The Annals of Statistics
1995, Vol. 23, No. 5, 1443–1473

NONPARAMETRIC ESTIMATION OF GLOBAL FUNCTIONALS AND A MEASURE OF THE EXPLANATORY POWER OF COVARIATES IN REGRESSION

BY KJELL DOKSUM[1] AND ALEXANDER SAMAROV[2]

University of California, Berkeley and Massachusetts Institute of Technology

In a nonparametric regression setting with multiple random predictor variables, we give the asymptotic distributions of estimators of global integral functionals of the regression surface. We apply the results to the problem of obtaining reliable estimators for the nonparametric coefficient of determination. This coefficient, which is also called Pearson's correlation ratio, gives the fraction of the total variability of a response that can be explained by a given set of covariates. It can be used to construct measures of nonlinearity of regression and relative importance of subsets of regressors, and to assess the validity of other model restrictions. In addition to providing asymptotic results, we propose several data-based bandwidth selection rules and carry out a Monte Carlo simulation study of finite sample properties of these rules and associated estimators of explanatory power. We also provide two real data examples.

1. Introduction. For regression experiments where the relationship between a random covariate vector $\mathbf{X}$ and a response variable Y does not necessarily follow either a linear or other specified parametric model, a natural measure of the strength of the relationship between $\mathbf{X}$ and Y is Pearson's correlation ratio

$$\eta^2 = \frac{\mathrm{Var}(m(\mathbf{X}))}{\mathrm{Var}(Y)}, \tag{1.1}$$

where $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$, $\mathbf{X} \in R^d$, $Y \in R^1$. The correlation ratio η^2 is based on the ANOVA decomposition

$$\mathrm{Var}(Y) = \mathrm{Var}(m(\mathbf{X})) + E(\sigma^2(\mathbf{X})), \tag{1.2}$$

where $\sigma^2(\mathbf{x}) = \mathrm{Var}(Y|\mathbf{X} = \mathbf{x})$, and thus gives the fraction of the variability of Y which is explained with the best predictor based on $\mathbf{X}$, $m(\mathbf{X})$, and can be interpreted as a nonparametric coefficient of determination or nonparametric

Received June 1993; revised February 1995.
[1]Partially supported by NSF Grants DMS-91-06752 and DMS-93-07403.
[2]Partially supported by NSF Grants DMS-90-01523 and DMS-93-06245 and by IFSRC at MIT.
AMS 1991 *subject classifications.* Primary 62J02; secondary 62G99.
Key words and phrases. Nonparametric R-squared, Pearson's correlation ratio, integral regression functionals, measure of subset importance, index of nonlinearity, bandwidth selection, cross-validation.

R-squared. It can also be defined via the extremal correlation property

$$\eta^2 = \mathrm{Corr}^2(m(\mathbf{X}), Y) = \sup_g \mathrm{Corr}^2(g(\mathbf{X}), Y), \tag{1.3}$$

where the supremum is taken over all real-valued functions $g(\mathbf{X})$ with finite second moments. Equation (1.3) is easily proved using the iterated expectation property and the Cauchy–Schwarz inequality.

The quantity η^2 can also be interpreted in terms of signal-to-noise ratio, which is usually defined as the variability (or energy) of the signal $\mathbf{X}$ over that of the noise $\varepsilon = Y - m(\mathbf{X})$:

$$\frac{\text{signal}}{\text{noise}} = \frac{\eta^2}{1 - \eta^2}.$$

Restricting for a moment attention to the case $d = \dim(\mathbf{X}) = 1$, we note that $\eta^2 = \eta^2_{xy}$ is an asymmetric measure. In fact, it is possible that $\eta^2_{xy} = 1$, while $\eta^2_{yx} < 1$. Asymmetry of η^2 reflects the fact that it is a regression rather than correlation measure of association. As such, it avoids some of the "pathologies" of ACE and the maximum correlation coefficient [see Rényi (1959), Breiman and Friedman (1985) and Buja (1990)].

The quantity η^2 is not a "strong" measure of association: while independence of X and Y clearly implies $\eta^2 = 0$, even $\max(\eta^2_{xy}, \eta^2_{yx}) = 0$ does not imply that X and Y are independent. In fact, it is possible that $\max(\eta^2_{xy}, \eta^2_{yx}) = 0$ while X and Y are functionally dependent: just consider a uniform distribution on the unit circumference. The point, of course, is that X and Y may be dependent not through the conditional means but in many other ways. On the other hand, if $Y = m(\mathbf{X}) + \varepsilon$ with $\mathbf{X}$ and ε independent, $\eta^2 = 0$ is equivalent to independence of $\mathbf{X}$ and Y.

Unlike "strong" measures of dependence, such as maximum correlation coefficient or Rényi's (1959) mean square contingency, η^2 is not invariant with respect to arbitrary transformations of $\mathbf{X}$ and Y [cf. Rényi (1959), Axiom E]. It is invariant, however, with respect to affine transformations of $\mathbf{X}$ and Y and one-to-one transformations of $\mathbf{X}$.

It is well known that,unless the dimensionality d is very small or n is very large, reliable nonparametric estimation of the regression function $m(\mathbf{x})$ is impossible because of the sparsity of the data, or what is known as the curse of dimensionality. The main idea of the paper is to use more accurately estimable integral functionals of $m(\mathbf{x})$ to assess the explanatory power of covariates and the validity of various restrictions on the model, which allow for one or other form of dimensionality reduction.

In the nonparametric setup, estimates of the correlation ratio η^2 are quite sensitive to values of $\mathbf{X}$ near the boundary of its support $S_{\mathbf{X}}$. By introducing a weight function $w(\mathbf{X})$ which is equal to 1 in the central part of $S_{\mathbf{X}}$ and is zero near the boundary of $S_{\mathbf{X}}$, we get a more "robust" measure that focuses on the explanatory power of $\mathbf{X}$ without being too sensitive to values near the boundary. We thus consider the weighted functional η^2_w, defined in (2.1), with

a bounded nonnegative weight function $w(\mathbf{x})$ which reduces the influence of outlying $\mathbf{X}$ values, or high leverage points.

We consider the problem of nonparametric estimation of η_w^2 as well as of closely related measures of nonlinearity and of the relative importance of subsets of covariates. The problem is essentially that of estimating a particular type of functional of the joint distribution F of $(\mathbf{X}, Y)$. Following Le Cam (1956), Hasminskii and Ibragimov (1979) and Bickel and Ritov (1988), we consider a "one-step" estimator based on the influence function of the functional in question. The advantages of this approach are discussed in the recent monograph by Bickel, Klaasen, Ritov and Wellner (1993), where many additional references are given. We also consider two other estimators of η_w^2 obtained by considering sample counterparts of the weighted versions of (1.1) and (1.3) with plugged-in nonparametric regression estimators [see (2.13) and (2.14)]. We refer to these three estimators as the one-step type, the regression variance type and the correlation type, respectively.

To obtain the asymptotic distribution of the three estimators of η_w^2 as well as of the measures of nonlinearity and subset importance, we prove a theorem which establishes the asymptotic normality at the root-n rate of the one-step estimates of functionals of the form $S(F) = E\phi(\mathbf{X}, m(\mathbf{X}))$, where ϕ is a smooth real-valued function on R^{d+1}. This result, together with Theorem 1 of Samarov (1993), is shown in Sections 2 and 6 to yield the asymptotic normality at the root-n rate of all our estimators.

Our estimators of η_w^2 depend on the bandwidth h of the kernel estimator $\hat{m}(\mathbf{X}_i)$ of regression $m(\mathbf{X}_i)$. In Section 3 we consider a data-dependent rule for selecting h based on the maximization of the correlation-type estimator with a "leave-one-out" regression estimator together with the usual cross-validation rule. We summarize in that section the results of Monte Carlo comparisons of the estimators and find that the one-step and correlation-type estimators are rather stable with respect to bandwidth choice while the regression variance-type estimator is quite unstable. The correlation-type estimator is slightly more stable than the one-step type, but the difference between them is very small for the relevant range of h's.

In Section 4 we consider estimators of the functionals measuring the extent of nonlinearity of regression and of the relative importance of subsets of covariates, and we show their asymptotic normality at the root-n rate. Section 5 contains two real data examples. In particular, we analyze the famous Boston housing data in terms of the explanatory power of selected covariates, the degree of nonlinearity and the relative importance of subsets of covariates. Proofs are given in Section 6.

The functional η^2 was considered by Pearson (1905), Kolmogorov (1933), Cramér (1945), Kruskal (1958), Rényi (1959), Kendall and Stuart (1962) and Rao (1973), among others. Estimation of global, or integral, smooth functionals has been considered in a large number of works from both theoretical and applied points of view. Important theoretical results are obtained in Koshevnik and Levit (1976), Levit (1978), Hasminskii and Ibragimov (1979), Ibragimov, Nemirovsky and Khasminskii (1986), Hall and Marron (1987),

Bickel and Ritov (1988), Donoho and Nussbaum (1990), Donoho and Liu (1991), Fan (1991) and Goldstein and Messer (1992), where, in various settings, conditions for $n^{1/2}$-consistency, asymptotic normality and efficiency are established for nonparametric estimators of functionals of the integral type; see also the recent book by Bickel, Klaasen, Ritov and Wellner (1993).

In applications, nonparametric functional estimation has been recently used to assess various aspects of parametric and nonparametric models [see, e.g., Azzalini, Bowman and Härdle (1989), Härdle and Stoker (1989), Joe (1989), Eubank and Spielgelman (1990), Kozek (1991), Staniswalis and Severini (1991), Abramson and Goldstein (1991), Robinson (1991) and Samarov (1993)]. Another recent area of application is efficient smoothing parameter selection [see Fan and Marron (1992) and Hall and Johnstone (1992)].

Estimation of the functional η^2 is closely related to the estimation of the residual variance $\tau^2 = E[\mathrm{Var}(Y|\mathbf{X})] = E(Y - m(\mathbf{X}))^2$ [see (1.2)]. Breiman and Meisel (1976) proposed estimators of τ^2 based on piecewise linear approximation of $m(\mathbf{x})$, when $\varepsilon_i = Y_i - m(\mathbf{X}_i)$ are i.i.d. $N(0, \tau^2)$. Gasser, Sroka and Jennen-Steinmetz (1986), Buckley, Eagleson and Silverman (1988) and Hall and Marron (1990) considered estimation of τ^2 for the case of fixed one-dimensional predictors. In these papers estimators of residual variance, based on spline and kernel regression estimators, are shown to be $n^{1/2}$-consistent and to satisfy certain optimality criteria. Estimation in the known or controlled design case may be considerably different from the estimation problem in our setting (see, e.g., Remark 2.2). The time series literature also examines the problem of estimating residual variance [see Skaug and Tjöstheim (1993), Cheng and Tong (1993) and Bhansali (1993)].

2. Estimation of explanatory power of covariates. In this section we discuss the problem of estimating η^2, based on a sample of i.i.d. observations $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ from a common distribution $F(\mathbf{x}, y)$, and we introduce estimators that, under suitable conditions, are asymptotically normal at the $\sqrt{n}$ rate. In order to avoid boundary effects and to have a measure of explanatory power for which $\sqrt{n}$ asymptotics would be available, we will include in the definition of η^2 a bounded nonnegative weight function $w(\mathbf{x})$ supported on a set where the density $f(\mathbf{x})$ of covariates $\mathbf{X}$ is bounded away from 0. Weight functions are invariably introduced when global measures of deviation (such as average square error or mean integrated square error) are used in order to avoid problems with the density $f(\mathbf{x})$ approaching zero and of large bias near the boundary of the support of $f(\mathbf{x})$ [cf. Marron and Härdle (1986), Härdle (1990) and Fan (1992)].

We thus consider the functional

$$\eta_w^2 = \frac{\int (m(\mathbf{x}) - \mu_{Y,w})^2 f(\mathbf{x}) w(\mathbf{x})\, d\mathbf{x}}{\sigma_{Y,w}^2}, \tag{2.1}$$

based on the ANOVA decomposition

$$\sigma_{Y,w}^2 = \int (m(\mathbf{x}) - \mu_{Y,w})^2 f(\mathbf{x}) w(\mathbf{x})\, d\mathbf{x} + \int (y - m(\mathbf{x}))^2 f(\mathbf{x}, y) w(\mathbf{x})\, d\mathbf{x}\, dy, \tag{2.2}$$

where

$$\mu_{Y,w} = \int w(\mathbf{x}) y f(\mathbf{x}, y)\, d\mathbf{x}\, dy, \qquad \sigma_{Y,w}^2 = \int (y - \mu_{Y,w})^2 f(\mathbf{x}, y) w(\mathbf{x})\, d\mathbf{x}\, dy,$$

and $f(\mathbf{x}, y)$ is the density of $F(\mathbf{x}, y)$.

Note that if we introduce $c_w = \int f(\mathbf{x}) w(\mathbf{x})\, d\mathbf{x}$ and the new densities

$$f_w(\mathbf{x}) = \frac{f(\mathbf{x}) w(\mathbf{x})}{c_w} \quad \text{and} \quad f_w(\mathbf{x}, y) = \frac{f(\mathbf{x}, y) w(\mathbf{x})}{c_w},$$

then on the set $\{\mathbf{x}: w(\mathbf{x}) \neq 0\}$, the conditional density of $Y|\mathbf{x}$ based on $f_w(\mathbf{x}, y)$ would remain the same as $Y|\mathbf{x}$ based on $f(\mathbf{x}, y)$, so that $m_w(\mathbf{x}) = m(\mathbf{x})$ and $\mathrm{Var}_w(Y|\mathbf{x}) = \mathrm{Var}(Y|\mathbf{x})$ on this set. Equation (2.2) is the ANOVA decomposition (1.2) for the density $f_w(\mathbf{x}, y)$. Using this weighted density, we can write η_w^2, similarly to (1.3), as

$$\eta_w^2 = \mathrm{Corr}_w^2(m(\mathbf{X}), Y). \tag{2.3}$$

Typically, the weight function $w(\mathbf{x})$ will be chosen to be one in the central part of the support of $f(\mathbf{x})$ and to descend smoothly to zero near the boundary of the support of $f(\mathbf{x})$. In this case the weighted η_w^2 is nearly indistinguishable from the unweighted η^2 given by (1.1). On the other hand, the weighted correlation ratio (2.1) is important in its own right because it gives a measure of explanatory power which is not sensitive to how $f(\mathbf{x})$ approaches zero near the boundary.

In the rest of the paper, we do not reflect in our notation the dependence of functionals and estimators on a weight function which will invariably be included in their definition (with the exception of Remark 2.4).

Estimation of $\eta^2 = \{\int w(\mathbf{x}) m^2(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x} - \mu_Y^2\}/\sigma_Y^2$ involves nonparametric estimation of the quadratic functional

$$T(F) = \int w(\mathbf{x}) m^2(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x}. \tag{2.4}$$

More generally, it will be useful to consider estimation of a class of functionals of the form

$$S(F) = E\phi(\mathbf{X}, m(\mathbf{X})), \tag{2.5}$$

with a real-valued smooth function ϕ, which includes the functional $T(F)$. Estimation of a somewhat bigger class of functionals, involving also derivatives of $m(\mathbf{x})$, was addressed in Samarov (1993) using a sample version of $S(F)$ with plugged-in kernel regression estimators. The asymptotic normality result in the present paper requires less stringent regularity conditions.

We consider here "one-step" estimators of $S(F)$ based on the influence function [cf. Bickel, Klaasen, Ritov and Wellner (1993)]. One-step estimators, originally based on the first Newton–Raphson iteration of the likelihood-type equations, are often used in parametric and nonparametric estimation to construct asymptotically efficient estimators. To construct a one-step estimator, we need to compute the influence function of $S(F)$, which is easily shown to be

$$\mathrm{IF}(\mathbf{x}, y; S, F) = D(\mathbf{x}, y, m(\mathbf{x})) - S(F),$$

where

$$D(\mathbf{x}, y, m(\mathbf{x})) = \phi(\mathbf{x}, m(\mathbf{x})) + \phi^{(1)}(\mathbf{x}, m(\mathbf{x}))(y - m(\mathbf{x})), \tag{2.6}$$

and $\phi^{(1)}(\mathbf{x}, t)$ is the partial derivative of ϕ with respect to its last argument. An interesting feature of the functionals (2.5) is that its influence function is centered with $S(F)$ itself, and so, there is no need here either to construct an initial estimator or to split the sample, as is done in many other situations where a one-step estimator is used [cf. Bickel and Ritov (1988)]. The one-step estimator of $S(F)$ then has the form

$$\hat{S}_n = n^{-1} \sum_{i=1}^{n} D(\mathbf{X}_i, Y_i, \hat{m}(\mathbf{X}_i)), \tag{2.7}$$

where $\hat{m}(\mathbf{X}_i)$ is a nonparametric regression estimator. In the following theorem, the proof of which is given in Section 6, we consider the estimator $\hat{S}_n$ with the "leave-one-out" kernel estimator defined as follows:

$$\hat{m}(\mathbf{X}_i) = \frac{(n-1)^{-1} \Sigma_{j \neq i} Y_i K_h(\mathbf{X}_j - \mathbf{X}_i)}{(n-1)^{-1} \Sigma_{j \neq i} K_h(\mathbf{X}_j - \mathbf{X}_i)} = \frac{\hat{g}(\mathbf{X}_i)}{\hat{f}(\mathbf{X}_i)}, \quad \text{say,} \tag{2.8}$$

where $\hat{g}(\mathbf{x})$ and $\hat{f}(\mathbf{x})$ are "leave-one-out" kernel estimators of $g(\mathbf{x}) = \int y f(\mathbf{x}, y)\, dy$ and $f(\mathbf{x})$, respectively, $K_h(\mathbf{u}) = h^{-d} K(\mathbf{u}/h)$, and a multivariate kernel $K(\cdot)$ and a bandwidth h are chosen in Conditions 4 and 5 of the theorem. We will be using "leave-one-out" estimators throughout this section without showing it in the notation; see Remark 2.9 and Section 3 on using "all-in" kernel estimators, and see Remark 2.4 and Section 3 on using other regression estimators.

THEOREM 2.1. *Consider the estimator $\hat{S}_n$ with the "leave-one-out" kernel estimator* (2.8), *and assume that the following conditions are satisfied*:

Condition 1. $E|Y|^4 < \infty$ *and* $\sigma^2(\mathbf{x}) = E((Y - m(\mathbf{X}))^2 | \mathbf{X} = \mathbf{x})$ *is bounded for* $\mathbf{x} \in S_{\mathbf{X}}$, *where* $S_{\mathbf{X}}$ *is the support of* $f(\mathbf{x})$.

Condition 2. *The function* $\phi(\mathbf{x}, m(\mathbf{x}))$ *is supported, as a function of* $\mathbf{x}$, *on an open convex set* $\Phi \in R^d$ *such that* $\inf_{\mathbf{x} \in \Phi} f(\mathbf{x}) \geq \delta$ *for some* $\delta > 0$; $\phi(\mathbf{x}, t)$ *is bounded and continuous in* $\mathbf{x}$ *and has bounded partial derivatives up to the order* 3 *with respect to* t *for* $\mathbf{x} \in \Phi$ *and* $t \in [-b, b]$, *for some* $b < \infty$.

Let k be an integer, $k \geq d$.

Condition 3. *Partial derivatives of $f(\mathbf{x})$ and $g(\mathbf{x})$ up to the order k satisfy a Lipschitz condition uniformly on $S_{\mathbf{X}}$.*

Condition 4. (a) $n^{1/2}h^{k+1} = o(1)$ *and* (b) $n^{1/2}h^d \to \infty$ *as* $n \to \infty$.

Condition 5. *The kernel $K(\mathbf{u})$ is a bounded function with support $\{\mathbf{u}: \|\mathbf{u}\| \leq 1\}$ such that*

$$K(-\mathbf{u}) = K(\mathbf{u}), \qquad \int K(\mathbf{u})\, d\mathbf{u} = 1,$$

and

$$\int K(\mathbf{u}) u_1^{l_1} u_2^{l_2} \cdots u_d^{l_d}\, d\mathbf{u} = 0 \quad \text{for } j = 1, \ldots, k,$$

where $j = l_1 + l_2 + \cdots + l_d$ and l_i are nonnegative integers.

Then the estimator (2.7) *is asymptotically linear, that is, as* $n \to \infty$,

$$\hat{S}_n - S(F) = \frac{1}{n} \sum_{i=1}^{n} D(\mathbf{X}_i, Y_i, m(\mathbf{X}_i)) - S(F) + o_p(n^{-1/2}). \tag{2.9}$$

REMARK 2.1. Conditions 3, 4(a) and 5 are used to control the bias. Note that Condition 4 implies that h decreases to zero faster than the optimal rate for curve estimation [cf. Stone (1982)], that is, we undersmooth to keep the bias of our estimator of the order $o(n^{-1/2})$. The variance remains of the order n^{-1} because of the additional averaging in (2.7). Note also that Condition 2 is weaker than the corresponding condition in Samarov (1993) and that the proof of Theorem 2.1 is considerably simpler than that of Theorem 1 of Samarov (1993) because there is no need to use a U-statistics projection argument for the one-step estimator.

For the special case of the functional $T(F)$, Theorem 2.1 implies the following corollary.

COROLLARY 2.1. *We assume the following condition.*

Condition 6. *The function $w(\mathbf{x})$ is a bounded continuous nonnegative weight function supported on an open convex set $\Psi \in R^d$ such that $\inf_{\mathbf{x} \in \Psi} f(\mathbf{x}) \geq \delta$ for some $\delta > 0$.*

Under conditions 1, 3, 4, 5 *and* 6, *the estimator*

$$\hat{T} = \frac{1}{n} \sum_{i=1}^{n} \left[2Y_i \hat{m}(\mathbf{X}_i) - \hat{m}^2(\mathbf{X}_i)\right] w(\mathbf{X}_i), \tag{2.10}$$

with $\hat{m}(\mathbf{X}_i)$ as in (2.8), *has the following expansion as* $n \to \infty$:

$$\begin{aligned} \hat{T} - T(F) = {} & \frac{1}{n} \sum_{i=1}^{n} w(\mathbf{X}_i)\left[2Y_i m(\mathbf{X}_i) - m^2(\mathbf{X}_i)\right] \\ & - T(F) + o_p(n^{-1/2}). \end{aligned} \tag{2.11}$$

REMARK 2.2. The expansion (2.11) implies that $\hat{T}$ is asymptotically normal with the asymptotic variance

$$4E(w^2(\mathbf{X})\sigma^2(\mathbf{X})m^2(\mathbf{X})) + \mathrm{Var}(w(\mathbf{X})m^2(\mathbf{X})).$$

When the design density is known, the asymptotic variance can be made smaller by the term $\mathrm{Var}(w(\mathbf{X})m^2(\mathbf{X}))$ [cf. Pastukhova and Khasminskii (1989)]. This term equals the error variance of the standard Monte Carlo method of approximate calculation of the integral $\int w(\mathbf{x})m^2(\mathbf{x})f(\mathbf{x})\,d\mathbf{x}$. This means that when the design density is known, the leading term of the error comes only from the noisy measurements of the regression function, while in the unknown design case, the Monte Carlo error of calculation of the integral with noiseless measurements is of the same order as the regression "measurement" error. When there is also a possibility of controlling the design, it may be possible to construct asymptotically efficient estimators of the functionals $T(F)$ and η^2 which do not require any smoothness of the regression function. For example, if we can perform repeated (e.g., two) measurements at the same design points, we can estimate the residual variance using the differences between those measurements, and use this estimate to estimate η^2 and $T(F)$.

The following proposition, which will be used several times in the rest of the paper, follows immediately from Samarov [(1993), Theorem 1] and (2.11).

PROPOSITION 2.1. *Let Conditions* 1, 3, 4, 5 *and* 6 *be satisfied, and, in addition, the weight function* $w(\mathbf{x})$ *satisfies the following condition*:

Condition 7: *Partial derivatives of* $w(\mathbf{x})$ *up to the order* k *satisfy a Lipschitz condition uniformly on the set* Φ *defined in Condition* 2.

Then, as $n \to \infty$, *we have the following*:

$$\text{(i)} \quad \frac{1}{n}\sum w(\mathbf{X}_i)\hat{m}^2(\mathbf{X}_i) = \frac{1}{n}\sum w(\mathbf{X}_i)\left[2Y_i m(\mathbf{X}_i) - m^2(\mathbf{X}_i)\right] + o_p(n^{-1/2});$$

$$\text{(ii)} \quad \frac{1}{n}\sum w(\mathbf{X}_i)\hat{m}(\mathbf{X}_i) = \frac{1}{n}\sum w(\mathbf{X}_i)Y_i + o_p(n^{-1/2});$$

$$\text{(iii)} \quad \frac{1}{n}\sum w(\mathbf{X}_i)\hat{m}(\mathbf{X}_i)Y_i = \frac{1}{n}\sum w_i(\mathbf{X}_i)\left[2Y_i m(\mathbf{X}_i) - m^2(\mathbf{X}_i)\right] + o_p(n^{-1/2}).$$

(Note that the summation over i here and below will always be from 1 to n, unless indicated otherwise.)

Writing η^2 as $(T(F) - \mu_Y^2)/\sigma_Y^2$ and using the estimator (2.10) of $T(F)$, together with the (weighted) sample mean $\bar{Y} = n^{-1}\Sigma Y_i w(\mathbf{X}_i)$ and variance $s_Y^2 = n^{-1}\Sigma(Y_i - \bar{Y})^2 w(\mathbf{X}_i)$, we obtain the following estimator of η^2, using

again the "leave-one-out" kernel estimator $\hat{m}(\mathbf{X}_i)$:

$$\hat{\eta}_1^2 = \frac{\sum w(\mathbf{X}_i)\left[2Y_i\hat{m}(\mathbf{X}_i) - \hat{m}^2(\mathbf{X}_i)\right] - n\overline{Y}^2}{ns_Y^2}. \tag{2.12}$$

Note that here and in the rest of the paper we use the weight function $w(\cdot)$ in $\overline{Y}$, s_Y^2 and in other sample moments without reflecting it in our notation.

REMARK 2.3. Note that $\hat{\eta}_1^2$ can be written as

$$\hat{\eta}_1^2 = 1 - \frac{\sum w(\mathbf{X}_i)\left(Y_i - \hat{m}(\mathbf{X}_i)\right)^2}{ns_Y^2} + \frac{(1-\overline{w})\overline{Y}^2}{s_Y^2},$$

$\overline{w} = n^{-1}\sum w(\mathbf{X}_i)$, which shows that $\hat{\eta}_1^2$ is a function of the weighted prediction mean square error used in cross-validation bandwidth selection procedures (cf. Section 3).

The next proposition states the asymptotic linearity of $\hat{\eta}_1^2$ together with two other estimators of η^2 which are based on the sample versions of (2.1) and (2.3), respectively, with the plugged-in kernel estimator (2.8):

$$\hat{\eta}_2^2 = \frac{n^{-1}\sum\left(\hat{m}(\mathbf{X}_i) - \overline{m}\right)^2 w(\mathbf{X}_i)}{s_Y^2} \tag{2.13}$$

and

$$\hat{\eta}_3^2 = \frac{\left[(1/n)\sum\left(\hat{m}(\mathbf{X}_i) - \overline{m}\right)\left(Y_i - \overline{Y}\right)w(\mathbf{X}_i)\right]^2}{(1/n)\sum\left(\hat{m}(\mathbf{X}_i) - \overline{m}\right)^2 w(\mathbf{X}_i)s_Y^2}, \tag{2.14}$$

where $\overline{m} = n^{-1}\sum\hat{m}(\mathbf{X}_i)w(\mathbf{X}_i)$.

PROPOSITION 2.2. *Let Conditions* 1, 3, 4 *and* 5 *be satisfied, and also assume that, for the estimator* $\hat{\eta}_1^2$, *Condition* 6 *holds and, for the estimators* $\hat{\eta}_2^2$ *and* $\hat{\eta}_3^2$, *Conditions* 6 *and* 7 *hold. Then all three estimators* $\hat{\eta}_j^2$, $j = 1, 2, 3$, *have the same first-order asymptotic expansion: as* $n \to \infty$,

$$n^{1/2}\left(\hat{\eta}_j^2 - \eta^2\right) = n^{-1/2}(1-\eta^2)\sum\left(e_i^2 - u_i^2\right)w(\mathbf{X}_i) + o_p(1), \tag{2.15}$$

where $e_i = (Y_i - \mu_Y)/\sigma_Y$ *is the standardized response and* $u_i = [Y_i - m(\mathbf{X}_i)]/\sigma_Y(1-\eta^2)^{1/2}$ *is the standardized residual.*

PROOF. Applying the identity

$$\frac{\hat{a}}{\hat{b}} = \frac{a}{b} + \hat{b}^{-1}\left[\hat{a} - a - (\hat{b} - b)\frac{a}{b}\right] \tag{2.16}$$

and the obvious expansions

$$\overline{Y}^2 - \mu_Y^2 = 2\mu_Y\left(\overline{Y} - \mu_Y\right) + o_p(n^{-1/2}),$$

$$s_Y^2 - \sigma_Y^2 = n^{-1}\sum w(\mathbf{X}_i)\left(Y_i - \mu_Y\right)^2 - \sigma_Y^2 + o_p(n^{-1/2}),$$

we obtain (2.15) for $j = 1$ from (2.11) and for $j = 2, 3$ using Proposition 2.1, after straightforward algebra. □

Proposition 2.2 implies that all three estimators $\hat{\eta}_j^2$ have the same first-order asymptotic properties. In Section 3 we compare their finite sample properties using Monte Carlo simulation and find that estimators $\hat{\eta}_1^2$ and $\hat{\eta}_3^2$ are much less sensitive to the choice of bandwidth than $\hat{\eta}_2^2$. The estimator $\hat{\eta}_3^2$ also has another advantage. While η^2 is always between 0 and 1, the estimators $\hat{\eta}_1^2$ and $\hat{\eta}_2^2$ may, with some small probability, be negative or greater than 1, respectively; $\hat{\eta}_3^2$ is clearly always between 0 and 1.

REMARK 2.4. In the case of a one-dimensional predictor X, it is possible to obtain results similar to Corollary 2.1 and Proposition 2.2 under somewhat weaker conditions and without including a weight function in the definition of the functionals; see Doksum and Samarov (1993) for details. Also, other nonparametric regression estimators, for example, $k - NN$, Yang–Stute [Yang (1981) and Stute (1984)] and splines [Silverman (1985)], or locally linear [Fan (1992)] estimators could be used in the above functional estimators and may, in fact, have some advantages over kernel estimators. The Monte Carlo simulations reported in Sections 3 and 4 provide results for kernel as well as locally linear estimates.

REMARK 2.5. Under additional regularity conditions (yet to be worked out), which would guarantee pathwise differentiability of the functional $T(F)$, Proposition 3.3.1 of Bickel, Klaasen, Ritov and Wellner (1993) together with our Corollary 2.1 implies that the estimator $\hat{T}$ will be asymptotically efficient in the class of regular estimates [see Bickel, Klaasen, Ritov and Wellner (1993)]. Similarly, under those conditions, $\hat{\eta}_j^2$, $j = 1, 2, 3$, will be asymptotically efficient estimators of η^2.

REMARK 2.6. One of the key questions in nonparametrics has always been [see, e.g., Pitman (1948) and Bickel, Klaasen, Ritov and Wellner (1993)]: how much efficiency is lost if a nonparametric procedure is used when the true model is, in fact, a simple parametric one, such as the normal model. We find that asymptotically there is no loss in using the nonparametric R-squared $\hat{\eta}_3^2$ over the parametric R-squared in the multivariate normal model. In terms of Bickel, Klaasen, Ritov and Wellner (1993), the extension from the normal model to the general nonparametric one is "free" and $\hat{\eta}_3^2$ is "adaptive." More specifically, it can readily be shown, using algebra and the δ-method, that, under the conditions of Proposition 2.2, the asymptotic distribution of $n^{1/2}(\hat{\eta}_3^2 - \eta^2)$ is the same as the asymptotic distribution of $n^{1/2}(\hat{\rho}_m^2 - \rho_m^2)$, where $\rho_m^2 = \eta^2$ is the correlation coefficient between $m(\mathbf{X})$ and Y (assuming m is known), and $\hat{\rho}_m^2$ is the squared sample correlation between $m(\mathbf{X})$ and Y. In the parametric setting where m is known and $m(\mathbf{X})$ and Y are bivariate normal, $\hat{\rho}_m^2$ is the MLE of ρ_m^2 and is asymptotically efficient. It follows then that, in this case, $\hat{\eta}_3^2$ is as efficient as $\hat{\rho}_m^2$; that is, $\hat{\eta}_3^2$, which does not require

that m is known, is as efficient as the most efficient estimate with m assumed known.

REMARK 2.7. Proposition 2.2 implies that $n^{1/2}(\hat{\eta}_j^2 - \eta^2)$ is asymptotically normal with mean zero and variance $(1 - \eta^2)^2 \operatorname{Var}[(e_1^2 - u_1^2)w(\mathbf{X}_1)]$. When $\eta^2 = 0$ or 1, this variance is zero and our result only implies degenerate normality. Thus we need to study the next term in the expansions to obtain a meaningful distributional convergence result. This is a nontrivial task which will be considered in our future work.

REMARK 2.8. Note that the sample nonparametric ANOVA decomposition does not quite hold:

$$
\begin{aligned}
s_Y^2 = {} & \frac{1}{n}\sum(\hat{\varepsilon}_i - \bar{\varepsilon})^2 w(\mathbf{X}_i) + \frac{1}{n}\sum(\hat{m}(\mathbf{X}_i) - \overline{m})^2 w(\mathbf{X}_i) \\
& + \frac{2}{n}\sum(\hat{\varepsilon}_i - \bar{\varepsilon})(\hat{m}(\mathbf{X}_i) - \overline{m})w(\mathbf{X}_i),
\end{aligned} \tag{2.17}
$$

where $\hat{\varepsilon}_i = Y_i - \hat{m}(\mathbf{X}_i)$ and $\bar{\varepsilon} = n^{-1}\sum\hat{\varepsilon}_i w(\mathbf{X}_i)$. By comparison, for the best linear predictor $\hat{m}_L(\mathbf{X})$, the ANOVA decomposition holds because of the orthogonality of the sample least squares. Even though s_Y^2 does not break up into its components in the general nonlinear model exactly, it does so asymptotically.

PROPOSITION 2.3. *Under the assumptions of Proposition* 2.1,

$$
\begin{aligned}
s_Y^2 = {} & \frac{1}{n}\sum(\hat{\varepsilon}_i - \bar{\varepsilon})^2 w(\mathbf{X}_i) + \frac{1}{n}\sum(\hat{m}(\mathbf{X}_i) - \overline{m})^2 w(\mathbf{X}_i) \\
& + o_p(n^{-1/2}).
\end{aligned} \tag{2.18}
$$

The proposition is proved by applying Proposition 2.1 to the cross product term in (2.17).

REMARK 2.9. In all of the above estimators of η^2, we have used "leave-one-out" kernel estimators because they are somewhat easier for the asymptotic analysis than "all-in" estimators [cf. Levit (1978) and Hall (1989)]. The asymptotic properties of the estimators $\hat{\eta}_j^2$, $j = 1, 2, 3$, stated in Proposition 2.2 remain unchanged, however, when the "all-in" kernel estimators are used. To prove this it is sufficient to show that, under the assumptions of Proposition 2.1, the replacement of the "one-out" kernel estimator with the "all-in" one in the left-hand sides of (2.11) and of the expressions in claims (i) and (ii) of Proposition 2.1 results in a change of the order $o_p(n^{-1/2})$; see Section 6 for the proof.

Note, however, that, with a "leave-one-out" estimator $\hat{m}(\mathbf{X}_i)$, the estimator $\hat{T}$ in (2.10) is easily shown to have negative bias:

$$
E(\hat{T}) = T(F) - \text{MISE}, \tag{2.19}
$$

where MISE is the familiar mean integrated square error of regression [see, e.g., Härdle (1990)], which explains the negative bias of the one-step-type estimator $\hat{\eta}_1^2$ in simulations reported in Section 3. The point here is that with a "leave-one-out" estimator we effectively estimate the mean square error of prediction $E[n^{-1}\Sigma(Y_i - \hat{m}(\mathbf{X}_i))^2]$, while η^2 is based on the residual variance $\tau^2 = E(Y - m(\mathbf{X}))^2$, the difference being equal to MISE. In the next section we study finite sample properties of both "one-out" and "all-in" estimators and find that in a number of examples a combination of the two gives better results.

3. Bandwidth selection and simulation results. There have been many papers in recent years addressing the problem of data-based bandwidth selection for nonparametric (mostly density or regression) curve estimation [e.g., Park and Marron (1990), Härdle (1990), Hall and Johnstone (1992) and references therein]. The methods are typically based either on some form of cross-validation or on plugging some preliminary nonparametric estimators into an asymptotic approximation to the integrated square error, or on a mixture of the two.

Much less is known about smoothing-parameter selection for integral functionals. For the estimation of the integrated square density, Schweder (1975), Jones and Sheather (1991) and Sheather, Hettsmansperger and Donald (1994) proposed bandwidth choice based on minimization of an approximate bias.

Recent works on optimal bandwidth selection for curves show that the behavior of those selectors importantly depends on how bandwidth is selected for estimators of integral functionals appearing in the asymptotic curve estimation error [see Hall and Johnstone (1992) and references therein].

As we have seen in Section 2, the estimators of integral functionals are, under certain conditions, $n^{1/2}$-consistent and asymptotically normal, so that the leading term in the MSE does not depend on the bandwidth, as long as it is chosen within certain limits; see Condition 5 of Theorem 2.1 and Remark 2.1. This means that the asymptotically optimal choice of h can only be made based on the higher-order terms, which makes this approach extremely cumbersome [cf. Härdle, Hart, Marron and Tsybakov (1992) and Skaug and Tjöstheim (1993)]. When a functional is not $n^{1/2}$-estimable, which would happen in our context, for example, when $m(\cdot)$ is not smooth enough, the bandwidth choice will affect the rate of convergence itself and thus will be even more important. Since in practice the precise degree of smoothness of the underlying curve is unknown, the data-based bandwidth selection should not depend on specific assumptions about it.

We consider here just one such choice and evaluate its performance using two Monte Carlo simulation examples. Other bandwidth selection rules are discussed in Doksum and Samarov (1993). Our proposal is based on the extremal property (1.3) of η^2: choose bandwidth as $h_{\mathrm{COR}} := \arg\max_h \hat{\eta}_3^2$, where $\hat{\eta}_3^2$, defined in (2.14), is based on the "leave-one-out" version of $\hat{m}$. In

TABLE 1
Values of the linear and nonparametric coefficients of determination for the bump model (3.1)

τ^2	$\frac{1}{4}$	1	4	16
η^2	0.947	0.816	0.526	0.217
ρ^2	0.445	0.384	0.247	0.102

the Monte Carlo experiments below, we compute $\hat{\eta}_3^2(h_{\mathrm{COR}})$ using the "all-in" regression estimators.

For comparison, we also consider the standard cross-validation method for curve estimation, which is equivalent to choosing the h that maximizes the "leave-one-out" version of the one-step estimator $\hat{\eta}_1^2$ (see Remark 2.3). We will refer to this bandwidth choice as h_{CV}.

EXAMPLE 3.1. Suppose that X and Y are related by the equation

$$Y = 2 - 5X + 5\exp\left\{-100\left(X - \tfrac{1}{2}\right)^2\right\} + \tau\varepsilon, \tag{3.1}$$

where X and ε are independent with respective distributions $U(0,1)$ and $N(0,1)$. This is the "bump" model [e.g., Härdle (1990)]. As can be seen from Table 1, we cover a reasonable range of values of η^2 and ρ^2, where ρ^2 is the usual Pearson correlation coefficient, by choosing τ^2 equal to $\frac{1}{4}$, 1, 4 and 16. It is easy to check that $\rho^2 = 0.4701\eta^2$.

Figure 1a shows a scatter plot of $n = 200$ points generated by this model with $\tau^2 = 1$ and the true regression curve $m(x)$. Figure 1b and c shows the results of a Monte Carlo study where 400 simulated samples of size $n = 200$ were generated from the bump model (3.1) with $\tau^2 = 1$. The results for $\tau^2 = \frac{1}{4}$, 4 and 16 were qualitatively similar. We have used kernel regression estimators with quartic kernel $K(u) = 15/16(1-u^2)^2 1\{|u| \le 1\}$ and weight function $w(x) = I[\hat{f}(\mathbf{x}) \ge b]$ with $b = 0.01$, which roughly corresponds to 5% trimming. Experiments with no trimming ($b = 0$) led to more unstable estimators, while larger trimming constants ($b = 0.02$) produced results roughly similar to those with $b = 0.01$ reported here.

Figure 1b shows medians over 400 realizations of four estimators of η^2, as a function of the bandwidth h. The bandwidth h is measured in units of the standard deviation of X. The four plotted estimators are the estimators $\hat{\eta}_j^2(h, \text{1-out})$, $j = 1,2,3$, with the "leave-one-out" kernel regression [see (2.12)–(2.14)], and the estimator $\hat{\eta}_1^2(h, \text{all-in})$ with the "all-in" kernel regression. Figure 1b shows that $\hat{\eta}_2(h, \text{1-out})$ is very sensitive to the choice of bandwidth, which, together with similar evidence from other experiments, makes it less attractive than the other two estimators. Figure 1b also shows that $\hat{\eta}_1^2(h, \text{1-out})$ underestimates the true value $\eta^2 = 0.8159$ for all h, as is to be expected from (2.19). We can also see that $\hat{\eta}_1^2(h, \text{all-in})$ is much larger than $\hat{\eta}_1^2(h, \text{1-out})$ for small h and approaches 1 as $h \to 0$, as is to be expected from

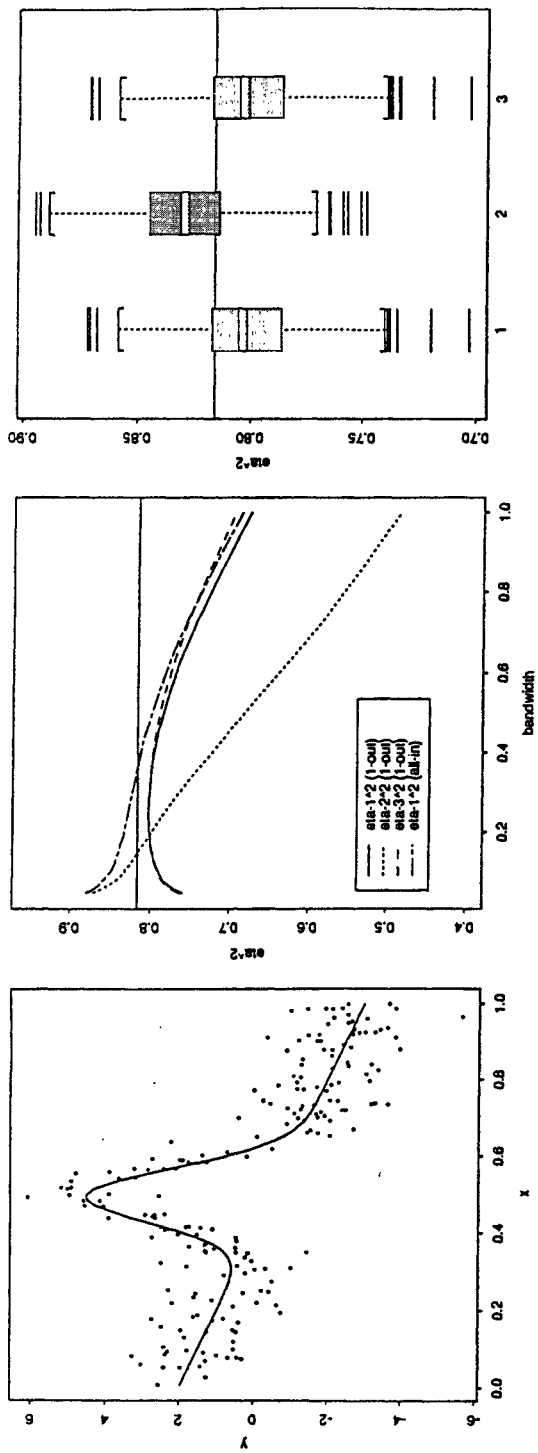

FIG. 1. *Simulation results for the bump model* (3.1) *with* $\tau = 1$ *and* $n = 200$: (a) *Scatterplot of one realization with the true regression curve* $m(x) = E(Y|x)$. (b) *Medians over* 400 *Monte Carlo trials of four different estimates of* η^2 *as functions of the bandwidth* h. *The bandwidth* h *is in units of the standard deviation of* X. *The estimates are based on the kernel estimates of* $m(x)$ *with quartic kernel. The horizontal solid line represents the true value of* $\eta^2 = 0.8159$. (c) *The boxplots of three estimates of* η^2 *with bandwidth selectors* h_{COR} *and* h_{CV} *over* 400 *Monte Carlo trials*: $1 = \hat{\eta}_3^2(h_{\mathrm{COR}}, 1\text{-}out)$; $2 = \hat{\eta}_3^2(h_{\mathrm{COR}}, all\text{-}in)$; $3 = \hat{\eta}_1^2(h_{\mathrm{CV}}, 1\text{-}out)$.

the overfitting aspect of the "all-in" estimate. Note that formula (6.8) implies that the estimator $\hat{\eta}_1^2(h, \text{all-in})$ will be greater than $\hat{\eta}_1^2(h, \text{1-out})$ for all h as long as the kernel is nonnegative, which is the case in our calculations. The estimates $\hat{\eta}_1^2(h, \text{1-out})$ and $\hat{\eta}_3^2(h, \text{1-out})$ are very close to each other and both are relatively stable as functions of h, with $\hat{\eta}_3$ being somewhat more stable.

Figure 1c shows boxplot summaries of the performance of estimates of η^2 with bandwidth selectors h_{COR} and h_{CV}. The first two boxplots both use the bandwidth choice h_{COR}: the one on the left is $\hat{\eta}_3^2(h_{\text{COR}}, \text{1-out})$, while the second one is $\hat{\eta}_3^2(h_{\text{COR}}, \text{all-in})$. The first tends to underestimate η^2, while the second one overestimates η^2. The rightmost boxplot is for $\hat{\eta}_1^2(h_{\text{CV}}, \text{1-out})$, which tends to underestimate η^2 and is quite similar to $\hat{\eta}_3^2(h_{\text{COR}}, \text{1-out})$.

The overall conclusion from this simulation experiment is that the "leave-one-out" versions of $\hat{\eta}_1^2$ and $\hat{\eta}_3^2$ are rather stable within a relatively wide range of bandwidths, so that some oversmoothing or undersmoothing should not cause a problem for these estimates. A significant variability of the selectors h_{CV} and h_{COR}, resulting from relatively flat maxima of $\hat{\eta}_1^2(h, \text{1-out})$ and $\hat{\eta}_3^2(h, \text{1-out})$, is not a drawback of these selectors but rather an indication of very desirable stability of the estimators. Since $\hat{\eta}_1^2$ and $\hat{\eta}_3^2$ are quite close to each other and $\hat{\eta}_3^2$ is a bit more stable, we will focus on $\hat{\eta}_3^2(h_{\text{COR}}, \text{1-out})$.

Note also that Figure 1c shows that $\hat{\eta}_3^2(h_{\text{COR}}, \text{1-out})$ underestimates η^2 by roughly the same amount that $\hat{\eta}_3^2(h_{\text{COR}}, \text{all-in})$ overestimates it, which suggests to try the average, or mixed, estimator

$$\hat{\eta}_3^2(\text{mixed}) = \tfrac{1}{2}\left[\hat{\eta}_3^2(h_{\text{COR}}, \text{1-out}) + \hat{\eta}_3^2(h_{\text{COR}}, \text{all-in})\right]. \tag{3.2}$$

We show next simulation results for this estimator and also compare it with the estimator $\hat{\eta}_3^2$ based on locally linear estimates $\tilde{m}(\mathbf{x})$ of $m(\mathbf{x})$ [see Fan (1992) for properties and bibliography]. To define $\tilde{m}(x)$, let $a(\mathbf{x})$ and $\mathbf{b}(\mathbf{x})$ be the values of a and $\mathbf{b}$ that minimize the locally linear weighted least squares criterion

$$S(a, \mathbf{b}) = \sum \left\{Y_j - \left[a + \mathbf{b}^T(\mathbf{X}_j - \mathbf{x})\right]\right\}^2 K\left(\frac{\mathbf{X}_j - \mathbf{x}}{h}\right), \tag{3.3}$$

where K is a kernel function, and define $\tilde{m}(\mathbf{x})$ to be equal to $a(\mathbf{x})$. As with the kernel estimates, one can consider "one-out" and "all-in" versions of $\tilde{m}(\mathbf{X}_i)$ depending on whether or not $(\mathbf{X}_i, Y_i)$ is left out of the sum in (3.3).

Table 2 gives the results of this Monte Carlo simulation, with locally linear and kernel regression estimators, respectively, where 500 simulated samples of size $n = 200$ were generated from the bump model (3.1), with $\tau^2 = \frac{1}{4}$, 1, 4 and 16, while Figure 2 shows the corresponding boxplots for $\tau^2 = \frac{1}{4}$ and 4. The results shown are for the same quartic kernel K and indicator weight function w with $b = 0.01$ as above. The results in Figure 2 show that, for this model and $h = h_{\text{COR}}$, $\hat{\eta}_3^2$(mixed) is much better than $\hat{\eta}_3^2$(all-in) and $\hat{\eta}_3^2$(1-out). This is also reflected in Table 2, which gives the bias, variance and mean squared errors of those estimates. The mixed estimate is much better than the other estimates in terms of mean squared error except in the case of the locally linear estimate when $\tau = 4$. In this one case, the one-out estimate

TABLE 2

Monte Carlo bias, variance and mean squared error (mse) (times 10^3) of the estimates of η^2 based on locally linear and kernel estimates of $m(x)$: the numbers in the table should be multiplied by 10^{-3} to get the correct values; the results are based on 500 simulations for each fixed τ in the bump model (3.1); the weight function used is $I[\hat{f}(x) \geq 0.01]$

	Locally linear			Kernel		
	Bias	**Var**	**mse**	**Bias**	**Var**	**mse**
			(a) $\tau = 0.5$			
One-out	−5.87	0.055	0.089	−6.6	0.070	0.113
All-in	4.50	0.047	0.068	4.03	0.062	0.079
Mixed	−0.68	0.0499	0.0504	−1.28	0.064	0.065
			(b) $\tau = 1$			
One-out	−12.00	0.572	0.716	−16.94	0.548	0.835
All-in	16.16	0.489	0.751	10.02	0.514	0.616
Mixed	2.08	0.517	0.522	−3.42	0.510	0.521
			(c) $\tau = 2$			
One-out	−34.35	2.23	3.41	−26.05	2.331	3.01
All-in	27.21	2.68	3.42	25.78	2.35	3.02
Mixed	−3.57	2.28	2.29	0.1	2.27	2.27
			(d) $\tau = 4$			
One-out	−31.68	2.54	3.54	−34.02	2.28	3.44
All-in	52.24	6.99	9.72	35.56	4.24	5.51
Mixed	10.28	3.55	3.65	0.76	2.78	2.78

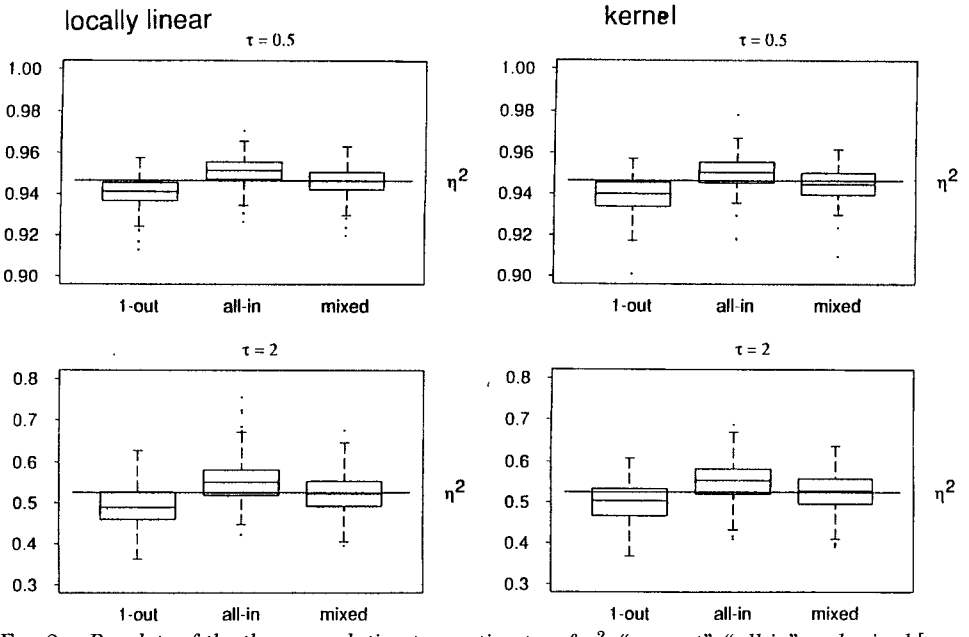

FIG. 2. *Boxplots of the three correlation-type estimates of η^2: "one-out", "all-in" and mixed [see (3.2)]. Results are based on 500 Monte Carlo samples of size 200 from model (3.1). The weight function is $w(x) = I[\hat{f}(x) \geq 0.01]$.*

is slightly better than the mixed. The locally-linear-based mixed estimate is better than the kernel-based mixed estimate when τ is 0.5, the kernel is better when τ is 4 and, when τ is 1 and 2, there is no difference.

In Doksum and Samarov (1993) we also report simulation results for evaluation of confidence intervals based on the asymptotic distribution of $\sqrt{n}(\hat{\eta}_3^2 - \eta^2)$ for this and other examples. The results show that even though the observed coverage probabilities approximate the nominal confidence levels reasonably well, the asymptotic theory tends to underestimate the actual variability. Other methods, such as bootstrap, may produce better estimates of variability. Fisher's variance stabilizing transformation leads to somewhat better results than the untransformed intervals.

EXAMPLE 3.2. Consider next the model

$$Y = 0.5 + 4X_1 + 4(X_2 - 0.5)^2 + 4X_3^{1/2} + \tau\varepsilon, \tag{3.4}$$

where X_1, X_2, X_3 and ε are independent; X_1, X_2 and X_3 are uniform on $[0, 1]$; and ε is $N(0, 1)$. For this model $\eta^2 = 104/(104 + 45\tau^2)$. As in Example 3.1, the mixed estimate $\hat{\eta}_3^2(h_{\mathrm{COR}})$ performed best; see Figure 3 and Table 3. Also, the estimate of η^2 based on the locally linear regression estimate performed better than the one based on the kernel regression estimate.

4. Explanatory power under restrictions; measures of nonlinearity and covariate subset importance. Many restrictions on the general regression model $m(\cdot)$ can be expressed in terms of $m(\cdot)$ belonging to a

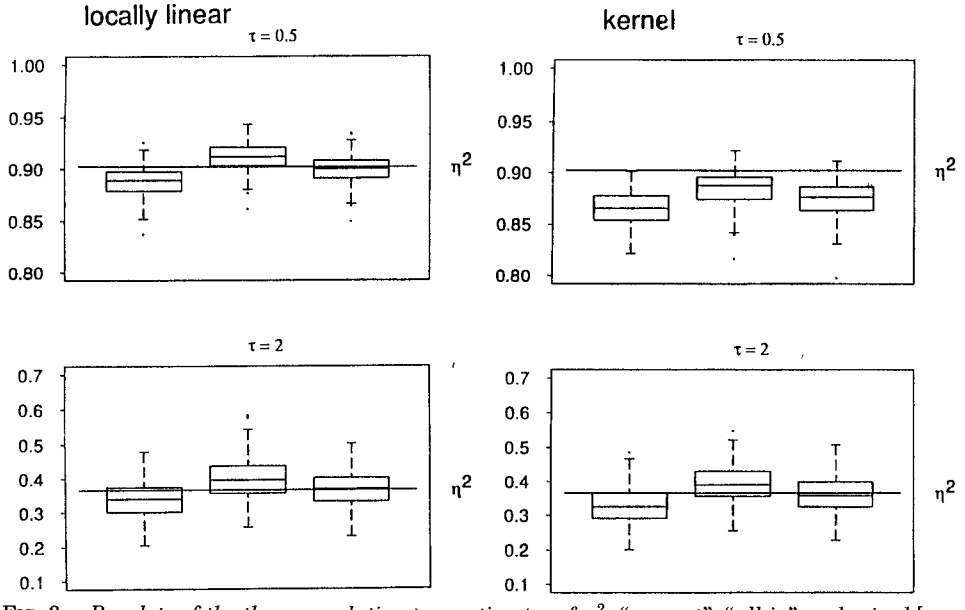

FIG. 3. *Boxplots of the three correlation-type estimates of* η^2: *"one-out", "all-in" and mixed* [*see* (3.2)]. *Results are based on* 200 *Monte Carlo samples of size* 200 *from model* (3.4). *The weight function is* $w(x) = I[\hat{f}(x) \geq 0.001]$.

TABLE 3
Bias, variance and mean squared error (times 10^3) for the model (3.4), based on 200 Monte Carlo trials, sample size $n = 200$ and trimming constant $b = 0.001$: the numbers in the table should be multiplied by 10^{-3} to get the correct values

	Locally linear			Kernel		
	Bias	**Variance**	**mse**	**Bias**	**Variance**	**mse**
			(a) $\tau = 0.5$			
One-out	−14.6	0.201	0.416	−38.2	0.325	1.179
All-in	9.2	0.170	0.257	−17.7	0.274	0.58
Mixed	−2.6	0.179	0.186	−27.9	0.287	1.07
			(b) $\tau = 1$			
One-out	−26.1	1.39	2.08	−43.7	1.44	3.36
All-in	15.4	1.41	1.65	4.0	1.15	1.17
Mixed	−5.3	1.36	1.38	−23.9	7.24	1.82
			(c) $\tau = 2$			
One-out	−29.2	2.70	3.55	−38.90	2.790	4.30
All-in	33.4	3.44	4.56	27.95	2.96	3.74
Mixed	2.0	2.80	2.80	−5.47	2.74	2.77
			(d) $\tau = 4$			
One-out	−29.1	1.59	2.44	−27.0	2.04	2.77
All-in	42.7	3.46	5.29	2.5	3.49	7.40
Mixed	6.7	2.07	2.11	17.7	2.40	2.72

subspace $F \subset L_2(f)$ of functions of $\mathbf{X}$ with finite second moments: (a) $m(\cdot)$ linear; (b) $m(\cdot)$ piecewise linear; (c) $m(\cdot)$ polynomial; (d) $m(\cdot)$ a function of a subset of variables; (e) $m(\cdot)$ additive; (f) $m(\cdot)$ a sum of ridge functions. The extent to which such restrictions hold or provide adequate approximations to $m(\mathbf{x})$ can be measured by estimating functionals of the form

$$\gamma_F = \frac{E(m(\mathbf{X}) - g_F(\mathbf{X}))^2}{E(Y - g_F(\mathbf{X}))^2}, \tag{4.1}$$

where $g_F(\mathbf{X}) = \arg\min_{g \in F} E(Y - g(\mathbf{X}))^2$, provided such a minimizer exists. It follows from the identity

$$E(Y - g(\mathbf{X}))^2 = E(Y - m(\mathbf{X}))^2 + E(m(\mathbf{X}) - g(\mathbf{X}))^2, \tag{4.2}$$

that if F is a linear subspace containing constants, then

$$E[m(\mathbf{X}) - g_F(\mathbf{X})]^2 \leq \mathrm{Var}(m(\mathbf{X})).$$

Combining this inequality with (4.2) written for $g = g_F$, we obtain from (4.1) that $0 \leq \gamma_F \leq \eta^2$. The functional γ_F measures the fraction of the residual variability left after fitting the best model from the subclass F. If F is a linear subspace containing constants, it can be easily verified, using the orthogonality $E[(m(\mathbf{X}) - g_F(\mathbf{X}))(Y - m(\mathbf{X}))] = 0$, that γ_F can be also written as

$$\gamma_F = \mathrm{Corr}^2(m(\mathbf{X}) - g_F(\mathbf{X}), Y - g_F(\mathbf{X})). \tag{4.3}$$

We will discuss here estimation of only two such functionals: measures of nonlinearity and relative importance of subsets of covariates.

4.1. *Assessing nonlinearity of regression.* According to (4.1) and (4.3), the nonlinearity index, which we will denote by γ_L, can be written as

$$\gamma_L = \frac{E\big(m(\mathbf{X}) - m_L(\mathbf{X})\big)^2}{E\big(Y - m_L(\mathbf{X})\big)^2} = \frac{E\big(m(\mathbf{X}) - m_L(\mathbf{X})\big)^2}{\sigma_Y^2(1-\rho^2)} \tag{4.4}$$

or as

$$\gamma_L = \mathrm{Corr}^2\big(m(\mathbf{X}) - m_L(\mathbf{X}), Y - m_L(\mathbf{X})\big), \tag{4.5}$$

where $m_L(\mathbf{X}) = \mu_Y + \sigma_{\mathbf{x}y}^T \Sigma_{\mathbf{x}}^{-1}(\mathbf{X} - \mu_{\mathbf{X}})$ is the best *linear* predictor of Y, $\mu_Y = E(Y)$, $\mu_{\mathbf{X}} = E(\mathbf{X})$, $\Sigma_{\mathbf{x}} = \mathrm{Var}(\mathbf{X})$, $\sigma_{\mathbf{x}y} = \mathrm{Cov}(\mathbf{X}, Y)$, $\sigma_Y^2 = \mathrm{Var}(Y)$ and the linear (population) coefficient of determination ρ^2 is defined as follows:

$$\rho^2 = \frac{\mathrm{Var}\big(m_L(\mathbf{X})\big)}{\mathrm{Var}(Y)} = \frac{\sigma_{\mathbf{x}y}^T \Sigma_{\mathbf{x}}^{-1} \sigma_{\mathbf{x}y}}{\sigma_Y^2}.$$

All the expectations, as usual, are taken here with respect to the weighted density $f_w(\mathbf{x}, y) = f(\mathbf{x}, y)w(\mathbf{x})/\int f(\mathbf{x})w(\mathbf{x})\,d\mathbf{x}$ and its marginals [cf. discussion below (2.2)].

Using the form of $m_L(\mathbf{X})$ and the fact that $\mathrm{Cov}(m(\mathbf{X}), m_L(\mathbf{X})) = \mathrm{Cov}(Y, m_L(\mathbf{X}))$, we can also write γ_L as

$$\gamma_L = \frac{\eta^2 - \rho^2}{1 - \rho^2}. \tag{4.6}$$

Sample versions of all three forms (4.4), (4.5) and (4.6) provide reasonable estimators of γ_L, the last one being easier for the asymptotic analysis since one can directly use the asymptotic linearity of estimators of η^2 established in Section 2. Let $\hat{\rho}^2 = s_{\mathbf{x}y}^T S_{\mathbf{x}}^{-1} s_{\mathbf{x}y}/S_Y^2$ be the estimator of ρ^2 obtained by replacing the (weighted) population moments in the definition of ρ^2 with the corresponding sample moments. Then, applying to $\hat{\rho}^2$ an argument similar to that given in the proof of Proposition 2.2, one can easily show that, provided the fourth moments of $\mathbf{X}$ and Y exist, $\Sigma_{\mathbf{x}}$ has full rank and $\rho^2 = \mathrm{Corr}^2(Y, m_L(\mathbf{X})) < 1$:

$$n^{1/2}\big(\hat{\rho}^2 - \rho^2\big) = n^{-1/2}(1-\rho^2)\sum_{i=1}^{n}\big(e_i^2 - u_{Li}^2\big)w(\mathbf{X}_i) + o_p(1), \tag{4.7}$$

where $e_i = (Y_i - \mu_Y)/\sigma_Y$ and $u_{Li} = (Y_i - m_L(\mathbf{X}_i))/\sigma_Y(1-\rho^2)^{1/2}$ [as usual, the weight $w(\cdot)$ is assumed to be included in calculation of all moments here].

Let $\hat{m}_{Li} = \hat{m}_L(\mathbf{X}_i) = \overline{Y} + s_{\mathbf{x}y}^T S_{\mathbf{x}}^{-1}(\mathbf{X}_i - \overline{\mathbf{X}})$ be the sample version of $m_L(\mathbf{X}_i)$, where all sample moments are calculated with weights $w_i = w(\mathbf{X}_i)$; and also write $\overline{m}_L = n^{-1}\sum_{i=1}^n \hat{m}_L(\mathbf{X}_i)w_i$ and $\hat{m}_i = \hat{m}(\mathbf{X}_i)$. Combining asymptotic linearity expansions for estimators of η^2 in Proposition 2.2, (4.7) and (2.16), we obtain the proof of the part of the following proposition concerning the estimator $\hat{\gamma}_{1L}$.

PROPOSITION 4.1. *Assume that the conditions of Proposition* 2.1 *are satisfied, that the fourth moment of* $\mathbf{X}$ *exists, that* Σ *has full rank and that* $\rho^2 = \rho_w^2 = \mathrm{Corr}_w^2(Y, m_L(\mathbf{X})) < 1$. *Then these three estimators of* γ_L,

$$\text{(i)} \qquad \hat{\gamma}_{1L} = \frac{\hat{\eta}^2 - \hat{\rho}^2}{1 - \hat{\rho}^2},$$

with $\hat{\eta}^2$ *one of* $\hat{\eta}_j^2$, $j = 1, 2$ *or* 3 [*see* (2.12)–(2.14)],

$$\text{(ii)} \qquad \hat{\gamma}_{2L} = n^{-1} \sum \left(\hat{m}_i - \hat{m}_{Li} - \overline{m} + \overline{m}_L\right)^2 w_i / s_Y^2 (1 - \hat{\rho}^2),$$

$$\text{(iii)} \qquad \hat{\gamma}_{3L} = \frac{\left[n^{-1}\Sigma(Y_i - \hat{m}_{Li})(\hat{m}_i - \hat{m}_{Li})w_i\right]^2}{\left[n^{-1}\Sigma(\hat{m}_i - \hat{m}_{Li} - (\overline{m} - \overline{m}_L))^2 w_i\right]\left[n^{-1}\Sigma(Y_i - \hat{m}_{Li})^2 w_i\right]},$$

all have the same first-order asymptotic expansion: as $n \to \infty$,

$$(4.8) \qquad n^{1/2}(\hat{\gamma}_{jL} - \gamma_L) = n^{-1/2}(1 - \gamma_L) \sum_{i=1}^{n} (u_{Li}^2 - u_i^2) w_i + o_p(1),$$

where u_i *is defined right after* (2.15).

The proof of (4.8) for the estimators $\hat{\gamma}_{2L}$ and $\hat{\gamma}_{3L}$ is given in Section 6.

As in Section 3, we prefer the "correlation-based" estimator $\hat{\gamma}_{3L}$ based on our experience with its finite sample behavior and the fact that, unlike estimators $\hat{\gamma}_{1L}$ and $\hat{\gamma}_{2L}$, it always takes values between 0 and 1.

Here we summarize Monte Carlo results for the nonlinearity index in the bump model, Example 3.1. We use both kernel and locally linear versions of $\hat{\gamma}_{3L}$ and in each case choose h to maximize the "leave-one-out" version of $\hat{\eta}_3^2$. The final estimate is the "all-in" version of $\hat{\gamma}_{3L}$ with this choice of bandwidth h. Table 4 indicates that when the standard deviation τ of the noise is small, the kernel estimate is slightly better than the locally linear estimate, while the locally linear estimate is much better when $\tau = 2$ and 4.

TABLE 4

Bias, variance and mean squared error (times 10^3*) of the estimated index of nonlinearity* $\hat{\gamma}_{3L}$ *for the bump model* (3.1): *the numbers in the table should be multiplied by* 10^{-3} *to get the correct values; the kernel version is computed for* 400 *Monte Carlo replicates while the locally linear version of* $\hat{\gamma}_{3L}$ *is computed for* 200 *Monte Carlo replicates, both with samples of size* 200

		Locally linear			Kernel		
		Bias	**Variance**	**mse**	**Bias**	**Variance**	**mse**
$\tau = 0.5$,	$\gamma_L = 0.904$	7.02	0.272	0.321	7.54	0.219	0.276
$\tau = 1$,	$\gamma_L = 0.701$	18.34	1.69	2.03	15.68	1.51	1.77
$\tau = 2$,	$\gamma_L = 0.369$	39.02	4.10	5.62	81.47	4.88	11.52
$\tau = 4$,	$\gamma_L = 0.128$	56.86	10.02	13.26	122.39	6.83	21.81

REMARK 4.1. Note that combining (4.4) and (4.6) one obtains another expression for η^2:

$$\eta^2 = \rho^2 + (1 - \rho^2)\mathrm{Corr}^2(m(\mathbf{X}) - m_L(\mathbf{X}), Y - m_L(\mathbf{X})),$$

the sample version of which leads to yet another estimator $\hat{\eta}^2$, which, unlike the estimators of η^2 considered earlier, satisfies the inequality $\hat{\rho}^2 \le \hat{\eta}^2 \le 1$.

4.2. *Measuring the relative importance of a subset of covariates.* In order to measure the importance of $\mathbf{X}_J = \{\mathbf{X}_j\colon\ j \in J\}$ relative to the full set of variables $\mathbf{X}$, we follow the pattern of (4.1) and define a measure of relative importance of $\mathbf{X}_J$ as

$$\begin{aligned}\gamma_J &= \frac{\eta^2 - \eta_J^2}{1 - \eta_J^2} \\ &= \frac{E(m(\mathbf{X}) - m_J(\mathbf{X}_J))^2}{E(Y - m_J(\mathbf{X}_J))^2} \\ &= \mathrm{Corr}^2(m(\mathbf{X}) - m_J(\mathbf{X}_J), Y - m_J(\mathbf{X}_J)).\end{aligned} \tag{4.9}$$

Note that the smaller values of γ_J correspond to the greater importance of $\mathbf{X}_J$. As in the previous subsection, we can take a sample version of any of the three expressions in (4.9) as an estimator of γ_J, and, for the same reasons as before, we prefer the third one:

$$\hat{\gamma}_J = \widehat{\mathrm{Corr}}^2(\hat{m}(\mathbf{X}) - \hat{m}_J(\mathbf{X}_J), Y - \hat{m}_J(\mathbf{X}_J)), \tag{4.10}$$

where $\widehat{\mathrm{Corr}}$ is the usual sample correlation in which as before we include the weight $w(\mathbf{X})$.

PROPOSITION 4.2. *Assume that the conditions of Proposition* 2.1 *are satisfied and the same conditions hold with* $\mathbf{X}$ *replaced by the subset of variables* $\mathbf{X}_J$. *Then*

$$\begin{aligned}&n^{1/2}(\hat{\gamma}_J - \gamma_J) \\ &\quad = n^{-1/2}(1 - \gamma_J)\sum_{i=1}^{n}(u_{Ji}^2 - u_i^2)w(\mathbf{X}_i) + o_p(1) \quad \text{as } n \to \infty,\end{aligned} \tag{4.11}$$

where $u_{Ji} = (Y_i - m_J(\mathbf{X}_{Ji}))/\sigma_y(1 - \eta_J^2)^{1/2}$.

The proof of the proposition is given in Section 6. We illustrate the use of $\hat{\gamma}_J$ in the next section.

5. Data examples. The estimators of η^2, γ_J and γ_L, used in this section are the correlation-type estimators $\hat{\eta}_3^2$, $\hat{\gamma}_J$ and $\hat{\gamma}_{3L}$.

EXAMPLE 5.1. We consider first the "food" data analyzed by Härdle (1990) where the expenditure on food in a household is linked to the net income of

the household. The sample size is $n = 7125$. We select the bandwidth which maximizes the "one-out" version of $\hat{\eta}^2$. Only the estimators based on kernel regression are considered. We find that the mixed estimate of the nonparametric correlation is $\hat{\eta}^2 = 0.433$, with a standard error of 0.012, while the "all-in" version of the nonlinearity index is $\hat{\gamma}_L^2 = 0.154$, with a standard error of 0.013. Thus after linear prediction there is 15.4% more variability to explain.

EXAMPLE 5.2. Next we consider the "Boston housing data" analyzed by Breiman and Friedman (1985). The response variable Y is the median value of a house in a given area, while the covariates we consider are as follows: X_1 is the average number of rooms per house in the area; X_2 is the percentage of the population of lower status in the area; and X_3 is the weighted distance to five Boston employment centers from houses in the area. The explanatory power of the three covariates is estimated as $\hat{\eta}^2(\text{mixed}) = 0.829$ by the kernel-based method, with a standard error of 0.019, and as $\hat{\eta}^2(\text{mixed}) = 0.838$ by the locally-linear-based method. By comparison, the ACE model with these three covariates [see Breiman and Friedman (1985)], that is, the "best" additive model with the "optimally" transformed response, has the coefficient of determination 0.812 (computed using the S-PLUS version of ACE).

Table 5 gives the result of an analysis of explanatory power, of subset importance and of nonlinearity. For each subset J of covariates it gives the estimated explanatory power $\hat{\eta}_J^2$, the measure of subset importance $\hat{\gamma}_J$ and the nonlinearity index $\hat{\gamma}_L^J$. The linear model coefficient of determination R^2 is

TABLE 5

The estimated subset explanatory power $\hat{\eta}_J^2$, measure of subset importance $\hat{\gamma}_J$ and the estimated linearity index $\hat{\gamma}_L^J$ for the covariates in the Boston housing data: the results are for the kernel-based estimate with $b = 0.001$; $\hat{\eta}_J^2$ is the mixed version of $\hat{\eta}_3^2$ with the bandwidth h_{COR}; $\hat{\rho}_J^2$ is the coefficient of determination R^2 for the indicated subset; standard errors are indicated in the parentheses underneath the values

Subset J	$\hat{\rho}_J^2$	$\hat{\eta}_J^2$	$\hat{\gamma}_J$	$\hat{\gamma}_L^J$
{Rooms}	0.484	0.570	0.648	0.197
		(0.046)	(0.034)	(0.031)
{% lower status} (LS)	0.541	0.679	0.554	0.320
		(0.028)	(0.040)	(0.045)
{Distance} (Dist.)	0.063	0.176	0.830	0.157
		(0.034)	(0.021)	(0.025)
{Rooms, % LS}	0.639	0.779	0.315	0.365
		(0.030)	(0.056)	(0.049)
{Rooms, Dist.}	0.496	0.575	0.621	0.277
		(0.051)	(0.042)	(0.047)
{% LS, Dist.}	0.562	0.724	0.395	0.505
		(0.020)	(0.057)	(0.033)
{Rooms, % LS, Dist.}	0.647	0.829		0.616
		(0.019)		(0.034)

included for comparison in the first column. The η^2 column shows that {Distance} explains 17.6% of the variability, while all three variables explain 82.9%. Since $\gamma_J = (\eta^2 - \eta_J^2)/(1 - \eta_J^2)$, small values of $\hat{\gamma}_J$ indicates strong subset importance. Thus the most important proper subset is {Rooms, % lower status}, while the least important subset is {Distance}. In fact, after prediction by {Distance}, there is 83.0% more variability to explain, while after prediction by {Rooms, % lower status} there is 31.5% more variability to explain.

The results in Table 5 show strong nonlinearity. For instance, for the subset $J = \{\%$ lower status$\}$, 32.0% of the residual variability left after subtracting out the variability explained by the linear predictor can be explained by the nonparametric predictor. Similarly, for J the subset of all three variables, 61.6% of the variability left after linear prediction can be explained by the nonparametric predictor.

6. Proofs. Throughout the proofs we will use the subscript i in place of the argument $\mathbf{X}_i$, so that $\hat{m}(\mathbf{X}_i) - m(\mathbf{X}_i)$, for example, will be written as $\hat{m}_i - m_i$ and $\sigma(\mathbf{X}_i)$ as σ_i.

We will need the following lemma, which can be checked by direct computation [cf. Prakasa Rao (1983)].

LEMMA 6.1. *Under Conditions* 3, 4, 5 *and* 6, *we have the following for the estimators* $\hat{f}$ *and* $\hat{g}$ *defined in* (2.8):

$$E\int w(\mathbf{x})\big(\hat{f}(\mathbf{x}) - f(\mathbf{x})\big)^2\, d\mathbf{x} = o(n^{-1/2}) \quad \text{as } n \to \infty.$$

If, in addition, Condition 1 *holds,*

$$E\int w(\mathbf{x})\big(\hat{g}(\mathbf{x}) - g(\mathbf{x})\big)^2\, d\mathbf{x} = o(n^{-1/2}).$$

PROOF OF THEOREM 2.1. The Taylor expansion of the influence function $D(\mathbf{x}, y, m(\mathbf{x}))$ in (2.7) gives

$$\begin{aligned}
&\hat{S}_n - \frac{1}{n}\sum_{i=1}^{n} D(\mathbf{X}_i, Y_i, m_i) \\
&\quad = \frac{1}{n}\sum_{i=1}^{n} D^{(1)}(\mathbf{X}_i, Y_i, m_i)(\hat{m}_i - m_i) \\
&\qquad + \frac{1}{2n}\sum_{i=1}^{n} D^{(2)}(\mathbf{X}_i, Y_i, \tilde{m}_i)(\hat{m}_i - m_i)^2 \\
&\quad = I_1 + I_2, \quad \text{say,}
\end{aligned} \tag{6.1}$$

where $D^{(1)}(\mathbf{x}, y, m) = \phi^{(2)}(\mathbf{x}, m)\varepsilon$ and $D^{(2)}(\mathbf{x}, y, m) = \phi^{(3)}(\mathbf{x}, m)\varepsilon - \phi^{(2)}(\mathbf{x}, m)$ are the first and second derivatives of D with respect to the argument m, $\phi^{(k)}$, $k = 1, 2, 3$, are the derivatives of ϕ with respect to m, $\varepsilon = y - m(\mathbf{x})$ and $\tilde{m}_i$ is between $\hat{m}_i$ and m_i.

Writing $\hat{m}_i - m_i$ as

$$\hat{m}_i - m_i = \frac{\hat{g}_i - \hat{f}_i m_i}{f_i} + \frac{(f_i - \hat{f}_i)(\hat{g}_i - \hat{f}_i m_i)}{f_i \hat{f}_i}, \tag{6.2}$$

we have

$$\begin{aligned} I_1 &= \frac{1}{n}\sum_i \phi^{(2)}(\mathbf{X}_i, m_i)\,\varepsilon_i \frac{\hat{g}_i - \hat{f}_i m_i}{f_i} \\ &\quad + \frac{1}{n}\sum_i \phi^{(2)}(\mathbf{X}_i, m_i)\,\varepsilon_i \frac{(f_i - \hat{f}_i)(\hat{g}_i - \hat{f}_i m_i)}{f_i \hat{f}_i} \\ &= I_{11} + I_{12}, \quad \text{say.} \end{aligned} \tag{6.3}$$

Now writing I_{11} as

$$I_{11} = \frac{1}{n(n-1)}\sum_i \sum_{j \neq i} \frac{\phi^{(2)}(\mathbf{X}_i, m_i)\,\varepsilon_i}{f_i}(Y_j - m_i)K_h(\mathbf{U}_{ij}),$$

where $\mathbf{U}_{ij} = \mathbf{X}_j - \mathbf{X}_i$, we have $EI_{11} = 0$ and

$$EI_{11}^2 = \frac{1}{n^2(n-1)^2}\sum_{i=1}^n \sum_{j \neq i} \sum_{l=1}^n \sum_{k \neq l} \left\{ E \frac{\phi^{(2)}(\mathbf{X}_i, m_i)\,\phi^{(2)}(\mathbf{X}_l, m_l)K_h(\mathbf{U}_{ij})K_h(\mathbf{U}_{lk})}{f_i f_l} \times \varepsilon_i \varepsilon_l (\varepsilon_j + m_j - m_i)(\varepsilon_l + m_l - m_k) \right\}.$$

Using the fact that $E[\varepsilon_i^{\nu_1}\varepsilon_j^{\nu_2}\varepsilon_l^{\nu_3}\varepsilon_k^{\nu_4}|\mathbf{X}^{(n)}] = 0$, when all indexes i, j, l, k are different, $1 \leq \nu_1 + \nu_2 + \nu_3 + \nu_4 \leq 4$ and at least one of the integers ν_r, $r = 1, 2, 3, 4$, is equal to 1, we have

$$EI_{11}^2 = \frac{1}{n^2(n-1)^2}\sum_{i=1}^n \sum_{j \neq i} \left\{ \frac{(\phi^{(2)}(\mathbf{X}_i, m_i))^2 K_h^2(\mathbf{U}_{ij})}{f_i^2}\left(\sigma_i^2 (m_i - m_j)^2 + \sigma_i^2 \sigma_j^2\right) + \frac{\phi^{(2)}(\mathbf{X}_i, m_i)\,\phi^{(2)}(\mathbf{X}_j, m_j)K_h^2(\mathbf{U}_{ij})}{f_i f_j}\sigma_i^2 \sigma_j^2 \right\}.$$

Using now that $|K_h(\mathbf{U}_{ij})| \leq C/h^d$ and that under Conditions 1, 2 and 3 the functions $\sigma_i^2(\mathbf{x})$, $1/f(\mathbf{x})$, $m(\mathbf{x})$ and $\phi^{(2)}(\mathbf{x}, m(\mathbf{x}))$ are bounded on the set Φ, we have $EI_{11}^2 = O((n(n-1)h^{2d})^{-1})$, which, together with Condition 4, implies that $EI_{11}^2 = o(n^{-1})$.

From the uniform consistency of the kernel density estimator $\hat{f}(\mathbf{x})$ for $\mathbf{x} \in \Phi$ [see, e.g., Prakasa Rao (1983)], we have, as $n \to \infty$,

$$\frac{1}{\inf_{\mathbf{x} \in \Phi}|\hat{f}(\mathbf{x})|} = O_p(1), \tag{6.4}$$

which, together with Conditions 1 and 3, implies that

$$\sup_{\mathbf{x} \in \Phi} |\hat{m}(\mathbf{x})| = O_p(1) \tag{6.5}$$

and

$$\sup_{\mathbf{x}\in\Phi} \left| \phi^{(k)}(\mathbf{x}, \hat{m}(\mathbf{x})) \right| = O_p(1), \qquad k = 1,2,3. \tag{6.6}$$

This implies that $I_{12} = O_p(1)J_1$, where

$$J_1 = \frac{1}{n}\sum_{i=1}^{n} v_i \frac{|\varepsilon_i|\,|f_i - \hat{f}_i|\left(|\hat{g}_i - g_i| + |\hat{f}_i - f_i|\,|m_i|\right)}{f_i},$$

where $v_i = v(\mathbf{X}_i)$ with some bounded continuous weight function $v(\mathbf{x})$ supported on Φ. Using conditioning on $\mathbf{X}^n$, we have

$$\begin{aligned} EJ_1 \le E\int & v(\mathbf{x})\sigma(\mathbf{x})\left|f(\mathbf{x}) - \hat{f}(\mathbf{x})\right| \\ & \times\left(\left|\hat{g}(\mathbf{x}) - g(\mathbf{x})\right| + \left|\hat{f}(\mathbf{x}) - f(\mathbf{x})\right|\left|m(\mathbf{x})\right|\right) d\mathbf{x}, \end{aligned} \tag{6.7}$$

and the Cauchy–Schwarz inequality together with Lemma 1 and boundedness of $\sigma(\mathbf{x})$ and $m(\mathbf{x})$ implies now that $EJ_1 = o(n^{-1/2})$, which in turn implies that $I_{12} = o_p(n^{-1/2})$.

Turning now to the term I_2 in (6.1), we have, using (6.6),

$$I_2 = \frac{O_p(1)}{n}\sum_{i=1}^{n} v_i\left(|\varepsilon_i| + |m_i - \tilde{m}_i| + 1\right)\left(\hat{m}_i - m_i\right)^2,$$

where $v_i = v(\mathbf{X}_i)$ and $v(\mathbf{x})$ is again a bounded continuous weight function supported on Φ. Using the fact that $\tilde{m}_i$ is between $\hat{m}_i$ and m_i, (6.5) and Condition 3, we have

$$I_2 = \frac{O_p(1)}{n}\sum_{i=1}^{n} v_i\left(|\varepsilon_i| + 1\right)\left(\hat{m}_i - m_i\right)^2.$$

Writing $\hat{m}_i - m_i$ as $(\hat{g}_i - \hat{f}_i m_i)/\hat{f}_i$ and using (6.4), we get

$$I_2 = \frac{O_p(1)}{n}\sum_{i=1}^{n} v_i\left(|\varepsilon_i| + 1\right)\left(\hat{g}_i - \hat{f}_i m_i\right)^2 = O_p(1)J_2.$$

Now, similarly to J_1 in (6.7), it can be shown, using Lemma 6.1, that $EJ_2 = o(n^{-1/2})$, which implies that $I_2 = o_p(n^{-1/2})$. The proof of the theorem is complete. □

REMARK 6.1. Note that, to justify the application of Theorem 1 of Samarov (1993) to the proof of Proposition 3.1 there and to the functionals in this paper, one should add the argument similar to that leading to formulas (6.4), (6.5) and (6.6) of this paper, and in the proof of step 1 of that theorem, replace a deterministic constant C with $O_p(1)$.

PROOF OF THE RESULTS CLAIMED IN REMARK 2.9. We will prove the claim for the estimator (2.10); for the other two expressions, it is proved with an almost

identical argument. Denote the weights of the "leave-one-out" kernel estimator

$$W_{jn}(\mathbf{X}_i) = \frac{K_h(\mathbf{X}_j - \mathbf{X}_i)}{\Sigma_{k \neq i} K_h(\mathbf{X}_k - \mathbf{X}_i)},$$

so that

$$\hat{m}(\mathbf{X}_i) = \sum_{j \neq i} W_{jn}(\mathbf{X}_i) Y_j.$$

The corresponding "all-in" estimator $\tilde{m}_i$ can be then written as

$$\tilde{m}(\mathbf{X}_i) = c_i \hat{m}(\mathbf{X}_i) + (1 - c_i) Y_i, \quad \text{where } c_i = \frac{1}{1 + W_{in}(\mathbf{X}_i)}$$

and

$$W_{in}(\mathbf{X}_i) = \frac{K_h(0)}{\Sigma_{k \neq i} K_h(\mathbf{X}_k - \mathbf{X}_i)} = \frac{K(0)}{(n-1) h^d \hat{f}(\mathbf{X}_i)}.$$

It is easy to check that the difference between the "one-out" and "all-in" estimators (2.10), that is,

$$\hat{T}_n = \frac{1}{n} \sum_{i=1}^{n} w_i \big(\hat{m}_i^2 + 2 \hat{m}_i \hat{\varepsilon}_i \big) \quad \text{and} \quad \tilde{T}_n = \frac{1}{n} \sum_{i=1}^{n} w_i \big(\tilde{m}_i^2 + 2 \tilde{m}_i \tilde{\varepsilon}_i \big),$$

can be written as

$$\tilde{T}_n - \hat{T}_n = \frac{1}{n} \sum_{i=1}^{n} w_i \big(1 - c_i^2\big) \hat{\varepsilon}_i^2. \tag{6.8}$$

Condition 4(b), the uniform consistency of the kernel estimator $\hat{f}(\mathbf{x})$ for $\mathbf{x} \in \Phi$ [already used in the proof of Theorem 2.1 just above (6.4)] and (6.4) imply that $\max_{i=1,n}(1 - c_i^2) = o_p(n^{-1/2})$. This together with the fact that $n^{-1} \Sigma_{i=1}^{n} w_i \hat{\varepsilon}_i^2 = O_p(1)$, which follows from Proposition 2.1, gives the claimed result $\tilde{T}_n - \hat{T}_n = o_p(n^{-1/2})$. □

PROOF OF PROPOSITION 4.1. The claim concerning the estimator $\hat{\gamma}_{1L}$ was proved just above Proposition 4.1. To prove (4.8) for the remaining two estimators, it is sufficient to show that

$$\hat{\gamma}_{2L} - \hat{\gamma}_{1L} = o_p(n^{-1/2}) \tag{6.9}$$

and

$$\hat{\gamma}_{3L} - \hat{\gamma}_{2L} = o_p(n^{-1/2}). \tag{6.10}$$

Using the definition of $\hat{m}_{Li} = \hat{m}_L(\mathbf{X}_i)$ given just above Proposition 4.1, we have

$$\begin{aligned}
&\sum \big(\hat{m}_i - \hat{m}_{Li} - (\overline{m} - \overline{m}_L)\big)^2 w_i \\
&\quad = \sum \big(\hat{m}_i - \overline{m}\big)^2 w_i + \sum \big(\hat{m}_{Li} - \overline{m}_L\big)^2 w_i \\
&\quad\quad - 2\sum \big(\hat{m}_i - \overline{m}\big)\big(\hat{m}_{Li} - \overline{m}_L\big) \qquad (6.11) \\
&\quad = \sum \big(\hat{m}_i - \overline{m}\big)^2 w_i + n s_{\mathbf{x}y}^T S_{\mathbf{X}}^{-1} s_{\mathbf{x}y} \\
&\quad\quad - 2 s_{\mathbf{x}y}^T S_{\mathbf{X}}^{-1} \sum \big(\hat{m}_i - \overline{m}\big)\big(\mathbf{X}_i - \overline{\mathbf{X}}\big) w_i.
\end{aligned}$$

Using the second claim of Proposition 2.1 and the expansion

$$n^{-1} \sum \hat{m}_i \mathbf{X}_i w_i = n^{-1} \sum Y_i \mathbf{X}_i w_i + o_p(n^{-1/2}),$$

which is obtained by the componentwise application of Theorem 2.1 to the functional $Ew(\mathbf{X})m(\mathbf{X})\mathbf{X}$, we get the following expression for the last term in (6.11):

$$\begin{aligned}
&2 s_{\mathbf{x}y}^T S_{\mathbf{X}}^{-1} \sum \big(\hat{m}_i - \overline{m}\big)\big(\mathbf{X}_i - \overline{\mathbf{X}}\big) w_i \\
&\quad = 2 s_{\mathbf{x}y}^T S_{\mathbf{x}}^{-1} \Big(\sum Y_i \mathbf{X}_i w_i - (2 - \overline{w}) \overline{Y}\, \overline{\mathbf{X}} \Big) + o_p(n^{1/2}) \\
&\quad = 2 n s_{\mathbf{x}y}^T S_{\mathbf{X}}^{-1} s_{\mathbf{x}y} + o_p(n^{1/2}),
\end{aligned}$$

and (6.9) follows.

Noting that $n^{-1}\Sigma(Y_i - \hat{m}_L(\mathbf{X}_i))^2 = s_y^2 - s_{\mathbf{x}y}^T S_{\mathbf{X}}^{-1} s_{\mathbf{x}y}$, we have

$$\begin{aligned}
&s_y^2 - s_{\mathbf{x}y}^T S_{\mathbf{X}}^{-1} s_{\mathbf{x}y} \\
&\quad = n^{-1} \sum \big(\hat{m}_i - \hat{m}_{Li} - (\overline{m} - \overline{m}_L)\big)^2 w_i + n^{-1} \sum \big(\hat{\varepsilon}_i - \overline{\varepsilon}\big)^2 w_i \\
&\quad\quad + 2 n^{-1} \sum \big(\hat{\varepsilon}_i - \overline{\varepsilon}\big)\big(\hat{m}_i - \hat{m}_{Li} - (\overline{m} - \overline{m}_L)\big) w_i.
\end{aligned}$$

The last expression can be rewritten, using the proof of (6.9), as

$$\begin{aligned}
s_y^2 - s_{\mathbf{x}y}^T S_{\mathbf{X}}^{-1} s_{\mathbf{x}y} &= n^{-1} \sum \big(\hat{m}_i - \overline{m}\big)^2 w_i - s_{\mathbf{x}y}^T S_{\mathbf{X}}^{-1} s_{\mathbf{x}y} \\
&\quad + o_p(n^{-1/2}) + n^{-1} \sum \big(\hat{\varepsilon}_i - \overline{\varepsilon}\big)^2 w_i \\
&\quad + 2 n^{-1} \sum \big(Y_i - \hat{m}_{Li}\big)\big(\hat{m}_i - \hat{m}_{Li} - (\overline{m} - \overline{m}_L)\big) w_i \\
&\quad - 2 n^{-1} \sum \big(\hat{m}_i - \hat{m}_{Li} - (\overline{m} - \overline{m}_L)\big)^2 w_i,
\end{aligned}$$

where we also used the fact that $\hat{\varepsilon}_i = (Y_i - \hat{m}_{Li}) - (\hat{m}_i - \hat{m}_{Li})$ and $\overline{Y} = \overline{m}_L$. It follows from Proposition 2.3 that the last two terms in the preceding expression are of order $o_p(n^{-1/2})$, which implies (6.10). □

PROOF OF PROPOSITION 4.2. The structure of the proof is very similar to that of Proposition 4.1. Write

$$\hat{\gamma}_{1J} = \frac{1}{n^{-1}\Sigma\big(\hat{\varepsilon}_{Ji} - \overline{\varepsilon}_J\big)^2 w_i} \left[\frac{1}{n} \sum \big(\hat{m}_i - \overline{m}\big)^2 w_i - n^{-1} \sum \big(\hat{m}_{Ji} - \overline{m}_J\big)^2 w_i \right]$$

and

$$\hat{\gamma}_{2J} = \frac{1}{n^{-1}\sum(\hat{\varepsilon}_{Ji} - \bar{\varepsilon}_J)^2 w_i} \frac{1}{n} \sum \left[\hat{m}_i - \hat{m}_{Ji} - (\overline{m} - \overline{m}_J)\right]^2 w_i,$$

where $\hat{\varepsilon}_{Ji} = Y_i - \hat{m}_J(\mathbf{X}_{Ji})$, $\bar{\varepsilon}_J = n^{-1}\sum \hat{\varepsilon}_{Ji} w_i$, $\hat{m}_{Ji} = \hat{m}_J(\mathbf{X}_{Ji})$ and $\overline{m}_J = n^{-1}\sum \hat{m}_J(\mathbf{X}_{Ji}) w_i$. The proposition follows from the following three facts, which are proved below:

$$n^{1/2}(\hat{\gamma}_{1J} - \gamma_J) = n^{-1/2}(1 - \gamma_J)\sum (u_{Ji}^2 - u_i^2) w_i + o_p(1); \tag{6.12}$$

$$\hat{\gamma}_{2J} - \hat{\gamma}_{1J} = o_p\left(\frac{1}{\sqrt{n}}\right); \tag{6.13}$$

$$\hat{\gamma}_J - \hat{\gamma}_{2J} = o_p\left(\frac{1}{\sqrt{n}}\right). \tag{6.14}$$

Applying Proposition 2.3 to $\mathbf{X}_J$, we can write $\hat{\gamma}_{1J} = (\hat{\eta}_2^2 - \hat{\eta}_{2J}^2)/(1 - \eta_{2J}^2 + o_p(n^{-1/2}))$, and (6.12) follows now from Proposition 2.2, applied to $\mathbf{X}$ and $\mathbf{X}_J$, and also using (2.16).

Next we have

$$\begin{aligned} &n^{-1} \sum \left[\hat{m}_i - \hat{m}_{Ji} - (\overline{m} - \overline{m}_J)\right]^2 w_i \\ &\quad = n^{-1} \sum (\hat{m}_i - \overline{m})^2 w_i - n^{-1} \sum (\hat{m}_{Ji} - \overline{m}_J)^2 w_i \\ &\qquad - 2n^{-1} \sum (\hat{m}_i - \overline{m} - (\hat{m}_{Ji} - \overline{m}_J))(\hat{m}_{Ji} - \overline{m}_J) w_i, \end{aligned} \tag{6.15}$$

where the last term, which we denote by I, can be written as

$$I = -2n^{-1} \sum \hat{m}_i \hat{m}_{Ji} w_i + 2n^{-1} \sum \hat{m}_{Ji}^2 w_i + 2(\overline{m} - \overline{m}_J)\overline{m}_J. \tag{6.16}$$

Applying to (6.16) the second claim of Proposition 2.1 and the expansions

$$\overline{m}_J = n^{-1} \sum Y_i w_i + o_p(n^{-1/2}), \tag{6.17}$$

$$n^{-1} \sum \hat{m}_J^2 w_i = n^{-1} \sum (2Y_i m_{Ji} - m_{Ji}^2) w_i + o_p(n^{-1/2}) \tag{6.18}$$

and

$$n^{-1} \sum \hat{m}_i \hat{m}_{Ji} w_i = n^{-1} \sum (2Y_i m_{Ji} - m_{Ji}^2) w_i + o_p(n^{-1/2}) \tag{6.19}$$

[the first two of which follow from Proposition 2.1 with the predictor $\mathbf{X}_J$ and the last of which is proved by repeating the argument of Theorem 1 of Samarov (1993) for the functional $Ew(\mathbf{X})m(\mathbf{X})m_J(\mathbf{X}_J)$], we obtain that $I = o_p(n^{-1/2})$, which implies (6.13).

Now, writing $\hat{\varepsilon}_{Ji} = \hat{\varepsilon}_i + \hat{m}_i - \hat{m}_{Ji}$, we have

$$\begin{aligned} &n^{-1} \sum (\hat{m}_i - \hat{m}_{Ji} - (\overline{m} - \overline{m}_J))(\hat{\varepsilon}_{Ji} - \bar{\varepsilon}_J) w_i \\ &\quad = n^{-1} \sum (\hat{m}_i - \hat{m}_{Ji} - (\overline{m} - \overline{m}_J))^2 w_i \\ &\qquad + n^{-1} \sum (\hat{m}_i - \hat{m}_{Ji} - (\overline{m} - \overline{m}_J))(\hat{\varepsilon}_i - \bar{\varepsilon}) w_i. \end{aligned} \tag{6.20}$$

It follows from Proposition 2.1 that

$$n^{-1}\sum(\hat{m}_i - \overline{m})(\hat{\varepsilon}_i - \overline{\varepsilon})w_i = o_p(n^{-1/2})$$

and

$$n^{-1}\sum(\hat{m}_{Ji} - \overline{m}_J)(\hat{\varepsilon}_{Ji} - \overline{\varepsilon}_J)w_i = o_p(n^{-1/2});$$

so, using formulas (6.15)–(6.20), we have

$$\begin{aligned}
&n^{-1}\sum(\hat{m}_i - \hat{m}_{Ji} - (\overline{m} - \overline{m}_J))(\hat{\varepsilon}_i - \overline{\varepsilon})w_i \\
&\quad = n^{-1}\sum(\hat{m}_{Ji} - \overline{m}_J)(\hat{\varepsilon}_i - \overline{\varepsilon})w_i + o_p(n^{-1/2}) \\
&\quad = -n^{-1}\sum(\hat{m}_{Ji} - \overline{m})\big(\hat{\varepsilon}_{Ji} - \overline{\varepsilon}_J - \hat{m}_i + \hat{m}_{Ji} + \overline{m} - \overline{m}_J\big)w_i + o_p(n^{-1/2}) \\
&\quad = n^{-1}\sum(\hat{m}_{Ji} - \overline{m})(\hat{m}_i - \hat{m}_{Ji} - (\overline{m} - \overline{m}_J))w_i + o_p(n^{-1/2}) \\
&\quad = -\tfrac{1}{2}I + o_p(n^{-1/2}) = o_p(n^{-1/2}),
\end{aligned}$$

and (6.14) follows. □

Acknowledgments. The authors would like to thank Sam Choi and Hongyu Zhao for helping with computing. The paper was in part written while Kjell Doksum was visiting the Department of Statistics, Harvard University, and the Dana Farber Cancer Institute, Division of Biostatistics, Harvard Medical School. We thank an Associate Editor and a referee for their useful comments and suggestions.

REFERENCES

ABRAMSON, I. and GOLDSTEIN, L. (1991). Efficient nonparametric testing by functional estimation. *J. Theoret. Probab.* **4** 137–159.

AZZALINI, A., BOWMAN, A. and HÄRDLE, W. (1989). On the use of nonparametric regression for model checking. *Biometrika* **76** 1–11.

BHANSALI, R. (1993). Estimation of the prediction error variance and an R^2 measure by autoregressive model fitting. *J. Time Ser. Anal.* **14** 125–146.

BICKEL, P. J., KLAASEN, C. A. J., RITOV, Y. and WELLNER, J. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Johns Hopkins Univ. Press.

BICKEL, P. and RITOV, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā Ser. A* **50** 381–393.

BREIMAN, L. and FRIEDMAN, J. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–598.

BREIMAN, L. and MEISEL, W. (1976). General estimates of the intrinsic variability of data in nonlinear regression models. *J. Amer. Statist. Assoc.* **71** 301–308.

BUCKLEY, M., EAGLESON, G. and SILVERMAN, B. (1988). The estimation of residual variance in nonparametric regression. *Biometrika* **75** 189–199.

BUJA, A. (1990). Remarks on functional canonical variables, alternating least squares methods and ACE. *Ann. Statist.* **18** 1032–1069.

CHENG, B. and TONG, H. (1993). On residual sums of squares in non-parametric autoregression. *Stochastic Process. Appl.* **48** 157–174.

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

CRAMÉR, H. (1945). *Mathematical Methods of Statistics.* Almquist and Wiksells, Uppsala.

DOKSUM, K. and SAMAROV, A. (1993). Global functionals and a measure of the explanatory power of covariates in nonparametric regression. Technical Report 255-93, IFSRC, Sloan School of Management, MIT.

DONOHO, D. and LIU, R. (1991). Geometrizing rates of convergence, II, III. *Ann. Statist.* **19** 633–667; 668–701.

DONOHO, D. and NUSSBAUM, M. (1990). Minimax quadratic estimation of a quadratic functional. *J. Complexity* **6** 290–323.

EUBANK, R. and SPIEGELMAN, C. (1990). Testing the goodness of fit of a linear model via nonparametric regression techniques. *J. Amer. Statist. Assoc.* **85** 387–392.

FAN, J. (1991). On the estimation of quadratic functionals. *Ann. Statist.* **19** 1273–1294.

FAN, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87** 998–1004.

FAN, J. and MARRON, J. S. (1992). Best possible constants for bandwidth selection. *Ann. Statist.* **20** 2057–2070.

FRIEDMAN, J. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

GASSER, T., SROKA, L. and JENNEN-STEINMETZ, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73** 625–633.

GOLDSTEIN, L. and MESSER, K. (1992). Optimal plug-in estimators for nonparametric functional estimation. *Ann. Statist.* **20** 1306–1328.

HALL, P. (1989). On projection pursuit regression. *Ann. Statist.* **17** 573–588.

HALL, P. and JOHNSTONE, I. (1992). Empirical functionals and efficient smoothing parameter selection. *J. Roy. Statist. Soc. Ser. B* **54** 475–530.

HALL, P. and MARRON, J. S. (1987). Estimation of integrated squared density derivatives. *Statist. Probab. Lett.* **6** 109–115.

HALL, P. and MARRON, J. S. (1990). The estimation of residual variance in nonparametric regression. *Biometrika* **77** 415–419.

HÄRDLE, W. (1990). *Applied Nonparametric Regression.* Cambridge Univ. Press.

HÄRDLE, W., HART, J., MARRON, J. and TSYBAKOV, A. (1992). Bandwidth choice for average derivative estimation. *J. Amer. Statist. Assoc.* **87** 227–233.

HÄRDLE, W. and STOKER, T. (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84** 986–995.

HASMINSKII, R. and IBRAGIMOV, I. (1979). On the nonparametric estimation of functionals. In *Proceedings of the Second Prague Symposium on Asymptotic Statistics* (P. Mandl and M. Hušková, eds.) 41–51. North-Holland, Amsterdam.

IBRAGIMOV, I., NEMIROVSKY, A. and KHASMINSKII, R. (1986). Some problems of nonparametric estimation in Gaussian white noise. *Theory Probab. Appl.* **31** 391–406.

JOE, H. (1989). Estimation of entropy and other functionals of a multivariate density. *Ann. Inst. Statist. Math.* **41** 683–697.

JONES, M. and SHEATHER, S. (1991). Using non-stochastic terms to advantage in kernel-based estimation of integrated density derivatives. *Statist. Probab. Lett.* **11** 511–514.

KENDALL, M. G. and STUART, A. (1962). *The Advanced Theory of Statistics* **2**. Hafner, New York.

KOLMOGOROV, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Springer, Berlin.

KOSHEVNIK, Y. and LEVIT, B. (1976). On the nonparametric analog of the information matrix. *Theory Probab. Appl.* **21** 759–774.

KOZEK, A. (1991). A nonparametric test of fit of a parametric model. *J. Multivariate Anal.* **37** 66–75.

KRUSKAL, W. H. (1958). Ordinal measures of association. *J. Amer. Statist. Assoc.* **53** 814–861.

LE CAM, L. (1956). On the asymptotic theory of estimation and testing hypotheses. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 129–156. Univ. California Press, Berkeley.

LEVIT, B. (1978). Asymptotically efficient estimation of nonlinear functionals. *Problems Inform. Transmission* **14** 204–209.

MARRON, J. S. and HÄRDLE, W. (1986). Random approximations to some measures of accuracy in nonparametric curve estimation. *J. Multivariate Anal.* **20** 91–113.

PARK, B. and MARRON, J. S. (1990). Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.* **85** 66–72.

PASTUKHOVA, YU. and KHASMINSKII, R. (1989). Estimation of nonlinear functionals from the regression function with the possibility of the regressor's design. *Problems Control Inform. Theory* **18** 65–77.

PEARSON, K. (1905). On the general theory of skew correlation and nonlinear regression. *Proc. Roy. Soc. London* **71** 303. (Draper's Research Memoires, Dulan & Co, Biometrics Series II.)

PITMAN, E. J. G. (1948). *Lecture Notes in Nonparametric Statistics*. Columbia Univ. Press.

PRAKASA RAO, B. L. S. (1983). *Nonparametric Functional Estimation*. Academic Press, New York.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York.

RÉNYI, A. (1959). On measures of dependence. *Acta. Math. Acad. Sci. Hungar.* **10** 441–451.

ROBINSON, P. (1991). Consistent nonparametric entropy-based testing. *Rev. Econom. Stud.* **58** 437–453.

SAMAROV, A. M. (1993). Exploring regression structure using non-parametric functional estimation. *J. Amer. Statist. Assoc.* **88** 836–849.

SCHWEDER, T. (1975). Window estimation of the asymptotic variance of rank estimators of location. *Scand. J. Statist.* **2** 113–126.

SHEATHER, S., HETTSMANSPERGER, T. and DONALD, M. (1994). Data based bandwidth selection for kernel estimators of the integral of $f^2(x)$. *Scand. J. Statist.* **21** 265–276.

SILVERMAN, B. (1985). Some aspects of spline smoothing approach to non-parametric curve fitting (with discussion). *J. Roy. Statist. Soc. Ser. B* **47** 1–52.

SKAUG, H. J. and TJÖSTHEIM, D. (1993). Nonparametric tests of serial independence. In *Developments in Time Series Analysis (in Honour of Maurice B. Priestly)* (T. S. Rao, ed.) 207–209. Chapman and Hall, London.

STANISWALIS, J. and SEVERINI, T. (1991). Diagnostics for assessing regression models. *J. Amer. Statist. Assoc.* **86** 684–692.

STONE, C. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

STUTE, W. (1984). Asymptotic normality of nearest neighbor regression function estimates. *Ann. Statist.* **12** 917–926.

YANG, S. (1981). Linear functions of concomitants of order statistics with applications to nonparametric estimation of a regression function. *J. Amer. Statist. Assoc.* **76** 658–662.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720

SLOAN SCHOOL OF MANAGEMENT
MASSACHUSETTS INSTITUTE
OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS 02139