On convex regression estimators

Néstor Aguilera* Liliana Forzani* Pedro Morin*

June 10, 2010

Abstract

A new nonparametric estimator of a convex regression function in any dimension is proposed and its convergence properties are studied. We start by using any estimator of the regression function and we *convexify* it by taking the convex envelope of a sample of the approximation obtained. We prove that the uniform rate of convergence of the estimator is maintained after the convexification is applied. The finite sample properties of the new estimator are investigated by means of a simulation study and the application of the new method is demonstrated in examples.

Keywords: approximation, convex regression, convexity, data-smoothing, nonparametric regression

1 Introduction

In the nonparametric regression model

$$Y_n = f(X_n) + e_n, \quad n = 1, 2, \ldots, \tag{1}$$

where $Y_n \in \mathbb{R}$, $X_n \in \mathbb{R}^d$ and e_n is an error term, it is not uncommon to have strong presumptions on properties of f—such as monotonicity, convexity or concavity—which should be taken into account.

Typical examples appear in economics (indirect utility, production or cost functions), medicine (dosage-response experiments) and biology (growth curves).

A much studied case is the instance of a monotone regression function for $d = 1$, estimated by using least squares (see, e.g., Brunk, 1955; Mukerjee, 1988, and Barlow et al., 1972 or Robertson et al., 1988 for a summary of this work). For convex (concave) regression Hildreth (1954) proposed to use convex least square estimates, and Hanson & Pledger (1976) proved their consistency. Algorithms for computing these estimates were developed by Wu (1982) and Fraser & Massam (1989), and the rate of convergence was derived by Mammen (1991). Later Groeneboom et al. (2001) derived the asymptotic distribution of the estimator at a fixed point of positive curvature. In all of these works the estimates hold pointwise.

Still in one dimension, one can avoid the complications of least squares techniques and use more conventional smoothing methods when f is convex (or

*Consejo Nacional de Investigaciones Científicas y Técnicas, and Universidad Nacional del Litoral, Argentina

concave), as shown by Birke & Dette (2007). Using the fact that a differentiable function is convex (concave) if the derivative is increasing (decreasing), they propose to first smooth the data using any constrained nonparametric estimate (kernel type, local polynomial, series or spline estimator), then compute the derivative of the smooth function thus obtained, which is isotonized and finally integrated to recover a convex estimation. As mentioned above, the isotonization of a function is something that has already been mastered in the non-parametric literature, and using those results the rates of convergence obtained by them are the usual in non-parametric regression.

Unfortunately this technique can only be used in one dimension and with smooth convex functions and cannot be extended to higher dimensions, since there is no such simple characterization of convexity in $\mathbb{R}^d$ for $d > 1$.

As far as we know, little has been done in higher dimensions. Siem et al. (2005) (see also Hoffmann et al., 2006) present a multivariate data smoothing method using a linear program (for the ℓ^1 and ℓ^∞ norms) or quadratic program (for the ℓ^2 norm). Shih et al. (2006) develop an approximation method based on multivariate adaptive regression splines (MARS). But none of these articles present convergence results.

We propose here a simple and fast method that can be used in any dimension and applied to any convex function, even if not too smooth. Like Birke and Dette, we start by using any approximating scheme on the data, but then we use a *convexification* step, consisting in taking the convex envelope of the approximating function just obtained. This last step can be done very quickly by current software such as QHULL (Barber et al., 1996), and the uniform rate of convergence of the approximation technique is maintained after the convexification is applied.

More precisely, we obtain uniform error estimates, and the rate of convergence of the convex estimator is the same as that of the original estimator, thereby showing that the convexification step adds basically no further errors to the estimating step.

The paper is organized as follows. In Section 2 we briefly review fundamental smoothing techniques. In Section 3 we show theoretical results on the convexification step, and how the error estimates for the convex estimate are derived from the smoothing step. Finally, in Section 4 we apply these techniques to approximate several problems in dimensions $d = 1$ and $d = 2$.

2 The smoothing step: review of the literature

As we have already pointed out, our method of convexification inherits the L^∞ rate of convergence from whichever smoothing process is chosen for the model (1). We think it is appropriate, then, to briefly review rates of convergence in L^∞-norm for some of the possible choices for such a process when no monotonicity or convexity assumptions are made on f.

Most of the approximation techniques with known rates of convergence are of the so called *smoothing* type, where a variable kernel is used, and we will focus our attention on these.

It should be noted that since there are many different schools and people involved, here we can give only partial references, leaving out several meaningful results available in the literature.

Perhaps the first ones to consider these problems were Devroye (1978) and Schuster & Yakowitz (1979). Devroye considered the Nadaraya-Watson regression estimator and proved the uniform convergence (without rates) for independent data, with fixed or random predictors belonging to $\mathbb{R}^d$, whereas Schuster and Yakowitz considered more general kernels in one dimension, establishing orders of convergence in probability. Later these results were extended by several authors, among them Bierens (1983) and Collomb (1984). They extended the result to non-independent data and Collomb was the first to give strong rates for uniform convergence. Further results on uniform convergence rates for different settings such as robust estimation and other kind of non-independent data were given by Collomb & Härdle (1986), Roussas (1990), Boente & Fraiman (1991), Truong & Stone (1992) and Tran (1993). Extensions to spline estimators were given by Eggermont & LaRiccia (2006), and to uniform choice of bandwidth by Einmahl & Mason (2005, 2000), Dony (2008), Dony & Einmahl (2006), Dony & Mason (2008), and Dony et al. (2006) (see also the references therein).

The asymptotic distribution of the maximal deviation between a non-parametric regression estimator and the true regression was first considered by Johnston (1982), extending to the regression context the results by Bickel & Rosenblatt (1973) and Rosenblatt (1976) on density estimation. For the case $d = 1$ and random predictors, Johnston showed—under some regularity assumptions—the L^∞ asymptotic distribution of the kernel regression estimator, which allowed him to give uniform confidence intervals for the regression estimator. This result was extended by Konakov & Piterbarg (1984) to other kernel estimators and by Härdle (1989) to general estimators defined implicitly, as for example M-smoothers and local polynomial estimators. As far as we know these results were not extended to higher dimensions or non-independent data.

3 A convex estimator and its convergence

Let us assume that the variables X_n in the model (1) take values on a bounded closed convex set $Q \subset \mathbb{R}^d$, and that $f \in \mathscr{C}$, where $\mathscr{C}$ is the set of (finite real valued) convex functions defined on Q.

Q need not be polyhedral, but assuming its boundary is smooth except for a finite set of "corners", in practice we may approximate it by a polyhedron. Thus, from now on, for simplicity we will assume that Q is a polyhedron, and therefore it is the convex hull of its finite set of vertices. In particular, we assume that Q is compact.

Let us assume that f_n is an estimator of f, defined in all of Q. To fix ideas, we may think that f_n is obtained by considering the points (X_i, Y_i), $i = 1, \dots, n$, by some procedure such as smoothing. Our purpose is to derive from f_n another estimator which is also convex.

To do so, we consider a finite set $\mathcal{M}_n \subset Q$ such that the convex hull of $\mathcal{M}_n$ is Q. The number of points in $\mathcal{M}_n$ need not be n and the points in $\mathcal{M}_n$ might be completely unrelated to $\{X_i : i \in \mathbb{N}\}$.

We now let $\mathscr{L}_n$ be the set of "convex functions below f_n on $\mathcal{M}_n$",

$$\mathscr{L}_n = \big\{\psi \in \mathscr{C} : \psi(x) \le f_n(x) \text{ for all } x \in \mathcal{M}_n\big\},$$

and define the *convex estimator* f_n^c, associated with the estimator f_n and the

set $\mathcal{M}_n$ by

$$f_n^c = \sup\{\psi : \psi \in \mathscr{L}_n\}. \tag{2}$$

Since $\mathcal{M}_n$ contains all the vertices of Q, it is easy to see that f_n^c is well defined on Q and that $f_n^c \in \mathscr{C}$. Furthermore, f_n^c is piecewise linear, determined by the maximum of hyperplanes. In particular:

Lemma 1. $f_n^c \in \mathscr{L}_n$.

As f_n^c is the "lower part" of the convex hull of the set $\{(x, f_n(x)) : x \in \mathcal{M}_n\}$, we may take advantage of any of a number of algorithms for finding convex hulls in $\mathbb{R}^d$. For instance, QHULL (Barber et al., 1996) finds the convex hull of a finite set of points in any number of dimensions, and is really fast for dimensions $d \leq 4$.

We are led to the following procedure for constructing a convex estimator f_n^c of f:

Procedure 2. *Given X_i and Y_i ($i = 1, 2, \dots$):*

Step 1. (Smoothing) Construct an estimator f_n of f, for instance through a smoothing procedure using the values X_i and Y_i for $i = 1, \dots, n$.

Step 2. (Grid of points) Choose $\delta_n > 0$ and $\mathcal{M}_n \subset Q$ so that any $x \in Q$ is the convex combination of points in $\mathcal{M}_n$ whose distance to x is not more than δ_n.

Step 3. (Convexification) Construct f_n^c as in (2), for instance by using a convex hull procedure such as QHULL.

In Figure 1 we represent the steps of the procedure with an example: in (a) we show the data and the resulting estimator f_n; in (b) we show the estimator and its values at the points of $\mathcal{M}_n$; in (c) we show the convex estimator f_n^c obtained from the values of f_n at $\mathcal{M}_n$; and in (d) we compare the original data and the convex estimator obtained.

We now show that if in the Procedure 2, f_n is a good approximation of f, then f_n^c is a good approximation of f provided it satisfies:

H-1. f is a continuous convex function defined on Q, with $\|f\|_{\text{Lip}} = L < \infty$, where

$$\|f\|_{\text{Lip}} = \sup\{|f(x) - f(y)|/|x - y| : x, y \in Q,\ x \neq y\}.$$

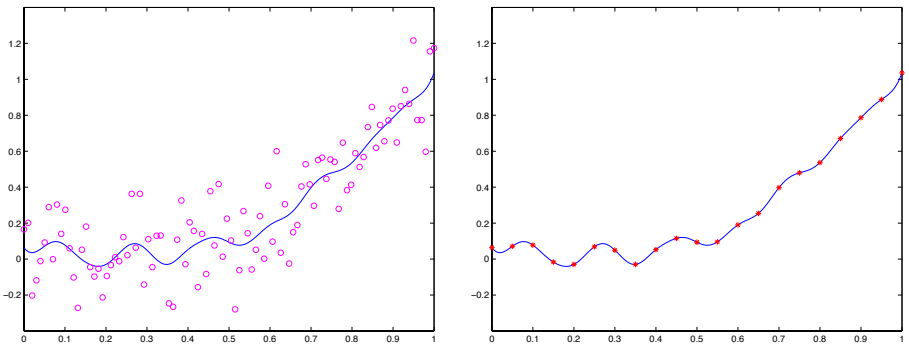
and $|x-y|$ denotes the (Euclidean) distance between x and y in $\mathbb{R}^d$. (Recall that convex functions on Q are locally Lipschitz, but here we require that f be uniformly Lipschitz in all of Q.)

Theorem 3. *Suppose f satisfies **H-1** and let f_n, δ_n, $\mathcal{M}_n$ and f_n^c be as in Procedure 2, with*

$$\sup\{|f_n(x) - f(x)| : x \in \mathcal{M}_n\} \leq \varepsilon_n. \tag{3}$$

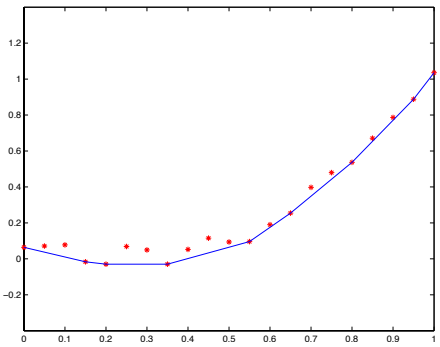
Then,

$$-\varepsilon_n \leq f_n^c(x) - f(x) \leq \varepsilon_n + L\delta_n \quad \text{for all } x \in Q.$$

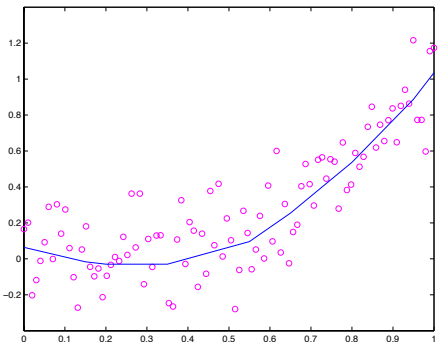


(a) Data and estimator f_n

(b) Estimator f_n and its values on $\mathcal{M}_n$



(c) Convex estimator f_n^c from f_n on $\mathcal{M}_n$



(d) Data and convex estimator f_n^c

Figure 1: Steps in constructing a convex estimator

Proof. Since f is convex and ε_n is a constant, the function $f - \varepsilon_n$ is convex. Moreover, $f(x) - \varepsilon_n \leq f_n(x)$ for all $x \in \mathcal{M}_n$ implies that $f - \varepsilon_n \in \mathscr{L}_n$, and by the definition of f_n^c in (2),

$$f(x) - \varepsilon_n \leq f_n^c(x) \quad \text{for all } x \in Q,$$

proving one inequality.

For the other inequality, consider $x \in Q$, and let $x_k \in \mathcal{M}_n$ and $\lambda_k \geq 0$, $k = 1, \ldots, d+1$, be such that

$$\sum_k \lambda_k x_k = x, \quad \sum_k \lambda_k = 1, \quad \text{and} \quad |x - x_k| \leq \delta_n \text{ for } k = 1, \ldots, d+1.$$

Then,

$$\begin{aligned} f_n^c(x) &\leq \sum_k \lambda_k \, f_n^c(x_k) && \text{since } f_n^c \in \mathscr{C}, \\ &\leq \sum_k \lambda_k \, f_n(x_k) && \text{by Lemma 1}, \\ &\leq \sum_k \lambda_k \, (f(x_k) + \varepsilon_n) && \text{by (3)}, \\ &= \left(\sum_k \lambda_k \, f(x_k) \right) + \varepsilon_n && \text{since } \sum_k \lambda_k = 1. \end{aligned}$$

Now, $\|f\|_{\text{Lip}} = L$ and $|x_k - x| < \delta_n$, and therefore

$$f(x_k) \leq f(x) + L\delta_n.$$

Hence, since $\lambda_k \geq 0$ and using again that $\sum_k \lambda_k = 1$, we conclude

$$f_n^c(x) \leq \left(\sum_k \lambda_k \, (f(x) + L\delta_n) \right) + \varepsilon_n = f(x) + L\delta_n + \varepsilon_n,$$

and the result follows. □

Remark. In the proof we have not used the finiteness of $\mathcal{M}_n$, and only the values of f_n on $\mathcal{M}_n$ are used.

Noticing that given $\delta_n > 0$ we may construct a finite set $\mathcal{M}_n$ with the property that any $x \in Q$ is a convex combination of points in $\mathcal{M}_n$ whose distance to x is no more than δ_n, we have:

Corollary 4. *If f satisfies **H-1**, given an estimator f_n of f and $\delta_n > 0$, we may find $\mathcal{M}_n$ and define f_n^c according to Procedure 2, so that*

$$\|f_n^c - f\|_\infty \leq \|f_n - f\|_\infty + L\delta_n.$$

Remark. In the extreme case where $f_n = f$ for all n, we have $\|f_n - f\|_\infty = 0$, but $\|f_n^c - f\|_\infty > 0$ in general (for instance, if $\mathcal{M}_n$ is finite and f is not piecewise linear).

Corollary 4 tells us that the convex estimator f_n^c obtained through the Procedure 2 inherits the approximation properties of the original estimator f_n,

and the rate of convergence is preserved or even bettered provided δ_n is small enough.

To illustrate this behavior, let us consider the following well-known types of convergence of a sequence of nonnegative random variables $(R_n)_n$ to 0, where $(r_n)_n$ is a bounded sequence of positive numbers (possibly converging to 0), and we have denoted by $\mathbb{P}$ the underlying probability measure:

T-1. For every $\varepsilon > 0$ there exists $M > 0$ such that $\sup_n \mathbb{P}(R_n > Mr_n) < \varepsilon$.

T-2. $\lim_{n\to\infty} \mathbb{P}(R_n > \varepsilon r_n) = 0$ for every $\varepsilon > 0$.

T-3. $R_n = O(r_n)$ or $R_n = o(r_n)$ a.s.

T-4. For every $\varepsilon > 0$, $\sum_{n=1}^{\infty} \mathbb{P}(R_n > \varepsilon r_n) < \infty$.

It is easy to see that:

Theorem 5. *If any of* ***T-1*** *through* ***T-4*** *holds for $R_n = \|f_n - f\|_\infty$, then it also holds for $R_n = \|f_n^c - f\|_\infty$, provided f satisfies* ***H-1*** *and f_n^c is constructed as in Corollary 4 with $\delta_n = o(r_n)$.*

For example, Tran (1993) shows:

Theorem 6. *For $j = 1, 2, \ldots$, let $\{(X_j, Y_j)\}_j$ be a strictly stationary sequence of random variables, where the X_j and the Y_j are $\mathbb{R}^d$-valued and $\mathbb{R}$-valued, respectively. Suppose $f(x) = \mathbb{E}(Y \mid X = x)$ is estimated by*

$$f_n(x) = \frac{1}{\#(I_n(x))} \sum_{i \in I_n(x)} Y_i \quad \text{for } x \in Q,$$

where $I_n(x) = \{i : 1 \le i \le n, |X_i - x| \le h_n\}$, and $h_n \approx (\log(n)/n)^{1/(d+2)}$.

Then, under appropriate assumptions (including adequate regularity conditions),

$$\|f_n - f\|_{L^\infty(Q)} = O(h_n) \quad a.s.$$

Tran's result gives a **T-3** type of convergence, and therefore (by Theorem 5) we have that under the same assumptions,

$$\|f_n^c - f\|_{L^\infty(Q)} = O(h_n) \quad \text{a.s.},$$

provided we take $\delta_n = o(h_n)$ in Corollary 4.

More elaborate types of convergence include exact asymptotic behavior. A very simple model might be, assuming X_n uniformly distributed on Q:

T-5. There exist a sequence $(d_n)_n$ converging to 0, and a random variable R such that

$$\mathbb{P}(r_n^{-1}(R_n - d_n) \le t) \to \mathbb{P}(R \le t),$$

for every $t \in \mathbb{R}$ at which $\mathbb{P}(R \le t)$ is continuous.

It is not possible in general to carry over this convergence from $R_n = \|f_n - f\|_\infty$ directly to $R_n = \|f_n^c - f\|_\infty$, as in general $\|f_n^c - f\|_\infty$ could be much smaller than $\|f_n - f\|_\infty$, and we cannot control $\|f_n - f\|_\infty$ solely in terms of

$\|f_n^c - f\|_\infty$ and $\|f\|_{\text{Lip}}$. Needless to say, by enlargening r_n we may transform a **T-5** type into, say, a **T-2** type of convergence.

Besides the interest in itself, the convergence of type **T-5** allows us to find uniform confidence bands for the regression curve, which is a practical concern. More precisely, if **T-5** is verified, for any α, $0 < \alpha < 1$, we may find optimal (or near optimal) s so that

$$\mathbb{P}(R_n \leq s) \geq 1 - \alpha. \tag{4}$$

If this inequality holds for $R_n = \|f_n - f\|_\infty$ and assuming f_n^c is constructed as in Corollary 4 with $\delta_n = o(1)$ for all n, then (4) is valid for $R_n = \|f_n^c - f\|_\infty$, albeit not with optimal s.

In other words, Corollary 4 allows us to convert a uniform confidence band for f_n of the form (4) into a (slightly different) uniform confidence band for f_n^c.

For instance, Johnston (1982, Theorem 2.1) shows:

Theorem 7. *Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a bivariate population, with X uniformly distributed in $Q = [0, 1]$, and consider the following estimator of $f(x) = \mathbb{E}(Y \mid X = x)$,*

$$f_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} Y_i \, K((x - X_i)/h_n), \tag{5}$$

where $h_n \approx n^{-\delta}$ for some δ, $1/5 < \delta < 1/3$, and K is a piecewise smooth density function with support in $[-A, A]$, $A > 1$.

Then, under appropriate regularity assumptions we have

$$\mathbb{P}\left((2\delta \log n)^{1/2} \left[\sup_{0 \leq x \leq 1} r_n^{-1}(x) \left(f_n(x) - f(x) \right) - d_n \right] < t \right) \to e^{-2 \exp(-t)},$$

where

$$r_n^2(x) = \frac{\int K^2(u)\, du \times \mathbb{E}(Y^2 \mid X = x)}{nh_n} \tag{6}$$

and $d_n = O\left((2\delta \log n)^{1/2}\right)$.

Confidence bands follow immediately (Johnston, 1982, Corollary 3.1):

Corollary 8. *Assuming Theorem 7 holds, an approximate $(1 - \alpha) \times 100\%$ confidence band is*

$$f_n(x) \pm r_n \left(d_n + c(\alpha)(2\delta \log n)^{-1/2} \right),$$

where $c(\alpha) = \log 2 - \log |\log(1-\alpha)|$ (for practical applications, one would estimate $\mathbb{E}(Y^2 \mid X = x)$ in (6)).

Theorem 7 and its corollary are still valid if instead of (5), f_n is a $M-$smoother estimator defined as a solution of

$$0 = \frac{1}{nh_n} \sum_{i=1}^{n} \psi(Y_i - f_n) \, K((x - X_i)/h_n),$$

with ψ a bounded monotone, antisymmetric real function (Härdle, 1989).

As a final remark, let us point out that we have only used that f_n approximates the Lipschitz convex function f, independently of whether f_n has been obtained through a smoothing procedure or any other approximation method.

4 Numerical results

In this section we report on some practical aspects of our algorithm and present some simulations and examples showing its performance.

4.1 Implementation

We implemented our algorithm using MATLAB. The smoothing step was done with local polynomials of degree 1 with Gauss's kernel, and for the convexification we used MATLAB's functions convhull (dimension 1) and convhulln (higher dimensions), which are based upon the QHULL algorithm described in Barber et al. (1996).

The bandwidth was chosen using cross-validation for the local-polynomial fitting at the data points. In the examples shown below, once the optimal bandwidth was chosen, the local-polynomial fitting function was computed at the same data points which were set a priori as design. Whenever the data points were not a priori designed, the local-polynomial fit was evaluated on a uniform grid having approximately the same number of points.

4.2 One dimensional simulations

In this section we briefly illustrate the finite sample properties of the convex estimate of the regression function by means of a simulation study. For this purpose we considered the same three examples presented in Birke & Dette (2007), namely,

$$
\begin{aligned}
f_1(x) &= e^{3(x-1)}, \\
f_2(x) &= \frac{16}{9}\left(x - \frac{1}{4}\right)^2, \\
f_3(x) &= \begin{cases} -4x+1 & \text{if } 0 \le x \le 1/4, \\ 0 & \text{if } 1/4 < x < 3/4, \\ 4x-3 & \text{if } 3/4 \le x, \end{cases}
\end{aligned}
$$

and $Q = [0, 1]$. Notice that even though the third function is just Lipschitz, all these functions satisfy the assumption **H-1**.

As in Birke & Dette (2007), we ran some simulations with $n = 100$ uniformly distributed design points for the explanatory variables and added a normal noise with standard deviation $\sigma = 0.1$ to the response variable.

In Figure 2 we display for each regression function five typical estimates obtained from different simulation runs observing a typical performance. The estimates for the two smooth functions f_1 and f_2 are comparable to the regressions obtained in Birke & Dette (2007), but our estimates of the nonsmooth regression function f_3 exhibit a much closer fit. This is an advantage of our method,

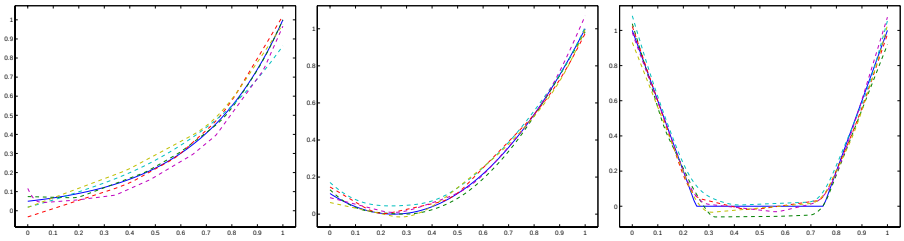


Figure 2: Regression functions f_1 (left), f_2 (middle), f_3 (right), and their estimates. Result of 5 simulations for each regression function, with sample size $n = 100$ and normal errors with $\sigma = 0.1$. The estimates are very reasonable, even for f_3, which is just Lipschitz, and not C^1.

which does not approximate the derivative of the regression function, and thus it demands less smoothness and approximates better non differentiable functions.

In the second part of this simulation study we investigated the mean square error, bias and variance of our convex estimate. For this we considered again the three regression functions in f_1, f_2, f_3 and computed—with 2000 simulation runs—the curves for the mean square error, squared bias and variance. The results shown in Figure 3 look very much alike those in Birke & Dette (2007), except for the ones related to f_3, where our estimator seems to be better. In this figure the mean square error, bias and variance of the estimator by local linear polynomials are represented by the dashed lines, while those quantities related to our convex estimator are represented by the solid lines.

Finally, in Figure 4 we show approximate 95% confidence bands for one estimate to each of the previous regression functions. We ran a simulation with 100 uniformly distributed design points for the explanatory variables and added a normal noise with $\sigma = 0.1$ to the response variable. In order to use the existing results on the width of the confidence bands from Johnston (1982, Corollary 3.1) (see also Theorem 7 and its corollary), the smoothing step was done with the formula

$$f_n(x) = \frac{1}{nh_n} \sum_{j=1}^{n} K((x - X_i)/h_n),$$

where $K(x) = \frac{3}{4}(1 - x^2)_+$ is Epanechnikov's Kernel. This regression formula has bad approximation properties at the endpoints of the interval, which explains the mild misfit observed there. The width of the band was 0.1392, 0.1382, 0.1628, for the estimate corresponding to f_1, f_2, and f_3, respectively.

4.3 Rabbits' data

We studied an example considered in Dudzinski & Mykytowycz (1961), who analyzed the relationship between age and eye lens weight for rabbits in Australia. This relationship is expected to be guided by a concave function. In this study, the dry weight of the eye lens was measured (in milligrams) for 71 free-living wild rabbits of known age (measured in days). A detailed description of the experiment and the data can be found in `http://www.statsci.org/data/oz/rabbit.html`. The data was analyzed by Ratkowsky (1983) using a parametric nonlinear growth

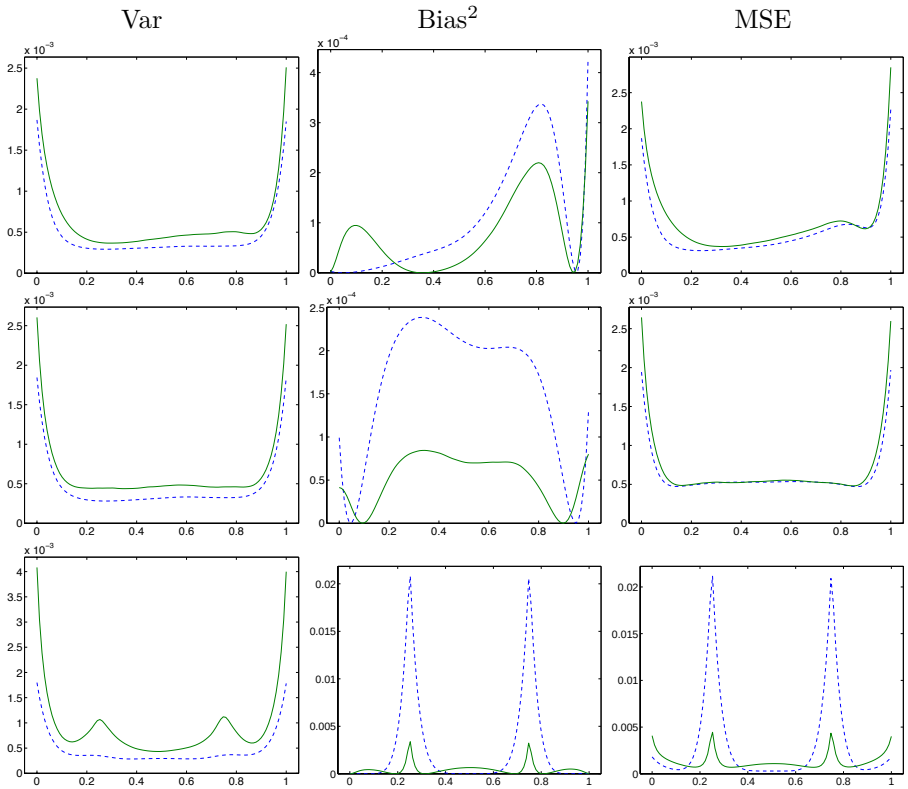


Figure 3: Variance (left), squared Bias (middle) and Mean Square Error (right) of our convex estimate (solid line) and of the local linear estimate (dashed line). These indicators were obtained with 2000 simulation runs, for f_1 (top), f_2 (middle) and f_3 (bottom), using 100 uniformly distributed design points for the explanatory variable, and normal error with $\sigma = 0.1$ for the observed variable. Only small differences between the local linear estimate and our convex estimate are observed. In some cases, our convex estimate is even better.

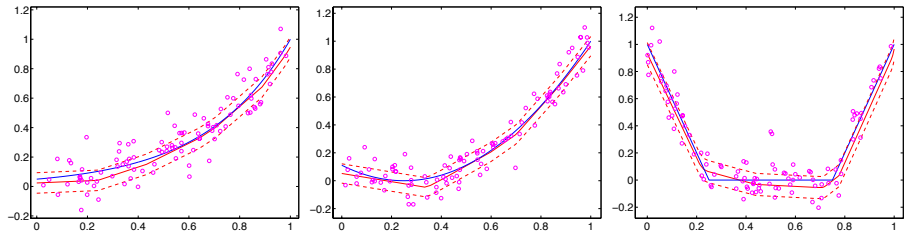


Figure 4: Approximate 95% confidence bands. In each plot we show the exact regression function, the estimate, the 95% confidence bands, and the data points for $N = 100$ design points, and normal error with $\sigma = 0.1$. The bands have width 0.1392, 0.1382, and 0.1628 for f_1 (left), f_2 (middle), f_3 (right), which were computed with the formula provided in Corollary 8.

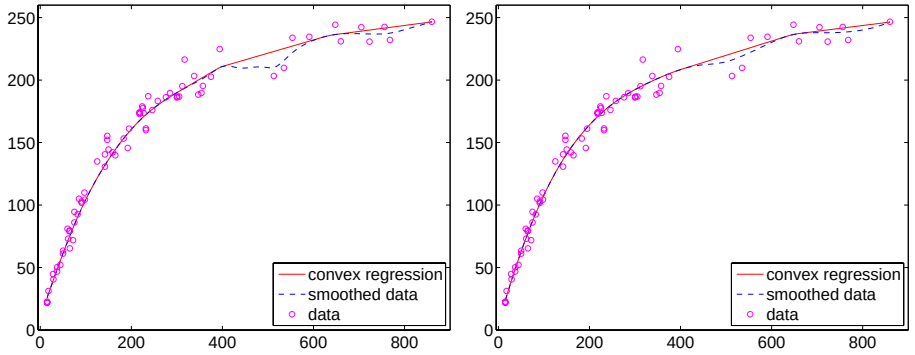


Figure 5: Convex regression of the rabbits' data. Dry weight of the eye lens (milligrams) versus age (days). Plot of the local polynomial smoothing (dashed blue), the convex estimate (solid red), and data points (magenta). The smoothing step is based on local polynomials of degree 1 (left) and degree 2 (right). The fit obtained is really excellent, with no essential difference between degree 1 and 2.

model, and by Birke & Dette (2007) with their non-parametric convex regression method. We used our method to obtain the concave regression, with the smoothing step performed with local polynomials of degree 1 and 2, and report the findings in Figure 5. In both cases, the bandwidth for the local polynomial smoothing was set using cross-validation, and the result of the smoothing step was evaluated at a uniform grid of 100 points. The convexification step yielded the estimated regression curves that can be observed in Figure 5 with an excellent fit to the data.

4.4 Two dimensional simulations

In this section we briefly illustrate the finite sample properties of the convex estimate of a regression function in two dimensions by means of a simulation study. For this purpose we considered the following convex regression function:

$$f(x_1, x_2) = \max\left\{2x_1^2 + x_2^2/2, 3x_1 + x_2\right\},$$

which is convex, and only Lipschitz. In Figure 6 we show the level curves of two estimated regression functions and the exact one in two simulations. We took uniform grids of 10×10, and 20×20 in each situation for the explanatory variable, and added normal error with $\sigma = 0.1$ to the value of $f(x_1, x_2)$ to emulate an observed variable. The level curves shown in the figure show a very good fit, even for a coarse grid of only 20×20 points.

In the second part of this simulation study we investigated the mean square error, bias and variance of our convex estimate. For this we considered again the same two dimensional regression function f and calculated by 2000 simulation runs the surfaces for the mean square error, squared bias and variance. The results depicted in Figure 7 show that the variance is concentrated on the boundary but is one order of magnitude smaller than the squared bias and the mean square error. These last two quantities are concentrated on the region of the domain where the regression function is not C^1.

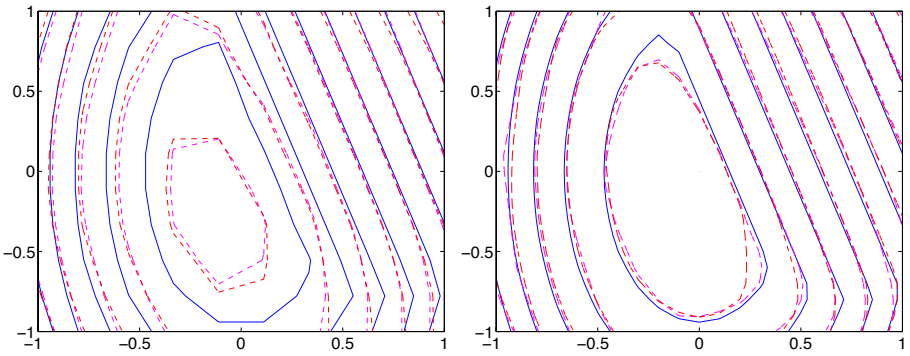


Figure 6: Level curves of two estimated regression functions (dashed red and magenta) and the exact one (solid blue) in two simulations with uniform design for the explanatory variables. One for a grid of 10×10 (left) points and another for a grid of 20×20 (right). The fit looks very good, even for a grid of only 20×20 points.

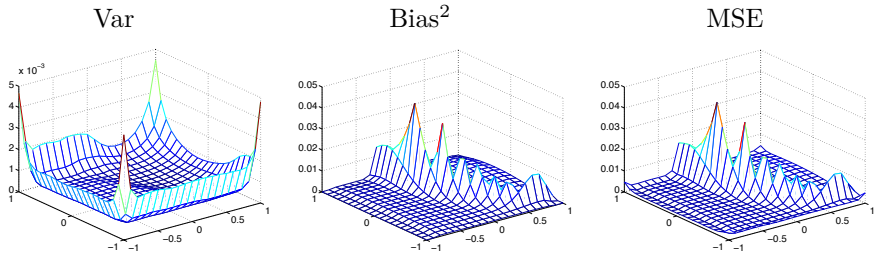


Figure 7: Variance (left), squared Bias (middle) and Mean Square Error (right) for the two dimensional simulation. These indicators were obtained with 2000 simulation runs, using 10×10 (left) and 20×20 uniformly distributed design points for the explanatory variable, and normal error with $\sigma = 0.1$ for the observed variable.

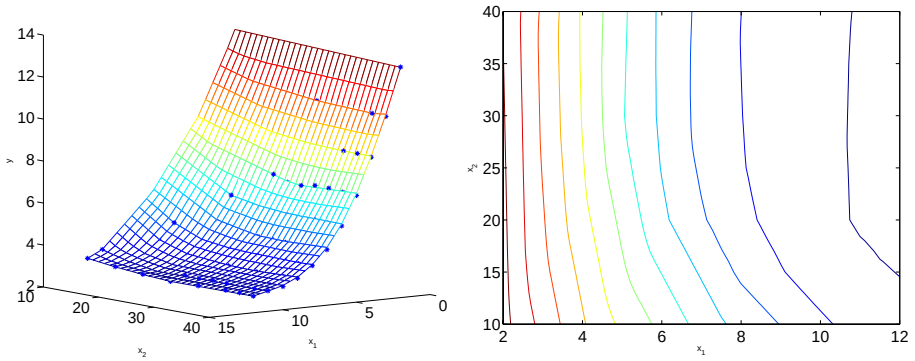


Figure 8: Pareto surface obtained as the convex regression of the data points (left). Contour curves of the convex graph (right) showing an excelent fit of the data (see also Figure 2 of Hoffmann et al., 2006).

4.5 Radiotherapy data

We studied a two dimensional example considered in Siem et al. (2005) (see also Hoffmann et al., 2006), who approximated the Pareto surface of a multiobjective optimization problem arising in the computation of the precise radiation dose. This Pareto surface is convex under certain conditions, and it should be computed from some Pareto points that can be measured from the patient. We obtained data from a patient of the Radboud University Nijmegen Medical Centre, in Nijmegen, the Netherlands. The data correspond to a multiobjective optimization problem with three objectives and contains 69 data points, which, due to measuring errors, are not convex.

By using our method we are able to smooth the data, obtaining a convex Pareto surface defined as a maximum of planes. This surface is initially defined on the convex hull of the X data, and we have extended it to a rectangular domain by considering the same maximum of planes.

In Figure 8 we show the data points together with the convex regression surface (left), and the contours of the convex regression (right), showing an excelent fit of the data (see also Figure 2 in Hoffmann et al., 2006).

Acknowledgements

We would like to thank A. Hoffmann and A. Siem for sharing with us the data of the Radboud University Nijhmegen Medical Centre, used in Section 4.5.

References

Barber, C. B., Dobkin, D. P. & Huhdanpaa, H. (1996). The quickhull algorithm for convex hulls. *ACM Trans. Math. Software* **22**, 469–483.

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. & Brunk, H. D. (1972). *Statistical inference under order restrictions. The theory and application of isotonic regression.* John Wiley & Sons, London-New York-Sydney. Wiley Series in Probability and Mathematical Statistics.

Bickel, P. J. & Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1**, 1071–1095.

Bierens, H. J. (1983). Uniform consistency of kernel estimators of a regression function under generalized conditions. *J. Amer. Statist. Assoc.* **78**, 699–707.

Birke, M. & Dette, H. (2007). Estimating a convex function in nonparametric regression. *Scand. J. Statist.* **34**, 384–404.

Boente, G. & Fraiman, R. (1991). Strong uniform convergence rates for some robust equivariant nonparametric regression estimates for mixing processes. *International Statistical Review* **59**, 355–372.

Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Ann. Math. Statist.* **26**, 607–616.

Collomb, G. (1984). Prédiction non paramétrique: étude de l'erreur quadratique du prédictogramme. *Statist. Anal. Données* **9**, 1–34.

Collomb, G. & Härdle, W. (1986). Strong uniform convergence rates in robust nonparametric time series analysis and prediction: kernel regression estimation from dependent observations. *Stochastic Process. Appl.* **23**, 77–89.

Devroye, L. (1978). The uniform convergence of the Nadaraya-Watson regression function estimate. *Canad. J. Statist.* **6**, 179–191.

Dony, J. (2008). *Nonparametric regression estimation.* PhD in Mathematical sciences, Free University of Brussels.

Dony, J. & Einmahl, U. (2006). Weighted uniform consistency of kernel density estimators with general bandwidth sequences. *Electron. J. Probab.* **11**, no. 33, 844–859 (electronic).

Dony, J., Einmahl, U. & Mason, D. M. (2006). Uniform in bandwidth consistency of local polynomial regression function estimators. *Austr. J. Statist.* **35**, 105–120.

Dony, J. & Mason, D. M. (2008). Uniform in bandwidth consistency of conditional U-statistics. *Bernoulli* **14**, 1108–1133.

Dudzinski, M. & Mykytowycz, R. (1961). The eye lens as an indicator of age in the wild rabbit in australia. *CSIRO Wildlife Research* **6**, 156–159.

Eggermont, P. P. B. & LaRiccia, V. N. (2006). Uniform error bounds for smoothing splines. In *High dimensional probability*, vol. 51 of *IMS Lecture Notes Monogr. Ser.* Inst. Math. Statist., Beachwood, OH, 220–237.

Einmahl, U. & Mason, D. M. (2000). An empirical process approach to the uniform consistency of kernel-type function estimators. *J. Theoret. Probab.* **13**, 1–37.

Einmahl, U. & Mason, D. M. (2005). Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.* **33**, 1380–1403.

Fraser, D. A. S. & Massam, H. (1989). A mixed primal-dual bases algorithm for regression under inequality constraints. Application to concave regression. *Scand. J. Statist.* **16**, 65–74.

Groeneboom, P., Jongbloed, G. & Wellner, J. A. (2001). Estimation of a convex function: characterizations and asymptotic theory. *Ann. Statist.* **29**, 1653–1698.

Hanson, D. L. & Pledger, G. (1976). Consistency in concave regression. *Ann. Statist.* **4**, 1038–1050.

Härdle, W. (1989). Asymptotic maximal deviation of M-smoothers. *J. Multivariate Anal.* **29**, 163–179.

Hildreth, C. (1954). Point estimates of ordinates of concave functions. *J. Amer. Statist. Assoc.* **49**, 598–619.

Hoffmann, A. L., Siem, A. Y. D., den Hertog, D., Kaanders, J. & H., H. (2006). Derivative-free generation and interpolation of convex Pareto optimal IMRT plans. *Physics in Medicine and Biology* **51**, 6349–6369.

Johnston, G. J. (1982). Probabilities of maximal deviations for nonparametric regression function estimates. *J. Multivariate Anal.* **12**, 402–414.

Konakov, V. D. & Piterbarg, V. I. (1984). On the convergence rate of maximal deviation distribution for kernel regression estimates. *J. Multivariate Anal.* **15**, 279–294.

Mammen, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.* **19**, 741–759.

Mukerjee, H. (1988). Monotone nonparameteric regression. *Ann. Statist.* **16**, 741–750.

Ratkowsky, D. (1983). *Nonlinear regression modeling.* Marcel Dekker Inc.

Robertson, T., Wright, F. T. & Dykstra, R. L. (1988). *Order restricted statistical inference.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Ltd., Chichester.

Rosenblatt, M. (1976). On the maximal deviation of k-dimensional density estimates. *Ann. Probability* **4**, 1009–1015.

Roussas, G. G. (1990). Nonparametric regression estimation under mixing conditions. *Stochastic Process. Appl.* **36**, 107–116.

Schuster, E. & Yakowitz, S. (1979). Contributions to the theory of nonparametric regression, with application to system identification. *Ann. Statist.* **7**, 139–149.

Shih, T. D., Chen, V. C. P. & Kim, S. B. (2006). Convex version of multivariate adaptive regression splines for optimization. *Proceedings of the 2006 IE Research Conference (Orlando, FL)* (preprint: http://students.uta.edu/dt/dts5878/convexMARS.pdf).

Siem, A. Y. D., den Hertog, D. & L., H. A. (2005). Multivariate convex approximation and least-norm convex data-smoothing. *CentER Discussion Paper* **2005-73**, 1–21.

Tran, L. T. (1993). Nonparametric function estimation for time series by local average estimators. *Ann. Statist.* **21**, 1040–1057.

Truong, Y. K. & Stone, C. J. (1992). Nonparametric function estimation involving time series. *Ann. Statist.* **20**, 77–97.

Wu, C.-F. (1982). Some algorithms for concave and isotonic regression. In *Optimization in statistics*, vol. 19 of *Stud. Management Sci.* North-Holland, Amsterdam, 105–116.

Corresponding author:
Liliana Forzani
Address: IMAL, Güemes 3450, 3000 Santa Fe, Argentina
e-mail: liliana.forzani@gmail.com