On Combining Independent Nonparametric Regression Estimators

by

Patrick D. Gerard
Experimental Statistics Unit, Box 9653
Mississippi State, MS 39762-9653

William R. Schucany
Department of Statistical Science
Southern Methodist University
Dallas, TX 75275-0332

Technical Report No. SMU/DS/TR-274

On Combining Independent Nonparametric Regression Estimators

Patrick D. GERARD
Experimental Statistics Unit, Box 9653, Mississippi State, MS 39762-9653, USA

William R. SCHUCANY
Department of Statistical Science, Southern Methodist University, Dallas, TX 75275-0332, USA

Abstract: Three estimators are investigated for linearly combining independent nonparametric regression estimators. Assuming fixed designs, the asymptotic mean squared errors and asymptotically optimal bandwidths are given for each estimator and compared. One estimator essentially ignores the differences in the sources and naively pools all of the data. The second utilizes individually optimized bandwidths and then estimates the best weights to combine them. The third estimator solves a general minimization problem and employs equal bandwidths and weights similar to those for combining unbiased estimators with unequal variances. It is found to be superior to the other two in most situations that would be encountered in practice.

Key Words:asymptotic optimality,bandwidth,kernel,local linear

1.Introduction

Suppose that we have data on measurements of cholesterol (y_i) and the weights (x_i) for some subjects. Consider estimating some unknown smooth relationship between them, often denoted by

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

Details of this model will be given in the next section. There are various ways to estimate the curve, $m(x)$, without forcing it to belong to some restrictive parametric class. For good discussions of nonparametric regression estimators using splines and kernel methods the reader should consult Eubank (1988), Müller (1988), Wahba (1990), or Härdle (1990). There is a rich literature on this general problem. Relatively little has been published on data sets from different sources.

Only a few journal articles can be found on the subject of combining independent nonparametric regression estimators. Härdle and Marron (1990) approach the problem from a semiparametric perspective. Hart and Wehrly (1986) do not assume independence of the observations within a group and do not take full advantage of independence if it is present. In this article, we propose and compare three estimators that linearly combine data from two sources, possibly with different variances, to form a nonparametric curve estimator. These estimators may be based on either kernel regression estimators or locally weighted linear regression estimators. The biases of these estimators depend upon the amount of smoothing, which is controlled by window width or bandwidth. To minimize the usual optimality criterion of asymptotic mean squared error (AMSE) requires a balancing of bias^2 and

variance. Both of these depend on the bandwidth. The details of these three estimators are derived in Section 2.

The first of these estimators, $\hat{m}_N(t)$, follows the naive approach of disregarding the fact that the data come from two sources and of proceeding with locally weighted linear regression as though one large dataset were available. That such an approach will be employed becomes increasingly likely as nonparametric curve estimators find their way into widely used computer packages. This naive estimator is equivalent, in terms of AMSE, to a linear combination of two nonparametric regression estimators, each employing a common bandwidth. The individual estimators have weights proportional to the product of their respective sample sizes and design densities. Since the variances are not involved in the weights, it is not possible for this estimator to reduce the influence of a more variable data set and thus it performs poorly when one group has a much larger variance.

The second of these estimators, $\hat{m}_O(t)$, is also a linear combination of estimators from each group. However, the asymptotically optimal bandwidth for each individual estimator is used and then the weights are obtained by minimizing the associated AMSE of that combination.

The third estimator, $\hat{m}_E(t)$, results from solving the more general problem of minimizing the AMSE of a linear combination of nonparametric regression estimators simultaneously with respect to the weighting factor and the two bandwidths employed by the estimators. Equal bandwidths and a weighting factor that is proportional to the inverse of the variance divided by the product of the sample size and design density yield a local minimum AMSE.

The AMSE of $\hat{m}_E(t)$ is never greater than that of $\hat{m}_N(t)$. Equality occurs when the ratio of the respective variance divided by the product of sample size and design density is unity. In particular, this includes the special balanced case of equal variances and an equal number of design points from the same design density for each group. Comparison of $\hat{m}_E(t)$ with $\hat{m}_O(t)$ is facilitated by the fact that the ratio of their respective AMSE's can be written as a function of the ratio specified above. Detailed comparisons of the three AMSE's are developed in Section 3. Since $\hat{m}_E(t)$ has a smaller AMSE for most situations seen in practice, it is recommended unless there is evidence of extreme imbalance in the variances, sample sizes or design densities of the two groups.

2. Methodology

2.1 Background

Nonparametric regression methods such as kernel regression and locally weighted linear regression have become a reasonable choice for data analysts to estimate a curve without specifying a functional form. The responses, y_i, are assumed to be related to the explanatory variables, x_i, by

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

with ε_i being independent, identically distributed random variables having zero mean and constant variance σ^2. The x_i are fixed values satisfying $x_i = F^{-1}\left(\frac{i}{n+1}\right)$, where $F(\cdot)$ is an absolutely continuous cumulative distribution function with corresponding density $f(\cdot)$ known as the design density. These values are typically taken to be in (0,1). Usually a certain amount of smoothness is assumed for $m(\cdot)$.

Gasser and Müller (1979) proposed

$$\hat{m}_{GM}(t) = \frac{1}{h}\sum_{i=1}^{n} y_i \int_{s_{i-1}}^{s_i} K\left(\frac{u-t}{h}\right)du,$$

with $s_0=0$, $s_i=(x_{i-1}+x_i)/2$, and $s_n=1$ as an estimator of $m(t)$. The weight function, $K(\cdot)$, is typically a second order kernel function supported on [-1,1] and h is a bandwidth governing the smoothness of the estimator. Larger values of h produce smoother curves, but the bias in $\hat{m}(t)$ is greater as a result.

Locally weighted linear regression estimators (Fan,1992) have gained popularity to some extent because of their improved performance in boundary regions (near the edges of their available data) compared to kernel estimators. The form of these "local linear" estimators is

$$\hat{m}_{LL}(t;h) = \frac{\sum_{i=1}^{n} K\left(\frac{z_i}{h}\right) z_i^2 \sum_{i=1}^{n} K\left(\frac{z_i}{h}\right) y_i - \sum_{i=1}^{n} K\left(\frac{z_i}{h}\right) z_i \sum_{i=1}^{n} K\left(\frac{z_i}{h}\right) z_i y_i}{\sum_{i=1}^{n} K\left(\frac{z_i}{h}\right) z_i^2 \sum_{i=1}^{n} K\left(\frac{z_i}{h}\right) - \left(\sum_{i=1}^{n} K\left(\frac{z_i}{h}\right) z_i\right)^2}, \qquad (2.1)$$

where $z_i=x_i-t$.

Performance of both of these estimators is typically gauged by AMSE. For interior estimation ($h < t < 1-h$) the AMSE is (Jones et. al.,1994),

$$AMSE\left[\hat{m}_{LL}(t;h)\right] = AMSE\left[\hat{m}_{GM}(t;h)\right] = \left[\frac{1}{2}k_2 m''(t)h^2\right]^2 + \frac{\sigma^2 Q}{nhf(t)},$$

where $k_2 = \int_{-1}^{1} u^2 K(u)du$ and $Q = \int_{-1}^{1} K^2(u)du$. The only assumptions required are that $n\to\infty$ and $h\to 0$ such that $nh\to\infty$ and some continuity of $m''(t)$. Consequently, our findings in this paper for local linear regression with weight function $K(\cdot)$ apply immediately for kernel regression with that same kernel.

When data are available from two sources, it may be reasonable to assume that the $j^{\underline{th}}$ observation in the $i^{\underline{th}}$ group, y_{ij}, is related to the corresponding explanatory variable , x_{ij}, through the same smooth mean function, i.e.,

$$y_{ij} = m(x_{ij}) + \varepsilon_{ij}, \qquad i = 1,2, j = 1,\ldots,n_i,$$

where the ε_{ij} are independent and identically distributed within the $i^{\underline{th}}$ group with zero mean and variance σ_i^2. Again, the x_{ij} are fixed variables satisfying

$$x_{ij} = F_i^{-1}\left(\frac{j}{n_i + 1}\right),$$

where $F_i(\cdot)$ is an absolutely continuous cumulative distribution function with corresponding density $f_i(\cdot)$. Three methods for estimating m(·) are derived in the next three sections.

Rao and Subrahmaniam (1971) investigate combining independent means and linear regressions. They limit their study to unbiased estimators, namely WLS and MINQUE. Their regression model entails different variances at each x_i rather than across sources. However, their more elementary problem of estimating a common mean has more potential relevance to our task of estimating $m(t)$. Even though Rao and Subrahmaniam (1971) investigate a different class of estimators, they find greater efficiency for the estimator that "smooths" the individual sample variances within each group (s_i^2). This is analogous to a general justification that is sometimes put forward for nonparametric regression, namely "borrowing" information about location from neighbors. We are not recommending shrinkage estimators of σ_i^2, although that may merit investigation.

The Associate Editor identified two papers that consider nonparametric curve estimation from data that are not all from a single source. Hart and Wehrly (1993) consider continuous time Gaussian processes with unknown covariance function. They establish consistency of cross validation by deleting one curve at a time. This is essentially the subject of Rice and Silverman (1991), although they address a practical application involving growth curves. Again, there are numerous groups of dependent data. Hence both papers address data structures that are substantially different than we do here. They have unknown covariance structure, but replication from a common population of curves. By contrast, we assume independent sequences of disturbances (ε_{ij}), but allow the possibility of different variances ($\sigma_1^2 \neq \sigma_2^2$).

2.2 A Naive Estimator

One possible estimator of $m(t)$ follows from naively disregarding the fact that the data are from two sources and proceeding with locally weighted linear regression as though only one "pooled" dataset were available. If the usual assumptions of $n_1 \to \infty$, $n_2 \to \infty$, and $h \to 0$ such that $n_1 h \to \infty$ and $n_2 h \to \infty$ are supplemented with $n_1/n_2 \to r$ $(0 < r < \infty)$, then using standard integral approximations and Taylor series expansions the expressions for the asymptotic bias and variance of this estimator, $\hat{m}_N(t)$, are

$$Bias[\hat{m}_N(t)] = \frac{1}{2}k_2 m''(t)h^2 + o(h^2)$$

and

$$Var[\hat{m}_N(t)] = \frac{Q}{h}\left[\frac{\sigma_1^2 n_1 f_1(t) + \sigma_2^2 n_2 f_2(t)}{(n_1 f_1(t) + n_2 f_2(t))^2}\right] + o\left(\frac{1}{n_1 h}\right).$$

Details of these derivations may be found in Gerard (1993).

Introducing notation for an equivalent variance, namely

$$\sigma_N^2 = \frac{\sigma_1^2 n_1 f_1(t) + \sigma_2^2 n_2 f_2(t)}{(n_1 f_1(t) + n_2 f_2(t))^2},$$

the asymptotically optimal bandwidth, h_N has the familiar form

$$h_N = \left[\frac{\sigma_N^2 Q}{(k_2 m''(t))^2}\right]^{1/5}.$$

This estimator is easily seen to be equivalent, in terms of AMSE, to

$$\hat{m}^*(t;h) = \frac{n_1 f_1(t)}{n_1 f_1(t) + n_2 f_2(t)}\hat{m}_1(t;h) + \frac{n_2 f_2(t)}{n_1 f_1(t) + n_2 f_2(t)}\hat{m}_2(t;h),$$

where $\hat{m}_i(t;h)$ is the locally weighted linear regression estimator (2.1) from the $i^{\underline{th}}$ group. As can be seen from the expression for $\hat{m}^*(t;h)$, this estimator essentially weights the individual estimators by the amount of data available to each and does not involve the variances σ_i^2.

2.3 Estimating $m(t)$ using Individual Optimal Bandwidths

A more reasonable method of estimation entails linearly combining estimators from each group, each of which employs the asymptotically optimal bandwidth for that respective group. The combination is selected that minimizes the AMSE of the resulting estimator. That is, first form the class of estimators

$$\hat{m}_O(t;c) = c\hat{m}_1(t;hopt_1) + (1-c)\hat{m}_2(t;hopt_2), \qquad i = 1,2,$$

where $hopt_i$ is the asymptotically optimal bandwidth for the $i^{\underline{th}}$ group,

$$hopt_i = \left[\frac{\sigma_i^2 Q}{(k_2 m''(t))^2 n_i f_i(t)} \right]^{1/5} \quad , i=1,2.$$

It follows that the value of c that minimizes $AMSE\left[\hat{m}_O(t;c)\right]$ is easily obtained by differentiation to be

$$c_O = \frac{5 - k^{2/5}}{5k^{4/5} - 2k^{2/5} + 5},$$

$$\text{where } k = \frac{\sigma_1^2}{n_1 f_1(t)} \Big/ \frac{\sigma_2^2}{n_2 f_2(t)}. \tag{2.2}$$

This is the ratio that was mentioned in the introduction. Use of this value of c yields $\hat{m}_O(t)=\hat{m}_O(t;c_O)$ and

$$AMSE\left[\hat{m}_O(t)\right] = AMSE\left[\hat{m}_2(t;hopt_2)\right]\left[1 - \frac{1}{5}\frac{\left(5 - k^{2/5}\right)^2}{\left(5k^{4/5} - 2k^{2/5} + 5\right)}\right].$$

This expression of the AMSE is especially useful in comparisons with the estimator introduced in the next section.

2.4 Equal Bandwidth Estimator

In the previous section, an estimator was derived by minimizing the AMSE of a linear combination of local linear regression estimators with each of the bandwidths fixed at their asymptotically optimal values. If, instead, the bandwidths as well as the weighting factor are allowed to vary, then a more general minimization problem arises. The next two theorems provide solutions to two such minimization problems. In the first, the resulting two bandwidths are equal, but all that can be guaranteed is a local minimum. In the second, by adding the constraint that the bandwidths are equal, the same solution can be shown to produce a global minimum. Both proofs may be found in the Appendix.

<u>Theorem 1</u>

Let $\hat{m}(t;c,h_1,h_2) = c\hat{m}_1(t;h_1) + (1-c)\hat{m}_2(t;h_2)$, where $\hat{m}_i(t;h_i)$ is the local linear regression estimator for the $i^{\underline{th}}$ group employing bandwidth h_i. The values

$$c = c_{opt} = \frac{\dfrac{\sigma_2^2}{n_2 f_2(t)}}{\dfrac{\sigma_1^2}{n_1 f_1(t)} + \dfrac{\sigma_2^2}{n_2 f_2(t)}} \tag{2.3}$$

and

$$h_1 = h_2 = h_E = \left[\frac{\dfrac{\sigma_1^2}{n_1 f_1(t)} \dfrac{\sigma_2^2}{n_2 f_2(t)} Q}{\left(\dfrac{\sigma_1^2}{n_1 f_1(t)} + \dfrac{\sigma_2^2}{n_2 f_2(t)} \right) \left(k_2 m''(t) \right)^2} \right]^{1/5} \tag{2.4}$$

produce a local minimum in $AMSE\left[\hat{m}(t;c,h_1,h_2)\right]$.

The next theorem considers the closely related minimization problem with the bandwidths required to be equal. The minimizers are the same as above and the minimum can be proven to be a global one.

<u>Theorem 2</u>

Let $\hat{m}(t;c,h) = c\hat{m}_1(t;h) + (1-c)\hat{m}_2(t;h)$, where $\hat{m}_i(t;h)$ is as in Theorem 1. The values of c and h given by (2.3) and (2.4), respectively produce a global minimum $AMSE\left[\hat{m}(t;c,h)\right]$.

Using these minimizers, our third estimator is denoted by

$$\hat{m}_E(t) = \frac{\dfrac{\sigma_2^2}{n_2 f_2(t)}}{\dfrac{\sigma_1^2}{n_1 f_1(t)} + \dfrac{\sigma_2^2}{n_2 f_2(t)}} \hat{m}_1(t;h_E) + \frac{\dfrac{\sigma_1^2}{n_1 f_1(t)}}{\dfrac{\sigma_1^2}{n_1 f_1(t)} + \dfrac{\sigma_2^2}{n_2 f_2(t)}} \hat{m}_2(t;h_E).$$

The AMSE surface as described in Theorem 1 is not convex and hence does not readily divulge a global minimum. The use of $\hat{m}_E(t)$ yields a local minimum. However, $\hat{m}_E(t)$ also represents the global minimization described in Theorem 2. That $\hat{m}_E(t)$ produces the global minimum AMSE in a constrained, though reasonable, class of estimators, as well as a local minimum in a more general class, lends credence to its use as a viable estimator. The search for smaller minima in the unconstrained problem consistently led us into regions with bandwidths that were outside the region of validity of the asymptotic expansions.

The asymptotic mean squared error of $\hat{m}_E(t)$ can be shown to be

$$AMSE\left[\hat{m}_E(t)\right] = \frac{5}{4}\left[\frac{\dfrac{\sigma_1^2}{n_1 f_1(t)}\dfrac{\sigma_2^2}{n_2 f_2(t)}}{\dfrac{\sigma_1^2}{n_1 f_1(t)}+\dfrac{\sigma_2^2}{n_2 f_2(t)}}Q\right]^{4/5}\left(k_2 m''(t)\right)^{2/5}$$

$$= AMSE\left[\hat{m}_2(t;h_2)\right]\left[\frac{k}{k+1}\right]^{4/5}, \tag{2.5}$$

where k is as in (2.2). The last form of $AMSE\left[\hat{m}_E(t)\right]$ will facilitate comparisons with $\hat{m}_O(t)$. Note that the dependence on m through $m''(t)$ cancels in ratios of these AMSE's.

The essence of $\hat{m}_E(t)$ is that equal bandwidths equalize the asymptotic biases of the two component estimators and then the value of c is the familiar one that minimizes asymptotic variance. This results in each term in $\hat{m}_E(t)$ having equal variance. An efficiency comparison of these three estimators, through ratios of their AMSE's is the subject of the next section.

3. Comparison of Estimators

In this section, asymptotic relative efficiency (ARE) comparisons of these estimators will be made through ratios of minimum AMSE's. If these ratios are to have the traditional sample size interpretation, then they need to be raised to the 5/4 power. Only one estimator is uniformly better than another. The "naive" is virtually dominated by the "equal bandwidth" approach, i.e. the minimized AMSE of $\hat{m}_N(t)$ is as least as large as that of $\hat{m}_E(t)$ for every k. This is easily seen since $\hat{m}_N(t)$ is asymptotically equivalent to a linear combination of local linear regression estimators using a common bandwidth but not necessarily equal to $\hat{m}_E(t)$. From Theorem 2, $\hat{m}_E(t)$ minimizes the AMSE of a class of estimators of this type. Thus $\hat{m}_N(t)$ is suboptimal unless its coefficients and common bandwidth happen to equal (2.3) and (2.4), respectively.

To simplify comparisons of $\hat{m}_N(t)$ to the nondominating $\hat{m}_O(t)$, take $n_1=n_2$ and $f_1(x)=f_2(x)=1$, the uniform density on [0,1]. Thus, each group has n equally spaced design points on [0,1]. The ratio of AMSE's in this case can be expressed as a function of k, which reduces to σ_1^2/σ_2^2. It follows that

$$\frac{AMSE\left[\hat{m}_N(t)\right]}{AMSE\left[\hat{m}_O(t)\right]} = \frac{(.3299)(k+1)^{4/5}}{\left[1-\dfrac{1}{5}\dfrac{\left(5-k^{2/5}\right)^2}{5k^{4/5}-2k^{2/5}+5}\right]} = ARE_1(k).$$

This expression is plotted as a function of k in Figure 1. Because of symmetry, only values of k between 0 an 1 need to be investigated. Note that $\hat{m}_N(t)$ has a smaller AMSE than $\hat{m}_O(t)$ for values of k near 1. Using Gauss-Newton methods to find roots of $ARE_1(k)=1$, we find that for $k > 1.6$ or $k < 1/1.6$, $\hat{m}_O(t)$ has the smaller ASME. Hence in this special case, pooling is more efficient unless one variance is 60% greater than the other.

Finally, the ARE of $\hat{m}_E(t)$ relative to $\hat{m}_O(t)$,

$$\frac{AMSE\left[\hat{m}_O(t)\right]}{AMSE\left[\hat{m}_E(t)\right]} = \frac{\left[1-\frac{1}{5}\frac{\left(5-k^{2/5}\right)^2}{5k^{4/5}-2k^{2/5}+5}\right]}{\left[\frac{k}{k+1}\right]^{4/5}} = ARE_2(k),$$

is plotted as a function of k in Figure 2. Again using Gauss-Newton methods to find roots of $ARE_2(k)=1$, it follows that $\hat{m}_E(t)$ has a smaller AMSE for $1/161.08 < k < 161.08$. For $k=1$, $AMSE\left[\hat{m}_O(t)\right]$ is 4.5% larger than $AMSE\left[\hat{m}_E(t)\right]$ and as $k \to 0$, $AMSE\left[\hat{m}_E(t)\right]$ is 4% larger than $AMSE\left[\hat{m}_O(t)\right]$. Note that in this comparison, the general form for k in (2.2) is applicable. In other words, this conclusion hold generally and not for the special uniform example.

We have some experience with these estimators on simulated finite samples. The observed relative efficiencies were in generally good agreement with those predicted by the ARE calculations. The simple MISE for $\hat{m}_E$ was consistently smaller than that of $\hat{m}_O$ for combinations of $m, f_1, f_2, \sigma_1, \sigma_2, n_1$, and n_2 that we investigated. The functions were linear combinations of exponentials and quadratic; the design densities were uniform and truncated exponential; the standard deviations 10% to 25% of the range of m; and the sample sizes were 100 and 200. A full report of that study, which incorporated data based bandwidth selection, is available from the authors.

4. Conclusions

Of the three estimators investigated for combining independent nonparametric regression estimators, $\hat{m}_E(t)$, an estimator with equal bandwidths and weights proportional to $\left(\frac{\sigma_i^2}{n_i f_i(t)}\right)^{-1}$ is superior in most instances that would be seen in practice. An analogous result for more than two samples should be reasonably straightforward. It is always at least as efficient as $\hat{m}_N(t)$, an estimator that ignores the fact that the data were from two sources. Additionally, it is more efficient than $\hat{m}_O(t)$, an estimator that optimally combines each individual estimator using its asymptotically optimal bandwidth, in virtually all cases that would be encountered in practice. Hence, unless there is a drastic difference in sample sizes, design densities, or variances, $\hat{m}_E(t)$ may be the recommended estimator.

As a practical matter, one must estimate the unknown bandwidths that minimize AMSE. The adaptive choice of good bandwidths is typically quite challenging. For the three types of estimators examined in this paper, bandwidth selection differs somewhat. Consequently, the comparisons of the three estimators with data-based bandwidths may produce somewhat different small sample efficiencies.

5. Appendix

5.1 Proof of Theorem 1

The asymptotic mean squared error of $\hat{m}(t;c,h_1,h_2)$ is

$$AMSE\left[\hat{m}(t;c,h_1,h_2)\right]=\left[\frac{1}{2}k_2 m''(t)(ch_1^2+(1-c)h_2^2)\right]^2+\frac{c^2\sigma_1^2 Q}{n_1 f_1(t)h_1}+\frac{(1-c)^2\sigma_2^2 Q}{n_2 f_2(t)h_2}.$$

Letting $\mathbf{s}'=(c,h_1,h_2)$, set

$$E(\mathbf{s})=AMSE\left[\hat{m}(t;c,h_1,h_2)\right].$$

Though it takes much tedious algebra, it can be shown that

$$\frac{\partial E(\mathbf{s})}{\partial s_i}=0, \qquad i=1,2,3,$$

is satisfied for $\mathbf{s}'_m=(c_{opt},h_E,h_E)$. Hence $\mathbf{s}_m$ is a candidate for producing a local minimum.

Define the Hessian matrix of partial derivatives as

$$\mathbf{P}=[p_{ij}]=\frac{\partial^2 E(\mathbf{s})}{\partial s_i \partial s_j}, \quad i,j=1,2,3,$$

and the matrix $\mathbf{D}(\mathbf{s}_m)=[d_{ij}(\mathbf{s}_m)]$ as the Hessian matrix $\mathbf{P}$ evaluated at $\mathbf{s}=\mathbf{s}_m$. Additionally, define

$$v_i=\frac{\sigma_i^2}{n_i f_i(t)}, \qquad i=1,2.$$

The first principal minor of $\mathbf{D}(\mathbf{s}_m)$ is strictly positive, since

$$d_{11}(\mathbf{s}_m) = 2Q^{4/5}(k_2 m''(t))^{2/5}\left[\frac{v_1^{4/5}}{\left(\frac{v_2}{v_1+v_2}\right)^{1/5}} + \frac{v_2^{4/5}}{\left(\frac{v_1}{v_1+v_2}\right)^{1/5}}\right] > 0.$$

The second principal minor of $\mathbf{D}(\mathbf{s}_m)$ is also strictly positive, as

$$d_{11}(\mathbf{s}_m)d_{22}(\mathbf{s}_m) - (d_{12}(\mathbf{s}_m))^2 =$$

$$Q^{6/5}(k_2 m''(t))^{8/5}\left[\begin{array}{l} 4v_1^{6/5} + \left(\frac{v_2}{v_1+v_2}\right)^{11/5} + 4\frac{v_2^{16/5}v_1^{1/5}}{(v_1+v_2)^{4/5}} \\ +5\left(\frac{v_1 v_2}{v_1+v_2}\right)^{6/5} + 11\frac{v_2^{11/5}v_1^{1/5}}{(v_1+v_2)^{6/5}} \end{array}\right] > 0.$$

The third principal minor of $\mathbf{D}(\mathbf{s}_m)$, is the determinant

$$\begin{array}{l} d_{11}(\mathbf{s}_m)d_{22}(\mathbf{s}_m)d_{33}(\mathbf{s}_m) - d_{11}(\mathbf{s}_m)(d_{23}(\mathbf{s}_m))^2 - (d_{12}(\mathbf{s}_m))^2 d_{33}(\mathbf{s}_m) \\ +2d_{12}(\mathbf{s}_m)d_{13}(\mathbf{s}_m)d_{23}(\mathbf{s}_m) - (d_{13}(\mathbf{s}_m))^2 d_{22}(\mathbf{s}_m) \end{array},$$

which is a very complicated expression. Using Maple® (Version V) on a Sun Workstation the expression simplifies to the strictly positive polynomial,

$$\frac{Q^{8/5}(k_2 m''(t))^{4/5} v_2^{8/5}}{(1+B)^{38/5}}\left[25B^6 + 150B^5 + 375B^4 + 500B^3 + 375B^2 + 150B\right],$$

where $B = \frac{v_2}{v_1}$.

Thus $\mathbf{s}=\mathbf{s}_m$ produces a local minimum AMSE as was to be shown.

5.2 Proof of Theorem 2.

The asymptotic mean squared error of $\hat{m}(t;c,h)$ is

$$AMSE\left[\hat{m}(t;c,h)\right] = \left[\frac{1}{2}k_2 m''(t)h^2\right]^2 + \frac{c^2\sigma_1^2 Q}{n_1 f_1(t)h} + \frac{(1-c)^2\sigma_2^2 Q}{n_2 f_2(t)h}.$$

Letting s=(c,h), the quantity to be minimized is

$$E(\mathbf{s}) = AMSE\left[\hat{m}(t;c,h)\right] = \left[\frac{1}{2}k_2 m''(t)h^2\right]^2 + \frac{c^2\sigma_1^2 Q}{n_1 f_1(t)h} + \frac{(1-c)^2\sigma_2^2 Q}{n_2 f_2(t)h}.$$

Though tedious, it is straightforward to show that $\mathbf{s}_m = (c_{opt}, h_E)$ satisfies

$$\frac{\partial E(\mathbf{s})}{\partial s_i} = 0, \qquad i = 1,2.$$

It remains to show that $E(\mathbf{s})$ is a convex function of s. The first term of $E(\mathbf{s})$ is obviously a convex function of h. Letting the last two terms of $E(\mathbf{s})$ equal $V(\mathbf{s})$, then

$$\frac{\partial^2 V(\mathbf{s})}{\partial s_1^2} = \frac{2\sigma_1^2 Q}{n_1 f_1(t)h} + \frac{2\sigma_2^2 Q}{n_2 f_2(t)h} > 0$$

and the determinant of the relevant Hessian matrix is

$$\frac{\partial^2 V(\mathbf{s})}{\partial s_1^2}\frac{\partial^2 V(\mathbf{s})}{\partial s_2^2} - \left(\frac{\partial^2 V(\mathbf{s})}{\partial s_1 \partial s_2}\right)^2 = \frac{4\sigma_1^2\sigma_2^2 Q^2}{n_1 f_1(t) n_2 f_2(t) h^4} > 0.$$

Thus $V(\mathbf{s})$ is convex and hence so is $E(\mathbf{s})$. Therefore s=$\mathbf{s}_m$ results in an absolute minimum AMSE and the proof is complete.

References

Eubank, R.L. (1988), *Spline Smoothing and Nonparametric Regression*, New York: Marcel Dekker.

Fan, J. (1992), Design-adaptive Nonparametric Regression, *The Journal of the American Statistical Association 87*, 998-1004.

Gasser, Th. and Müller, H.G. (1979), Kernel Estimation of Regression Functions, in *Smoothing Techniques for Curve Estimation* (Th. Gasser and M. Rosenblatt, eds.), 23-68. Heidelberg: Springer.

Gerard, P.D. (1993), *Combining Independent Nonparametric Regression Estimators,* Ph.D. Dissertation, Department of Statistical Science, Southern Methodist University.

Härdle, W. (1990), *Applied Nonparametric Regression,* New York: Oxford University Press.

Härdle, W. and Marron, J.S. (1990), Semiparametric Comparison of Regression Curves, *The Annals of Statistics 18*, 63-89.

Hart, J.D. and Wehrly, T.E. (1986), Kernel Regression Estimation Using Repeated Measurements Data, *Journal of the American Statistical Association 81*, 1080-1088.

Hart, J.D. and Wehrly, T.E. (1993), Consistency of Cross-Validation When the Data are Curves, *Stochastic Processes and Their Applications 45*, 351-361.

Jones,M.C., Davies, S.J., and Park,B.U. (1994), Versions of Kernel-type Regression Estimators, *Journal of the American Statistical Association 89*, 825-832.

Müller, H.G. (1988), *Nonparametric Regression Analysis of Longitudinal Data*, New York: Springer-Verlag.

Rao, J.N.K. and Subrahmaniam, K. (1971), Combining Independent Estimators and Estimation in Linear Regression with Unequal Variances, *Biometrics 27*, 971-990.

Rice, J.A. and Silverman, B.W. (1991), Estimating the Mean and Covariance Structure Nonparametrically When the Data are Curves, *Journal of the Royal Statistical Society B 53*, 233-244.

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia: SIAM.

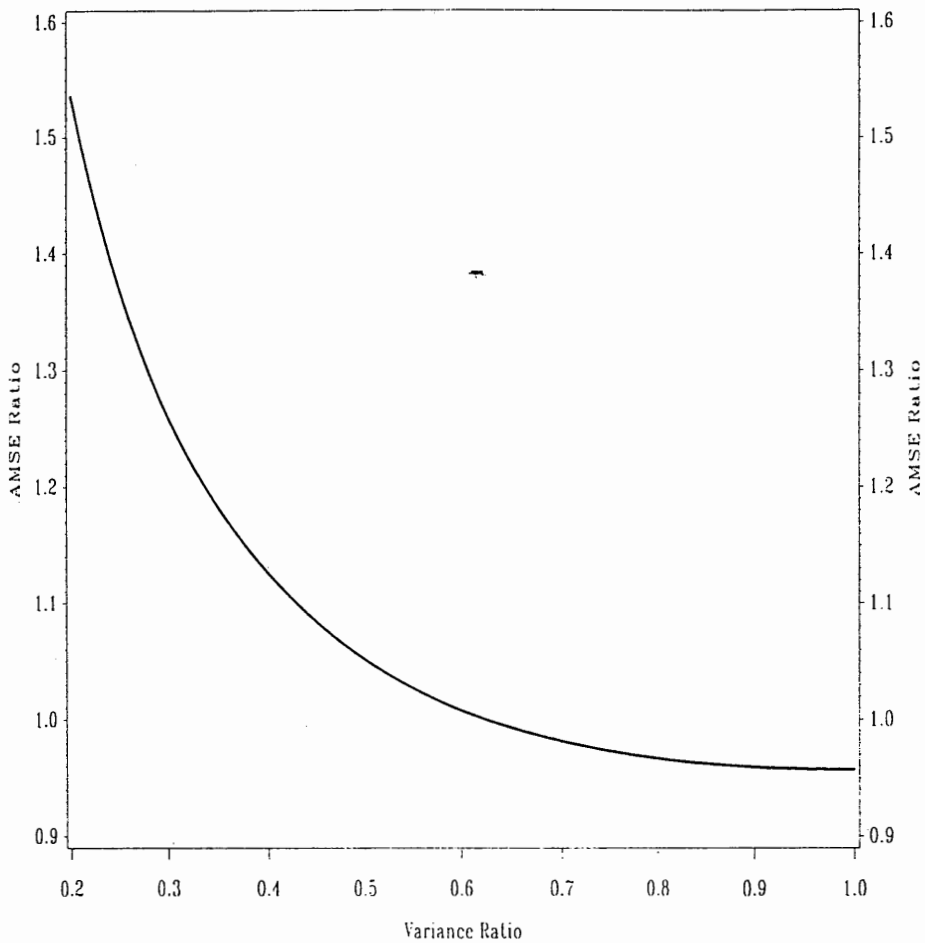

Figure 1. Asymptotic Relative Efficiency of $\hat{m}_O$ relative to the naive $\hat{m}_N$ as a function of $k = \sigma_1^2 / \sigma_2^2$.

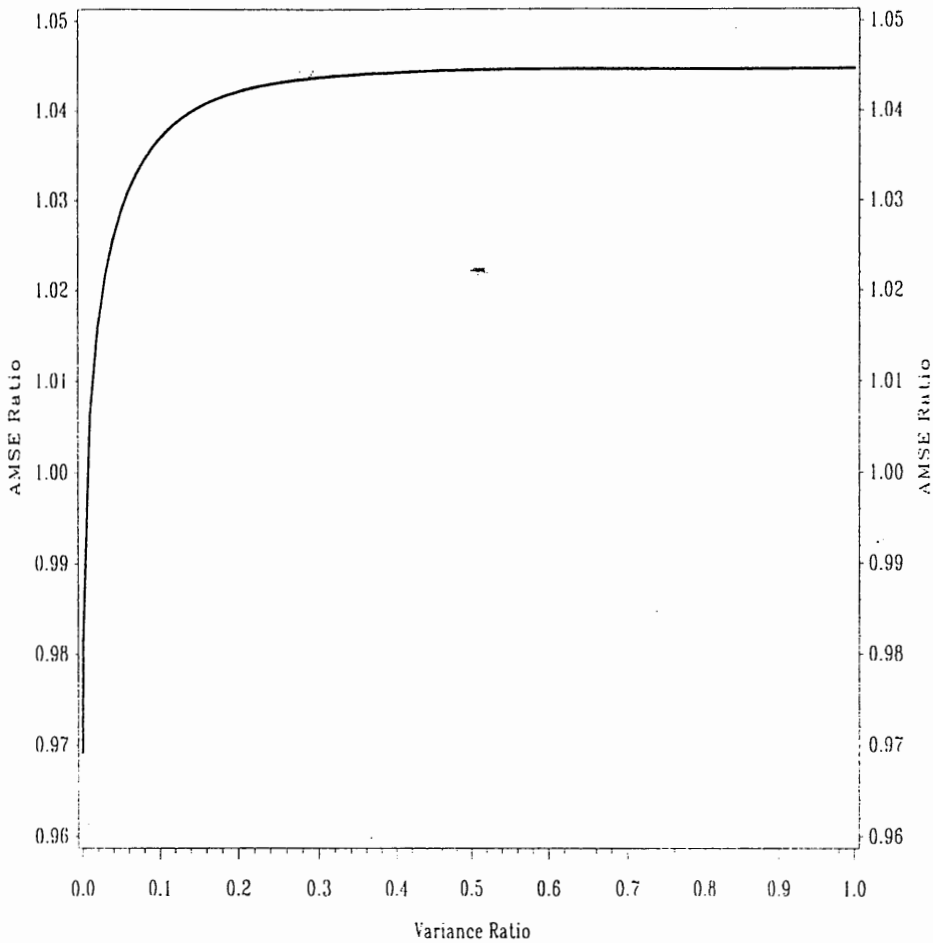

Figure 2. Asymptotic Relative Efficiency of $\hat{m}_E$ relative to $\hat{m}_O$ as a function of $k = \sigma_1^2 / \sigma_2^2$.