

# LARGE SAMPLE PROPERTIES OF SHAPE RESTRICTED REGRESSION ESTIMATORS WITH SMOOTHNESS ADJUSTMENTS

Jayanta Kumar Pal and Michael Woodroffe

*The University of Michigan*

**Running headline :** Asymptotic properties of smooth monotone estimators.

*Abstract:* The isotonic regression problem with a smoothness penalty is considered. The shape-restricted smooth estimator was characterized as a solution to a set of recurrence relations in Tantiyaswasdikul and Woodroffe [16]. Using a related Green's function, the estimator can be represented as a kernel regression estimator. Under regularity conditions on the underlying regression function, asymptotic normality of the estimator is established for a large range of choices of the tuning parameter.

*Key words and phrases:* Asymptotic normality, Green's function, Monotonic functions, Smoothing Spline.

## 1 Introduction

There are many regression problems for which the underlying function is known to be monotone; for example, dose-response experiments in biology and the modeling of disease incidences as function of the toxicity level etc. Industry examples like the effect of temperature on the strength of steel are also available. The least squares estimator for this problem is widely known, and described by Robertson, Wright and Dykstra(1989) in [14]. Unfortunately, this estimator lacks smoothness, and its non-normal asymptotic behavior complicates its use. We refer to Wright in [17] for the derivation and Groeneboom et al in [2] for its distribution. Accordingly, there has been recent interest in combining smoothness and monotonicity. Friedman et al in [1], Mukherjee in [9] and Mammen in [5] were early contributors, and Ramsey in [12] is a recent method to counter the problem. Hall et al in [3] provide a recent review with references. Here we follow the approach of Tantiyaswasdikul and Woodroffe in [16] (referred as TW henceforth),

who proposed a penalized least square estimator.

In Section 2, we review the derivation of TW and give several properties. Section 3 contains the main results of this paper. The estimator of TW can be described as an approximate solution to a differential equation with boundary values. It is shown that their estimator can be approximated by a kernel estimator, using the Green's function for a closely related boundary value problem as a kernel. Many asymptotic properties of the estimator follows as consequences. Section 4 contains simulation results and an example. There is some precedence for the use of Green's functions to approximate splines in the absence of shape restriction : Rice and Rosenblatt in [13], Silverman in [15], Messer in [7] and Nychka in [10] notable among them. These ideas are modified to allow for shape restriction.

## 2 The Smoothing Spline

Consider a regression problem,

$$Y_i = \phi(t_{i,n}) + \epsilon_i, \quad i = 1, \dots, n \quad (1a)$$

where  $t_{i,n}$  are pre-specified design points on  $[0, 1]$  with  $0 < t_{1,n} < \dots < t_{n,n} < 1$ , and  $\phi$  is a non-decreasing function. Here we can restrict ourselves to the unit interval, without loss of generality. We suppose that  $\epsilon_i$  are mean zero independent and identically distributed random errors, with a moment-generating function finite on some neighborhood of 0. From now on, we will denote  $t_{i,n}$  as  $t_i$  only, to avoid notational complications.

Let  $\omega$  be the uniform distribution on  $t_1, \dots, t_n$ , and let  $g$  be a piecewise constant function for which  $g(t_k) = Y_k$  for  $k = 1, \dots, n$ . To combine smoothness with shape restrictions in a weighted manner, TW chooses to minimize a penalized least-squares criterion of the form

$$\psi(f) = \int_0^1 (g - f)^2 d\omega + \alpha \int_0^1 f'(t)^2 dt.$$

### 2.1 Characterization of the estimator

Let  $H$  be the set of all absolutely continuous functions  $h$  for which  $h' \in L^2[0, 1]$ . Let  $H^+ = \{h \in H : h' \geq 0 \text{ a.e.}\}$ , the set of non-decreasing  $h \in H$ . Let  $F(x) =$

$\int_0^x f(t)d\omega(t)$ , and  $G(x) = \int_0^x g(t)d\omega(t)$ . Further, the positive part of any real  $x$  is denoted as  $x_+ = x1_{x \geq 0}$ .

**Theorem 1.** *Necessary and sufficient conditions for  $f \in H^+$  to minimize  $\psi$  on  $H^+$  are that  $F(1) = G(1)$  and,*

$$\alpha f' = (F - G)_+ \text{ a.e.} \quad (1)$$

TW used this characterization to derive an algorithm to compute the estimate. Let,  $F_0 = G_0 = 0$  and for  $k = 1, \dots, n$ ,

$$f_k = f(t_k), \quad F_k = \frac{1}{n}(f_1 + \dots + f_k), \quad G_k = \frac{1}{n}(Y_1 + \dots + Y_k).$$

Then the condition (1) reduces to,

$$\alpha f'(t) = (F_k - G_k)_+ \text{ for } t_k \leq t < t_{k+1} \quad (2)$$

and also  $F_n = G_n$  by virtue of the first condition in the theorem. Consequently,  $f$  is a continuous non-decreasing piecewise linear function with

$$f_k = f_{k-1} + \frac{1}{\alpha}(F_{k-1} - G_{k-1})_+(t_k - t_{k-1}) \quad (3)$$

for  $k = 1, \dots, n$ . Relation (3) determines  $f_1, \dots, f_n$  for a given  $f_0$ , and  $f_0$  is determined from the condition  $F_n = G_n$  (with the aid of Property A below). Several properties of the estimator are needed and mentioned below.

For  $c \in \mathbb{R}$ , let  $f_0(\alpha, c) = c$ , and define

$$\begin{aligned} f_k(\alpha, c) &= f_{k-1}(\alpha, c) + \frac{1}{\alpha}(F_{k-1}(\alpha, c) - G_{k-1})_+(t_k - t_{k-1}) \\ F_k(\alpha, c) &= \frac{1}{n}(f_1(\alpha, c) + f_2(\alpha, c) + \dots + f_k(\alpha, c)) \end{aligned}$$

We need to solve for real  $c$ , which satisfies the equation ' $F_n(\alpha, c) = G_n$ '. We denote the solutions as  $c_\alpha$ ,  $\hat{f}_k(\alpha) = f_k(\alpha, c_\alpha)$  and  $\check{F}_k(\alpha) = F_k(\alpha, c_\alpha)$  respectively. The following lemma ensures that a unique solution exists. The proof can be found in the Appendix.

**Lemma 1.** *The following properties can be derived from (3).*

*Property A: For fixed  $\alpha$ ,  $f_k(\alpha, c)$  is strictly increasing in  $c$ , is continuous in  $c$ , and goes to  $-\infty$  and  $\infty$  as  $c$  drifts to  $-\infty$  and  $\infty$ . Consequently, so does  $F_k(\alpha, c)$ . Therefore,  $F_n(\alpha, c) = G_n$  has a unique solution  $c_\alpha$ . Moreover,*

$$\min_{1 \leq k \leq n} \frac{nG_k}{k} \leq c_\alpha \leq G_n.$$

*Property B:* For each  $c \in \mathcal{R}$ , both  $f_k(\alpha, c)$  and  $F_k(\alpha, c)$  are non-increasing in  $\alpha$ .

*Property C:*  $c_\alpha$  is non-decreasing in  $\alpha$ .

Property A enables us to set up a bisection search algorithm to compute the final estimate in an iterative procedure. The next lemma investigates the shape of the solution. We denote the LSE by  $\vec{f}$ . We recall that it is the left hand slope of the greatest convex minorant (we call it  $\tilde{G}$  with knot values  $\tilde{G}_k$ ), of the cumulative sum diagram  $G$ . The proofs are outlined in the Appendix.

**Lemma 2.** *The estimator satisfies the following :*

*Property D:*  $\check{F}_k(\alpha)$  is non-decreasing in  $\alpha$ .

*Property E:* For all  $k = 1, \dots, n$  and all  $\alpha$ , we have,  $\check{F}_k(\alpha) \geq \tilde{G}_k$ .

*Property F:* For all  $\alpha$  and  $t$  between 0 and 1,  $\vec{f}(t_1) \leq \hat{f}(\alpha, t) \leq \vec{f}(t_n)$ .

*Property G:* For equally spaced design points, (i.e.  $t_i = i/n$ ),  $(\check{F}(\alpha, t) - \tilde{G}(t))_- = O_p(\frac{1}{n})$  uniformly in  $\alpha$  and  $t$ , where  $x_- = x1_{x \leq 0}$ .

Finally, we refer to the main result of Pal and Woodroffe [11], which shows that the cumulative sum diagram  $G$ , and its greatest convex minorant  $\tilde{G}$  are close. Under the assumption that the true regression function  $f$  is strictly increasing with derivative bounded away from 0, it is clear from Theorem 1 of [11] that,

$$\max_{1 \leq k \leq n} |\tilde{G}_k - G_k| = O_P\left(\left(\frac{\log n}{n}\right)^{\frac{2}{3}}\right).$$

We will refer to this result in the future.

### 3 Asymptotic properties of the estimator

The main result is presented in this section. The dependence on  $n$  becomes important, and there is a slight change in the notation. Henceforth,  $c = c_\alpha$  is that value of  $c$  identified in Property A;  $\hat{f}$  is the resulting estimator; and  $\hat{F}(t) = \int_0^t \hat{f}(s) ds$ .

### 3.1 Green's Function

Since the characterization equation (3) of the estimator does not yield a closed form representation of the estimator, it seems impossible to compute or approximate its bias, variance, mean square error etc. theoretically. However, we can proceed by replacing that difference equation by an analogous differential equation which fortunately has a closed form solution.

Consider the differential equation,

$$\alpha F''(t) = F(t) - H(t) \quad 0 \leq t \leq 1 \quad (4)$$

with boundary conditions  $F(0) = 0$  and  $F(1) = A$ . We assume that  $H$  is absolutely continuous with derivative  $h$ . Letting  $\beta = 1/\sqrt{\alpha}$ , the homogeneous equation  $(\alpha F'' - F) = 0$  has solutions  $e^{\pm\beta t}$ , and the corresponding Green's Function is,

$$K_\alpha(t, s) = \frac{1}{2}\beta e^{-\beta|t-s|} \quad \text{for } 0 \leq t \leq 1.$$

To solve the differential equation with boundary conditions, let,

$$F_0(t) = \int_0^1 K_\alpha(t, s)H(s)ds \quad 0 \leq t \leq 1.$$

Then,  $\alpha F_0'' = F_0 - H$ . To satisfy the boundary conditions, let,

$$F(t) = c_0(\beta)e^{-\beta t} + c_1(\beta)e^{-\beta(1-t)} + F_0(t).$$

The values of  $c_0(\beta)$  and  $c_1(\beta)$  can be evaluated from the boundary conditions as

$$c_0(\beta) = \frac{F_0(1) - A - F_0(0)e^\beta}{e^\beta - e^{-\beta}},$$

$$c_1(\beta) = \frac{(A - F_0(1))e^\beta + F_0(0)}{e^\beta - e^{-\beta}}.$$

from which it follows that  $|c_0(\beta)| + |c_1(\beta)| \leq 6\|H\| + 4A$ , for  $\beta \geq 1$ , where  $\|H\| = \sup_{0 \leq t \leq 1} |H(t)|$ . It is quite important that this is the unique solution to the Differential Equation (4) with the given boundary conditions.

Define, for all  $l \in L^1$ ,

$$\mathcal{K}_\alpha l(t) = \int_0^1 K_\alpha(t, s)l(s)ds.$$

**Lemma 3.** *In case of equally spaced design variables [i.e.  $t_i = i/n$ ],  $\|\hat{F} - \check{F}\| = O_P(1/n)$ .*

*Proof :* For any  $0 \leq x \leq 1$  and  $n$ , let  $k = \lfloor nx \rfloor$ . Then,

$$\begin{aligned} |\hat{F}(x) - \check{F}(x)| &= \left| \int_0^x \hat{f}(t) d\omega(t) - \int_0^x \hat{f}(t) dt \right| \\ &= \left| \sum_{i=1}^k \hat{f}\left(\frac{i}{n}\right) \frac{1}{n} - \frac{1}{n} \sum_{i=1}^{k-1} \hat{f}\left(\frac{i}{n}\right) - \frac{1}{2n} \hat{f}\left(\frac{k}{n}\right) - \frac{1}{2} \left(x - \frac{k}{n}\right) \left(\hat{f}\left(\frac{k}{n}\right) + \hat{f}(x)\right) \right| \\ &\leq \frac{1}{2n} \sup |\hat{f}|. \end{aligned}$$

Moreover,  $\sup |\hat{f}| \leq |\vec{f}(t_n)| + |\vec{f}(t_1)| = O_p(1)$ , since, the LSE  $\vec{f}$  is a stochastically bounded estimator. Hence, the lemma follows.  $\square$

From now on, we will consider uniformly spaced points only.

The next proposition allows us to represent  $\hat{F}$  as the sum of a convolution of  $K_\alpha$  (defined in Section 3.1) with the greatest convex minorant  $\tilde{G}$  and a remainder term that is of smaller order.

**Proposition 1.** *Under the assumptions stated in Section 2,*

$$\hat{F}(t) = \mathcal{K}_\alpha \tilde{G}(t) + \mathcal{K}_\alpha R(t) + c_0(\beta) e^{-\beta t} + c_1(\beta) e^{\beta(t-1)}$$

where both  $c_0$  and  $c_1$  are stochastically bounded functions of  $\beta$ , and  $\|R\| = O_P(n^{-\frac{2}{3}}(\log n)^{\frac{2}{3}})$ .

*Proof:* **Property G** implies that,  $\|(\check{F} - \tilde{G}) - (\check{F} - \tilde{G})_+\| \leq O_p(1/n)$ . Combining that with (3), we get,

$$\begin{aligned} \|\alpha \hat{F}'' - (\hat{F} - \tilde{G})\| &= \|(\check{F} - G)_+ - (\hat{F} - \tilde{G})\| \\ &= \|(\check{F} - G)_+ - (\check{F} - \tilde{G}) - (\hat{F} - \check{F})\| \\ &\leq \|(\check{F} - G)_+ - (\check{F} - \tilde{G})_+\| + \|(\hat{F} - \check{F})\| + O_p\left(\frac{1}{n}\right) \\ &\leq \|\tilde{G} - G\| + O_p\left(\frac{1}{n}\right) \\ &= O_P\left(n^{-\frac{2}{3}}(\log n)^{\frac{2}{3}}\right). \end{aligned}$$

Let,  $R = (\hat{F} - \tilde{G}) - \alpha \hat{F}''$ . Then,  $\|R\| = O_P(n^{-\frac{2}{3}}(\log n)^{\frac{2}{3}})$  and  $\hat{F}$  satisfies,

$$\alpha \hat{F}'' - \hat{F} = -\tilde{G} - R = -H \text{ (say).}$$

Hence, using the uniqueness of the solution of Section 3.1,

$$\begin{aligned}\hat{F}(t) &= \int_0^1 K_\alpha(t, s)H(s)ds + c_0(\beta)e^{-\beta t} + c_1(\beta)e^{\beta(t-1)} \\ &= \int_0^1 K_\alpha(t, s)\tilde{G}(s)ds + \int_0^1 K_\alpha(t, s)R(s)ds + c_0(\beta)e^{-\beta t} + c_1(\beta)e^{\beta(t-1)},\end{aligned}$$

where  $c_0$  and  $c_1$  are defined as in Section 3.1 with  $A = G_n$ . By Marshall's lemma,  $\|\tilde{G} - F\| \leq \|G - F\| \rightarrow 0$ , since  $F$  is a convex function. (Refer to Robertson, Wright and Dykstra [14] (Page 329) for a statement of Marshall's lemma.) Therefore,  $\|\tilde{G}\|$  is bounded. So,  $H$  is bounded in  $n$  and  $t$  and boundedness of  $c_0(\beta)$  and  $c_1(\beta)$  follows. The second term of the above representation is of order  $n^{-\frac{2}{3}}(\log n)^{\frac{2}{3}}$ , since  $|\int_0^1 K_\alpha(t, s)R(s)ds| \leq \|R\| \int_0^1 K_\alpha(t, s)ds$ . The proposition follows.  $\square$

To get the analogous representation for  $\hat{f}$ , we need to define a few variables and functions related to the true regression function  $\phi$ , which will give us the bias and random components of the estimator. The cumulative regression functions are also required in this context. We define,

$$\phi_k = \phi(t_k), \quad \bar{\Phi}(x) = \int_0^x \phi(t)d\omega(t), \quad \Phi(x) = \int_0^x \phi(t)dt.$$

The next function has to be defined through a characterizing differential equation analogous to the original equation (4).

**Proposition 2.** *Suppose that the true regression function  $\phi$  is twice continuously differentiable. Then, there is a function  $\tau_\alpha$  which minimizes*

$$\int_0^1 (\tau - \phi)^2 dt + \alpha \int_0^1 \{\tau'\}^2 dt.$$

*among all functions and also satisfies the approximation,  $\tau_\alpha(t) = \phi(t) + \alpha\phi''(t) + o(\alpha)$ .*

*Proof:* We consider the minimization problem as a problem in Calculus of variation. Euler's equation gives the differential equation,  $\alpha\tau'' = \tau - \phi$ . Using the same Green's function technique as in Section 3.1,  $\tau_\alpha(t) = \mathcal{K}_\alpha\phi(t)$  satisfies the equation. However  $\mathcal{K}_\alpha\phi(t) = \phi(t) + \alpha\phi''(t) + o(\alpha)$  (It is a special case of Equation (6.4) in Theorem 2.2 by Nychka, [10]). The proposition follows.  $\square$

Now we are in a state, where we can derive the crucial representation of the estimator obtained in Section 2.

**Theorem 2.** *The regression estimator  $\hat{f}$  admits the representation as,*

$$\hat{f}(t) = \tau_\alpha(t) + \frac{\beta}{2n} \sum_{i=1}^n e^{-\beta|t-t_i|} \epsilon_i + O_P(n^{-\frac{2}{3}} \log n) \beta + e^{-\beta t(1-t)} O_P(\beta).$$

*uniformly in  $\alpha$  and in  $t \in (0, 1)$ .*

*Proof :* The Lebesgue integral  $\hat{F}$  can be written as in Proposition 1. Differentiating pointwise we get,

$$\begin{aligned} \hat{f}(t) &= \int_0^1 \frac{\partial}{\partial t} K_\alpha(t, s) \tilde{G}(s) ds + \int_0^1 \frac{\partial}{\partial t} K_\alpha(t, s) R(s) ds \\ &\quad - \beta e^{-\beta t} c_0(\beta) + \beta e^{-\beta(1-t)} c_1(\beta) \\ &= \tau_\alpha(t) + \int_0^1 \frac{\partial}{\partial t} K_\alpha(t, s) [G(s) - \bar{\Phi}(s)] ds + V_1(t) + V_2(t), \end{aligned}$$

where,

$$V_1(t) = \int_0^1 \frac{\partial}{\partial t} K_\alpha(t, s) (\tilde{G}(s) + R(s) - \Phi(s) - G(s) + \bar{\Phi}(s)) ds,$$

and

$$V_2(t) = -\beta e^{-\beta t} c_0(\beta) + \beta e^{-\beta(1-t)} c_1(\beta) - \frac{1}{2} \beta e^{-\beta(1-t)} \Phi(1) = e^{-\beta t(1-t)} O_P(\beta).$$

However,

$$\begin{aligned} |V_1(t)| &\leq \frac{1}{2} \|\tilde{G} - G + R - \Phi + \bar{\Phi}\| \int_0^1 \beta^2 e^{-\beta|t-s|} ds \\ &\leq \beta [\|\tilde{G} - G + R\| + \|\Phi - \bar{\Phi}\|]. \end{aligned}$$

Since  $\|\Phi - \bar{\Phi}\| = o(1/n)$ , and both  $\|\tilde{G} - G\|$  and  $\|R\|$  are  $O_p(n^{-2/3}(\log n)^{2/3})$ , we deduce,  $|V_1(t)| = O_P(n^{-2/3}(\log n)^{2/3})\beta$ . Finally,

$$\begin{aligned} \int_0^1 \frac{\partial}{\partial t} K_\alpha(t, s) [G(s) - \bar{\Phi}(s)] ds &= - \int_0^1 \frac{\partial}{\partial s} K_\alpha(t, s) [G(s) - \bar{\Phi}(s)] ds \\ &= -K_\alpha(t, 1) [G(1) - \bar{\Phi}(1)] + \int_0^1 K_\alpha(t, s) [dG(s) - d\bar{\Phi}(s)] \\ &= -\frac{\beta}{2n} e^{-\beta(1-t)} \sum_{i=1}^n \epsilon_i + \frac{\beta}{2n} \sum_{i=1}^n e^{-\beta|t-t_i|} \epsilon_i \\ &= O_P\left(\frac{\beta e^{-\beta(1-t)}}{\sqrt{n}}\right) + \frac{\beta}{2n} \sum_{i=1}^n e^{-\beta|t-\frac{i}{n}|} \epsilon_i. \end{aligned}$$

That completes the detail of the representation.  $\square$

*Remark :* Theorem 2 implies that the constrained estimator is approximately a kernel regression estimator by employing the Laplace Kernel function. Here, The tuning parameter  $\alpha$  plays a role similar to the bandwidth  $h$ . The asymptotic mean  $\tau_\alpha$  has a bias, that we seek to make negligible. We need  $\alpha$  to be reasonably small to ensure that. On the other hand, we do not want to let  $\beta$  grow rapidly, as that will inflate the random component. As a balance, we compromise  $\alpha$  to be in an admissible range.

**Corollary 1.** *Let  $\alpha$  satisfy the property  $\alpha n^{2/3} \rightarrow \infty$  but  $\alpha n^{2/5} \rightarrow 0$ . Suppose also that, the true regression function  $\phi$  is twice continuously differentiable with bounded second derivative. Then for  $t \in (0, 1)$ ,*

$$\sqrt{\frac{n}{\beta}}[\hat{f}(t) - \phi(t)] \Rightarrow N\left[0, \frac{\sigma^2}{4}\right]. \quad (5)$$

However, if  $\alpha n^{2/5} \rightarrow K \geq 0$  then,  $\sqrt{\frac{n}{\beta}}[\hat{f}(t) - \phi(t)] \Rightarrow N[K^{5/4}\phi''(t), \frac{\sigma^2}{4}]$ .

*Proof:* Define  $U_\alpha(t) = \frac{\beta}{2n} \sum_{i=1}^n e^{-\beta|t - \frac{i}{n}|} \epsilon_i$ . For a fixed  $t$ , this is a sequence of sums of a triangular array. We invoke the Lindeberg-Lévy Central Limit Theorem to verify that,

$$\sqrt{\frac{n}{\beta}}U_\alpha(t) \Rightarrow N\left[0, \frac{\sigma^2}{4} \int_{-\infty}^{\infty} e^{-2|u|} du\right] = N\left[0, \frac{\sigma^2}{4}\right].$$

The Lindeberg's condition is easily satisfied since  $\epsilon_i$ 's have finite moment generating function.

Moreover, from Proposition 2,

$$\sqrt{\frac{n}{\beta}}[\tau_\alpha(t) - \phi(t)] = \sqrt{\frac{n}{\beta}}[\alpha\phi''(t) + o(\alpha)] = K^{5/4}\phi''(t) + o(1).$$

The remainder terms  $V_1$  and  $V_2$  are  $o_P(1)$  for the admissible range of  $\alpha$ . The corollary follows.  $\square$

*Remark :* The case  $\alpha = 0$  guides us back to the LSE, which has a non-normal asymptotic distribution. The choice  $\alpha = n^{-2/3}$  yields the slowest rate of convergence (it is also at the boundary) in the limit, that of the MLE. Then,  $\beta$  is of the same order as the number of jump points of the MLE. A comparison to the theorem 2.2 by Nychka in [10] yields an approximately similar bias and a variance

twice that of the unconstrained spline. However, the asymptotic normality is an important feature here. The conditions on the magnitude of  $\alpha$  is more stringent than that of [10], a price paid for the penalization.

## 4 Simulation Results and applications

Before applying the technique to real data, we investigate the estimator's performance in simulations, especially for small sample size. Different underlying regression functions are considered : exponential (convex); sinusoidal (in its concave range); and two polynomials  $3x^2 - 2x^3$  (flat ends and inflexion) and  $2x^2 - x$  (violates the shape-restriction). Errors are generated from  $N(0,.01)$ , Student's  $t$  (with 4 degrees of freedom and scaled down by .1) and Beta (with parameters 3,2 and centralized). The TW estimators are compared with the LSE and the unconstrained smoothing splines (with optimal smoothing), graphically and numerically.

### 4.1 Optimal choice of $\alpha$

As shown in proposition 2, the estimator has a bias  $\alpha f''(t)$ . Though negligible for large  $n$ , it has to be accounted for in small sample. To counter that, we will follow the technique used in the selection of bandwidth in Kernel estimation to find out the optimal values of  $\alpha$ . Now,

$$MSE(x) = \frac{\beta\sigma^2}{4n} + \alpha^2 f''(x)^2. \Rightarrow IMSE = \frac{\beta\sigma^2}{4n} + \alpha^2 \int_0^1 f''(x)^2 dx.$$

So, the optimum value of  $\alpha$  is,

$$\alpha = \left[ \frac{16n \int_0^1 f''(x)^2 dx}{\sigma^2} \right]^{-\frac{2}{5}}.$$

Unfortunately, that depends on both  $\sigma^2$  and the underlying regression function  $f$ . Now,  $\sigma^2$  can be consistently estimated, as observed by Meyer and Woodroffe in [8]. However, since our method yields piecewise linear function, it is unable to estimate  $f''$ . To plug in an estimate of  $\int_0^1 f''(x)^2 dx$ , one can think of using a Kernel substitute. However, we prefer an easier route by pretending the regression function to be quadratic over the range  $[0, 1]$ , i.e. it has a constant second

derivative. Under the model  $f(t) = a + bt + ct^2$  we estimate  $\hat{c}$  using simple linear regression, and observe that  $\int_0^1 f''(x)^2 dx = 4c^2$ . Finally, we select our estimated smoothing parameter as,

$$\hat{\alpha} = \left[ \frac{64n\hat{c}^2}{\hat{\sigma}^2} \right]^{-\frac{2}{5}}.$$

The following table compares the approximate IMSE of the smooth estimators with data-driven  $\alpha$ , alongside that of the optimal  $\alpha$ , the LSE, and the linear smoothing spline with the smoothing parameter taken as its optimal value. We also provide the Monte Carlo standard error to give an idea of how much variation the different estimates have across simulations. ( We have 1000 MC simulations for all of them.) The sample size taken is  $n = 100$ . Clearly, the optimal smoothing would have worked substantially better than the LSE, and even the data-driven adaptive smoothing has a remarkable effect on the overall IMSE. As expected, the estimator fails in comparison to the unconstrained spline when the shape restriction is violated. When the true function is monotone, the TW estimator performs a little better than the unconstrained spline. The reason lies in the fact that the constraint preempts the estimator to have too many ups and downs, and therefore restricts it to have too much variation. The unconstrained spline does not share the property. This fact will also be corroborated by the graphical plots shown later.

## 4.2 Graphical comparisons

Figure 1 shows us the relative performance of the estimators for different regression functions and error distributions. The sample size is 100, but the signal-to-noise ratio is diminished to a scale of .2, to discern the plots through a cursory glance. Clearly, the smooth estimator alleviates the spiking problem, as well as reducing the roughness of the LSE. A function violating the shape-constraints ( $2x^2 - x$ ) over  $[0, 1]$  is included for comparison. Table 3 gives us the optimal values of  $\alpha$  and the IMSE for these individual curves. We compare the mean square error of the estimators at specific points, for different sample sizes. Figure 2 shows that the optimally smooth isotonic estimator always outperforms the LSE, and so does the adaptive smooth estimator in most cases. Though they have a bias that the LSE doesn't have, the variance is much lower. However, for

Mean function	Error distribution	TW optimal	TW adaptive	LSE (PAVA)	unconstrained spline(optimal)
$e^x$	$N(0, .01)$	.0102	.0227	.0234	.0113
	$.1t_4$	.0168	.0314	.0413	.0187
	$\beta(3, 2) - .6$	.0296	.0450	.0721	.0352
$\sin(\pi x/2)$	$N(0, .01)$	.0087	.0131	.0194	.0112
	$.1t_4$	.0151	.0192	.0343	.0194
	$\beta(3, 2) - .6$	.0260	.0285	.0644	.0347
$3x^2 - 2x^3$	$N(0, .01)$	.0097	.0186	.0207	.0127
	$.1t_4$	.0168	.0229	.0362	.0220
	$\beta(3, 2) - .6$	.0277	.0303	.0653	.0387
$2x^2 - x$	$N(0, .01)$	.0106	.0231	.0198	.0138
	$.1t_4$	.0175	.0320	.0347	.0236
	$\beta(3, 2) - .6$	.0306	.0457	.0635	.0426

Table 1: *The approximate IMSE obtained using Monte-Carlo averages of Riemann sums, for sample size  $n = 100$  and number of replications  $M = 1000$  for all combinations of error and mean function.*

points close to the ends and a skew error, the adaptive smooth estimator fares worse than the LSE. If the shape restriction is violated, the unconstrained spline works much better than the TW estimators. However, if the assumption holds, there is little to choose between them.

### 4.3 Analysis of ASA cars data

As an illustration of our method we use part of the "cars" data from the 1983 ASA Data Exposition. These data are available at the StatLib Internet site (<http://lib.stat.cmu.edu/datasets/cars.data>) at Carnegie Mellon University. Here, the covariate  $X$  are the engine outputs of different models of cars, in horsepower, and the response  $Y$  are their fuel efficiencies, to be studied as function of the engine outputs. Several methods of constrained as well as unconstrained smoothing have been applied to the data in Mammen, Marron, Turlach and Wand in [6]. Since logically more powerful engines require more fuel, a decreasing smooth function should be fitted to the observations. In Figure 3, we fit a non-smooth usual isotonic estimator (LSE), alongside the smooth isotonic estimator, and an unconstrained smoothing spline, using the data itself to design the bandwidth

Mean function	Error distribution	TW optimal	TW adaptive	LSE (PAVA)	unconstrained spline(optimal)
$e^x$	$N(0, .01)$	.0051	.0067	.0077	.0073
	$.1t_4$	.0098	.0114	.0239	.0145
	$\beta(3, 2) - .6$	.0176	.0232	.0295	.0233
$\sin(\pi x/2)$	$N(0, .01)$	.0054	.0071	.0075	.0074
	$.1t_4$	.0099	.0116	.0183	.0175
	$\beta(3, 2) - .6$	.0173	.0214	.0319	.0231
$3x^2 - 2x^3$	$N(0, .01)$	.0052	.0065	.0077	.0074
	$.1t_4$	.0115	.0186	.0239	.0158
	$\beta(3, 2) - .6$	.0189	.0256	.0323	.0248
$2x^2 - x$	$N(0, .01)$	.0050	.0068	.0074	.0072
	$.1t_4$	.0115	.0202	.0231	.0172
	$\beta(3, 2) - .6$	.0186	.0276	.0314	.0250

Table 2: The Monte Carlo Standard error of the IMSE, for sample size  $n = 100$  and number of replications  $M = 1000$  for all combinations of error and mean function.

Mean function	Error distribution	$\alpha$	IMSE(TW optimal)	IMSE(LSE)	IMSE(unconstrained spline)
$\sin(\pi x/2)$	$N(0, .01)$	.0044	.0059	.0200	.0080
$3x^2 - 2x^3$	$\beta(3, 2) - .6$	.0053	.0134	.0548	.0227
$2x^2 - x$	$N(0, .01)$	.0027	.0229	.0359	.0244
$e^x$	$\beta(3, 2) - .6$	.0040	.0534	.0853	.0655

Table 3: The value of the adaptive choice of  $\alpha$  and the approximate IMSE for the curves shown in Figure 1.

according to our method described above.

### Acknowledgment

We acknowledge Dr. Moulinath Banerjee and Dr. Jiayang Sun for their useful comments.

## 5 Appendix

### Proof of lemma 2:

Property A: The final inequality of Property A is a restatement of Lemma 4 in [16]. The rest of it is clear.  $\square$

Property B: For any  $\alpha < \beta$ ,  $f_0(\alpha, c) = c = f_0(\beta, c)$ . Suppose, inductively, that  $f_j(\alpha, c) \geq f_j(\beta, c)$  for  $j = 0, 1, \dots, k-1$  for some  $k$ . Then  $F_{k-1}(\alpha, c) \geq F_{k-1}(\beta, c)$ , and consequently,

$$\begin{aligned} f_k(\alpha, c) &= f_{k-1}(\alpha, c) + \frac{1}{\alpha}(F_{k-1}(\alpha, c) - G_{k-1})_+(t_k - t_{k-1}) \\ &\geq f_{k-1}(\beta, c) + \frac{1}{\beta}(F_{k-1}(\beta, c) - G_{k-1})_+(t_k - t_{k-1}) \\ &= f_k(\beta, c) \end{aligned}$$

and the result follows.  $\square$

Property C: If  $\alpha < \beta$ , then  $F_n(\alpha, c_\beta) \geq F_n(\beta, c_\beta) = G_n$ , and therefore,  $c_\alpha \leq c_\beta$ .  $\square$

**Proof of lemma 2:**

Property D: Let  $\alpha < \beta$ . Clearly,  $\check{F}_1(\alpha) = c_\alpha/n \leq c_\beta/n = \check{F}_1(\beta)$  and  $\check{F}_n(\alpha) = \check{F}_n(\beta) = G_n$ . Suppose,  $\check{F}_j(\alpha) > \check{F}_j(\beta)$  for some  $j$  between 2 and  $n-1$ , and let  $k$  be the smallest such  $j$ . Then  $\hat{f}_k(\alpha) > \hat{f}_k(\beta)$ . Therefore,

$$\begin{aligned} \hat{f}_{k+1}(\alpha) &= \hat{f}_k(\alpha) + \frac{1}{\alpha}(\check{F}_k(\alpha) - G_k)_+(t_{k+1} - t_k) \\ &> \hat{f}_k(\beta) + \frac{1}{\beta}(\check{F}_k(\beta) - G_k)_+(t_{k+1} - t_k) \\ &= \hat{f}_{k+1}(\beta) \end{aligned}$$

and  $\check{F}_{k+1}(\alpha) > \check{F}_{k+1}(\beta)$ . Proceeding like this, we get,  $\check{F}_n(\alpha) > \check{F}_n(\beta)$ , a contradiction to our assertion.  $\square$

Property E: Clearly,  $\check{F}_1(\alpha) = c_\alpha/n \geq \min_{1 \leq k \leq n} G_k/k = \check{G}_1$  and  $\check{F}_n(\alpha) = G_n = \check{G}_n$ . As above, suppose,  $\check{F}_j(\alpha) < \check{G}_j$  for some  $j$  between 2 and  $n-1$ , and let  $k$  be the smallest such  $j$ . Then, as  $\check{G}_k \leq G_k$  for all  $k$ ,

$$\begin{aligned} \hat{f}_{k+1}(\alpha) &= \hat{f}_k(\alpha) + \frac{1}{\alpha}(\check{F}_k(\alpha) - G_k)_+(t_{k+1} - t_k) \\ &= \hat{f}_k(\alpha) \\ &= n[\check{F}_k(\alpha) - \check{F}_{k-1}(\alpha)] \\ &< n[\check{G}_k - \check{G}_{k-1}] \\ &\leq n[\check{G}_{k+1} - \check{G}_k] \quad (\text{as } \check{G} \text{ is convex}). \end{aligned}$$

Therefore,

$$\begin{aligned}\check{F}_{k+1}(\alpha) &= \check{F}_k(\alpha) + \frac{\hat{f}_{k+1}(\alpha)}{n} \\ &< \tilde{G}_{k+1}.\end{aligned}$$

Consequently,  $\hat{f}_{k+2}(\alpha) = \hat{f}_{k+1}(\alpha) = \hat{f}_k(\alpha) \leq n[\tilde{G}_{k+1} - \tilde{G}_k] \leq n[\tilde{G}_{k+2} - \tilde{G}_{k+1}]$  (as above) and  $\check{F}_{k+2}(\alpha) < \tilde{G}_{k+2}$  as well. Proceeding in the same line, we conclude,  $\check{F}_n(\alpha) < \tilde{G}_n$ , a contradiction.  $\square$

Property F: Clearly, for all  $t$  and  $\alpha$ ,  $\hat{f}(\alpha, t) \geq c_\alpha \geq \vec{f}(t_1)$ , establishing the first inequality. The other inequality follows by symmetry.  $\square$

Property G: The following lemma shows that the MLE itself is stochastically bounded.

**Lemma 4.** *Both  $\vec{f}(t_1)$  and  $\vec{f}(t_n)$  are bounded w.p. 1.*

*Proof:*

$$\begin{aligned}\vec{f}(t_1) &= \min_{i \geq 1} \frac{Y_1 + \dots + Y_i}{i} \\ &\geq f(t_1) + \min_{i \geq 1} \frac{\epsilon_1 + \dots + \epsilon_i}{i} \\ &\geq f(0+) + \min_{i \geq 1} \frac{\epsilon_1 + \dots + \epsilon_i}{i}\end{aligned}$$

We know that,  $(\epsilon_1 + \dots + \epsilon_n)/n \rightarrow 0$  w.p.1, using Strong Law of Large Numbers. Hence,  $\min_{i \geq 1} (\epsilon_1 + \dots + \epsilon_i)/i > -\infty$  w.p. 1. Consequently,  $\vec{f}(t_1) > -\infty$  w.p. 1. Using symmetry with respect to the mean function, we get,  $\vec{f}(t_n) < \infty$  w.p. 1. The lemma follows.  $\square$

Now, for  $t_k \leq t < t_{k+1}$ ,  $\check{F}(\alpha, t) = \check{F}_k(\alpha)$ , and  $\tilde{G}(t)$  lies between  $\tilde{G}_k$  and  $\tilde{G}_{k+1}$ . Hence,  $|\tilde{G}(t) - \tilde{G}_k| \leq |\tilde{G}_k - \tilde{G}_{k+1}|$ . However,  $|\tilde{G}_k - \tilde{G}_{k+1}| = \vec{f}(t_k)(t_{k+1} - t_k) = \vec{f}(t_k)/n$ , and hence is  $O_p(1/n)$  uniformly in  $t$ , by lemma 4. The property follows by using Property E and observing that the previous bound is also uniform in  $\alpha$ .  $\square$

## References

- [1] Friedman, J. and Tibshirani, Rob (1984). The monotone Smoothing of Scatterplots. *Technometrics* 26, 3, August 1984.

- [2] Groeneboom, Piet and Wellner, Jon A. (2001). Computing Chernoff's distribution. *Journal of Computational & Graphical Statistics* June 2001, Vol. 10, I2, P 388.
- [3] Hall, Peter and Huang, Li-Shan (2001). Nonparametric kernel regression subject to monotonicity constraints. *Annals of Statistics* 29, no.3, 624-647.
- [4] Kiefer, J. and Wolfowitz, J. (1976). Asymptotically Minimax Estimation of concave and convex Distribution Functions. *Z. Wahrsch. verw. Gebiete*
- [5] Mammen, Enno (1991). Estimating a Smooth Monotone Regression Function. *Annals of Statistics* 19, 2, 724-740.
- [6] Mammen, E., Marron, J.S., Turlach, B.A. and Wand, M.P. A General Projection Framework for Constrained Smoothing. *Statistical Science* 2001, 16, 3, 232-248.
- [7] Messer, K. (1991). A Comparison of a Spline estimate to its equivalent Kernel Estimate. *Annals of Statistics* 19, 2, 817-829.
- [8] Meyer, Mary and Woodroffe, M.B. (2000). On the degrees of freedom of Shape restricted Regression. *The Annals of Statistics*. 28, 4, 1083-1104.
- [9] Mukherjee, Hari (1988). Monotone Non-parametric Regression. *The Annals of Statistics*. 16, 2, 741-750.
- [10] Nychka, Douglas (1995). Splines as Local Smoothers. *The Annals of Statistics*. 23, 4, 1175-1197.
- [11] Pal, Jayanta and Woodroffe, M. B. (2004). On the distance between the Cumulative Sum Diagram and its Greatest Convex Minorant under unequally spaced design points. *University of Michigan Department of Statistics Technical Report 412*. Submitted to *Scandinavian Journal of Statistics*. Also available at [www.stat.lsa.umich.edu/~jpal/kwext.pdf](http://www.stat.lsa.umich.edu/~jpal/kwext.pdf)
- [12] Ramsay, J. O. (1998). Estimating Smooth Monotone Functions. *Journal of the Royal Statistical Society. Series B*. 60, 2, 365-375.
- [13] Rice, J. and Rosenblatt, M. (1983) Smoothing Splines: regression, derivatives and deconvolution. *Annals of Statistics*. 11 141-156.

- [14] Robertson, Tim, Wright, F. T. and Dykstra, R. L. (1987). Order Restricted Statistical Inference. *John Wiley and Sons*
- [15] Silverman, B. (1984). Spline Smoothing: the equivalent variable Kernel Method. *Annals of Statistics* 12 898-916.
- [16] Tantiyaswasdikul, Chim and Woodroffe, M. B. (1994). Isotonic Smoothing Splines under Sequential Designs. *Journal of Statistical Planning and Inference*. 38, 1, January 1994, 75-87.
- [17] Wright, F. T. (1981). The asymptotic behavior of monotone regression estimates. *The Annals of Statistics*. 9, 443-448.
- [18] Wright, Ian and Wegman, Edward (1980). Isotonic, Convex and Related Splines. *The Annals of Statistics*. 8, 5, 1023-1035.

Department of Statistics, University of Michigan

E-mail: (jpal@umich.edu)

Department of Statistics, University of Michigan

E-mail: (michaelw@umich.edu)

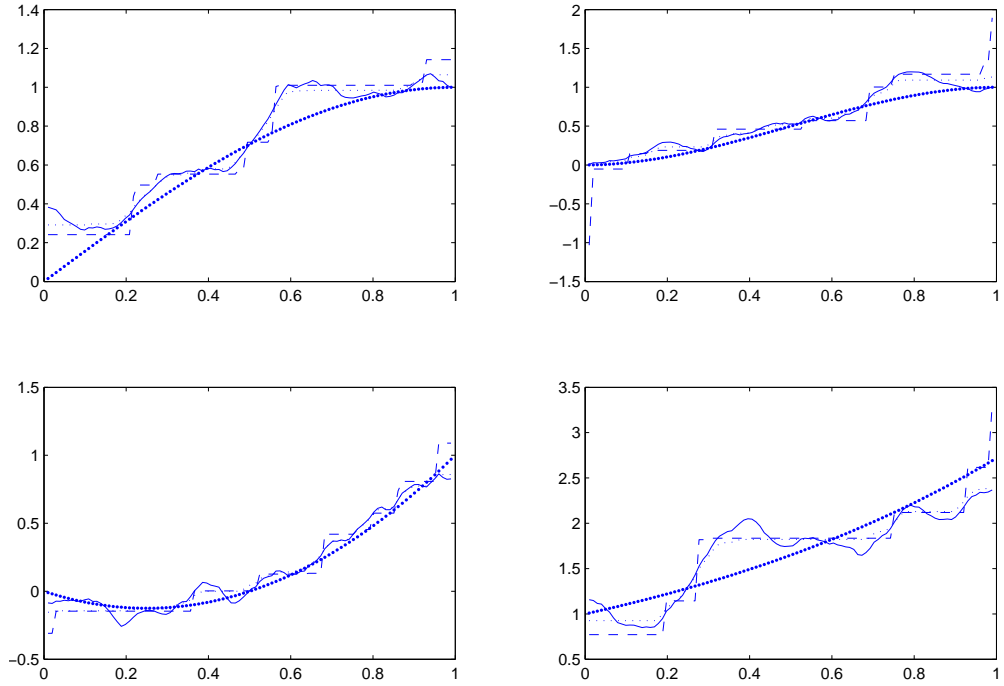


Fig. 1: The plot of the smooth estimators alongwith the LSE over  $[0, 1]$  for sample size 100. The bold dotted line is the true function, whereas the dashed, dotted and solid lines represent the LSE, the optimally smooth estimator and the unconstrained smooth estimator. The regression functions are  $\sin(\pi x/2)$ ,  $3x^2 - 2x^3$ ,  $2x^2 - x$  and  $e^x$  respectively. The errors are generated from Normal (left panels) and Beta (right panels). The values of alpha and the IMSE for each plot are mentioned in Table 3.

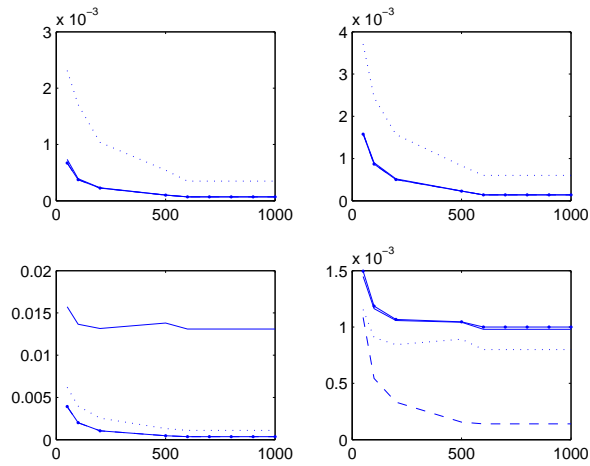


Fig. 2: The mean square error of the smooth estimators alongwith the MLE. The MSE for the MLE is the dotted line, whereas the dash-dotted, dashed and the solid lines represent the MSE for smooth estimators with optimal smoothing and data-estimated optimal smoothing, and the unconstrained spline respectively. The chosen mean function-error distribution-point of interest combinations are  $e^x - N(0, .01) - .25$ ,  $\sin(\pi x/2) - .1t_4 - .5$ ,  $3x^2 - 2x^3 - \beta(3, 2) - .6 - .75$  and  $2x^2 - x - N(0, .01) - .5$  respectively. As the first three plots show, there is no difference between the constrained and the unconstrained smoothing splines when the shape restriction is present.

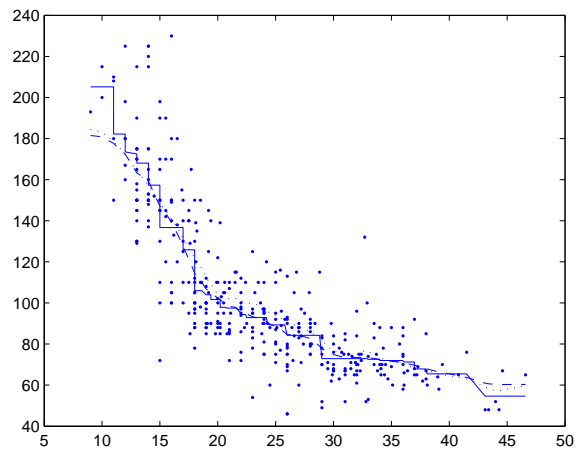


Fig. 3: Scatterplot of the fuel efficiency as a function of engine output. The PAVA estimator is the solid line, and the dashed line is the smooth estimator with optimally chosen smoothing parameter. The dotted line indicates a smoothing spline without the shape-restriction.