**Weierstraß-Institut**

**für Angewandte Analysis und Stochastik**

im Forschungsverbund Berlin e.V.

Preprint ISSN 0946 – 8633

# Spatially adaptive local likelihood modeling via stagewise aggregation

Denis Belomestny [1] and Vladimir Spokoiny [2]

submitted: 19th Januar 2005

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

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

No. 1000
Berlin 2005

2000 *Mathematics Subject Classification.* 62G05,62G07, 62G08, 62H30 .

*Key words and phrases.* adaptive weights, local likelihood, exponential family, density estimation, classification.

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**

The paper presents a new method of spatially adaptive local likelihood estimation for a broad class of nonparametric models, including e.g. the regression, Poisson and binary response model. Given a sequence of local likelihood estimates which we call ”weak” estimates, the proposed method yields a new aggregated estimate whose pointwise risk does not exceed the smallest risk among all “weak” estimates up to some logarithmic multiplier. We establish a number of important theoretical results concerning optimality of the aggregated estimate and show a good performance of the procedure in simulated and real life examples.

# 1 Introduction

This paper offers a new method of spatially adaptive nonparametric estimation based on aggregating a family of local likelihood estimates. Local likelihood approach was intensively discussed last years, see e.g. Tibshirani and Hastie (1987), Staniswalis (1989), Loader (1996). We refer to Fan, Farmen and Gijbels (1998) for a nice and detailed overview of local maximum likelihood approach and related literature. In particular, the suggested method is very general and applies to many statistical models in a unified way. Similarly to usual nonparametric smoothing in regression or density framework, an important issues for local likelihood modeling is the choice of localization (smoothing) parameters. Different types of model selection techniques based on the asymptotic expansion of the local likelihood are mentioned in Fan, Farmen and Gijbels (1998) which includes global and variable bandwidth selection. However, the performance of estimators based on bandwidth selection is often rather unstable, see e.g. Breiman (1996). This suggests that in some cases, the attempt to identify the true local model is not necessarily the right thing to do. One approach to reduce variability in adaptive estimation is model mixing or aggregation. Yang (2004), Catoni (2001) among other suggested global aggregated procedures that achieves the best estimation risks over the family of given “weak” estimates. Nemirovski (2000), Juditsky and Nemirovski (2000) developed for the regression set-up the aggregation procedures that achieves a risk within a multiple of $log(n)/n$ of the smallest risk in the class of all convex combinations of "weak" estimates. Tsybakov (2003) discussed the asymptotic minimax rate for aggregation. Aggregation for density estimation has been investigated by Li and Barron (1999), Tsybakov (2005). A pointwise aggregation has not been yet considered to the best of our knowledge.

We propose a new approach towards local likelihood modelling which is based on the idea of the spatial (pointwise) aggregation of a family of local likelihood estimates ("weak" estimates) $\widetilde{\theta}^{(k)}$. The main idea is, given the sequence $\{\widetilde{\theta}^{(k)}\}$ to construct in a data driven way the "optimal" aggregated estimate $\widehat{\theta}(x)$ separately at each point $x$. "Optimality" means that this estimate satisfies some kind of oracle inequality, that is, its pointwise risk does not exceed the smallest pointwise risk among all "weak" estimates up to a logarithmic multiple.

Our algorithm can be roughly described as follows. Let $\{\widetilde{\theta}^{(k)}(x)\}$, $k = 1, \ldots, K$, be a "nested" sequence of weak local likelihood estimates at a point $x$ ordered due to decreasing variability. A new aggregated estimate of $\theta(x)$ is constructed sequentially by mixing previously constructed aggregated estimate $\widehat{\theta}^{(k-1)}$ with the current "weak" estimate $\widetilde{\theta}^{(k)}$:

$$\widehat{\theta}^{(k)} = \gamma_k \widetilde{\theta}^{(k)} + (1-\gamma_k)\widehat{\theta}^{(k-1)},$$

where the mixing parameter $\gamma_k$ (which may depend on the point $x$) is defined using a measure of statistical difference between $\widehat{\theta}^{(k-1)}$ and $\widetilde{\theta}^{(k)}$. In particular, $\gamma_k$ is equal to zero if $\widehat{\theta}^{(k-1)}$ lies outside the confidence interval around $\widetilde{\theta}^{(k)}$. In view of the sequential and poinwise nature of the algorithm, the suggested procedure is called *Spatial Stagewise Aggregation* (SSA). An important feature of the procedure proposed is that it is very simple and transparent and applies in a unified manner for a big family of different models like Gaussian, binary, Poisson regression, density estimation, classification etc. The procedure does not require any splitting of the sample as many other aggregation procedures do, cf. Yang (2004). The SSA procedure can be easily studied theoretically. We establish precise nonasymptotic "oracle" results which apply under very mild conditions in a rather general set-up. We also show that the oracle property automatically implies spatial adaptivity of the proposed estimate.

The paper is organized as follows. Section 2 describes the considered model and our setup: varying coefficient exponential family. Section 2.3 presents some useful exponential inequalities for the lack of fit statistic in context of local likelihood estimation. A detailed description of the proposed method is given in Section 3. Applications to regression, density estimation and classification are discussed in Sections 4, 5, 6 respectively. Theoretical properties of the aggregation procedure are presented in Section 7. Finally, some technical assertions and proofs about the varying coefficient exponential family are collected in Section 8.

# 2 Local likelihood modeling

This section describes the considered model and states the problem. Suppose we are given independent random data $Z_1, \ldots, Z_n$ of the form $Z_i = (X_i, Y_i)$. Here every $X_i$ means 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 $\theta$ which may depend on the location $X_i$, $\theta = \theta(X_i)$. We illustrate this set-up by means of the few examples.

**Example 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 $\theta$ and i.i.d. Gaussian errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

**Example 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, $\boldsymbol{P}(Y_i = 1 \mid X_i = x) = \theta(x)$ and $\boldsymbol{P}(Y_i = 0 \mid X_i = x) = 1 - \theta(x)$. Such models arise in many econometric applications, they are widely used in classification and digital imaging.

**Example 3. (Inhomogeneous Poisson model)** Suppose that every $Y_i$ is valued in the set $\mathbb{N}$ of nonnegative integer numbers and $\boldsymbol{P}(Y_i = k \mid X_i = x) = \theta^k(x) e^{-\theta(X_i)}/k!$, that is, $Y_i$ follows a Poisson distribution with parameter $\theta = \theta(x)$. This model is commonly used in the queueing theory, it occurs in positron emission tomography, it also serves as an approximation of the density model, obtained by a binning procedure.

All the given examples are particular cases of the varying coefficient exponential family model, see Section 2.2 for more details. Some further examples can be found in Fan, Farmer and Gijbels (1998).

Now we present a formal definition for our model. Let $\mathcal{P} = (P_\theta, \theta \in \varTheta)$ be a family of probability measures on $\mathcal{Y}$ where $\varTheta$ 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 the density $p(\cdot, \theta(x))$ for some unknown function $\theta(x)$ on $\mathcal{X}$. The aim of the data-analysis is to infer on this function $\theta(x)$.

In the parametric setup, when the parameter $\theta$ does not depend on the location, that is, the distribution of every "observation" $Y_i$ coincides with $P_\theta$ for some $\theta \in \varTheta$ the parameter $\theta$ can be well estimated by the parametric maximum likelihood method:

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

In the nonparametric varying coefficient framework, one usually applies the local likelihood approach which is based on the assumption that the parameter $\theta$ is constant only within some neighborhood of every point $x$ in the "feature" space $\mathcal{X}$. This leads to considering a local model concentrated in some neighborhood of the point $x$.

## 2.1 Localization

We use *localization by weights* as a general method to describe a local model. Let, for a fixed $x$, a nonnegative weight $w_i = w_i(x) \le 1$ be assigned to the observations $Y_i$ at $X_i$, $i = 1, \ldots, n$. The weights $w_i(x)$ determine a local model corresponding to the point $x$ in the sense that, 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

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

We mention two possible ways 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 $\boldsymbol{l}_i = \rho(x, X_i)/h$ where $h$ is a bandwidth, $\rho(x, X_i)$ is the Euclidean distance between $x$ and the design point $X_i$ and $K_{\mathrm{loc}}$ is a *location kernel.*

*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 classification by $k$-nearest neighbor method or in the regression tree approach.

We do not assume any special structure for the weights $w_i(x)$, that is, any configuration of weights is allowed. In what follows we will identify a local model in $x$ by the set $W(x) = \{w_1(x), \dots, w_n(x)\}$ and denote

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

## 2.2 Local likelihood estimation for an exponential family model

The examples given above can be considered as particular cases of local exponential family distributions. This means that all measures $P_\theta$ from this family are dominated by a $\sigma$-finite measure $P$ on $\mathcal{Y}$ and 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 $\Theta$ and $p(y)$ is some nonnegative function on $\mathcal{Y}$. The parameter $\theta$ is defined by the equations $\int p(y,\theta)P(dy) = 1$ and $\boldsymbol{E}_\theta Y = \int y p(y,\theta)P(dy) = \theta$ which implies the relation $B'(\theta) = \theta C'(\theta)$.

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 $I(\theta) := \boldsymbol{E}_\theta |p'_\theta(Y,\theta)/p(Y,\theta)|^2$ satisfy

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

Table 1 provides the Kullback-Leibler distance $\mathcal{K}(\theta, \theta')$ for the examples from Section 2.

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

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

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

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

Table 1: $\mathcal{K}(\theta, \theta')$ and $I(\theta)$ for the examples from Section 2.

| Model | $\mathcal{K}(\theta, \theta')$ | $I(\theta)$ |
|---|---|---|
| Gaussian regression | $(\theta - \theta')^2/(2\sigma^2)$ | $\sigma^{-2}$ |
| Bernoulli model | $\theta \log(\theta/\theta') + (1-\theta)\log\{(1-\theta)/(1-\theta')\}$ | $\theta^{-1}(1-\theta)^{-1}$ |
| Poisson model | $\theta \log(\theta/\theta') - (\theta - \theta')$ | $1/\theta$ |

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

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

## 2.3 Exponential Inequalities for the Lack of Fit Statistic

Here we present some exponential inequalities for the "lack of fit statistic" $L(W, \widetilde{\theta}, \theta)$ which apply for arbitrary weights and arbitrary sample size.

We assume some regularity of the considered parametric family $\mathcal{P}$.

**(A1)** $\mathcal{P} = (P_\theta, \theta \in \Theta \subseteq I\!R)$ is an exponential family with a one-dimensional parameter.

**(A2)** $\Theta$ is compact and the Fisher information $I(\theta)$ fulfills

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

Our first result can be regarded as a nonasymptotic local version of the Wilks theorem.

**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 for any $z > 0$*

$$\boldsymbol{P}(L(W, \widetilde{\theta}, \theta) > z) = \boldsymbol{P}\left(N\mathcal{K}(\widetilde{\theta}, \theta) > z\right) \le 2e^{-z}.$$

**Remark 1.** The local likelihood estimate $\widetilde{\theta}$ does not change if all the weights $w_i$ are multiplied by the same constant $c$, see (2.1). However, the lack of fit statistic $L(W, \widetilde{\theta}, \theta)$ will be multiplied by this constant. The result of Theorem 2.1 continues to apply after this multiplication provided that the condition $\max_i w_i \le 1$ still holds. The strongest result corresponds to the case with $\max_i w_i = 1$.

**Remark 2.** Condition A2 ensures that the Kullback-Leibler divergence $\mathcal{K}$ fulfills $\mathcal{K}(\theta', \theta) \le I|\theta' - \theta|^2$ for any point $\theta'$ in a neighborhood of $\theta$, 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 $\theta$. Indeed, under homogeneity, the true parameter value $\theta$ 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 $w_j$. In this case the "lack of fit statistic" between the local likelihood estimate $\widetilde{\theta}$ and the corresponding mean value $\overline{\theta} := \boldsymbol{E}\widetilde{\theta} = N^{-1}\sum_{i=1}^{n} w_i\theta_i$ with $\theta_i = \theta(X_i)$ 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 and A2, then there is* $\alpha \ge 0$ *depending on* $\varkappa$ *only such that for every* $z > 0$

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

More details and proofs can be found in Section 9.

## 3 Description of the method

Let a point $x \in \mathcal{X}$ be fixed and let $\{\widetilde{\theta}^{(k)}(x),\, k = 1, \ldots, K\}$ be a sequence of local likelihood estimates of $\theta = \theta(x)$ of the type

$$\widetilde{\theta}^{(k)}(x) = \sum_{i=1}^{n} w_i^{(k)} Y_i \Big/ \sum_{j=1}^{n} w_j^{(k)}, \quad w_i^{(k)} = w_i^{(k)}(x) \in [0, 1].$$

We say that the sequence $\{\widetilde{\theta}^{(k)}\}$ is *strictly nested*, if

**(A3)** for some constants $\nu_*, \nu^*$ with $0 < \nu_* \le \nu^* < 1$, the values $N_k = \sum_{j=1}^{n} w_j^{(k)}$ satisfy for every $2 \le k \le K$

$$\nu_* \le N_{k-1}/N_k \le \nu^*.$$

Some typical examples of strictly nested sets of estimates are given below in Section 3.1.

**Remark 1.** Due to Theorems 2.1 and 2.2 the value $1/N_k$ measures the variability of the estimate $\widetilde{\theta}^{(k)}$ in the homogeneous or nearly homogeneous cases. The condition A3 means that variability of the estimates $\widetilde{\theta}^{(k)}$ decreases with $k$.

Given the set of strictly nested "weak" estimates $\widetilde{\theta}^{(k)} = \widetilde{\theta}^{(k)}(x)$, we consider a larger class of their convex combinations $\widehat{\theta}$:

$$\widehat{\theta} = \sum_{k=1}^{K} \alpha_k \widetilde{\theta}^{(k)}, \quad \alpha_1 + .... + \alpha_K = 1, \quad \alpha_k \geq 0,$$

where the mixing coefficients $\alpha_k$ which may depend on the point $x$. We aim at constructing a new estimate $\widehat{\theta}$ in this class which performs as good as the best one in the original family $\{\widetilde{\theta}^{(k)}(x)\}$. This estimate is computed sequentially via the following algorithm.

1. **Initialization**:

$$\widehat{\theta}^{(1)} = \widetilde{\theta}^{(1)}.$$

2. **Stagewise aggregation**: For $k = 2, ..., K$

$$\widehat{\theta}^{(k)} := \gamma_k \widetilde{\theta}^{(k)} + (1 - \gamma_k)\widehat{\theta}^{(k-1)},$$

with the mixing parameter $\gamma_k$ defined for some $\lambda > 0$ and a kernel $K_{\mathrm{ag}}(\cdot)$ as

$$\gamma_k = K_{\mathrm{ag}}\bigl(\boldsymbol{m}^{(k)}/\lambda\bigr), \qquad \boldsymbol{m}^{(k)} := N_k \mathcal{K}(\widetilde{\theta}^{(k)}, \widehat{\theta}^{(k-1)})$$

**3. Final Estimate**: $\widehat{\theta} = \widehat{\theta}^{(K)}$.

The idea behind the procedure is quite simple. We start with the "weakest" estimate $\widetilde{\theta}^{(1)}$ having the smallest degree of locality but the largest variability of order $1/N_1$. Next we consider estimates with larger values $N_k$. Every next estimate $\widetilde{\theta}^{(k)}$ is compared with the previously constructed estimate $\widehat{\theta}^{(k-1)}$. If the difference is not significant then the new estimate $\widehat{\theta}^{(k)}$ basically coincides with $\widetilde{\theta}^{(k)}$. Otherwise the procedure essentially keeps the previous value $\widehat{\theta}^{(k-1)}$. For measuring the difference between estimates, we apply the penalty $\boldsymbol{m}^{(k)} := N_k \mathcal{K}(\widetilde{\theta}^{(k)}, \widehat{\theta}^{(k-1)})$ as explained in Remark 2.

**Remark 2.** If $K_{\mathrm{ag}}(\cdot)$ is the uniform kernel on $[0, 1]$ then $\gamma_k$ is either zero or one depending on the value $\boldsymbol{m}^{(k)}$. This easily yields by induction arguments that the final estimate coincides with one of the "weak" estimates $\widetilde{\theta}^{(k)}$. In this case our method can be considered as a pointwise model selection method.

## 3.1 Examples of sequences of local likelihood estimates

A sequence of "weak" local likelihood estimates at point $x$ is uniquely defined by the weights $w_i^{(k)}(x)$, $k = 1, ..., K$. We use mainly two weighting schemes, corresponding to two possible localization methods: localization by a kernel and localization by a $k$-nearest neighbor window.

In the case of kernel weights we employ Epanechnikov kernel $K_{\mathrm{loc}}(x) = (1 - x^2)_+$ and define weights at point $x$ as

$$w_i^{(k)}(x) := K_{\mathrm{loc}}(\rho(x, X_i)/h_k), \quad k = 1, ..., K,$$

where $h_k$ is an exponentially increasing sequence of bandwidths with $h_k/h_{k-1} = a$. Here $h_1$ and $a$ can be treated as parameters of the procedure. It easy to see that the above choice of $h_k$ delivers an exponentially increasing sequence of $N_k$ under usual condition on the design $X_1, \ldots, X_n$. Such kind of local likelihood sequences is efficient only in the case of a low dimensional design space $\mathcal{X}$.

For a given $k$, a $k$-NN window $U(x)$ is taken to contain $k$ nearest neighbors of the point $x$. In this case

$$w_i^{(k)}(x) := \mathbf{1}\big(\rho(x, X_i) \le \rho_{(k)}\big)$$

where $\rho_{(1)} \le \rho_{(2)} \le \ldots \le \rho_{(n)}$ is the ordered sequence of the distances $\rho_i := \rho(x, X_i)$. A sequence of integer numbers $k_j = [a^{j-1}k_1]$, $j = 1, ..., K$ with some fixed initial number $k_1$ uniquely determines an exponentially increasing sequence $\{N_j\}$. Local likelihood estimates with the $k$-NN localization scheme are particulary interesting for the classification problem in high dimensions.

Sometimes a hybrid scheme with $w_i^{(k)}(x) = K_{\mathrm{loc}}(\rho_i/\rho_{(k)})$ can be useful.

## 3.2 Choice of parameters

**Kernel $K_{\mathrm{ag}}$:** The kernel $K_{\mathrm{ag}}$ should satisfy $0 \le K_{\mathrm{ag}} \le 1$ and should be supported on $[0, 1]$. Our default choice is the triangle kernel $K_{\mathrm{ag}}(u) = (1 - u)_+$.

**Parameters defining the weighting scheme:** The initial bandwidth and initial number of nearest neighbors should be reasonable small. In most examples we fix small natural $k_1$ and select $h_1 = c/n$ with some $c$ ensuring that every ball with center $X_i$ and radius $h_1$ contain at least $k_1$ points. The parameter $a$ controls the growth rate of the local neighborhoods. It should be selected to provide that the mean number of points inside a ball $\mathcal{B}_{h_k}(x)$ with radius $h_k$ grows exponentially with $k$ for some factor $a_{grow} > 1$. If $X_i$ are from $I\!R^d$, then in the case of kernel weights the parameter $a$ can be taken as $a = a_{grow}^{1/d}$. For the $k$-NN weights we just take $a = a_{grow}$. Our default choice is $a_{grow} = 1.25$. Any value in the range $[1.1, 1.3]$ can be taken as well. The maximal bandwidth $h_K$ can be taken large so that every ball $\mathcal{B}_{h_K}(x)$ contains the whole sample for the last iteration $K$. The geometric grow of the parameter $h$ or of the number of nearest neighbors ensures that the total number of iterations is typically bounded by $C\log(n)$ for some fixed constant $C$.

**Parameter $\lambda$:** The most important parameter of the procedure is $\lambda$ which scales the statistical penalty $\boldsymbol{m}^{(k)}$. Small values of $\lambda$ lead to overpenalization and a high variability of the resulting estimate. Large values of $\lambda$ may result in loss of adaptivity of the method and oversmoothing. In some sense this parameter is similar to the wavelet threshold applied in a nonlinear wavelet transform.

A reasonable way to define the parameter $\lambda$ for specific applications is based on the "monotonicity condition". This condition means that in a homogeneous situation $\theta(X_i) \equiv \theta$, the mixing parameter $\gamma_k$ is close to one for each $1 \le k \le K$. This would lead to an aggregated estimate $\widehat{\theta}$ which essentially coincides with $\widetilde{\theta}^{(K)}$. Therefore, one can adjust the parameter $\lambda$ simply selecting by Monte-Carlo simulations the minimal value of $\lambda$ providing a prescribed probability of getting $\gamma_K \approx 1$ for parametric model $\theta(x) \equiv \theta$. A theoretical justification is given by Proposition 7.1, that claims that the choice $\lambda = C_\lambda \log n$ with a sufficiently large $C_\lambda$ yields the "monotonicity" condition whatever the parameter $\theta$ or the sample size $n$ is.

Note that at the end of the iteration process the strong overlapping of the models $W^{(k)}$ and $W^{(k-1)}$ causes a high correlation between the estimates $\widetilde{\theta}^{(k)}$ and $\widehat{\theta}^{(k-1)}$. This suggests to take a relatively large value of $\lambda$ in the beginning and decrease it with iterations until a lower bound, say $\lambda_0$ is reached. This leads to the following proposal: $\lambda_k = \max\{\lambda_1 - \lambda_2 \log h^{(k)}, \lambda_0\}$ for some $\lambda_0, \lambda_1$ and $\lambda_2$. Our default choice which is used in all examples below is $\lambda_1 = 3$ and $\lambda_0 = 0.05\lambda_1$.

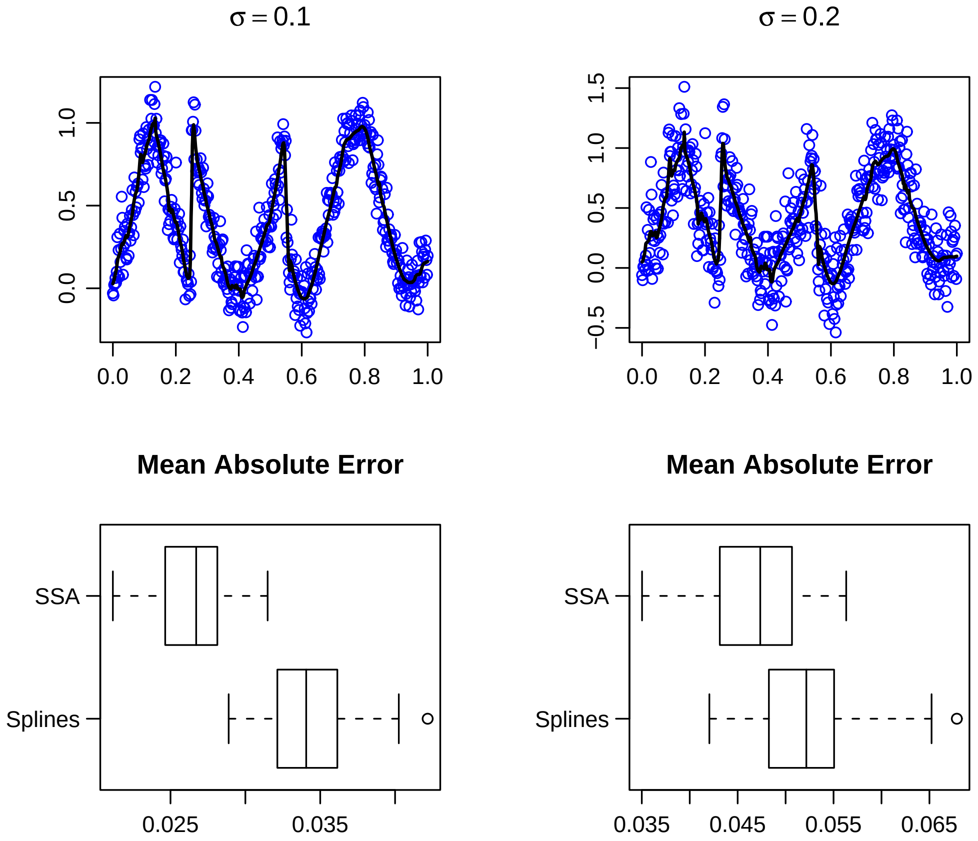

Figure 1: Simulated data sets together with SSA estimates (top row) and Box-Plots of MAE for SSA and penalized cubic smoothing splines (bottom row) for Example 1.

# 4 Application to regression

This section illustrates how the SSA procedure can be used in the univariate regression set-up by means of two simulated examples. The data are generated as $(X_i, Y_i)$ with $Y_i = f(X_i) + \sigma\varepsilon_i$ for $i = 1, \ldots, n$. The sample size is $n = 500$. The points $X_i$ are equidistant on $(0, 1)$. Errors $\varepsilon_i$ are i.i.d. standard Gaussian. The error variance $\sigma^2$ is unknown and estimated from the data.

For comparison we use a penalized cubic smoothing spline, with smoothing parameter determined by generalized cross validation. See Heckman and Ramsey (2000) for details.

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

$$f_1(x) = \begin{cases} 8x & x < 0.125, \\ 2 - 8x & 0.125 \le x < 0.25, \\ 44(x-0.4)^2 & 0.25 \le x < 0.55, \\ 0.5\cos(6\pi(x-0.775) + 0.5 & 0.55 \le x. \end{cases}$$

The upper row of Figure 1 shows plots of the first data set for $\sigma = 0.1$ and $0.2$, respectively, together with the estimate obtained by SSA with default parameters and $h_K = 1$. The bottom row reports the results in form of box-plots of Mean Absolute Error (MAE) obtained for the two procedures in 500 simulation runs.

**Example 2.** In a second example we consider the following smooth function

$$f_2(x) = \sin\left(\frac{2.4\pi}{x+0.2}\right), \qquad x \in [0,1].$$

Figure 2 shows the results for the function $f_2$.

In both examples SSA clearly outperforms penalized smoothing splines in terms of global mean averaged risk.

# 5 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 SSA method. Here we consider the case of small or moderate $d$, e.g. $d \le 3$.

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). Let the interval $[0,1]$ be split into $M$ equal disjoint intervals of length $\delta = 1/M$. Then the cube $[0,1]^d$ can be split into $n = M^d$ nonoverlapping small cubes with the side length $\delta$, 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 $i$th cube $J_i$. The pairs $(X_i, Y_i)$ for $i = 1, \ldots, n$ can be viewed as new observations.

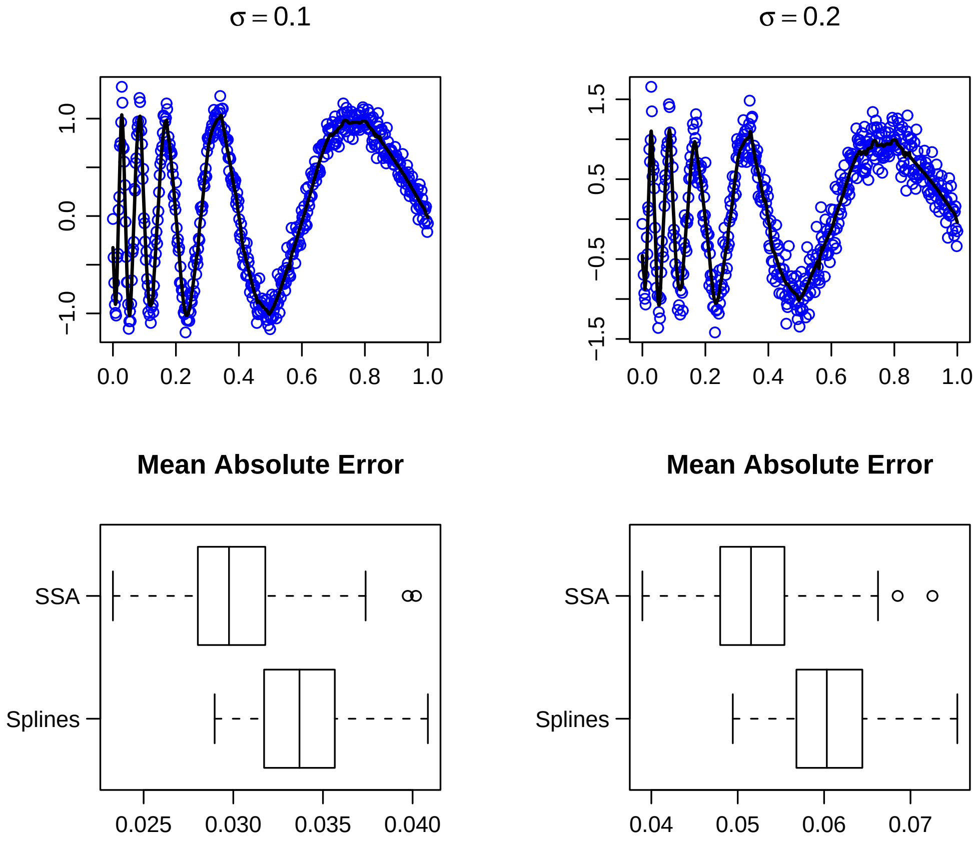

Figure 2: One sample and the SSA estimate (top), Box-Plots of MAE for SSA and penalized cubic smoothing splines (bottom) for Example 2.

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 $\theta_i = Lp_i = LP(J_i)$.

If the value $\theta_i$ has been estimated by $\widetilde{\theta}_i$ then the target density $f$ is estimated at $X_i$ as $\widetilde{f}(X_i) = n\widetilde{\theta}_i / \sum_{j=1}^{n} \widetilde{\theta}_j$.

For estimating the values $\theta_i$ from the "observations" $(X_i, Y_i)$ we apply the SSA procedure with the local Poisson family from Example 3. In addition to the standard parameter set, we need to specify the bin length $\delta$. A reasonable choice is $\delta = c/K$ where $K$ is the smallest integer satisfying $K^d \geq L$ and $c \leq 1$. The procedure applies even if $c$ is small and many bin counts $Y_i$ are zero. For comparison we also computed the kernel density estimates (KDE) with Epanechnikov kernel and the bandwidth minimizing the estimated Mean Absolute Error (MAE).

**Example 1.** We test our procedure for two univariate normal mixture densities taken from the set of 15 densities provided by Marron and Wang (1992). We generate in each case $n = 500$ observations. In the upper row of Figure 3 we show

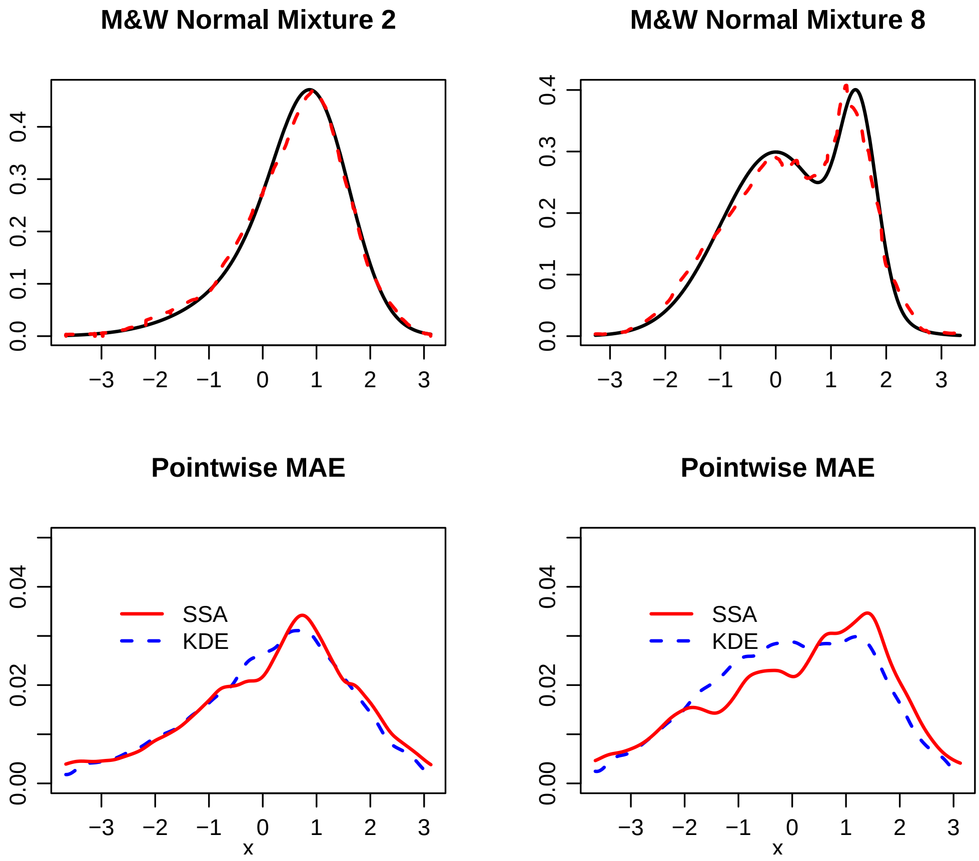

Figure 3: Top: SSA estimates from 500 observations (dashed line) and true density curves (solid line). Bottom: pointwise MAE for SSA and the kernel density estimates based on 500 simulations.

typical realizations of the densities estimates by SSA (dashed line) obtained from 500 observations using a regular grid with interval-length $\delta = 0.001$ and range $(-4.1, 4.1)$. The true densities (solid line) are given for comparison. The maximal bandwidth was chosen $h_K = 3$. The plots in the bottom row show the pointwise mean average error (MAE) for SSA and kernel density estimates.

**Example 2.** In this example we consider Old Faithful Geyser data (Azzalini and Bowman, 1990), $(x_t, y_t)$, where $x_t$ measures the waiting time between successive eruptions of the geyser, and $y_t$ measures the duration of the subsequent eruption. Figures 4(left) and 4(center) displays histograms of these two variables. It is worth noting that the both are certainly non-normal. The common feature of interest is the presence of two modes. One group of eruptions is only 2 minutes in duration, while the other averages over 4 minutes in duration. Likewise, the waiting time between eruptions clusters into two groups, one less than an hour and the other greater than one hour. The distribution of eruption durations appears to be a mixture of two normal densities, but the distribution of the waiting times appears more complicated.

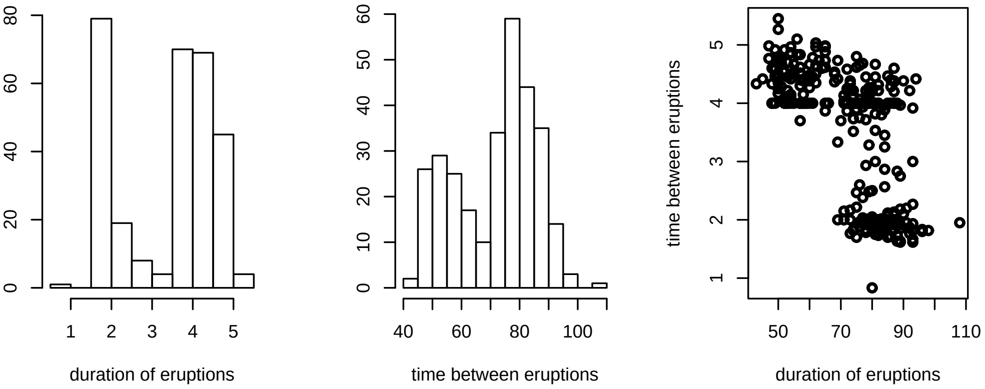

Figure 4: Marginal histograms (left and center) and the scatter plot (right) for the Old Faithful Geyser data set.

Figure 4(right) presents the scatter diagrams of $(x_t, y_t)$. The important feature of the underlying distribution is the presence of three modes. One can also easily recognize two well-separated clusters on Figure 4(right), short waiting periods are associated with long eruption durations. It is therefore desirable that the estimate of the density preserves the above features. Upper panel of Figure 5 shows respectively contour and perspective plots of the density estimate obtained by SSA procedure. The bottom row shows the same graphs for the estimate obtained by 2D Binned Kernel Density Estimation procedure (KernSmooth package in R) with suggested bandwidths. We see that the SSA density estimate underpin very well the three mode structure of the underlying data and separates two clusters, while the KDE looses the cluster structure.

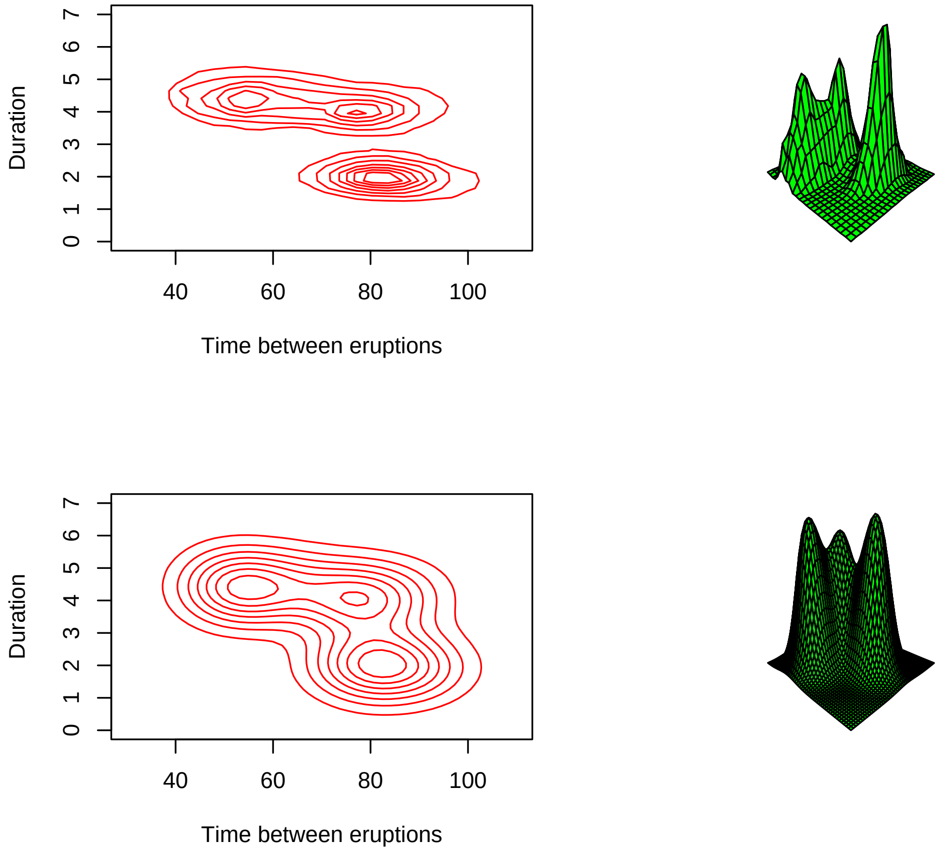

Figure 5: Contour and perspective plots of density estimates obtained by SSA procedure (top) and Bivariate Kernel Smoothing (bottom).

# 6 Application to classification

One observes a training sample $(X_i, Y_i)$, $i = 1, \ldots, n$, with $X_i$ valued in a 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 $\theta_i = \theta(X_i)$, that is, $\boldsymbol{P}(Y_i = 0 | X_i) = 1 - \theta(X_i)$ and $\boldsymbol{P}(Y_i = 1 | X_i) = \theta(X_i)$. The "ideal" Bayes discrimination rule is $\rho(x) = \mathbf{1}\left(\theta(x) \geq 1/2\right)$. Since the function $\theta(x)$ is usually unknown it is replaced by its estimate $\widehat{\theta}$. If the distribution of $X_i$ within the class $k$ has density $f_k$ then

$$\theta_i = \pi_1 f_1(X_i) / (\pi_0 f_0(X_i) + \pi_1 f_1(X_i)).$$

where $\pi_k$ the prior probability of $k$th population $k = 0, 1$.

Nonparametric methods of estimating the function $\theta$ are typically based on local averaging. Two typical examples are given by the $k$-nearest neighbor ($k$-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 of $\theta(x)$ is defined by averaging the observations $Y_i$ over $\mathcal{D}_k(x)$:

$$\widetilde{\theta}^{(k)}(x) = k^{-1} \sum_{X_i \in \mathcal{D}_k(x)} Y_i .$$

The definition of the kernel estimate of $\theta(x)$ involves a univariate kernel function $K(\cdot)$ and the bandwidth $h$:

$$\widetilde{\theta}^{(h)}(x) = \sum_{i=1}^{n} K\left(\frac{\rho(x, X_i)}{h}\right) Y_i \Big/ \sum_{i=1}^{n} K\left(\frac{\rho(x, X_i)}{h}\right).$$

Both methods require the choice of a smoothing parameter.The SSA method can be viewed as an extension of both methods using the aggregation idea. Namely, for estimating the function $\theta$ at the points $X_1, \ldots, X_n$ we can directly apply the SSA procedure to the sequence of $k$-NN (resp. kernel) estimates with an exponentially increasing number of nearest neighbors (resp. bandwidth).

**Example 1.** In this example we consider the classification problem for two class with densities $f_0(x)$ and $f_1(x)$ given by two component normal mixtures:

$$\begin{aligned} f_0(x) &= 0.2\phi(x; (-1, 0), 0.5\boldsymbol{I}_2) + 0.8\phi(x; (1, 0), 0.5\boldsymbol{I}_2) \\ f_1(x) &= 0.5\phi(x; (0, 1), 0.5\boldsymbol{I}_2) + 0.5\phi(x; (0, -1), 0.5\boldsymbol{I}_2) \end{aligned}$$

where $\phi(\cdot; \mu, \Sigma)$ is the density of multivariate normal distribution with mean vector $\mu$ and covariance matrix $\Sigma$, $\boldsymbol{I}_2$ means the $2 \times 2$ unit matrix. We run 10 simulations with 100 observations for each class in the training set and another 100 in the testing set. SSA procedure was implemented with kernel weights and parameters $h_0 = 0.1$, $h_K = 3$. Two other classification methods, $k$-NN and kernel classifiers, are applied to the same data set. Figure 6 shows the dependence of the misclassification error for these classifiers on the corresponding smoothing parameters. The misclassification errors for SSA and for the Bayes classifiers are given for comparison.

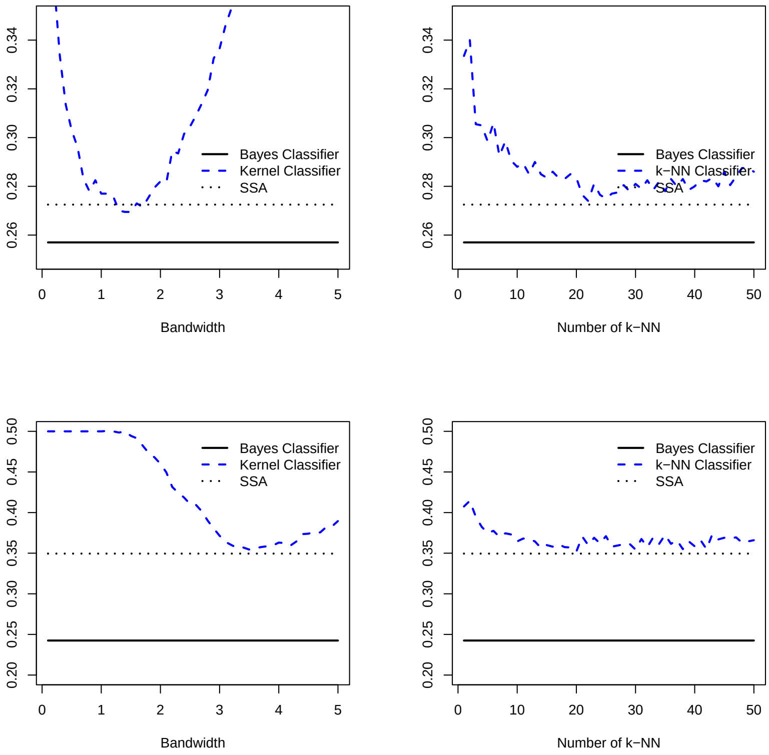

Figure 6: Misclassification errors as a functions of the main smoothing parameters for $k$-NN (right) and kernel (left) classifiers. SSA and Bayes misclassification errors are given as a reference lines. Top: Example 1 (dimension 2). Bottom: Example 2 (dimension 10).

**Example 2.** We now consider the same example but with added $8$ independent $\mathcal{N}(0,1)$ distributed nuisance components, so that now $X_i = (X_i^1, .., X_i^{10})$ with

$$(X_i^1, X_i^2) \sim f_{\text{class}(i)}, \quad (X_i^3, .., X_i^{10}) \sim \mathcal{N}((\underbrace{0, ..., 0}_{8}), \boldsymbol{I}_8).$$

The SSA procedure is applied using $k$-NN weights with the number of nearest neighbors exponentially increasing from $5$ to $50$. The results are given in the bottom row of Figure 6. We observe that although the quality of the SSA classifier has substantially decreased compared to the dimension independent Bayes error rate, it performs as good as the best $k$-NN or kernel classifier.

**Example 3.** [BUPA liver disorders] We consider the dataset sampled by BUPA Medical Research Ltd. It consists of $7$ variables and $345$ observed vectors. The subjects are single male individuals. The first $5$ variables are measurements taken

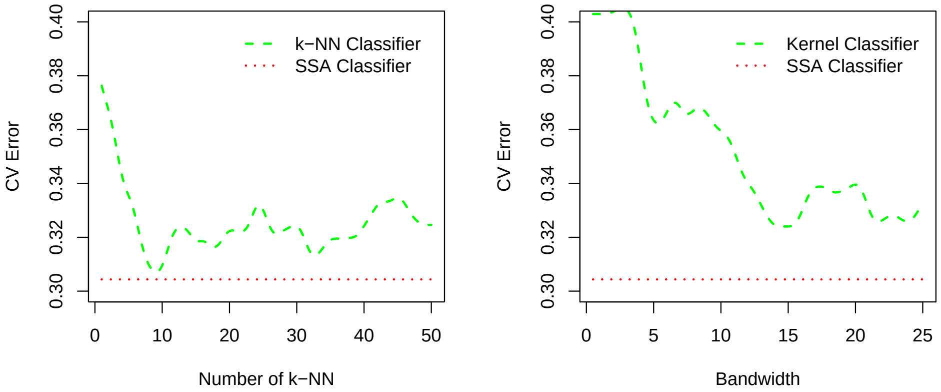

Figure 7: One-leave-out cross-validation errors as a functions of the main smoothing parameters for $k$-NN (right) and kernel (left) classifiers. The dotted line describes the error of SSA classifier.

by blood tests that are thought to be sensitive to liver disorders and might arise from excessive alcohol consumption. The sixth variable is a sort of selector variable. The seventh variable is the label indicating the class identity. Among all the observations, there are $145$ people belonging to the liver-disorder group (corresponding to selector number $2$) and $200$ people belonging to the liver-normal group. The BUPA liver disorder data set is notoriously difficult for classifying with the usual error rates about $30\%$. We apply SSA, $k$-NN and kernel classifiers to tackle this problem. In SSA procedure the hybrid weighting scheme (see section 4.1) was employed with number of $k$-NN ranging from $2$ to $30$. Figure 7 shows the corresponding one-leave-out cross-validation errors for above methods. One can see that the SSA method is uniformly better than kernel or $k$-NN classifiers.

# 7 Some theoretical properties of SSA

This section discusses some important theoretical properties of the proposed aggregating procedure. In particular we establish the "oracle" result which claims that the aggregated estimates is up to a log-factor as good as the best estimate among the considered family $\{\widetilde{\theta}^{(k)}\}$ of weak estimates. As a corollary we show rate optimality of the procedure on the smoothness function classes.

The "oracle" result is in its turn a consequence of two important properties of the aggregated estimate $\widehat{\theta}$: "monotonicity" and "stability". "Monotonicity" can be viewed as the oracle result in the homogeneous situation. In this case the oracle choice would be the estimate with the largest value $N_k$, that is, the last estimate $\widetilde{\theta}^{(K)}$ in the family $\{\widetilde{\theta}^{(k)}\}$. The "monotonicity" property means that at every step of the procedure the new estimate $\widetilde{\theta}^{(k)}$ will be taken with the weight $\gamma_k$ close to one, and hence, the aggregated estimate $\widehat{\theta}^{(k)}$ is close to the local likelihood estimate $\widetilde{\theta}^{(k)}$. The "monotonicity" property can be naturally extended to a nearly homogeneous case, for all steps $k$ for which the mean value $\overline{\theta}^{(k)} = \sum_i w_i^{(k)}\theta_i / \sum_i w_i^{(k)}$ is still close to the true value $\theta = \theta(x)$. The "monotonicity" ensures that the quality of estimation improves and confidence bounds for $\widehat{\theta}^{(k)}$ become tighter as the number of iterations increases provided that the near homogeneity is not violated. Finally, the "stability" property ensures that the quality gained under local homogeneity due to "monotonicity" will be kept for the final estimate.

Throughout this chapter $\rho$ stands for the root of equation

$$K_{\mathrm{ag}}(\rho) = (1-\nu^*)/(1-\nu^*/2), \tag{7.1}$$

where $\nu^*$ comes from A3.

## 7.1 Behavior under homogeneity

First we consider the homogeneous situation with the constant parameter value $\theta(x) = \theta$ and present some sufficient condition for the "monotonicity" result.

**Proposition 7.1.** *Assume A1 through A3 and let $\theta(\cdot) \equiv \theta$. Let the parameter $\lambda$ of the procedure fulfill $\lambda = C_\lambda \log(n)$ with a constant $C_\lambda$ such that with $\mu = 2$*

$$C_\lambda \ge \rho^{-1}\varkappa^2\bigl(\sqrt{\mu/\nu_*} + 1\bigr)^2. \tag{7.2}$$

*Then the last step estimate $\widehat{\theta} = \widehat{\theta}^{(K)}$ fulfills*

$$\boldsymbol{P}\bigl(\mathcal{K}(\widehat{\theta},\theta) > \mu\log(n)/N_K\bigr) \le 2K/n. \tag{7.3}$$

*Proof.* Define

$$\mathcal{A}^{(k)} = \{N_k\,\mathcal{K}(\widetilde{\theta}^{(k)},\theta) \le \log(n)\}.$$

Theorem 2.1 applied with $z = \log(n)$ yields in the homogeneous situation

$$\boldsymbol{P}\bigl(N_k\mathcal{K}(\widetilde{\theta}^{(k)},\theta) > \log(n)\bigr) \le 2e^{-\log(n)}.$$

Therefore $\boldsymbol{P}(\mathcal{A}^{(k)}) \geq 1 - 2/n$. This obviously implies that $\boldsymbol{P}\bigl(\mathcal{A}\bigr) \geq 1 - 2K/n$ where $\mathcal{A}$ is the intersection of the sets $\mathcal{A}^{(k)}$: $\mathcal{A} = \bigcap_{k \leq K} \mathcal{A}^{(k)}$. We now show that $\mathcal{K}(\widehat{\theta}^{(k)}, \theta) \leq \mu \log(n)/N_k$ on $\mathcal{A}$ for all $k \leq K$ which implies the assertion. The proof utilizes the following simple "metric like" property of $\mathcal{K}^{1/2}(\cdot,\cdot)$.

**Lemma 7.2.** *Under condition A2 it holds for every pair* $\theta', \theta''$

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

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

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

*Proof.* Introduce the new parameter $v = C(\theta)$ and define $D(v) = B(\theta) = B(C^{-1}(v))$. For any $v_1 \leq v_2$ it holds

$$\mathcal{K}(v_1, v_2) = D(v_2) - D(v_1) - (v_2 - v_1)D'(v_1) = 0.5|v_2 - v_1|^2 D''(\widetilde{v})$$

where $\widetilde{v} \in [v_1, v_2]$ and $D''(v) = 1/I(\theta)$ and the results easily follow from A2. □

We prove the statement by induction in $k$. By definition, it holds on $\mathcal{A}$ for $\widehat{\theta}^{(1)} = \widetilde{\theta}^{(1)}$ that $\mathcal{K}(\widehat{\theta}^{(1)}, \theta) \leq \mu \log(n)/N_1$. Now suppose that $\mathcal{K}(\widehat{\theta}^{(k-1)}, \theta) \leq \mu \log(n)/N_{k-1}$. We show that this and the condition $\mathcal{K}\bigl(\widetilde{\theta}^{(k)}, \theta\bigr) \leq \log(n)/N_k$ imply $\mathcal{K}(\widehat{\theta}^{(k)}, \theta) \leq \mu \log(n)/N_k$. The mixing penalty $\boldsymbol{m}^{(k)} = N_k \mathcal{K}(\widetilde{\theta}^{(k)}, \widehat{\theta}^{(k-1)})$ fulfills on $\mathcal{A}$ due to Lemma 7.2, Assumption A3 and (7.1)

$$\begin{aligned} \boldsymbol{m}^{(k)} &\leq N_k \varkappa^2 \bigl(\mathcal{K}^{1/2}\bigl(\widehat{\theta}^{(k-1)}, \theta\bigr) + \mathcal{K}^{1/2}\bigl(\widetilde{\theta}^{(k)}, \theta\bigr)\bigr)^2 \\ &\leq N_k \varkappa^2 \bigl(\sqrt{\mu \log(n)/N_{k-1}} + \sqrt{\log(n)/N_k}\bigr)^2 \\ &\leq \varkappa^2 \bigl(\sqrt{\mu/\nu_*} + 1\bigr)^2 \log(n) \leq \rho\lambda. \end{aligned}$$

This yields $\gamma_k = K_{\mathrm{ag}}(\boldsymbol{m}^{(k)}/\lambda) \geq K(\rho) \geq (1 - \nu_*)/(1 - \nu_*/2)$. Convexity of the Kullback-Leibler divergence $\mathcal{K}(u, v)$ w.r.t. the first argument implies on $\mathcal{A}$

$$\begin{aligned} \mathcal{K}(\widehat{\theta}^{(k)}, \theta) &= \mathcal{K}(\gamma_k \widetilde{\theta}^{(k)} + (1 - \gamma_k)\widehat{\theta}^{(k-1)}, \theta) \\ &\leq \gamma_k \mathcal{K}(\widetilde{\theta}^{(k)}, \theta) + (1 - \gamma_k)\mathcal{K}(\widehat{\theta}^{(k-1)}, \theta) \\ &\leq \gamma_k \log(n)/N_k + \mu(1 - \gamma_k)\log(n)/N_{k-1} \\ &\leq \mu \log(n)\bigl(\gamma_k/2 + (1 - \gamma_k)/\nu_*\bigr)/N_k \\ &\leq \mu \log(n)/N_k \end{aligned}$$

as required. □

**Remark 1.** The aggregated estimate $\widehat{\theta}^{(k)}$ is a convex combination of the first $k$ "weak" estimates and can be represented in the form

$$\widehat{\theta}^{(k)} = \sum_{l=1}^{k} \alpha_l^{(k)} \widetilde{\theta}^{(l)}, \quad \alpha_1^{(k)} + ... + \alpha_k^{(k)} = 1,$$

where $\alpha_l^{(k)} = \gamma_l \prod_{j=l+1}^{k}(1 - \gamma_j)$. Under homogeneity every coefficient $\gamma_l$ exceeds with a high probability the value $(1 - \nu_*)/(1 - \nu_*/2)$. Therefore, for any fixed $l$, the mixing coefficient $\alpha_l^{(k)}$ exponentially decreases as $k$ increases and the estimate $\widehat{\theta}^{(k)}$ behaves as an exponential smoothing of $\widetilde{\theta}^{l}$ for $l \leq k$.

## 7.2 Behavior under local homogeneity

In the case of local homogeneity we also have "monotonicity" as long as the "bias" measured by $\mathcal{K}^{1/2}(\overline{\theta}^{(k)}, \theta)$ with $\overline{\theta}^{(k)} := N_k^{-1} \sum_{i=1}^{n} w_i^{(k)} \theta_i$ remains sufficiently small.

**Proposition 7.3.** *Assume A1 through A3. Let $k^* \leq K$ be such that*

$$\max_{1 \leq k \leq k^*} \mathcal{K}^{1/2}(\overline{\theta}^{(k)}, \theta) \leq \delta \sqrt{\log(n)/N_{k^*}} \tag{7.4}$$

*for some $\delta > 0$. Let, a constant $\mu$ fulfill*

$$\sqrt{\mu/2} \geq \varkappa\sqrt{1+\alpha} + \varkappa\delta \tag{7.5}$$

*where $\alpha$ is defined in Theorem 2.2. Let the parameter $\lambda$ of the procedure fulfill $\lambda = C_\lambda \log(n)$ with some constant $C_\lambda$ such that*

$$C_\lambda \geq \rho^{-1}\varkappa^2\bigl(\sqrt{1+\alpha} + \sqrt{\mu/\nu_*} + \delta\bigr)^2 . \tag{7.6}$$

*Then it holds*

$$\boldsymbol{P}\left(\mathcal{K}(\widehat{\theta}^{(k^*)}, \theta) > \mu \log(n)/N_{k^*}\right) \leq 2k^*/n. \tag{7.7}$$

*Proof.* Define for $k \leq k^*$

$$\mathcal{A}^{(k)} = \{N_k \mathcal{K}(\widetilde{\theta}^{(k)}, \overline{\theta}^{(k)}) \leq (1+\alpha)\log(n)\}$$

where $\alpha$ is a constant from Theorem 2.2 and depending on $\varkappa$ from Assumption A2 only. Let also $\mathcal{A} = \bigcap_{k \leq k^*} \mathcal{A}^{(k)}$. Theorem 2.2 yields

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

Therefore $\boldsymbol{P}(\mathcal{A}) \ge 1 - 2k^*/n$. Now we check by induction that $\mathcal{K}(\widehat{\theta}^{(k^*)}, \theta) \le \mu \log(n)/N_k$ on $\mathcal{A}$ for all $k \le k^*$.

Theorem 2.2 implies (7.7) for the initial weak estimate $\widetilde{\theta}^{(1)}$. Suppose for some $k \le k^*$ that $\mathcal{K}(\widehat{\theta}^{(k-1)}, \theta) \le \mu \log(n)/N_{k-1}$. It holds on the set $\mathcal{A}^{(k)}$ by Lemma 7.2 and (7.5)

$$\begin{aligned} N_k \mathcal{K}\bigl(\widetilde{\theta}^{(k)}, \theta\bigr) &\le N_k\, \varkappa^2 \Bigl(\mathcal{K}^{1/2}\bigl(\widetilde{\theta}^{(k)}, \overline{\theta}^{(k)}\bigr) + \mathcal{K}^{1/2}\bigl(\overline{\theta}^{(k)}, \theta\bigr)\Bigr)^2 \\ &\le N_k\, \varkappa^2 \Bigl(\sqrt{(1+\alpha)\log(n)/N_k} + \delta\sqrt{\log(n)/N_k}\Bigr)^2 \\ &\le 0.5\mu \log(n). \end{aligned}$$

The use of (7.6) yields in the similar way

$$\begin{aligned} \boldsymbol{m}^{(k)} &\le N_k\, \varkappa^2 \Bigl(\mathcal{K}^{1/2}\bigl(\widetilde{\theta}^{(k)}, \overline{\theta}^{(k)}\bigr) + \mathcal{K}^{1/2}\bigl(\widehat{\theta}^{(k-1)}, \theta\bigr) + \mathcal{K}^{1/2}\bigl(\overline{\theta}^{(k)}, \theta\bigr)\Bigr)^2 \\ &\le N_k\, \varkappa^2 \log(n) \Bigl(\sqrt{(1+\alpha)/N_k} + \sqrt{\mu/N_{k-1}} + \delta\sqrt{1/N_k}\Bigr)^2 \le \rho\lambda. \end{aligned}$$

This in turn yields $\gamma_k = K_{\mathrm{st}}(\boldsymbol{m}^{(k)}/\lambda) \ge (1-\nu_*)/(1-\nu_*/2)$ and (7.7) follows in the same line as (7.3) in the proof of Proposition 7.1. □

## 7.3 Stability

Under the local homogeneity it holds with high probability $\mathcal{K}(\widehat{\theta}^{(k)}, \theta) \le \mu \log(n)/N_k$ as far as the estimation bias $\mathcal{K}^{1/2}(\overline{\theta}^{(k)}, \theta)$ remains small. If the bias starts to increase after first $k$ iterations, then the important *stability* property of the procedure is that the quality of estimation of order $(\log(n)/N_k)^{1/2}$ gained by the estimate $\widehat{\theta}^{(k)}$ will not be lost at further iterations.

**Proposition 7.4.** *Under A1 and A2, it holds for every* $k \le K$

$$\mathcal{K}\bigl(\widehat{\theta}^{(k)}, \widehat{\theta}^{(k-1)}\bigr) \le \lambda/N_k. \tag{7.8}$$

*Moreover, under A1 through A3, it holds for every* $k'$ *with* $k < k' \le K$

$$\mathcal{K}\bigl(\widehat{\theta}^{(k')}, \widehat{\theta}^{(k)}\bigr) \le c\lambda/N_k \tag{7.9}$$

*with* $c = \varkappa^2(1/\sqrt{\nu^*} - 1)^{-2}$.

**Remark 2.** 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. This property follows just from the construction of the procedure.

*Proof.* By convexity of the Kullback-Leibler divergence $\mathcal{K}(u,v)$ w.r.t. the first argument

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

If $\mathcal{K}\bigl(\widetilde{\theta}^{(k)}, \widehat{\theta}^{(k-1)}\bigr) \ge \lambda/N_k$, then $\gamma_k = 0$ and (7.8) follows. Now, Assumption A2, Lemma 7.2 and Proposition 7.4 yield

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

The use of Assumption A3 leads to the bound

$$\mathcal{K}^{1/2}\bigl(\widehat{\theta}^{(k')}, \widehat{\theta}^{(k)}\bigr) \le \varkappa\bigl(\lambda/N_k\bigr)^{1/2} \sum_{l=k+1}^{k'} (\nu^*)^{(l-k)/2} \le \varkappa\sqrt{\nu^*}(1-\sqrt{\nu^*})^{-1}\bigl(\lambda/N_k\bigr)^{1/2}$$

which proves (7.9). □

**Theorem 7.5.** *Let, for some* $k \le K$, *the estimate* $\widehat{\theta}^{(k)}$ *fulfill*

$$\mathcal{K}\bigl(\widehat{\theta}^{(k)}, \theta\bigr) \le \mu \log(n)/N_k$$

*with some constant* $\mu$. *Then it holds for the final estimate* $\widehat{\theta} = \widehat{\theta}^{(K)}$

$$\mathcal{K}\bigl(\widehat{\theta}, \theta\bigr) \le c' \log(n)/N_k,$$

*where* $c' = \varkappa^2\left(\sqrt{cC_\lambda} + \sqrt{\mu}\right)^2$ *with* $c = \varkappa^2(1/\nu^* - 1)^{-2}$ *and* $C_\lambda = \lambda/\log(n)$.

*Proof.* Proposition 7.4 and Lemma 7.2 imply

$$\mathcal{K}\bigl(\widehat{\theta}, \theta\bigr) \le \varkappa^2\left(\mathcal{K}^{1/2}\bigl(\widehat{\theta}^{(k)}, \theta\bigr) + \mathcal{K}^{1/2}\bigl(\widehat{\theta}, \widehat{\theta}^{(k)}\bigr)\right)^2 \le \varkappa^2\left(\sqrt{cC_\lambda} + \sqrt{\mu}\right)^2 \log(n)/N_k$$

and the assertion follows. □

Combining Theorem 7.5 and Proposition 2.2 yields the so called "oracle" inequality

**Corollary 7.6.** *The following inequality holds for some constant* $C \equiv C(\varkappa, \nu^*, \nu_*)$:

$$\boldsymbol{P}\left(\mathcal{K}^{1/2}(\widehat{\theta}, \theta) > C \min_{l}\left(\max_{k\le l} \mathcal{K}^{1/2}(\overline{\theta}^{(k)}, \theta) + \sqrt{\log(n)/N_l}\right)\right) \le 2K/n.$$

**Remark 3.** The first term $\max_{k\le l} \mathcal{K}^{1/2}(\overline{\theta}^{(k)}, \theta)$ in the "oracle" bound can be viewed as the upper bound for the bias of the estimate $\widetilde{\theta}^{(k)}$ while the second term $\sqrt{\log(n)/N_l}$ bounds the stochastic component of $\widetilde{\theta}^{(k)}$, so that the sum bounds the risk of the estimate $\widetilde{\theta}^{(k)}$, cf. Theorem 2.2. Therefore, the risk of the aggregated estimate corresponds to the minimal possible risk among the family $\{\widetilde{\theta}^{(k)}\}$. Lepski, Mammen and Spokoiny (1997) established a similar result in the regression setup for the pointwise adaptive Lepski procedure.

## 7.4 Rate of estimation under smoothness conditions on $\theta(\cdot)$ . Spatial adaptivity

Here we consider the case when $\theta(\cdot)$ satisfies some smoothness conditions in a neighborhood of a fixed point $x \in \mathcal{X} \subseteq I\!R^d$ . More precisely, we assume that the error of the local constant approximation of $\theta(\cdot)$ by $\theta(x)$ within this neighborhood is sufficiently small. In addition we impose some mild regularity condition on the design $X_1, \ldots, X_n$ . We show that under these assumptions the results of Proposition 7.3 and Theorem 7.5 lead to the rate of estimation $(\log(n)/n)^{1/(2+d)}$ which coincides up to a log-factor with the classical nonparametric rate of estimation corresponding to the smoothness degree one.

Let $x$ be fixed and $\theta = \theta(x)$ , $\theta_i = \theta(X_i)$ for all $i$ . Assume the following condition

**(A4)** For some $\eta > 0$ , the function $\theta(\cdot)$ fulfills

$$\mathcal{K}^{1/2}(\theta_i, \theta) \le L\eta \qquad \forall X_i \in \mathcal{B}_\eta(x).$$

Here $\mathcal{B}_h(x)$ means the ball of radius $h$ centered at $x$ . In addition, we assume that for every $k$ the weights $w_i^{(k)} = w_i^{(k)}(x)$ defining the estimate $\widetilde{\theta}^{(k)}(x)$ are supported on $\mathcal{B}_{h_k}(x)$ where the sequence of bandwidths $h_k$ grows exponentially and the number $N_k$ of design points in $\mathcal{B}_{h_k}(x)$ is nearly proportional to its volume.

**(A5)** There exists a sequence $\{h_k\}$ with $h_k = a h_{k-1}$ for some $a > 1$ such that the weights $\{w_i^{(k)}(x)\}$ satisfy

$$w_i^{(k)}(x) = 0 \quad \text{if} \quad X_i \notin \mathcal{B}_{h_k}(x).$$

**(A6)** For some positive constants $\varkappa_1 \le \varkappa_2$ and any $k \le K$ it holds

$$\varkappa_1 \le N_k/(n h_k^d) \le \varkappa_2.$$

**Theorem 7.7.** *Let* $\overline{h} = \bigl(L^2 n/\log(n)\bigr)^{-1/(2+d)}$ . *Let Assumptions A1, A2, A4, A5 and A6 be fulfilled with* $\eta = c\overline{h}$ *with some* $0 < c \le 1$ *and* $h_1 \le \eta \le h_K$ . *If the parameter* $\lambda$ *fulfills* $\lambda \ge C_\lambda \log n$ , *with* $C_\lambda = C_\lambda(\varkappa, \nu_*)$ , *then*

$$\boldsymbol{P}\Bigl(\mathcal{K}^{1/2}(\widehat{\theta}, \theta) > C_1 L^{d/(2+d)} (\log(n)/n)^{1/(2+d)}\Bigr) \le 2K/n \tag{7.10}$$

*where* $C_1$ *means a fixed constant depending on* $\varkappa$ , $\nu_*$ , $\nu^*$ , $\varkappa_1$ , $\varkappa_2$ , $\eta$ *and* $a$ *only.*

**Remark 4.** The rate of estimation given in Theorem 7.7 coincides with the optimal rate of estimation for the function smoothness class of degree one 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 $B^1_{p,q}$.

**Remark 5.** If the weights $w_i^{(k)}$ satisfy $\sum_i (X_i - x) w_i(x) \approx 0$ then the rate result can be extended to Besov function classes $B^s_{p,q}$ with $s \in [1,2]$. The latter condition on the weights is easy to check for the case of kernel weights for a regular design and a symmetric continuous kernel.

*Proof.* Let $h_1$ fulfill $h_1 \le c\overline{h}$ with some constant $c \le 1$ which we specify below. If $k$ is a maximal index such that $h_k \le c\overline{h}$ then A4 implies for every $X_i \in \mathcal{B}_{h_k}(x)$ that $\mathcal{K}^{1/2}\bigl(\theta_i, \theta\bigr) \le Lc\overline{h}$. The use of A6 yields

$$N_k \mathcal{K}\bigl(\theta_i, \theta\bigr) \le L^2 \varkappa_2 c^{2+d} n \overline{h}^{2+d} = \varkappa_2 c^{2+d} \log(n).$$

Convexity of $\mathcal{K}(\cdot, \theta)$ implies (7.4) for $\delta = (\varkappa_2 c^{2+d})^{1/2}$ and all $k^* \le k$. If $c$ is sufficiently small then Theorem 7.3 applies yielding with a high probability the following accuracy of estimating $\theta$ by $\widehat{\theta}^{(k)}$:

$$\mathcal{K}\bigl(\widehat{\theta}^{(k)}, \theta\bigr) \le \mu \log(n)/N_k \le \frac{\mu \log(n)}{\varkappa_1 n h_k^d}.$$

Since $h_k \ge c\overline{h}/a$, it holds with some fixed constant $C_2$ that

$$\mathcal{K}^{1/2}\bigl(\widehat{\theta}^{(k)}, \theta\bigr) \le C_2 L^{2d/(2+d)} \bigl(\log(n)/n\bigr)^{2/(2+d)}.$$

By Theorem 7.5, the same rate of estimation holds for the final estimate $\widehat{\theta}$. □

# 8 Some exponential bounds for exponential families

This section presents 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 2.3.

We assume the observation $Y_i$ to be $P_{\theta_i}$-distributed with $\theta_i$ 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 $\theta,\theta'$. The local MLE $\widehat{\theta}$ is given as $\widehat{\theta} = \sum_{i=1}^{n} w_i Y_i / \sum_{i=1}^{n} w_i$. We use the representation $\widehat{\theta} = S/N$ with $S = \sum_{i=1}^{n} w_i Y_i$, $N = \sum_{i=1}^{n} w_i$ and denote $\overline{\theta} = N^{-1} \sum_{i=1}^{n} w_i \theta_i$.

The result given below bounds in probability the expression $L(W,\widehat{\theta},\overline{\theta})$. It is convenient to introduce the parameter $v = C(\theta)$ and define $\overline{v} = C(\overline{\theta})$ and $D(v) = B(\theta) = B(C^{-1}(v))$. Since $C'(\theta) > 0$, the new parameter $v$ is uniquely defined. By simple analysis $D'(v) = \theta = C^{-1}(v)$ and $D''(v) = 1/C'(\theta) = 1/I(\theta) = 1/I(C^{-1}(v))$. Moreover, $\mathcal{K}(v_1,v_2) = D(v_2) - D(v_1) - (v_2 - v_1)D'(v_1)$ is the Kullback-Leibler distance between two parametric distributions corresponding to the parameters $v_1$ and $v_2$. In what follows we use the notation $q(u|v) = \mathcal{K}(v,v+u) = D(v+u) - D(v) - uD'(v)$.

**Theorem 8.1.** *Let the Fisher information $I(\theta) = C'(\theta)$ be positive on $\Theta$. For a given $z \ge 0$, let $\mathcal{U}(W,z)$ be the set of solutions $u$ of equation $q(u|\overline{v}) = \int_0^u x D''(\overline{v} + x)dx = z/N$. If there is some $\alpha > 0$ such that for all $\mu \in (0,1]$ and all $u \in \mathcal{U}(W,z)$*

$$q(\pm w_\ell \mu u | v_\ell) \le (1+\alpha) w_\ell \mu^2 q(u|\overline{v}), \qquad \ell = 1,\ldots,n, \tag{8.1}$$

*then*

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

**Remark 1.** The condition (8.1) can be easily checked in many particular situations. We give two typical examples. The first one corresponds to the homogeneous case when all $v_i$ coincide with their mean $\overline{v}$. Then (8.1) is fulfilled automatically with $\alpha = 0$. Indeed the function $q(\cdot|v)$ satisfies $q'(u|v) = D'(v+u) - D'(v)$ and $q''(u|v) = D''(v+u) = 1/I(C^{-1}(v+u)) > 0$ and thus, it is convex. Since also $q(0|v) = 0$, it holds $q(wa|v) \le wq(a|v)$ for every $w \in [0,1]$ and every $a$ implying (8.1) 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 and A2. The Taylor expansion yields that $q(wu|v) = D(v+wu) - D(v) - wuD'(v) = 1/2\, w^2 u^2 D''(v + \tau wu)$ for some $\tau \in [0,1]$. Under condition A1 $\varkappa^{-2} \le D''(v)/D''(\overline{v}) \le \varkappa^2$ for all $v$ and one easily gets for every $u \in \mathcal{U}(W,z)$ that $u^2 \le 2zI^*/N$. Therefore, the condition (8.1) is easy to check for $1 + \alpha = \varkappa^2$ which yields Theorem 2.2

as corollary of Theorem 8.1. Moreover, only the local variability of the Fisher information $I(\theta)$ on the support of the local model $W$ is important so the value $\alpha$ is typically close to zero.

**Proof of Theorem 8.1** The log-likelihood ratio can be rewritten for the new parameter $v$ as

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

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

**Lemma 8.2.** *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

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

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

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

or, equivalently,

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

Therefore

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

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

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

for some $v_* < \overline{v}$. □

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

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

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

**Lemma 8.3.** *For every $u$ and every $z$*

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

*Moreover, if $u$ fulfills (8.1) 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 $\mu$

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

The independence of the $Y_\ell$'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 $\log \int e^{v_\ell y - D(v_\ell)} P(dy) = 0$ and $\theta_\ell = D'(v_\ell)$ yield

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

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

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

This inequality applied with $\mu = (1+\alpha)^{-1}$ and (8.1) imply

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

Similarly

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

and the lemma follows. □

# Concluding remarks

The paper offers a new approach to aggregating a set of "weak" estimates in a pointwise sequential manner. The proposed procedure is very natural and appealing and it applies in a unified way to many different statistical models and problems. We established a number of remarkable theoretical properties of this procedure including the oracle result and spatial adaptivity. The results are stated under very mild conditions on the models and apply in a nonasymptotic way. The procedure also demonstrates a very reasonable numerical performance in all the simulated examples we considered. In particular, it outperforms the other smoothing methods we tried. Its practical implementation and application to many practical problems is straightforward and does not require a fine tuning of the parameters. The procedure applies in a multidimensional case for an arbitrary design.

A small limitation of the proposed procedure is the simplest method of local likelihood estimation based on the local constant approximation. A more sophisticated local polynomial approximation can deliver better results in the case of estimating a smooth function. An extension of the method to the local polynomial regression is straightforward. However, the general local likelihood approach is more difficult to study because the closed form solution of the local likelihood problem is not available. This extension is to be considered elsewhere.

# References

[1] Azzalini, A. and Bowman, A.W. (1990). A Look at Some Data on the Old Faithful Geyser. *Applied Statistics*, **39** 357365.

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

[3] Breiman, L. (1996). Stacked regression. *Machine Learning*, **24** 49–64.

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

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

[6] Catoni, O. (1999). Üniversaläggregation rules with exact bias bounds. Preprint.

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

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

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

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

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

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

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

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

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

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

[17] Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric estimation. *Ann. Statist.*, **28** 682–712.

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

[19] 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.

[20] Li, J. and Barron, A. (1999). Mixture density estimation. In. S.A. Sola, T.K. Leen, and K.R. Mueller, editors, *Advances in Neural Inforamtion proceedings systems* **12**

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

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

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

[24] Marron, J. S., Wand, M. P. (1992). Exact mean integrated squared error. *Ann. Statist.* **20** no. 2, 712–736

[25] Rigollet, Ph. and Tsybakov, A.(2005) Linear and convex aggregation of density estimators. Manuscript.

[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.

[29] Tsybakov, A. (2003) Optimal rates of aggregation. Computational Learning Theory and Kernel Machines. B.Scholkopf and M.Warmuth, eds. *Lecture Notes in Artificial Intelligence*, **2777** Springer, Heidelberg, 303-313.

[30] Yang, Y. (2001). Adaptive regression for mixing. *Journal of the American Statistical Association*, **96** 574–588.

[31] Yang, Y. (2004). Aggregating regression procedures to improve performance. *Bernoulli* **10** no. 1, 25–47