Improved Estimators in Nonparametric Regression Problems

Linda H. ZHAO

Linear estimators of multivariate means are considered. Generalizations of some well-known theorems about admissibility of linear estimators are given. The results then are applied to show that commonly used kernel-type estimators in nonparametric regression problems can be constructively improved in a simple way. An asymptotic result is described that gives a quantitative measure of the maximum improvement to be gained in certain situations. A theoretical bound shows that gains are achievable in the relative risk of up to 58.6% (rectangular kernel) or 29.2% (Epanechnikov kernel). Some examples of smaller sample size are also investigated, and these show relative risk gains ranging up to 18% in realistic settings.

KEY WORDS: Admissibility; Kernel estimators; Nonparametric regression.

1. INTRODUCTION

Consider the ordinary multivariate means problems. Observe $y_1, \ldots, y_n$, and assume that Y has mean θ and variance $\sigma^2\boldsymbol{\Sigma}$, where $Y = (y_1, \ldots, y_n)', \theta = (\theta_1, \ldots, \theta_n)', \boldsymbol{\Sigma}$ is known, and σ^2 is unknown. Linear estimators $\hat{\theta} = \mathbf{M}Y$ of θ are investigated under the quadratic loss

$$L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)'\mathbf{D}(\hat{\theta} - \theta),$$

where $\mathbf{D}$ is a given symmetric positive definite matrix. Cohen (1966) showed that a linear estimator $\hat{\theta} = \mathbf{M}Y$ is admissible among all linear estimators if and only if $\mathbf{M}\boldsymbol{\Sigma}$ is symmetric and all the eigenvalues of $\mathbf{M}$ are in [0, 1] (see also Rao 1976). Given an inadmissible linear estimator we will constructively find a class of better linear estimators. Cohen (1966) gave a particular better estimator when $\mathbf{D} = \boldsymbol{\Sigma}^{-1}$.

Regression analysis is a general statistical tool that has been widely used in virtually every area of statistical applications. In particular, nonparametric regression techniques have been developing rapidly in last 20 years or so. Here observe $y_1, \ldots, y_n$, and assume that

$$y_i = \mathbf{m}(x_i) + \varepsilon_i, \qquad i = 1, \ldots, n$$

where x_i's are fixed points, ε_i's are independent $E(\varepsilon_i) = 0$, and $\text{var}(\varepsilon_i) = \sigma^2$; that is, $\mathbf{m}(x)$ is the mean function of Y given X. There is no preassumption on the functional form of $\mathbf{m}$. The goal is to estimate $\mathbf{m}(x)$.

Several nonparametric methods have been proposed for the above problem (see, e.g., Fan 1993; Härdle 1990; Hastie and Tibshirani 1990; Zhao 1993). Many of these are linear methods, in the sense that the estimator of $\mathbf{m}$ is given by

$$\hat{\mathbf{m}}(x) = \sum_{i=1}^{n} W_n(x; x_i)y_i, \tag{1}$$

where the weight functions $W_n(x; x_i)$ are independent of y. I show that such methods can often be improved by applying the construction mentioned earlier.

In Section 5 I examine the improvement obtainable in three data analytic nonparametric regression situations. The first involves the analysis of Canadian income data discussed by Chu and Marron (1991); the second, an analysis of geyser data treated by Hall and Turlach (1997). For these data, I investigate the improvement available by my method when applied to the interpolation estimator proposed by Hall and Turlach. Finally, I apply the method to some simulated data. Here I can precisely measure the relative reduction in overall quadratic risk. The reductions obtained in these and similar examples vary from about 18% down to about 5% at the desirable bandwidths. For very small bandwidths, corresponding to low bias–high variance estimators, the improvement is much less (e.g., only 1%). The interested reader can go directly to this section to see the nature of the examples and of the attainable improvement.

Section 2 presents the general theory of improved linear estimators. Section 3 shows how this theory can often be applied in nonparametric regression. Section 4 presents some asymptotic results as $n \to \infty$. In brief, the maximum asymptotic improvement on three common estimators involving the Epanechnikov kernel can be as great as 29%. These same three methods involving a rectangular kernel can be asymptotically improved as much as 58.6%.

2. A CLASS OF BETTER LINEAR ESTIMATORS IN MULTIVARIATE MEAN PROBLEMS

Observe $Y = (y_1, \ldots, y_n)'$ and assume that Y has mean θ and variance $\sigma^2\boldsymbol{\Sigma}$, with known nonsingular matrix $\boldsymbol{\Sigma}$. In many applications one makes the additional assumption that Y is normally distributed, but that assumption is not needed here. I want to estimate θ under the weighted quadratic risk function. For an estimator $\hat{\theta}$, the risk function is

$$R(\hat{\theta}, \theta) = E_\theta(\hat{\theta} - \theta)'\mathbf{D}(\hat{\theta} - \theta) \tag{2}$$

for the given positive definite matrix $\mathbf{D}$.

Linda H. Zhao is Assistant Professor, Department of Statistics, University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *lzhao@stat.wharton.upenn.edu*).

Journal of the American Statistical Association
March 1999, Vol. 94, No. 445, Theory and Methods

An estimator $\hat{\theta}_1$ is better than $\hat{\theta}_2$ if

$$\begin{aligned} R(\hat{\theta}_1, \theta) &\leq R(\hat{\theta}_2, \theta) \quad \forall\, \theta \\ &< \qquad\qquad\quad \text{for some } \theta. \end{aligned}$$

An estimator $\hat{\theta}$ is inadmissible if some other estimator is better. An estimator is admissible if it is not inadmissible.

I am particularly interested in linear estimators $\hat{\theta} = \mathbf{M}Y$ for some $\mathbf{M}$. Some simple calculations yield that for $\hat{\theta} = \mathbf{M}Y$,

$$R(\hat{\theta}, \theta) = \sigma^2 \operatorname{tr}(\mathbf{M}'\mathbf{D}\mathbf{M}) + \theta'(I - \mathbf{M})'\mathbf{D}(I - \mathbf{M})\theta. \quad (3)$$

I refer to $\sigma^2 \operatorname{tr}(\mathbf{M}'\mathbf{D}\mathbf{M})$ as $\operatorname{var}(\mathbf{M})$ and to $\theta'(\mathbf{M} - I)'\mathbf{D}(\mathbf{M} - I)\theta$ as $\operatorname{bias}^2(\mathbf{M}, \theta)$.

This general problem is structurally invariant under nonsingular transformations. Hence without loss of generality I assume for now that $\mathbf{\Sigma} = I$ and $\sigma^2 = 1$. Remark 2 and Theorem 3 at the end of this section provide statements that show the effect of arbitrary $\mathbf{\Sigma}$. (Transform Y to $Y^{(1)} = \mathbf{A}Y$. Then $Y^{(1)}$ has mean $\theta^{(1)} = \mathbf{A}\theta$, etc. Let $\hat{\theta}^{(1)} = \mathbf{A}\hat{\theta}$ and $\mathbf{D}^{(1)} = \mathbf{A}^{-1'}\mathbf{D}\mathbf{A}^{-1}$. Then $R(\hat{\theta}, \theta) = R_{\mathbf{D}^{(1)}}(\hat{\theta}^{(1)}, \theta^{(1)})$. Choose $\mathbf{A}$ so that $\mathbf{\Sigma}^{(1)} = I$.)

Cohen (1966) investigated the case where $\mathbf{D} = I$. If $\mathbf{M}$ is either not symmetric or is symmetric but has some eigenvalues outside of [0, 1] then he explicitly constructed a better linear estimator. In the following two theorems I generalize Cohen's formula for $\mathbf{D} = I$ and also provide formulas for $\mathbf{D} \neq I$.

Theorem 1. Let $\mathbf{\Sigma} = I$. Consider a linear estimator $\hat{\theta}$ such that

$$\hat{\theta} = \mathbf{M}Y$$

for some $\mathbf{M}$. If $\mathbf{M}$ is asymmetric or $\mathbf{M}$ is symmetric but maxeig $\mathbf{M} > 1$, then there exists $0 \leq \gamma_0 < 1$ and a family of linear estimators, denoted by $\tilde{\theta}_\gamma$, with a symmetric matrix $\mathbf{G}_\gamma, \gamma_0 \leq \gamma \leq 1$, such that

$$\tilde{\theta}_\gamma = \mathbf{G}_\gamma Y \quad (4)$$

satisfies

$$R(\tilde{\theta}, \theta) < R(\hat{\theta}, \theta) \quad \forall\, \theta. \quad (5)$$

The formula for γ_0 is given in (13) and (14), and

$$\mathbf{G}_\gamma = I - \gamma\mathbf{D}^{-1/2}[\mathbf{D}^{1/2}(I - \mathbf{M})'\mathbf{D}(I - \mathbf{M})\mathbf{D}^{1/2}]^{1/2}\mathbf{D}^{-1/2}. \quad (6)$$

Furthermore, if $\gamma = 1$, then

$$\operatorname{bias}^2(\mathbf{G}_1, \theta) = \operatorname{bias}^2(\mathbf{M}, \theta) \quad \forall\, \theta$$

$$\operatorname{var}(\mathbf{G}_1) < \operatorname{var}(\mathbf{M}); \quad (7)$$

if $\gamma = \gamma_0$, then

$$\operatorname{bias}^2(\mathbf{G}_{\gamma_0}, \theta) \leq \operatorname{bias}^2(\mathbf{M}, \theta) \quad \forall\, \theta$$

$$\operatorname{var}(\mathbf{G}_{\gamma_0}) \leq \operatorname{var}(\mathbf{M})$$

$$(\text{if } \gamma_0 > 0, \text{ then } \operatorname{var}(\mathbf{G}_{\gamma_0}) = \operatorname{var}(\mathbf{M})); \quad (8)$$

and if $\gamma_0 < \gamma < 1$, then

$$\operatorname{bias}^2(\mathbf{G}_\gamma, \theta) \leq \operatorname{bias}^2(\mathbf{M}, \theta) \quad \forall\, \theta$$

$$\operatorname{var}(\mathbf{G}_\gamma) < \operatorname{var}(\mathbf{M}). \quad (9)$$

In (8) and (9) there is strict inequality in the bias^2 statements whenever the right side is not 0.

The following standard lemma, also used by Cohen (1966), is needed.

Lemma 1. For any square matrix $\mathbf{A}$,

$$\operatorname{tr}(\mathbf{A}'\mathbf{A})^{1/2} \geq \operatorname{tr}(\mathbf{A}),$$

and strict inequality holds if $\mathbf{A}$ is asymmetric or if $\mathbf{A}$ is symmetric and mineig$(\mathbf{A}) < 0$.

Proof of Theorem 1.

I first prove the assertion involving an asymmetric $\mathbf{M}$.

1. By the definition in (6), it is trivial to see that $\mathbf{G}_\gamma$ is symmetric.

2. To prove (7), observe that from (6),

$$(I - \mathbf{G}_1)'\mathbf{D}(I - \mathbf{G}_1) = (I - \mathbf{M})'\mathbf{D}(I - \mathbf{M}), \quad (10)$$

and

$$\begin{aligned} &\operatorname{tr}(\mathbf{D}\mathbf{G}_1) = \operatorname{tr}(\mathbf{D}) \\ &\quad - \operatorname{tr}[(\mathbf{D}^{1/2}(I - \mathbf{M})\mathbf{D}^{1/2})'(\mathbf{D}^{1/2}(I - \mathbf{M})\mathbf{D}^{1/2})]^{1/2}. \end{aligned} \quad (11)$$

Because M is asymmetric, so is $\mathbf{D}^{1/2}(I - \mathbf{M})\mathbf{D}^{1/2}$. Apply Lemma 1 to (11) to get

$$\begin{aligned} \operatorname{tr}(\mathbf{D}\mathbf{G}_1) &< \operatorname{tr}(\mathbf{D}) - \operatorname{tr}(\mathbf{D}^{1/2}(I - \mathbf{M})\mathbf{D}^{1/2}) \\ &= \operatorname{tr}(\mathbf{D}) - \operatorname{tr}(\mathbf{D}) + \operatorname{tr}(\mathbf{D}\mathbf{M}) \\ &= \operatorname{tr}(\mathbf{D}\mathbf{M}). \end{aligned} \quad (12)$$

From (10),

$$\mathbf{D} - \mathbf{G}_1'\mathbf{D} - \mathbf{D}\mathbf{G}_1 + \mathbf{G}_1'\mathbf{D}\mathbf{G}_1 = \mathbf{D} - \mathbf{M}'\mathbf{D} - \mathbf{D}\mathbf{M} + \mathbf{M}'\mathbf{D}\mathbf{M}.$$

Hence (12) yields $\operatorname{tr}(\mathbf{G}_1'\mathbf{D}\mathbf{G}_1) < \operatorname{tr}(\mathbf{M}'\mathbf{D}\mathbf{M})$. This together with (10) leads to (7).

3. To prove (8) and (9), begin by noting that

$$(I - \mathbf{G}_\gamma)'\mathbf{D}(I - \mathbf{G}_\gamma) = \gamma^2(I - \mathbf{M})'\mathbf{D}(I - \mathbf{M}).$$

Hence, because $\gamma < 1$,

$$\theta'(I - \mathbf{G}_\gamma)'\mathbf{D}(I - \mathbf{G}_\gamma)\theta \leq \theta'(I - \mathbf{M})'\mathbf{D}(I - \mathbf{M})\theta$$

with inequality whenever the right side is not 0. Let

$$\begin{aligned} g(\gamma) &= \operatorname{tr}(\mathbf{G}_\gamma'\mathbf{D}\mathbf{G}_\gamma) \\ &= \gamma^2 \operatorname{tr}((I - \mathbf{M})'\mathbf{D}(I - \mathbf{M})) + c_1\gamma + c_0, \end{aligned} \quad (13)$$

where $c_1 = -2 \operatorname{tr}((\mathbf{D}^{1/2}(I - \mathbf{M})'\mathbf{D}(I - \mathbf{M})\mathbf{D}^{1/2})^{1/2})$ and $c_0 = \operatorname{tr}(\mathbf{D})$ are functions of $\mathbf{M}$ and $\mathbf{D}$. Also, $\operatorname{tr}((I - \mathbf{M})'\mathbf{D}(I - \mathbf{M})) > 0$, because $(I - \mathbf{M})'\mathbf{D}(I - \mathbf{M})$ is at least positive semidefinite. This implies that $g(\gamma)$ is an upward parabola (see Fig. 1).

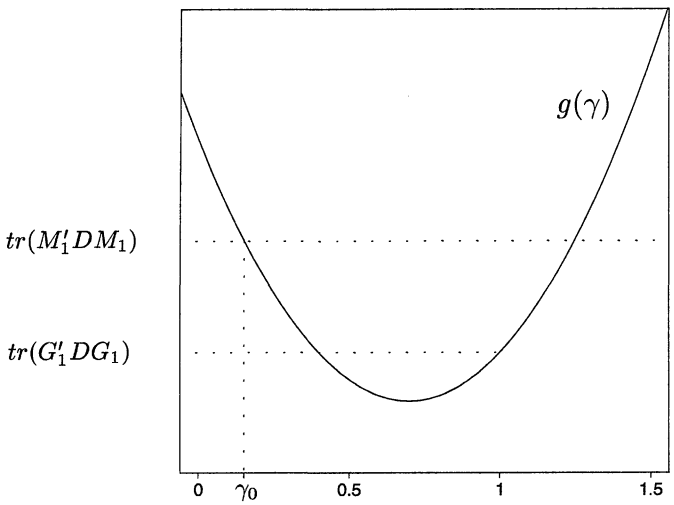


Figure 1. Graph of g(γ).

From step 2,

$$g(1) = \operatorname{tr}(\mathbf{G}_1'\mathbf{D}\mathbf{G}_1) < \operatorname{tr}(\mathbf{M}'\mathbf{D}\mathbf{M}).$$

This guarantees that the smallest root of $g(\gamma) = \operatorname{tr}(\mathbf{M}'\mathbf{D}\mathbf{M})$ will be less than 1. Take

$$\gamma_0 = \max\{0, \text{ the smaller root of } g(\gamma) = \operatorname{tr}(\mathbf{M}'\mathbf{D}\mathbf{M})\}. \tag{14}$$

Then

$$0 \le \gamma_0 < 1.$$

If $\gamma = \gamma_0$, then

$$\operatorname{tr}(\mathbf{G}_{\gamma_0}'\mathbf{D}\mathbf{G}_{\gamma_0}) \le \operatorname{tr}(\mathbf{M}'\mathbf{D}\mathbf{M}),$$

which leads to (8). If

$$\gamma_0 < \gamma < 1,$$

then

$$\operatorname{tr}(\mathbf{G}_{\gamma}'\mathbf{D}\mathbf{G}_{\gamma}) < \operatorname{tr}(\mathbf{M}'\mathbf{D}\mathbf{M}),$$

which leads to (9).

Now suppose that $\mathbf{M}$ is symmetric and maxeig $\mathbf{M} > 1$. Then mineig $\mathbf{D}^{1/2}(I - \mathbf{M})\mathbf{D}^{1/2} < 0$. Hence (12) holds because of the last assertion in Lemma 1, and the remainder of the proof is the same.

Remark. For asymmetric $\mathbf{M}$, Cohen (1966) has the same result in the special case when $\mathbf{D} = I$ and $\gamma = 1$. He also gives an improvement when $\mathbf{D} = I$ and $\mathbf{M}$ is symmetric with maxeig $\mathbf{M} > 1$, but there his improvement is different from ours.

Numerical implementation of this estimator is simple and direct. One need only construct the quadratic equation in (14) to solve for γ_0. The estimator (6) is then a direct matrix computation that involves producing the square root of the bracketed matrix in (6). This can be done quickly with standard software for sample sizes in the hundreds. For larger values of n, it may be helpful to make-use of special properties of M, such as the fact that in nonparametric regression it is often a banded matrix. Similar comments apply to the estimator in Theorem 2.

Notice that the improved estimators always have the maximum eigenvalue ≤ 1. In the case where the minimum eigenvalue < 0, one can get further improvement via the following theorem.

Theorem 2. Let $\mathbf{\Sigma} = I$. Consider a linear estimator $\hat{\theta}$ such that

$$\hat{\theta} = \mathbf{M}Y.$$

Suppose that M is symmetric and mineig$(\mathbf{M}) < 0$. Denote the orthogonal decomposition of $\mathbf{M}$ as $\mathbf{M} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}'$ with $\lambda_1 < 0, \ldots \lambda_r < 0 \le \lambda_{r+i}, i = 1, \ldots, n - r$. Let $U = \operatorname{diag}(u_i)$, where $u_i = |\lambda_i|, i = 1, \ldots r$ and $u_i = 0$ otherwise. Define

$$H_\rho = \mathbf{M} + \rho\mathbf{D}^{-1}\mathbf{Q}U\mathbf{Q}' \tag{15}$$

for $0 < \rho < \rho^*$, where ρ^* is as in (21). Then $\tilde{\theta}_\rho = \mathbf{H}_\rho Y$ dominates $\hat{\theta}$. Also, $\tilde{\theta}_\rho$ dominates $\hat{\theta}$ when $\rho = \rho^*$ so long as $\rho_1 \ne \rho_2$, where ρ_1 and ρ_2 are defined in (18) and (20).

Proof.

$$\begin{aligned}
&\operatorname{bias}^2(\mathbf{M}, \theta) - \operatorname{bias}^2(\mathbf{H}_\rho, \theta) \\
&\quad = \theta'[2\rho\mathbf{Q}(I - \mathbf{\Lambda})U\mathbf{Q}' - \rho^2\mathbf{Q}U\mathbf{Q}'\mathbf{D}^{-1}\mathbf{Q}U\mathbf{Q}']\theta \\
&\quad = \theta'\mathbf{Q}[2\rho(I - \mathbf{\Lambda})U - \rho^2 U\mathbf{Q}'\mathbf{D}^{-1}\mathbf{Q}U]\mathbf{Q}'\theta \qquad (16) \\
&\quad = \rho^2\theta'\mathbf{Q}U^{1/2}(I + U)^{1/2}\left\{\frac{2}{\rho} - \mathbf{T}\right\}(I + U)^{1/2}U^{1/2}\mathbf{Q}'\theta, \qquad (17)
\end{aligned}$$

where $\mathbf{T} = (I + U)^{-1/2}U^{1/2}\mathbf{Q}'\mathbf{D}^{-1}\mathbf{Q}U^{1/2}(I + U)^{-1/2}$, which is a symmetric positive semidefinite rank r matrix with nonzero entries only in the upper left $r \times r$ quadrant. Take

$$\rho_1 = \frac{2}{\operatorname{maxeig}(\mathbf{T})} > 0; \tag{18}$$

then (17) will be greater than or equal to 0 if $\rho < \rho_1$. The equality holds only if $Q\theta = (0, \ldots, \theta_{r+1}, \ldots, \theta_n)$.

For the variance terms,

$$\begin{aligned}
&\operatorname{var}(\mathbf{M}) - \operatorname{var}(\mathbf{H}_\rho) \\
&\quad = -2\rho \operatorname{tr}(\mathbf{\Lambda}U) - \rho^2 \operatorname{tr}(U\mathbf{Q}'\mathbf{D}^{-1}\mathbf{Q}U) \\
&\quad > 0 \quad \text{for} \quad 0 < \rho < \rho_2, \qquad (19)
\end{aligned}$$

where

$$\rho_2 = \frac{2\operatorname{tr}(U^2)}{\operatorname{tr}(U\mathbf{Q}'\mathbf{D}^{-1}\mathbf{Q}U)} > 0. \tag{20}$$

Let

$$\rho^* = \min\{\rho_1, \rho_2\} > 0. \tag{21}$$

This complete the proof.

When $\mathbf{D} = I$ and mineig$(\mathbf{M}) < 0, \rho_2 = 2, \rho_1 > 2$, and the maximum improvement in the special cases when $\rho = 1$ or $\rho = 2$ can be characterized.

Corollary 1. Assume that $\mathbf{D} = I$. Consider a linear estimator $\hat{\theta} = \mathbf{M}Y$, such that $\mathbf{M}$ is symmetric and $\lambda_0 = \text{mineig}(\mathbf{M}) < 0$. $\mathbf{M} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}'$ for some orthogonal matrix $\mathbf{Q}$ and

$$\mathbf{\Lambda} = \begin{pmatrix} -\mathbf{\Lambda}_1 & 0 \\ 0 & \mathbf{\Lambda}_2 \end{pmatrix}$$

with $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ diagonal, where $\mathbf{\Lambda}_1$ is $r \times r$, $\mathbf{\Lambda}_1 > 0, \mathbf{\Lambda}_2 \geq 0$. Assume that $|\lambda_0| \leq 1$; then

$$\mathbf{H}_1 = \mathbf{Q}\begin{pmatrix} 0 & 0 \\ 0 & \mathbf{\Lambda}_2 \end{pmatrix}\mathbf{Q}', \tag{22}$$

$$\mathbf{H}_2 = \mathbf{Q}\begin{pmatrix} \mathbf{\Lambda}_1 & 0 \\ 0 & \mathbf{\Lambda}_2 \end{pmatrix}\mathbf{Q}'. \tag{23}$$

Let $\hat{\theta}^+$ and $\hat{\theta}_{abs}$ be the linear estimators with matrices $\mathbf{H}_1$ and $\mathbf{H}_2$. Both estimators improve on $\hat{\theta}$. The maximum relative improvements are

$$\max_{\theta} \frac{R(\hat{\theta},\theta) - R(\hat{\theta}^+,\theta)}{R(\hat{\theta},\theta)} = \max\left\{1 - \frac{\sum \lambda_i^{+2}}{\sum \lambda_i^2}, 1 - \frac{1}{(1-\lambda_0)^2}\right\} \tag{24}$$

and

$$\max_{\theta} \frac{R(\hat{\theta},\theta) - R(\hat{\theta}_{abs},\theta)}{R(\hat{\theta},\theta)} = 1 - \left(\frac{1+\lambda_0}{1-\lambda_0}\right)^2, \tag{25}$$

where

$$\lambda_i^+ = \begin{cases} 0 & \lambda_i < 0 \\ \lambda_i & \lambda_i \geq 0. \end{cases}$$

Proof. Improvement of $\hat{\theta}^+$ and $\hat{\theta}_{abs}$ over $\hat{\theta}$ follows from Theorem 2, because here $\rho_2 = 2$ and $\rho_1 > 2$.

To prove (24), note that

$$\begin{aligned}
&\max_{\theta} \frac{R(\hat{\theta},\theta) - R(\hat{\theta}^+,\theta)}{R(\hat{\theta},\theta)} \\
&= \max_{\theta}\left\{\frac{\text{tr}(\mathbf{M}^2) - \text{tr}(\mathbf{M}^{+2}) + \theta'((I-\mathbf{M})^2 - (I-\mathbf{M}^+)^2)\theta}{\text{tr}(\mathbf{M}^2) + \theta'(I-\mathbf{M})^2\theta}\right\} \\
&= \max\left\{1 - \frac{\text{tr}(\mathbf{M}^{+2})}{\text{tr}(\mathbf{M}^2)}, \max_{\theta \neq 0}\left(1 - \frac{\theta'(I-\mathbf{M}^+)^2\theta}{\theta'(I-\mathbf{M})^2\theta}\right)\right\} \\
&= \max\left\{1 - \frac{\sum \lambda_i^{+2}}{\sum \lambda_i^2}, 1 - \min_i\left(\frac{(1-\lambda_i^+)^2}{(1-\lambda_i)^2}\right)\right\} \\
&= \max\left\{1 - \frac{\sum \lambda_i^{+2}}{\sum \lambda_i^2}, 1 - \frac{1}{(1-\lambda_0)^2}\right\}.
\end{aligned}$$

Note that the expression $1 - (\sum \lambda_i^{+2} / \sum \lambda_i^2)$ gives the improvement in the variance term when using $\hat{\theta}^+$. Equation (25) can be proven similarly.

Remark 1. In Theorem 2, $\mathbf{H}_\rho$ may not be symmetric when $\mathbf{D}$ is not a multiple of the identity. One may use $(\mathbf{H}_\rho)_\gamma$, as described in Theorem 1. This matrix is symmetric and has $\text{maxeig}(\mathbf{H}_\rho)_\gamma \leq 1$. If mineig $(\mathbf{H}_\rho)_\gamma < 0$, then one may repeat the procedure of Theorem 2 to get further improvement.

Remark 2. When Y has variance of $\sigma^2\mathbf{\Sigma}$ and $\mathbf{\Sigma}$ is known, one can use the structural invariance of the problem to obtain similar explicit results. For a linear estimator $\hat{\theta} = \mathbf{G}Y$, the risk function can be written as

$$\begin{aligned}
R(\hat{\theta},\theta) &= E_\theta(\mathbf{G}Y - \theta)'\mathbf{D}(\mathbf{G}Y - \theta) \\
&= E_{\theta^{(1)}}(\mathbf{\Sigma}^{-1/2}\mathbf{G}\mathbf{\Sigma}^{1/2}Y^{(1)} - \theta^{(1)})' \\
&\quad \times \mathbf{\Sigma}^{1/2}\mathbf{D}\mathbf{\Sigma}^{1/2}(\mathbf{\Sigma}^{-1/2}\mathbf{G}\mathbf{\Sigma}^{1/2}Y^{(1)} - \theta^{(1)}),
\end{aligned}$$

where $\theta^{(1)} = \mathbf{\Sigma}^{-1/2}\theta, Y^{(1)} = \mathbf{\Sigma}^{-1/2}Y$ with $E(Y^{(1)}) = \theta^{(1)}$ and $\text{var}(Y^{(1)}) = \sigma^2 I$. So $\mathbf{G}Y$ is admissible if and only if $\mathbf{\Sigma}^{-1/2}\mathbf{G}\mathbf{\Sigma}^{1/2}$ is symmetric and all of the eigenvalues of $\mathbf{G}$ are between 0 and 1. Furthermore, $\mathbf{\Sigma}^{-1/2}\mathbf{G}\mathbf{\Sigma}^{1/2}$ is symmetric if and only if $\mathbf{G}\mathbf{\Sigma}$ is symmetric. This leads to the following general theorem.

Theorem 3. Consider a linear estimator $\hat{\theta} = \mathbf{G}Y$. Let $\sigma^2\mathbf{\Sigma}$ denote the covariance matrix of Y, with $\mathbf{\Sigma} \neq I$. If $\mathbf{G}\mathbf{\Sigma}$ is not symmetric or the $\text{mineig}(\mathbf{G}) < 0$ or $\text{maxeig}(\mathbf{G}) > 1$, then we can construct a class of linear estimators that are better than $\hat{\theta}$.

Proof. As noted, set $\mathbf{M} = \mathbf{\Sigma}^{-1/2}\mathbf{G}\mathbf{\Sigma}^{1/2}$ and replace $\mathbf{D}$ by $\mathbf{\Sigma}^{1/2}\mathbf{D}\mathbf{\Sigma}^{1/2}$ in Theorems 1 and 2.

3. APPLICATIONS IN NONPARAMETRIC REGRESSION

In this section the nonparametric regression model is considered. Observe $y_1, \ldots, y_n$ and assume that $y_i = \mathbf{m}(x_i) + \varepsilon_i, i = 1, \ldots, n$. Here $x_i \in I = (\alpha, \beta) \subset \mathcal{R}, y_i \in \mathcal{R}$, and $\mathbf{m}$ is the mean function of Y given X; that is,

$$\mathbf{m}(x) = E(Y|x), \tag{26}$$

and ε is observational error. Assume that

$$\varepsilon_i \text{ are independent with mean 0 and variance } \sigma^2. \tag{27}$$

The unknown function $\mathbf{m}$ will be estimated. This model does not put any restriction on the functional form of $\mathbf{m}$. For most applications, assumptions are made concerning the smoothness of $\mathbf{m}$, but such assumptions are not required for the following discussion.

The idea of kernel smoothing in density estimation can be traced back to Rosenblatt (1956) and Parzen (1962). This was adapted to the nonparametric regression problem by Nadaraya (1964) and Watson (1964), who proposed the following scheme.

Take a symmetric function $K(x)$ such that $\int K(x)\,dx = 1$. Let

$$K_{h_n}(\cdot) = \frac{1}{h_n}K\left(\frac{\cdot}{h_n}\right). \tag{28}$$

For given h_n, define the weight at x given to y_i as

$$W_i(x) = \frac{K_{h_n}(x - x_i)}{\sum_l K_{h_n}(x - x_l)}. \tag{29}$$

(Assume that the denominator is positive, which will normally be true in applications.) Then the Nadaraya–Watson estimator is the weighted average of y_i given by

$$\hat{\mathbf{m}}_{\mathrm{NW}}(x) = \sum_i W_i(x) y_i. \tag{30}$$

Here K is called the kernel; h_n, the bandwidth.

Various kernel functions have been used in general. For example, one may take the rectangular kernel

$$K(x) = \frac{1}{2} I_{[-1,1]}(x),$$

which produces the local average. The normal (0, 1) density function

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

is also frequently used. The Epanechnikov (1969) kernel,

$$K(x) = .75(1 - x^2) I_{[-1,1]}(x),$$

which has certain optimality properties, is sometimes preferred (see, e.g., Donoho and Liu 1991; Donoho, Liu, and MacGibbon 1990; Sacks and Ylvisaker 1981).

Although the choice of K is relevant, it is more crucial which h_n is used. The choices of h_n are beyond the scope of this article. I assume here that h_n is fixed, but see the remark following Corollary 2.

Another popular kernel-based estimator was proposed by Gasser and Müller (1979). Assume, with no loss of generality, that the x_i have been arranged in nondecreasing order. Let $s_{i-1} \le x_i \le s_i, i = 1, \ldots, n$, according to some predetermined scheme; for example, take $s_0 = -\infty, s_n = \infty$, and $s_j = (x_j + x_{j+1})/2, j = 1, \ldots, n-1$. Then define

$$W_i(x) = \int_{s_{i-1}}^{s_i} K_{h_n}(x - t)\, dt \tag{31}$$

with K_{h_n} as in (28). The Gasser–Müller estimator is the weighted average with weights defined in (31). It is denoted by

$$\hat{\mathbf{m}}_{\mathrm{GM}}(x) = \sum_i W_i(x) y_i. \tag{32}$$

A third popular kernel-based technique is locally weighted polynomial regression, as proposed by Cleveland (1979) and Stone (1977). For simplicity, I discuss only locally weighted linear regression. Consider the weighted linear regression that finds the minimum solution to

$$\min_{a,b} \sum_i (y_i - (a + b(x_i - x))^2 K\left(\frac{x_i - x}{h_n}\right). \tag{33}$$

Suppose that $\hat{a}$ and $\hat{b}$ are the minimizer of (33). Then $\hat{a}$ will be used to estimate $m(x)$. Denote $\hat{a}$ by $\hat{\mathbf{m}}_{\mathrm{LL}}$. Simple calculation gives the form of $\hat{\mathbf{m}}_{\mathrm{LL}}$ as

$$\hat{\mathbf{m}}_{\mathrm{LL}}(x) = \sum_i w_i y_i,$$

$$\text{where} \quad w_i = \frac{\alpha_i}{\sum \alpha_j}, \quad \text{with}$$

$$\alpha_i = K\left(\frac{x - x_i}{h_n}\right)(s_{n,2}(x) - (x - x_i) s_{n,1}(x))$$

and

$$s_{n,l}(x) = \sum_i K\left(\frac{x - x_i}{h_n}\right)(x - x_i)^l, \qquad l = 1, 2. \tag{34}$$

Properties of the foregoing estimators have also been given by Fan (1993) and Hastie and Loader (1993). More details about the three popular nonparametric regression methods mentioned in this section have been provided by Hastie and Tibshirani (1990) and Härdle (1990).

When an estimator $\hat{\mathbf{m}}$ is linear, as earlier, it can be written as

$$\hat{\mathbf{m}}(x) = \sum_i w_i(x) y_i \tag{35}$$

for some $w_i(x)$, which is a function of $\{x_i\}$ and x only.

One useful measure of the accuracy of an estimator is its weighted mean integrated squared error (WMISE), defined as

$$\mathrm{WMISE}(\mathbf{m}, \hat{\mathbf{m}}) = E_{\mathbf{m}}\left(\int_a^b (\hat{\mathbf{m}}(x) - \mathbf{m}(x))^2 \lambda(x)\, dx\right), \tag{36}$$

where $\lambda(x) \ge 0$ is a weight function. The discrete version of (36) is

$$R_{\mathbf{A}}(\mathbf{m}, \hat{\mathbf{m}}) = \sum E(\mathbf{m}(x_i) - \hat{\mathbf{m}}(x_i))^2 d_{ii}, \tag{37}$$

which is most often used with $d_{ii} > 0$. [If the x_i are equally spaced, then the choice $d_{ii} = \lambda(x_i)/n$ makes (37) a good approximation of (36).]

Let

$$\vec{\mathbf{m}} = \begin{pmatrix} m(x_1) \\ \vdots \\ m(x_n) \end{pmatrix} \qquad \hat{\vec{\mathbf{m}}} = \begin{pmatrix} \hat{m}(x_1) \\ \vdots \\ \hat{m}(x_n) \end{pmatrix}$$

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{D} = (d_{ij}) = \mathrm{diag}(d_{ii}). \tag{38}$$

By the foregoing definition, $\mathbf{D} > 0$. Notice that estimating $\mathbf{m}(x)$ in nonparametric problems under $R_{\mathbf{A}}$ is equivalent to estimating $\vec{\mathbf{m}}$ in a multivariate mean problem under the same weighted squared error loss. Results about the latter problem, especially the results in the previous section, then can be applied to the nonparametric regression problems.

In the previous section I gave a simple method for improving a linear estimator whose matrix is either not symmetric or has some negative eigenvalues. Next I show that all three commonly used estimators mentioned earlier in this section are linear but with asymmetric matrices and thus can be improved in the sense that $R_{\mathbf{A}}$ can be reduced.

Theorem 4. The Nadaraya–Watson, Gasser–Müller, and locally weighted linear estimators are all linear but with asymmetric matrices.

Proof. The linearity is a direct result of the definition in (30), (32), and (34). To see the asymmetricity, let $\mathbf{M}_{\mathrm{NW}}$ denote the matrix corresponding to the Nadaraya–Watson estimator; that is, $\hat{\mathbf{m}}_{\mathrm{NW}} = \mathbf{M}_{\mathrm{NW}}Y$. By (28)–(30),

$$\mathbf{M}_{\mathrm{NW}}(i,j) = \frac{K_{h_n}(x_i - x_j)}{\sum_l K_{h_n}(x_i - x_l)} \tag{39}$$

and

$$\mathbf{M}_{\mathrm{NW}}(j,i) = \frac{K_{h_n}(x_j - x_i)}{\sum_l K_{h_n}(x_j - x_l)}. \tag{40}$$

Note that $K_{h_n}(x_i - x_j) = K_{h_n}(x_j - x_i)$ because of the symmetry of K, but $\sum_l K_{h_n}(x_i - x_l) \neq \sum_l K_{h_n}(x_j - x_l)$ for some i and j. This inequality will occur if $x_i \neq x_j$ and x_i is sufficiently close to an endpoint of the interval (α, β). It will also frequently occur if the x's are not equally spaced. This shows that $\mathbf{M}_{\mathrm{NW}}$ is asymmetric.

I can also explicitly write the other two matrices. Let $\mathbf{M}_{\mathrm{GM}}$ and $\mathbf{M}_{\mathrm{LL}}$ denote the matrices corresponding to Gasser–Müller and locally weighted linear estimators. Then, by (31)–(32),

$$\mathbf{M}_{\mathrm{GM}}(i,j) = \int_{s_{i-1}}^{s_i} K_{h_n}(x_j - t)\,dt. \tag{41}$$

By (34),

$$\mathbf{M}_{\mathrm{LL}}(i,j) = \frac{K\left(\frac{x_i - x_j}{h_n}\right)\left(s_{n,2}(x_i) - (x_i - x_j)s_{n,1}(x_i)\right)}{\sum_l K\left(\frac{x_i - x_l}{h_n}\right)\left(s_{n,2}(x_i) - (x_i - x_l)s_{n,1}(x_i)\right)}, \tag{42}$$

where

$$s_{n,m} = \sum_l K\left(\frac{x_i - x_l}{h_n}\right)(x_i - x_l)^m, \qquad \mathbf{m} = 1, 2.$$

The asymmetry arguments for these matrices are similar to the argument for $\mathbf{M}_{\mathrm{NW}}$; I omit the details. Note that whenever $s_{n,j} \neq 0$, one should expect $\mathbf{M}_{\mathrm{LL}}(i,j) \neq \mathbf{M}_{\mathrm{LL}}(j,i)$ for $i \neq j$.

As a corollary of Theorem 4, the three commonly used estimators can be improved according to Theorem 1.

Corollary 2. The Nadaraya–Watson, Gasser–Müller, and locally weighted linear estimators can be constructively improved in the sense that $R_{\mathbf{A}}$ can be reduced.

Proof. In view of Theorem 4, Theorem 1 can be applied to explicitly construct new matrices by formula (6).

Remarks. The new estimators can be improved further by Theorem 2 (15) if there are some negative eigenvalues.

The previous considerations apply to kernel estimators for which the bandwidth, h_n, is not dependent on Y. In practice, one often uses a data-dependent h_n, such as determined by some type of cross-validation. A heuristically promising approach in such situations would be to apply the methodology of this article as if $h_n = \hat{h}_n$—that is, after computing $\hat{h}_n$ substitute $\hat{h}_n$ in (39), (41), and (42) as appropriate, and then apply the methodology of Theorems 1 and 2 to construct new estimates. This appears to work well in the various simulations that I have conducted, but I have not yet found theoretical proof for its desirability.

4. ASYMPTOTIC COMPARISONS

This section gives a theorem that describes the asymptotic maximum relative improvement (as $n \to \infty$) for two commonly used kernel or weight functions.

Theorem 5. Suppose that $\{x_i\}$ are equally spaced on $[-1, 1]$. Let $\mathbf{M} = (m_{ij})$ denote the matrix corresponding to the Nadaraya–Watson, Gasser–Müller, or locally weighted linear regression estimator. When n is sufficiently large and h_n is sufficiently small and the rectangular or Epanechnikov kernels are used, then the estimators can be improved. Especially, if $\mathbf{D} = I$, then the maximum improvement (%) as $n \to \infty$ and $h_n \to \infty, nh_n \to \infty$ will be at least 32.5 for $\mathbf{H}_1$ and 58.6 for $\mathbf{H}_2$ for the rectangular kernel, and 15.24 for $\mathbf{H}_1$ and 29.22 for $\mathbf{H}_2$ for the Epanechnikov kernel, where $\mathbf{H}_1$ and $\mathbf{H}_2$ are defined in Corollary 1.

Proof. I give a sketch of the proof.

1. Consider the Nadaraya–Watson estimator. Under the foregoing assumption, $\mathbf{M}$ is almost symmetric because, by (28)–(30),

$$\mathbf{m}_{ij} = \frac{\frac{1}{nh}K\left(\frac{x_i - x_j}{h}\right)}{\sum_j \frac{1}{nh}K\left(\frac{x_i - x_j}{h}\right)} \approx \frac{\frac{1}{nh}K\left(\frac{x_i - x_j}{h}\right)}{\int \frac{1}{h}K\left(\frac{x_i - t}{h}\right)dt}$$
$$= \frac{1}{nh}K\left(\frac{x_i - x_j}{h}\right).$$

So Theorem 2 can be applied if there are some negative eigenvalues. The same asymptotic formula holds for the other two methods.

2. Notice that the linear transformation $\mathbf{M}$: $\mathbf{R}^n \to \mathbf{R}^n$ is asymptotically equivalent to the linear operator L: $\mathcal{L}_2 \to \mathcal{L}_2$, defined as

$$L(g)(x) = \frac{1}{h}\int K\left(\frac{x - t}{h}\right)g(t)\,dt, \qquad g \in \mathcal{L}_2.$$

To find the limiting eigenvalues of $\mathbf{M}$ when $nh_n \to \infty$, it suffices to find the eigenvalues of L.

3. The periodic eigenfunctions of L are $\sin(k\pi t)$ and $\cos(k\pi t)$ with eigenvalues

$$\int \frac{1}{h}K\left(\frac{t}{h}\right)\cos(k\pi t)\,dt, \qquad k = 1, \ldots.$$

This can be shown by taking $g = \sin(k\pi x)$

$$\frac{1}{h}\int_{-1}^{1} K\left(\frac{x - t}{h}\right)\sin(k\pi t)\,dt$$
$$= \frac{1}{h}\int_{-1}^{1} K\left(\frac{t}{h}\right)\sin(k\pi x - k\pi t)\,dt$$

$$= \frac{1}{h}\int_{-1}^{1} K\left(\frac{t}{h}\right)\cos(k\pi t)\,dt\sin(k\pi x).$$

This can be done similarly for $g = \cos(k\pi x)$.

4. Consider the rectangular kernel; that is,

$$K(x) = \frac{1}{2}I_{[-1,1]}(x).$$

The eigenvalues are

$$\begin{aligned}\lambda_k &= \frac{1}{2}\int_{-h}^{h}\frac{1}{h}\cos(k\pi t)\,dt \\ &= \frac{\sin(k\pi h)}{k\pi h}, \qquad k = 1,\ldots\end{aligned}$$

Let

$$\lambda_0 = \min_{\omega}\left\{\frac{\sin\omega}{\omega}\right\} = \frac{\sin\omega_0}{\omega_0} = -.217\ldots$$

with $\omega_0 = 4.493\ldots$. Let $h_n \to 0$ with $nh_n \to \infty$ and let $\mathbf{m}_n(t) = \sin(\omega_0 t/h_n)$.

Let $\mathbf{m}_n^*$ denote the normalized vector of means:

$$\begin{aligned}(\mathbf{m}_n^*)_i &= \sqrt{\frac{2}{n}}E(Y_i|x_i = 2i/n - 1) \\ &= \sqrt{\frac{2}{n}}\mathbf{m}_n\left(\frac{\omega_0}{h_n}(2i/n - 1)\right).\end{aligned}$$

The normalization, $\sqrt{2/n}$, is chosen so that

$$\begin{aligned}\|\mathbf{m}_n^*\|^2 &= \frac{2}{n}\sum\mathbf{m}_n^2\left(\frac{\omega_0}{h_n}\left(\frac{2i}{n} - 1\right)\right) \\ &\to \int_{-1}^{1}\mathbf{m}_n^2\left(\frac{\omega_0 t}{h_n}\right)dt = 1.\end{aligned}$$

Then, by steps 1 and 3,

$$\begin{aligned}(\mathbf{M}\mathbf{m}_n^*)_i &= \sum_j \mathbf{m}_{ij}(\mathbf{m}_n^*)_j \\ &= \frac{\sin\omega_0}{\omega_0}(\mathbf{m}_n^*)_i + o\left(\frac{1}{\sqrt{n}}\right)\end{aligned}$$

as $n \to \infty$, with the $o(1)$ term being uniform in i for $nh_n < i < n(1-h_n)$. Furthermore, for any i, $|\mathbf{M}(\mathbf{m}_n^*)_i| \le 2/n$, because $\mathbf{M}(\mathbf{m}_n^*)_i$ is a weighted average of terms of magnitude at most $\sqrt{2/n}$.

5. It follows from the bounded convergence theorem that $\mathbf{m}_n^*$ is asymptotically an eigenvector of $\mathbf{M}$ in the sense that $\mathbf{M}\mathbf{m}_n^* \to \lambda_0\mathbf{m}_n^*$. More precisely,

$$\begin{aligned}&\left\|\left(\mathbf{M} - \sum_0 I\right)\mathbf{m}_n^*\right\|^2 \\ &\quad\le \left(\sum_{i\notin(nh_n, n(1-h_n))}\left(\frac{2}{n}\right)\right) + \frac{2}{n}\sum_{i=1}^{n}(o(1)^2) \to 0.\end{aligned}$$

From this it follows that the minimum eigenvalue of $\mathbf{M}$, say $\lambda^{(n)}$, satisfies $\limsup\lambda^{(n)} \le \lambda_0$. (Actually, one can show further that $\lambda^{(n)} \to \lambda_0$.)

Now formulas (24) and (25) can be applied to find that the maximum percentage improvements will be

$$\mathbf{H}_1\colon\ 1 - \left(\frac{1}{1+\omega_0}\right)^2 = 32.5\%$$

and

$$\mathbf{H}_2\colon\ 1 - \left(\frac{1-\omega_0}{1+\omega_0}\right)^2 = 58.6\%.$$

The arguments for the Gasser–Miller and locally weighted linear regression estimators are entirely similar, and yield the same improvements.

The proof for the Epanechnikov kernel is analogous, except that one must now define $\lambda_0 = \min_\omega\{3\sin\omega/\omega^3 - \omega\cos\omega)\} = -.08617\ldots$, with the minimum attained at $\omega_0 = 5.763\ldots$.

Remarks. The foregoing maximum improvements occur when h_n and $\mathbf{m}(t) = \mathbf{m}_n(t)$ are as defined in the proof of the theorem and $\sigma^2/(nh_n) \to 0$. This is an uncommon sequence of mean functions, and h_n would not be a desirable bandwidth if this were known to be the situation. For practical situations and well-chosen bandwidths, one should expect much smaller percentage improvements from the application of Theorem 2. In some of those situations, however, significant improvements also are available from using Theorem 1. Some examples are given in the next section.

For the rectangular and Epanechnikov estimators in Theorem 5, some asymptotic improvement is always available. It follows from the foregoing proof and Corollary 1 that one can always use $\mathbf{H}_1$ to improve on the variance term without increasing the bias^2 term. The variance improvement when using $\mathbf{H}_1$, is $1 - \Sigma\lambda_i^{+2}/\Sigma\lambda_i^2$. The limiting value of this can be calculated. For the rectangular kernel, $\lambda_k = [\sin(k\pi h)]/k\pi h$ as shown in step 4 from the foregoing proof. So the asymptotic improvement in variance as $h_n \to 0$ is

$$\begin{aligned}1 - \Sigma\lambda_i^{+2}/\Sigma\lambda_i^2 &= 1 - \frac{\sum\left(\frac{\sin^+(k\pi h_n)}{k\pi h_n}\right)^2}{\sum\left(\frac{\sin(k\pi h_n)}{k\pi h_n}\right)^2} \\ &\to 1 - \frac{\int_0^\infty\left(\frac{\sin^+(x)}{x}\right)^2}{\int_0^\infty\left(\frac{\sin(x)}{x}\right)^2} \quad \text{as} \quad h_n \to 0 \\ &= .067.\end{aligned}$$

Similarly, the improvement for the Epanechnikov kernel is .7%.

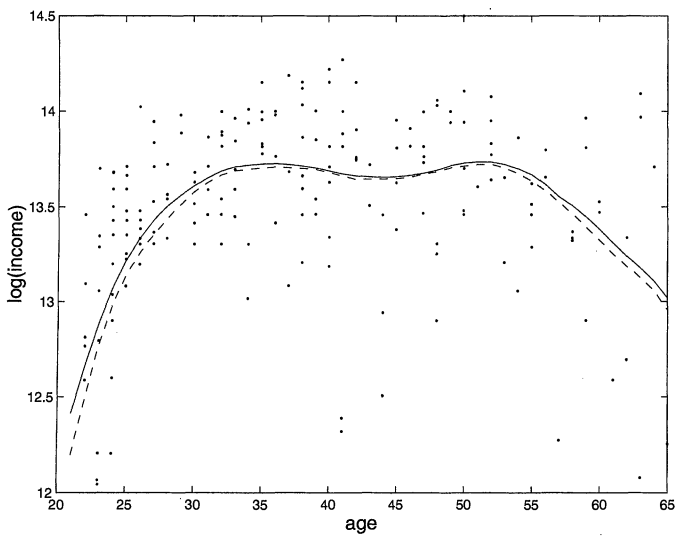
5. EXAMPLES FOR MODERATE *n*

The preceding asymptotic improvements are based only on the negative eigenvalues of $\mathbf{M}$, through Theorem 2 and Corollary 1. They do not take into account the improvement due to asymmetry of $\mathbf{M}$, as in Theorem 1. For realistic,

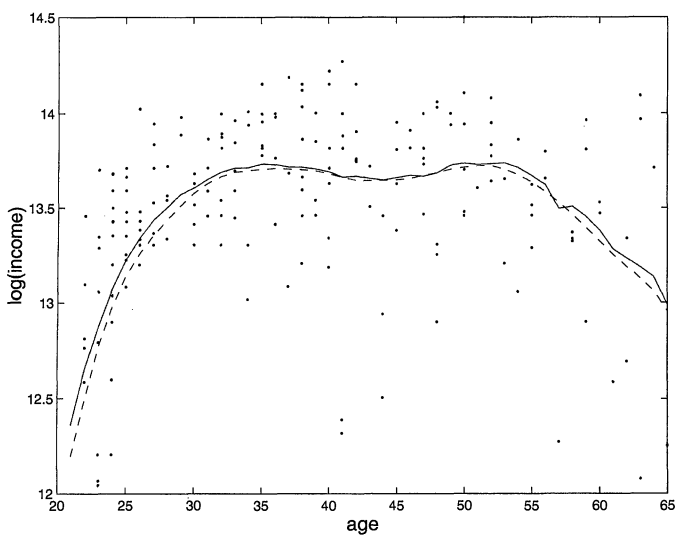
moderate values of n, this improvement can be substantial. The degree of available improvement depends on the placement of $\{x_i\}$, on the details of the estimation method used, and on $\mathbf{m}(x)$. As no general numerical statement seems feasible, I give some examples to provide some idea of what may occur.

Example 1. Consider the income data collected in Canada in 1985 and discussed by Chu and Marron (1991). Figure 2 shows a plot of this data together with a local linear fit based on the Epanechnikov kernel with bandwidth $h = 8$ and the plots of two improved estimators discussed here. In the following I assume the model (26)–(27).

The covariance matrix $\mathbf{\Sigma}_1 = \text{diag}(1/n_i)$, where n_i is the number of observations corresponding to x_i. Take $\mathbf{D}_1 = \text{diag}(n_i)$. Applying the transformation in Theorem 3 yields $\mathbf{D} = \mathbf{\Sigma} = I$. One can proceed so as to emphasize reduction in variance (option A) or reduction in bias^2 (option B).



(a)



(b)

Figure 2. (Example 1: Scatterplots of the Income Data. – – –, local linear fits with h = 8; ———, obtained with variance reduction (a) and with Bias^2 reduction (b).

Option A. To achieve the largest reduction of the variance part of the risk, one can begin by applying (6) with $\gamma = 1$. This results in a 9.49% decrease in the variance of the estimator and no decrease in the bias^2 term.

Because the resulting matrix $\mathbf{G}$ has some negative eigenvalues, the estimator can be further improved by applying (15). For the choice that gives maximum improvement in the variance term in general, use $\rho = \rho_2/2$ if $\rho_2/2 < \rho_1$. When $\mathbf{D} = I, \rho = 1$, because $\rho_2/2 = 1 < \rho_1$.

Iterating this two-step process leads to a symmetric matrix with all its eigenvalues in [0, 1]. Compared to the original estimator, the variance is reduced by 9.88%. Applying Corollary 1 yields an asymptotic figure for the maximum reduction in the bias^2 term of 14.24%. The actual reduction in bias^2 is probably much less than this theoretical maximum.

Option B. For a maximum reduction in bias^2 term, one should use $\gamma = \gamma_0$ in (6) and $\rho = \rho^*$ in (15). After only one step, the resulting matrix is already symmetric with all the eigenvalues in [0, 1]. The asymptotic maximum reduction in bias^2 is 16.54%, and the variance is the same as the original.

Discussion. In Figure 2 note that the kernel estimator (dashed curve) gives a visually smooth curve as a consequence of bandwidth choice based on a prior notion that the curve should be smooth. The variance-reducing estimator (solid curve) in option A inherits the smoothness of the kernel estimator, which it dominates in risk. This appears to also be the case in other examples investigated, and there are heuristic grounds for thinking that it will usually be so.

On the other hand, the curve (solid one) in option B appears much more jagged than the other two. This is not surprising in view of the fact that the emphasis in its construction is on reduction of bias. Nevertheless, as a consequence of the theory, even if the underlying population curve is smooth, the (jagged) estimator produced here dominates the original smoother kernel estimator.

Because it is visually more pleasing, option A should be preferred to option B for situations in which plots of the data are an important part of the output.

Example 2. Consider the geyser data treated by Hall and Turlach (1997), as obtained from Härdle (1991). This dataset has the property that the predictor (x) variables are very unevenly spaced and most of them are concentrated in two clusters, one located in approximately the lower $\frac{1}{3}$ of their range and the other in approximately the upper $\frac{2}{5}$ of their range.

Hall and Turlach used this dataset to investigate the performance of a modified kernel-based estimator they propose that is particularly suited for data with very unevenly spaced predictor variables. The sample size of the dataset is 274. Hall and Turlach first produced a "benchmark" estimate for the regression curve from the entire dataset. They then produced subsamples of size $n = 25$ by sampling without

replacement from all of the data. Nonparametric regression estimators were computed using Hall and Turlach's method from each of the samples of size 25 and were compared to the benchmark curve.

Hall and Turlach's estimator is linear in our sense. They used a fixed bandwidth of $h = 1.96$ for the kernel on which their procedure is based, and the remaining procedure they suggested is then linear.

I also took subsamples of size 25 and computed the Hall and Turlach estimator. Then I computed my improved estimator via option B (= maximum bias^2 improvement). A typical result is shown in Figure 3. Figure 3(a) shows the scatterplot and the benchmark fit; 3(b) shows for a typical subsample of size 25 the estimator produced by Hall and Turlach's method and the improved estimator produced by a two-step implementation in my option B.

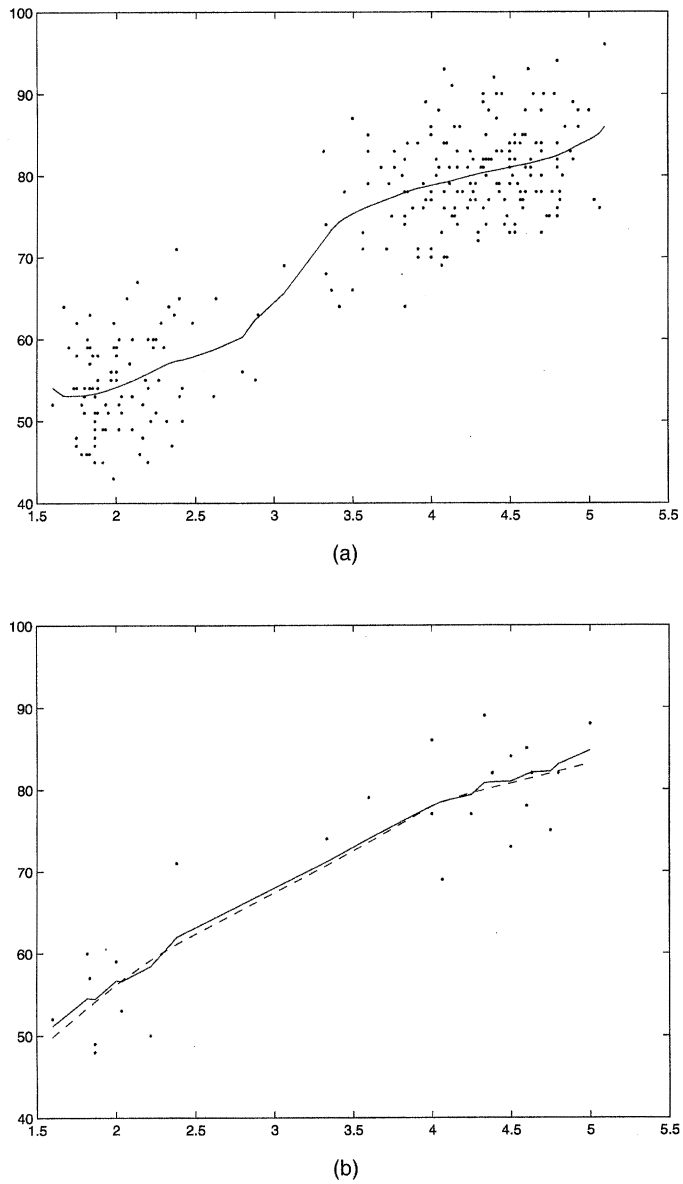


Figure 3. Example 2: Scatterplots of the Geyser Data Together With the Local Linear Fit, h = .5 (a), and of a Subsample Size 25 Together with Hall and Turlach's Fitted Line (– – –) and the Improved One (———), h = 1.96.

I repeated this subsampling procedure 50 times. Each time I computed the quadratic fit to the benchmark of the Hall–Turlach estimator and of my modification of it after two steps, and compared these values to find the relative reduction in quadratic loss. The relative reductions varied from 13.2% to 18.1%, with a mean of 15.05% and a standard error of .95%.

Example 3. I also calculated the relative improvement in a number of special examples with known choices of $\mathbf{m}(\cdot)$. The results varied greatly, of course, depending on the choice of $\mathbf{m}$, the sample size, the sample distribution of the independent variables, and the bandwidth h. Across the range of examples, I found relative improvements ranging from negligible amounts up to about 20%. Because of the extremely wide variety of potential examples and the sensitivity of the relative improvement results to detailed features of these examples, it seems most useful to just present the results from one rather typical example.

Figure 4 summarizes this example. The sample size was 50, and the independent variables were a random sample from a normal distribution with mean .75 and variance 1. Their histogram is shown in Figure 4(a). The function $\mathbf{m}$ for this example has three modes and is given by the equation

$$\mathbf{m}(x) = \frac{1}{((x-.03)^2+.01)} + \frac{1}{((x-.9)^2+.03)} + \frac{1}{((x-1.6)^2+.02)} - 6.$$

Figure 4(b) shows this underlying function, $\mathbf{m}$. For this example, σ was chosen to be $\sigma = 15$. Figure 4(b) also shows a typical scatterplot for this situation. Figure 4(c) shows the basic result for this situation. This plot shows the relative reduction in risk obtained for various values of h. The results shown are from using the improved estimator with $\gamma = \gamma_0$ in (6) and then $\rho = \rho^*$ in (15), which emphasizes

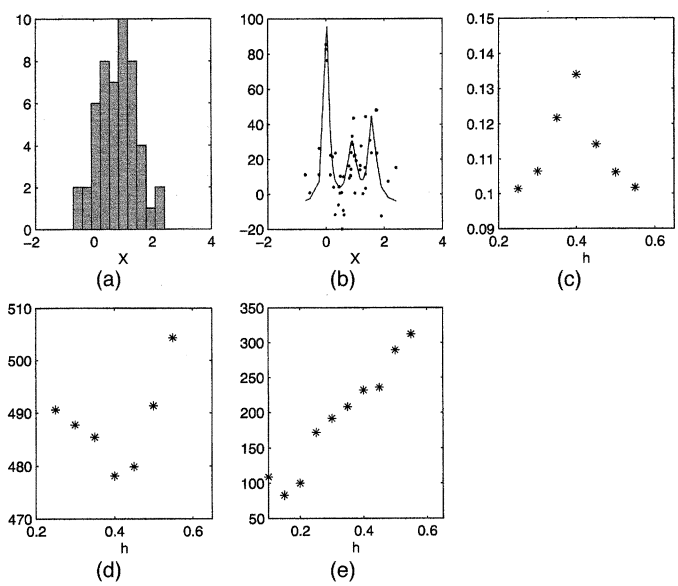


Figure 4. Results for a Typical Simulation as Described in Example 3. (a) Histogram of X; (b) underlying function; (c) proportional reduction at various h; (d) cross-validation; (e) risk.

bias reduction. Relative reduction with $\gamma = 1$ was much smaller here—in the range of .5% to 2%. I emphasize that these reductions are calculated from the algebraic formulas in Theorems 1 and Theorem 2. They depend on the location of independent variables, as shown by the histogram in Figure 4(a); they do not depend on the specific y values indicated in Figure 4(b). Suitable bandwidth choices for this example range from about .25 to about .45. For those values, the relative reduction shown in Figure 4(c) ranges from 10% to 13.5%. To confirm this as the suitable range of bandwidth, Figure 4(d) shows the cross-validation plot from the S-PLUS locfit function corresponding to the data pictured in Figure 4(b). The cross-validated bandwidth choice here is $h = .4$. The oracle bandwidth for this example would be $h = .15$. This can be seen from the risk function shown in Figure 4(e).

[Received March 1996. Revised July 1998.]

REFERENCES

Chu, C., and Marron, S. (1991), "Choosing a Kernel Regression Estimator," *Statistical Science*, 6, 404–436.

Cleveland, W. S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836.

Cohen, A. (1966), "All Admissible Linear Estimators of the Mean Vector," *Annals of Mathematical Statistics*, 37, 458–463.

Donoho, D. L. (1994), "Statistical Estimation and Optimal Recovery," *The Annals of Statistics*, 22, 238–270.

Donoho, D. L., and Liu, R. (1991), "Geometrizing Rates of Convergence III," *The Annals of Statistics*, 19, 668–701.

Donoho, D. L., Liu, R., and MacGibbon, B. (1990), "Minimax Risk Over Hyperrectangles and Implications," *The Annals of Statistics*, 18, 1416–1437.

Epanechnikov, V. A. (1969), "Nonparametric Estimates of a Multivariate Probability Density," *Theory and Probability Applications*, 14, 153–158.

Fan, J. (1993), "Local Linear Regression Smoothers and Their Minimax Efficiencies," *The Annals of Statistics*, 21, 196–216.

Gasser, T., and Müller, H. G. (1979), "Kernel Estimation of Regression Function," in *Smoothing Techniques for Curve Estimation*, eds. Gasser and Rosenblatt, Heidelberg: Springer-Verlag.

Hall, P., and Turlach, B. A. (1997), "Interpolation Methods for Adapting to Sparse Design in Nonparametric Regression," *Journal of the American Statistical Association*, 92, 466–472.

Härdle, W. K. (1990), *Applied Nonparametric Regression*, Cambridge, MA: Cambridge University Press.

Hastie, T., and Loader, C. (1993), "Local Regression: Automatic Kernel Carpentry," *Statistical Science*, 8, 120–143.

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, London: Chapman and Hall.

Nadaraya, E. A. (1964), "On Estimating Regression," *Theory of Probability and Applications*, 10, 186–190.

Parzen, E. (1962), "On the Estimation of a Probability Density and Mode," *Annals of Mathematical Statistics*, 33, 1065–1076.

Rao, C. (1976), "Estimation of Parameters in a Linear Model," *Annals of Statistics*, 4, 1023–1037.

Rosenblatt, M. (1956), "Remarks on Some Nonparametric Estimates of a Density Function," *Annals of Mathematical Statistics*, 27, 832–835, 1956.

Sacks, J., and Ylvisaker, D. (1981), "Asymptotically Optimum Kernels for Density Estimation at a Point," *Annals of Statistics*, 9, 334–346.

Stone, C. J. (1977), "Consistent Nonparametric Regression," *Annals of Statistics*, 5, 595–620.

Watson, G. S. (1964), "Smooth Regression Analysis," *Sankhyā*, Ser. A, 26, 359–372.

Zhao, L. H. (1993), "Frequentist and Bayesian Aspects of Some Nonparametric Estimation Problems," Ph.D. thesis, Cornell University.