LA-UR- 08-6365

Approved for public release; distribution is unlimited.

Title:	Recursive Bias Estimation for High Dimensional Regression Smoothers
Author(s):	Pierre-Andre Cornillon, INRA Nicolas Hengartner, 186926, CCS-3 Eric Matzner-Lober, 222816, UHB, France
Intended for:	Publication

Los Alamos
NATIONAL LABORATORY
EST. 1943

Form 836 (7/06)

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. We apply our method to simulated and real data and show that our method compares favorably with existing procedure.

1. Introduction. Regression is a fundamental data analysis tool for uncovering functional relationships between pairs of observations $(X_i, Y_i), i = 1, \ldots, n$. 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 of Y given X. This results in nonparametric regression estimators [e.g. 15, 21, 33]. The vector of predicted values $\widehat{Y}_i$ at the observed covariates X_i from a nonparametric regression is called a regression smoother, or simply a smoother, because the predicted values $\widehat{Y}_i$ are less variable than the original observations Y_i.

Over the past thirty years, numerous smoothers have been proposed: running-mean smoother, running-line smoother, bin smoother, kernel based smoother, spline regression smoother, smoothing splines smoother, locally weighted running-line smoother, just to mention a few. We refer to Buja

AMS 2000 subject classifications: 62G08

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

et al. [6], Eubank [14], Fan and Gijbels [15], Hastie and Tibshirani [21] for more in depth treatments of univariate regression smoothers and to Cleveland and Devlin [8] for multiple smoothers.

An important property of smoothers is that they do not require a rigid (parametric) specification of the regression function. That is, we model the pairs (X_i, Y_i) as

$$Y_i = m(X_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where $m(\cdot)$ is an unknown smooth function. The disturbances ε_i are independent mean zero and variance σ^2 random variables that are independent of the covariates X_i, $i = 1, \ldots, n$. To help our discussion on smoothers, we rewrite Equation (1.1) compactly in vector form by setting $Y = (Y_1, \ldots, Y_n)^t$, $m = (m(X_1), \ldots, m(X_n))^t$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^t$, to get

$$Y = m + \varepsilon. \tag{1.2}$$

Finally we write $\widehat{m} = \widehat{Y} = (\widehat{Y}_1, \ldots, \widehat{Y}_n)^t$, the vector of fitted values from the regression smoother at the observations. 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 is denoted 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 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. The parameter λ is the bandwidth for kernel smoother, the span size for running-mean smoother, bin smoother, and the penalty factor λ for spline smoother.

Much has been written on how to select an appropriate smoothing parameter, see for example Simonoff [33]. Ideally, we want to choose the smoothing parameter λ to minimize the expected squared prediction error. But without explicit knowledge of the underlying regression function, the mean squared prediction error can not be computed directly. Instead, one relies on estimates of the mean squared prediction error using Stein Unbiased Risk Estimate [34] or Cross-Validation [26].

It is well known in multivariate analysis that the distance between typical covariates increases with increasing dimensions d of the covariates X. The resulting sparseness of the covariates, often called *the curse of dimensionality*,

forces one to use larger smoothing parameters in higher dimensions, which in term leads to more 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. Instead one often focus on fitting structurally constrained regression models, such as additive models [21] and multiplicative models [28].

This paper takes a different approach. Instead of focusing on selecting the tuning parameters λ of the smoother, or equivalently the size of the local neighborhood over which the responses are averaged, we fix the smoothing parameter to some reasonably large value to ensure that smoother averages the responses over large neighborhoods. This resulting *over-smooth* of the data has a small variance but a substantial bias. We then proceed to *correct* the initial smoother by subtracting from it an estimate of its bias obtained by smoothing the residuals from the initial fit. If we smooth the residuals with the same smoother that we used to smooth the data, we do not change the size of the local neighborhood we are averaging over. As a result, this bias correction partially circumvents the root cause of the curse of dimensionality.

Since the estimate of the bias is itself biased, there is potentially a benefit to iterating the bias correction step. We can let the data tell us the desirable number of iterations of bias correction by minimizing an estimate of the prediction error, obtained by cross-validation or generalized cross-validation, for example.

We show in this paper that the behavior of the sequence of iteratively bias corrected smoother depends on the spectrum of $I - S_\lambda$. For some commonly used smoothers, such as Gaussian kernel regression smoothers and smoothing splines, the bias of the iteratively bias corrected smoothers converge to zero. But for other common smoothers, such as the nearest neighbor smoother and kernel regression smoothers with an Epanechnikov kernel, the bias of the iterative bias corrected smoother diverges.

This approach has the potential to work very well when the true underlying regression function is smooth. When applied to simulated data, our method leads to smoothers that whose mean squared prediction error is up to 30% smaller than the mean squared prediction of additive models and MARS. The good predictions observed on simulated data is also realized on real data. When applied to the Los Angeles Ozone data, our method of iterative bias reduction produces a smoother with substantially lower mean squared prediction error, around 18% smaller, than the mean squared prediction error of competing smoothers proposed in the literature.

From a historical perspective, the idea of estimating the bias from resid-

uals to correct a pilot estimator of a regression function goes back to the concept of *twicing* introduced by Tukey [35] to estimate bias of misspecified multivariate regression models. More recently, 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]. The correspondence between L_2-boosting and our iterative bias correction procedure follows from the representation of the bias corrected smoother presented in Section 2 and the expression found in Bühlmann and Yu [4]. This new interpretation for the L_2 boosting algorithm as iterative bias corrections was alluded to in Ridgeway [30]'s discussion of Friedman et al. [18] paper on the statistical interpretation of boosting. The idea of iterative debiasing regression smoothers is also present in Breiman [3] in the context of the *bagging* algorithm.

Finally, while this paper focus on linear smoothers to estimate bias, it is possible to apply the same idea with nonlinear bias reduction techniques. For example, one can use the multiplicative bias correction technique of Burr et al. [7], Hengartner and Matzner-Løber [22] that preserve the sign of the pilot smoother.

This paper is structure as follows. We start in Section 2 by presenting two approaches to bias estimation for linear smoothers. The first is based on the plug-in method while the second focuses on smoothing residuals. We give conditions under which both approaches produce the same estimate for the bias. Identifying the smoothing matrix corresponding to the k-times bias corrected smoother allows us to describe qualitatively the behavior of the sequence of iteratively bias-corrected smoothers in terms of the spectrum of $I - S$. In Section 3, we study the behavior of the iterative bias-corrected smoother based on commonly used multivariate smoothers: kernel smoothers, K-nearest neighbor smoothers and smoothing splines. We prove, and show by example, that not all smoothers are suitable for to be used by our iterative bias reduction technique. In particular, we prove that the iterative bias correction of nearest neighbor smoothers produces a sequence of smoothers that behave erratically after a small number of iterations, and eventually diverges. Another class of smoothers that are not suitable for our iterative bias correction scheme are kernel smoothers based on kernels that are not positive definite.

The iterative bias correction scheme is controlled by two *parameters*: the smoothness of the initial smoother, and the number of bias correction iterations. We discuss the choice of both these parameters in Section 4. The simulations in Section 5 show that combining a GCV based stopping rule

to the iterative bias reduction algorithm seems to work well. It stops early when the sequence of iterated bias corrected smoothers misbehaves, and otherwise takes advantage of the bias reduction. Our simulation compares optimum smoothers and optimum iterative bias corrected smoothers (using generalized cross validation) for general smoothers without knowledge of the underlying regression function. We conclude that the optimal iterative bias corrected smoother outperforms the optimal smoother.

Finally, the proofs are gathered in the Appendix.

2. Bias estimation. This section introduces our bias corrected linear smoother and characterizes the qualitative behavior of the sequence of smoothers obtained through iterative bias correction.

2.1. *Bias Corrected Linear Smoothers.* Recall the multivariate nonparametric regression model in vector form (1.2) from Section 1

$$Y = m + \varepsilon,$$

where the errors ε are independent, have mean zero and constant variance σ^2, and are independent of the covariates $X = (X_1, \ldots, X_n)$, $X_j \in \mathbb{R}^d$. Linear smoothers can be written as

$$\widehat{m}_1 = SY, \tag{2.1}$$

where S is an $n \times n$ smoothing matrix. Typical smoothing matrices are contractions, so that $\|SY\| \leq \|Y\|$, and as a result the associated smoother SY is called a shrinkage smoother (see for example Buja et al. [6]). Let I be the $n \times n$ identity matrix. The linear smoother (2.1) has bias

$$B(\widehat{m}_1) = \mathbb{E}[\widehat{m}_1|X] - m = (S - I)m \tag{2.2}$$

and variance

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

respectively.

There are at least two approaches to estimate the bias (2.2). A first estimator for the bias is obtained by plugging in an estimator $\widehat{m} = S_2 Y$ for the unknown regression function m into the expression (2.2) for the bias of the estimator $\widehat{m}_1 = S_1 Y$. This produces

$$\begin{aligned} \hat{b}_1 &= (S_1 - I)\hat{m} \\ &= (S_1 - I)S_2 Y. \end{aligned}$$

Correcting the pilot smoother $\widehat{m}_1$ for its bias produces

$$\widehat{m}_2 \;=\; S_1Y + (I - S_1)S_2Y,$$

which is itself a linear smoother. Repeating the bias correction step k times, leads to the linear smoother

$$\widehat{m}_k \;=\; S_1Y + (I - S_1)S_2Y + \cdot + (I - S_1)(I - S_2)\cdots S_kY, \tag{2.3}$$

whose associated smoothing matrix is simplified in the following theorem.

THEOREM 2.1 (Plug-in estimator). *After k iterations, estimator (2.3) can be explicitly written as*

$$\widehat{m}_k \;=\; [I - (I - S_1)(I - S_2)\cdots(I - S_k)]Y. \tag{2.4}$$

A second estimator for the bias is obtained by observing that the residuals $R_1 = Y - \widehat{m}_1 = (I - S_1)Y$ have expected value $\mathbb{E}[R_1|X] = m - \mathbb{E}[\widehat{m}_1|X] = (I - S_1)m = -B(\widehat{m}_1)$. This suggests estimating the bias by smoothing the negative residuals

$$\tilde{b}_1 := -S_2R_1 = -S_2(I - S_1)Y. \tag{2.5}$$

Correcting the pilot smoother $\widehat{m}_1$ with the latter bias estimate produces the smoother

$$\widehat{m}_2 \;=\; S_1Y + S_2(I - S_1)Y.$$

Iterating the bias reduction step k times leads to the linear smoother

$$\widehat{m}_k \;=\; S_1Y + S_2(I - S_1)Y + \cdots + S_k(I - S_{k-1})\cdots(I - S_1)Y. \tag{2.6}$$

The next theorem provides a compact representation for its smoothing matrix

THEOREM 2.2 (Residual smoothing estimator). *After k iterations, estimator (2.6) can be explicitly written as*

$$\widehat{m}_k \;=\; [I - (I - S_k)(I - S_{k-1})\cdots(I - S_1)]Y. \tag{2.7}$$

Theorems 2.1 and 2.2 show that in general, the two sequences of smoothers are not the same unless the smoothing matrices $S_1, \ldots, S_k$ commute. An important special case of the latter is when $S_1 = S_2 = \cdots = S$, in which case both (2.1) and (2.2) reduce to the following corollary.

COROLLARY 2.3. *If the same smoother S is used to smooth the data and to estimate the bias, then the k^{th} iterated bias corrected linear smoother $\widehat{m}_k$ can be explicitly written as*

$$\begin{aligned} \widehat{m}_k &= S[I + (I-S) + (I-S)^2 + \cdots + (I-S)^{k-1}]Y \\ &= [I - (I-S)^k]Y = S_k Y. \end{aligned} \tag{2.8}$$

Remark 1 If the smoother S is a projection (as is the case for bin smoothers and regression splines), then the estimated bias

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

and hence $S_k = S$ for all k.

Remark 2 In the univariate case, smoothers of the form (2.8) arise from the L_2 boosting algorithm with convergence factor $\mu_k \equiv 1$ studied by Bühlmann and Yu [4] when S is a smoothing spline. This provides a new statistical interpretation for L_2 boosting. Breiman [3] noted a similar interpretation for the bagging algorithm applied to the residuals of nonparametric smoothers.

2.2. *Predictive smoothers.* As defined by (2.1), smoothers predict the conditional expectation of responses at the design points. It is interesting to extend regression smoothers to produce predictions at arbitrary locations. Such an extension enables us to assess and compare the quality of various smoothers in terms of how well they predict new observations.

To this end, recall that the prediction of a linear smoother S at an arbitrary location x can be written 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.

To extend the iterative bias corrected smoother $\widehat{m}_k$ defined in 2.8, we write

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

and predict $m(x)$ by

$$\widehat{m}_k(x) = S(x)^t \widehat{\beta}_k. \tag{2.9}$$

This formulation is computationally advantageous because the vector of weights $S(x)$ only needs to be computed once, and the iterative bias correction scheme leads to the sequential update rule

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

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

2.3. *Properties of iterative bias corrected smoothers.* The squared bias and variance of the k^{th} iterated bias corrected smoother $\widehat{m}_k$ (2.8) are

$$\begin{aligned} B^2(\hat{m}_k) &= m^t \left((I-S)^k\right)^t (I-S)^k m \\ \mathrm{V}(\hat{m}_k) &= \sigma^2 (I-(I-S)^k)\left((I-(I-S)^k)\right)^t, \end{aligned}$$

respectively. It follows that the qualitative behavior of the sequence of iterative bias corrected smoothers $\widehat{m}_k$ is determined by the spectrum of $I-S$. The next theorem collects the various convergence results for sequence of iterated bias corrected linear smoothers.

THEOREM 2.4. *Suppose that the singular values $\lambda_j = \lambda_j(I-S)$ of $I-S$ satisfy*

$$-1 < \lambda_j < 1 \quad \textit{for} \quad j = 1, \ldots, n. \tag{2.10}$$

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 does not imply that the smoother S itself is a shrinkage smoother as defined by Buja et al. [6]. Conversely, not all shrinkage estimators satisfy the condition (2.10) of the theorem. In the next section, 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 will fail. 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 over-corrects the bias of the smoothers, which results in a divergent sequence of smoothers.

We conclude this section by noting that iterating the bias correction scheme to reach the limiting smoother $\widehat{m}_\infty$ is not desirable, for either $\widehat{m}_\infty = Y$ or $\|\widehat{m}_\infty\| = \infty$. However, since each iteration decreases the bias at the cost of increased variance, a suitably selected estimator from the sequence $\{\widehat{m}_k\}$ is likely to improve upon the initial smoother $\widehat{m}_1$.

3. Bias reduction for classical smoothers. This section is devoted to understanding the behavior of the iterative bias reduction schema using classical smoothers, which in light of Theorem 2.4, depends on the magnitude of the singular values of the matrix $I - S$.

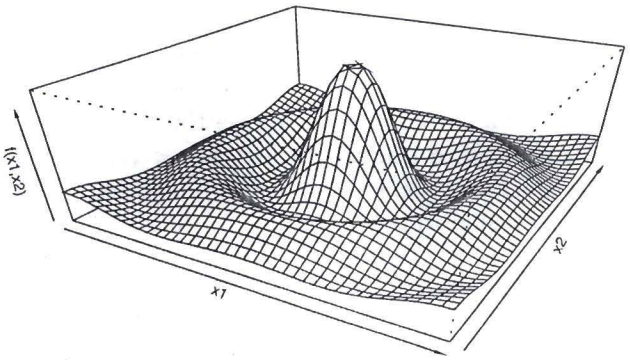

FIG 1. *"Mexican hat" function 3.1 on* $[-10, 10] \times [-10, 10]$

3.1. *Pedagogical example.* Throughout this section, we illustrate the theoretical results by applying the iterative bias reduction scheme using various common smoothers on the same simulated bivariate regression example. In that example, we sample the well known "Mexican hat" (see Figure 1) ,

$$m(x_1, x_2) = 10\frac{\sin(\sqrt{x_1^2+x_2^2})}{\sqrt{(x_1^2+x_2^2)}}. \tag{3.1}$$

at 100 points taken on the regular grid $\{-9.5, -8.5, \ldots, 8.5, 9.5\}^2$. The disturbances are mean zero Gaussian with variance producing a signal to noise ratio of five.

3.2. *Projection type smoothers.* We start our discussion by noting that iterative bias reduction a projection type smoothers is of no interest because residuals $(I-S)Y$ are orthogonal to smoother SY. It follows that the smoothed residuals $S(I-S)Y=0$, and as a result, $\widehat{m}_k = \widehat{m}_1$ for all k.

3.3. *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.

3.3.1. *Spectrum of kernel smoothers.* To apply Theorem 2.4, 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

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

This shows that 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 less or equal to one. Thus, in light of Theorem 2.4, 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 3.1. *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 $X_1, \ldots, X_n$ 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 $x_1, \ldots, x_m$, the matrix

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

is positive definite. We refer to Schwartz [31] 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.

3.3.2. *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 the symmetric matrix $A = D^{1/2}\mathbb{K}D^{1/2}$ is symmetric as $A = P_A \Lambda_A P_A^t$, with P_A the orthonormal of eigenvectors and Λ_A diagonal matrix of associated eigenvalues leads to 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.

3.3.3. *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 3.1 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 2 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.

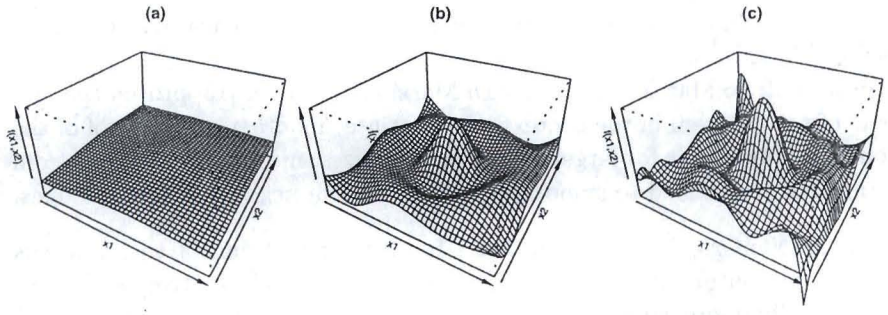

FIG 2. *Gaussian kernel smoother of the function $m(x_1, x_2)$ from $n = 100$ equidistributed points on $[-10, 10] \times [-10, 10]$, evaluated on a regular grid with $k = 1$ iteration (a), 50 iterations (b) and 10000 iterations (c).*

3.3.4. *Kernel smoothers with Uniform and Epanechnikov kernels.* The uniform and the Epanechnikov kernels are not positive definite. Theorem 3.1 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 3.2 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 3.2. *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 lack a suitability of Epanechnikov kernel smoothers for the iterated bias correction scheme is illustrated in the numerical example shown in Figure 3. 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.

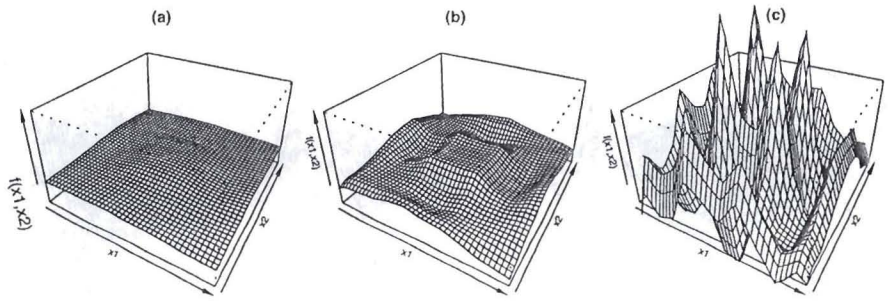

FIG 3. *Epanechnikov kernel smoother of the function $m(x_1, x_2)$ from $n = 100$ equidistributed points on $[-10, 10] \times [-10, 10]$, evaluated on a regular grid with $k = 1$ iteration (a), 5 iterations (b) and 25 iterations (c).*

3.4. *K-nearest neighbor smoother.* The associated smoothing matrix of the K-nearest neighbor smoother is $S_{ij} = 1/K$when X_j belongs to the K-nearest neighbor of X_i and $S_{ij} = 0$ otherwise. While this smoother enjoys

many desirable properties, it is not well suited for the iterative bias correction scheme because the matrix $I - S$ has singular values larger than one. The next theorem gives conditions on the configurations of the design points that leads to negative singular for the smoothing matrix.

THEOREM 3.3. *Let S be the smoothing matrix of the K nearest neighbor smoother with $K \geq 3$. Let $\mathcal{N}_i = \{X_j : X_j$ is a nearest neighbor of $X_i\}$, the set of nearest neighbors of X_i. If there exists at least one neighboring set $\mathcal{N}_i$ such that there exists $X_j, X_k \in \mathcal{N}_i$ for which*

$$X_i \in \mathcal{N}_j, \quad X_i \in \mathcal{N}_k \text{ and } X_k \notin \mathcal{N}_j \text{ and } X_j \notin \mathcal{N}_k,$$

then at least one singular value of S is negative.

The proof of the theorem is found in the appendix. A consequence of theorems 3.3 and 2.4, is that the sequence of iterative bias corrected K-nearest neighbor smoothers divergent, and hence should not be used in practice.

We confirm this behavior numerically. Using the same data as before, we apply the iterative bias reduction algorithm to the K-nearest neighbor smoother. We start with a pilot K-nearest neighbor smoother with $K = 20$ (see Figure 4 panel (a)) that produces a very smooth nearly constant surface as K is a significant fraction of the sample size $n = 100$. Already after five iterations, the smoother deteriorates (panel (b)) and exhibits an erratic behavior after ten iterations (panel (c)).

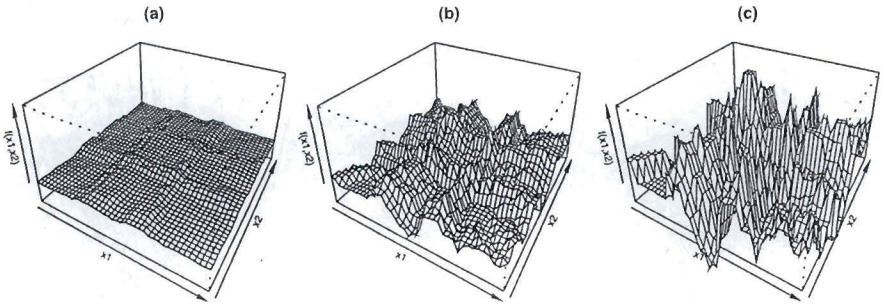

FIG 4. *K-nearest neighbor smoother of the function $m(x_1, x_2)$ given $n = 100$ points equidistributed on $[-10, 10] \times [-10, 10]$ evaluated on a regular grid with $k = 1$ iteration (a), 5 iterations (b) and 10 iterations (c).*

3.5. *Smoothing spline smoother.* It is well known that the univariate spline smoother is a symmetric smoothing matrix S whose eigenvalues lie

between 0 and 1. To extend smoothing splines to multivariate settings, one can use the fact that polynomial smoothing splines can be recasted in the framework of reproducing kernel Hilbert space, as well as thin-plate splines see Gu [20]. Recall briefly that univariate cubic smoothing splines can be described as minimizers of

$$\min_{f\in\mathcal{H}} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \|f\|_w^2, \tag{3.2}$$

where $\mathcal{H}$ is a reproducing kernel Hilbert space (here $C^{(2)}$ equiped with a suitable reproducing kernel) and $\|f\|_w^2$ is a squared (semi)norm in $\mathcal{H}$. Using this framework, modeling with several covariate can be done easily with tensor product of reproducing kernel Hilbert space. The resulting smoothing matrix S remains symmetric and has eigenvalues between 0 and 1 Gu [20], p. 61. In light of Theorem 2.4, the iterative bias correction scheme based on smoothers S from splines (in both univariate and multivariate settings) leads to fitted value $\hat{m}_k$ that converge to Y. The resulting sequence of smoothers is stable as opposed to those based on nearest neighbor, Epanechnikov or uniform kernel. Moreover, bias and variance can be expressed as a function of eigenvalues as in Bühlmann and Yu [4] who studied the behavior of the L_2 boosting algorithm for univariate cubic smoothing splines.

4. Parameter selection. The iterative bias reduction scheme requires the user to supply two *parameters*: the bandwidth of the smoother and the number of iterations of bias correction. The choice for both these parameters is discussed in this section.

4.1. *Selecting the bandwidth.* 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.

4.2. *Data driven selection of the number of bias reduction steps.* Theorem 2.4 in Section 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. The possibility of such improvements was shown in the numerical examples of the previous section. 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 [29], Akaike Information Criteria (AIC), Akaike [1], Bayesian Information Criterion (BIC), Schwarz [32], 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 the optimum number of iterations k of the iterative bias correction algorithm by minimizing estimates of the expected squared prediction error of the smoothers over some pre-specified set $\mathcal{K} = \{1, 2, \ldots, M\}$ for the number of iterations.

We rely on the expansive literature on model selection to provide insight into the statistical properties of stopped bias corrected smoother. Theorem 3.2 of Li [27] 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 $\hat{m}$ is unbiased, which is rarely hold in practice and which is particularly not true in our context.

Extensive simulations, both in the univariate and the multivariate settings Cornillon et al. [9] have shown that not only is CV and GCV

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

are computationally more efficient, and application of these criteria lead to better final smoothers.

5. Simulations and real example. In this section, we show, via simulations, that our proposed iterative correction procedure works well for

both simulated and real data. Due to their prevalence in the literature, we include a discussion of how our methods compares in the context of univariate smoothing splines. We further 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.

5.1. *Univariate case.* One of the most common smoother used in univariate function estimation is the smoothing spline. The aim of that section is to show how the iterative bias correction scheme using smoothing splines compare to a classical smoothing spline estimation. In order to do so, we compare for different sample size $n = 50$, 100 and 500, the iterative smoother using two different starting points and two different stopping rules (GCV and Cross Validation) with the smoothing spline estimator obtained by the function smooth.spline in R. To this end, we consider three different 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 between 0 and 1, 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%.

5.2. *Multivariate case.* Here, we focus on the multivariate case, that is $X \in \mathbb{R}^d$, $d > 1$, and consider multivariate Gaussian kernel smoothers. Statistical lore discourages using fully nonparametric methods in higher dimensions as the resulting 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.

The aim of this section is to 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]. To this end, 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.

MARS	Add.	$BR_{ddl=1.50}$	$BR_{ddl=1.20}$	$BR_{ddl=1.10}$	$BR_{ddl=1.05}$
			$n = 50$		
4.298	3.577	3.551 (31)	3.122 (290)	3.179 (1740)	3.226 (14140)
			$n = 100$		
3.746	3.314	2.239 (42)	2.014 (589)	1.967 (5084)	1.970 (42930)
			$n = 200$		
3.354	2.753	1.842 (57)	1.747 (960)	1.707 (8780)	1.680 (76820)

The same features as those found in univariate simulations are found. First, the smoother the pilot estimator is, the bigger the number of iterations chosen by GCV is. Second, the smoother the pilot estimator is, the better is 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 pilot smoother less smooth.

5.3. *Los Angeles Ozone Data.* We consider the classical data set of ozone concentration in the Los Angeles basin which has been previously considered by many authors (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 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 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 componentwise additive model

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

where the component $\hat{f}^{[k],(j)}$ 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 are in the 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 Squared Error
L_2Boost with component-wise spline	17.78
additive model (backfitted with R)	17.44
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	14.74
1.1 initial DDL per variable and **64** iterations	14.74
1.2 initial DDL per variable and **15** iterations	14.78
1.5 initial DDL per variable and **3** iterations	14.97

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 15% over other multivariate methods.

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 [30] 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 to be used for boosting. We show that many weak learners, such as the k-nearest neighbor smoother and some kernel smoothers, are not stable under iterated bias estimation. Our results extend and complement the recent results of Di Marzio and Taylor [12].

Iterating the bias correction scheme until convergence is not desirable. Better smoothers result if one stops the iterative scheme. Our simulations and application to real data show that our method performs well in higher dimensions, 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 2.3 To show 2.8, let $\Sigma = I + (I-S) + \cdots + (I-S)^{k-1}$. The conclusion follows from a telescoping sum argument applied to

$$S\Sigma = \Sigma - (I-S)\Sigma = I - (I-S)^k.$$

Proof of Theorem 2.4

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

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

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

Proof of Theorem 3.3 Let us consider the K-nn smoother the matrix S

is of general term

$$S_{ij} = \frac{1}{K} \quad \text{if} \quad X_j \in K\text{-nn}(X_i).$$

In order to bound the singular values of $(I-S)$, consider the eigen values of $(I-S)(I-S)'$ which are the square of the singular values of $I-S$. Since $A = (I-S)(I-S)'$ is symmetric, we have for any vector u that

$$\lambda_n \leq \frac{u'Au}{u'u} \leq \lambda_1. \tag{A.1}$$

Let us find a vector u such that $u'Au > u'u$. First notice that $A = I - S - S' - SS'$. Thus we have that

$$A_{ii} = 1 - \frac{1}{K}.$$

Second, to bound $A_i ij$, we need to consider three cases:

1. If X_i belongs to the K-nn of X_j and vice versa, then $S_{ij} = S'_{ji} = 1/K$. This does not mean that all the K-nn neighbor of X_i are the same as those of X_j, but if it is the case, then $(SS')_{ij} \leq K/K^2$ and otherwise in the pessimistic case, we bound $(SS')_{ij} \geq 2/K^2$. It therefore follows that

$$2/K^2 - \frac{2}{K} \leq \quad A_{i,j} \quad \leq \frac{K}{K^2} - \frac{2}{K} = -\frac{1}{K}.$$

2. If X_i belongs to the K-nn of X_j $S_{ij} = 1/K$ but X_j does not belong to the K-nn of X_i then $S'_{ji} = 0$. There is at a maximum of $K-1$ points that are in the K-nn of X_i and in the K-nn of X_j so $(SS')_{ij} \leq (K-1)/K^2$. In the pessimistic case, there is only one point, which leads to the bound

$$\frac{1}{K^2} - \frac{1}{K} \leq \quad A_{i,j} \quad \leq \frac{K-1}{K^2} - \frac{1}{K} \leq -\frac{1}{K^2}.$$

3. If X_i does not belong to the K-nn of X_j $S_{ij} = 0$ and X_j does not belong to the K-nn of X_i then $S'_{ji} = 0$. However there are potentially as many as $K-2$ points that are in the K-nn of X_i and in the K-nn of X_j. In that case

$$0 \leq A_{ij} \quad \leq \quad \leq \frac{K-2}{K^2}.$$

Choose three points E, F and G in X_i such that

$$\begin{aligned} E \in K\text{-nn}(F) \quad &\text{and} \quad F \in K\text{-nn}(E) \\ F \in K\text{-nn}(G) \quad &\text{and} \quad G \in K\text{-nn}(E) \\ E \notin K\text{-nn}(G) \quad &\text{or} \quad G \notin K\text{-nn}(E). \end{aligned}$$

an tool example is given in the next picture for bivariate random variables in the unit square with $K = 5$.

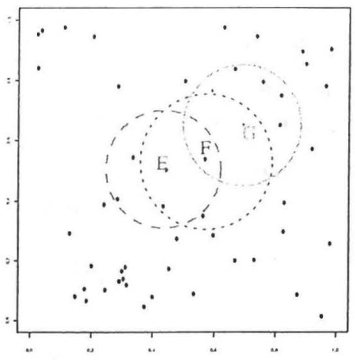

FIG 5. *Example of the configuration of the points E, F and G that lead to a negative eigenvalue of the smoothing matrix for $K = 5$.*

Next, consider the vector u of $\mathbb{R}^n$ that is zero every where except at position e corresponding at point E (respectively f and g) where its value is -1 (respectively 2 and -1). For this choice, we expand $u'Au$ to get

$$\begin{aligned} u'Au &= A_{e,e} + 4A_{f,f} + A_{g,g} - 4A_{e,f} - 4A_{f,g} + 2A_{e,g} \\ &= 6 - \frac{6}{K} - 4A_{e,f} - 4A_{f,g} + 2A_{e,g}. \end{aligned}$$

With the choice of E, F and G, we have

$$u'Au \geq 6 + \frac{2}{K} + 2A_{e,g}.$$

The latter shows that $u'Au > u'u$ whenever

$$A_{e,g} > -\frac{1}{K},$$

which is always true with the choice of points E and G.

Proof of Theorem 3.1 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^tAu where the individual entries A_{ij} of matrix A are equal to

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

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

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

and its Fourier inverse by

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

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

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

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

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

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

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

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

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

We bound the first integral

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

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 $\int_{-\infty}^{\infty} |\widehat{\varphi}(t)|^2 \hat{K}(t)dt < 0$ 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 3.2 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 3.1). 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

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

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.

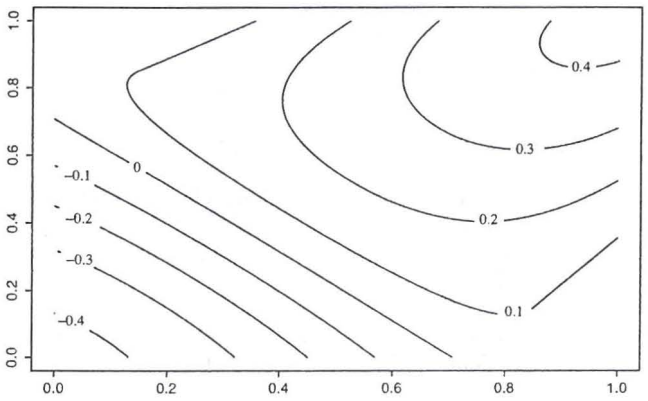

FIG 6. *Contour of an upper bound of* $det(\mathbb{K}[3])$ *as a function of* (x, y).

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] T. Burr, N. Hengartner, E. Matzner-Løber, and S. Mayers. Smoothing low resolution spectral data. *IEEE Transactions on Nuclear Detection (submitted)*, 2008.
[8] W. Cleveland and S. Devlin. Locally weighted regression : an approach to regression analysis by local fitting. *J. Amer. Stat. Ass.*, 83:596–610, 1988.
[9] P.-A. Cornillon, N. Hengartner, and E. Matzner-Løber. Recursive bias estimation and l_2 boosting. Technical report, ArXiv:0801.4629, 2008.
[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 E. Matzner-Løber. Asymptotic unbiased density estimators. *ESAIM*, 2008.
[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. Li. From stein's unbiased risk estimate to the method of generalized cross validation. *Ann. Statist.*, 13:1352–1377, 1985.
[27] 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.
[28] O. Linton and J. Nielsen. A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 82:93–100, 1995.
[29] C. L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.
[30] G. Ridgeway. Additive logistic regression: a statistical view of boosting: Discussion. *Ann. of Statist.*, 28:393–400, 2000.
[31] L. Schwartz. *Analyse IV applications à la théorie de la mesure.* Hermann, Paris, 1993.

[32] G. Schwarz. Estimating the dimension of a model. *Annals of statistics*, 6:461–464, 1978.
[33] J. Simonoff. *Smoothing Methods in Statistics*. Springer, New York, 1996.
[34] C. Stein. Estimation of the mean of a multivariate distribution. *Annals of Statistics*, 9:1135–1151, 1981.
[35] J. Tukey. *Explanatory Data Analysis*. Addison-Wesley, 1977.

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