bigl(\boldsymbol{l}_{ij}^{(0)}\bigr)$ with $\boldsymbol{l}_{ij}^{(0)} = \bigl|\rho(X_i, X_j)/h^{(0)}\bigr|^2$ for a small bandwidth $h^{(0)}$, cf. (2.1). 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 $\theta(X_i) = \theta(X_j)$ using the estimates $\widehat{\theta}_i^{(k-1)}$ and $\widehat{\theta}_j^{(k-1)}$ obtained in the previous iteration. These weights are then used to compute new improved estimates $\widehat{\theta}_i^{(k)}$ due to (2.1).

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 $\widehat{f}^{(k-1)}(X_j)$ of the regression function f at two different points for checking the hypothesis of homogeneity $f(X_i) = f(X_j)$. Here we utilize the result of Theorem 2.2. The basic idea is to check that the estimate $\widehat{\theta}_j^{(k-1)}$ belongs to the confidence interval of the estimate $\widehat{\theta}_i^{(k-1)}$. More precisely, we apply the quantity

$$T_{ij}^{(k)} = N_i^{(k-1)} \mathcal{K}\bigl(\widehat{\theta}_i^{(k-1)}, \widehat{\theta}_j^{(k-1)}\bigr) \tag{3.1}$$

as a "test statistic", that is, when computing the new weight $w_{ij}^{(k)}$ at the next iteration step we check for a large value of $T_{ij}^{(k)}$.

3.1 Definition of weights

For every pair (i,j), the weight $w_{ij}^{(k)}$ at the kth iteration is computed on the base of two quantities: a *location penalty* $\boldsymbol{l}_{ij}^{(k)} = |\rho(X_i,X_j)/h^{(k)}|^2$ and a *statistical penalty* $\boldsymbol{s}_{ij}^{(k)} = T_{ij}^{(k)}/\lambda$, see (3.1). Here λ is a parameter of the procedure which can be treated as the critical value for the test statistic $T_{ij}^{(k)}$. It is natural to require that each of these two penalties has an independent influence. This leads to considering the product

$$w_{ij}^{(k)} = K_{\mathrm{loc}}\bigl(\boldsymbol{l}_{ij}^{(k)}\bigr) K_{\mathrm{st}}\bigl(\boldsymbol{s}_{ij}^{(k)}\bigr), \tag{3.2}$$

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

During iteration the location penalty for any fixed two points X_i, X_j will be relaxed because of the growing bandwidth $h^{(k)}$ leading to a free extension under homogeneity. At the same time, the statistical penalty becomes more and more sensitive as the local "sample size" $N_i^{(k)}$ grows leading to separation of regions with different parameter values.

3.2 Control of stability using a "memory" step

The adaptive weights $W_i^{(k)} = \{w_{ij}^{(k)}\}$ defined in (3.2) lead to the local likelihood estimate

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

If the local parametric assumption continues to hold in $U_i^{(k)}$ then this new estimate improves the previous step estimate $\widehat{\theta}_i^{(k-1)}$ 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 into 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 $\widetilde{\theta}_i^{(k)}$ is compared with the previous one $\widehat{\theta}_i^{(k-1)}$. If the difference is significant, the new estimate $\widetilde{\theta}_i^{(k)}$ is forced towards the last estimate $\widehat{\theta}_i^{(k-1)}$, that is, the estimate $\widehat{\theta}_i^{(k)}$ is defined as $\widehat{\theta}_i^{(k)} = \eta_i \widetilde{\theta}_i^{(k)} + (1-\eta_i)\widehat{\theta}_i^{(k-1)}$. The parameter η_i measuring the difference between two estimates is again defined by checking the 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:

$$\eta_i = K_{\mathrm{me}}(\boldsymbol{m}_i^{(k)}), \qquad \boldsymbol{m}_i^{(k)} = \tau^{-1}\overline{N}_i^{(k)}\mathcal{K}\bigl(\widetilde{\theta}_i^{(k)}, \widehat{\theta}_i^{(k-1)}\bigr).$$

Here K_{me} is some kernel, $\overline{N}_i^{(k)} = \sum_j K_{\mathrm{loc}}\bigl(\boldsymbol{l}_{ij}^{(k)}\bigr)$ is the "volume" of the scrolled local neighborhood at the step k. Sometimes it is useful to fix the minimal memory effect given by some value $\eta_0 \in (0,1)$. This leads to the definition $\eta_i = (1-\eta_0)K_{\mathrm{me}}(\boldsymbol{m}_i^{(k)})$. The values η_0, τ are parameters of the procedure.

3.3 The procedure

This section presents a description of the procedure. Important ingredients of the method are: the kernels K_{loc}, K_{st} and K_{me}, the parameters λ, τ and η_0, the initial bandwidth $h^{(0)}$, the factor $a > 1$ and the maximal bandwidth h^*. The choice of the parameters is discussed in Section 3.4. The procedure reads as follows:

1. Initialization: select the parameters λ, τ, η_0 $h^{(0)}$, a and h^*. Define for all i, j $w_{ij}^{(0)} = K_{\mathrm{loc}}(|\rho(X_i, X_j)/h^{(0)}|^2)$. Compute the initial estimates $\widehat{\theta}_i^{(0)} = S_i^{(0)}/N_i^{(0)}$ with $S_i^{(0)} = \sum_j w_{ij}^{(0)} Y_j$ and $N_i^{(0)} = \sum_j w_{ij}^{(0)}$. Set $k = 1$, $h^{(1)} = a h^{(0)}$.

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

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

$$\begin{aligned} \boldsymbol{l}_{ij}^{(k)} &= \Bigl|\rho(X_i, X_j)/h^{(k)}\Bigr|^2, \\ \boldsymbol{s}_{ij}^{(k)} &= \lambda^{-1} T_{ij}^{(k)} = \lambda^{-1} N_i^{(k-1)} \mathcal{K}\bigl(\widehat{\theta}_i^{(k-1)}, \widehat{\theta}_j^{(k-1)}\bigr). \end{aligned} \tag{3.3}$$

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

- **Estimation:** Compute for every i

$$\begin{aligned} \widetilde{\theta}_i^{(k)} &= S_i^{(k)}/\widetilde{N}_i^{(k)} \\ \eta_i &= (1-\eta_0)K_{\mathrm{me}}\bigl(\tau^{-1}\overline{N}_i^{(k)}\mathcal{K}\bigl(\widetilde{\theta}_i^{(k)},\widehat{\theta}_i^{(k-1)}\bigr)\bigr) \end{aligned} \tag{3.4}$$

with $\widetilde{N}_i^{(k)} = \sum_{j=1}^n w_{ij}^{(k)}$, $\overline{N}_i^{(k)} = \sum_{j=1}^n \overline{w}_{ij}^{(k)}$ and $S_i^{(k)} = \sum_{j=1}^n w_{ij}^{(k)} Y_j$.

Define the new local MLE estimate $\widehat{\theta}_i^{(k)}$ of θ_i and the value $N_i^{(k)}$ as:

$$\widehat{\theta}_i^{(k)} = \eta_i \widetilde{\theta}_i^{(k)} + (1-\eta_i)\widehat{\theta}_i^{(k-1)}, \qquad N_i^{(k)} = \eta_i \widetilde{N}_i^{(k)} + (1-\eta_i)N_i^{(k-1)}.$$

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

Remark 3.1. In some cases, e.g. Bernoulli and Poisson distributions, local MLE can obtain a value of zero at the border of the parameter space. This may lead to an infinite Kullback-Leibler distance. To avoid such behavior we replace the Kullback-Leibler distance $\mathcal{K}(\theta,\theta')$ by $\mathcal{K}(\theta,(1-\nu)\theta'+\nu\theta)$ with $\nu = 1/(2\overline{N}_i)$ when estimating in point X_i . The effect of this modification vanishes as $\overline{N}_i$ tends to ∞ .

3.4 Choice of parameters

This section briefly discusses the impact of every parameter of the procedure.

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 $K_{\mathrm{loc}}(u) = K_{\mathrm{me}}(u) = (1-u)_+$. Our numerical results indicate that similarly to standard local polynomial smoothing the particular choice of kernels does not significantly affect the performance of the method.

Initial bandwidth $h^{(0)}$, parameter a and maximal bandwidth h^* : The initial bandwidth $h^{(0)}$ should be reasonably small. We select $h^{(0)}$ such that every ball $U_i^{(0)}$ with center X_i and radius $h^{(0)}$ contains only the design point X_i . The parameter a controls the growth rate of the local neighborhoods for every point X_i . It should be selected to provide that the mean number of points inside a ball $U_i^{(k)}$ with radius $h^{(k)}$ grows exponentially with k with some factor $a_{grow} > 1$. If X_i are from $I\!\!R^d$, then the parameter a can be taken as $a = a_{grow}^{1/d}$. Our default choice is $a_{grow} = 1.25$.

The maximal bandwidth h^* can be taken large so that every ball $U_i^{(k)}$ contains the whole sample for the last iteration k and the location penalty nearly vanishes. However, the parameter h^* can be used to bound the numerical complexity of the procedure. The geometric growth of the parameter h ensures that the total number of iterations is typically bounded by $C\log(n)$ for some fixed constant C.

Parameters λ, τ, η_0: The important parameter of the procedure is λ which scales the statistical penalty $\boldsymbol{s}_{ij}$. A small value of λ leads to overpenalization and therefore unstable performance of the method in a homogeneous situation. A large value of λ may result in a loss of sensitivity to discontinuities. A reasonable way to select the parameter λ for a specific application is based on the condition of free extension, which we also call the "propagation condition". We discuss this choice in the next section.

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 $\widetilde{\theta}_i^{(k)}$ and $\widehat{\theta}_i^{(k-1)}$. This suggests that the value of τ can be decreased during iteration. 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. The parameter η_0 controls the memory effect within the iteration process. Our default choice is $\eta_0 = 0.25$.

3.5 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 selected as the minimal value of λ that, in case of a homogeneous (parametric) model $\theta(x) \equiv \theta$, provides a prescribed probability to obtain the global model at the end of the iteration process. The value can be adjusted by Monte-Carlo simulations. A theoretical justification is given by Theorem 5.1 in Section 5.1, that claims that the choice $\lambda = C\log(n)$ with a sufficiently large C ensures the "propagation" condition whatever the parameter θ is.

Our numerical results indicate that an increase of the sample size does not necessarily require to increase λ. Therefore, we utilize as default the constant value $\lambda = t_\alpha(\chi_1^2)$, that is, the α-quantile of the χ^2 distribution with one degree of freedom that relies on the asymptotic distribution of every test statistic $T_{ij}^{(k)}$. The value α depends on the specified exponential family.

4 Application examples

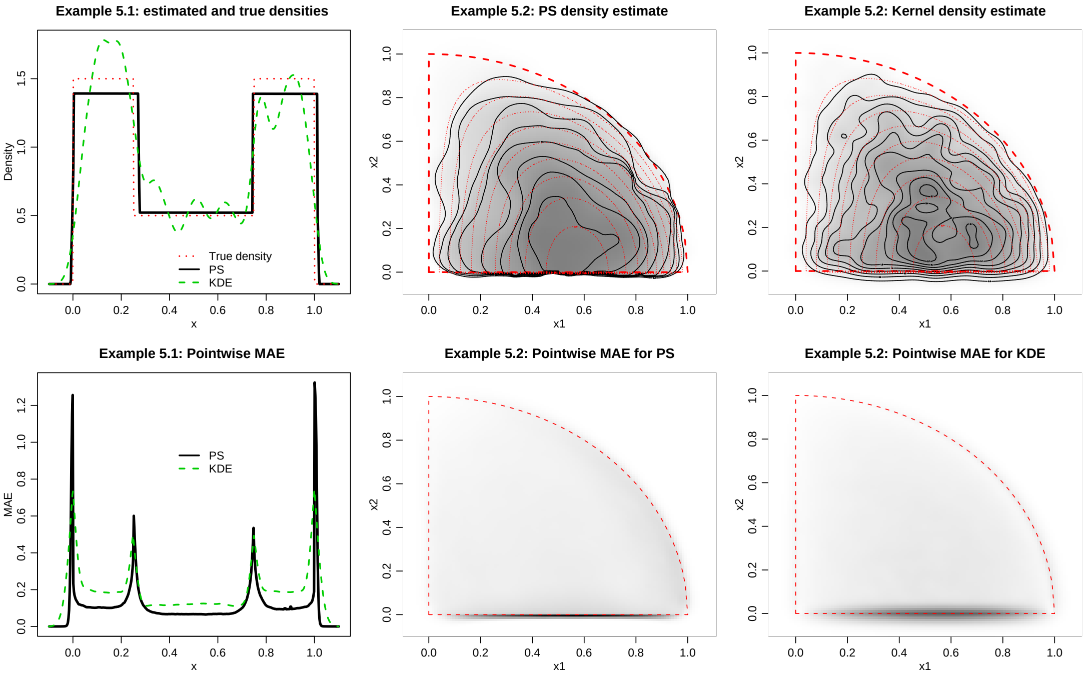
In this section we consider two possible applications of the proposed PS procedure to nonparametric density estimation and classification. Applications to local constant volatility estimation are rigorously discussed in Polzehl and Spokoiny (2004). Polzehl and Spokoiny (2002) explains how a similar procedure can be applied to tail index estimation.

4.1 Application to nonparametric density estimation

Suppose that observations $Z_1, \ldots, Z_L$ are sampled independently from some unknown distribution P on $I\!R^d$ with density $f(x)$. The problem of adaptive estimation of f can be successfully attacked by the PS method. Here we consider a small or moderate d, e.g. $d \le 3$. The case $d > 3$ can be handled as well but requires a different preprocessing.

Without loss of generality we suppose that the observations are located in the cube $[0,1]^d$. We do not assume that f is compactly supported or that f is bounded away from zero on $[0,1]^d$. As a first step we apply a *binning* procedure, see e.g. Fan and Marron (1994) or Fan and Gijbels (1996). Let the interval $[0,1]$ be split into M equal disjoint intervals of length $\delta = 1/M$. Then the cube $[0,1]^d$ can be split into $n = M^d$ nonoverlapping small cubes with side length δ, which we denote by $J_1, \ldots, J_n$. Let X_i be the center point of the cube J_i and let Y_i be the number of observations lying in the ith cube J_i. The pairs (X_i, Y_i) for $i = 1, \ldots, n$ can be viewed as new observations. The joint distribution of $Y_1, \ldots, Y_n$ is described by the multinomial law. This model can be very well approximated by the Poisson model with independent observations Y_i having Poisson distribution with intensity parameter $\theta_i = Lp_i$ where $p_i = P(J_i)$. If the value θ_i has been estimated by $\widehat{\theta}_i$ then the target density f is estimated at X_i as $\widehat{f}(X_i) = \delta^{-d}\widehat{\theta}_i / L$ or as $\widehat{f}(X_i) = \delta^{-d}\widehat{\theta}_i / \sum_{j=1}^n \widehat{\theta}_j$.

For estimating the values θ_i from the “observations” Y_i we apply the PS procedure with the local Poisson family from Example 2.4. In addition to the standard parameter set, we need to specify the bin length δ. A reasonable choice is $\delta = c/K$ where K is the smallest integer satisfying $K^d \ge L$ and $c \le 1$. The procedure applies even if c is small and many bin counts Y_i are zero. Using a small c reduces the discretization error but


Example 5.1: estimated and true densities

Example 5.2: PS density estimate

Example 5.2: Kernel density estimate

Example 5.1: Pointwise MAE

Example 5.2: Pointwise MAE for PS

Example 5.2: Pointwise MAE for KDE

Figure 4.1: Density estimation: Estimates for typical realizations and simulation results.

increases the “sample size” n and therefore, the computational effort by factor c^{-d}.

We use two simulated examples to illustrate the performance of the method. For comparison we also computed kernel density estimates (KDE) with a Gaussian kernel and the bandwidth minimizing the Mean Absolute Error (MAE).

Example 4.1. We generate $n = 200$ observations from the univariate distribution with density $f(x) = 1.5 \cdot I_{\{0 \le x < 0.25\}} + 1.5 \cdot I_{\{0.75 \le x \le 1\}} + 0.5 \cdot I_{\{0.25 \le x < 0.75\}}$.

In the upper left of Figure 1 we provide one typical realization of density estimates using PS (solid line) and KDE (dashed line). The PS-estimate was obtained using a regular grid with interval-length $\delta = 0.0025$ and range $(-.2, 1.2)$. The true density (dotted line) is given for comparison. The maximal bandwidth was chosen $h^* = 5$. The parameters were set to $\alpha = .93$ and $\tau_0 = 1.5$. The lower left plot shows the pointwise MAE for both estimates obtained from 500 simulations. The integrated mean absolute error for PS is 0.141 compared to 0.232 for KDE with optimal bandwidth.

Example 4.2. We generate $n = 2500$ observations from the 2-dimensional density $f(x_1, x_2) = 7.5 \cdot x_1(1 - x_1^2 - x_2^2)_+ I_{\{x_1 \ge 0, x_2 \ge 0\}}$. This density possesses a discontinuity along the axis $x_2 = 0$ and discontinuities of the first derivative along the line $x_1 = 0$ and the boundary of the unit disk. The central upper plot of Figure 1 provides a gray-value image of the estimated density. Contour lines for both the estimate and the true density (dotted) are shown additionally. Results were obtained using a 2-dimensional grid with 140×140 cells on $(-.2, 1.2) \times (-.2, 1.2)$, i.e. with a bin width $\delta = .01$. The maximal bandwidth was set to $h^* = 0.2$. We employed parameters $\alpha = .99$ and $\tau_0 = .5$. The external contour can be interpreted as the estimated support of the density. The quality of the estimate of the density support is very good along the edge $x_2 = 0$ and only slightly worse along the axis $x_1 = 0$ and the boundary of the unit circle where the density is continuous. The central plot in the bottom of Figure 1 provides the pointwise MAE as obtained from 500 simulations. The left row gives the corresponding information for the kernel density estimate (KDE) with optimal (for MAE) bandwidth. We observe an integrated mean absolute error of 0.140 for the PS estimate compared to 0.154 for KDE with optimal bandwidth. The pointwise MSE plot shows that the PS method does very well outside of the density support while the KDE oversmooths near the boundary. As expected, KDE performs slightly better in the region of regularity of the density f.

4.2 Application to classification

Let (X_i, Y_i), $i = 1, \ldots, n$ be a training sample, with X_i valued in an Euclidean space $\mathcal{X} = I\!R^d$ with known class assignment $Y_i \in \{0, 1\}$. Our objective is to construct a discrimination rule assigning every point $x \in \mathcal{X}$ to one of the two classes. The classification problem can be naturally treated in the context of a binary response model. It is assumed that each observation Y_i at X_i is a Bernoulli r.v. with parameter $p(X_i)$, that is, $\boldsymbol{P}(Y_i = 0) = 1 - p(X_i)$ and $\boldsymbol{P}(Y_i = 1) = p(X_i)$. The "ideal" Bayes discrimination rule is $\rho(x) = \mathbf{1}\left(p(x) \geq \pi_1\right)$ where π_1 is the prior probability of the class one. Usually $\pi_1 = 1/2$. Since the function $p(x)$ is typically unknown it is replaced by its estimate $\widehat{p}$.

Nonparametric methods of estimating the function p are based on local averaging. Two typical examples are given by the k-nearest neighbors (k-NN) estimate and the kernel estimate. For a given k, define for every point x in $\mathcal{X}$ the subset $\mathcal{D}_k(x)$ of the design $X_1, \ldots, X_n$ containing the k nearest neighbors of x. Then the k-NN estimate $\widetilde{p}_k(x)$ of $p(x)$ is defined by averaging the observations Y_i over $\mathcal{D}_k(x)$ while the kernel estimate $\widetilde{p}_h(x)$ utilizes the kernel weights $K\left(\rho^2(x, X_i)/h^2\right)$ with a univariate kernel function $K(t)$ and the bandwidth h:

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

Both methods require the choice of a smoothing parameter.

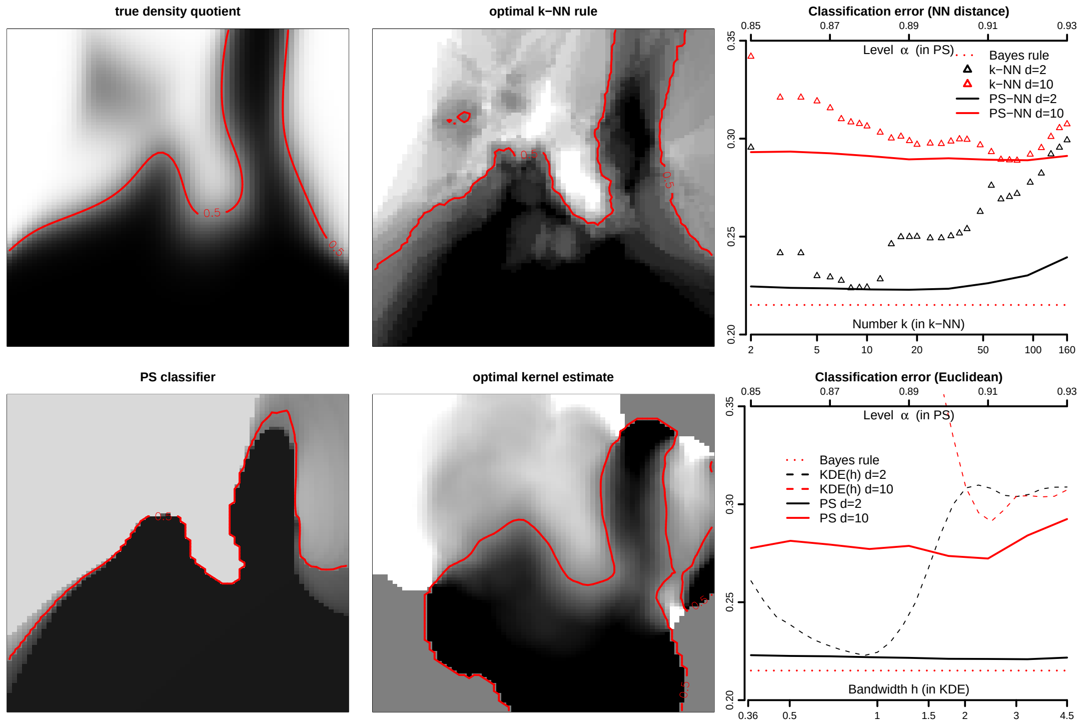


Figure 4.2: Classification rules obtained by the optimal Bayes decision, the best k-nearest neighbor rule, adaptive weights smoothing (PS) and the best rule based on kernel estimation.

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

Example 4.3. To illustrate the behavior of PS in this context we use the data of a simulated two-dimensional discriminant analysis example from Hastie et.al. (2001, p. 13). The data consist of 200 training observations, 100 from each class. The probability densities for each class are mixtures of Gaussians, see Hastie et.al. (2001, p. 17) for details. Figure 2 illustrates the classification rules for the ideal Bayes rule, the k-nearest neighbor rule with optimal $k = 8$, the classification rule obtained by PS with $\alpha = .92$, $\tau_0 = 1$ and $h^* = 10$, and the classification rule obtained by the kernel estimate using an Epanechnikov kernel with optimal bandwidth $h = 0.9$. In each case the estimated, or true, function $p(x)$ are provided together with the 0.5-contour line defining the classification rule.

Additionally a 10-dimensional data set has been created adding 8 i.i.d $U(-1,1)$ nuisance components to each point X_i. The right row of Figure 2 shows graphs of error rates for $d = 2$ and $d = 10$, as functions of the main smoothing parameter for the rules defined by k-nearest neighbor and PS (based on a Nearest Neighbor distance) (top) and kernel estimation and PS (using a Euclidean distance) (bottom). Error rates are obtained by classification of 6831 points in predictor space. Numerical integration with respect to class probabilities and Monte Carlo integration are used in the 2D- and 10D-case, respectively. The ideal Bayes risk is given for a comparison.

The PS procedure produces the lowest classification errors between the three methods. Low values are obtained over a wide range for α, with our default setting, $\alpha = .92$, being slightly conservative. The choice of a smoothing parameter for the other methods is rather critical, with optimal values strongly depending on d.

5 Some important properties of the PS estimate

This section discusses some properties of the proposed PS 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 the "oracle" quality of the resulting estimate and, as a consequence, rate optimality over Besov function classes. Finally we show that the procedure separates every two nearly homogeneous regions with significantly different parameter values. To simplify the exposition, we assume the value of the parameter η_0 describing the minimal memory effect (see (3.4)) is set to zero. Extension to the case of an arbitrary $\eta_0 < 1/2$ is straightforward. We also suppose that the location kernel K_{loc} is supported on $[0,1]$. We start from the "propagation result" for the homogeneous case with the constant parameter value $\theta(x) = \theta$.

5.1 One step propagation under homogeneity

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 and $w_{ij}^{(k)} \approx \overline{w}_{ij}^{(k)}$. 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 the varying coefficient regular exponential family satisfying A1, which is in agreement with all our examples.

Define for every i by $U_i^{(k)}$ the ball of radius $h^{(k)}$ with the center at X_i. Let also $\overline{N}_i^{(k)} = \sum_j \overline{w}_{ij}^{(k)} = \sum_j K_{\mathrm{loc}}(\rho^2(X_i, X_j)/|h^{(k)}|^2)$. This quantity can be treated as the sample size for the local model with nonadaptive kernel weights corresponding to the bandwidth $h^{(k)}$. We assume that the values $\overline{N}_i^{(k)}$ grow with k but not too fast, and also 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

$$\nu_1 \le \overline{N}_i^{(k-1)} / \overline{N}_i^{(k)} \le \nu.$$

(A3) There exists a constant $\omega^{(k)}$ such that for all X_i, X_j with $\rho(X_i, X_j) \le h^{(k)}$

$$(\overline{N}_i^{(k)} / \overline{N}_j^{(k)})^{1/2} \le \omega^{(k)}.$$

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

Remark 5.1. Assumption S0 can be provided using the standard splitting technique, that is, by splitting the original samples into few non overlapping subsamples, cf. Bickel et al (1998, p. 45). 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 $\widehat{\theta}_i^{(k)}$ 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 $\widehat{\theta}_i^{(k-1)}$. Our numerical results nicely confirm that the "propagation" property continues to hold even if the same sample is used at every iteration. It is also worth mentioning that Assumption S0 is only used for proving the "propagation" property.

Under the above conditions and homogeneity of the function $\theta(\cdot)$, 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 likelihood estimation scheme.

The initial estimates $\widehat{\theta}_i^{(0)}$ are obtained by the standard local likelihood method with kernel weights $\overline{w}_{ij}^{(0)}$. By Theorem 2.1 the values $\overline{N}_i^{(0)}\mathcal{K}(\widehat{\theta}_i^{(0)},\theta)$ can with high probability be uniformly bounded by $2\log(n)$. We continue by induction. Assume that after $k-1$ iterations, the following conditions are fulfilled with a high probability for every i

$$N_i^{(k-1)} \ge 0.5\overline{N}_i^{(k-1)}, \qquad \overline{N}_i^{(k-1)}\mathcal{K}(\widehat{\theta}_i^{(k-1)},\theta) \le \mu\log(n), \qquad \widetilde{N}_i^{(k)} \ge \nu\overline{N}_i^{(k)} \tag{5.1}$$

for $\mu \ge 4$. Now we show that the similar result continues to hold for the kth iteration.

Define ρ such that $\min\{K_{\mathrm{me}}(\rho), K_{\mathrm{st}}(\rho)\} = \nu$ where ν is shown in A3.

Theorem 5.1. *Suppose that $\theta(\cdot) \equiv \theta$. 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*

$$C_\tau \ge 3\mu\varkappa^2/(\rho\nu_1), \qquad C_\lambda \ge \varkappa^2\mu(1+\omega^{(k)})^2/\rho \tag{5.2}$$

with $\mu \ge 4$. If the condition (5.1) meets, then there exists a random set $\mathcal{A}^{(k)}$ such that $\boldsymbol{P}(\mathcal{A}^{(k)}) \ge 1 - 2/n$, and it holds on $\mathcal{A}^{(k)}$

$$\overline{N}_i^{(k)}\mathcal{K}(\widehat{\theta}_i^{(k)},\theta) \le \mu\log(n), \qquad N_i^{(k)} \ge \overline{N}_i^{(k)}/2. \tag{5.3}$$

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

$$\min_{X_j \in U_i^{(k)}} K_{\mathrm{st}}(\boldsymbol{s}_{ij}^{(k+1)}) \ge \nu, \quad \widetilde{N}_i^{(k+1)} \ge \nu\overline{N}_i^{(k+1)}. \tag{5.4}$$

Proof. Define $\mathcal{A}^{(k)} = \{\widetilde{N}_i^{(k)}\mathcal{K}\bigl(\widetilde{\theta}_i^{(k)},\theta\bigr) \le 0.5\mu\log(n),\ \forall i\}$. We now apply a general exponential bound from Theorem 2.1: with $z = 0.5\mu\log(n)$ for every i

$$\boldsymbol{P}\left(\widetilde{N}_i^{(k)}\mathcal{K}\bigl(\widetilde{\theta}_i^{(k)},\theta\bigr) > 0.5\mu\log(n)\right) \le 2e^{-0.5\mu\log(n)}.$$

Therefore $\boldsymbol{P}(\mathcal{A}^{(k)}) \ge 1 - 2ne^{-0.5\mu\log(n)} \ge 1 - 2/n$ provided that $\mu \ge 4$. We now show that the assertions of the theorem are fulfilled on the set $\mathcal{A}^{(k)}$.

In the proof we use the following simple lemma.

Lemma 5.2. *Under condition A1 it holds for every* $\theta, \theta', \theta'' \in \Theta$

$$\mathcal{K}^{1/2}(\theta', \theta'') \le \varkappa \mathcal{K}^{1/2}(\theta', \theta) + \varkappa \mathcal{K}^{1/2}(\theta'', \theta).$$

Also, for any sequence $\theta_0, \theta_1, \ldots, \theta_m$ *,*

$$\mathcal{K}^{1/2}(\theta_0, \theta_m) \le \varkappa \sum_{l=1}^{m} \mathcal{K}^{1/2}(\theta_{l-1}, \theta_l).$$

Proof. The reparametrization $\upsilon = C(\theta)$ and $D(\upsilon) = B(\theta)$ is useful, see the proof of Theorem 6.1 in the next section for more details. For any $\upsilon_1 \le \upsilon_2$ it holds

$$\mathcal{K}(\upsilon_1, \upsilon_2) = D(\upsilon_2) - D(\upsilon_1) - (\upsilon_2 - \upsilon_1) D'(\upsilon_1) = 0.5 |\upsilon_2 - \upsilon_1|^2 D''(\widetilde{\upsilon})$$

where $\widetilde{\upsilon} \in [\upsilon_1, \upsilon_2]$ and $D''(\upsilon) = 1/I(\theta)$ and the results easily follow from A1. □

For every i, the memory penalty $\boldsymbol{m}_i^{(k)} = \tau^{-1} \overline{N}_i^{(k)} \mathcal{K}\bigl(\widetilde{\theta}_i^{(k)}, \widehat{\theta}_i^{(k-1)}\bigr)$ fulfills on $\mathcal{A}^{(k)}$ by Lemma 5.2, Assumption A2, (5.1) and (5.2)

$$\begin{aligned}
\boldsymbol{m}_i^{(k)} &\le \tau^{-1} \overline{N}_i^{(k)} \varkappa^2 \bigl\{ \mathcal{K}^{1/2}\bigl(\widehat{\theta}_i^{(k-1)}, \theta\bigr) + \mathcal{K}^{1/2}\bigl(\widetilde{\theta}_i^{(k)}, \theta\bigr) \bigr\}^2 \\
&\le \tau^{-1} \overline{N}_i^{(k)} \varkappa^2 \bigl\{ \bigl(\mu \log(n) / \overline{N}_i^{(k-1)}\bigr)^{1/2} + \bigl(0.5 \mu \log(n) / \widetilde{N}_i^{(k)}\bigr)^{1/2} \bigr\}^2 \\
&\le C_\tau^{-1} \nu_1^{-1} \mu \varkappa^2 \bigl(1 + \sqrt{0.5}\bigr)^2 \le \rho.
\end{aligned}$$

This yields $\eta_i = K_{\mathrm{me}}(\boldsymbol{m}_i^{(k)}) \ge \nu \ge 2/3$. By definition $\widehat{\theta}_i^{(k)} = \eta_i \widetilde{\theta}_i^{(k)} + (1 - \eta_i) \widehat{\theta}_i^{(k-1)}$. Convexity of the Kullback-Leibler divergence $\mathcal{K}(u, v)$ w.r.t. the first argument implies

$$\begin{aligned}
\mathcal{K}(\widehat{\theta}_i^{(k)}, \theta) &= \mathcal{K}(\eta_i \widetilde{\theta}_i^{(k)} + (1 - \eta_i) \widehat{\theta}_i^{(k-1)}, \theta) \\
&\le \eta_i \mathcal{K}(\widetilde{\theta}_i^{(k)}, \theta) + (1 - \eta_i) \mathcal{K}(\widehat{\theta}_i^{(k-1)}, \theta) \\
&\le \bigl(0.5 \eta_i / \widetilde{N}_i^{(k)} + (1 - \eta_i) / \overline{N}_i^{(k-1)}\bigr) \mu \log(n) \\
&\le \bigl(0.5 \eta_i / \nu + (1 - \eta_i) / \nu_1\bigr) \mu \log(n) / \overline{N}_i^{(k)} \le \mu \log(n) / \overline{N}_i^{(k)}
\end{aligned}$$

because of $\eta_i \ge 2/3$ and $\nu \ge \nu_1 \ge 2/3$. Further, these bounds and (5.1) imply

$$\begin{aligned}
N_i^{(k)} &= \eta_i \widetilde{N}_i^{(k)} + (1 - \eta_i) N_i^{(k-1)} \ge \eta_i \nu \overline{N}_i^{(k)} + 0.5 (1 - \eta_i) \overline{N}_i^{(k-1)} \\
&\ge (\eta_i + 0.5(1 - \eta_i)) \nu_1 \overline{N}_i^{(k)} \ge 0.5 \overline{N}_i^{(k)}.
\end{aligned}$$

Hence, (5.3) is proved.

By definition $T_{ij}^{(k+1)} = N_i^{(k)} \mathcal{K}(\widehat{\theta}_i^{(k)}, \widehat{\theta}_j^{(k)})$. Lemma 5.2, Assumptions A1 and A3, (5.3) and the inequality $N_i^{(k)} \le \overline{N}_i^{(k)}$ yield on the set $\mathcal{A}^{(k)}$ for every pair i, j with $X_j \in U_i^{(k)}$

$$\begin{aligned}
T_{ij}^{(k+1)} &\le \varkappa^2 \overline{N}_i^{(k)} \Bigl(\mathcal{K}^{1/2}(\widehat{\theta}_i^{(k)}, \theta) + \mathcal{K}^{1/2}(\widehat{\theta}_j^{(k)}, \theta) \Bigr)^2 \\
&\le \varkappa^2 \mu \log(n) \bigl(1 + \bigl(\overline{N}_i^{(k)} / \overline{N}_j^{(k)}\bigr)^{1/2}\bigr)^2 \le \varkappa^2 \mu \log(n) (1 + \omega^{(k)})^2.
\end{aligned}$$

Therefore, on $\mathcal{A}^{(k)}$, it holds for every considered pair i, j in view of (5.2)

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

and $K_{\mathrm{st}}(\boldsymbol{s}_{ij}^{(k+1)}) \ge \nu$. This also yields $\widetilde{N}_i^{(k+1)} \ge \nu \overline{N}_i^{(k+1)}$ for all i and (5.4) follows. □

A sequential application of the result of Theorem 5.1 yields the following conclusion for the last step estimate $\widehat{\theta}_i$ under homogeneity:

Corollary 5.3. *Let the conditions of Theorem 5.1 be fulfilled for every iteration k with $\mu \ge 4$. Then the last step estimate $\widehat{\theta}_i = \widehat{\theta}_i^{(k^*)}$ fulfills*

$$\boldsymbol{P}\bigl(\overline{N}_i^{(k^*)} \mathcal{K}(\widehat{\theta}_i, \theta) > \mu \log(n)\bigr) \le 2k^*/n.$$

If $h_{\max}$ 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 . Since also $\mathcal{K}(\theta', \theta) \approx I(\theta)|\theta' - \theta|^2/2$, this result claims the root-n consistency of the estimate $\widehat{\theta}_i$.

5.2 One step propagation under local homogeneity of $\theta(\cdot)$

Here we extend the propagation result to the case when $\theta(\cdot)$ is not constant but can be well approximated by a constant 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 provided that the local neighborhoods $U_i^{(k)}$ remain restricted to this region of local homogeneity. Due to Theorem 5.1 the estimate $\widehat{\theta}_i^{(k)}$ under homogeneity of $\theta(\cdot)$ satisfies with a high probability the condition $\overline{N}_i^{(k)} \mathcal{K}(\widehat{\theta}_i^{(k)}, \theta_i) \le \mu \log(n)$ which means the accuracy of estimation of order $\bigl(\log(n)/\overline{N}_i^{(k)}\bigr)^{1/2}$. We aim to show that if the error of local approximation of the function $\theta(\cdot)$ in the neighborhood $U_i^{(k)}$ of X_i is smaller in order than $(\overline{N}_i^{(k)})^{-1/2}$, then the result continues to hold.

In the contrary to the previous section where the assertion of Theorem 5.1 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 $\theta(\cdot)$ and hence the rate of estimation may vary from point to point. For every $X_i \in \mathcal{U}^{(k)}$, we measure the variability of the function θ in a neighborhood $U_i^{(k)}$ by the maximum of $\mathcal{K}^{1/2}(\theta_j, \theta_i)$ over $X_j \in U_i^{(k)}$. Our "local homogeneity" condition means that this variability is not larger in order than the variability of the estimate $\widehat{\theta}_i^{(k)}$.

(A4) For every $X_i \in \mathcal{U}^{(k)}$ and every $X_j \in U_i^{(k)}$, it holds with a constant $\delta_i^{(k)} \ge 0$

$$\mathcal{K}^{1/2}(\theta_j, \theta_i) \le \delta_i^{(k)} \bigl(\log(n)/\overline{N}_i^{(k)}\bigr)^{1/2}.$$

Denote $\overline{\mathcal{U}}^{(k)} = \bigcup_{X_i \in \mathcal{U}^{(k)}} U_i^{(k)}$. Similarly to the homogeneous case we assume that after $k-1$ iterations, the following conditions are fulfilled with a high probability:

$$N_i^{(k-1)} \geq 0.5 \overline{N}_i^{(k-1)}, \quad \overline{N}_i^{(k-1)} \mathcal{K}(\widehat{\theta}_i^{(k-1)}, \theta) \leq \mu \log(n) \quad \widetilde{N}_i^{(k)} \geq \nu \overline{N}_i^{(k)}, \tag{5.5}$$

for all $X_i \in \overline{\mathcal{U}}^{(k)}$. Here ν is from Assumption A3 and μ is some fixed value satisfying

$$\sqrt{0.5\mu} \geq \varkappa \sqrt{2(1+\alpha)/\nu} + \varkappa \delta_i^{(k)} \tag{5.6}$$

with the constant α from Theorem 2.2 which depends on $\varkappa$ only.

Theorem 5.4. *Let, for the step k of the procedure, Assumptions A1 through A4, S0 be fulfilled 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*

$$C_\tau \geq 3\mu\varkappa^2/(\rho\nu_1), \qquad C_\lambda \geq \varkappa\bigl(\delta_i^{(k)} + \sqrt{\mu}(1+\omega^{(k)})\bigr)^2/\rho. \tag{5.7}$$

where μ fulfills (5.6). If (5.5) meets for this μ, then there exists a random set $\mathcal{A}^{(k)}$ such that $\boldsymbol{P}(\mathcal{A}^{(k)}) \geq 1 - 2/n$, and it holds on $\mathcal{A}^{(k)}$ for every $X_i \in \mathcal{U}^{(k)}$

$$\overline{N}_i^{(k)} \mathcal{K}(\widehat{\theta}_i^{(k)}, \theta_i) \leq \mu \log(n), \quad N_i^{(k)} \geq \overline{N}_i^{(k)}/2. \tag{5.8}$$

In addition, on $\mathcal{A}^{(k)}$ for every $X_i \in \mathcal{U}^{(k)}$

$$\min_{X_j \in U_i^{(k)}} K_{\mathrm{st}}(\boldsymbol{s}_{ij}^{(k+1)}) \geq \nu, \quad \widetilde{N}_i^{(k+1)} \geq \nu \overline{N}_i^{(k+1)}. \tag{5.9}$$

Proof. Define

$$\mathcal{A}^{(k)} = \{\widetilde{N}_i^{(k)} \mathcal{K}(\widetilde{\theta}_i^{(k)}, \boldsymbol{E}\widetilde{\theta}_i^{(k)}) \leq 2(1+\alpha)\log(n), \forall X_i \in \overline{\mathcal{U}}^{(k)}\}.$$

Here $\boldsymbol{E}\widetilde{\theta}_i^{(k)}$ stands for $\sum_j w_{ij}^{(k)} \theta_j / \sum_j w_{ij}^{(k)}$ and α is shown in Theorem 2.2. This theorem yields for every i

$$\boldsymbol{P}\left(\widetilde{N}_i^{(k)} \mathcal{K}(\widetilde{\theta}_i^{(k)}, \boldsymbol{E}\widetilde{\theta}_i^{(k)}) > 2(1+\alpha)\log(n)\right) \leq 2e^{-2\log(n)} = 2n^{-2}.$$

Therefore $\boldsymbol{P}(\mathcal{A}^{(k)}) \geq 1 - 2n \cdot n^{-2} \geq 1 - 2/n$.

Now we check that the assertions of the theorem are satisfied on $\mathcal{A}^{(k)}$. First we bound the estimation error of $\widetilde{\theta}_i^{(k)}$ for $X_i \in \mathcal{U}^{(k)}$. Since $\boldsymbol{E}\widetilde{\theta}_i^{(k)}$ is a convex combination of θ_j for $X_j \in U_i^{(k)} \subseteq \overline{\mathcal{U}}^{(k)}$ and $\mathcal{K}(\theta, \theta')$ is a convex function w.r.t. θ, it holds on the set $\mathcal{A}^{(k)}$ by Lemma 5.2 and Assumption A4 in view of (5.6)

$$\begin{aligned}
\overline{N}_i^{(k)} \mathcal{K}(\widetilde{\theta}_i^{(k)}, \theta_i) &\leq \overline{N}_i^{(k)} \varkappa^2 \Bigl(\mathcal{K}^{1/2}(\widetilde{\theta}_i^{(k)}, \boldsymbol{E}\widetilde{\theta}_i^{(k)}) + \mathcal{K}^{1/2}(\boldsymbol{E}\widetilde{\theta}_i^{(k)}, \theta_i)\Bigr)^2 \\
&\leq \overline{N}_i^{(k)} \varkappa^2 \Bigl\{\bigl(2(1+\alpha)\log(n)/\widetilde{N}_i^{(k)}\bigr)^{1/2} + \delta_i^{(k)}\bigl(\log(n)/\overline{N}_i^{(k)}\bigr)^{1/2}\Bigr\}^2 \\
&\leq \varkappa^2 \log(n) \Bigl(\sqrt{2(1+\alpha)/\nu} + \delta_i^{(k)}\Bigr)^2 \leq 0.5\mu\log(n).
\end{aligned}$$

Now (5.8) follows in the same line as (5.3) in the proof of Theorem 5.1.

For every $X_i \in \mathcal{U}^{(k)}$ and $X_j \in U_i^{(k)}$, by Lemma 5.2, (5.7) on $\mathcal{A}^{(k)}$ holds

$$\begin{aligned} T_{ij}^{(k+1)} &\le \varkappa^2 N_i^{(k)} \Bigl(\mathcal{K}^{1/2}(\widehat{\theta}_i^{(k)}, \theta_i) + \mathcal{K}^{1/2}(\widehat{\theta}_j^{(k)}, \theta_j) + \mathcal{K}^{1/2}(\theta_j, \theta_i) \Bigr)^2 \\ &\le \varkappa^2 \overline{N}_i^{(k)} \Bigl\{ \bigl(\mu \log(n)/\overline{N}_i^{(k)}\bigr)^{1/2} + \bigl(\mu \log(n)/\overline{N}_j^{(k)}\bigr)^{1/2} + \delta_i^{(k)} \bigl(\log(n)/\overline{N}_i^{(k)}\bigr)^{1/2} \Bigr\}^2 \\ &\le \varkappa^2 \log(n) \bigl(\sqrt{\mu} + \omega^{(k)} \sqrt{\mu} + \delta_i^{(k)} \bigr)^2 \le C_\lambda \log(n) \rho. \end{aligned}$$

Thus, on $\mathcal{A}^{(k)}$ holds $\boldsymbol{s}_{ij}^{(k+1)} = \lambda^{-1} T_{ij}^{(k+1)} \le \rho$ and $K_{\mathrm{st}}(\boldsymbol{s}_{ij}^{(k+1)}) \ge \nu$. This yields (5.9) and also $\widetilde{N}_i^{(k+1)} \ge \nu \overline{N}_i^{(k+1)}$ for all $X_i \in \mathcal{U}^{(k)}$. □

We present one corollary of Theorem 5.4.

Corollary 5.5. *Let, with a fixed k, Assumptions S0 and A1 through A4, (5.6) and (5.7) 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 $\widehat{\theta}_i^{(k)}$ fulfills*

$$\boldsymbol{P}\Bigl(\max_{X_i \in \mathcal{U}^{(k)}} \bigl| \overline{N}_i^{(k)} \mathcal{K}(\widehat{\theta}_i^{(k)}, \theta_i) \bigr| > \mu \log(n) \Bigr) \le 2k/n.$$

5.3 Control of stability by the memory step

Due to Theorem 5.4, a small error of the local constant approximation of $\theta(\cdot)$ in a vicinity of a point X_i ensures the propagation condition for the local models $W_i^{(k)}$. This particularly means that the k step estimate $\widehat{\theta}_i^{(k)}$ delivers the accuracy of estimation of order $\bigl(\log(n)/\overline{N}_i^{(k)}\bigr)^{1/2}$. Now we consider the situation when the local neighborhoods $U_i^{(k)}$ become too large during the iteration process and A4 is not fulfilled any more. Of course, propagation cannot be stated in this case and is not what we long for 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 property is almost a direct consequence of the construction of the "memory" step. Namely, the following proposition holds.

Proposition 5.6. *Under A1, for every i and every k, it holds*

$$\overline{N}_i^{(k)} \mathcal{K}\bigl(\widehat{\theta}_i^{(k)}, \widehat{\theta}_i^{(k-1)}\bigr) \le \tau. \tag{5.10}$$

Moreover, under A1 and A2, it holds for every $k' > k$ with $c_1 = \varkappa^2 \nu \bigl(1 - \sqrt{\nu}\bigr)^{-2}$:

$$\overline{N}_i^{(k)} \mathcal{K}\bigl(\widehat{\theta}_i^{(k')}, \widehat{\theta}_i^{(k)}\bigr) \le c_1 \tau. \tag{5.11}$$

Remark 5.2. An interesting feature of this result is that it is fulfilled with probability one. In fact, it follows just from the construction of the procedure. Assumption S0 is not required for the proof.

Proof. By definition of $\widehat{\theta}_i^{(k)} = \eta_i \widetilde{\theta}_i^{(k)} + (1-\eta_i)\widehat{\theta}_i^{(k-1)}$ with $\eta_i = K_{\mathrm{me}}\bigl(\tau^{-1}\overline{N}_i^{(k)}\mathcal{K}(\widetilde{\theta}_i^{(k)}, \widehat{\theta}_i^{(k-1)})\bigr)$. Convexity of the Kullback-Leibler divergence $\mathcal{K}(u,v)$ w.r.t. the first argument implies

$$\mathcal{K}\bigl(\widehat{\theta}_i^{(k)}, \widehat{\theta}_i^{(k-1)}\bigr) \le \eta_i \mathcal{K}\bigl(\widetilde{\theta}_i^{(k)}, \widehat{\theta}_i^{(k-1)}\bigr).$$

If $\mathcal{K}\bigl(\widetilde{\theta}_i^{(k)}, \widehat{\theta}_i^{(k-1)}\bigr) \ge \tau/\overline{N}_i^{(k)}$, then $\eta_i = 0$ and (5.10) follows.

Now, Assumption A1, Lemma 5.2 and Proposition 5.6 yield

$$\mathcal{K}^{1/2}\bigl(\widehat{\theta}_i^{k'}, \widehat{\theta}_i^{(k)}\bigr) \le \varkappa \sum_{l=k+1}^{k'} \mathcal{K}^{1/2}\bigl(\widehat{\theta}_i^{(l)}, \widehat{\theta}_i^{(l-1)}\bigr) \le \varkappa \sum_{l=k+1}^{k'} \bigl(\tau/\overline{N}_i^{(l)}\bigr)^{1/2}.$$

The use of Assumption A2 leads to the bound

$$\mathcal{K}^{1/2}\bigl(\widehat{\theta}_i^{(k')}, \widehat{\theta}_i^{(k)}\bigr) \le \varkappa\bigl(\tau/\overline{N}_i^{(k)}\bigr)^{1/2}(\nu^{1/2} + \ldots + \nu^{(k'-k)/2}) \le \varkappa\sqrt{\nu}\bigl(1-\sqrt{\nu}\bigr)^{-1}\bigl(\tau/\overline{N}_i^{(k)}\bigr)^{1/2}$$

which proves (5.11). □

The next theorem states the desirable "stability" property of the procedure. It follows directly from Proposition 5.6 with the use of Lemma 5.2.

Theorem 5.7. *Let A1 and A2 hold for all k. Let, for some k and some i,*

$$\overline{N}_i^{(k)}\mathcal{K}\bigl(\widehat{\theta}_i^{(k)}, \theta_i\bigr) \le \mu \log(n)$$

for some constant μ. Then it holds for the final estimate $\widehat{\theta}_i$

$$\overline{N}_i^{(k)}\mathcal{K}\bigl(\widehat{\theta}_i, \theta_i\bigr) \le c \log(n)$$

with $c = \varkappa^2\bigl(\sqrt{c_1 C_\tau} + \sqrt{\mu}\bigr)^2$, $C_\tau = \tau/\log(n)$ and c_1 from Proposition 5.6.

Remark 5.3. Corollary 5.5 and Theorem 5.7 imply the "oracle" property of the procedure. Indeed, if local homogeneity holds in a neighborhood of radius $h^{(k)}$ around X_i, then the local likelihood estimate with such "oracle" bandwidth delivers the accuracy of order $\bigl(\log(n)/\overline{N}_i^{(k)}\bigr)^{1/2}$, and a similar results holds for the adaptive estimate $\widehat{\theta}_i$.

5.4 Rate of estimation under smoothness conditions on $\theta(\cdot)$

Here we consider the case when $\theta(\cdot)$ satisfies some smoothness conditions in a neighborhood of a fixed point x. This means that the error of the local constant approximation of $\theta(\cdot)$ by $\theta(x)$ is sufficiently small. In addition we require some design regularity in the neighborhood of x. We show that the result of Theorems 5.4 and 5.7 lead in such a situation to the classical nonparametric rate of estimation $n^{-1/(2+d)}$ (up to a log multiplier) corresponding to the smoothness degree one.

Let a point $x = X_i$ be fixed. Define $\overline{h}^{(k)} = h^{(1)} + \ldots + h^{(k)}$ for $k \geq 1$ 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 $\overline{h}^{(k)} \leq h^{(k)}/(1-a^{-1})$. To ensure the quality of estimation of the function $\theta(\cdot)$ at the point X_i we assume some smoothness of $\theta(\cdot)$ and also some design regularity in the neighborhood $\mathcal{B}_i^{(k)}$ for some sufficiently large k.

(A4s) It holds $\mathcal{K}^{1/2}(\theta_j, \theta_{j'}) \leq L h^{(k)}$ for all $X_j, X_{j'} \in \mathcal{B}_i^{(k)}$ such that $|X_j - X_{j'}| \leq h^{(k)}$.

(A5) For a fixed k and every $X_i \in \mathcal{B}^{(k)}$, it holds for some constants $\nu_2 \leq \nu_3$

$$\nu_2 \leq \overline{N}_i^{(k)} / \bigl(n|h^{(k)}|^d\bigr) \leq \nu_3 .$$

Theorem 5.8. *Define $\check{h} = \bigl(L^2 n/\log(n)\bigr)^{-1/(2+d)}$ 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, (5.6) and (5.7) are satisfied for every $k' \leq k$ with $\delta_i^{(k')} = \delta$. Then*

$$\boldsymbol{P}\Bigl(\mathcal{K}^{1/2}\bigl(\widehat{\theta}_i, \theta_i\bigr) > C_1 L^{d/(2+d)} (\log(n)/n)^{1/(2+d)}\Bigr) \leq 2k^*/n \tag{5.12}$$

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

Proof. Set $c_0 = \bigl(\delta^2/\nu_3\bigr)^{1/(2+d)}$ and take k such that $h^{(k)}$ is the largest bandwidth that fulfills $h^{(k)} \leq c_0 \check{h}$. 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 $\overline{h}^{(k)}$. For any two points $X_j, X_{j'} \in \mathcal{B}_i^{(k)}$ with $|X_j - X_{j'}| \leq h^{(k)}$, the smoothness condition A4s clearly implies $\mathcal{K}^{1/2}(\theta_j, \theta_{j'}) \leq L h^{(k)}$. The use of $h^{(k)} = c\check{h}$ with $c \leq c_0$ yields

$$\overline{N}_j \mathcal{K}\bigl(\theta_j, \theta_{j'}\bigr) \leq L^2 \nu_3 |h^{(k)}|^{2+d} n = L^2 \nu_3 c^{2+d} \check{h}^{2+d} n = \nu_3 c^{2+d} \log(n) \leq \delta^2 \log(n)$$

and A4 holds true for the step k with $\delta_i^{(k)} = \delta$ 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 $h^{(k'+1)} + \ldots + h^{(k)}$. Therefore, Theorem 5.4 applies yielding with a high probability the following accuracy of estimating θ_i by $\widehat{\theta}_i^{(k)}$ under A4s and A5:

$$\mathcal{K}\bigl(\widehat{\theta}_i^{(k)}, \theta_i\bigr) \leq \mu \log(n) / \overline{N}_i^{(k)} \leq \frac{\mu \log(n)}{\nu_2 n |h^{(k)}|^d} \leq C_1 L^{2d/(2+d)} \bigl(\log(n)/n\bigr)^{2/(2+d)}$$

with some fixed constant C_1 depending on c and the other constants from Assumptions A2–A5. By Theorem 5.7, the same rate of estimation holds for the final estimate $\widehat{\theta}_i$. □

Remark 5.4. The rate of estimation given in Theorem 5.8 coincides with the optimal rate of estimation for the considered smoothness class 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 in the Sobolev and Besov function classes $B^1_{p,q}$.

5.5 Separation property

All results presented earlier discussed the propagation property and its consequences on the quality of estimation. In this section we state one more result indicating some benefits of the adaptive weights scheme. We show that the propagation stops when the local constant approximation does not provide a reasonable accuracy. More precisely, we consider the case of two different nearly homogeneous regions, and fixed two points X_{i_1} and X_{i_2}, one from every region. We assume that for both points the propagation holds until some step k which leads to the accuracy of estimation $\mathcal{K}(\widehat{\theta}^{(k)}_{i_m}, \theta_{i_m}) \le \mu_m \log(n)/\overline{N}^{(k)}_{i_m}$ for $m = 1, 2$. We show that if $\mathcal{K}(\theta_{i_1}, \theta_{i_2}) > C \log(n)/\overline{N}^{(k)}_{i_1}$ for a sufficiently large C then the procedure assigns a zero weight $w^{(k')}_{i_1 i_2}$ for all $k' \ge k$.

Theorem 5.9. *Let the statistical kernel K_{st} have a compact support on $[0, A]$ for some $A > 0$. Let, at step k, A1 and A3 be fulfilled and for two points X_{i_1} and X_{i_2} hold $\overline{N}^{(k)}_{i_m}\mathcal{K}\bigl(\widehat{\theta}^{(k)}_{i_m}, \theta_{i_m}\bigr) \le \mu_m \log(n)$ with some constants μ_m for $m = 1, 2$. Let also $N^{(k)}_{i_1} \ge b\overline{N}^{(k)}_{i_1}$ for some $b > 0$. If*

$$\overline{N}^{(k)}_{i_1}\mathcal{K}\bigl(\theta_{i_1}, \theta_{i_2}\bigr) > \varkappa^2\bigl(\sqrt{\mu_1} + \sqrt{\omega^{(k)}\mu_2} + \sqrt{C_\lambda A/b}\bigr)^2 \log(n)$$

then $w^{(k+1)}_{i_1 i_2} = 0$. Moreover, if $\mathcal{K}(\theta_{i_1}, \theta_{i_2}) > C \log(n)/\overline{N}^{(k)}_{i_1}$ for some fixed C, then for every $k' > k$ the condition $N^{(k')}_{i_1} \ge b\overline{N}^{(k')}_{i_1}$ implies also $w^{(k')}_{i_1 i_2} = 0$.

Proof. It suffices to show that $\boldsymbol{s}^{(k)}_{i_1 i_2} = \lambda^{-1} N^{(k)}_i \mathcal{K}(\widehat{\theta}^{(k)}_{i_1}, \widehat{\theta}^{(k)}_{i_2}) > A$. By Lemma 5.2

$$\begin{aligned}
\mathcal{K}^{1/2}\bigl(\widehat{\theta}^{(k)}_{i_1}, \widehat{\theta}^{(k)}_{i_2}\bigr) &\ge \varkappa^{-1}\mathcal{K}^{1/2}\bigl(\theta_{i_1}, \theta_{i_2}\bigr) - \mathcal{K}^{1/2}\bigl(\widehat{\theta}^{(k)}_{i_1}, \theta_{i_1}\bigr) - \mathcal{K}^{1/2}\bigl(\widehat{\theta}^{(k)}_{i_2}, \theta_{i_2}\bigr) \\
&\ge \varkappa^{-1}\mathcal{K}^{1/2}\bigl(\theta_{i_1}, \theta_{i_2}\bigr) - \sqrt{\mu_1 \log(n)/\overline{N}^{(k)}_{i_1}} - \sqrt{\mu_2 \log(n)/\overline{N}^{(k)}_{i_2}} \\
&\ge \varkappa^{-1}\mathcal{K}^{1/2}\bigl(\theta_{i_1}, \theta_{i_2}\bigr) - \sqrt{\mu_1 \log(n)/\overline{N}^{(k)}_{i_1}} - \sqrt{\mu_2 \log(n)\omega^{(k)}/\overline{N}^{(k)}_{i_1}}.
\end{aligned}$$

This and the inequality $N^{(k)}_{i_1} \ge b\overline{N}^{(k)}_{i_1}$ yield

$$\lambda^{-1} N^{(k)}_{i_1}\mathcal{K}\bigl(\widehat{\theta}^{(k)}_{i_1}, \widehat{\theta}^{(k)}_{i_2}\bigr) \ge bC_\lambda^{-1}\Bigl(\varkappa^{-1}\sqrt{\overline{N}^{(k)}_{i_1}\mathcal{K}(\theta_{i_1}, \theta_{i_2})/\log(n)} - \sqrt{\mu_1} - \sqrt{\omega^{(k)}\mu_2}\Bigr)^2 \ge A$$

and the first assertion follows. The second one can be easily shown by involving the result of Proposition 5.6. □

6 Some exponential bounds for exponential families

We present some general results for the local exponential family model. The considered exponential family $\mathcal{P} = (P_\theta\,, \theta \in \Theta \subseteq I\!R)$ is described by the functions $C(\theta)$ and $B(\theta)$, with $p(y,\theta) = dP_\theta/dP(y) = p(y)\exp\left(C(\theta)y - B(\theta)\right)$ and $E_\theta Y = \int y p(y,\theta) dP(y) = \theta$ for all $\theta \in \Theta$, see Section 2.1.

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

It is convenient to introduce the parameter $\upsilon = C(\theta)$ and define $\overline{\upsilon} = C(\overline{\theta})$ and $D(\upsilon) = B(\theta) = B(C^{-1}(\upsilon))$. Since $C'(\theta) > 0$, the new parameter υ is uniquely defined. By simple analysis $D'(\upsilon) = \theta = C^{-1}(\upsilon)$ and $D''(\upsilon) = 1/C'(\theta) = 1/I(\theta) = 1/I(C^{-1}(\upsilon))$. Moreover, $\mathcal{K}(\upsilon_1,\upsilon_2) = D(\upsilon_2) - D(\upsilon_1) - (\upsilon_2 - \upsilon_1)D'(\upsilon_1)$ is the Kullback-Leibler distance between two parametric distributions corresponding to the parameters υ_1 and υ_2. In what follows we use the notation $q(u|\upsilon) = \mathcal{K}(\upsilon, \upsilon + u) = D(\upsilon + u) - D(\upsilon) - uD'(\upsilon)$.

Theorem 6.1. *Let the Fisher information $I(\theta) = C'(\theta)$ be positive on Θ. For a given $z \geq 0$, let $\mathcal{U}(W,z)$ be the set of solutions u of equation $q(u|\overline{\upsilon}) = \int_0^u x D''(\overline{\upsilon}+x)dx = z/N$. If there is some $\alpha > 0$ such that for all $\mu \in (0,1]$ and all $u \in \mathcal{U}(W,z)$*

$$q(\pm w_\ell\, \mu u | \upsilon_\ell) \leq (1+\alpha) w_\ell\, \mu^2 q(u|\overline{\upsilon}), \qquad \ell = 1,\ldots,n, \tag{6.1}$$

then

$$\boldsymbol{P}\bigl(L(W,\widetilde{\theta},\overline{\theta}) > z\bigr) = \boldsymbol{P}\bigl(N\mathcal{K}(\widetilde{\theta},\overline{\theta}) > z\bigr) \leq 2e^{-z/(1+\alpha)}.$$

Remark 6.1. The condition (6.1) can be easily checked in many particular situations. We give two typical examples. The first one corresponds to the homogeneous case when all υ_i coincide with their mean $\overline{\upsilon}$. Then (6.1) is fulfilled automatically with $\alpha = 0$. Indeed the function $q(\cdot|\upsilon)$ satisfies $q'(u|\upsilon) = D'(\upsilon + u) - D'(\upsilon)$ and $q''(u|\upsilon) = D''(\upsilon + u) = 1/I(C^{-1}(\upsilon + u)) > 0$ and thus, it is convex. Since also $q(0|\upsilon) = 0$, it holds

$q(wa|v) \leq wq(a|v)$ for every $w \in [0,1]$ and every a implying (6.1) with $\alpha = 0$ and arbitrary u. This special case was already stated as a separate result in Theorem 2.1.

The second special case was mentioned in Theorem 2.2. Assume A1. The Taylor expansion yields that $q(wu|v) = D(v+wu) - D(v) - wuD'(v) = 1/2\, w^2u^2D''(v+\tau wu)$ for some $\tau \in [0,1]$. Under condition A1 $\varkappa^{-2} \leq D''(v)/D''(\overline{v}) \leq \varkappa^2$ for all v and one easily obtains $u^2 \leq 2zI^*/N$ for every $u \in \mathcal{U}(W,z)$.Therefore, the condition (6.1) is easy to check for $1+\alpha = \varkappa^2$ which yields Theorem 2.2 as corollary of Theorem 6.1. Moreover, only the local variability of the Fisher information $I(\theta)$ on the support of the local model W is important so the value α is typically close to zero.

Proof of Theorem 6.1 The log-likelihood ratio can be rewritten for the new parameter v as

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

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

Lemma 6.2. *For given z, there exist two values $v^* > \overline{v}$ and $v_* < \overline{v}$ such that*

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

Proof. It holds

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

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

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

or, equivalently,

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

Therefore

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

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

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

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

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

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

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

Lemma 6.3. *For every u and every z*

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

Moreover, if u fulfills (6.1) then

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

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

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

The independence of the Y_ℓ's implies

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

The equalities $\log \int e^{\upsilon_\ell y - D(\upsilon_\ell)} P(dy) = 0$ and $\theta_\ell = D'(\upsilon_\ell)$ yield

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

for every $a \ge 0$ and every $\ell \le n$. Therefore

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

This inequality applied with $\mu = (1+\alpha)^{-1}$ and (6.1) imply

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

Similarly

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

and the lemma follows. □

7 Conclusion and outlook

This paper presents a new method of adaptive nonparametric estimation based on the idea of *propagation* and *separation* using *adaptive weights*. An important feature of the proposed PS procedure is that it applies to many different statistical problems in a unified way. In many cases its adjustment to the particular situation is trivial. For all the examples in this paper, we essentially applied the same procedure. Sometimes, a preliminary model (data) transformation is required, as in density estimation. The procedure can automatically handle jumps and discontinuities of the underlying function. The procedure allows for arbitrary dimensionality of $\mathcal{X}$. This makes it feasible to apply the procedure to e.g. image denoising or estimation of a multivariate density and to use it in case of a multidimensional explanatory vector X_i. The PS procedure is computationally straightforward and the numerical complexity can be easily controlled by restricting the largest bandwidth h^*, see PS2000 for details.

We also establish some remarkable theoretical results about the properties of the resulting estimate. In particular, it is spatially adaptive and achieves the optimal rate of convergence over Besov function classes. The results are stated and proved in a precise nonasymptotic form and they apply for a broad class of nonparametric statistical models. Our results heavily rely on the general exponential bound for exponential families, see Theorem 6.1 and its corollary which seem to be of independent interest.

References

[1] Bickel, P.J., C.A.J. Klaassen, Y. Ritov and J.A. Wellner (1998). *Efficient and Adaptive Estimation for Semiparametric Models*, 1998, Springer.

[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] Fan, J., Farmen, M. and and Gijbels, I. (1998). Local maximum likelihood estimation and Inference. *J. Royal Statist. Soc.* Ser. B, **60**, 591–608.

[5] Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Royal Statist. Soc. B* **57** 371–394.

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

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

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

[9] Fan, J. and Zhang, J. (2004). Sieve empirical likelihood ratio tests for nonparametric functions. *Ann. Statist.* **32** 1858-1907.

[10] Friedman, J.H. (2001). Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **29** no. 5, 11891232.

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

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

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

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

[15] 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.

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

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

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

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

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

[21] Polzehl, J. and Spokoiny, V. (2003). Image denoising: pointwise adaptive approach. *Annals of Statistics*, **31** 30–57.

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

[23] Polzehl, J. and Spokoiny, V. (2004b). Varying coefficient GARCH versus local constant volatility modeling. Comparison of the predictive power. WIAS-Preprint No. 977, 2004.

[24] Polzehl, J. and Spokoiny, V. (2004c). Spatially adaptive regression estimation: Propagation-separation approach. WIAS-Preprin No. 998.

[25] Qiu, P. (1998). Discontinuous regression surface fitting. *Ann. Statist.* **26** no. 6, 2218–2245.

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

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

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