Recursive Bias estimation for multivariate regression smoothers

Pierre-André Cornillon, Nicolas W. Hengartner, Eric Matzner-Løber

ESAIM: PS 18 (2014) 483–502
DOI: 10.1051/ps/2013046

ESAIM: Probability and Statistics
www.esaim-ps.org

RECURSIVE BIAS ESTIMATION FOR MULTIVARIATE REGRESSION SMOOTHERS

Pierre-André Cornillon[1], N.W. Hengartner[2] and E. Matzner-Løber[3]

Abstract. This paper presents a practical and simple fully nonparametric multivariate smoothing procedure that adapts to the underlying smoothness of the true regression function. Our estimator is easily computed by successive application of existing base smoothers (without the need of selecting an optimal smoothing parameter), such as thin-plate spline or kernel smoothers. The resulting smoother has better out of sample predictive capabilities than the underlying base smoother, or competing structurally constrained models (MARS, GAM) for small dimension ($3 \leq d \leq 7$) and moderate sample size $n \leq 1000$. Moreover our estimator is still useful when $d > 10$ and to our knowledge, no other adaptive fully nonparametric regression estimator is available without constrained assumption such as additivity for example. On a real example, the Boston Housing Data, our method reduces the out of sample prediction error by 20%. An R package **ibr**, available at CRAN, implements the proposed multivariate nonparametric method in R.

Mathematics Subject Classification. 62G07, 62G20.

Received February 20, 2012. Revised May 22, 2013.

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 multivariate predictor 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 exogenous variable x and then estimate locally the conditional expectation $m(x) = \mathbb{E}[Y|X = x]$. This results in nonparametric regression estimators. We refer the interested reader to [10] for a more in depth treatment of various classical regression smoothers. Operationally, the vector of n fitted values at $X_1, \ldots, X_n$ from linear smoothers can be written as $\widehat{\mathrm{m}} = SY$, where S is a $n \times n$ smoothing matrix which depends on observations $X_1, \ldots, X_n$ and on a tuning parameter λ (which we suppress for ease of notation). The tuning parameter 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 explanatory variable over which the responses are averaged.

Keywords and phrases. nonparametric regression, smoother, kernel, thin-plate splines, stopping rules.

[1] IRMAR, UMR 6625, Univ. Rennes 2, 35043 Rennes, France.

[2] Stochastics Group, Los Alamos National Laboratory, Los Alamos, NM 87545, USA.

[3] Lab. Mathématiques Appliquées, Agrocampus Ouest et Univ. Rennes 2, 35043 Rennes, France. eml@uhb.fr

Article published by EDP Sciences

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 almost interpolates the data. For example, the tuning parameter λ is the bandwidth for kernel smoother, the span size for running-mean smoother, the scalar that governs the smoothness penalty term for Thin Plate Splines (TPS)...

It is well known that given n uniformly distributed points in the unit ball $\{x \in \mathbb{R}^d : \|x\| \leq 1\}$, the expected number of points that are covered by a ball centered at the origin with radius $\varepsilon < 1$, is $n\varepsilon^d$. This is to say that covariates in high dimension 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 to balance bias squared and variance does not alleviate this problem.

The challenge of nonparametric estimation in high dimension is also reflected in the optimal rate of convergence. Specifically, when the regression function m mapping $\mathbb{R}^d$ to $\mathbb{R}$ belongs to some finite smoothness functional classes (Hölder, Sobolev, Besov) the optimal mean squared error rate of convergence is $n^{-2\nu/(2\nu+d)}$ where ν is the smoothing index. As a result, common wisdom suggest avoiding all general nonparametric smoothing in moderate dimensions (say $d > 5$) and focus instead on fitting structurally constrained regression models, such as additive [19] and projection pursuit models [14]. The popularity of additive models stems in part from the interpretability of the individual estimated additive components, and from the fact that the estimated regression function converges to the best additive approximation of the true regression function at the optimal univariate mean squared error rate of $n^{-2\nu/(2\nu+1)}$. While additive models do not estimate the true underlying regression function, one hopes for the approximation error to be small enough so that for moderate sample sizes, the prediction mean square error of the additive model is less than the prediction error of a fully nonparametric regression model.

The impact of the curse of dimensionality is lessened for very smooth regression functions. For regression functions with $\nu = 2d$ continuous derivatives, the optimal rate is $n^{-4/5}$, a value recognized as the optimal 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. Since the pioneer work of [22], adaptive nonparametric estimation became a major topic in mathematical statistics, see for example [17]. Adaptive nonparametric estimator can be achieve either by direct estimation (see Lepski's method and related papers) or by aggregation of different procedures, see [31]. This paper presents a practical and simple nonparametric multivariate smoothing procedure that adapts to the underlying smoothness of the true regression function. Our estimator is easily computed by successive application of existing smoothers, such as TPS or kernel smoother.

Section 2 introduces our procedure and motivates it as repeated corrections to the bias of a smoother, where the number of corrections is chosen by Generalized Cross-Validation (GCV). Section 3 applies the iterative bias reduction procedure to multivariate TPS smoothers. TPS smoothers have attractive theoretical properties that facilitate proofs of adaptation to the unknown smoothness of our procedure. However, implementation of the TPS is limited by the need of the sample size to be larger than the size of its parametric component. The latter grows exponentially with the dimension of the covariates d. For practical considerations, we consider, in Section 4, the iterative bias reduction procedure using kernel smoothers and nearest neighbor smoothers. We provide both positive and negative results showing that the desirable properties of iterative bias correction scheme are not universal. In particular, we show that iterative bias correction of nearest neighbor smoothers, and kernel smoothers whose kernel are not positive definite do not enjoy the desirable behavior of TPS. The simulation results presented in Section 5 show that for moderate dimensions of the covariates (*e.g.* $3 \leq d \leq 7$), and sample sizes ranging from $n = 50$ to $n = 800$, our iterated smoother has significantly smaller prediction error than the base smoother with using an "optimal smoothing" parameter. We end this section with the prediction of the classical Boston housing data set ($n = 506$ and $d = 13$). The interested reader can download an R implementation of our procedure with optimized computations for moderate sample size [6]. Finally, the proofs are gathered in the Appendix A.

2. Iterative bias reduction

This section presents the general iterative bias reduction framework for linear regression smoothers and shows that the resulting smoother, when combined with GCV, adapts to the underlying smoothness of the regression function. The advantage of our smoother is its simplicity: we only need to repeatedly estimate the current bias. Suppose that the pairs $(X_i, Y_i) \in \mathbb{R}^d \times \mathbb{R}$ are related through the regression model

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

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 $Y = (Y_1, \ldots, Y_n)^T$ (where ^{T} denotes the matrix transpose), $\mathrm{m} = (m(X_1), \ldots, m(X_n))^T$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$, to get

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

Linear smoothers can be written as

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

where S is an $n \times n$ smoothing matrix and $\widehat{\mathrm{m}} = (\widehat{Y}_1, \ldots, \widehat{Y}_n)^T$, denotes the vector of fitted values. Let I be the $n \times n$ identity matrix. The bias of the linear smoother (2.3), conditionally on the observed values of the covariates $X_1^n = (X_1, \ldots, X_n)$, is

$$\mathbb{E}\left[\widehat{\mathrm{m}} | X_1^n\right] - \mathrm{m} = (S - I)\mathrm{m} = -\mathbb{E}\left[(I - S)Y | X_1^n\right]. \tag{2.4}$$

2.1. Bias reduction of linear smoothers

Let us start with an initial estimator $\hat{m}_1 = S_1 Y$. Expression (2.4) suggests that the bias can be estimated by smoothing the negative residuals $-\mathrm{r}_1 = -(Y - \widehat{\mathrm{m}}_1) = -(I - S)Y$ or alternatively by plugging in an estimator $\tilde{\mathrm{m}}_1 = SY$ for m into the expression of the bias (2.4). In both cases, correcting the pilot smoother $\widehat{\mathrm{m}}_1$ by subtracting the estimated bias yields a *bias corrected* smoother $\widehat{\mathrm{m}}_2$. Since $\widehat{\mathrm{m}}_2$ is itself a linear smoother, it is possible to correct its bias as well. Repeating the bias reduction step $k - 1$ times produces two kind of linear smoother. If we consider using a possibly different smoothing matrices S_k at each iteration k we have:

- The k-times bias corrected smoother obtained by smoothing the current residuals:

$$\begin{aligned} \widehat{\mathrm{m}}_k &= S_1 Y + S_2 (I - S_1) Y + \ldots + S_k (I - S_{k-1}) \ldots (I - S_1) Y \\ &= \left[I - (I - S_k)(I - S_{k-1}) \ldots (I - S_1)\right] Y. \end{aligned} \tag{2.5}$$

- The k-times bias corrected smoother obtained by plugging in an estimator of m in the bias:

$$\begin{aligned} \widehat{\mathrm{m}}_k &= S_1 Y + (I - S_1) S_2 Y + \ldots + (I - S_1)(I - S_2) \ldots S_k Y \\ &= \left[I - (I - S_1)(I - S_2) \ldots (I - S_k)\right] Y. \end{aligned} \tag{2.6}$$

While in general, these two estimates for the bias lead to distinct bias corrected smoothers (2.5) and (2.6), they are identical when the same smoothing matrix is used at every step of the procedure. Taking $S = S_1 = S_2 = \ldots = S_k$, both the plug-in estimator and the residual smoothing estimator agree and the k-times bias corrected smoother can be written as

$$\begin{aligned} \widehat{\mathrm{m}}_k &= \widehat{\mathrm{m}}_0 + \widehat{\mathrm{b}}_1 + \ldots + \widehat{\mathrm{b}}_k \qquad (2.7) \\ &= S\left[I + (I - S) + (I - S)^2 + \ldots + (I - S)^{k-1}\right] Y \\ &= \widehat{\mathrm{m}}_{k-1} + S\mathrm{r}_{k-1} \\ &= \left[I - (I - S)^k\right] Y. \qquad (2.8) \end{aligned}$$

This closed form shows that the behavior of the sequence of iterative bias corrected smoothers $\widehat{\mathrm{m}}_k$ is governed by the spectrum of $I-S$. If the eigenvalues λ_j of $I-S$ are in $[0,1)$ then as k tends to infinity, the bias converges to 0 and the variance increases to $n\sigma^2$ and the resulting smoother will interpolate the data. A reduction in the bias of an estimator increases its variance. Thus for bias correction procedures to be practical, we need to start with a pilot smoother that oversmooths (and hence is heavily biased), and the question of "when to stop the iterative bias reduction procedure" is addressed in Section 2.3.

In the univariate case, smoothers of the form (2.7) arise from the L_2-boosting algorithm with a symmetric base smoother S and a convergence factor μ_k equal to one, see [13] for a definition of this factor. Thus we can interpret the L_2-boosting algorithm as an iterative bias reduction procedure fitting residuals with the same smoother at each 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 [27] to estimate the bias of misspecified multivariate regression models. [26], in the fixed equispaced design in $[0,1]$ twiced the kernel estimator and obtained

$$\widehat{m}_{SM}(x) = \frac{2}{n}\sum_{i=1}^{n} K_h\left(x-x_i\right) Y_i - \frac{1}{n}\sum_{i=1}^{n} K_h\left(x-x_i\right)\sum_{j=1}^{n} K_h\left(x_i-x_j\right) Y_j.$$

The last term is an approximation of $K*K$ (the convolution of K by K). They prove that $K_2 = 2K_h - K_h * K_h$ is a higher order kernel. Using [1] idea of iterating that procedure replacing K for example by K_2 will lead to $K_h + K_2 - K_h * K_2$ which is again a higher order kernel. Iterating that procedure with an initial kernel of order r, one will obtain at step k and $(k-1)r$ order kernel. So iterating in equidistant fixed design our procedure could be seen as choosing the order of the kernel. However, this result does not generalize to non-equidistant or random design. That is, the iterative bias reduction estimator is not equivalent to using higher order kernels.

The idea of iterative debiasing regression smoothers is also present in [2] 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 's discussion of the paper on the statistical interpretation of boosting of [15].

Bühlmann and Yu [4] presented the statistical properties of the L_2-boosted univariate smoothing splines and show that if the true unknown function belongs to a Sobolev space of order μ, the L_2-boosted smoother $\widehat{\mathrm{m}}_k$ achieves the optimal mean squared error convergence rate. We are generalizing their results to the multidimensional case directly without proposing additive boosting.

Di Marzio and Taylor [9] describe the behavior of univariate kernel smoothers after a single bias-correction iteration. They show that the bias is decreased by 2-steps Nadaraya−Watson estimator with its optimal bandwidth compared to the pilot estimator (1 step) with optimal bandwidth. Moreover, whenever the number of steps k is chosen, an optimal bandwidth have to be found.

2.2. Properties of iterative bias corrected smoothers

The squared bias and variance of the kth iterated bias corrected smoother $\widehat{\mathrm{m}}_k$ given in (2.8) are

$$\begin{aligned}\mathbb{E}\left([\widehat{\mathrm{m}}_k|X_1^n]-\mathrm{m}\right)\mathbb{E}\left([\widehat{\mathrm{m}}_k|X_1^n]-\mathrm{m}\right)^T &= (I-S)^k\mathrm{m}\;\mathrm{m}^T\left((I-S)^k\right)^T\\ \operatorname{var}\left(\widehat{\mathrm{m}}_k|X_1^n\right) &= \sigma^2\left(I-(I-S)^k\right)\left(\left(I-(I-S)^k\right)\right)^T.\end{aligned}$$

This shows that the behavior of the sequence of iterative bias corrected smoothers $\widehat{\mathrm{m}}_k$ can be related to the spectrum of $I-S$: if the eigenvalues λ_j of $I-S$ are between 0 and 1, the bias will decrease to 0 as n increases. Not all linear smoothers satisfy the condition on the spectrum of $I-S$. In Section 4, we give examples of common smoothers for which $\lambda_j > 1$, and show numerically that for these shrinkage smoothers, the iterative bias correction scheme fails.

The number of iterations of the bias correction scheme is analogous to smoothing parameters of more classical smoothers: For small numbers of iterations, the smoother is very smooth, becoming increasingly wiggly as the

number of iterations increases, to ultimately interpolate the data. Smoothers at either extreme (oversmoothing or interpolating the data) may have large prediction errors, and the presumptions is that along the sequence of bias corrected smoothers, there will be smoothers that have significantly smaller prediction errors. In Section 3, we show no only that this fact holds for TPS, but that there exists smoothers in that sequence that "adapts to the unknown smoothness" of the regression function and achieves the optimal rate of convergence. Since standard TPS smoothers are not adaptive, this demonstrates the usefulness of iterative bias correction.

2.3. Data-driven selection of the number of steps

The choice of the number of iterations is crucial since each iteration of the bias correction algorithm reduces the bias and increases the variance. Often a few iterations of the bias correction scheme will improve upon the pilot smoother. This brings up to 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 Akaike Information Criteria (AIC), modified AIC [21], Bayesian Information Criterion (BIC), cross-validation, L-fold cross-validation, Generalized cross validation [8], and data splitting. 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 $\mathcal{K}_n = \{1, 2, \ldots, M_n\}$ for the number of iterations. Extensive simulations of the above mentioned model selection criteria, both in the univariate and the multivariate settings [5] have shown that GCV

$$\hat{k}_{\mathrm{GCV}} = \arg\min_{k \in \mathcal{K}_n} \left\{ \log \widehat{\sigma_k}^2 - 2 \log \left(1 - \frac{\mathrm{trace}\,(S_k)}{n} \right) \right\}$$

is a good choice, both in terms of computational efficiencies and of producing good final smoothers and asymptotic results (*cf.* Thm. 3.2). At each iteration, $\widehat{\sigma_k}^2$ corresponds to the estimated variance of the current residuals. Strongly related to the number of iteration is the smoothness of the pilot smoother, since the smoother the pilot is, the bigger is the number of iteration. One has to be sure that the pilot smoother oversmooths. We will discuss that point in the simulation part, since it depends on the type of smoother (TPS, kernel).

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

We study the statistical properties of the iterative bias reduction of multivariate TPS smoothers. Suppose the unknown function m from $\mathbb{R}^d \to \mathbb{R}$ belongs to the Sobolev space $\mathcal{H}^{(\nu)}(\Omega) = \mathcal{H}^{(\nu)}$, where ν is an unknown integer such that $\nu > d/2$ and Ω is an open bounded subset of $\mathbb{R}^d$. Given a smoothing parameter λ, the thin-plate smoother of degree ν_0 minimizes on $\mathcal{H}^{(\nu)}$ [see 16, 30]

$$\sum_{i=1}^{n} \left(Y_i - f\left(X_i\right) \right)^2 + \lambda \left[\sum_{\substack{i_1, \ldots, i_d \geq 0 \\ i_1 + \ldots + i_d \leq \nu_0}} \int_{\mathbb{R}^d} \left| \frac{\partial^{i_1 + \ldots + i_d}}{\partial x_{i_1} \ldots \partial x_{i_d}} f(x) \right|^2 \mathrm{d}x \right]. \tag{3.1}$$

The first part of the functional to be minimized controls the data fitting while the second part, controls the smoothness. TPS are an attractive class of multivariate smoothers for two reasons: first, a closed form solution of (3.1) can be found and it is a linear smoother see [16], and second, the eigenvalues of the smoothing matrix are approximatively known [28].

3.1. Numerical example

The eigenvalues of the associated smoothing matrix lie between zero and one. In light of Section 2.2, the sequence of bias corrected TPS 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

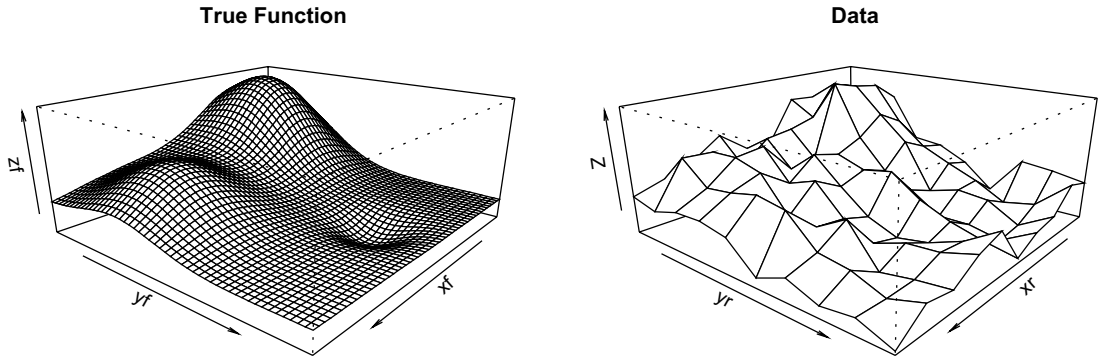

FIGURE 1. True regression function $m(x_1, x_2)$ (3.2) on the square $[0,1] \times [0,1]$ used in our numerical examples and a sample of size 100 with errors and a sample of 100 points.

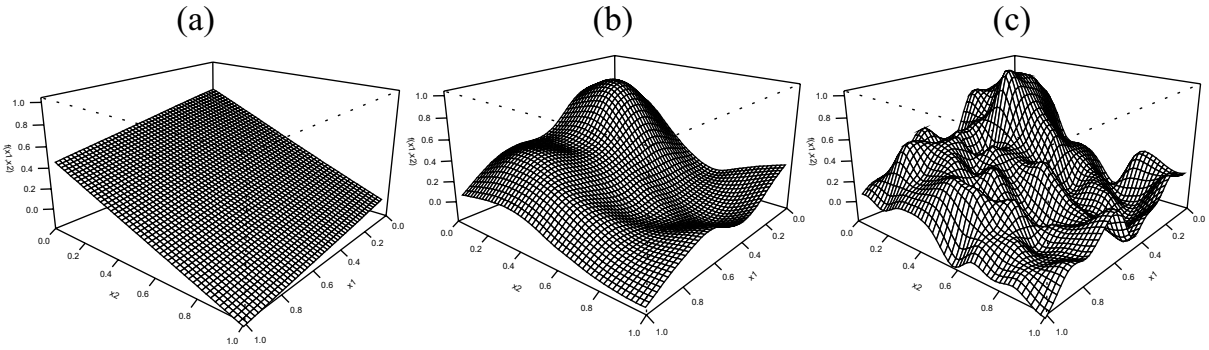

FIGURE 2. TPS regression smoothers from 100 noisy observations from (3.2) (see Fig. 1) evaluated on a regular grid on $[0,1] \times [0,1]$. Panel (a) shows the pilot smoother, panel (b) graphs the bias corrected smoother after 500 iterations and panel (c) graphs the smoother after 50 000 iterations of the bias correction scheme.

steps, the resulting bias corrected smoother will be a good estimate for the true underlying regression function. This behavior is confirmed numerically in the following pedagogical example of a bivariate regression problem: Figure 1 graphs Wendelberger's test function [29]

$$m(x,y) = \frac{3}{4}\exp\left(-\frac{(9x-2)^2+(9y-2)^2}{4}\right) + \frac{3}{4}\exp\left(-\frac{(9x+1)^2}{49} + \frac{(9y+1)^2}{10}\right) + \frac{1}{2}\exp\left(-\frac{(9x-7)^2+(9y-3)^2}{4}\right) - \frac{1}{5}\exp\left(-(9x-4)^2-(9y-7)^2\right) \tag{3.2}$$

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). At iteration $k = 500$ (or 499 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. This example shows the importance of suitably selecting the number of bias correction iterations.

3.2. Adaptation to smoothness of the regression function

Let Ω be an open set of $\mathbb{R}^d$ satisfying an uniform cone condition and having a Lipschitz boundary see [28]. 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 $\nu_0 \leq \nu$ which is symmetric and admits $M_0 = \binom{\nu_0+d-1}{\nu_0-1}$ eigenvalues equal to one (in practice we will take the smallest possible value $\nu_0 = \lfloor d/2 \rfloor + 1$) and fix the smoothing parameter $\lambda_0 > 0$ to some reasonably large value. Our next Theorem states that there exists a number of iterations $k = k(n)$, depending on the sample size, for which the resulting estimate $\widehat{\mathrm{m}}_k$ achieves the optimal 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 optimal rate of convergence.

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

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

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 $\widehat{\mathrm{m}}_1 = SY$ is obtained with S a TPS of degree ν_0, with $\lfloor d/2 \rfloor + 1 \leq \nu_0 < \nu$ and a fixed smoothing parameter $\lambda_0 > 0$ not depending on the sample size n, then there is an optimal number of iterations $k(n)$ such that the resulting smoother $\widehat{\mathrm{m}}_k$ satisfies

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

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

Obviously, the hypothesis on the design implies that Ω is bounded. This hypothesis is fulfilled for example for uniform design see [4]. 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. Rate optimality of the smoother $\widehat{\mathrm{m}}_k$ 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. Selecting the optimal number of iterations is important and we prove that result with GCV criterion using Theorem 3.2 of [23].

Theorem 3.2. *Suppose that the hypothesis on Ω and the design of Theorem 3.1 are fulfilled and the initial estimator is the same as in Theorem 3.1. Let $\hat{k}_{\mathrm{GCV}} \in \mathcal{K}_n = \{1, \ldots, \lfloor n^\gamma \rfloor\}$, $1 \leq \gamma \leq (2\nu_0)/d$, denote the index in the sequence of bias corrected smothers whose associated smoother minimize the generalized cross-validation criteria. Suppose that the noise ε in (2.1) has finite $4q$th absolute moment, where $q > \gamma(2\nu/d + 1)$, that is, $\mathbb{E}[|\varepsilon|^{4q}] < \infty$. Then as the sample size n grows to infinity,*

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

The moment condition is satisfied for Gaussian or subgaussian errors.

4. Iterative bias reduction of kernel and K-nearest neighbor smoothers

The matrix S of TPS is symmetric and has eigenvalues in $(0, 1]$ see for example [28]. In particular, the first $M_0 = \binom{\nu_0+d-1}{\nu_0-1}$ eigenvalues are all equal to one, corresponding to the parametric component of the smoothing spline. The sample size n needs to be at least M_0, and since from Theorem 3.1 we want $\nu_0 > d/2$, it follows that M_0 grows exponentially fast in the number of covariates d. In particular the dimension of the parametric component is 5, 28, 165, 1001 for $d = 4$, 6, 8, 10, respectively, and more generally, M_0 grows like $3^{d/2} \times (3/2)^d$ for large d. This feature limits the practical usefulness of TPS smoothers. For example, the regression model in Section 5 for the Boston housing data set that has 13 covariates can not be fit with a TPS because its sample size $n = 506 < 27\,500 \approx M_0$.

A possible solution is to approximate the TPS smoother with a kernel smoother, with an appropriate kernel. In this section, we discuss kernel based smoothers and we give a necessary and sufficient condition on the kernel that ensures that the iterative bias correction scheme is well behaved.

4.1. Nearest neighbor smoother

Our first result is that K-nearest neighbor (KNN) smoothers are not suited for the iterative bias reduction scheme (L_2 boosting) because the matrix $I - S$ has eigenvalues larger than one. This result is somewhat surprising, as nearest neighbor classifiers are the weak classifiers of choice for boosting algorithm of machine learning. Recall that the smoothing matrix of the K-nearest neighbor smoother has entries $S_{ij} = 1/K$ when X_j belongs to the K-nearest neighbor of X_i, and $S_{ij} = 0$ otherwise. By definition X_i does not belong to its K-nearest neighbor, so that $S_{ii} = 0$. It follows that the trace of S is zero, and since S is a stochastic matrix, it has at least one eigenvector with eigenvalue equal to one. It follows that there exists an eigenvector having a negative eigenvalue, which implies that the spectrum of $I - S$ is not contained in the unit interval $[0, 1]$.

Lemma 4.1. *Let S be the smoothing matrix of the K nearest neighbor smoother with $K \geq 1$. Then S has at least one negative eigenvalue.*

4.2. Kernel type smoothers

The matrix S of kernel 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 $d_h(x, y) = \sqrt{\sum_{j=1}^d (x_j - y_j)^2/h_j^2}$, where $h = (h_1, \ldots, h_d)$ denotes the bandwidth vector, gives rise to elliptic neighborhoods. While the smoothing matrix S is not symmetric, it has a real spectrum. 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

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

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 results of Section 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 4.2. *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

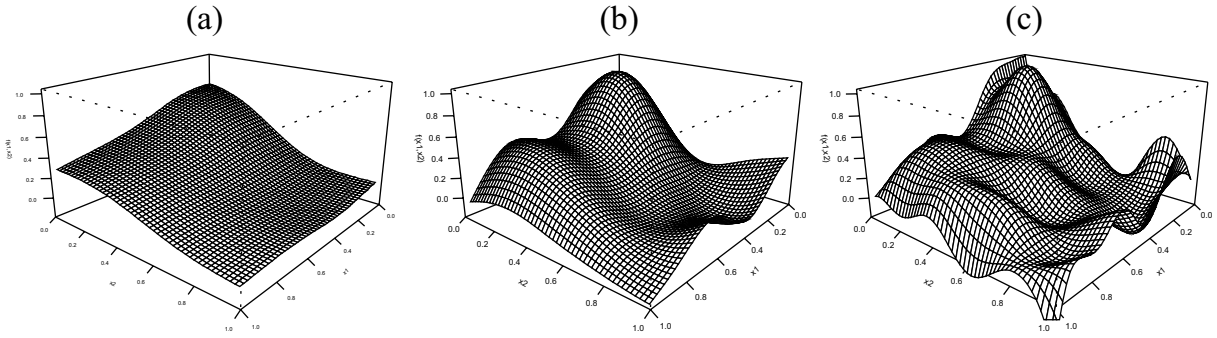

FIGURE 3. Gaussian kernel smoother of $m(x_1, x_2)$ from $n = 100$ equidistributed points on $[0,1] \times [0,1]$, evaluated on a regular grid with (a) $k = 1$, (b) 50 and (c) 10 000 iterations.

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 4.3. 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-definite function, that is, for any finite set of points $x_1, \ldots, x_m$, the matrix $\mathbb{K}$ is positive definite. We refer to [25] for a detailed study of positive definite functions.

Remark 4.4. [9] 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 eigenvalues.

The Gaussian and triangular kernels are positive definite kernels (they are the Fourier transform of a finite positive measure, [11]). In light of Theorem 4.2, the iterative bias correction of Nadaraya–Watson kernel smoothers with these kernels produces a sequence of well behaved 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)) produce a fit that is visually similar to the original function. Continued bias corrections then slowly degrade the fit as the smoother starts to over-fit the data. Continuing the bias correction scheme will eventually lead to a smoother that interpolates the data. This example hints at the potential gains that can be realized by suitably selecting the number of bias correction steps.

The uniform and the Epanechnikov kernels are not positive definite. Theorem 4.2 states that for large enough samples, we expect with high probability that $I-S$ has at least one eigenvalue larger than one. When this occurs, the sequence of iterative bias corrected smoothers will behave erratically and eventually diverge. Lemma 4.5 below strengthens this result by giving an explicit condition on the configurations of the design points for which the largest eigenvalue of $I - S$ is always larger than one.

Lemma 4.5. *Denote by $\mathcal{N}_i$ the following set: $\{X_j : K(d_h(X_j, X_i)) > 0\}$.*

If there exists a set $\mathcal{N}_i$ which contains (at least) two points X_j, X_k different of 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 $\mathcal{N}_i$ that contains (at least) two points X_j, X_k different of X_i that satisfy

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

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

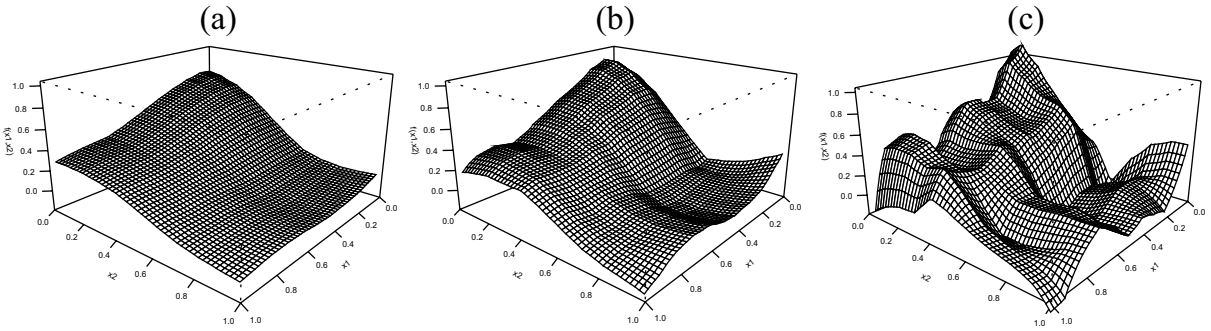

FIGURE 4. Epanechnikov kernel smoother of $m(x_1, x_2)$ from $n = 100$ equidistributed points on $[0, 1] \times [0, 1]$, evaluated on a regular grid with (a) $k = 1$, (b) 5 and (c) 25 iterations.

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 function 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.

5. Simulations and a real example

This section presents the results of a modest simulation study to compare the empirical mean squared error

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(\hat{m}\left(X_i\right) - m\left(X_i\right) \right)^2 \tag{5.1}$$

of our procedure to its competitors for two functions, in dimensions $d = 3$, 5, 7 and sample sizes $n = 50$, 100, 200, 500, 800, with a noise to signal ratio of 10%. In order to exploit our theoretical result, the pilot smoother has to oversmooth otherwise the pilot smoother will have very small bias and our iterative debiasing procedure has no more justification. So starting with a small λ will lead to zero or a small number of iterations. Oppositely, starting with a big λ will normally lead to a large number of iterations. We decide in this section to use the values by default in the ibr R-package.

The TPS is governed by a single parameter λ that weights the contribution of the roughness penalty. For estimating a d-valued regression function, the parametric component is $M_0 = \binom{\nu_0+d-1}{\nu_0-1}$ and we choose λ such that the initial degree of freedom of the pilot smoother equals 1.5 M_0. The implementation for the kernel smoother is different since we could choose a different bandwidth for each explanatory variables. We choose one bandwidth for each explanatory variable X_i such as the effective degree of freedom for the one-dimensional smoothing matrix related to X_i has a trace equal to 1.1 (more degree than a constant but less than a linear model). For such values, the pilot smoothers always oversmooth.

Our simulations was designed to allow us to investigate three aspects: first, compare the performance of the TPS with smoothing parameter selected by GCV with the IBR smoother using a TPS with a large smoothing parameter. We expect that adaptation of our method will translate into a better performance of our smoother over the optimal TPS smoother. Second, to compare the performance between ibr smoother using either TPS and kernel based smoothers. Since kernel smoothers do not have a parametric component (which may, or may not, be needed to fit the data), we believe that kernel smoothers use more effectively their degree of freedom, which translates into better performance. Third, we want to compare the performance of fully nonparametric smoothers and additive smoothers. While with additive models we estimate an approximation of the true

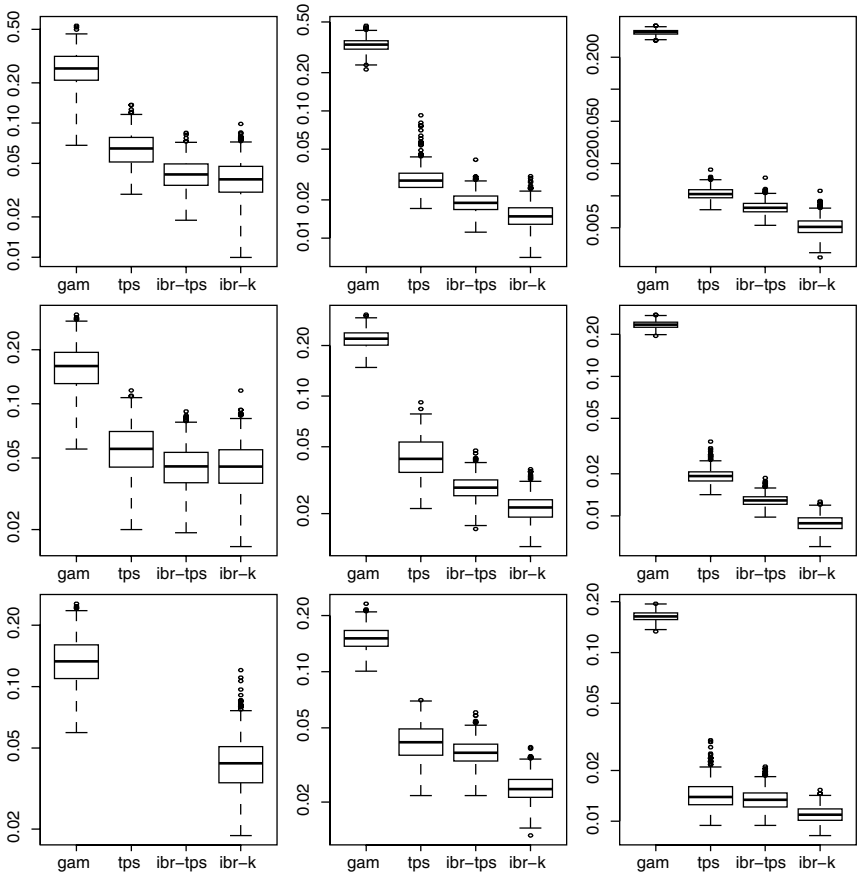

FIGURE 5. Boxplot of Mean Squared Error (MSE) of smoothers for the regression functions (from top to bottom) of three variables $\sin(2\pi(x_1x_2)^{1/2}) + \cos(2\pi(x_2x_3)^{1/2})$, five variables $\sin(2\pi(x_1x_2x_3)^{1/3}) + \cos(2\pi(x_3x_4x_5)^{1/3})$ and seven variables $\sin(2\pi(x_1x_2x_3x_4)^{1/4}) + \cos(2\pi(x_4x_5x_6x_7)^{1/4})$, and of sample size (from left to right) of $n = 50$, 200, 800. Each panel shows the boxplot of the MSE of a GAM smoother, TPS smoother, ibr with TPS smoother and ibr with kernel smoother.

regression function, it is generally believed that the approximation error of an additive model is smaller than the estimation error of a fully multivariate smoother even for dimensions for small sample sizes, *e.g.* $n = 50$, 100, and moderate dimensions of the covariates, *e.g.* $d = 5$. The results are summarized in Table 1 and Figure 5.

Figure 5 shows nine panels each containing the boxplots of the MSE from 500 simulations, on a logarithmic scale on the y-axis. Moving from top to bottom ranges the regression functions from the function of three variables $\sin(2\pi(x_1x_2)^{1/2}) + \cos(2\pi(x_2x_3)^{1/2})$, to the function of five variables $\sin(2\pi(x_1x_2x_3)^{1/3}) + \cos(2\pi(x_3x_4x_5)^{1/3})$ and to the function of seven variables $\sin(2\pi(x_1x_2x_3x_4)^{1/4}) + \cos(2\pi(x_4x_5x_6x_7)^{1/4})$. All the covariates are i.i.d. uniforms on the interval $(1, 2)$. Moving from left to right changes the sample size from $n = 50$, 200, 800. Within each panel, the boxplot of MSE is shown, in the order from left to right, of additive models using the function **gam** from the R package **mgcv**, TPS with optimal smoothing parameter using the function **Tps** from the R package **fields**, iterative bias reduction with TPS smoother using the function **ibr** from the **ibr** R package and iterative bias reduction with kernel smoothers, using again the **ibr** function. For reasons explained in Section 4, no TPS smoothers can be evaluated for the $d = 7$, $n = 50$ panel.

TABLE 1. Ratio of median MSE over 500 simulations of a smoother and the median MSE over 500 simulations of the kernel based IBR smoother. The smoothers, from left to right, are Generalized Additive Model (GAM), TPS with optimally selected smoothing parameter (tps), TPS based ibr (ibr-tps) and kernel based ibr (ibr-k).

function	n	gam	tps	ibr-tps	ibr-k
$x_1x_2x_3$	50	2.59	1.63	1.39	1
	100	4.59	1.89	1.58	1
	200	8.38	2.14	1.73	1
	500	17.9	2.56	2.08	1
	800	27.4	2.82	2.39	1
$\sin(2\pi(x_1x_2)^{1/2}) + \cos(2\pi(x_2x_3)^{1/2})$	50	6.72	1.70	1.09	1
	100	12.0	1.80	1.19	1
	200	22.3	1.91	1.27	1
	500	46.2	1.99	1.45	1
	800	67.3	2.04	1.51	1
$x_1x_2x_3x_4x_5$	50	2.16	1.60	1.47	1
	100	3.83	1.42	1.39	1
	200	6.64	1.28	1.24	1
	500	13.17	1.24	1.22	1
	800	19.44	1.26	1.23	1
$\sin(2\pi(x_1x_2x_3)^{1/3}) + \cos(2\pi(x_3x_4x_5)^{1/3})$	50	3.62	1.26	1	1
	100	6.32	1.76	1.15	1
	200	10.0	1.95	1.31	1
	500	18.6	2.06	1.38	1
	800	26.5	2.18	1.46	1
$x_1x_2x_3x_4x_5x_6x_7$	50	2.05	–	–	1
	100	3.11	–	–	1
	200	5.26	3.53	3.17	1
	500	9.85	2.46	2.45	1
	800	13.8	2.07	2.07	1
$\sin(2\pi(x_1x_2x_3x_4)^{1/4}) + \cos(2\pi(x_4x_5x_6x_7)^{1/4})$	50	3.16	–	–	1
	100	4.38	–	–	1
	200	6.43	1.78	1.57	1
	500	11.1	1.37	1.31	1
	800	14.9	1.27	1.22	1

Figure 5 shows that a fully nonparametric smoother is always preferred to an additive smoother, even for relative small sample sizes and moderate dimensions.

In extensive simulations [7], we observe that this qualitative conclusion holds over a wide variety of regression functions. Generally, as expected, the TPS with optimal smoothing parameter has a somewhat worse performance than the TPS ibr smoother. And finally, the kernel based ibr smoother is slightly better than the TPS based ibr smoother, especially in higher dimensions.

Table 1 gives further insight into the performance of the various smoothers. Our table presents the ratio of the median MSE (in 500 simulation runs) of various smoothers to the median MSE of the kernel based ibr smoother. Since all the entries are larger than one, we conclude that kernel based ibr consistently outperforms the other smoothing procedures over the range of sample size, number of covariates and regression functions we considered in our study.

The improvement over a GAM model ranges from 100% to 6000%. This reinforces our conclusions that fully nonparametric regressions are practical for moderately large number of covariates, even for sample sizes as small as $n = 50$. The other notable observation is that the values in the ibr-tps column are always less than those in

TABLE 2. Predicted mean Squared Error on test observations for Boston housing data.

Method	Mean predicted squared error
Multivariate regression	20.09
L_2Boost with component-wise spline	9.59
additive model (backfitted with R)	11.77
Projection pursuit (with R)	12.64 (4)
MARS (with R)	10.54
ibr kernel with 1.1 initial DDL per variable and 1230 iterations	7.35

the tps column, showing that consistently, the TPS based ibr smoother has better performance than TPS with optimal smoothing parameter. In our simulation study, the typical improvement is of 20%.

5.1. Boston housing data

We apply our method on the Boston housing data. This dataset, created by [18] has been extensively to showcase the performance and behavior of nonparametric multivariate smoothers, see for example [3] and more recently by [9]. The data contains 13 explanatory variables describing each of 506 census tracts in the Boston area taken from the 1970 census, together with 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 compare our method with the MARS algorithm of [12] as implemented in the R package **mda**, with projection pursuit regression (function **ppr**), additive models using the backfitting algorithm of [19] as implemented in the R package **mgcv**, and additive Boosting [4] from the R package **mboost**. The predicted mean squared error is estimated by randomly splitting 30 times the data into training sets (size $n = 350$) and testing sets ($n = 156$). We summarize the results of our analysis in the following table:

Table 2 again supports our claim that the fully multivariate method presented in the paper leads to a reduction of more than 30% in the prediction mean squared error over competing state-of-the-art multivariate smoothing methods. A similar comparison for responses on the logarithmic scale reveals the even larger reduction of 40% in the prediction mean squared error. Since our fully nonparametric regression smoother has substantially smaller prediction error over additive linear models and low-order interaction models, we conclude that there exist higher order interactions in that data that are significant.

6. CONCLUSION

This paper introduces a fully multivariate regression smoother for estimating the regression function m obtained by successive bias correction from a very smooth (biased) pilot smoother. We show that the resulting smoother is adaptive to the underlying smoothness (see Thms. 3.1 and 3.2). This adaptation to the underlying smoothness partially mitigates the effect from the curse of dimensionality in many practical examples, and make it practical to use fully nonparametric smoother in moderate dimensions, even for smaller sample sizes.

As in L_2 boosting, the proposed iterative bias correction scheme needs a weak learner as a base smoother S, but all weak learners are not suitable. For instance, Epanechnikov kernel smoothers are not interesting (see Thm. 4.2). We further 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. Finally, the R package **ibr** available at CRAN implements the proposed multivariate nonparametric method in R.

Appendix A.

We are omitting the proof of Theorem 3.1 since it is an multivariate extension of the proof given in [4] using results of [28] about the eigen decomposition of S.

Proof of Theorem 3.2. We show that conditions **(A.1)** to **(A.7)** given by [23], in Theorem 3.2 are satisfied. To make the proof self contained, we recall briefly these conditions:

(A.1) $\lim_{n\to\infty} \sup_{k\in\mathcal{K}_n} \lambda(S_k) < \infty$;
(A.2) $E(\varepsilon^{4q}) < \infty$;
(A.3) $\sum_{k\in\mathcal{K}_n} (nR_n(k))^{-q} \to 0$; where $R_n(k) = \mathbb{E}(\|\mathrm{m}_n - \widehat{\mathrm{m}}_{k,n}\|^2)/n$;
(A.4) $\inf_{k\in\mathcal{K}_n} n^{-1}\|\widehat{\mathrm{m}}_{k,n} - \mathrm{m}_n\|^2 \to 0$, in probability;
(A.5) for any sequence $\{k_n \in \mathcal{K}_n\}$ such that $n^{-1}\mathrm{trace}(S_{k_n}S_{k_n}^T) \to 0$ we have $\frac{\{n^{-1}\mathrm{trace}(S_{k_n})\}^2}{n^{-1}\mathrm{trace}(S_{k_n}S_{k_n}^T)} \to 0$;
(A.6) $\sup_{k\in\mathcal{K}_n} n^{-1}\mathrm{trace}(S_k) \leq \gamma_1$ for some $1 > \gamma_1 > 0$;
(A.7) $\sup_{k\in\mathcal{K}_n} \{n^{-1}\mathrm{trace}(S_k)\}^2/\{n^{-1}\mathrm{trace}(S_kS_k^T)\} \leq \gamma_2$ for some $1 > \gamma_2 > 0$.

Conditions (A.1) to (A.4)
The eigenvalues of S_k (denoted $\lambda_j(S_k), 1 \leq j \leq n$ or λ_j for brevity) are between 0 and 1 for all n. The first $M_0 = \binom{\nu_0+d-1}{\nu_0-1}$ eigenvalues are equal to one and the remaining $\lambda_j, M_0 < j \leq n$ are strictly less than 1 and greater than 0. Thus the condition **(A.1)** is fulfilled.

To fulfill condition **(A.3)** we need to calculate $\sum_{k\in\mathcal{K}_n} nR_n(k)^{-q}$, where q is an integer to be found, $\mathrm{m}_n = (m(X_1), \ldots, m(X_n))^T$ and $\widehat{\mathrm{m}}_{k,n} = S_kY$. Using Theorem 3.1 we have that for an optimal choice of k, $R_n(k) = O(n^{d/(2\nu+d)})$. Let us choose $\mathcal{K}_n$ such that its cardinal is of order n^γ $(1 \leq \gamma \leq (2\nu_0)/d)$, the order of an upper bound of $\sum_{k\in\mathcal{K}_n} nR_n(k)^{-q}$ is $n^{\gamma-\frac{qd}{2\nu+d}}$. To have **(A.3)** fulfilled we need that $\gamma - \frac{qd}{2\nu+d} < 0$, that is $q > \gamma(2\nu/d+1)$.

Condition **(A.4)** is satisfied because of Theorem 3.1.

Conditions **(A.5)** to **(A.7)** are related the trace of the matrices S_k and S_k^2. Recall that

$$\frac{1}{n}\mathrm{trace}(S_k) = \frac{1}{n}\left(M_0 + \sum_{j=M_0+1}^{n} \left[1 - (1-\lambda_j)^k\right]\right),$$

where $M_0 = \binom{\nu_0+d-1}{\nu_0-1}$ is the number of equal to 1 and the remaining $n - M_0$ eigenvalues λ_j are less than 1, bigger than 0 and decreasing. We have for all k, that

$$1 \geq \frac{1}{n}\mathrm{trace}(S_k) \geq \frac{1}{n}\mathrm{trace}(S_k^2) \geq \left(\frac{1}{n}\mathrm{trace}(S_k)\right)^2.$$

It follows that both $\mathrm{trace}(S_k)$ and $\mathrm{trace}(S_k^2)$ are increasing with k, and we have $\lim_{k\to\infty}\left(\frac{1}{n}\mathrm{trace}(S_k)\right) = 1$ and $\lim_{k\to\infty}\frac{1}{n}\mathrm{trace}(S_k^2) = 1$. When all the eigenvalues of S_k equal 1 the corresponding smoother interpolates and it is statistically inappropriate, let's fix $\max_{k\in\mathcal{K}_n} = n^\gamma$ for some $\gamma < \alpha_0$, which will enable us to stop the iteration step before reaching to interpolation. Thanks to [28], we have, when the smoothing parameter is $\lambda_0 > 0$, the following approximation:

$$\lambda_j \approx \frac{1}{1+\lambda_0 j^{\alpha_0}}, \quad \alpha_0 = \frac{2\nu_0}{d} > 1.$$

Let us write

$$(1-\lambda_j)^k \approx \left[\frac{\lambda_0 j^{\alpha_0}}{1+\lambda_0 j^{\alpha_0}}\right]^k = \left(1+\lambda_0^{-1}j^{-\alpha_0}\right)^{-k},$$

from which it follows that

$$\begin{aligned}\frac{1}{n}\text{trace}(S_k) &\approx \frac{1}{n}M_0 + \frac{1}{n}\sum_{j=M_0+1}^{n}\left(1-\left[1+\lambda_0^{-1}j^{-\alpha_0}\right]^{-k}\right)\\ &\approx \frac{1}{n}M_0 + \frac{1}{n}\sum_{j=M_0+1}^{n} g_k(j).\end{aligned}$$

For fixed k and $j_n < n$, we have that

$$\begin{aligned}g_k(j_n) &= 1-\left[1+\lambda_0^{-1}j_n^{-\alpha_0}\right]^{-k}\\ &= 1-\exp\left[-k\ln\left(1+\lambda_0^{-1}j_n^{-\alpha_0}\right)\right]\\ &\approx 1-\exp\left[-k\lambda_0^{-1}j_n^{-\alpha_0}\right].\end{aligned}$$

Let us consider the case where j_n tends to infinity and $-k_n j_n^{-\alpha_0}$ tends to zero. Thus when n grows to infinity, $\forall j \geq j_n$ we have the following approximation for $g_k(j)$:

$$g_k(j) \approx kj^{-\alpha_0}\lambda_0^{-1}. \tag{A.1}$$

Order of an upper bound of $\text{trace}(S_k)/n$.
For all $k \in \mathcal{K}_n$, we have when n grows to infinity:

$$\begin{aligned}\frac{1}{n}\text{trace}(S_k) &\approx \frac{M_0}{n} + \frac{1}{n}\sum_{j=M_0+1}^{j_n} g_k(j) + \frac{1}{n}\sum_{j=j_n+1}^{n} g_k(j)\\ &\lessapprox \frac{j_n}{n} + \frac{1}{n}\int_{j_n}^{n} g_k(j)dj\\ &\lessapprox \frac{j_n}{n} + \frac{kj_n^{1-\alpha_0}}{n(\alpha_0-1)}\lambda_0^{-1}\end{aligned}$$

with the last approximation which follows from equation (A.1). Choosing j_n such that $kj_n^{-\alpha_0}$ tends to zero, we have an upper bound for $\text{trace}(S_k)/n$ of order $\frac{j_n}{n}$.

Order of a lower bound of $\text{trace}(S_k^2)/n$.
For all $k \in \mathcal{K}_n$, we have when n grows to infinity that

$$\begin{aligned}\frac{1}{n}\text{trace}(S_k^2) &\approx \frac{M_0}{n} + \frac{1}{n}\sum_{j=M_0+1}^{j_n} g_k^2(j) + \frac{1}{n}\sum_{j=j_n+1}^{n} g_k^2(j)\\ &\geq \frac{M_0}{n} + \frac{g_k^2(j_n)}{n}(j_n - M_0) + \frac{1}{n}\int_{j_n+1}^{n+1} g_k^2(j)dj\\ &\approx \frac{M_0}{n} + \frac{j_n - M_0}{n}g_k^2(j_n) + \frac{k^2\lambda_0^{-2}}{n}\frac{(j_n+1)^{-2\alpha_0+1}}{2\alpha_0-1} - \frac{k^2\lambda_0^{-2}}{n}\frac{(n+1)^{-2\alpha_0+1}}{2\alpha_0-1}\end{aligned}$$

with the last approximation which follows from equation (A.1).

Again, choosing j_n such that $kj_n^{-\alpha_0}$ tends to zero, we get that a lower bound for $\frac{1}{n}\text{trace}(S_k^2)$ is either of order $\frac{1}{n}$ or of order $\frac{k^2 j_n^{-2\alpha_0+1}}{n}$ if $k^2 j_n^{-2\alpha_0+1} \to \infty$.

Condition (A.5) and (A.7)
For $n > M_0$, we have at least two different eigenvalues for S_k. Thus the empirical variance of the eigenvalues of S_k is positive *i.e.* $\frac{1}{n}\sum_{i=1}^{n}(\lambda_i - \bar{\lambda})^2 > 0$. This implies that for given $n > M_0$

$$(\text{trace}(S_k)/n)^2(\text{trace}(S_k^2)/n)^{-1} < 1.$$

For n growing to infinity, let us show that $(\text{trace}(S_k)/n)^2(\text{trace}(S_k^2)/n)^{-1}$ tends to 0 for all $k \in \mathcal{K}_n$ (implying condition **(A.5)**).

Let us partition the grid $\mathcal{K}_n$ in two parts: $\mathcal{K}_n^{(1)} = \{1, \ldots, \lfloor n^{\frac{1}{10}} \rfloor - 1\}$ and $\mathcal{K}_n^{(2)} = \{\lfloor n^{\frac{1}{10}} \rfloor, \ldots, n^{\gamma}\}$. Consider the following two cases:

- For $k \in \mathcal{K}_n^{(1)}$;
  choose $j_n = n^{\frac{1}{10}}$, so $k_n j_n^{-\alpha_0}$ tends to zero, thus, the previous calculated order can be used and we get an upper bound for $(\text{trace}(S_k)/n)^2(\text{trace}(S_k^2)/n)^{-1}$ of order $\frac{\frac{j_n^2}{n^2}}{\frac{1}{n}} = \frac{j_n^2}{n}$. Thus,

  $$(\text{trace}(S_k)/n)^2(\text{trace}(S_k^2)/n)^{-1} \to 0.$$

- For $k \in \mathcal{K}_n^{(2)}$;
  for a given k, we have that $k = O(n^{\beta_1})h(n)$ with $h(n) = o(n^{\beta_3}),\ \forall \beta_3 > 0$. We have that $\frac{1}{10} \leq \beta_1 \leq \gamma < \alpha_0$. We further assume that:
  (a) $k j_n^{-\alpha_0} \to 0$
  (b) $k^2 j_n^{-2\alpha_0+1} \to \infty$.
  These conditions are satisfied for

  $$j_n = n^{\beta_2}, \text{ with } \beta_2 = \frac{1}{\alpha_0}\Big(\frac{\alpha_0 - \beta_1}{\eta(2\alpha_0 - 1)} + \beta_1\Big) \text{ and } \eta = \max(10\alpha_0, 3)$$

  Indeed, condition (a) is obviously satisfied. Let us verify condition (b):

  $$\begin{aligned} 2\beta_1 - 2\beta_2\alpha_0 + \beta_2 &= 2\beta_1 - 2\frac{\alpha_0 - \beta_1}{\eta(2\alpha_0 - 1)} - 2\beta_1 + \frac{1}{\alpha_0}\left(\frac{\alpha_0 - \beta_1}{\eta(2\alpha_0 - 1)} + \beta_1\right) \\ &= \frac{\alpha_0 - \beta_1}{\eta(2\alpha_0 - 1)}\frac{1 - 2\alpha_0}{\alpha_0} + \frac{\beta_1}{\alpha_0} = \frac{\beta_1 + \beta_1\eta - \alpha_0}{\eta\alpha_0} > 0, \end{aligned}$$

  with the last inequality following from the fact that $\beta_1 \geq 1/10$ and $\eta \geq 10\alpha_0 > 10\alpha_0 - \frac{1}{10}$.
  Using the previous calculated order, we get that the upper bound of $(\text{trace}(S_k)/n)^2(\text{trace}(S_k^2)/n)^{-1}$ is of order $\frac{j_n^{2\alpha_0+1}}{nk^2}$. This quantity tends to 0 because

  $$\begin{aligned} (2\alpha_0 + 1)\beta_2 - 1 - 2\beta_1 &= \beta_2 + 2\frac{\alpha_0 - \beta_1}{\eta(2\alpha_0 - 1)} + 2\beta_1 - 1 - 2\beta_1 = \frac{\alpha_0 - \beta_1}{\alpha_0\eta(2\alpha_0 - 1)} + \frac{\beta_1}{\alpha_0} + 2\frac{\alpha_0 - \beta_1}{\eta(2\alpha_0 - 1)} - 1 \\ &= \frac{\alpha_0 - \beta_1}{\alpha_0\eta}\left(\frac{1}{2\alpha_0 - 1} + \frac{2\alpha_0}{2\alpha_0 - 1}\right) + \frac{\beta_1 - \alpha_0}{\alpha_0} = \frac{\alpha_0 - \beta_1}{\alpha_0\eta}\left(\frac{2\alpha_0 + 1}{2\alpha_0 - 1} - \eta\right) < 0 \end{aligned}$$

  with the last inequality following from the fact that $\frac{2\alpha_0+1}{2\alpha_0-1} < 3$ (as $\alpha_0 > 1$) and $\eta > 3$.

Let us denote $k_n^* = \arg\max_{k \in \mathcal{K}_n}(\text{trace}(S_k)/n)^2(\text{trace}(S_k^2)/n)^{-1}$ (for a given n, $\mathcal{K}_n$ is finite). For all $k \in \mathcal{K}_n$ we have the following limit: $(\text{trace}(S_k)/n)^2(\text{trace}(S_k^2)/n)^{-1} \to 0$. It implies that $(\text{trace}(S_{k_n^*})/n)^2(\text{trace}(S_{k_n^*}^2)/n)^{-1} \to 0$. It exists a finite $N_{k_n^*}$ such that for all $n > N_{k_n^*}$ $(\text{trace}(S_{k_n^*})/n)^2(\text{trace}(S_{k_n^*}^2)/n)^{-1}$ will be less than 1/2. This implies that condition **(A.7)** hold for all $n > N_{k_n^*}$. For all $n \leq N_{k_n^*}$, n is finite (and $k \in \mathcal{K}_n$ too)

and $(\mathrm{trace}(S_k)/n)^2(\mathrm{trace}(S_k^2)/n)^{-1} \leq \max_{k\in\mathcal{K}_n, n\leq N_{k_n^*}}(\mathrm{trace}(S_k)/n)^2(\mathrm{trace}(S_k^2)/n)^{-1}) < 1$. Thus the condition **(A.7)** holds.

Condition (A.6)
The n eigenvalues are non-increasing. Take $j_n = n\zeta$ with ζ fixed and less than one. We have that the maximal value of the mean of the trace which occurs at $k_n = n^\gamma$ is bounded by

$$\frac{1}{n}\mathrm{trace}(S_k) \leq \frac{j_n}{n} + \frac{(n-j_n)}{n} k_n j_n^{-\alpha_0}.$$

We can easily show that the last quantity is less than a given value smaller than 1. The aim of setting k_n equal to n^γ is to ensure that at the border of grid $\mathcal{K}_n$, the smoother is not the identity, *i.e.* we are not interpolating the data. When the smoother is too close to the identity matrix, conditions **(A.6)** and **(A.7)** are not longer fulfilled. Moreover, being very close to identity is not interesting from a statistical viewpoint.

Proof of Theorem 4.2. For notational simplicity, 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

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

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 (and its Fourier inverse) by

$$\hat{\varphi}(t) = \int \mathrm{e}^{-2i\pi t x}\varphi(x)\mathrm{d}x \qquad \hat{\varphi}_{\mathrm{inv}}(t) = \int \mathrm{e}^{2i\pi t x}\varphi(x)\mathrm{d}x.$$

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

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

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

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

is integrable and by virtue of Parseval's identity

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

Using the following version of Parseval's identity see [11]

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

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

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

Consider the following vector

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

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

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

We bound the first integral

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

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

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

By virtue of the dominated convergence theorem, the value of the last integral converges to $\int_{-\infty}^{\infty}|\widehat{\varphi}(t)|^2\hat{K}(t)\mathrm{d}t < 0$ 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. The conclusion of the theorem follows.

Proof of Proposition 4.5. 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

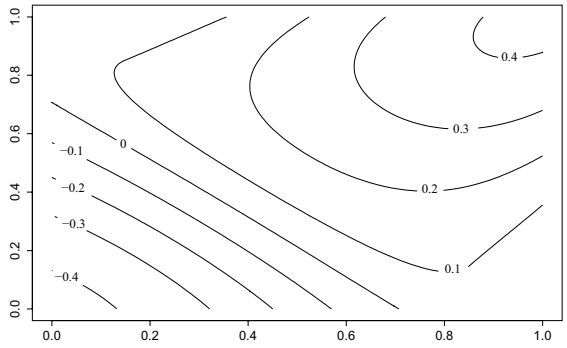

FIGURE 6. Contour of an upper bound of $\det(\mathbb{K}[3])$ as a function of (x, y).

vectors. For example, if the usual Euclidean distance is used, we have $d_h^2(X_i, X_j) = \sum_{l=1}^{d}(X_{il} - X_{jl})^2/h_l^2$. 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 Thm. 4.2). Recall that if $\mathbb{K}$ is semidefinite positive then all its principal minor (see [20], p. 398) are nonnegative. In particular, we can show that A is not semidefinite positive by producing a 3×3 principal minor with negative determinant. 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

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

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 $d_h(X_{i_1}, X_{i_2}) = x \le 1$ and $d_h(X_{i_2}, X_{i_3}) = y \le 1$. Using triangular inequality, we have

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

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.

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

References

[1] B. Abdous, Computationally efficient classes of higher-order kernel functions. *Can. J. Statist.* **23** (1995) 21–27.

[2] L. Breiman, *Using adaptive bagging to debias regressions*. Technical Report 547, Dpt of Statist., UC Berkeley (1999).

[3] L. Breiman and J. Friedman, Estimating optimal transformation for multiple regression and correlation. *J. Amer. Stat. Assoc.* **80** (1995) 580–598.

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

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

[6] P.-A. Cornillon, N. Hengartner and Matzner-Løber, *ibr: Iterative Bias Reduction*. CRAN (2010). http://cran.r-project.org/web/packages/ibr/index.html.

[7] P.-A. Cornillon, N. Hengartner, N. Jégou and Matzner-Løber, Iterative bias reduction: a comparative study. *Statist. Comput.* (2012).

[8] P. Craven and G. Wahba, Smoothing noisy data with spline functions. *Numer. Math.* **31** (1979) 377–403.

[9] M. Di Marzio and C. Taylor, On boosting kernel regression. *J. Statist. Plan. Infer.* **138** (2008) 2483–2498.

[10] R. Eubank, Nonparametric regression and spline smoothing. Dekker, 2nd edition (1999).

[11] W. Feller, An introduction to probability and its applications, vol. 2. Wiley (1966).

[12] J. Friedman, Multivariate adaptive regression splines. *Ann. Statist.* **19** (1991) 337–407.

[13] J. Friedman, Greedy function approximation: A gradient boosting machine. *Ann. Statist.* **28** (1189–1232) (2001).

[14] J. Friedman and W. Stuetzle, Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** (817–823) (1981).

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

[16] C. Gu, Smoothing spline ANOVA models. Springer (2002).

[17] L. Gyorfi, M. Kohler, A. Krzyzak and H. Walk, A Distribution-Free Theory of Nonparametric Regression. Springer Verlag (2002).

[18] D. Harrison and D. Rubinfeld, Hedonic prices and the demand for clean air. *J. Environ. Econ. Manag.* (1978) 81–102.

[19] T. Hastie and R. Tibshirani, Generalized Additive Models. Chapman & Hall (1995).

[20] R.A. Horn and C.R. Johnson, Matrix analysis. Cambridge (1985).

[21] C. Hurvich, G. Simonoff and C.L. Tsai, Smoothing parameter selection in nonparametric regression using and improved akaike information criterion. *J. Roy. Stat. Soc. B* **60** (1998) 271–294.

[22] O. Lepski, Asymptotically minimax adaptive estimation. I: upper bounds. optimally adaptive estimates. *Theory Probab. Appl.* **37** (1991) 682–697.

[23] K.-C. Li, Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *Ann. Statist.* **15** (1987) 958–975.

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

[25] L. Schwartz, Analyse IV applications à la théorie de la mesure. Hermann (1993).

[26] W. Stuetzle and Y. Mittal, Some comments on the asymptotic behavior of robust smoothers, in *Smoothing Techniques for Curve Estimation*, edited by T. Gasser and M. Rosenblatt. Springer-Verlag (1979) 191–195.

[27] J. Tukey, Explanatory Data Analysis. Addison-Wesley (1977).

[28] F. Utreras, Convergence rates for multivariate smoothing spline functions. *J. Approx. Theory* (1988) 1–27.

[29] J. Wendelberger, *Smoothing Noisy Data with Multivariate Splines and Generalized Cross-Validation*. PhD thesis, University of Wisconsin (1982).

[30] S. Wood, Thin plate regression splines. *J. R. Statist. Soc. B* **65** (2003) 95–114.

[31] Y. Yang, Combining different procedures for adaptive regression. *J. Mult. Analysis* **74** (2000) 135–161.