LA-UR- 09-01573

Approved for public release;
distribution is unlimited.

Title: Recursive Bias Estimation for high dimensional regression smoothers

Author(s): Nicolas Hengartner
Pierre-Andre Cornillon
Eric Matzner-Lober

Intended for: Publication in the annals of statistics

Submitted to the Annals of Statistics

RECURSIVE BIAS ESTIMATION FOR HIGH DIMENSIONAL REGRESSION SMOOTHERS

By Pierre-André Cornillon, Nicolas Hengartner and Eric Matzner-Løber

Montpellier SupAgro, Los Alamos National Laboratory and University Rennes 2

In multivariate nonparametric analysis, sparseness of the covariates also called curse of dimensionality, forces one to use large smoothing parameters. This leads to biased smoother. Instead of focusing on optimally selecting the smoothing parameter, we fix it to some reasonably large value to ensure an over-smoothing of the data. The resulting smoother has a small variance but a substantial bias. In this paper, we propose to iteratively correct of the bias initial estimator by an estimate of the latter obtained by smoothing the residuals. We examine in details the convergence of the iterated procedure for classical smoothers and relate our procedure to L_2-Boosting. For multivariate thin plate spline smoother, we proved that our procedure adapts to the correct and unknown order of smoothness for estimating an unknown function m belonging to $\mathcal{H}^{(}\nu)$ (Sobolev space where m should be bigger than $d/2$). We apply our method to simulated and real data and show that our method compares favorably with existing procedures.

1. Introduction.

Regression is a fundamental data analysis tool for uncovering functional relationships between pairs of observations (X_i, Y_i), $i = 1, \dots, n$. The traditional approach specifies a parametric family of regression functions to describe the conditional expectation of the response variable Y given the independent multivariate variables $X \in \mathbb{R}^d$, and estimates the free parameters by minimizing the squared error between the predicted values and the data. An alternative approach is to assume that the regression function varies smoothly in the independent variable x and estimate locally the conditional expectation $m(x) = \mathbb{E}[Y|X = x]$. This results in nonparametric regression estimators. We refer the interested reader to [e.g. 6, 7, 14, 15, 21, 35, 38] for a more in depth treatment of various classical regression smoothers. The vector of predicted values $\widehat{Y}_i$ at the observed covariates X_i from a nonparametric regression is called a regression smoother,

AMS 2000 subject classifications: 62G08

Keywords and phrases: nonparametric regression, smoother, kernel, nearest neighbor, smoothing splines, stopping rules

or simply a smoother, because the predicted values $\widehat{Y}_i$ are less variable than the original observations Y_i.

Operationally, linear smoothers can be written as

$$\widehat{m} = S_\lambda Y,$$

where S_λ is a $n \times n$ smoothing matrix. Smoothing matrices S_λ typically depend on a tuning parameter, which we denote by λ, that governs the tradeoff between the smoothness of the estimate and the goodness-of-fit of the smoother to the data, by controlling the effective size of the local neighborhood of the exploratory variable over which the responses are averaged. We parameterize the smoothing matrix such that large values of λ will produce very smooth curves while small λ will produce a more wiggly curve that wants to interpolate the data. For example, the tuning parameter λ is the bandwidth for kernel smoother, the span size for running-mean smoother, the number of nearest neighbors for k-nearest neighbor smoothers, and the scalar that governs the relative importance of sum of squared errors and the smoothness penalty term.

It is well known that given n uniformly distributed points in the cube $[-1, 1]^d$, the expected number of points that are covered by a ball centered at the origin with radius $\varepsilon < 1$, scales as $n\varepsilon^d$. This is to say that covariates in high dimensions are typically sparse. This phenomenon is sometimes called *the curse of dimensionality*. As a consequence, nonparametric smoothers must average over larger neighborhoods, which in turn produces more heavily biased smoothers. Optimally selecting the smoothing parameter does not alleviate this problem, and therefore, the common wisdom is to avoid general nonparametric smoothing in higher dimension and focus instead on fitting structurally constrained regression models, such as additive models [21, 22, 27].

The impact of the curse of dimensionality is lessened for very smooth regression functions. For example, regression functions with $2d$ continuous derivatives have minimax mean squared error of $n^{-4/5}$, a value recognized as the minimax mean squared error of estimates for twice differentiable univariate regression functions. The difficulty is that in practice, the smoothness of the regression function is typically unknown. Nevertheless, there are large potential gains (in terms of rates of convergence) if one considers multivariate smoothers that adapt to the smoothness of the regression function.

This paper presents a simple and intuitive procedure based on repeated application of classical multivariate linear smoothers to construct a smoother that adapts to the smoothness of the regression function. Our smoother is constructed as iteratively, starting from a pilot soother that over-smooths the data. That is, the pilot smoother has has a small variance at the cost of carrying a substantial bias. That bias can be estimated by smoothing the residuals from the pilot smoother, possibly using the same smoothing matrix, and subtracted from the pilot smoother. The bias estimation and bias correction steps can be iterated to generate a sequence of bias corrected smoothers. Section 2 discuss the behavior of that sequence, and we give conditions on the smoothing matrix which ensures convergence of that sequence of smoothers to the vector of responses Y. We propose to select a smoother from that sequence that minimizes an estimate of the prediction error, such as calculated by cross-validation or generalized cross-validation.

In Section 3, we show that this procedure applied to multivariate thin splines adapts to the smoothness of the regression function. For practical considerations, we sometimes prefer to use kernel based smothers instead of thin spline smoothers. In Section 4, we give conditions on the smoothing kernel that guarantees good behavior of the sequence of iterative bias corrected smoothers.

Beyond the nice theoretical properties of our estimator, we show in both simulated and real data that our smoother significantly improves on the prediction mean square errors over popular competing multivariate nonparametric regression models, including additive models, projection pursuit regression and MARS. For example, prediction mean squared error for the Los Angeles ozone data set ??, using our fully nonparametric smoother on eight explanatory variables, is at least 13% smaller than the competing current state-of-the-art smoothers. The gains are even more impressive for the Boston housing data ??, where the prediction mean squared error of our fully nonparametric smoother using thirteen explanatory variable is 30% smaller than its competitors.

Finally, the proofs are gathered in the Appendix.

2. Iterative bias reduction.

This section presents the general iterative bias reduction framework for linear regression smoothers.

2.1. *Preliminaries.* Suppose that the pairs $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ are related through the nonparametric regression model

$$Y_i \quad = \quad m(X_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $m(\cdot)$ is an unknown smooth function, and the disturbances ε_i are independent mean zero and variance σ^2 random variables that are independent of all the covariates $(X_1, \ldots, X_n)$. It is helpful to rewrite Equation (2.1) in vector form by setting $Y = (Y_1, \dots, Y_n)^t$, $m = (m(X_1), \dots, m(X_n))^t$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^t$, to get

$$Y = m + \varepsilon. \tag{2.2}$$

Linear smoothers can be written as

$$\widehat{m}_1 = SY, \tag{2.3}$$

where S is an $n \times n$ smoothing matrix and $\widehat{m} = \widehat{Y} = (\widehat{Y}_1, \dots, \widehat{Y}_n)^t$, denotes the vector of fitted values. Typical smoothing matrices are contractions, by virtue that the fitted values have smaller norm than the raw data, that is $||SY|| \leq ||Y||$. We refer to Buja et al. [6] for in depth discussion of such *shrinkage smoothers.*

Let I be the $n \times n$ identity matrix. The bias of the linear smoother (2.3)

$$\begin{aligned} B(\widehat{m}_1) &= \mathbb{E}[\widehat{m}_1|X] - m = (S - I)m && (2.4)\\ &= -\mathbb{E}[(I - S)Y], && (2.5)\end{aligned}$$

and its variance is

$$V(\widehat{m}_1|X) = SS'\sigma^2,$$

respectively.

2.2. *Bias reduction of Linear Smoothers.* The expression (2.5) for the bias suggests that it can be estimated by smoothing the negative residuals $-R_1 = -(Y - \widehat{m}_1) = -(I - S_1)Y$. That is,

$$\hat{b}_1 := -S_2R_1 = -S_2(I - S_1)Y \tag{2.6}$$

estimates the bias. Correcting the pilot smoother $\widehat{m}_1$ by subtracting $\hat{b}_1$ from the pilot smoother $\widehat{m}_1$ yields a *bias corrected* smoother

$$\begin{aligned} \widehat{m}_2 &= S_1Y + S_2(I - S_1)Y \\ &= (S_1 + S_2(I - S_1))Y. \end{aligned}$$

Since $\widehat{m}_2$ is itself a linear smoother, it is possible to corrected its bias as well. Repeating the bias reduction step k times produces to the linear smoother

$$\widehat{m}_k = S_1Y + S_2(I - S_1)Y + \cdots + S_k(I - S_{k-1}) \cdots (I - S_1)Y. \tag{2.7}$$

PROPOSITION 2.1 (Residual smoothing estimator). *After k iterations, the bias corrected estimator (2.7) can be explicitly written as*

$$\widehat{m}_k = [I - (I - S_k)(I - S_{k-1}) \cdots (I - S_1)]Y. \tag{2.8}$$

Remark. An alternative approach is to estimate the bias by plugging in an estimator $\tilde{m} = S_2 Y$ for the regression function m into the expression of the bias (2.4). This produces the estimator

$$\hat{b} = (S_1 - I)S_2 Y$$

for the bias. Iteratively correcting the bias using that estimate produces, after k steps, the linear smoother

$$\begin{aligned} \hat{m}_k &= S_1 Y + (I - S_1)S_2 Y + \cdot + (I - S_1)(I - S_2)\cdots S_k Y \\ &= [I - (I - S_1)(I - S_2)\dots(I - S_k)]Y. \end{aligned} \tag{2.9}$$

While in general, the these two estimates for the bias lead to distinct bias corrected smoothers (2.7) and (2.9), these two smoothers are identical when the same smoothing matrix is used at every step of the procedure. The resulting k^{th} iterated bias corrected smoother becomes

$$\hat{m}_k = [I - (I - S)^k]Y. \tag{2.10}$$

In the univariate case, smoothers of the form (2.10) arise from the L_2-boosting algorithm when setting the convergence factor μ_k of that algorithm to one. Thus we can interpret the L_2-boosting algorithm as an iterative bias reduction procedure. Breiman [3] noted a similar interpretation for the bagging algorithm applied to the residuals of nonparametric smoothers. From that interpretation, it follows that the L_2-boosting of projection smoothers, as is the case for bin smoothers and regression splines, is ineffective since the estimated bias

$$\hat{b} = S(I - S)Y = 0.$$

Bühlmann and Yu [4] present the statistical properties of the L_2-boosted univariate smoothing splines, while **?**] describes the behavior of univariate kernel smoothers after a single bias-correction iteration.

From a historical perspective, the idea of estimating the bias from residuals to correct a pilot estimator of a regression function goes back to the concept of *twicing* introduced by Tukey [36] to estimate bias of misspecified multivariate regression models. The idea of iterative debiasing regression smoothers is also present in Breiman [3] in the context of the *bagging* algorithm. More recently, the interpretation of the L_2-boosting algorithm as an iterative bias correction scheme was alluded to in Ridgeway [31]'s discussion of Friedman et al. [18] paper on the statistical interpretation of boosting. Finally, Di Marzio and Taylor [12] studied one-step bias correction of univariate kernel regression smoothers, and showed that it corresponded to making on iteration of the L_2 boosting algorithm of Bühlmann and Yu [4].

2.3. *Predictive smoothers.* As defined by (2.3), smoothers predict the conditional expectation of responses only at the design points. It is useful to extend regression smoothers to enable predictions at arbitrary locations $x \in \mathbb{R}^d$ of the covariates. Such an extension allows us to assess and compare the quality of various smoothers by how well the smoother predicts new observations.

To this end, write the prediction of the linear smoother S at an arbitrary location x as

$$\hat{m}(x) = S(x)^t Y,$$

where $S(x)$ is a vector of size n whose entries are the weights for predicting $m(x)$. The vector $S(x)$ is readily computed for many of the smoothers used in practice.

Next, writting the iterative bias corrected smoother $\widehat{m}_k$ as

$$\begin{aligned} \widehat{m}_k &= \widehat{m}_0 + \widehat{b}_1 + \cdots + \widehat{b}_k \\ &= S[I + (I-S) + (I-S)^2 + \cdots + (I-S)^{k-1}]Y \\ &= S\widehat{\beta}_k, \end{aligned}$$

it follows that we can write predict $m(x)$ by

$$\widehat{m}_k(x) = S(x)^t \widehat{\beta}_k. \tag{2.11}$$

This formulation is computationally advantageous because the vector of weights $S(x)$ only needs to be computed once, while the iterative bias correction scheme leads to the sequential update rule for the coefficients $\widehat{\beta}_k$

$$\widehat{\beta}_k = \widehat{\beta}_{k-1} + R_k,$$

where $R_k = Y - \widehat{m}_k$ is the residual vector from the previous fit.

2.4. *Convergence properties of iterative bias corrected smoothers.* The squared bias and variance of the k^{th} iterated bias corrected smoother $\widehat{m}_k$ (2.10) are

$$B^2(\hat{m}_k) \;\; = \;\; m^t \left((I-S)^k\right)^t (I-S)^k m$$

and

$$\mathrm{V}(\hat{m}_k) \;\; = \;\; \sigma^2(I-(I-S)^k)\left((I-(I-S)^k)\right)^t,$$

respectively. This shows that the qualitative behavior of the sequence of iterative bias corrected smoothers $\widehat{m}_k$ can be related to the spectrum of $I-S$. The next theorem collects the various convergence results for sequence of iterated bias corrected linear smoothers.

THEOREM 2.2. *Suppose that the singular values λ_j of $I - S$ satisfy*

$$-1 < \lambda_j < 1 \quad \textit{for} \quad j = 1, \dots, n. \tag{2.12}$$

Then we have that

$$\begin{aligned} &\|\hat{b}_k\| < \|\hat{b}_{k-1}\| \quad \textit{and} \quad \lim_{k\to\infty} \hat{b}_k = 0, \\ &\|R_k\| < \|R_{k-1}\| \quad \textit{and} \quad \lim_{k\to\infty} R_k = 0, \\ &\lim_{k\to\infty} \widehat{m}_k = Y \quad \textit{and} \quad \lim_{k\to\infty} \mathbb{E}[\|\widehat{m}_k - m\|^2] = n\sigma^2. \end{aligned}$$

Conversely, if $I - S$ has a singular value $|\lambda_j| > 1$, then

$$\lim_{k\to\infty} \|\hat{b}_k\| = \lim_{k\to\infty} \|R_k\| = \lim_{k\to\infty} \|\widehat{m}_k\| = \infty.$$

The assumption that for all j, the singular values $-1 < \lambda_j(I - S) < 1$ implies that $I-S$ is a contraction, so that $\|(I-S)Y\| < \|Y\|$. This condition however does not imply that the smoother S is itself a shrinkage smoother as defined by Buja et al. [6]. Conversely, not all shrinkage smoothers satisfy condition (2.12) of the theorem. In Section 5, we will give examples of common shrinkage smoothers for which $|\lambda_j(I - S)| > 1$, and show numerically that for these shrinkage smoothers, the iterative bias correction scheme fails. The reason of this failure lies with the fact that $\hat{b}_k$ overestimates the true bias b_k, and hence the iterative bias corrected smoother repeatedly overcorrects for the bias of the smoothers, which results in a divergent sequence of smoothers.

2.5. *Data-driven selection of the number of bias reduction steps.* Theorem 2.2 states that the limit of the sequence of iterated bias corrected smoothers is either the raw data Y or has norm $\|\hat{Y}_\infty\| = \infty$. It follows that iterating the bias correction algorithm until convergence is not desirable. However, since each iteration of the bias correction algorithm reduces the bias and increases the variance, often a few iteration of the bias correction scheme will improve upon the pilot smoother. This brings up the important question of how to decide when to stop the iterative bias correction process.

Viewing the latter question as a model selection problem suggests stopping rules for the number of iterations based on Mallows' C_p [28], Akaike Information Criteria (AIC), Akaike [1], Bayesian Information Criterion (BIC), Schwarz [33], cross-validation, L-fold cross-validation, and Generalized cross validation Craven and Wahba [10], and data splitting Hengartner et al. [23].

Each of these data-driven model selection methods estimate an optimum number of iterations k of the iterative bias correction algorithm by minimizing estimates for the expected squared prediction error of the smoothers over some pre-specified set $\mathcal{K} = \{1, 2, \ldots, M_n\}$ for the number of iterations.

We rely on the extensive literature on model selection to provide insight into the statistical properties of stopped bias corrected smoother. In particular, Theorem 3.2 of Li [26] describes the asymptotic behavior of the generalized cross-validation (GCV) stopping rule applied to smoothers. Results on the finite sample performance for data splitting for arbitrary smoothers is given in Theorem 1 of Hengartner et al. [23]. In nonparametric smoothing, the AIC criteria has a noticeable tendency to select more iterations than needed, leading to a final smoother $\widehat{m}_{\widehat{k}_{AIC}}$ that typically undersmooths the data. As a remedy, Hurvich et al. [25] introduced a corrected version of the AIC under the simplifying assumption that the nonparametric smoother $\widehat{m}$ is unbiased, which is rarely hold in practice and which is particularly not true in our context.

Extensive simulations of the above mentioned model selection criteria, both in the univariate and the multivariate settings [8] have shown that GCV

$$\hat{k}_{GCV} = \operatorname*{argmin}_{k \in \mathcal{K}} \left\{ \log \widehat{\sigma}^2 - 2 \log \left(1 - \frac{\operatorname{trace}(S_k)}{n} \right) \right\}$$

is a good choice, both in terms of computational efficiencies and of producing good final smoothers.

3. Iterative bias reduction of multivariate thin-plate splines smoothers.

In this section, we study the statistical properties of the iterative bias reduction of multivariate thin-plate spline smoothers. Given a smoothing parameter λ, the thin-plate smoother of degree ν_0 minimizes

(3.1)

$$\min_f \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \left[\sum_{\substack{i_1, \ldots, i_d = 0 \\ i_1 + \cdots + i_d \leq \nu_0}} \int_{\mathbb{R}^d} \left| \frac{\partial^{i_1 + \cdots + i_d}}{\partial x_{i_1} \ldots \partial x_{i_{\nu_0}}} f(x) \right|^2 dx \right].$$

Thin-plate smoothing splines are attractive class of multivariate smoothers for two reasons: First, the solution of (3.1). once cast within a Reproducing Kernel Hilbert Space (RKHS) framework see Gu [20], is numerically tractable and second, the eigenvalues of the smoothing matrix are approximatively known (c.f. Utreras [37]).

3.1. *Numerical Example.* It is easy to establish that the eigenvalues of the associated smoothing matrix lie between zero and one. In light of Theorem 2.2, the sequence of bias corrected thin-plate spline smoothers, starting from a pilot that oversmooths the data, will converge to an interpolant of the raw data. As a result, we anticipate that after some suitable number of bias correction steps, the resulting bias corrected smoother will be a good estimate for the true underlying regression function.

FIG 1. *True regression function $m(x_1, x_2)$ (3.2) on the square $[-10, 10] \times [-10, 10]$ used in our numerical examples.*

This behavior is confirmed numerically in the following pedagogical example of a bivariate regression problem: Figure 1 graphs Wendelberger's test function Wendelberger [39]

$$
\begin{aligned}
m(x_1, x_2) \;=\;& \frac{3}{4}\exp\left\{-((9x-2)^2+(9y-2)^2)/4\right\} + \\
&+\frac{3}{4}\exp\left\{-((9x+1)^2/49+(9y+1)^2/10)\right\} + \\
&+\frac{1}{2}\exp\left\{-((9x-7)^2+(9y-3)^2)/4)\right\} - \\
&-\frac{1}{5}\exp\left\{-((9x-4)^2+(9y-7)^2)\right\} \qquad (3.2)
\end{aligned}
$$

that is sampled at 100 locations on the regular grid $\{0.05, 0.15, \ldots, 0.85, 0.95\}^2$. The disturbances are mean zero Gaussian with variance producing a signal to noise ratio of five. Figure 2 shows the evolution of the bias corrected smoother, starting from a nearly linear pilot smoother in panel (a). After

500 iterative bias reduction steps, the smoother shown in panel (b) is visually close to the original regression function. Continuing the bias correction scheme will eventually lead to a smoother that interpolates the raw data. To illustrate this, we show the bias corrected smoother after 50000 iterations in panel (c). Notice how noisy that estimator is, compared to the one in panel (b). This example shows the importance of suitably selecting the number of bias correction iterations.

FIG 2. *Thin-plate spline regression smoothers from* 100 *noisy observations from 3.2 (see Figure 1) evaluated on a regular grid on* $[-10,10]\times[-10,10]$. *Panel (a) shows the pilot smoother, panel (b) graphs the bias corrected smoother after 500 iterations and panel (c) graphs the smoother after 50000 iterations of the bias correction scheme.*

3.2. *Adaptation to smoothness of the regression function.* Let Ω be an open bounded subset of $\mathbb{R}^d$ and suppose that the unknown regression function m belongs to the Sobolev space $\mathcal{H}^{(\nu)}(\Omega)=\mathcal{H}^{(\nu)}$, where ν is an integer such that $\nu > d/2$. Let S denote the smoothing matrix of a thin-plate spline of order $\nu_0 \leq \nu$ (in practice we will take the smallest possible value $\nu_0 = \lceil d/2 \rceil$) and fix the smoothing parameter $\lambda_0 > 0$ to some reasonably large value. Our next theorem states that there exists a number of bias reduction steps $k = k(n)$, depending on the sample size, for which the resulting estimate $\widehat{m}_k$ achieves the minimax rate of convergence. In light of that theorem, we expect that an iterative bias corrected smoother, with the number of iterations selected by GCV, will achieve the minimax rate of convergence.

THEOREM 3.1. *Assume that the design* $X_i \in \Omega$, $i = 1,\dots,n$ *satisfies the following assumption: Define*

$$h_{max}(n) = \sup_{x\in\Omega}\inf_{i=1,\dots,n} |x - X_i|, \text{ and } h_{min}(n) = \min_{i\neq j} |X_i - X_j|,$$

and assume that there exists a constant $B > 0$ *such that*

$$\frac{h_{max}(n)}{h_{min}(n)} \leq B \quad \forall n.$$

Suppose that the true regression function $m \in \mathcal{H}^{(\nu)}$.

If the initial estimator $\hat{m}_1 = SY$ is obtain with S a thin-plate spline of degree ν_0, with $\lceil d/2 \rceil \leq \nu_0 < \nu$ and a fixed smoothing parameter $\lambda_0 > 0$ not depending on the sample size n, then there is an optimal number of bias reduction steps $k(n)$ such that the resulting smoother $\hat{m}_k$ satisfies

$$\mathbb{E}\left[\left(\frac{1}{n}\sum_{j=1}^{n}(\hat{m}_k(X_j) - m(X_j)\right)^2\right] = O\left(\frac{1}{n^{2\nu/(2\nu+d)}}\right),$$

which is the optimal minimax rate of convergence for $m \in \mathcal{H}^{(\nu)}$.

Remark. Rate optimality of the smoother $\hat{m}_k$ is achieved by suitable selection of the number of bias correcting iterations, while the smoothing parameter λ_0 remains unchanged. That is, the effective size of the neighborhoods the smoother averages over remains constant.

THEOREM 3.2. *Suppose that the noise ε in (2.1) has finite tenth moment, that is, $\mathbb{E}[\varepsilon^{10}] < \infty$. Let $\hat{k}_{GCV} \in \mathcal{K}_n = \{1, \ldots, n^\alpha\}$, $\alpha \geq 1$, denote the index in the sequence of bias corrected smothers whose associated smoother minimize the generalized cross-validation criteria. Then as the sample size n grows to infinity,*

$$\frac{\|\hat{m}_{\hat{k}_{GCV}} - m\|^2}{\inf_{k\in\mathcal{K}_n}\|\hat{m}_k - m\|^2} \longrightarrow 1, \quad \textit{in probability.}$$

While adaptation of the L_2-boosting algorithm applied to univariate smoothing splines was proven by [4], the application of bias reduction to achieve adaptation to the smoothness of multivariate regression function has not been previously exploited. The practical importance of our procedure is revealed in both our simulation study and our analysis of classical multivariate test datasets in Section 5. In both instances, our method makes substantially better predictions over state-of-the-art structural models such as additive regression smoothing, MARS, and projection pursuit regression.

4. Iterative Bias reduction of Kernel smoothers. The smoothing matrix S of thin plate spline is symmetric and has eigenvalues in $(0, 1]$ are known (see for example [37]). In particular, the first $M_0 = \binom{\nu_0+d-1}{\nu_0-1}$ eigenvalues are all equal to one. This feature limits the practical usefulness of thin plate spline smoothers in iterative bias correction schemes.

Recall that our procedures requires a heavily biased pilot smoother. One measure of the smoothness (and hence of the bias) of the pilot smoother is the trace of the smoothing matrix, often called the effective degree of freedom of the smoother, with large effective degrees of freedom associated with noisy smoothers and small effective degrees of freedom associated with smooth smoothers. In light of Theorem ??, we want $\nu_0 > d/2$. It follows that the effective degree of freedom increases with the dimension. In particular the effective degree of freedom is at least $5, 28, 165, 1001$ for $d = 4, 6, 8, 10$, respectively.

A possible resolution to this problem is to approximate the thin plate spline smoother with a kernel smoother, with an appropriate kernel. See Messer [29], Silverman [34] for example. In this section, we discuss kernel based smoothers in general, and we give a necessary and sufficient condition on the kernel that ensures that the iterative bias correction scheme is well behaved. We supplement our theorems with numerical examples of both good and bad behavior of our scheme.

4.1. *Kernel type smoothers.* The smoothing matrix S of Nadaraya kernel type estimators has entries $S_{ij} = K(d_h(X_i, X_j))/\sum_k K(d_h(X_i, X_j))$. where $K(.)$ is typically a symmetric function in $\mathbb{R}$ (e.g., Uniform, Epanechnikov, Gaussian), and $d_h(x, y)$ is a weighted distance between two vectors $x, y \in \mathbb{R}^d$. The particular choice of the distance $d(\cdot, \cdot)$ determines the shape of the neighborhood. For example, the weighted Euclidean norm

$$d_h(x, y) = \sqrt{\sum_{j=1}^{d} \frac{(x_j - y_j)^2}{h_j^2}},$$

where $h = (h_1, \ldots, h_d)$ denotes the bandwidth vector, gives rise to elliptic neighborhoods.

4.2. *Spectrum of kernel smoothers.* To apply Theorem 2.2, we need to characterize the spectrum of $I - S$. While the smoothing matrix S is not symmetric, it has a real spectrum. To see this, write $S = D\mathbb{K}$, where $\mathbb{K}$ is symmetric matrix with general element $\mathbb{K}_{ij} = K(d_h(X_i, X_j))$ and D is diagonal matrix with elements $D_{ii} = 1/\sum_j K(d_h(X_i, X_j))$. If q is an eigenvector of S associated to the eigenvalue λ, then

$$Sq = D\mathbb{K}q = D^{1/2}\left(D^{1/2}\mathbb{K}D^{1/2}\right)D^{-1/2}q = \lambda q,$$

and hence

$$\left(D^{1/2}\mathbb{K}D^{1/2}\right)\left(D^{-1/2}q\right) = \lambda\left(D^{-1/2}q\right).$$

Hence the symmetric matrix $A = D^{1/2}\mathbb{K}D^{1/2}$ has the same spectrum as S. Since S is row-stochastic, all its eigenvalues are bounded by one. Thus, in light of Theorem 2.2, we seek conditions on the kernel K to ensure that its spectrum is non-negative. Necessary and sufficient conditions on the smoothing kernel K for S to have a non-negative spectrum are given in the following theorem.

THEOREM 4.1. *If the inverse Fourier-Stieltjes transform of a kernel $K(\cdot)$ is a real positive finite measure, then the spectrum of the Nadaraya-Watson kernel smoother lies between zero and one.*

Conversely, suppose that $X_1, \dots, X_n$ are an independent n-sample from a density f (with respect to Lebesgue measure) that is bounded away from zero on a compact set strictly included in the support of f. If the inverse Fourier-Stieltjes transform of a kernel $K(\cdot)$ is not a positive finite measure, then with probability approaching one as the sample size n grows to infinity, the maximum of the spectrum of $I - S$ is larger than one.

Remark 1: The assumption that the inverse Fourier-Stieltjes transform of a kernel $K(\cdot)$ is a real positive finite measure is equivalent to the kernel $K(\cdot)$ being positive a definite function, that is, for any finite set of points $x_1, \dots, x_m$, the matrix

$$\begin{pmatrix} K(0) & K(d_h(x_1,x_2)) & K(d_h(x_1,x_3)) & \dots & K(d_h(x_1,x_m)) \\ K(d_h(x_2,x_1)) & K(0) & K(d_h(x_2,x_3)) & \dots & K(d_h(x_2,x_m)) \\ \vdots & & & & \vdots \\ K(d_h(x_m,x_1)) & K(d_h(x_m,x_2)) & K(d_h(x_m,x_3)) & \dots & K(0) \end{pmatrix}$$

is positive definite. We refer to Schwartz [32] for a detailed study of positive definite functions.

Remark 2: Di Marzio and Taylor Di Marzio and Taylor [13] proved the first part of the theorem in the context of univariate smoothers. Our proof of the converse shows that for large enough sample sizes, most configurations from a random design lead to smoothing matrix S with negative singular values.

4.2.1. *Numerical implementation.* Iterative smoothing of the residuals can be computationally burdensome. To derive an alternative, and computationally more efficient representation of the iterative bias corrected smoother, observe that

$$\begin{aligned}\hat{m}_k &= [I-(I-S)^k]Y \\ &= [I-(D^{1/2}D^{-1/2}-D^{1/2}D^{1/2}\mathbb{K}D^{1/2}D^{-1/2})^k]Y \\ &= [I-D^{1/2}(I-D^{1/2}\mathbb{K}D^{1/2})^k D^{-1/2}]Y \\ &= D^{1/2}[I-(I-A)^k]D^{-1/2}Y.\end{aligned}$$

Writing

$$A = D^{1/2}\mathbb{K}D^{1/2} = P_A\Lambda_A P_A^t,$$

where P_A is the orthonormal matrix of eigenvectors and Λ_A diagonal matrix of their associated eigenvalues, we obtain a computationally efficient representation for the smoother

$$\hat{m}_k = D^{1/2}P_A[I-(I-\Lambda_A)^k]P_A^t D^{-1/2}Y.$$

Note that the eigenvalue decomposition of A needs only to be computed once, and hence leads to a fast implementation for calculating the sequence of bias corrected smoothers.

4.2.2. *Example of Gaussian kernel smoother.* The Gaussian and triangular kernels are positive definite kernels (they are the Fourier transform of a finite positive measure Feller [16]). In light of Theorem 4.1 the iterative bias correction of Nadaraya-Watson kernel smoothers with these kernels produces a sequence of well behavior smoother.

The anticipated behavior of iterative bias correction for Gaussian kernel smoothers is confirmed in our numerical example. Figure 3 shows the progression of the sequence of bias corrected smoothers starting from a very smooth surface (see panel (a)) that is nearly constant. Fifty iterations (see panel (b)) produces a fit that is visually similar to the original function. Continued bias corrections then then slowly degrades the fit as the smoother starts to over-fit the data. Panel (c) show the smoother after 10000 iterations. Continuing the bias correction scheme will eventually lead to a smoother that interpolates the data. This examples hints at the potential gains that can be realized by suitably selecting the number of bias correction steps.

FIG 3. *Gaussian kernel smoother of the function $m(x_1, x_2)$ from $n = 100$ equidistributed points on $[-10, 10] \times [-10, 10]$, evaluated on a regular grid with $k = 1$ iteration (a), 50 iterations (b) and 10000 iterations (c).*

4.2.3. *Kernel smoothers with Uniform and Epanechnikov kernels.* The Uniform and the Epanechnikov kernels are not positive definite. Theorem 4.1 states that for large enough samples, we expect with high probability that $I - S$ has at least one eigenvector larger than one. When this occurs, the sequence of iterative bias corrected smoothers will behave erratically and eventually diverge. Proposition 4.2 below strengthens this result by giving an explicit condition on the configurations of the design points for which the largest singular value of $I - S$ is always larger than one.

PROPOSITION 4.2. *Denote by $\mathcal{N}_i = \{X_j : K(d_h(X_j, X_i)) > 0\}$ the the set of distinctive points in the neighbors of X_i.*

If there exists a set $\mathcal{N}_i$ such that $|\mathcal{N}_i| \geq 3$ that contains points $X_j, X_k \neq X_i$ such that $d_h(X_i, X_j) < 1$, $d_h(X_i, X_k) < 1$ and $d_h(X_j, X_k) > 1$, then the smoothing matrix S for the Uniform kernel smoother has at least one negative eigenvalue.

If there exits a set $\mathcal{N}_i$ such that $|\mathcal{N}_i| \geq 3$ that contains points $X_j, X_k \neq X_i$ that satisfy

$$d_h(X_j, X_k) > \min\{d_h(X_i, X_j), d_h(X_i, X_k)\},$$

then the smoothing matrix S for the Epanechnikov kernel smoother has at least one negative eigenvalue.

Remark. The proof of the proposition is readily adapted to multivariate kernel smoothers whose kernel are defined as the product of univariate kernel in each of the components.

The failure of the iterated bias correction scheme using Epanechnikov kernel smoothers is illustrated in the numerical example shown in Figure 4. As for the Gaussian smoother, the initial smoother (panel (a)) is nearly constant. After five iterations (panel (b)) some of the features of the *Mexican hat* become visible. Continuing the bias corrections scheme produces an unstable smoother. Panel (c) shows that after only 25 iterations, the smoother becomes noisy. Nevertheless, when comparing panel (a) with panel (b), we see that some improvement is possible from a few iterations of the bias reduction scheme.

FIG 4. *Epanechnikov kernel smoother of the function* $m(x_1, x_2)$ *from* $n = 100$ *equidistributed points on* $[-10, 10] \times [-10, 10]$, *evaluated on a regular grid with* $k = 1$ *iteration (a), 5 iterations (b) and 25 iterations (c).*

5. Simulations and real example. In this section, we show that our proposed iterative correction procedure works well for both simulated and real data. To provide a baseline for comparison, we first study the performance of iterative bias corrected univariate smoothing splines. We then proceed to show that our method has desirable finite sample properties in the multivariate setting and compares advantageously when applied to the well known the Los Angeles Ozone data. All the numerical examples were computed using the **ibr** R-package [9], freely available at URL: http://www.uhb.fr//sc_sociales/labstats/EML/ibr_0.01.tar.gz.

5.1. *Selecting the smoothing parameter.* An important question is how to chose the bandwidth of smoother. We know that for bias reduction to be effective, we want to use a large bandwidth that oversmooths the responses, as such pilot smoothers will be heavily biased. As a general rule, the larger the bandwidth, the more biased the pilot smoother will be and the more iterations of the bias reduction scheme will be required to obtain a "good" smoother. Otherwise, the method is generally robust to the choice of the bandwidth.

The bandwidth in each component of the covariate depends on its scale. It is common to first rescale the data before selecting the bandwidth. In our numerical experiments, we found it preferable to leave the scales unchanged, and to select the bandwidth based on the effective degree of freedom (trace of the smoothing matrix) of the univariate smoother in each of the components, with typical values for the degree of freedom we ranging from 1.05 to 1.2. A further advantage of the latter choice is that there is no explicit reference to sample size.

5.2. *Univariate case.* Smoothing splines are a popular smoothers with good asymptotic and finite sample behavior. Here, we show the benefits of applying the iterative bias correction scheme to smoothing splines. To do so, we compare the iterative smoother using two different starting points, two different stopping rules (GCV and Cross Validation) and three sample size $n = 50$, 100 and 500. We calculate the smoothing spline with the R-function smooth.spline, which we apply to data from the following three regression functions

$$\begin{aligned} m_1(x) &= \sin(5\pi x) \\ m_2(x) &= 1 - 48x + 218x^2 - 315x^3 + 145x^4 \\ m_3(x) &= \exp(x - \frac{1}{3})\{x < \frac{1}{3}\} + \exp[-2(x - \frac{1}{3})]\{x \geq \frac{1}{3}\}. \end{aligned}$$

The explanatory variable X is a Uniform[0, 1] distributed random variable, an error (Gaussian or Student 5) with variance such that the signal to noise ratio is 80%. For 100 replications, we calculate on a finite grid in [0, 1] the quadratic error between the true function and the proposed estimate. Table (1) reports the median over the 100 replications of the ratio of the error obtained but the iterative estimator and the smoothing spline estimator.

TABLE 1

Median over 100 simulations of the number of iterations and median of the ratio of the MSE obtained by the iterative debiasing estimation and the MSE obtained by the smoothing splines smoother for $n = 50$ data points.

error	$\hat{k}_{1\mathrm{GCV}}$	$S_{\hat{k}_{1\mathrm{GCV}}}$	$\hat{k}_{2\mathrm{GCV}}$	$S_{\hat{k}_{2\mathrm{GCV}}}$	$\hat{k}_{1\mathrm{CV}}$	$S_{\hat{k}_{1\mathrm{CV}}}$	$\hat{k}_{2\mathrm{CV}}$	$S_{\hat{k}_{2\mathrm{CV}}}$
Function $m_1(x) = \sin(5\pi x)$								
Gaussian	4077	0.86	65	0.88	4191	0.84	88	0.83
Student	4115	0.87	70	0.88	4853	0.84	96	0.84
Function $m_2(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$								
Gaussian	1219	1.09	21	1.12	1339	1.07	27	1.10
Student	1307	1.11	22	1.13	1714	1.07	30	1.09
Function $m_3(x) = \exp(x - \frac{1}{3})\{x < \frac{1}{3}\} + \exp[-2(x - \frac{1}{3})]\{x \geq \frac{1}{3}\}$								
Gaussian	135	0.93	3	0.93	138	0.92	3	0.93
Student	147	0.95	3	0.97	156	0.94	3	0.94

Each entry in the table reports the median number of iterations and the median of the ratio of the MSE obtained by the iterative debiasing estimation and the MSE obtained by the smoothing splines smoother for $n = 50$ data points. As expected. larger smoothing parameter of the initial smoother requires more iterations of the iterated algorithm to reach its optimum. Interestingly, the selected smoother starting with a very smooth smoother, has slightly smaller mean squared error. In some cases, the iterative bias correction has smaller mean squared error than the *one-step* smoother, with improvements ranging from 5% to 15%.

5.3. *Multivariate case.* Here, we focus on the multivariate case $X \in \mathbb{R}^d$, $d > 1$, and consider multivariate Gaussian kernel smoothers. Statistical lore discourages using fully nonparametric methods in higher dimensions because estimators suffer from the curse of dimensionality. Instead, of focuses on estimating structurally constrained regressions models, such as additive models, multiplicative models, or multivariate tensor product of spline basis in low dimension such as MARS that have better statistical properties at the cost of possible misspecification error.

Next we show via simulations that the iterative bias correction scheme using a fully nonparametric regression smoother compares advantageously to the MARS algorithm of Friedman [17] and additive models using the backfitting algorithm of Hastie and Tibshirani [21]. For illustration, we consider fitting the following test function

$$m(x) \quad = \quad 10 \sin(\pi x_1 x_2) + 20(x_3 - .5)^2 + 10x_4 + 5x_5.$$

previously used by [17]. As in that paper, the covariate are independent Uniform distributions in each of the five variables, and Gaussian disturbances with small variance[1] were added to the response surface $m(x)$.

For each sample size ($n = 50, 100, 200$), we generate the data as above, use 90% of the data as a training set and predict the remaining 10% with the R package **mda** for MARS and R package **mgcv** for the additive model $m_1(x_1) + \cdots + m_5(x_5)$ without interaction. We compare the prediction mean square error of these methods with our iterative bias reduction scheme using a Gaussian kernel regression smoother with three choices of bandwidths chosen such that the effective degree of freedom for each covariate is 1.05, 1.1, and 1.2. The results we report in Table 2 are over 100 replications of the simulation.

[1] the variance is such that the signal to noise ratio is 95%.

TABLE 2

Median over 100 simulations. with the median of iteration between parenthesis.

	$n = 50$	$n = 100$	$n = 200$
MARS	4.298	3.746	3.354
Add.	3.577	3.314	2.753
PPR	6.925 (1)	6.213	4.411
$\text{BR}_{ddl=1.50}$	3.551 (31)	2.239 (42)	1.842 (57)
$\text{BR}_{ddl=1.20}$	3.122 (290)	2.014 (589)	1.747 (960)
$\text{BR}_{ddl=1.10}$	3.179 (1740)	1.967 (5084)	1.707 (8780)
$\text{BR}_{ddl=1.05}$	3.226 (14140)	1.970 (42930)	1.680 (76820)
TPSi	3.89 (431)	2.483 (361)	1.988 (287)

The pattern is similar to that seen in univariate simulations are found. First, the smoother the pilot estimator, the bigger the number of iterations chosen by GCV. Second, the smoother the pilot estimator, the better the results. But here, for very small datasets ($n = 50$ data points and $d = 5$ variables) the smoothest pilot estimator (df 1.05) tried leads to results that are worse than the second smoothest (df 1.10). MSE obtained with $n = 100$ using these pilot estimators are nearly the same, whereas MSE obtained with $n = 200$ shows that the smoothest leads to the best results. Simulations in univariate settings with very wiggly curve (not shown in this paper) have shown similar results: if the pilot smoother is too smooth (near to the constant), it cannot capture the whole unknown function as well as a less smooth pilot smoother.

5.4. *Los Angeles Ozone Data.* We consider the classical data set of ozone concentration in the Los Angeles basin which is as a standard dataset comparing the performance of multivariate smoothers (Breiman [2], Bühlmann and Yu [4, 5]). The sample size of the data is $n = 330$ and the number of explanatory variables $d = 8$. We use a multivariate Gaussian kernel and select the bandwidth so that the univariate smoother in each of the variables has the same trace, i.e., the same effective degree of freedom. These are chosen equal to 1.05, 1.1, 1.2 and 1.5 in order to investigate the influence of such parameter. We compare our iterative bias procedure with Mars using R package **mda**, with additive models estimation using R package **mgcv** and L_2-Boosting proposed by Bühlmann and Yu [4], which we recall here. Multivariate L_2-Boosting proposed by Bühlmann and Yu [4] leads to component-wise additive model

$$\hat{m}_k^{\text{boost}} = \hat{\mu} + \sum_{j=1}^{d} \hat{f}^{[k],(j)}(x_j),$$

where the component $\acute{f}^{[k],(j)}$ is obtained by choosing the univariate smoother $S_{\lambda_j}(X_j)$ which leads to the best improvement in smoothing the residuals of previous iteration $k-1$.

The estimate mean squared prediction error is obtain by randomly splitting the data into 297 training and 33 test observations and averaging 50 times over such random partitions. We use the same configuration as Bühlmann and Yu [4] and reporting theirs results we obtain the following table :

TABLE 3
Predicted mean Squared Error on test observations of ozone data for different methods.

Method	Mean Predicted Squared Error
L_2Boost with component-wise spline	17.78
additive model (backfitted with R)	17.44
Projection pursuit (with R)	16.89
MARS (with R)	17.49
iterative bias reduction with GCV stopping rule	
and multivariate Gaussian kernel with	
1.05 initial DDL per variable and **297** iterations	14.85
1.1 initial DDL per variable and **64** iterations	14.83
1.2 initial DDL per variable and **15** iterations	14.86
1.5 initial DDL per variable and **3** iterations	14.98

We can see (table 3) that, as in univariate setting, the smoother the pilot estimator is, the better the final estimation is, at the cost of increasing computation time. The combination of iterated of GCV and bias corrected estimator leads to a diminution of more than 12% over other multivariate methods.

5.5. *Boston housing data.* We apply our method on the Boston housing data. This dataset, cearted by [?] has been extensively analyzed see for example [?] or more recently by [13]. The data contains 506 census tracts in the Boston area taken from 1970 census and each instance has 13 explanatory variables (1 is binary and the explanatory variable is the median value of owner-occupied homes in $1000's. The sample size of the data is $n = 506$ and the number of explanatory variables $d = 13$. We use here a multivariate Gaussian kernel and select each individual bandwidth in order to have the same degree of freedom by variable. These are chosen equal to 1.05, 1.1, 1.2 and 1.5 in order to investigate the influence of such parameter. We compare our iterative bias procedure with a classical multivariate regression model, with Mars using R package **mda**, with additive models estimation using R package **mgcv** and with projection pursuit regression.

The estimate mean squared prediction error is obtain by randomly splitting the data into 350 training and 156 test observations and averaging 50 times over such random partitions. We obtain the following table :

TABLE 4
Predicted mean Squared Error on test observations of Boston housin data for different methods.

Method	Mean Predicted Squared Error
additive model (backfitted with R)	11.77
Projection pursuit (with R)	11.98
MARS (with R)	10.54
Multivariate regression	20.09
iterative bias reduction with GCV stopping rule and multivariate Gaussian kernel with	
1.1 initial DDL per variable and **1230** iterations	7.25
1.2 initial DDL per variable and **253** iterations	6.75
1.5 initial DDL per variable and **29** iterations	6.97

We can see (table 4) that, the combination of of GCV and iterated bias corrected estimator leads to a diminution of more than 30% of the prediction mean squared error over other multivariate methods (and more than 40% reduction if we transforme the Y on the log scale).

6. Discussion. In this paper, we make the connection between iterative bias correction and the L_2 boosting algorithm, thereby providing a new interpretation for the latter. A link between bias reduction and boosting was suggested by Ridgeway [31] in his discussion of the seminal paper Friedman et al. [18], and explored in Di Marzio and Taylor [11, 12] for the special case of kernel smoothers. In this paper, we show that this interpretation holds for general linear smoothers.

It was surprising to us that not all smoothers were suitable for boosting. Our results extend and complement the recent results of Di Marzio and Taylor [12]. We show that many weak learners, such as the k-nearest neighbor smoother and some kernel smoothers, are not stable under iterated bias estimation. We conjecture that positive defined kernels can equivalent kernels of an appropriate L-spline Ramsay and Heckman [30].

Iterating the bias correction scheme until convergence is *not* desirable. Better smoothers result if one stops the iterative scheme. Both in our simulations and application to real data, we show good performance of our method for high dimensional smoothers, even for moderate sample sizes.

As a final remark, note that one does not need to keep the same smoother throughout the iterative bias correcting scheme. We conjecture that there are advantages to using weaker smoothers later in the iterative scheme, and shall investigate this in a forthcoming paper.

APPENDIX A: APPENDIX

Proof of Theorem 2.2

$$\begin{aligned} \|\hat{b}_k\|^2 &= \| -(I-S)^{k-1}SY\|^2 \\ &= \|(I-S)(I-S)^{k-2}SY\|^2 \leq \|(I-S)\|^2\|\hat{b}_{k-1}\|^2 \\ &\leq \|\hat{b}_{k-1}\|^2, \end{aligned}$$

where the last inequality follows from the assumptions on the spectrum of $I-S$. Similarly, one shows that

$$\|R_k\|^2 = \|(I-S)^kY\|^2 \leq \|I-S\|^2\|R_{k-1}\|^2 < \|R_{k-1}\|^2.$$

Proof of Theorem 3.1 Let $\nu_0 < \nu$ and fix the smoothing parameter λ_0. Define $S = S_{\nu_0,\lambda_0}$. The eigen decompostion of S (Utreras, 1988) gives

$$\lambda_1 = \cdots = \lambda_{M_0} = 1 \quad \text{and} \quad \lambda_j \approx \frac{1}{1+\lambda_0 j^{2\nu_0/d}},$$

where $M_0 = C_{d+\nu_0-1}^{\nu_0-1}$. Let us evaluate the variance:

$$V(\hat{m}_k, \lambda_0, \nu_0) = \sigma^2\frac{M_0}{n} + \frac{\sigma^2}{n}\sum_{j=M_0+1}^{n}\left[\left(1-(1-\frac{1}{1+\lambda j^{2\nu_0/d}}\right)^{k+1}\right]^2.$$

Choose J_n in $j = M_0, \ldots, n$, and split the sum in two parts. Then bound the summand of the first sum by one to get

$$V(\hat{m}_k, \lambda_0, \nu_0) \leq \sigma^2\frac{M_0}{n} + \sigma^2\frac{J_n - M_o}{n} + \frac{\sigma^2}{n}\sum_{j=J_n+1}^{n}\left[\left(1-(1-\frac{1}{1+\lambda j^{2\nu_0/d}}\right)^{k+1}\right]^2.$$

As the function $1-(1-u)^k \leq ku$ for $u \in [0,1]$, we have

$$\begin{aligned} V(\hat{m}_k, \lambda_0, \nu_0) &\leq \sigma^2\frac{J_n}{n} + (k+1)^2\sum_{j=J_n+1}^{n}\left(\frac{1}{1+\lambda j^{2\nu_0/d}}\right)^2 \\ &\leq \sigma^2\frac{J_n}{n} + (k+1)^2\sum_{j=J_n+1}^{n}\frac{1}{\lambda^2 j^{4\nu_0/d}}. \end{aligned}$$

Bounding the sum by the integral and evaluate the latter, one has

$$V(\hat{m}_k, \lambda_0, \nu_0) \quad \leq \quad \sigma^2 \frac{J_n}{n} + (k+1)^2 \frac{\sigma^2}{n} \frac{1}{\lambda^2 (4\nu_0/d - 1)} J_n^{-4\nu_0/d+1}.$$

If we want to balance the two terms of the variance, one has to choose the following number of iterations $K_n = O(J_n^{2\nu_0/d})$. For such a choice the variance is of order

$$V(\hat{m}_k, \lambda_0, \nu_0) \quad = \quad O\left(\frac{J_n}{n}\right).$$

Let us evaluate the squared bias of $\hat{m}_k$. Recall first the decomposition of $S_{\nu_0,\lambda_0} = P_{\nu_0} \Lambda P'_{\nu_0}$ and denote by $\mu_{j,\nu_0} = [P'_{\nu_0}]_j m$ the coordinate of m in the eigen vector space of S_{ν_0,λ_0}.

$$\begin{aligned} b(\hat{m}_k, \lambda_0, \nu_0) &= \frac{1}{n} \sum_{j=1}^{n} (1 - \lambda_j)^{2k+2} \mu_{j,\nu_0}^2 \\ &= \frac{1}{n} \sum_{j=M_0+1}^{j_n} (1 - \lambda_j)^{2k+2} \mu_{j,\nu_0}^2 + \frac{1}{n} \sum_{j=j_n+1}^{n} (1 - \lambda_j)^{2k+2} \mu_{j,\nu_0}^2 \end{aligned}$$

If m belongs to $\mathcal{H}^{(\nu)}$ it belongs to and $\mathcal{H}^{(\nu_0)}$ and we have the following relation

$$\text{(A.1)} \qquad \frac{1}{n} \sum_{j=M_0+1}^{n} j^{2\nu_0/d} \mu_{j,\nu_0}^2 \leq M < \infty.$$

and with the following bound $\lambda_j > 0$, we obtain that the first term if bounded by say M':

$$\begin{aligned} b(\hat{m}_k, \lambda_0, \nu_0) &\leq M' + \frac{1}{n} \sum_{j=j_n+1}^{n} j^{-2\nu/d} j^{2\nu/d} \mu_{j,\nu_0}^2 \\ b(\hat{m}_k, \lambda_0, \nu_0) &\leq M' + j_n^{-2\nu/d} \frac{1}{n} \sum_{j=j_n+1}^{n} j^{2\nu/d} \mu_{j,\nu_0}^2 \end{aligned}$$

Using the same type of bound as in equation (A.1) we get

$$b(\hat{m}_k, \lambda_0, \nu_0) \quad \leq \quad M' + j_n^{-2\nu/d} M''.$$

Thus the bias is of order $O(j_n^{-2\nu/d})$.

Balancing the squared bias and the variance lead to the choice

$$J_n \quad = \quad O(n^{1/(1+2\nu/d)})$$

and we obtain the desired optimal rate.

Proof of Theorem 4.1 For pedagogical reasons, we present the proof in the univariate case. Let $X_1, \ldots, X_n$ is an i.i.d. sample from a density f that is bounded away from zero on a compact set strictly included in the support of f. Consider without loss of generality that $f(x) \geq c > 0$ for all $|x| < b$.

We are interested in the sign of the quadratic form $u^t A u$ where the individual entries A_{ij} of matrix A are equal to

$$A_{ij} \quad = \quad \frac{K_h(X_i - X_j)}{\sqrt{\sum_l K_h(X_i - X_l)}\sqrt{\sum_l K_h(X_j - X_l)}}.$$

Recall the definition of the scaled kernel $K_h(\cdot) = K(\cdot/h)/h$. If v is the vector of coordinate $v_i = u_i/\sqrt{\sum_l K_h(X_i - X_l)}$ then we have $u^t A u = v^t \mathbb{K} v$, where $\mathbb{K}$ is the matrix with individual entries $K_h(X_i - X_j)$. Thus any conclusion on the quadratic form $v^t \mathbb{K} v$ carry on to the quadratic form $u^t A u$. To show the existence of a negative eigenvalue for $\mathbb{K}$, we seek to construct a vector $U = (U_1(X_1), \ldots, U_n(X_n))$ for which we can show that the quadratic form

$$U^t \mathbb{K} U = \sum_{j=1}^{n} \sum_{k=1}^{n} U_j(X_j) U_k(X_k) K_h(X_j - X_k)$$

converges in probability to a negative quantity as the sample size grows to infinity. We show the latter by evaluating the expectation of the quadratic form and applying the weak law of large number. Let $\varphi(x)$ be a real function in L_2, define its Fourier transform

$$\hat{\varphi}(t) \quad = \quad \int e^{-2i\pi t x} \varphi(x) dx$$

and its Fourier inverse by

$$\hat{\varphi}_{inv}(t) \quad = \quad \int e^{2i\pi t x} \varphi(x) dx.$$

For kernels $K(\cdot)$ that are real symmetric probability densities, we have

$$\hat{K}(t) \quad = \quad \hat{K}_{inv}(t).$$

From Bochner's theorem, we know that if the kernel $K(\cdot)$ is not positive definite, then there exists a bounded symmetric set A of positive Lebesgue measure (denoted by $|A|$), such that

$$\hat{K}(t) < 0 \quad \forall t \in A. \tag{A.2}$$

Let $\widehat{\varphi}(t) \in L_2$ be a real symmetric function supported on A, bounded by B (i.e. $|\widehat{\varphi}(t)| \leq B$). Obviously, its inverse Fourier transform

$$\varphi(x) = \int_{-\infty}^{\infty} e^{-2\pi i x t} \widehat{\varphi}(t) dt$$

is integrable and by virtue of Parceval's identity

$$\|\varphi\|^2 = \|\widehat{\varphi}\|^2 \leq B^2 |A| < \infty.$$

Using the following version of Parceval's identity [see 16, p.620]

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \varphi(x)\varphi(y) K(x-y) dx dy = \int_{-\infty}^{\infty} |\widehat{\varphi}(t)|^2 \hat{K}(t) dt,$$

which when combined with equation (A.2), leads us to conclude that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \varphi(x)\varphi(y) K(x-y) dx dy < 0.$$

Consider the following vector

$$U = \frac{1}{nh} \begin{bmatrix} \frac{\varphi(X_1/h)}{f(X_1)} \mathbb{I}(|X_1| < b) \\ \frac{\varphi(X_2/h)}{f(X_2)} \mathbb{I}(|X_2| < b) \\ \vdots \\ \frac{\varphi(X_n/h)}{f(X_n)} \mathbb{I}(|X_n| < b) \end{bmatrix}.$$

With this choice, the expected value of the quadratic form is

$$\begin{aligned} \mathbb{E}[Q] &= \mathbb{E}\left[\sum_{j,k=1}^{n} U_j(X_j) U_k(X_k) K_h(X_j - X_k) \right] \\ &= \frac{1}{n} \int_{-b}^{b} \frac{1}{f(s)h^2} \varphi(s/h)^2 K_h(0) ds \\ &\quad + \frac{n^2 - n}{n^2} \int_{-b}^{b} \int_{-b}^{b} \frac{1}{h^2} \varphi(s/h)\varphi(t/h) K_h(s-t) ds dt \\ &= I_1 + I_2. \end{aligned}$$

We bound the first integral

$$\begin{aligned} I_1 &= \frac{K_h(0)}{nh^2} \int_{-b}^{b} \frac{\varphi(s/h)^2}{f(s)} ds \\ &\leq \frac{K_h(0)}{nch} \int_{-b/h}^{b/h} \varphi(u)^2 du \\ &\leq \frac{B^2 |A| K(0)}{ch^2} n^{-1}. \end{aligned}$$

Observe that for any fixed value h, the latter can be made arbitrarily small by choosing n large enough. We evaluate the second integral by noting that

$$\begin{aligned} I_2 &= \left(1-\frac{1}{n}\right)h^{-2}\int_{-b}^{b}\int_{-b}^{b}\varphi(s/h)\varphi(t/h)K_h(s-t)dsdt \\ &= \left(1-\frac{1}{n}\right)h^{-2}\int_{-b}^{b}\int_{-b}^{b}\varphi(s/h)\varphi(t/h)\frac{1}{h}K\left(\frac{s}{h}-\frac{t}{h}\right)dsdt \\ &= \left(1-\frac{1}{n}\right)h^{-1}\int_{-b/h}^{b/h}\int_{-b/h}^{b/h}\varphi(u)\varphi(v)K(u-v)dudv. \qquad \text{(A.3)} \end{aligned}$$

By virtue of the dominated convergence theorem, the value of the last integral converges to $\int_{-\infty}^{\infty}|\hat{\varphi}(t)|^2\hat{K}(t)dt < 0$ as h goes to zero. Thus for h small enough, (A.3) is less than zero, and it follows that we can make $\mathbb{E}[Q] < 0$ by taking $n \geq n_0$, for some large n_0. Finally, convergence in probability of the quadratic form to its expectation is guaranteed by the weak law of large numbers for U statistics [see 19, for example]. The conclusion of the theorem follows.

Proof of Proposition 4.2 To handle multivariate case, let each component h_j of the vector h be larger than the minimum distance between three consecutive points, and denote by $d_h(X_i, X_j)$ the distance between two vectors related to the vector chosen by the user. For example, if the usual Euclidean distance is used, we have

$$d_h^2(X_i, X_j) = \sum_{l=1}^{d}\left(\frac{X_{il}-X_{jl}}{h_l}\right)^2 .$$

The multivariate kernel evaluated at X_i, X_j can be written as $K(d_h(X_i, X_j))$ where K is univariate. We are interested in the sign of the quadratic form $u^t\mathbb{K}u$ (see proof of Theorem 4.1). Recall that if $\mathbb{K}$ is semidefinite then all its principal minor [see 24, p.398] are nonnegative. In particular, we can show that A is non-positive definite by producing a 3×3 principal minor with negative determinant. To this end, take the principal minor $\mathbb{K}[3]$ obtained by taking the rows and columns (i_1, i_2, i_3). The determinant of $\mathbb{K}[3]$ is

$$\begin{aligned} det(\mathbb{K}[3]) = \; & K(d_h(0))\left[K(d_h(0))^2 - K(d_h(X_{i_3}, X_{i_2}))^2\right] \\ & -K(d_h(X_{i_2}, X_{i_1})) \times \\ & [K(d_h(0))K(d_h(X_{i_2}, X_{i_1})) - K(d_h(X_{i_3}, X_{i_2}))K(d_h(X_{i_3}, X_{i_1}))] \\ & +K(d_h(X_{i_3}, X_{i_1})) \times \\ & [K(d_h(X_{i_2}, X_{i_1}))K(d_h(X_{i_3}, X_{i_2})) - K(d_h(0))K(d_h(X_{i_3}, X_{i_1}))] . \end{aligned}$$

Let us evaluate this quantity for the Uniform and Epanechnikov kernels.

Uniform kernel. Choose 3 points in $\{X_i\}_{i=1}^n$ with index i_1, i_2, i_3 such that

$$d_h(X_{i_1}, X_{i_2}) < 1, \quad d_h(X_{i_2}, X_{i_3}) < 1, \quad \text{and} \quad d_h(X_{i_1}, X_{i_3}) > 1.$$

With this choice, we readily calculate

$$det(\mathbb{K}[3]) \;=\; 0 - K_h(0)\left[K_h(0)^2 - 0\right] - 0 < 0.$$

Since a principal minor of $\mathbb{K}$ is negative, we conclude that $\mathbb{K}$ and A are not semidefinite positive.

Epanechnikov kernel. Choose 3 points $\{X_i\}_{i=1}^n$ with index i_1, i_2, i_3, such that $d_h(X_{i_1}, X_{i_3}) > \min(d_h(X_{i_1}, X_{i_2}); d_h(X_{i_2}, X_{i_3}))$ and set $d_h(X_{i_1}, X_{i_2}) = x \leq 1$ and $d_h(X_{i_2}, X_{i_3}) = y \leq 1$.

Using triangular inequality, we have

$$\begin{aligned} det(\mathbb{K}[3]) \;<\; & 0.75(0.75^2 - K(y)^2) - K(x)(0.75K(x) - K(y)K(\min(x,y))) \\ & -K(\min(x,y))K(x)K(y) - 0.75K(x+y)^2 \end{aligned}$$

The right hand side of this equation is a bivariate function of x and y. Numerical evaluations of that function show that small x and y leads to negative value of this function, that is the determinant of $\mathbb{K}[3]$ can be negative.

FIG 5. *Contour of an upper bound of $det(\mathbb{K}[3])$ as a function of (x, y).*

Thus a principal minor of $\mathbb{K}$ is negative, and as a result, $\mathbb{K}$ and A are not semidefinite positive.

REFERENCES

[1] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and B. F. Csaki, editors, *Second international symposium on information theory*, pages 267–281, Budapest, 1973. Academiai Kiado.

[2] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[3] L. Breiman. Using adaptive bagging to debais regressions. Technical Report 547, Department of Statistics, UC Berkeley, 1999.

[4] P. Bühlmann and B. Yu. Boosting with the l_2 loss: Regression and classification. *J. Amer. Statist. Assoc.*, 98:324–339, 2003.

[5] P. Bühlmann and B. Yu. Sparse boosting. *J. Macine Learning Research*, 7:1001–1024, 2006.

[6] A. Buja, T. Hastie, and R. Tibshirani. Linear smoothers and additive models. *Ann. of Statist.*, 17:453–510, 1989.

[7] W. Cleveland and S. Devlin. Locally weighted regression : an approach to regression analysis by local fitting. *J. Amer. Stat. Ass.*, 83:596–610, 1988.

[8] P.-A. Cornillon, N. Hengartner, and E. Matzner-Løber. Recursive bias estimation and l_2 boosting. Technical report, ArXiv:0801.4629, 2008.

[9] P.-A. Cornillon, N. Hengartner, and Matzner-Løber. ibr: Iterative bias reduction multivariate smoothing. *submitted to the Journal of Statistical Software*, 2009.

[10] P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31:377–403, 1979.

[11] M. Di Marzio and C. Taylor. Boosting kernel density estimates: a bias reduction technique ? *Biometrika*, 91:226–233, 2004.

[12] M. Di Marzio and C. Taylor. Multiple kernel regression smoothing by boosting. *submitted*, 2007.

[13] M. Di Marzio and C. Taylor. On boosting kernel regression. *to appear in JSPI*, 2008.

[14] R. Eubank. *Spline Smoothing and Nonparametric Regression*. Marcel Dekker, New-York, 1988.

[15] J. Fan and I. Gijbels. *Local Polynomial Modeling and Its Application, Theory and Methodologies*. Chapman et Hall, New York, 1996.

[16] W. Feller. *An introduction to probability and its applications*, volume 2. Wiley, New York, 1966.

[17] J. Friedman. Multivariate adaptive regression splines. *Ann. Statist.*, 19:337–407, 1991.

[18] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Ann. of Statist.*, 28:337–407, 2000.

[19] W. Grams and R. Serfling. Convergence rates for u-statistics and related statistics. *Annals of Statistics*, 1:153–160, 1973.

[20] C. Gu. *Smoothing spline ANOVA models*. Springer, New-York, 2002.

[21] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman & Hall, 1995.

[22] N. Hengartner and S. Sperlich. Rate optimal estimation with the integration method in the presence of many covariates. *Journal of multivariate anolysis*, 95(2):246–272, 2005.

[23] N. Hengartner, M. Wegkamp, and E. Matzner-Løber. Bandwidth selection for local linear regression smoothers. *JRSS B*, 64:1–14, 2002.

[24] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge, New York, 1985.

[25] C. Hurvich, G. Simonoff, and C. L. Tsai. Smoothing parameter selection in nonparametric regression using and improved akaike information criterion. *J. R. Statist. Soc. B*, 60:271–294, 1998.

[26] K.-C. Li. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics*, 15:958–975, 1987.

[27] O. Linton and J. Nielsen. A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 82:93–100, 1995.

[28] C. L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

[29] K. Messer. A comparison of a spline estimate to its equivalent kernel estimate. *Ann. of Statist.*, 19:817–829, 1991.

[30] J. O. Ramsay and N. Heckman. Some theory for l-spline smoothing. Technical report, McGill University, 1996.

[31] G. Ridgeway. Additive logistic regression: a statistical view of boosting: Discussion. *Ann. of Statist.*, 28:393–400, 2000.

[32] L. Schwartz. *Analyse IV applications à la théorie de la mesure*. Hermann, Paris, 1993.

[33] G. Schwarz. Estimating the dimension of a model. *Annals of statistics*, 6:461–464, 1978.

[34] B. Silverman. Spline smoothing: the equivalent variable kernel method. *Annals of statistics*, 12:898–916, 1984.

[35] J. Simonoff. *Smoothing Methods in Statistics*. Springer, New York, 1996.

[36] J. Tukey. *Explanatory Data Analysis*. Addison-Wesley, 1977.

[37] F. Utreras. Convergence rates for multivariate smoothing spline functions. *Journal of Approximation Theory*, pages 1–27, 1988.

[38] G. Wahba. *Spline models for observational data*. SIAM, Philadelphia, 1990.

[39] J. Wendelberger. Smoothing noisy data with multivariate splines and generalized cross-validation. *Ph.D thesis, University of Wisconsin*, 1982.

ADDRESS OF P-A CORNILLON
UMR ASB - MONTPELLIER SUPAGRO
34060 MONTPELLIER CEDEX 1
E-MAIL: pierre-andre.cornillon@supagro.inra.fr

ADDRESS OF N. HENGARTNER
LOS ALAMOS NATIONAL LABORATORY,
NW. USA
E-MAIL: nickh@lanl.gov

ADDRESS OF E. MATZNER-LØBER
STATISTICS, IRMAR UMR 6625.
UNIV. RENNES 2,
35043 RENNES, FRANCE
E-MAIL: eml@uhb.fr