Automatic and asymptotically optimal data sharpening for nonparametric regression

Fang Yao[a], Thomas C.M. Lee[b,c,*]

[a]Department of Statistics, University of Toronto, 100 St. George Street, Toronto, Ontario, Canada M5S 3G3
[b]Department of Statistics, Chinese University of Hong Kong, Shatin, N.T., Hong Kong
[c]Colorado State University, USA

ARTICLE INFO

ABSTRACT

In this article we consider data-sharpening methods for nonparametric regression. In particular modifications are made to existing methods in the following two directions. First, we introduce a new tuning parameter to control the extent to which the data are to be sharpened, so that the amount of sharpening is adaptive and can be tuned to best suit the data at hand. We call this new parameter the sharpening parameter. Second, we develop automatic methods for jointly choosing the value of this sharpening parameter as well as the values of other required smoothing parameters. These automatic parameter selection methods are shown to be asymptotically optimal in a well defined sense. Numerical experiments were also conducted to evaluate their finite-sample performances. To the best of our knowledge, there is no bandwidth selection method developed in the literature for sharpened nonparametric regression.

1. Introduction

Data sharpening are data pre-processing procedures that can be applied to enhance the performances of certain standard and relatively simple estimation methods. They aim to produce pre-processed data in such a way that when these pre-processed data are fed to a simple estimation method, the final estimation results are improved relative to the case when the original raw data were used. One of the earlier applications of data sharpening was for probability density estimation (e.g., see Samiuddin and el Sayyad, 1990; Choi and Hall, 1999; Hall and Minnotte, 2002). Since then other data-sharpening methods have also been developed for nonparametric regression (Choi et al., 2000), hazard rate estimation (Claeskens and Hall, 2002) and spectral density estimation (Yao and Lee, 2007). In addition, data sharpening has also been applied to nonparametric estimation subject to constraints (e.g., see Braun and Hall, 2001; Hall and Kang, 2005). It has been shown theoretically that such data-sharpening methods are capable of reducing the estimation bias to a higher order, while at the same time only inflate the variance by a constant factor.

In this article we study data-sharpening methods in the context of nonparametric regression, for which pioneering work was done by Choi et al. (2000). In Choi et al. (2000) the authors propose three different sharpening strategies for the Nadaraya–Watson estimator. The first strategy sharpens the explanatory variable, the second strategy sharpens the response variable, while the last strategy sharpens simultaneously both the explanatory and response variables. The exact formulae used by these three sharpening strategies were designed and motivated by careful and large-sample theoretical considerations. Motivated by our previous work on spectral density estimation (Yao and Lee, 2007), in the present paper we develop two modifications of the above sharpening strategies and investigate their empirical performances. The first modification is concerned with the sharpening of

* Corresponding author at: Department of Statistics, Chinese University of Hong Kong, Shatin, N.T., Hong Kong. Tel.: +852 26097941; fax: +852 26035188.
E-mail addresses: fyao@utstat.toronto.edu (F. Yao), tlee@sta.cuhk.edu.hk (T.C.M. Lee).

the response variable. Unlike those "fixed-amount" sharpening formulae examined by Choi et al. (2000), we allow the amount of sharpening to be adjustable so that it could be tuned to adapt best to the data at hand. We achieve this via the introduction of a new tuning parameter, which we shall call the *sharpening parameter*. Our second modification is, for each sharpening strategy, the proposal of an automatic method for jointly selecting all the relevant tuning parameters, including the sharpening parameter and the bandwidth(s) that control the amount of smoothing. In the regression setting we are unaware of any bandwidth selection method developed in the literature for data sharpening, not to mention the non-existence of any method for choosing the sharpening parameter. Therefore, these new parameter selection methods greatly enhance the applicability and practicality of data sharpening for nonparametric regression.

The new parameter selection method was developed using the idea of Stein's unbiased risk estimation (SURE, Stein, 1981). That is, an unbiased estimator for the L_2 risk between the true function and the regression curve estimator is first constructed and then the parameters are chosen as the joint minimizer of this risk estimator obtained from the sharpened data. We have investigated the theoretical properties of our proposed unbiased risk parameter selection methods. By adopting and modifying a technique of Li (1987), we were able to show that these selection methods are asymptotically optimal in the following sense: the ratio of the L_2 loss between the true and the estimated curve to the corresponding minimal possible loss converges to 1 in probability. See Theorem 2 for a precise description of this result.

We have also conducted a numerical experiment to evaluate the practical performances of the above three different sharpening strategies when combined with automatic parameter selection. Based on the empirical results, we recommend sharpening only the response variable. It is because it provides a best compromise between statistical and computational efficiencies.

Before we proceed we highlight the major differences between the current work and the work conducted in Yao and Lee (2007). First, only regularly-spaced data and one sharpening strategy are considered in Yao and Lee (2007) while the current work extends to irregularly-spaced data and multiple sharpening strategies. Second, these two pieces of work study different random structures: the current one studies additive errors with moment constraints, whereas for Yao and Lee (2007) the common distribution of the multiplicative errors is assumed to be completely known (standard exponential). Last and most importantly, the present work studies the consistency properties of the risk estimators while no theoretical study is offered by Yao and Lee (2007) in this aspect.

The rest of this article is organized as follows. Section 2 provides background material and defines our data-sharpening formulae. Automatic selection methods for choosing the sharpening parameter and bandwidths are described in Section 3. Theoretical properties and practical performances of our sharpening methods are given, respectively, in Sections 4 and 5. Lastly technical details are deferred to Appendix A.

2. Adaptive data-sharpening estimators

Suppose we observe n independently and identically distributed (i.i.d.) observations $\{(X_i, Y_i)\}_{i=1}^n$ that were generated from a bivariate population (X, Y). Our interest is to estimate the regression function $g(x) = E(Y|X=x)$ that satisfies the model assumption

$$Y = g(X) + \varepsilon, \tag{1}$$

where the independent random error ε is zero-mean with variance $\text{Var}(\varepsilon) = \sigma^2$. We also assume that the marginal density of X (i.e., the design density) has a compact support.

Let $K(\cdot)$ be a nonnegative kernel function and h be a bandwidth. The Nadaraya–Watson estimator of $g(x)$ is defined as

$$\hat{g}(x) = \sum_{j=1}^n W_j(x; h) Y_j, \tag{2}$$

where

$$W_j(x; h) = \frac{K_h(X_j - x)}{\sum_{k=1}^n K_h(X_k - x)} \quad \text{with } K_h(X_j - x) = \frac{1}{h} K\left(\frac{X_j - x}{h}\right). \tag{3}$$

Notice that $\hat{g}(x)$ can be interpreted as a weighted average of the Y_j's with weights $W_j(x; h)$'s sum to unity.

Following Choi et al. (2000), we investigate three different sharpening strategies. The first strategy sharpens only the explanatory variable X. It aims to cluster the design points X_i's closer together at regions for which the design density is high, and also to separate them further apart at regions that have low design density. As illustrated by Choi et al. (2000), the goal for this strategy is to obtain an improved estimate of the design density, which will in turn lead to a better estimate for $g(x)$. The second strategy keeps the design points unchanged and instead adjusts the response variable Y. For any (X_i, Y_i) that is believed to be near a local maximum, this second strategy aims to increase the value of Y_i with the hope that this will offset the downward bias typically caused by the local averaging operation around local maxima. Similarly, it also aims to decrease the value of Y_i if this Y_i is believed to be near a local minimum. Lastly, the third sharpening strategy combines the above two strategies together; that is, both the explanatory and response variables will be adjusted.

Our adaptive sharpening strategies modify those in Choi et al. (2000) in the following ways. First, in Choi et al. (2000) a single bandwidth is used for sharpening the explanatory variable X as well as for smoothing the response variable Y, while we propose using two different bandwidths, h_x and h_y, for executing these two different tasks. Second, for the sharpening of Y, we introduce

an additional parameter α, the sharpening parameter mentioned before, to control the amount of sharpening. We have developed methods for automatically choosing the values of h_x, h_y and α.

Sharpening of explanatory variable X : Let h_x be the bandwidth for sharpening the explanatory variable X The "sharpened" design points are defined as

$$\widehat{X}_i = \sum_{j=1}^{n} W_j(X_i; h_x)X_j, \quad i = 1, \ldots, n, \tag{4}$$

where $W_j(X_i; h_x)$ was given in (3). With these sharpened design points, one could obtain a data-sharpening estimator for $g(\cdot)$ by smoothing $\{(\widehat{X}_i, Y_i)\}_{i=1}^{n}$ with a second bandwidth h_y. Denote the resulting estimator as $\tilde{g}^X(\cdot)$. It admits the expression

$$\tilde{g}^X(x) = \sum_{j=1}^{n} \widetilde{W}_j(x; h_y)Y_j \quad \text{with } \widetilde{W}_j(x; h_y) = \frac{K_{h_y}(\widehat{X}_j - x)}{\sum_{k=1}^{n} K_{h_y}(\widehat{X}_k - x)}. \tag{5}$$

Notice that two bandwidths (h_x and h_y) are required for computing $\tilde{g}^X(\cdot)$, and that the weights $\widetilde{W}$'s depends implicitly on h_x through the $\widehat{X}_j$'s. Also notice that $\tilde{g}^X(\cdot)$ is essentially the same as the "unsharpened" estimator $\hat{g}(\cdot)$, except now the raw data (X_i, Y_i)'s are replaced by the sharpened data $(\widehat{X}_i, Y_i)$'s. In below we shall refer $\tilde{g}^X(\cdot)$ as the " *X*-only estimator" of $g(\cdot)$.

Sharpening of response variable Y : The second data-sharpening estimator for $g(\cdot)$ is defined as follows. Denote the unsharpened fitted value of Y_i as $\widehat{Y}_i$; that is, $\widehat{Y}_i = \hat{g}(X_i)$ with $\hat{g}(\cdot)$ defined in (2). The sharpened response variable is defined as

$$\widetilde{Y}_i = (1 + \alpha)Y_i - \alpha\widehat{Y}_i, \quad i = 1, \ldots, n, \tag{6}$$

where $0 \leq \alpha \leq 1$ is a tuning parameter that controls the degree of sharpening to Y_i. We shall call α the *sharpening parameter*. When $\alpha = 1$ the above sharpening formula reduces to the corresponding sharpening formula studied by Choi et al. (2000), while no sharpening is done when $\alpha = 0$. Now the resulting data-sharpening estimator $\tilde{g}^Y(\cdot)$ for $g(\cdot)$ can be obtained by smoothing $\{(X_i, \widetilde{Y}_i)\}_{i=1}^{n}$ with bandwidth h_y:

$$\tilde{g}^Y(x) = \sum_{j=1}^{n} W_j(x; h_y)\widetilde{Y}_j, \tag{7}$$

where $W_j(x; h_y)$ was defined in (3). Note that $\tilde{g}^Y(\cdot)$ requires both h_y and α to be pre-specified. We shall call $\tilde{g}^Y(\cdot)$ the "*Y*-only estimator" for $g(\cdot)$.

Sharpening of both explanatory and response variables X and Y : Lastly we present the data-sharpening estimator for $g(\cdot)$ for the case when both X and Y are sharpened. A more consistent notation for such an estimator is $\tilde{g}^{X,Y}(\cdot)$, but for simplicity we remove the superscripts and denote this estimator as $\tilde{g}(\cdot)$. This data-sharpening estimator is defined similarly to the unsharpened estimator $\hat{g}(\cdot)$, except that (X_i, Y_i)'s are now replaced by $(\widehat{X}_i, \widetilde{Y}_i)$'s:

$$\tilde{g}(x) = \sum_{j=1}^{n} \widetilde{W}_j(x; h_y)\widetilde{Y}_j, \tag{8}$$

where $\widehat{X}_i, \widetilde{Y}_i$ and $\widetilde{W}_j(x; h_y)$ were defined in (4), (6) and (5), respectively. Observe that for $\tilde{g}(\cdot)$ three parameters (h_x, h_y, α) are needed to be pre-selected. In sequel $\tilde{g}(\cdot)$ is referred as the "*X*-and-*Y* estimator" for $g(\cdot)$.

One can speculate that the introduction of the sharpening parameter in the last two situations leads to a more flexible and data-adaptive sharpening strategy. Indeed, the simulation results to be reported below suggest that this adaptive sharpening strategy performs very well in practice.

3. Automatic parameter selection

In practice the calculation of any of the above three versions of data-sharpening estimators for $g(\cdot)$ requires the selection of all or some of the following three tuning parameters: h_x, h_y and α. This section presents automatic selection methods for choosing their values. These methods are based on the unbiased risk estimation approach of Stein (1981). That is, the parameters are chosen as the minimizers of an unbiased estimator of an appropriate risk function measuring the distance between the true and the estimated $g(\cdot)$. We begin by defining the risk function for each of the three versions of data-sharpening estimators. We need the following notation to proceed. Let $\mathcal{X}_n = (X_1, \ldots, X_n)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$. For $i = 1, \ldots, n$, write $g_i = g(X_i)$, $\tilde{g}_i^X = \tilde{g}^X(X_i)$, $\tilde{g}_i^Y = \tilde{g}^Y(X_i)$ and $\tilde{g}_i = \tilde{g}(X_i)$. Finally set $\boldsymbol{g} = (g_1, \ldots, g_n)^T$, $\tilde{\boldsymbol{g}}^X = (\tilde{g}_1^X, \ldots, \tilde{g}_n^X)^T$ $\tilde{\boldsymbol{g}}^Y = (\tilde{g}_1^Y, \ldots, \tilde{g}_n^Y)^T$ and $\tilde{\boldsymbol{g}} = (\tilde{g}_1, \ldots, \tilde{g}_n)^T$.

For the *X*-only estimator, conditioned on $\mathcal{X}_n$, the corresponding risk function depends on the bandwidths h_x and h_y, and is defined as

$$R_X(h_x, h_y) = \frac{1}{n}E\left\{\sum_{i=1}^{n}(g_i - \tilde{g}_i^X)^2 \|\mathcal{X}_n\right\} = \frac{1}{n}E(\|\boldsymbol{g} - \tilde{\boldsymbol{g}}^X\|^2 |\mathcal{X}_n), \tag{9}$$

where $\| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$ is the Euclidean norm associated with the corresponding inner product. Ideally h_x and h_y should be chosen as the joint minimizer of $R_X(h_x, h_y)$. However, in practice this is not possible, as the true risk $R_X(h_x, h_y)$ is an unknown quantity. To overcome this issue, we shall construct an unbiased estimator for $R_X(h_x, h_y)$ and select h_x and h_y as the joint minimizer of such an unbiased risk estimator.

In a similar fashion, the risk functions for the *Y*-only and the *X*-and-*Y* data-sharpening estimators are defined, respectively, as

$$R_Y(h_y, \alpha) = \frac{1}{n} E \left\{ \sum_{i=1}^{n} (g_i - \tilde{g}_i^Y)^2 | \mathscr{X}_n \right\} = \frac{1}{n} E(\| \mathbf{g} - \tilde{\mathbf{g}}^Y \|^2 | \mathscr{X}_n) \tag{10}$$

and

$$R(h_x, h_y, \alpha) = \frac{1}{n} E \left\{ \sum_{i=1}^{n} (g_i - \tilde{g}_i)^2 | \mathscr{X}_n \right\} = \frac{1}{n} E(\| \mathbf{g} - \tilde{\mathbf{g}} \|^2 | \mathscr{X}_n). \tag{11}$$

For clarity, denote the residual sum of squares for the *X*-only estimator $\tilde{g}^X(\cdot)$ as

$$\text{RSS}_X(h_x, h_y) = \sum_{i=1}^{n} (Y_i - \tilde{g}_i^X)^2.$$

Similarly the residual sum of squares for the other two estimators $\tilde{g}^Y(\cdot)$ and $\tilde{g}(\cdot)$ are denoted as

$$\text{RSS}_Y(h_y, \alpha) = \sum_{i=1}^{n} (Y_i - \tilde{g}_i^Y)^2 \quad \text{and} \quad \text{RSS}(h_x, h_y, \alpha) = \sum_{i=1}^{n} (Y_i - \tilde{g}_i)^2,$$

respectively. In the next theorem, we present our unbiased estimators for the above three risk functions (9)–(11).

Theorem 1. *Conditioned on $\mathscr{X}_n$,*

$$\widehat{R}_X(h_x, h_y) = \frac{\text{RSS}_X(h_x, h_y)}{n} + \frac{\sigma^2}{n} \sum_{i=1}^{n} \{2\widetilde{W}_i(X_i; h_y) - 1\}, \tag{12}$$

$$\widehat{R}_Y(h_y, \alpha) = \frac{\text{RSS}_Y(h_y, \alpha)}{n} + \frac{\sigma^2}{n} \sum_{i=1}^{n} \left[2 \left\{ (1 + \alpha) W_i(X_i; h_y) - \alpha \sum_{j=1}^{n} W_j(X_i; h_y) W_i(X_j; h_y) \right\} - 1 \right] \tag{13}$$

and

$$\widehat{R}(h_x, h_y, \alpha) = \frac{\text{RSS}(h_x, h_y, \alpha)}{n} + \frac{\sigma^2}{n} \sum_{i=1}^{n} \left[2 \left\{ (1 + \alpha) \widetilde{W}_i(X_i; h_y) - \alpha \sum_{j=1}^{n} \widetilde{W}_j(X_i; h_y) W_i(X_j; h_y) \right\} - 1 \right], \tag{14}$$

are unbiased estimators of the risk functions $R_X(h_x, h_y)$, $R_y(h_y, \alpha)$ and $R(h_x, h_y, \alpha)$, respectively.

If the noise variance σ^2 is known, then the tuning parameters (h_x, h_y, α) can be chosen as the joint minimizer of the corresponding risk estimator. If σ^2 is unknown, methods are available for obtaining an asymptotically unbiased and $\sqrt{n}$-consistent estimate $\hat{\sigma}^2$ that is independent of the tuning parameters (h_x, h_y, α) (e.g., see Buckley et al., 1988; Hall et al., 1990; Hall and Marron, 1990). It is straightforward to see that the asymptotic unbiasedness of the risk estimators in Theorem 1 still holds if the unknown σ^2 is replaced by such an independent estimate $\hat{\sigma}^2$.

If the noise variance σ^2 is known or an estimate $\hat{\sigma}^2$ can be obtained independently of the tuning parameters, it is straightforward to see that the minimization of the risk estimators $\hat{R}_X(h_x, h_y)$, $\hat{R}_Y(h_y, \alpha)$ and $\hat{R}(h_x, h_y, \alpha)$ in Theorem 1 are equivalent to the minimization of

$$\widetilde{R}_X(h_x, h_y) = \frac{\text{RSS}_X(h_x, h_y)}{n} + \frac{2\sigma^2}{n} \sum_{i=1}^{n} \widetilde{W}_i(X_i; h_y), \tag{15}$$

$$\widetilde{R}_Y(h_y, \alpha) = \frac{\text{RSS}_y(h_y, \alpha)}{n} + \frac{2\sigma^2}{n} \sum_{i=1}^{n} \left\{ (1 + \alpha) W_i(X_i; h_y) - \alpha \sum_{j=1}^{n} W_j(X_i; h_y) W_i(X_j; h_y) \right\} \tag{16}$$

and

$$\widetilde{R}(h_x, h_y, \alpha) = \frac{\text{RSS}(h_x, h_y, \alpha)}{n} + \frac{2\sigma^2}{n} \sum_{i=1}^{n} \left\{ (1 + \alpha) \widetilde{W}_i(X_i; h_y) - \alpha \sum_{j=1}^{n} \widetilde{W}_j(X_i; h_y) W_i(X_j; h_y) \right\}, \tag{17}$$

respectively. In practice, for the three data-sharpening estimators $\tilde{g}^X(\cdot)$, $\tilde{g}^Y(\cdot)$ and $\tilde{g}(\cdot)$, we propose choosing their tuning parameters as, respectively, the joint minimizer of $\widetilde{R}_X(h_x, h_y)$, $\widetilde{R}_Y(h_y, \alpha)$ and $\widetilde{R}(h_x, h_y, \alpha)$. In all our numerical work below σ^2 is assumed unknown and is estimated by the fourth order σ^2 estimator proposed by Hall et al. (1990), which admits the form

$$\hat{\sigma}^2 = \frac{1}{n-4} \sum_{k=1}^{n-4} \left(\sum_{j=0}^{4} d_j Y_{j+k} \right)^2,$$

with $(d_0, d_1, d_2, d_3, d_4) = (0.2708, -0.0142, 0.6909, -0.4858, -0.4617)$.

4. Theoretical properties

This section summarizes our theoretical findings of the proposed parameter selection procedures that minimize the risk estimators defined in (15)–(17). More specifically, we will establish that these parameter selection procedures are asymptotically optimal in a well defined sense, as stated in (21)–(23) below.

For technical feasibility, we shall assume that the risk estimators (15)–(17) are minimized over the discrete index sets $\mathscr{A}_n^X$, $\mathscr{A}_n^Y$ and $\mathscr{A}_n$, respectively. In other words, say for $\widetilde{R}_X(h_x, h_y)$ in (15), its joint minimizer is restricted to be an element of $\mathscr{A}_n^X$, where $\mathscr{A}_n^X$ can be seen as a two-dimensional gridded values of (h_x, h_y). One could always keep increasing the grid density of $\mathscr{A}_n^X$ so that $\mathscr{A}_n^X$ is dense in $\mathscr{R}^2$. Therefore, practically speaking, when the grid density is high enough, no difference would be made no matter if $\widetilde{R}_X(h_x, h_y)$ was minimized over $\mathscr{A}_n^X$ or $\mathscr{R}^2$. Similar comments apply to both $\mathscr{A}_n^Y$ and $\mathscr{A}_n$. In below, we shall denote the order of the cardinalities of $\mathscr{A}_n^X$, $\mathscr{A}_n^Y$ and $\mathscr{A}_n$ as, respectively, n^{δ_X}, n^{δ_Y} and n^{δ}. That is, $|\mathscr{A}_n^X| = O(n^{\delta_X})$, $|\mathscr{A}_n^Y| = O(n^{\delta_Y})$ and $|\mathscr{A}_n| = O(n^{\delta})$ for some $\delta_X, \delta_Y, \delta > 0$.

To continue, define, respectively, the loss functions for the three versions of the sharpening estimators as

$$L_X(h_x, h_y) = \frac{1}{n} \sum_{i=1}^{n} (g_i - \tilde{g}_i^X)^2 = \frac{1}{n} \|\boldsymbol{g} - \tilde{\boldsymbol{g}}^X\|^2, \tag{18}$$

$$L_Y(h_y, \alpha) = \frac{1}{n} \sum_{i=1}^{n} (g_i - \tilde{g}_i^Y)^2 = \frac{1}{n} \|\boldsymbol{g} - \tilde{\boldsymbol{g}}^Y\|^2 \tag{19}$$

and

$$L(h_x, h_y, \alpha) = \frac{1}{n} \sum_{i=1}^{n} (g_i - \tilde{g}_i)^2 = \frac{1}{n} \|\boldsymbol{g} - \tilde{\boldsymbol{g}}\|^2. \tag{20}$$

Let $(\hat{h}_x, \hat{h}_y) \in \mathscr{A}_n^X$, $(\hat{h}_y, \hat{\alpha}) \in \mathscr{A}_n^Y$ and $(\hat{h}_x, \hat{h}_y, \hat{\alpha}) \in \mathscr{A}_n$ be, respectively, the minimizers of (15)–(17). That is, they are the parameters selected by the proposed parameter selection procedures. Then our proposed selection procedures are asymptotically optimal in the following sense:

$$\frac{L_X(\hat{h}_x, \hat{h}_y)}{\inf_{(h_x, h_y) \in \mathscr{A}_n^X} L_X(h_x, h_y)} \xrightarrow{P} 1, \tag{21}$$

$$\frac{L_Y(\hat{h}_y, \hat{\alpha})}{\inf_{(h_y, \alpha) \in \mathscr{A}_n^Y} L_Y(h_y, \alpha)} \xrightarrow{P} 1 \tag{22}$$

and

$$\frac{L(\hat{h}_x, \hat{h}_y, \hat{\alpha})}{\inf_{(h_x, h_y, \alpha) \in \mathscr{A}_n} L(h_x, h_y, \alpha)} \xrightarrow{P} 1, \tag{23}$$

for any sequence of design points $\mathscr{X}_n$. These asymptotic optimality for the proposed procedures are established in Theorem 2 below. Similar definitions for asymptotic optimality have also been studied by previous authors in different contexts, for both parametric and nonparametric model selection problems. For examples, see Craven and Wahba (1979) on generalized cross-validation and Li (1986, 1987) on Mallow's C_L and (generalized) cross-validation.

Now we present the required assumptions for establishing (21)–(23). Define the $n \times n$ matrices $\boldsymbol{W} = \{W_j(X_i; h_y)\}_{1 \le i,j \le n}$ and $\widetilde{\boldsymbol{W}} = \{\widetilde{W}_j(X_i; h_y)\}_{1 \le i,j \le n}$. From (5), (7) and (8), the so-called "smoother matrices" for obtaining the sharpened estimators $\tilde{\boldsymbol{g}}^X$, $\tilde{\boldsymbol{g}}^Y$ and $\tilde{\boldsymbol{g}}$ are, respectively, $\boldsymbol{M}_X = \widetilde{\boldsymbol{W}}$, $\boldsymbol{M}_Y = (1 + \alpha)\boldsymbol{W} - \alpha \boldsymbol{W}^2$ and $\boldsymbol{M} = (1 + \alpha)\widetilde{\boldsymbol{W}} - \alpha \widetilde{\boldsymbol{W}} \boldsymbol{W}$. That is, $\tilde{\boldsymbol{g}}^X = \boldsymbol{M}_X \boldsymbol{Y}$, $\tilde{\boldsymbol{g}}^Y = \boldsymbol{M}_Y \boldsymbol{Y}$ and $\tilde{\boldsymbol{g}} = \boldsymbol{M} \boldsymbol{Y}$. Denote the maximum singular value of an arbitrary matrix $\boldsymbol{A}$ by λ_A. The assumptions required for establishing the asymptotic

optimality of $\widetilde{R}_X(h_x, h_y)$ are: for any sequence of design points $\mathscr{X}_n$,

(A1) $\lim \sup_{n\to\infty} \sup_{(h_x,h_y)\in\mathscr{A}_n^X} \lambda_{\boldsymbol{M}_X} < \infty$,

(A2) $\sum_{(h_x,h_y)\in\mathscr{A}_n^X} \{nR_X(h_x, h_y)\}^{-m_X} \to 0$ as $n \to \infty$,

(A3) $E(\varepsilon^{4m_X}) < \infty$,

for some $m_X > 0$ satisfying certain condition discussed below. Assumption (A1) is natural, and in fact if $\lambda_{\boldsymbol{M}_X} > 1$, then $\tilde{\boldsymbol{g}}^X$ is inadmissible and dominated by some other linear estimators (Cohen, 1966). To understand (A2), one first notes that in nonparametric regression with the sample size n goes to infinity, the optimal risk $R_X(h_x, h_y)$ is typically of order $n^{-(1-\delta_X')}$ for some $\delta_X' > 0$. If the cardinality of $\mathscr{A}_n^X$ is of polynomial order n^{δ_X} for some $\delta_X > 0$, one can always find an $m_X > \delta_X/\delta_X'$ so that (A2) is satisfied. The last assumption (A3) is just a standard moment condition on the error distribution.

Analogously, the assumptions for establishing the asymptotic optimality of $\widetilde{R}_Y(h_y, \alpha)$ and $\widetilde{R}(h_x, h_y, \alpha)$ are: for any sequence of design points $\mathscr{X}_n$,

(A1*) $\lim \sup_{n\to\infty} \sup_{(h_y,\alpha)\in\mathscr{A}_n^Y} \lambda_{\boldsymbol{M}_Y} < \infty$,

(A2*) $\sum_{(h_y,\alpha)\in\mathscr{A}_n^Y} \{nR_Y(h_y, \alpha)\}^{-m_Y} \to 0$ as $n \to \infty$,

(A3*) $E(\varepsilon^{4m_Y}) < \infty$,

and

(A1†) $\lim \sup_{n\to\infty} \sup_{(h_x,h_y,\alpha)\in\mathscr{A}_n} \lambda_{\boldsymbol{M}} < \infty$,

(A2†) $\sum_{(h_x,h_y)\in\mathscr{A}_n} \{nR(h_x, h_y, \alpha)\}^{-m} \to 0$ as $n \to \infty$,

(A3†) $E(\varepsilon^{4m}) < \infty$,

for some $m_Y > 0$ and $m > 0$ that have certain lower bounds similar to m_X.

Theorem 2. *If the noise variance σ^2 is known, then, for any sequence of design points $\mathscr{X}_n$,*

- $\widetilde{R}_X(h_x, h_y)$ *is asymptotically optimal under assumptions* (A1)–(A3); *i.e.,* (21) *holds for* $(\tilde{h}_x, \tilde{h}_y) = \operatorname{arginf}_{(h_x,h_y)\in\mathscr{A}_n^X} \widetilde{R}_X(h_x, h_y)$,

- $\widetilde{R}_Y(h_y, \alpha)$ *is asymptotically optimal under assumptions* (A1*)–(A3*); *i.e.,* (22) *holds for* $(\tilde{h}_y, \tilde{\alpha}) = \operatorname{arginf}_{(h_y,\alpha)\in\mathscr{A}_n^Y} \widetilde{R}_Y(h_y, \alpha)$, *and*

- $\widetilde{R}(h_x, h_y, \alpha)$ *is asymptotically optimal under assumptions* (A1†)–(A3†); *i.e.,* (23) *holds for* $(\tilde{h}_x, \tilde{h}_y, \tilde{\alpha}) = \operatorname{arginf}_{(h_x,h_y,\alpha)\in\mathscr{A}_n} \widetilde{R}(h_x, h_y, \alpha)$.

If the noise variance is unknown and is replaced by an independent $\sqrt{n}$-consistent estimate $\hat{\sigma}^2$, then the conclusions in Theorem 2 also hold with the following additional assumptions:

(A4) $\sup_{(h_x,h_y)\in\mathscr{A}_n^X} n^{-3/2} R_X^{-1}(h_x, h_y) \operatorname{tr}(\boldsymbol{M}_X) \longrightarrow 0$;

(A4*) $\sup_{(h_y,\alpha)\in\mathscr{A}_n^Y} n^{-3/2} R_Y^{-1}(h_y, \alpha) \operatorname{tr}(\boldsymbol{M}_Y) \longrightarrow 0$;

(A4†) $\sup_{(h_x,h_y,\alpha)\in\mathscr{A}_n} n^{-3/2} R^{-1}(h_x, h_y, \alpha) \operatorname{tr}(\boldsymbol{M}) \longrightarrow 0$.

Corollary 1. *If the noise variance σ^2 is unknown and is replaced by an independent $\sqrt{n}$-consistent estimate $\hat{\sigma}^2$, then the asymptotic optimality of $\widetilde{R}_X(h_x, h_y)$, $\widetilde{R}_Y(h_y, \alpha)$ and $\widetilde{R}(h_x, h_y, \alpha)$ still holds for any sequence of design points $\mathscr{X}_n$ under the assumption sets* (A1)–(A4), (A1*)–(A4*) *and* (A1†)–(A4†), *respectively.*

5. Numerical performance

Two sets of numerical experiments were conducted to evaluate the empirical properties of the proposed methodology.

5.1. The settings of Choi et al. (2000)

In this first set of experiments we compare the relative performances amongst the unsharpened estimator $\hat{g}(\cdot)$ defined by (2) and all other sharpened estimators: the sharpened X-only estimator $\tilde{g}^X(\cdot)$ defined by (5), the sharpened Y-only estimator $\tilde{g}^Y(\cdot)$ defined by (7), and the sharpened X-and-Y estimator $\tilde{g}(\cdot)$ defined by (8). The bandwidth h for the unsharpened estimator $\hat{g}(\cdot)$ was automatically chosen as the minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{g}(X_i)\}^2 + \frac{\sigma^2}{n}\sum_{i=1}^{n}\{2W_i(X_i; h) - 1\},$$

Table 1
MSE values for each combination of (q, λ) together with standard errors in parentheses obtained from the numerical experiments conducted in Section 5.1.

(q, λ)	Unsharpened	X-only	Y-only	X-and-Y
(1,0.1)	10.84 (0.41)	10.41 (0.41)	9.32 (0.38)	9.20 (0.38)
(1,0.5)	15.67 (0.58)	14.81 (0.57)	13.41 (0.50)	13.15 (0.51)
(1,0.9)	23.52 (1.03)	22.55 (1.04)	23.05 (1.07)	22.00 (1.07)
(2,0.1)	9.98 (0.29)	9.69 (0.28)	8.55 (0.26)	8.51 (0.27)
(2,0.5)	10.25 (0.28)	9.92 (0.29)	8.88 (0.24)	8.75 (0.26)
(2,0.9)	12.66 (0.41)	12.29 (0.41)	11.50 (0.40)	11.34 (0.40)
(4,0.1)	9.15 (0.27)	8.88 (0.26)	7.85 (0.23)	7.76 (0.22)
(4,0.5)	8.69 (0.24)	8.47 (0.24)	7.57 (0.20)	7.54 (0.20)
(4,0.9)	10.37 (0.28)	10.67 (0.31)	9.70 (0.31)	9.75 (0.31)

For clarity, these values are multiplied by 100. Recall that (q, λ) are parameters of the mixture density that generated the design points x: $\lambda B(4, q) + (1 - \lambda)U[0, 1]$, where $B(\cdot, \cdot)$ and $U[0, 1]$ are the beta and the standard uniform distributions, respectively.

which is an unbiased estimator of the risk function

$$\frac{1}{n}E\left[\sum_{i=1}^{n}\{g_i - \hat{g}(X_i)\}^2 | \mathscr{X}_n\right]$$

(e.g., see Rice, 1984). We are aware that, in terms of empirical performance, there are more reliable kernel-based smoothing methods other than this unsharpened estimator $\hat{g}(\cdot)$. However, as $\hat{g}(\cdot)$ can be treated as an unsharpened version of the three sharpened estimators proposed above, we adopted it here for the purpose of evaluating the improvement, if any, that could be gained by using data sharpening.

For the three data-sharpening estimators, their tuning parameters were chosen, respectively, as the joint minimizer of $\widetilde{R}_X(h_x, h_y)$, $\widetilde{R}_Y(h_y, \alpha)$ and $\widetilde{R}(h_x, h_y, \alpha)$. In all cases the noise variance σ^2 was assumed unknown and was estimated by the fourth order estimator proposed by Hall et al. (1990). Although this variance estimator may suffer from finite-sample bias problems (Seifert and Gasser, 1993), it performed satisfactory in our numerical work.

We follow Choi et al. (2000) and adopted the following experimental settings. The target regression function was

$$g(x) = 2 - 5x + 2.5k \exp\{-200k(x - 0.5)^2\},$$

which is a linear trend with a Gaussian peak centered at $x = 0.5$. We used $k = 2$. The design density for x was a mixture density of a beta distribution $B(p, q)$ with parameters p and q and the uniform distribution $U[0, 1]$ on $[0, 1]$: $\lambda B(4, q) + (1 - \lambda)U[0, 1]$, where $0 \leq \lambda \leq 1$ is the mixing probability. Three values of q and three values of λ were used: $q = 1, 2$ and 4, and $\lambda = 0.1, 0.5$ and 0.9. The sample size was $n = 100$ and the noise was zero-mean Gaussian with $\sigma^2 = 0.5^2$. To reduce the edge effects, we followed Choi et al. (2000) and placed additional designed points distributed uniformly in $[-0.5, 0] \cup [1, 1.5]$. The following kernel function was used: $K(x) = \frac{3}{4}(1 - x^2), x \in [0, 1]$. It is the optimal kernel of order $(0, 2)$ derived in Gasser et al. (1985).

For each combination of (q, λ), 200 artificial data sets were generated. Then the unsharpened and the three data-sharpening methods were applied to each simulated data set to obtain estimates for $g(x)$. For the unsharpened estimate $\hat{g}(\cdot)$, the following was calculated as an approximation to the mean-squared-error (MSE) $\int_0^1 \{g(x) - \hat{g}(x)\}^2 dx$:

$$\text{MSE}(\hat{g}) = \frac{1}{201}\sum_{i=0}^{200}\left\{g\left(\frac{1}{200}\right) - \hat{g}\left(\frac{1}{200}\right)\right\}^2.$$

Similar MSE values were also calculated for the three sharpening estimators $\tilde{g}^X(\cdot)$, $\tilde{g}^Y(\cdot)$ and $\tilde{g}(\cdot)$. The means and standard errors of these calculated MSE values are tabulated in Table 1.

Paired t-tests were applied to test if the difference between the averaged MSE values of any two estimation methods is significant or not. The significance level used was $\frac{5}{4}\% = 1.25\%$. Based on the test results the four estimation methods were ranked in the following manner. If the averaged MSE value of a method is significantly less than the remaining three, it will be assigned a rank 1. If the averaged MSE value of a method is significantly larger than one but less than two methods, it will be assigned a rank 2, and similarly for ranks 3 and 4. Methods having non-significantly different averaged MSE values will share the same averaged rank. The resulting rankings are tabulated in Table 2.

From Tables 1 and 2, one could see that, in terms of statistical significance, $\tilde{g}^X(\cdot)$ never gave a worse averaged MSE value than $\hat{g}(\cdot)$, $\tilde{g}^Y(\cdot)$ never gave a worse averaged MSE value than $\tilde{g}^X(\cdot)$, and $\tilde{g}(\cdot)$ never gave a worse averaged MSE value than $\hat{g}^Y(\cdot)$. Also, the overall averaged t-test rankings for $\hat{g}(\cdot)$, $\tilde{g}^X(\cdot)$, $\tilde{g}^Y(\cdot)$ and $\tilde{g}(\cdot)$ are, respectively, 3.72, 3.06, 1.72 and 1.50. Judging from these results it seems that the X-and-Y data-sharpening estimator $\tilde{g}(\cdot)$ is the most preferable estimator. However, as $\tilde{g}(\cdot)$ requires the selection of three tuning parameters (h_x, h_y, α) (i.e., need to solve a three-dimensional minimization problem), the Y-only data-sharpening estimator $\tilde{g}^Y(\cdot)$ is a viable alternative. It is because $\tilde{g}^Y(\cdot)$ only requires the choosing of two parameters (h_y, α) and it gave very comparable performance relative to $\tilde{g}(\cdot)$.

Table 2
Pairwise *t*-test rankings for the unsharpened and sharpening estimates for the numerical experiments conducted in Section 5.1.

(q, λ)	Unsharpened	X-only	Y-only	X-and-Y
(1, 0.1)	4	3	1.5	1.5
(1, 0.5)	4	3	1.5	1.5
(1, 0.9)	2.5	2.5	2.5	2.5
(2, 0.1)	4	3	1.5	1.5
(2, 0.5)	4	3	2	1
(2, 0.9)	4	3	1.5	1.5
(4, 0.1)	4	3	2	1
(4, 0.5)	3.5	3.5	1.5	1.5
(4, 0.9)	3.5	3.5	1.5	1.5
Averaged	3.72	3.06	1.72	1.50

Table 3
Medians of the relative absolute distances between various ideal and selected parameters for the numerical experiments conducted in Section 5.1.

(q, λ)	Unsharpened h	X-only		Y-only		X-and-Y		
		h_x	h_y	h_y	α	h_x	h_y	α
(1,0.1)	0.31	0.50	0.25	0.29	0.00	0.53	0.29	0.00
(1,0.5)	0.38	0.59	0.33	0.33	0.00	0.50	0.32	0.11
(1,0.9)	0.36	0.60	0.33	0.33	0.20	0.50	0.29	0.25
(2,0.1)	0.30	0.60	0.25	0.29	0.00	0.67	0.25	0.00
(2,0.5)	0.27	0.50	0.25	0.29	0.00	0.40	0.25	0.00
(2,0.9)	0.25	0.50	0.25	0.20	0.00	0.50	0.17	0.00
(4,0.1)	0.26	0.67	0.25	0.29	0.00	0.67	0.20	0.00
(4,0.5)	0.19	0.40	0.20	0.17	0.00	0.50	0.17	0.00
(4,0.9)	0.15	0.59	0.00	0.17	0.00	0.50	0.17	0.00

The following has been done for assessing the qualities of the selected parameters. For each simulated data set and for the *Y*-only estimator, we obtained the corresponding ideal bandwidth h_y^{ideal} and ideal sharpening parameter α^{ideal} that jointly minimize MSE($\hat{g}$). These ideal values are of course not obtainable in practice. We then calculated the following absolute distances between these ideal values with those $(\hat{h}_y, \hat{\alpha})$ that were automatically selected by the corresponding proposed selection method:

$$\left| \frac{\hat{h}_y - h_y^{\mathrm{ideal}}}{\hat{h}_y} \right| \quad \text{and} \quad \left| \frac{\hat{\alpha} - \alpha^{\mathrm{ideal}}}{\hat{\alpha}} \right|.$$

Finally, for each combination of (q, λ), the median values of the above two distances were calculated. Similar median values for the unsharpened, *X*-only, and *X*-and-*Y* estimators were also computed. The reason that the medians instead of the means were calculated was that, for some data sets the values of the selected or the ideal values of (h_y, α) were very close to zero. This led to numerical instability when the above distances were being averaged, and the use of median bypasses this issue.

The above median values are tabulated in Table 3, and from which one could see that the proposed selection methods performed very well for selecting the sharpening parameter α, and reasonably well for the *Y*-direction bandwidth h_y.

For each simulated data set we have also calculated the MSE value when the ideal parameter values were used in place of those selected values. Denote this "ideal MSE" value as $\mathrm{MSE}^{\mathrm{ideal}}$ and the one from the selected parameters as $\widehat{\mathrm{MSE}}$. With this we computed the following relative reduction of MSE value: $(\widehat{\mathrm{MSE}} - \mathrm{MSE}^{\mathrm{ideal}})/\widehat{\mathrm{MSE}}$. The averaged MSE reductions for all combinations of (q, λ) are listed in Table 4. Loosely, about 10–30% further reduction in MSEs is possible if the ideal parameter values were known.

To visually evaluate the qualities of various estimated curves, we randomly selected a data set generated with $q = 4$ and $\lambda = 0.5$, and computed the corresponding unsharpened curve estimate and the three different versions of sharpened estimates. This data set, the true regression function, and all the curve estimates are displayed in Fig. 1. To further facilitate the visual inspection of the behaviors of various sharpening strategies around the central peak of the true regression function, zoomed-in plots for $X \in [0.44, 0.56]$ are given in Fig. 2. From these plots one could see that, for both the *Y*-only and *X*-and-*Y* estimators, the resulting estimated curves were pulled closer towards the true regression function.
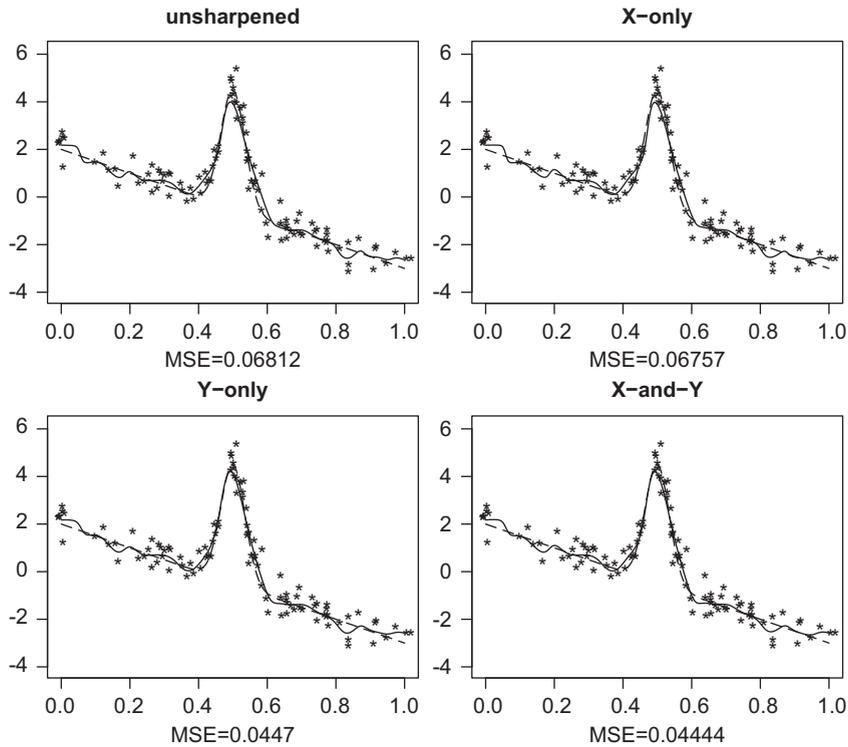
Fig. 1. Plots of the observations (asterisks), true regression function (broken lines) and various estimated curves (solid lines) for the numerical experiments conducted in Section 5.1. The MSE value of each estimated curve is listed at the bottom of each panel.
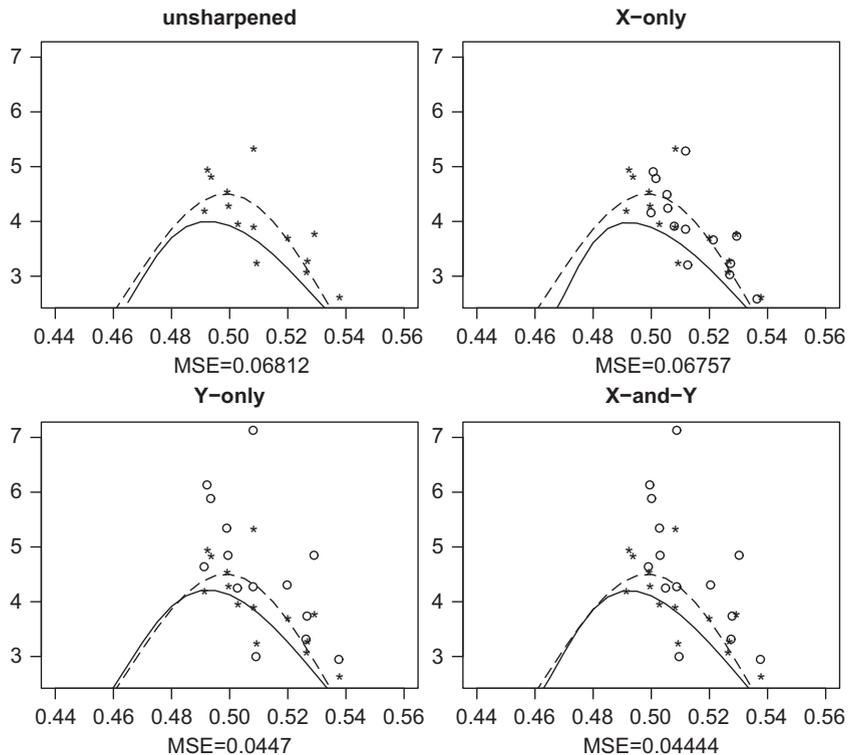

Fig. 2. Zoomed-in plots of Fig. 1 for $X \in [0.44, 0.56]$. In each panel the true regression function and the estimated curve are represented as broken and solid lines, respectively. Raw observations and sharpened data are denoted as asterisks and circles, respectively. One could see that, for both the *Y*-only and *X*-and-*Y* estimators, the resulting estimated curves were pulled closer towards the true regression function.

Table 4
MSE reduction values (in %), together with standard errors in parentheses, if the ideal parameter values were used instead of those selected by the proposed automatic methods for the numerical experiments conducted in Section 5.1.

(q, λ)	Unsharpened	X-only	Y-only	X-and-Y
(1,0.1)	21.6 (1.06)	23.1 (0.97)	23.1 (1.07)	25.9 (1.03)
(1,0.5)	29.3 (1.11)	31.4 (1.09)	31.3 (1.16)	35.0 (1.14)
(1,0.9)	20.1 (1.14)	23.6 (1.11)	26.8 (1.24)	30.0 (1.22)
(2,0.1)	19.9 (1.02)	22.2 (0.98)	20.8 (1.03)	24.6 (1.06)
(2,0.5)	17.0 (1.04)	19.4 (0.91)	18.8 (0.98)	21.2 (0.95)
(2,0.9)	10.9 (0.83)	14.1 (0.79)	12.4 (0.88)	17.1 (0.91)
(4,0.1)	17.3 (0.98)	20.0 (0.96)	19.1 (1.02)	22.6 (0.99)
(4,0.5)	8.9 (0.72)	11.4 (0.70)	9.9 (0.71)	13.9 (0.75)
(4,0.9)	3.5 (0.37)	11.4 (0.72)	6.2 (0.56)	12.6 (0.81)

Table 5
MSE values for each combination of (q, λ) together with standard errors in parentheses obtained from the numerical experiments conducted in Section 5.2.

(q, λ)	Test function 1		Test function 2	
	Unsharpened	Y-only	Unsharpened	Y-only
(1,0.1)	10.56 (0.34)	7.22 (0.30)	2.60 (0.08)	2.19 (0.07)
(1,0.5)	6.61 (0.41)	3.05 (0.14)	3.34 (0.11)	3.09 (0.11)
(1,0.9)	36.56 (3.17)	16.61 (1.51)	7.00 (0.34)	6.49 (0.36)
(2,0.1)	10.75 (0.35)	7.73 (0.33)	2.56 (0.08)	2.21 (0.07)
(2,0.5)	5.75 (0.42)	2.65 (0.14)	2.62 (0.08)	2.22 (0.07)
(2,0.9)	25.65 (1.48)	10.55 (0.67)	4.27 (0.16)	3.45 (0.14)
(4,0.1)	10.85 (0.31)	7.73 (0.31)	2.45 (0.07)	2.06 (0.06)
(4,0.5)	5.47 (0.26)	2.43 (0.10)	2.41 (0.06)	2.06 (0.06)
(4,0.9)	22.24 (1.33)	8.80 (0.60)	3.50 (0.11)	2.90 (0.11)

For clarity, these values are multiplied by 100.

5.2. Two additional test functions

In order to demonstrate the usefulness of the sharpening methodology, the above-recommended Y-only estimator was further tested with two additional test functions.

Test function 1: $g(x) = 1 - 48x^2 + 218x^2 - 315x^3 + 145x^4$
Test function 2: $g(x) = \sin(8x - 4) + 2\exp\{-16(4x - 2)^2\}$.

These two functions have been used by previous authors (e.g., Ruppert et al., 1995; Fan and Gijbels, 1996).

The design points were generated in the same manner as before, and the same nine combinations of (q, λ) were tested. The noise standard deviation σ was set to the value that the resulting signal-to-noise ratio $\|g\|/\sigma$ was 3. For each generated data set, both the unsharpened and the Y-only estimators were applied to estimate the unknown test function, and the corresponding MSE values were computed. The averages and standard errors of the MSE values are provided in Table 5. These values show that the Y-only estimator never performed worse than the unsharpened estimator. In some situations the averaged MSE values of the Y-only estimator are even less than 50% of those from the unsharpened estimator.

5.3. Local linear regression

When comparing to the classical Nadaraya–Watson estimator, it is known that the local linear regression estimator has superior bias properties (e.g., Fan and Gijbels, 1996). Therefore it is interesting to compare data sharpening with local linear regression.

A new set of simulations was conducted using the same experimental setup as in Section 5.1. Three estimators were studied: the proposed Y-only estimator, local linear regression (without sharpening), and local linear regression with the Y-only sharpening strategy. The bandwidths for the later two estimators were chosen using Stein's unbiased risk estimation approach. Two major empirical conclusions can be drawn from the numerical results. First, the unsharpened local linear regression estimator gave inferior performance to both Y-only sharpening estimators. Second, the Y-only local linear regression did not outperform the Y-only Nadaraya–Watson estimator. This seems to suggest that the comparative bias advantage of local linear regression disappears when data sharpening is employed. For brevity, these simulation results are omitted but can be obtained from the authors.

6. Conclusion

In this paper we have modified the data-sharpening technique for nonparametric regression in two directions. First, through the introduction of the sharpening parameter α, we allow the data to be sharpened to different extents. Second, using the idea of unbiased risk estimation, we have developed methods for simultaneously choosing the values of all the tuning parameters required for the computations of the sharpened estimators. We have also shown that these parameter selection methods are asymptotically optimal in a well defined sense. In addition we have conducted numerical experiments to demonstrate the superiority of the sharpened estimators when comparing to their unsharpened counterparts. From the simulation results we recommend the estimator that sharpens only the response variable as the "best" estimator, in the sense that it provides a good compromise between statistical performance and computational speed.

Acknowledgments

Appendix A. Proofs

Proof of Theorem 1. We first show the unbiasedness of $\widehat{R}_X(h_x, h_y)$. Note that

$$E\{\mathrm{RSS}_X(h_x, h_y)|\mathscr{X}_n\} = E\left\{\sum_i (Y_i - \tilde{g}_i^X)^2 |\mathscr{X}_n\right\} = \sum_i E\{Y_i^2 - 2Y_i\tilde{g}_i^X + (\tilde{g}_i^X)^2 |\mathscr{X}_n\},$$

$E(Y_i|\mathscr{X}_n) = g_i$ and $E(Y_i^2|\mathscr{X}_n) = g_i^2 + \sigma^2$, where $g_i = g(X_i)$ and $\boldsymbol{g} = (g_1, \ldots, g_n)^T$. For brevity, in the sequel we suppress the dependence of expectations on the design points $\mathscr{X}_n$. We also use the notation "$\sum_i$" for "$\sum_{i=1}^n$" unless specified otherwise. From (5) we have

$$\begin{aligned}
E(Y_i\tilde{g}_i^X) &= E\left\{Y_i \sum_j \widetilde{W}_i(X_j; h_y)Y_j\right\} \\
&= \widetilde{W}_i(X_i; h_y)(g_i^2 + \sigma^2) + \sum_{j \neq i} \widetilde{W}_j(X_i; h_y)g_i g_j \\
&= \sum_j \widetilde{W}_j(X_i; h_y)g_i g_j + \widetilde{W}_i(X_i; h_y)\sigma^2 \\
&= g_i E(\tilde{g}_i^X) + \widetilde{W}_i(X_i; h_y)\sigma^2.
\end{aligned}$$

Then we have

$$\begin{aligned}
E(Y_i - \tilde{g}_i^X)^2 &= g_i^2 + \sigma^2 - 2\{g_i E(\tilde{g}_i^X) - \widetilde{W}_i(X_i; h_y)\sigma^2\} + E\{(\tilde{g}_i^X)^2\} \\
&= E\{(g_i - \tilde{g}_i^X)^2\} - \{2\widetilde{W}_i(X_i; h_y) - 1\}\sigma^2
\end{aligned}$$

and

$$E\{\mathrm{RSS}_X(h_x, h_y)\} = nR_x(h_x, h_y) - \sigma^2 \sum_i \{2\widetilde{W}_i(X_i; h_y) - 1\},$$

which leads to the unbiasedness of $\widehat{R}_X(h_x, hy)$.

We next show the unbiasedness of $R_Y(h_x, hy)$. From (2) and (7), one has $\widetilde{Y}_i = (1 + \alpha)Y_i - \alpha\sum_j W_j(X_i; h_y)Y_j$ and

$$\tilde{g}_i^Y = (1 + \alpha) \sum_j W_j(X_i; h_y)Y_j - \alpha \sum_{j,k} W_j(X_i; h_y)W_k(X_j; h_y)Y_k.$$

For convenience, write $W_{ij} = W_j(X_i; h_y)$. Then we have

$$E(Y_i \tilde{g}_i^Y) = E\left\{(1+\alpha)(g_i + \varepsilon_i)\sum_j W_{ij}(g_j + \varepsilon_j)\right\} - \alpha E\left\{(g_i + \varepsilon_i)\sum_{j,k} W_{ij}W_{jk}(g_k + \varepsilon_k)\right\}$$

$$= (1+\alpha)\left(g_i \sum_j W_{ij}g_j + \sigma^2 W_{ii}\right) - \alpha\left\{\sum_j W_{ij}W_{ji}(g_i^2 + \sigma^2) + g_i \sum_j \sum_{k \neq i} W_{ij}W_{jk}g_k\right\}$$

$$= \left\{(1+\alpha)W_{ii} - \alpha \sum_j W_{ij}W_{ji}\right\}\sigma^2 + g_i\left\{(1+\alpha)\sum_j W_{ij}g_j - \alpha \sum_{j,k} W_{ij}W_{jk}g_k\right\}$$

$$= \left\{(1+\alpha)W_{ii} - \alpha \sum_j W_{ij}W_{ji}\right\}\sigma^2 + g_i E(\tilde{g}_i^Y).$$

Therefore

$$E\{(Y_i - \tilde{g}_i^Y)^2\} = g_i^2 + \sigma^2 - 2\left[\left\{(1+\alpha)W_{ii} - \alpha \sum_j W_{ij}W_{ji}\right\}\sigma^2 + g_i E(\tilde{g}_i^Y)\right] + E\{(\tilde{g}_i^Y)^2\}$$

$$= E\{(g_i - \tilde{g}_i^Y)^2\} - \left[2\left\{(1+\alpha)W_{ii} - \alpha \sum_j W_{ij}W_{ji}\right\} - 1\right]\sigma^2$$

and

$$E\{\text{RSS}_Y(h_y, \alpha)\} = nR_y(h_y, \alpha) - \sigma^2 \sum_i \left[2\left\{(1+\alpha)W_{ii} - \alpha \sum_j W_{ij}W_{ji}\right\} - 1\right].$$

Thus $\widehat{R}_Y(h_y, \alpha)$ is an unbiased estimator of $R_y(h_y, \alpha)$. By analogy to $\widehat{R}_Y(h_y, \alpha)$, one can easily show the unbiasedness of $\widehat{R}(h_x, h_y, \alpha)$.   □

Proof of Theorem 2. In here we establish the asymptotic optimality of $\widetilde{R}_y$, while the proofs for the other two versions are essentially the same. First note that the smoother matrix for computing $\tilde{\boldsymbol{g}}^Y$ is $\boldsymbol{M}_Y = (1+\alpha)\boldsymbol{W} - \alpha\boldsymbol{W}^2$, where $\boldsymbol{W} = (W_{ij})_{1 \leq i,j \leq n}$, which implies that the diagonal elements of $\boldsymbol{M}_Y = (m_{ij})_{1 \leq i,j \leq n}$ is $m_{ii} = (1+\alpha)W_{ii} - \alpha\sum_j W_{ij}W_{ji}$. Then one can express $\widetilde{R}_Y(h_y, \alpha) = (1/n)(\text{RSS}_Y + 2\sum_i m_{ii}\sigma^2)$. Write $\boldsymbol{\epsilon} = (\varepsilon_i, \ldots, \varepsilon_n)^T$ and $\boldsymbol{A} = \boldsymbol{I} - \boldsymbol{M}_Y$, where $\boldsymbol{I}$ is the $n \times n$ identity matrix. We have

$$n\widetilde{R}_Y(h_y, \alpha) = \|\boldsymbol{Y} - \boldsymbol{g} + \boldsymbol{g} - \tilde{\boldsymbol{g}}^Y\|^2 + 2\sum_i m_{ii}\sigma^2$$

$$= \|\boldsymbol{\epsilon}\|^2 + nL_Y(h_y, \alpha) + 2\langle\boldsymbol{\epsilon}, \boldsymbol{g} - \tilde{\boldsymbol{g}}^Y\rangle + 2\sum_i m_{ii}\sigma^2$$

$$= \|\boldsymbol{\epsilon}\|^2 + nL_Y(h_y, \alpha) + 2\langle\boldsymbol{\epsilon}, \boldsymbol{A}\boldsymbol{g}\rangle + 2\left(\sum_i m_{ii}\sigma^2 - \langle\boldsymbol{\epsilon}, \boldsymbol{M}_Y\boldsymbol{\epsilon}\rangle\right). \tag{24}$$

Since $\|\boldsymbol{\epsilon}\|^2$ does not depend on (h_y, α), in order to prove (22), it is sufficient to show that, for any sequence of design points $\mathscr{X}_n$,

$$\sup_{\mathscr{A}_n^Y} |\langle\boldsymbol{\epsilon}, \boldsymbol{A}\boldsymbol{g}\rangle|/\{nR_Y(h_y, \alpha)\} \xrightarrow{P} 0, \tag{25}$$

$$\sup_{\mathscr{A}_n^Y} |\sum_i m_{ii}\sigma^2 - \langle\boldsymbol{\epsilon}, \boldsymbol{M}_Y\boldsymbol{\epsilon}\rangle|/\{nR_Y(h_y, \alpha)\} \xrightarrow{P} 0 \tag{26}$$

and

$$\sup_{\mathscr{A}_n^Y} |L_Y(h_y, \alpha)/R_Y(h_y, \alpha) - 1| \xrightarrow{P} 0. \tag{27}$$

To show (25), we apply Chebyshev's inequality: for any $\delta > 0$, one has

$$P\left\{\sup_{\mathscr{A}_n^Y} \frac{|\langle \boldsymbol{\epsilon}, \boldsymbol{A}\boldsymbol{g}\rangle|}{nR_Y(h_y, \alpha)} > \delta\right\} \leq \frac{1}{\delta^{2m_Y}} \sum_{\mathscr{A}_n^Y} \frac{E(\langle \boldsymbol{\epsilon}, \boldsymbol{A}\boldsymbol{g}\rangle^{2m_Y})}{n^{2m_Y} R_Y(h_y, \alpha)^{2m_Y}}$$

$$\leq \frac{C}{\delta^{2m_Y}} \sum_{\mathscr{A}_n^Y} \frac{\|\boldsymbol{A}\boldsymbol{g}\|^{2m_Y}}{n^{2m_Y} R_Y^{2m_Y}(h_y, \alpha)}, \tag{28}$$

for some constant $C > 0$, by observing (A3) and applying Theorem 2 of Whittle (1960). Since $nR_Y(h_y, \alpha) = \|\boldsymbol{A}\boldsymbol{g}\|^2 + E(\|\boldsymbol{M}_Y\boldsymbol{\epsilon}\|^2) \geq \|\boldsymbol{A}\boldsymbol{g}\|^2$ and (A2), the right-hand side of (28) is bounded by $C\delta^{-2m_Y}\sum_{\mathscr{A}_n^Y}\{nR_Y(h_y, \alpha)\}^{-m_Y} \to 0$. Thus (25) is proved.

Eq. (26) can be shown analogously by observing

$$E\langle \boldsymbol{\epsilon}, \boldsymbol{M}_Y\boldsymbol{\epsilon}\rangle = E\left(\sum_i m_{ii}\varepsilon_i^2\right) = \sigma^2 \sum_i m_{ii}$$

and

$$nR_Y(h_y, \alpha) \geq E(\|\boldsymbol{M}_Y\boldsymbol{\epsilon}\|^2) = E\left(\sum_i m_{ii}\varepsilon_i^2\right) = \sigma^2 \sum_i m_{ii} \geq \sigma^2 \mathrm{tr}(\boldsymbol{M}_Y^T\boldsymbol{M}_Y).$$

Then (26) follows by (A2) and (A3) and again Theorem 2 of Whittle (1960),

$$P\left\{\sup_{\mathscr{A}_n^Y} \frac{|\sum_i m_{ii}\sigma^2 - \langle \boldsymbol{\epsilon}, \boldsymbol{M}_Y\boldsymbol{\epsilon}\rangle|}{nR_Y(h_y, \alpha)} > \delta\right\} \leq \frac{1}{\delta^{2m_Y}} \sum_{\mathscr{A}_n^Y} \frac{E\{(\langle \boldsymbol{\epsilon}, \boldsymbol{M}_Y\boldsymbol{\epsilon}\rangle - E\langle \boldsymbol{\epsilon}, \boldsymbol{M}_Y\boldsymbol{\epsilon}\rangle)^{2m_Y}\}}{n^{2m_Y} R^{2m_Y}(h_y, \alpha)}$$

$$\leq \frac{C'}{\delta^{2m_Y}} \sum_{\mathscr{A}_n^Y} \frac{\mathrm{tr}(\boldsymbol{M}_Y^T\boldsymbol{M}_Y)}{n^{2m_Y} R_Y^{2m_Y}(h_y, \alpha)}$$

$$\leq \frac{C'}{\delta^{2m_Y}\sigma^2} \sum_{\mathscr{A}_n^Y} \frac{1}{n^{m_Y} R_Y^{m_Y}(h_y, \alpha)} \to 0.$$

To show (27), one notes

$$L_Y(h_y, \alpha) - R_Y(h_y, \alpha) = \frac{1}{n}\{\|\boldsymbol{M}_Y\boldsymbol{\epsilon}\|^2 - E(\|\boldsymbol{M}_Y\boldsymbol{\epsilon}\|^2) - 2\langle \boldsymbol{A}\boldsymbol{g}, \boldsymbol{M}_Y\boldsymbol{\epsilon}\rangle\}.$$

Then it is sufficient to show

$$\sup_{\mathscr{A}_n^Y} |\langle \boldsymbol{A}\boldsymbol{g}, \boldsymbol{M}_Y\boldsymbol{\epsilon}\rangle|/\{nR_Y(h_y, \alpha)\} \xrightarrow{P} 0 \tag{29}$$

and

$$\sup_{\mathscr{A}_n^Y} |\|\boldsymbol{M}_Y\boldsymbol{\epsilon}\|^2 - E(\|\boldsymbol{M}_Y\boldsymbol{\epsilon}\|^2)|/\{nR_Y(h_y, \alpha)\} \xrightarrow{P} 0, \tag{30}$$

which is similar to the proofs of (25) and (26). Observing

$$\langle \boldsymbol{A}\boldsymbol{g}, \boldsymbol{M}_Y\boldsymbol{\epsilon}\rangle = \langle \boldsymbol{\epsilon}, \boldsymbol{M}_Y^T A\boldsymbol{g}\rangle \quad \text{and} \quad \|\boldsymbol{M}_Y^T A\boldsymbol{g}\|^2 \leq \lambda_{\boldsymbol{M}_Y}\|A\boldsymbol{g}\|^2,$$

that lead to (29), while

$$\|\boldsymbol{M}_Y\boldsymbol{\epsilon}\|^2 = \langle \boldsymbol{\epsilon}, \boldsymbol{M}_Y^T\boldsymbol{M}_Y\boldsymbol{\epsilon}\rangle \quad \text{and} \quad \mathrm{tr}(\boldsymbol{M}_Y^T\boldsymbol{M}_Y\boldsymbol{M}_Y^T\boldsymbol{M}_Y) \leq \lambda_{\boldsymbol{M}_Y}^2 \mathrm{tr}(\boldsymbol{M}_Y^T\boldsymbol{M}_Y),$$

complete the proof of (30). $\quad\square$

Proof of Corollary 1. The identity (24) holds with σ^2 replaced by $\hat{\sigma}^2$, and (25) and (27) are still valid. We only need to verify (26) with σ^2 substituted by $\hat{\sigma}^2$. Since the estimate $\hat{\sigma}^2$ does not depend on the tuning parameters h_y and α, it is sufficient to show

$$|\hat{\sigma}^2 - \sigma^2| \sup_{\mathscr{A}_n^Y} \sum_i m_{ii}/\{nR_Y(h_y, \alpha)\} \xrightarrow{P} 0.$$

It is obvious that the assumption (A4*) leads to the asymptotic optimality of $\widetilde{R}_Y(h_y, \alpha)$, given that the estimate $\hat{\sigma}^2$ is $\sqrt{n}$-consistent. $\quad\square$

References

Braun, W.J., Hall, P., 2001. Data sharpening for nonparametric inference subject to constraints. Journal of Computational and Graphical Statistics 10, 786–806.

Buckley, M.J., Eagleson, G.K., Silverman, B.W., 1988. The estimation of residual variance in nonparametric regression. Biometrika 75, 189–199.

Choi, E., Hall, P., 1999. Data sharpening as a prelude to density estimation. Biometrika 86, 941–947.

Choi, E., Hall, P., Rousson, V., 2000. Data sharpening methods for bias reduction in nonparametric regression. The Annals of Statistics 28, 1339–1355.

Claeskens, G., Hall, P., 2002. Data sharpening for hazard rate estimation. Australian & New Zealand Journal of Statistics 44, 277–283.

Cohen, A., 1966. All admissible linear estimators of the mean vector. Annals of Mathematical Statistics 37, 458–463.

Craven, P., Wahba, G., 1979. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Numerical Mathematics 31, 377–403.

Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and its Applications. Chapman & Hall, London.

Gasser, T., Müller, H.-G., Mammitzsch, V., 1985. Kernels for nonparametric curve estimation. Journal of the Royal Statistical Society Series B 47, 238–252.

Hall, P., Kang, K.-H., 2005. Unimodal kernel density estimation by data sharpening. Statistica Sinica 15, 73–98.

Hall, P., Kay, J.W., Titterington, D.M., 1990. Asymptotically optimal difference-based estimation of variance in nonparametric regression. Biometrika 77, 521–528.

Hall, P., Marron, J.S., 1990. On variance estimation in nonparametric regression. Biometrika 77, 415–419.

Hall, P., Minnotte, M.C., 2002. High order data sharpening for density estimation. Journal of the Royal Statistical Society Series B 64, 141–157.

Li, K.C., 1986. Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. The Annals of Statistics 14, 1101–1112.

Li, K.C., 1987. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. The Annals of Statistics 15, 958–975.

Rice, J.A., 1984. Bandwidth choice for nonparametric regression. The Annals of Statistics 12, 1215–1230.

Ruppert, D., Sheather, S.J., Wand, M.P., 1995. An effective bandwidth selector for local least squares regression. Journal of the American Statistical Association 90, 1257–1270.

Samiuddin, M., el Sayyad, G.M., 1990. On nonparametric kernel density estimates. Biometrika 77, 865–874.

Seifert, B., Gasser, T., 1993. Nonparametric estimation of residual variance revisited. Biometrika 80, 373–383.

Stein, C.M., 1981. Estimation of the mean of a multivariate normal distribution. The Annals of Statistics 9, 1135–1151.

Whittle, P., 1960. Bounds for the moments of linear and quadratic forms in independent variables. Theory of Probability and its Applications 5, 302–305.

Yao, F., Lee, T.C.M., 2007. Spectral density estimation using sharpened periodograms. IEEE Transactions on Signal Processing 55, 4711–4716.