LA-UR- 09-01177

Approved for public release;
distribution is unlimited.

Title: Recursive Bias Estimation and ~~L_2~~ L_2 Boosting

Author(s):

Nicolas Hengartner
Pierre-Andre Cornillon
Eric Matzner-Lober

Intended for: Proceedings of COLT 2009

Los Alamos
NATIONAL LABORATORY
EST. 1943

Recursive Bias Estimation and L_2 Boosting

Pierre-André Cornillon
Montpellier SupAgro
cornillon@supagro.inra.fr

Nicolas Hengartner
Los Alamos National Laboratory
nickh@lanl.gov

Eric Matzner-Løber
University Rennes
eml@uhb.fr

Abstract

This paper presents a general iterative bias correction procedure for nonparametric regression smoothers. This bias reduction schema is shown to correspond operationally to the L_2 Boosting algorithm, which provides a new statistical interpretation for L_2 Boosting. Controlling the number of bias correction steps is necessary to avoid over-fitting the data. For multivariate thin plate spline regression smoother ~~with~~ ~~the~~ the number of iterations is selected using cross-validation, our bias corrected smoother adapts to smoothness of the underlying regression function m. We show the excellent finite sample performance of our smoother (available as an R package) over existing state-of-the-art multivariate regression procedures on both simulated and real data sets.

1 Introduction

Regression is a fundamental data analysis tool for relating a univariate response variable Y to a multivariate predictor $X \in \Re^d$ from the observations $(X_i, Y_i), i = 1, \dots, n$. Traditional nonparametric regression use the assumption that the regression function varies smoothly in the independent variable x to locally estimate the conditional expectation $m(x) = E[Y|X = x]$. The resulting vector of predicted values $\widehat{Y}_i$ at the observed covariates X_i 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.

Linear smoothers are linear in the response variable Y and are operationally written as

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

where S_λ is a $n \times n$ smoothing matrix. The smoothing matrix S_λ typically depends on a tuning parameter which we denote by λ, and 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 are associated to smoothers that averages over larger neighborhood and produce very smooth curves, while small λ are associated to smoothers that average over smaller neighborhood to 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.

Ideally, we want to choose the smoothing parameter λ to minimize the expected squared prediction error. There is a vast literature on how to do this (approximatively) without explicit knowledge of the underlying regression function, see for example ([Sim96]). This paper takes a different approach. Instead of optimally selecting the tuning parameter λ, we fix it to some reasonably large value that ensures that the resulting smoothers *oversmooths* the data so that the resulting smoother will have a relatively small variance but a substantial bias, and focus on correcting that bias. Our approach to bias correction rests on the observation that the conditional expectation of the $-R = -(Y - \widehat{Y})$, given X, is the bias of the smoother. This provides us with the opportunity to estimate the bias by smoothing the residuals R. The bias of the original smoother can be partially corrected by subtracting from it the estimated bias. This bias correction can be iteratively applied, producing a sequence of iterative bias corrected smoothers that are formally defined in Section 2.

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 ([Tuk77]) to estimate bias from model misspecification in multivariate regression. Obviously, one can iteratively repeat the bias correction step until the increase to the variance from the bias correction outweighs the magnitude of the reduction in bias, leading to an iterative bias correction.

Another iterative nonparametric function estimation method, seemingly unrelated to bias reduction, is Boosting. Boosting was introduced as a machine learning algorithm for combining multiple weak learners by averaging their weighted predictions ([Sch90, Fre95]). The good performance of the Boosting algorithm on a variety of datasets stimulated statisticians to understand it from a statistical point of view. In his seminal paper, [Bre98] shows how Boosting can be interpreted as a gradient descent method. This view of Boosting was reinforced by [Fri01]. Adaboost, a popular variant of the Boosting algorithm, can be understood as a method for fitting an additive model ([FHT00]) and recently [EHJT04] made a connection between L_2 Boosting and Lasso for linear models. But connections between iterative bias reduction and Boosting can be made. In the context of nonparametric density estimation, [DMT04] have shown that one iteration of the Boosting algorithm reduced the bias of the initial estimator in a manner similar to the multiplicative bias reduction methods ([HG95, JLN95, HML08]). In the follow-up paper ([DMT07]), they extend their results to the nonparametric regression setting and show that one step of the Boosting algorithm applied to an oversmooth effects a bias reduction. As expected, the decrease in the bias comes at the cost of an increase in the variance of the corrected smoother.

It is well known in multivariate data 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 and focus instead on fitting structurally constrained regression models, such as additive models [HT95, LN95]. In this paper, we depart from the classical multivariate structural regression models, and focus instead on very estimating smooth fully nonparametric regression functions. To exploit optimally the smoothness of the regression function, we shall consider procedures that adapt to the smoothness of the true regression function. We show in Section 3 that iteratively correcting for the bias, together with a suitable stopping rule, leads to smoothers that adapt to the smoothness of the true regression function and converge at the minimax rate of convergence.

Practical considerations leads us to consider kernel based smoothers, and its variant, nearest neighbor smoothers. Not all kernel smoothers are suitable for the iterative bias reduction procedure.

Beyond the nice theoretical properties of our estimator, we show in Section 5 using both simulated and real data sets, that our iterative bias corrected smoother significantly improves on the prediction mean square errors over popular competing multivariate nonparametric regression models, including additive models, projection pursuit regression and MARS. For example, prediction mean squared error for the Boston housing data [BF85], using our fully nonparametric smoother on 13 explanatory variables, is at least 40% smaller than the competing current state-of-the-art smoothers.

2 Iterative bias reduction

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

2.1 Preliminaries

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

$$Y_i \quad = \quad m(X_i) + \varepsilon_i, \quad i = 1, \dots, n, \qquad (1)$$

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. It is helpful to rewrite Equation (1) in vector form by setting $Y = (Y_1, \dots, Y_n)^t$, $m = (m(X_1), \dots, m(X_n))^t$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^t$, to get

$$Y \quad = \quad m + \varepsilon. \qquad (2)$$

Linear smoothers can be written as

$$\widehat{m}_1 = S_1 Y, \qquad (3)$$

where S_1 is an $n \times n$ smoothing matrix and $\widehat{m} = \widehat{Y} = (\widehat{Y}_1, \dots, \widehat{Y}_n)^t$, denotes the vector of fitted values. Typical smoothing matrices are contractions, by virtue that the fitted values have smaller norm than the raw data, that is $\|SY\| \leq \|Y\|$. We refer to [BHT89] for in depth discussion of such *shrinkage smoothers*.

The linear smoother (3) has bias

$$B(\widehat{m}_1) = E[\widehat{m}_1|X] - m = (S_1 - I)m$$

and variance

$$V(\widehat{m}_1|X) = S_1 S_1' \sigma^2,$$

respectively. A natural question is "how can one estimate the bias?" To answer this question, observe that the residuals $R_1 = Y - \widehat{m}_1 = (I - S_1)Y$ have expected value $E[R_1|X] = m - E[\widehat{m}_1|X] = (I - S_1)m = -B(\widehat{m}_1)$. This suggests estimating the bias by smoothing the negative residuals

$$\widehat{b}_1 := -S_2 R_1 = -S_2(I - S_1)Y.$$

For simplicity, we assume that the same smoother $S = S_1 = S_2$ is used. Note that the resulting estimate for the bias is zero whenever the smoothing matrix S is a projection, as is the case for linear regression, bin smoothers and regression splines. But since most smoothers are not projections, we an opportunity to correct for the bias of the pilot smoother $\widehat{m}_1$ by subtracting from it $\hat{b}_1$, which yields a *bias corrected* smoother

$$\begin{aligned} \widehat{m}_2 &= SY + S(I - S)Y \\ &= (I - (I - S)^2)Y. \end{aligned}$$

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

$$\begin{aligned} \widehat{m}_k &= SY + S(I - S)Y + \cdots + S(I - S)^{k-1}Y \\ &= (I - (I - S)^k)Y. \end{aligned}$$

More recently, [DMT07] 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 [BY03]. 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 [BY03]. This new interpretation for the L_2 boosting algorithm as iterative bias corrections was alluded to in [Rid00]'s discussion of [FHT00] paper on the statistical interpretation of boosting.

As defined by (3), smoothers predict the conditional expectation of responses only at the design points. It is useful to extend regression smoothers to enable predictions at arbitrary locations $x \in \mathbb{R}^d$ of the covariates. Such an extension allows us to assess and compare the quality of various smoothers by how well the smoother predicts new observations. To this end, write the prediction of the linear smoother S at an arbitrary location x as

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

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

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

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

to conclude that

$$\widehat{m}_k(x) = S(x)^t \widehat{\beta}_k \tag{4}$$

predicts $m(x)$.

The mean squared error of the k^{th} iterated bias corrected linear smoother $\widehat{m}_k$ (??) is

$$\begin{aligned} MSE(\hat{m}_k) &= m^t \left((I-S)^k\right)^t (I-S)^k m \\ &\quad + \sigma^2 (I - (I-S)^k) \left((I - (I-S)^k)\right)^t. \end{aligned}$$

The qualitative behavior of the sequence of iterative bias corrected smoothers $\widehat{m}_k$ depends on the spectrum of S. it is easily shown that the sequence of iteratively bias corrected smoothers $\widehat{m}_k$ convergent whenever the singular values of the smoothing matrix S lie between zero and two. In that case, the limit $\lim_{k\to\infty} \widehat{m}_k = Y$. Otherwise, $\lim_{k\to\infty} \|\widehat{m}_k\| = \infty$.

It follows that iterating the bias correction algorithm until convergence is not desirable. However, since each iteration of the bias correction algorithm reduces the bias and increases the variance, often a few iteration of the bias correction scheme will improve upon the pilot smoother. This brings up the important question of how to decide when to stop the iterative bias correction process.

Viewing the latter question as a model selection problem suggests stopping rules for the number of iterations based on Mallows' C_p, Akaike Information Criteria (AIC), Bayesian Information Criterion (BIC), cross-validation, L-fold cross-validation, and Generalized cross validation. Each of these data-driven model selection methods estimate an optimum number of iterations k of the iterative bias correction algorithm by minimizing estimates for the expected squared prediction error of the smoothers over some pre-specified set $\mathcal{K} = \{1, 2, \ldots, M_n\}$ for the number of iterations. We rely on the expansive literature on model selection to provide insight into the statistical properties of stopped bias corrected smoother. In particular, Theorem 3.2 of [Li87] describes the asymptotic behavior of the generalized cross-validation (GCV) stopping rule applied to smoothers.

3 Iterative bias reduction of multivariate thin-plate splines smoothers

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

$$\min_f \sum_{i=1}^n (Y_i - f(X_i))^2 \quad + \lambda\Gamma(\nu_0), \tag{5}$$

where

$$\Gamma(\nu_0) \;=\; \sum_{\substack{i_1,\ldots,i_d = 0 \\ i_1 + \ldots + i_d \le \nu_0}} \int_{\mathbb{R}^d} \left| \frac{\partial^{i_1+\ldots+i_d}}{\partial x_{i_1} \ldots \partial x_{i_{\nu_0}}} f(x) \right|^2 dx.$$

Thin-plate smoothing splines are attractive class of multivariate smoothers for two reasons: First, the solution of (5), once cast within a Reproducing Kernel Hilbert Space (RKHS) framework [Gu02], is numerically tractable and second, the eigenvalues of the smoothing matrix are approximatively known (c.f. [Utr88]).

3.1 Numerical Example

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

Figure 1: True regression function $m(x_1, x_2)$ (6) on the square $[-10, 10] \times [-10, 10]$ used in our numerical examples.

This behavior is confirmed numerically in the following pedagogical example of a bivariate regression problem: Figure 1 graphs Wendelberger's test function [Wen82]

$$\begin{aligned} m(x_1, x_2) &= \frac{3}{4} \exp\left\{ -((9x-2)^2 + (9y-2)^2)/4 \right\} \\ &\quad + \frac{3}{4} \exp\left\{ -((9x+1)^2/49 + (9y+1)^2/10) \right\} \\ &\quad + \frac{1}{2} \exp\left\{ -((9x-7)^2 + (9y-3)^2)/4) \right\} \\ &\quad - \frac{1}{5} \exp\left\{ -((9x-4)^2 + (9y-7)^2) \right\} \end{aligned} \tag{6}$$

that is sampled at 100 locations on the regular grid $\{0.05, 0.15, \ldots, 0.85, 0.95\}^2$. The disturbances are mean zero Gaussian with variance producing a signal to noise ratio of five. Figure 2 shows the evolution of the bias corrected smoother, starting from a nearly linear pilot smoother in panel (a). After 500 iterative bias reduction steps, the smoother shown in panel (b) is visually close to the original regression function.

Continuing the bias correction scheme will eventually lead to a smoother that interpolates the raw data. To illustrate this, we show the bias corrected smoother after 50000 iterations in panel (c). Notice how noisy that estimator is, compared to the one in panel (b). This numerical examples hints to the potential gains realizable by suitably selecting the number of bias correction steps.

Figure 2: Thin-plate spline regression smoothers from 100 noisy observations from 6 (see Figure 1) evaluated on a regular grid on $[-10, 10] \times [-10, 10]$. Panel (a) shows the pilot smoother, panel (b) graphs the bias corrected smoother after 500 iterations and panel (c) graphs the smoother after 50000 iterations of the bias correction scheme.

3.2 Adaptation to smoothness of the regression function

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

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

$$h_{max}(n) = \sup_{x\in\Omega} \inf_{i=1,\ldots,n} |x - X_i|,$$

and

$$h_{min}(n) = \min_{i\neq j} |X_i - X_j|,$$

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

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

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

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

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

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

The proof is postponed to the appendix.

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

Theorem 2 *Suppose that the error admits a finite moment of order 10, than $\hat{k}_{GCV}$ selected by GCV on a grid $\mathcal{K}_n = \{1, \ldots, n^{1+\alpha}\}$ where α could be bigger than one, then*

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

Good approximation properties of the L_2-boosting algorithm applied to univariate smoothing splines, similarly to our Theorem 1, was proven by [BY03]. Theorem 2 states that, using cross-validation to select the number of bias correction steps, we can achieve adaptation to the smoothness of the underlying true regression function.

The application of bias reduction to maximally exploit the smoothness of multivariate regression function has not been previously exploited. The practical benefits of this adaptation is revealed in both our simulation study and our analysis of classical multivariate test datasets. In both instances, our method makes substantially better out of sample predictions over state-of-the-art structural models that include additive regression smoothing, MARS, and projection pursuit regression. The statistical properties of the iterative corrected thin-plate splines are tractable, they have several drawbacks that limit their practical usefulness. An advantage of smoothing splines smoothers is that the eigenvalues of the smoothing matrix are well known. The matrix S is symmetric and all the eigen values are in $[0, 1]$ and the first $M_0 = \binom{\nu_0+d-1}{\nu_0-1}$ eigen values are equal to 1 see for example [Utr88].

4 Kernel Based Smoothers

Our iterative procedure needs a pilot smoother S_λ with substantial bias. As our procedure is adaptive (theorem 1), we use thin plate spline with the lowess possible order ν_0 which must be bigger than $d/2$. We need to choose the smoothing parameter λ such as the pilot smoother S_λ oversmooths the data. The degree of freedom (i.e the trace of S_λ) of the pilot estimator must be small. The minimum degree of freedon of the pilot smoother is $M_0 = \binom{\nu_0+d-1}{\nu_0-1}$ and regarding to the sample size, the pilot smoother is usually not smooth. For the real Los Angeles data set, $n = 330$ and $d = 8$, one should normally take $\nu_0 = 5$ and the initial ddl will be bigger than 495. So, for practical used, if the sample size is limited (less than 500), it better to use a kernel smoother.

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. We will use a Gaussian kernel because the corresponding spectrum of $I - S$ is less than 1. All kernels do not share that property ([PCML09b]). 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.

4.1 Good behavior of Gaussian kernel smoothers.

4.2 Failure of uniform and Epanechnikov kernel smoothers

Not all kernels smoothers are well behaved under the iterative bias correction scheme. In particular, we can show that kernel smoothers based on either the uniform or the Epanechnikov kernel leads to a sequence $\hat{m}_k$ whose norm diverges with the number of bias correction steps k.

Theorem 3 *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.

5 Real example : 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 ([Bre96, BY03, BY06]). 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, freely available as an R package **IBR** [PCML09a], with Mars using R package **mda**, with additive models estimation using R package **mgcv**, projection pursuit regression using R function *ppr* and L_2-Boosting proposed by [BY03], which we recall here. Multivariate L_2-Boosting proposed by [BY03] leads to component-wise additive model

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

where the component $\hat{f}^{[k],(j)}$ 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 [BY03] and reporting theirs results we obtain the following table :

Table 1: Predicted Mean Squared Error on test observations of ozone data for different methods.

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

We can see (table 1) that, as in univariate setting, the smoother the pilot estimator is, the better the final estimation is, at the cost of increasing computation time. The combination of iterated of GCV and bias corrected estimator leads to a diminution of more than 12% over other multivariate methods.

We will also present the Boston housing data and others classical data sets during the presentation.

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 [Rid00] in his discussion of the seminal paper [FHT00], and explored in [DMT04, DMT07] 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. 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 [DMT07].

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.

A Appendix

Proof of Theorem 1 Let $\nu_0 < \nu$ and fix the smoothing parameter λ_0. Define $S = S_{\nu_0,\lambda_0}$. The eigen decompostion of S (Utreras, 1988) gives

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

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

$$\begin{aligned} & V(\hat{m}_k, \lambda_0, \nu_0) \\ = \; & \sigma^2 \frac{M_0}{n} + A_{M_0} \end{aligned}$$

where

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

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

$$V(\hat{m}_k, \lambda_0, \nu_0) \leq \sigma^2 \frac{M_0}{n} + \sigma^2 \frac{J_n - M_o}{n} + A_{J_n}.$$

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

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

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

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

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

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

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

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

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

$$\frac{1}{n} \sum_{j=M_0+1}^{n} j^{2\nu_0/d} \mu_{j,\nu_0}^2 \leq M < \infty. \tag{7}$$

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

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

Using the same type of bound as in equation (7) we get

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

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

Balancing the squared bias and the variance lead to the choice

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

and we obtain the desired optimal rate.

Proof of Theorem 3. 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 $u^t K u$ (see proof of Theorem **??**). Recall that if K is semidefinite then all its principal minor 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 $K[3]$ obtained by taking the rows and columns (i_1, i_2, i_3). The determinant of $K[3]$ is readily computed 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

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

With this choice, we readily calculate

$$det(K[3]) \quad = \quad 0 - K_h(0)\left[K_h(0)^2 - 0\right] - 0 < 0.$$

Since a principal minor of K is negative, we conclude that 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

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

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

Figure 3: Contour of an upper bound of $det(K[3])$ as a function of (x, y).

Thus a principal minor of K is negative, and as a result, K and A are not semidefinite positive.

References

[BF85] L. Breiman and J. Friedman. Estimating optimal transformation for multiple regression and correlation. *Journal of the American Statistical Association*, 80:580–598, 1985.

[BHT89] A. Buja, T. Hastie, and R. Tibshirani. Linear smoothers and additive models. *Ann. of Statist.*, 17:453–510, 1989.

[Bre96] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

[Bre98] L. Breiman. Arcing classifier (with discussion). *Ann. of Statist.*, 26:801–849, 1998.

[BY03] P. Bühlmann and B. Yu. Boosting with the l_2 loss: Regression and classification. *J. Amer. Statist. Assoc.*, 98:324–339, 2003.

[BY06] P. Bühlmann and B. Yu. Sparse boosting. *J. Macine Learning Research*, 7:1001–1024, 2006.

[DMT04] M. Di Marzio and C. Taylor. Boosting kernel density estimates: a bias reduction technique ? *Biometrika*, 91:226–233, 2004.

[DMT07] M. Di Marzio and C. Taylor. Multiple kernel regression smoothing by boosting. *submitted*, 2007.

[EHJT04] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression (with discussion). *Ann. of Statist.*, 32:407–451, 2004.

[FHT00] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Ann. of Statist.*, 28:337–407, 2000.

[Fre95] Y. Freund. Boosting a weak learning algorithm by majority. *Information and Computation*, 121:256–285, 1995.

[Fri01] J. Friedman. Greedy function approximation: A gradient boosting machine. *Ann. Statist.*, 28:1–141, 2001.

[Gu02] Chong Gu. *Smoothing spline ANOVA models*. Springer, New-York, 2002.

[HG95] N.L. Hjort and I.K. Glad. Nonparametric density estimation with a paramteric start. *Ann. Statist.*, 23:882–904, 1995.

[HML08] N. Hengartner and E. Matzner-Løber. Asymptotic unbiased density estimators. *ESAIM*, 2008.

[HT95] T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models*. Chapman & Hall, 1995.

[JLN95] M.C. Jones, O. Linton, and J.P. Nielsen. A simple and effective bias reduction method for kernel density estimation. *Biometrika*, 82:327–338, 1995.

[Li87] Ker-Chau 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.

[LN95] O. Linton and J.P. Nielsen. A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 82:93–100, 1995.

[PCML09a] N. Hengartner P.A. Cornillon and E. Matzner-Lober. Ibr: Iterative bias reduction smoothers with r. *URL: http://www.uhb.fr//sc_sociales/labstats/EML/ibr_0.01.tar.*, 2009.

[PCML09b] N. Hengartner P.A. Cornillon and E. Matzner-Lober. Recursive bias estimation for high dimensional regression smoothers. *Submitted*, 2009.

[Rid00] G. Ridgeway. Additive logistic regression: a statistical view of boosting: Discussion. *Ann. of Statist.*, 28:393–400, 2000.

[Sch90] R.E Schapire. The strength of weak learnability. *Machine learning*, 5:197–227, 1990.

[Sim96] J.S. Simonoff. *Smoothing Methods in Statistics*. Springer, New York, 1996.

[Tuk77] J.W. Tukey. *Explanatory Data Analysis*. Addison-Wesley, 1977.

[Utr88] F. Utreras. Convergence rates for multivariate smoothing spline functions. *Journal of Approximation Theory*, pages 1–27, 1988.

[Wen82] J. Wendelberger. Smoothing noisy data with multivariate splines and generalized cross-validation. *Ph.D thesis, University of Wisconsin*, 1982.