CHAPTER 1

LINEARLY UNBIASED ESTIMATION OF CONDITIONAL MOMENT AND CORRELATION FUNCTIONS

Hans-Georg Müller

Department of Statistics
University of California, Davis, CA, U.S.A

E-mail: mueller@wald.ucdavis.edu

We consider a random-design regression model with vector-valued observations and develop nonparametric estimation of smooth conditional moment functions in the predictor variable. This includes estimation of higher order mixed moments and also functionals of the moments, such as conditional covariance, correlation, variance, and skewness functions. Our asymptotic analysis targets the limit distributions. We find that some seemingly reasonable procedures do not reproduce the identity or other linear functions without undesirable bias components, i.e., they are *linearly biased*. Alternative *linearly unbiased* estimators are developed which remedy this bias problem without increasing the variance. A general linearly unbiased estimation scheme is introduced for arbitrary smooth functionals of moment functions.

Key words: Covariance function; Moment functional; Identity reproducing estimation; Local linear fitting; Mean squared error; Nonparametric regression; Skewness; Smoothing; Variance function.

1. Introduction

We consider the situation of a nonparametric regression model with a random predictor X and a vector of dependent variables $Y \in \Re^p$, $p \geq 1$. It is assumed that one observes a sample of n pairs (X_i, Y_i), $i = 1, ..., n$, of independent and identically distributed (i.i.d.) bivariate data, drawn from a joint distribution $F(u, v)$. Extending the basic problems of estimating the mean regression function $E(Y|X = x)$ for univariate responses Y [Fan and Gijbels (1996), Wand and Jones (1995)] or of estimating a variance function $\text{var}(Y|X = x)$ [Müller and Stadtmüller (1993)], we consider estimation of mixed conditional moment functions of the type

$$\mu_\alpha(x) = E(Y^\alpha|X = x) = E(Y_1^{\beta_1} Y_2^{\beta_2} ... Y_p^{\beta_p}|X = x), \quad (1)$$

where $\alpha = (\beta_1, ..., \beta_p)$ is a multi-index of nonnegative integers and $Y = (Y_1, ..., Y_p)^T$. Given a number of $k \geq 1$ such conditional moments $\mu_{\alpha_1}, ..., \mu_{\alpha_k}$, and a smooth mapping G from $\Re^k$ to $\Re$, our main object of interest is the functional

$$g(x) = G\left(\mu_{\alpha_1}(x), ..., \mu_{\alpha_k}(x)\right). \quad (2)$$

Interest in estimating the function $g(\cdot)$ is motivated by the following examples. (Whenever $p = 1$, we write Y for Y_1.)

Example 1: *Conditional Moment Function.*
For $k = 1$, $p = 1$, $\alpha_1 = 1$, and $G(x) = x$, one has the *conditional moment function*

$$\mu_\ell(x) = E(Y^\ell|X = x), \quad (3)$$

which includes the classical regression function for $\ell = 1$.

Example 2: *Conditional Variance Function.*
For $k = 2$, $p = 1$, $\alpha_1 = 2$, $\alpha_2 = 1$ and $G(x, y) = x - y^2$, we obtain the *conditional variance function*

$$\begin{aligned} g(x) = v(x) = var(Y|X = x) &= E(Y^2|X = x) - \{E(Y|X = x)\}^2 \\ &= \mu_2(x) - (\mu_1(x))^2. \end{aligned} \quad (4)$$

Example 3: *Conditional Skewness Function.*
The choices $k = 3$, $p = 1$, $\alpha_1 = 1$, $\alpha_2 = 2$, $\alpha_3 = 3$ and $G(x_1, x_2, x_3) = (x_3 - 3x_1x_2 + 2x_1^3)/((x_2 - x_1^2)^{3/2})$ lead to the *conditional skewness function*

$$g(x) = s(x) = E\{(Y - \mu(x))^3|X = x\} = \left\{\frac{\mu_3 - 3\mu_2\mu_1 + 2\mu_1^3}{(\mu_2 - \mu_1^2)^{3/2}}\right\}(x). \quad (5)$$

Example 4: *Conditional Covariance Function.*
For $k = 3$, $p = 2$, $\alpha_1 = (1, 1)$, $\alpha_2 = (1, 0)$, $\alpha_3 = (0, 1)$ and $G(x_1, x_2, x_3) = x_1 - x_2x_3$, we arrive at the *conditional covariance function*

$$g(x) = v_{12}(x) = E(Y_1Y_2|X = x) - E(Y_1|X = x)E(Y_2|X = x). \quad (6)$$

Example 5: *Conditional Correlation Function.*
For $k = 5$, $p = 2$, $\alpha_1 = (1,1)$, $\alpha_2 = (1,0)$, $\alpha_3 = (0,1)$, $\alpha_4 = (2,0)$, $\alpha_5 = (0,2)$ and $G(x_1, x_2, x_3, x_4, x_5) = (x_1 - x_2x_3)/((x_4 - x_2^2)(x_5 - x_3^2))^{1/2}$, we obtain the *conditional correlation function*

$$g(x) = \rho_{12}(x) = \frac{v_{12}(x)}{\{v_{11}(x)v_{22}(x)\}^{1/2}}, \tag{7}$$

where $v_{k\ell}(x) = E(Y_kY_\ell|X = x) - E(Y_k|X = x)E(Y_\ell|X = x)$.

Conditional moment functions, especially for the first moment, are naturally of interest in applications of nonparametric regression. The variance function has long been recognized as a valuable tool in applied statistics. Applications include the construction of confidence regions, adjustments of least squares estimators in parametric regression models to heteroscedasticity, local bandwidth selection, and volatility modeling [see, e.g., Dette and Munk (1998), Eubank and Thomas (1993), Fan and Yao (1998), Müller and Stadtmüller (1987) and Picard and Tribouley (2000)].

For example, the estimation of a conditional skewness function will be of particular interest in cases where the skewness of responses Y changes sign for varying predictors x, in which case an overall skewness estimate is less meaningful. In an obvious manner, conditional curtosis functions, conditional cumulant functions and conditional moment generating functions can be defined along the same lines. Conditional covariance and correlation functions are of particular interest. These functions are relevant whenever the response is multivariate, and when the relation between any two response variables changes as a predictor variable varies.

Another type of conditional correlation function has been introduced by Bjerve and Doksum (1993) and was further analyzed in Doksum et al. (1994) and Doksum and Samarov (1995). The topic of these papers is measuring the local strength of a relation between a univariate dependent variable Y and a univariate predictor variable X, which is modeled as varying with predictor value. In contrast, our focus here is on the dependency of the correlation between two response variables, conditional on the level of a predictor variable. This dependency could be characterized as a conditional partial correlation function, which we will model nonparametrically in the following.

A criterion which has been shown to be of practical as well as theoretical interest for discriminating between various possible function estimates focuses on whether an estimator can reproduce a linear function with a zero leading (first order) bias term. This is a desirable feature, as then bias is controlled irrespective of the locations and design of the predictors. Different aspects of this property with regard to the comparison of specific smoothers have been pointed out by various authors, among them Chu and Marron (1991), Jennen-Steinmetz and Gasser (1987), Jones and Park (1994), and Müller (1997). Specifically, the notion of an *identity reproducing function estimator* was introduced in Müller and Song (1993), where it was shown how given function estimators can be modified by incorporating an identity reproducing transformation in such a way that they become identity reproducing. Additional results along these lines were obtained by Mammen and Marron (1997) and Park, Kim and Jones (1997). A defining feature is that scatterplot data (X_i, X_i), fed into an identity-reproducing estimator of $E(Y|X = x)$, will return the identity function as function estimate, which is of course the true underlying function in this situation.

We formalize here the notion of vanishing leading bias terms in the following way: We call a function estimator $\hat{g}(x)$ of $g(x)$ *linearly unbiased*, if the leading term of its asymptotic bias is proportional to the second derivative $g^{(2)}(x)$ and does not involve any further terms depending on $g(\cdot)$ or the joint distribution $F(\cdot, \cdot)$ of X and Y. This notion is motivated by several appealing properties of linearly unbiased estimators: not only do they reproduce the identity function, but they also are associated with a bias structure that is predictable from the curve estimate itself, since bias depends only on the second derivative of the function g that is to be estimated; in particular the bias does not depend on properties of the underlying design such as the density of the design points. For example, for Nadaraya-Watson quotient type kernel estimators of the mean regression function it is well known that these estimators are linearly biased, while convolution type kernel estimators and local polynomial smoothers are linearly unbiased [Bhattacharya and Müller (1993)].

The paper is organized as follows. In Section 2, we provide further details on linearly unbiased curve estimators. We then state the main results and obtain a general construction for linearly unbiased estimates of conditional moment functionals in Section 3. In Section 4, we provide examples of linearly unbiased estimators, including estimators for skewness, covariance and correlation functions. A simulation example is included in Section 5, while proofs, auxiliary results and assumptions can be found in Section 6.

2. Preliminaries on linearly unbiased curve estimators

A generalized version of the concept of linear unbiasedness is *r-th order polynomial unbiasedness*, which would imply that the leading term of the asymptotic bias is proportional to $g^{(r+1)}(x)$. Examples of such estimators are provided by local polynomial fitting of polynomials of degree $(r+1)$, applied for estimating functions that are $(r+1)$ times continuously differentiable, and also by convolution kernel estimators using kernels of order $(r+1)$ [see Gasser, Müller and Mammitzsch (1985)]. When targeting shapes other than linear or polynomial, other notions of target unbiasedness might be of interest, extending the concept of polynomial unbiasedness to other functional shapes which one desires to estimate without leading bias terms. For the sake of simplicity, we consider here only the case $r = 1$, corresponding to linear unbiasedness.

Note that linear or first order unbiasedness implies that the bias, when estimating linear functions $g(\cdot)$, corresponds to a relatively small remainder term. This is of course a highly desirable requirement in nonparametric curve estimation, since parametric estimates based on the assumption of a linear underlying function will be unbiased when the underlying function to be estimated is indeed linear. If this is not the case, however, then such parametric estimates will be inconsistent. The idea is that the price one pays in terms of bias in nonparametric estimation, which is much more flexible than parametric modeling and yields consistent estimates as long as the underlying function is smooth, should be kept reasonably small when estimating functions with common parametric shapes.

If linear unbiasedness is not satisfied, curve estimates will show unpredictable systematic deviations that are often dependent on the underlying design, whereas under linear unbiasedness, the leading bias term depends only on the local curvature of the function to be estimated. This bias can then be at least roughly assessed from the estimated function. For the special case of estimating a variance function, the need for a bias correction of this sort has been recognized in Ruppert et al. (1997) and accordingly included in their estimation procedure by dividing smoothed squared residuals by a constant.

The analysis in Ruppert et al. (1997) also provides one of many examples of the commonly adopted conditional approach, where one focuses on the behavior of the conditional mean squared errors $E\{(\hat{g}(X) - g(X))^2|X_1, ..., X_n\}$. While such conditional measures of performance are valuable in their own right, they can only provide partial reassurance to a user who encounters new data with different designs. For instance, it is well known, and indeed corresponds to a practical problem, that unconditional mean squared error does not even exist for local polynomial smoothers, including the Nadaraya-Watson kernel estimator, while it does exist for convolution type kernel estimators (Seifert and Gasser (1996) discuss these issues in great detail). Unconditional asymptotics for local polynomial fitting and related estimation methods can still be achieved by discarding moment based criteria such as mean squared error and instead focussing on asymptotic bias and variance, defined as the bias and variance obtained from an asymptotic limiting distribution; this is the approach we adopt here. In fact, a prime example where the large sample limit of conditional bias and variance differs from asymptotic bias and variance defined in this sense is provided by local polynomial fitting.

Assume a sample of i.i.d. random vectors $(X_i, Y_i) \in \Re^{p+1}$, $i = 1, ..., n$, is given, and consider the problem of estimating the conditional moment function $\mu_\alpha(x) = E(Y^\alpha | X = x)$ for a fixed x in the domain of g. Given a sequence of bandwidths $b > 0$ and a kernel or weight function $K \geq 0$, we define kernel weights

$$w_i(x) = (nb)^{-1} K\{b^{-1}(x - X_i)\}, \tag{8}$$

usually assuming that for the sequence of bandwidths $b \to 0, \quad nb \to \infty$ as $n \to \infty$, and that the kernel K is a square integrable probability density function with finite variance that is centered around 0.

The most common estimators are linear smoothers of the form

$$\hat{\mu}_\alpha(x) = \sum_{i=1}^{n} W_i(x) Y_i^\alpha, \tag{9}$$

i.e., weighted averages of the responses, where the $W_i(\cdot)$ are weight functions which characterize a particular smoothing method. Linear smoothers include splines and kernel estimators. For *Nadaraya-Watson kernel estimators* which are of quotient type, the smoothing weights W_i are explicitly given by

$$W_{i,NW}(x) = w_i(x) / \sum_{j=1}^{n} w_j(x), \tag{10}$$

leading to estimates $\hat{\mu}_{\alpha,NW}$.

A second form of kernel estimation is local linear fitting by weighted least squares. Here the smoothing weight functions W_i are obtained by solving the weighted least squares problem (compare Fan and Gijbels (1996))

$$\hat{\mu}_{\alpha,LS}(x) = \arg\min_{a_0} \left[\min_{a_1} \left\{ \sum_{i=1}^{n} w_i(x) \left[Y_i^{\alpha} - (a_0 + a_1 (X_i - x)) \right]^2 \right\} \right], \quad (11)$$

for which the explicit smoothing weights are found to be

$$W_{i,LS}(x) = \frac{w_i(x)}{\sum_{j=1}^n w_j(x)} - \frac{\sum_{j=1}^n w_j(x)(X_j - x)}{\sum_{j=1}^n w_j(x)} \times \left[\frac{w_i(x)(X_i - x)\sum_{j=1}^n w_j(x) - w_i(x)\sum_{j=1}^n w_j(x)(X_j - x)}{\sum_{j=1}^n w_j(x)\sum_{j=1}^n w_j(x)(X_j - x)^2 - (\sum_{j=1}^n w_j(x)(X_j - x))^2} \right]. \quad (12)$$

One obtains the asymptotic distributions of both of these estimates under suitable regularity conditions (compare Bhattacharya and Müller (1993)) as

$$(nb)^{1/2} \left[\hat{\mu}_{\alpha}(x) - \mu_{\alpha}(x) \right] \to \mathcal{N}(B, V) \quad (13)$$

in distribution. Here the expression for the asymptotic variance is the same for both estimators $\hat{\mu}_{\alpha,NW}$ and $\hat{\mu}_{\alpha,LS}$, and is given by

$$V = f_X^{-1}(x) \left[\mu_{2\alpha}(x) - (\mu_{\alpha}(x))^2 \right] \int K^2(u)\, du, \quad (14)$$

where $f_X(\cdot)$ denotes the marginal density of X. However, the asymptotic bias terms B differ between Nadaraya-Watson and local least squares estimators. Assuming $nb^5 \to d^2$ for a constant $d \geq 0$, we find for the asymptotic bias term B_{NW} of Nadaraya-Watson kernel estimators and for the bias term B_{LS} for local linear fitting that

$$B_{NW} = \frac{d}{2} \frac{\mu_{\alpha}^{(2)}(x) f_X(x) + 2\mu_{\alpha}^{(1)}(x) f_X^{(1)}(x)}{f_X(x)} \int K(u) u^2\, du \quad (15)$$

and

$$B_{LS} = \frac{d}{2} \mu_{\alpha}^{(2)}(x) \int K(u) u^2\, du. \quad (16)$$

We conclude that Nadaraya-Watson estimators $\hat{\mu}_{\alpha,NW}$ are linearly biased for μ_{α}, while local linear estimators $\hat{\mu}_{\alpha,LS}$ are linearly unbiased. This is not entirely surprising, given that Nadaraya-Watson kernel estimators can be derived as the local weighted least squares solutions of fitting local constants, which naturally leads to less flexible biases as compared to fitting local least squares lines.

3. Main results on linearly unbiased estimation

For the following, we assume that regularity conditions (C1)-(C8), listed in section 6, are in force. Relevant for the formulation of the following result on the estimation of functionals of moment functions is the condition $nb^5 \to d^2$, as $n \to \infty$ for a $d > 0$. We consider functionals (2), $g(x) = G\left(\mu_{\alpha_1}(x), ..., \mu_{\alpha_k}(x)\right)$, as described in the Introduction. In order to estimate $g(x)$, a natural approach are the plug-in estimators

$$\hat{g}(x) = G\{\hat{\mu}_{\alpha_1,LS}(x), ..., \hat{\mu}_{\alpha_k,LS}(x)\}. \tag{17}$$

As the following result demonstrates, these estimators generally do not possess the desirable property of linear unbiasedness. Specifically, writing $\mu = (\mu_{\alpha_1}, ..., \mu_{\alpha_k})$, and using the abbreviations $c_B = \frac{1}{2}d\int K(u)u^2du$ and $c_V = \int K^2(u)du$, we have the following.

Theorem 1: Under regularity conditions (C1)-(C8),

$$(nb)^{1/2}\{\hat{g}(x) - g(x)\} \to \mathcal{N}(\tilde{B}, \tilde{V})$$

in distribution, where

$$\tilde{B} = c_B \sum_{m=1}^{k} \left\{ \frac{dG}{dx_m}|_\mu \frac{d^2}{d^2x}\mu_{\alpha_m}(x) \right\},$$

$$\tilde{V} = \frac{c_V}{f_X(x)} \left(\sum_{l,m=1}^{k} \frac{dG}{dx_l}|_\mu \frac{dG}{dx_m}|_\mu \right) \{\mu_{\alpha_l+\alpha_m}(x) - \mu_{\alpha_l}(x)\mu_{\alpha_m}(x)\}.$$

The proof of this and the next theorem can be found in Section 6. As a simple illustration of this result, consider the conditional covariance function (Example 4, eq.(6))

$$v_{12}(x) = E(Y_1Y_2|X = x) - E(Y_1|X = x)E(Y_2|X = x).$$

The plug-in estimator is $\hat{v}_{12}(x) = \hat{\mu}_{11}(x) - \hat{\mu}_{10}(x)\hat{\mu}_{01}(x)$ and Theorem 1 leads to

$$(nb)^{1/2}\{\hat{v}_{12}(x) - v_{12}(x)\} \to \mathcal{N}(\tilde{B}_{12}, \tilde{V}_{12})$$

in distribution, where

$$\begin{aligned}
\tilde{B}_{12} &= c_B(\mu_{11}^{(2)} - \mu_{01}\mu_{10}^{(2)} - \mu_{01}^{(2)}\mu_{10})(x), \\
\tilde{V}_{12} &= \frac{c_V}{f_X(x)}\{\mu_{22} - \mu_{11}^2 + \mu_{01}^2(\mu_{02} - \mu_{01}^2) + \mu_{10}^2(\mu_{20} - \mu_{10}^2) \\
&\quad - 2\mu_{01}(\mu_{21} - \mu_{11}\mu_{10}) - 2\mu_{10}(\mu_{12} - \mu_{11}\mu_{01}) \\
&\quad + 2\mu_{10}\mu_{01}(\mu_{11} - \mu_{10}\mu_{01})\}(x).
\end{aligned}$$

As

$$v_{12}^{(2)}(x) = \{\mu_{11}^{(2)} - (\mu_{10}^{(2)}\mu_{01} + 2\mu_{10}^{(1)}\mu_{01}^{(1)} + \mu_{10}\mu_{01}^{(2)})\}(x),$$

this estimator is found to be linearly biased. The problem is that in the asymptotic bias $\tilde{B}_{12}$ the term $-2\mu_{10}^{(1)}\mu_{01}^{(1)}(x)$ is missing.

In order to achieve linear unbiasedness for the general case, one needs to target the asymptotic bias term

$$\begin{aligned} B^* &= c_B\, g^{(2)}(x) \\ &= c_B \sum_{m=1}^{k} \left\{ \frac{dG}{dx_m} |_\mu \frac{d^2}{d^2x}\mu_{\alpha_m}(x) \right\} \\ &\quad + c_B \sum_{l,m=1}^{k} \frac{d^2G}{dx_l dx_m} |_\mu \frac{d}{dx}\mu_{\alpha_l}(x) \frac{d}{dx}\mu_{\alpha_m}(x). \end{aligned} \tag{18}$$

Therefore the problem of linear bias in estimates $\hat{g}$ arises because the second summand in the desirable asymptotic bias term B^* is missing from the actual bias term $\tilde{B}$.

To remedy this problem, we introduce the bias correction term,

$$\hat{\Delta}(x) = \frac{1}{2} s_x^2 \sum_{l,m=1}^{k} \left\{ \frac{d^2G}{dx_l dx_m} |_{\hat{\mu}}\ \hat{\delta}_{\alpha_l}(x)\hat{\delta}_{\alpha_m}(x) \right\}, \tag{19}$$

where

$$\hat{\delta}_\alpha(x) = \frac{\sum_{i=1}^n w_i(x)(X_i - x)Y_i^\alpha \sum_{i=1}^n w_i(x) - \sum_{i=1}^n w_i(x)(X_i - x)\sum_{i=1}^n w_i(x)Y_i^\alpha}{\sum_{i=1}^n w_i(x)(X_i - x)^2 \sum_{i=1}^n w_i(x) - \left[\sum_{i=1}^n w_i(x)(X_i - x)\right]^2} \tag{20}$$

and

$$s_x^2 = \frac{\sum_{i=1}^n w_i (X_i - x)^2}{\sum_{i=1}^n w_i} - \left\{ \frac{\sum_{i=1}^n w_i (X_i - x)}{\sum_{i=1}^n w_i} \right\}^2. \tag{21}$$

We then propose the bias corrected estimators

$$\hat{g}^*(x) = \hat{g}(x) + \hat{\Delta}(x) = G(\hat{\mu}_{\alpha_1}, ..., \hat{\mu}_{\alpha_k}) + \hat{\Delta}(x). \tag{22}$$

Interestingly, this bias correction has no effect on the variance, and the modified estimator $\hat{g}^*$ is justified by the following result.

Theorem 2: Under regularity conditions (C1)-(C8),

$$(nb)^{1/2}\{\hat{g}^*(x) - g(x)\} \to \mathcal{N}(B^*, V^*)$$

in distribution, where

$$B^* = c_B g^{(2)}(x), \quad V^* = \tilde{V}.$$

Continuing the example of the covariance function, the bias-corrected estimator is given by

$$\hat{v}^*_{12}(x) = \hat{\mu}_{12}(x) - \hat{\mu}_{10}(x)\hat{\mu}_{01}(x) - s_x^2\hat{\delta}_{01}(x)\hat{\delta}_{10}(x).$$

According to Theorem 2, this estimator has the desirable asymptotic bias term $B^*_{12} = c_B v_{12}^{(2)}(x)$ and therefore is linearly unbiased.

4. Variance, skewness and correlation function estimation

We resume the discussion of the examples given in the Introduction in the light of Theorem 2. Multi-indices are replaced by single indices if only one coordinate of the response vector Y needs to be considered; for simplicity the first coordinate is chosen by default. Regarding the simple moment functions (3), $\mu_\ell(x) = E(Y_1^\ell | X = x)$, we find that $\hat{\delta} \equiv 0$, so that according to Theorem 2, the plug-in estimators $\hat{\mu}_{\alpha,LS}$ defined in (11), (12) are linearly unbiased. This can also be seen from the fact that G is the identity function in this case. Indeed, applying Theorem 1 yields an asymptotic normal distribution with

$$B_\ell = c_B \mu_\ell^{(2)}(x), \qquad V_\ell = \frac{c_V}{f_X(x)}[\mu_{2\ell}(x) - \{\mu_\ell(x)\}^2]. \tag{23}$$

We note that in the general case, the asymptotic quantities B^*, V^* or $\tilde{B}$, $\tilde{V}$ are useful for the construction of approximate asymptotic confidence intervals for estimates $\hat{g}^*(x)$ (22) or $\hat{g}(x)$ (17). There exist various possibilities for the actual construction of such asymptotic pointwise confidence intervals. Common approaches include to simply ignore the asymptotic bias, centering the asymptotic confidence intervals symmetrically around the curve estimates, or to use undersmoothing for the construction of confidence intervals, so as to justify that bias becomes asymptotically negligible and can be safely ignored. Other approaches are to approximate the asymptotic variance in the limiting normal distribution by the square root of mean squared error, while centering the intervals around the curve estimates, in an effort to make the intervals wider to account for the bias, or bootstrapping (see, e.g., Claeskens and van Keilegom (2003), Eubank and Speckman (1993), Hall (1992) and Picard and Tribouley (2000)).

Continuing Example 2 concerning the nonparametric estimation of variance functions $v(x) = Var(Y_1|X = x)$ (4), the plug-in estimator is

$$\hat{v}(x) = \hat{\mu}_{2,LS}(x) - \{\hat{\mu}_{1,LS}(x)\}^2. \tag{24}$$

With the local weighted least squares smoothing weight functions $W_{i,LS}(x)$ (12) we may write

$$\begin{aligned} \hat{v}(x) &= \sum_{i=1}^{n} W_{i,LS}(x) Y_{1i}^2 - \left\{ \sum_{i=1}^{n} W_{i,LS}(x) Y_{1i} \right\}^2 \\ &= \sum_{i=1}^{n} W_{i,LS}(x) \{Y_{1i} - \hat{\mu}(x)\}^2, \end{aligned} \tag{25}$$

so that this estimator is seen to be equivalent to smoothing squared residuals obtained from an initial local linear fit.

Another proposal for variance function estimation that is closely related to the work of Doksum et al. (1994) and Doksum and Samarov (1995) is to estimate the variance function by using the error mean square as in classical regression, but now formed from properly weighted residuals within the local smoothing window. The starting point is the well-known classical formula for the mean square due to error, $MSE = (\tilde{s}_y^2 - \hat{b}_1^2 \tilde{s}_x^2)(n-1)/(n-2)$ in simple linear regression, where $\tilde{s}_y^2 = (n-1)^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2$, $\tilde{s}_x^2 = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ and $\hat{b}_1$ is the least squares estimate of the slope parameter of the regression line. Since $E(MSE) = \sigma^2 = \mathrm{var}(Y_i)$, this relationship can be exploited for estimating a local variance function. Localizing and introducing weights, this motivates the estimate

$$\hat{v}_{MSE}(x) = \frac{\sum_{i=1}^{n} w_i(x) Y_i^2}{\sum_{i=1}^{n} w_i(x)} - \left\{ \frac{\sum_{i=1}^{n} w_i(x) Y_i}{\sum_{i=1}^{n} w_i(x)} \right\}^2 - \hat{\delta}_1(x)^2 s_x^2, \tag{26}$$

with $w_i, \hat{\delta}_1$ and s_x^2 as defined in (8), (20) and (21).

Finally, the bias-corrected version of estimator (25) is obtained as

$$\hat{v}^*(x) = \hat{v}(x) + \hat{\Delta}(x) = \hat{\mu}_2(x) - \{\hat{\mu}_1(x)\}^2 - s_x^2 \{\hat{\delta}_1(x)\}^2. \tag{27}$$

We note that for all three estimators $\hat{v}$, $\hat{v}_{MSE}$ and $\hat{v}^*$, we obtain the same asymptotic variance term

$$V_V = f_X^{-1}(x) c_V \{\mu_4 - \mu_2^2 + 4(\mu_2 - \mu_1^2)\mu_1^2 - 4\mu_1(\mu_3 - \mu_1\mu_2)\}(x). \tag{28}$$

For $\hat{v}$ and $\hat{v}^*$ this follows from Theorems 1 and 2, and for $\hat{v}_{MSE}$ it is shown in Section 6. For these estimators, a more interesting comparison concerns the leading asymptotic bias terms. According to Theorems 1, 2, we find for $\hat{v}$ and $\hat{v}^*$,

$$\tilde{B} = c_B\{v^{(2)} + 2\mu_1^{(1)2}\}(x), \qquad B^* = c_B v^{(2)}, \tag{29}$$

the latter being the desired expression for linearly unbiased estimation. As seen in Section 6, the asymptotic bias term for $\hat{v}_{MSE}$ is

$$B_{MSE} = c_B\{v^{(2)} + 2v^{(1)}\frac{f_X^{(1)}}{f_X}\}(x). \tag{30}$$

The estimator $\hat{v}_{MSE}$ thus is seen to suffer from the drawback that the bias depends on the marginal density f_X which means artificial curvature in the curve estimates may be introduced by just replacing a uniform marginal distribution of the predictor variable X by a normal marginal distribution.

Continuing now the discussion of the conditional skewness function (5) in Example 3, the straightforward "plug-in" estimate is

$$\hat{s}(x) = \left\{ \frac{\hat{\mu}_{3,LS} - 3\hat{\mu}_{2,LS}\hat{\mu}_{1,LS} + 2\hat{\mu}_{1,LS}^3}{\left(\hat{\mu}_{2,LS} - \hat{\mu}_{1,LS}^2\right)^{3/2}} \right\}(x).$$

This estimate is linearly biased. According to Theorem 2, the linearly unbiased skewness function estimate is obtained by introducing an additional bias correction,

$$\begin{aligned}\hat{s}^*(x) = \hat{s}(x) + \frac{(3/2)s_x^2}{(\hat{\mu}_2 - \hat{\mu}_1^2)^{7/2}} &\Big[\left(\hat{\mu}_2\hat{\mu}_3 + 4\hat{\mu}_1^2\hat{\mu}_3 - 5\hat{\mu}_2^2\hat{\mu}_1\right)\hat{\delta}_1^2\Big] \\ &+ \left(\hat{\mu}_2^2 + 4\hat{\mu}_2\hat{\mu}_1^2 - 5\hat{\mu}_1\hat{\mu}_3\right)\hat{\delta}_1\hat{\delta}_2 - \frac{1}{4}\left(3\hat{\mu}_1\hat{\mu}_2 + 2\hat{\mu}_1^3 - 5\hat{\mu}_3\right)\hat{\delta}_2^2 \\ &+ 2\Big[\left(\hat{\mu}_1\hat{\mu}_2 - \hat{\mu}_1^3\right)\hat{\delta}_1\hat{\delta}_3 - \left(\hat{\mu}_2 - \hat{\mu}_1^2\right)\hat{\delta}_2\hat{\delta}_3\Big],\end{aligned}$$

where the argument x has been omitted in terms on the r.h.s., and all $\hat{\mu}_\ell = \hat{\mu}_{\ell,LS}(x)$. Analogously one can construct linearly unbiased estimators for kurtosis functions, etc.

Of some interest for applications is Example 5, the estimation of the conditional correlation function (7). The plug-in estimate is

$$\hat{\rho}_{12}(x) = \frac{\hat{v}_{12}(x)}{\{\hat{v}_{11}(x)\hat{v}_{22}(x)\}^{1/2}} = \frac{\hat{\mu}_{11} - \hat{\mu}_{10}\hat{\mu}_{01}}{[\{\hat{\mu}_{20} - \hat{\mu}_{10}^2\}\{\hat{\mu}_{02} - \hat{\mu}_{01}^2\}]^{1/2}},$$

where $\hat{\mu}_{lm} = \hat{\mu}_{lm,LS}(x)$. This estimate is linearly biased. The construction of the bias correction requires estimation of the mixed partial derivatives $\frac{\partial G}{\partial x_l \partial x_m}$, $1 \le l, m \le 5$ of $G(x_1, x_2, x_3, x_4, x_5) = (x_1 - x_2x_3)/\{(x_4 - x_1^2)(x_5 -$

$x_3^2)\}^{1/2}$. These derivatives are easiest calculated by using a package that includes symbolic calculus. According to Theorem 2, the linearly unbiased estimator is found to be, setting $A = \{\hat{\mu}_{(20),LS}(x) - \hat{\mu}^2_{(10),LS}(x)\}^{-1/2}$, $B = \{\hat{\mu}_{(02),LS}(x) - \hat{\mu}^2_{(01),LS}(x)\}^{-1/2}$ and omitting arguments x on the r.h.s.,

$$\begin{aligned}
\hat{\rho}^*_{12}(x) &= \hat{\rho}_{12}(x) \\
&+ \frac{1}{2}s_x^2[2A^3B\hat{\mu}_{(10)}\hat{\delta}_{(11)}\hat{\delta}_{(10)} + 2AB^3\hat{\mu}_{(01)}\hat{\delta}_{(11)}\hat{\delta}_{(01)} - A^3B\hat{\delta}_{(11)}\hat{\delta}_{(20)} \\
&- A^3B\hat{\delta}_{(11)}\hat{\delta}_{(02)} + A^5B\{-3\hat{\mu}_{(10)}\hat{\mu}_{(01)}\hat{\mu}_{(20)} + \hat{\mu}_{(11)}(2\hat{\mu}^2_{(10)} + \hat{\mu}_{(20)})\}\hat{\delta}^2_{(10)} \\
&+ 2A^3B^3(\hat{\mu}_{(11)}\hat{\mu}_{(10)}\hat{\mu}_{(01)} - \hat{\mu}_{(02)}\hat{\mu}_{(20)})\hat{\delta}_{(10)}\hat{\delta}_{(01)} \\
&+ A^5B\{-3\hat{\mu}_{(11)}\hat{\mu}_{(10)} + \hat{\mu}_{(01)}(2\hat{\mu}^2_{(10)} + \hat{\mu}_{(20)})\}\hat{\delta}_{(10)})\hat{\delta}_{(20)} \\
&+ A^3B^3(-\hat{\mu}_{(11)}\hat{\mu}_{(10)} + \hat{\mu}_{(01)}\hat{\mu}_{(20)})\hat{\delta}_{(10)}\hat{\delta}_{(02)} \\
&+ A^5B\{-3\hat{\mu}_{(10)}\hat{\mu}_{(01)}\hat{\mu}_{(02)} + \hat{\mu}_{(11)}(2\hat{\mu}^2_{(01)} + \hat{\mu}_{(02)})\}\hat{\delta}^2_{(01)} \\
&+ A^3B^3(-\hat{\mu}_{(11)}\hat{\mu}_{(01)} + \hat{\mu}_{(10)}\hat{\mu}_{(02)})\hat{\delta}_{(01)}\hat{\delta}_{(20)} \\
&+ A^5B\{-3\hat{\mu}_{(11)}\hat{\mu}_{(01)} + \hat{\mu}_{(10)}(2\hat{\mu}^2_{(01)} + \hat{\mu}_{(02)})\}\hat{\delta}_{(01)}\hat{\delta}_{(02)} \\
&+ \frac{3}{4}A^5B(\hat{\mu}_{(11)} - \hat{\mu}_{(10)}\hat{\mu}_{(01)})\hat{\delta}^2_{(20)} + \frac{1}{2}A^3B^3(\hat{\mu}_{(11)} - \hat{\mu}_{(10)}\hat{\mu}_{(01)})\hat{\delta}_{(20)}\hat{\delta}_{(02)} \\
&+ \frac{3}{4}AB^5(\hat{\mu}_{(11)} - \hat{\mu}_{(10)}\hat{\mu}_{(01)})\hat{\delta}^2_{(02)}].
\end{aligned}$$

5. A simulation example

The finite sample behavior of estimators $\hat{v}$ (24), $\hat{v}_{MSE}$ (26) and $\hat{v}^*$ (27) was investigated in a small scale simulation study. Since according to (28), these estimators behave identically with respect to asymptotic variance, and as the bias behavior is the focus of this paper, only the bias was investigated. This was done graphically by averaging estimated variance functions for the various methods over 200 Monte Carlo runs.

According to (29) and (30), asymptotic biases for $\hat{v}, \hat{v}_{MSE}$ and $\hat{v}^*$ are determined by the behavior of $v^{(2)} + 2\mu_1^{(1)^2}$, $v^{(2)} + 2v^{(1)}f^{(1)}/f$, respectively. The following example was used to investigate the influence of these terms and Monte Carlo runs were made assuming that $n = 50$, 250, 1250 observations were available.

Pseudo-random pairs (X_i, Y_i), $i = 1, ..., n$ were generated according to

$$X_i = \sqrt{0.25}Z_i, \quad Y_i = (X_i + 2)^2 + Z_i\sigma_i,$$

with

$$\sigma_i = 0 \text{ for } X_i \leq -0.5, \text{ and } \sigma_i = (X_i + 0.5)^{1/2} \text{ for } X_i > -0.5,$$

where Z_i are independent standard normal pseudo random numbers. The relevant functions are seen to be

$$\mu_1(x) = (x+2)^2, \quad \mu_1^{(1)}(x) = 2(x+2),$$
$$v(x) = (x+0.5), \quad v^{(1)} f^{(1)}/f = -8x, \, x \geq -0.5.$$

Several bandwidths and also other variance functions were chosen, and the results were qualitatively the same for a broad range of cases. We report the results for the bandwidth $b = 0.4$ and sample size $n = 250$.

A typical data sample is shown in Figure 1, while Figure 2 displays the average curve estimates from 200 Monte Carlo runs for the three estimators considered along with the target variance function which is linear in this example.

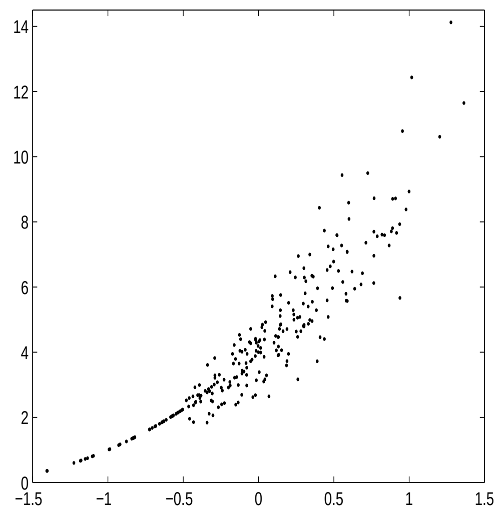

Fig. 1. Sample points (X_i, Y_i) for $n = 250$.

As expected, the asymptotic bias term $2\mu_1^{(1)^2}$ is so large that as a consequence the estimators $\hat{v}$ are unacceptable. The differences between $\hat{v}_{MSE}$ and $\hat{v}^*$ are more subtle. It is obvious that $\hat{v}_{MSE}$ exhibits an upward bias for small predictor levels below 0, and a downward bias for larger predictor levels above 0. This behavior is expected from the asymptotic bias expression (30). The linearly unbiased estimator $\hat{v}^*$ has a very small, more or less constant bias. It emerges as the clearly preferred estimator, as predicted by theory.

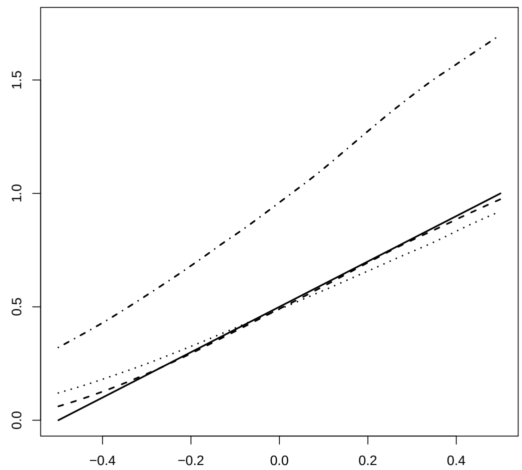

Fig. 2. Variance function estimators averaged over 200 Monte Carlo runs for $n = 250$ data pairs: Target variance function (solid), linearly biased estimators $\hat{v}$ (24) (dash-dotted), $\hat{v}_{MSE}$ (26) (dotted) and linearly unbiased estimator $\hat{v}^*$ (27) (dashed).

6. Technical details and proofs

6.1. *Assumptions and auxiliary results*

We first list the necessary regularity conditions. For a given x in the domain of the predictor variable X, let $N(x)$ be a neighborhood of x. We denote convergence in distribution by $\xrightarrow{\mathcal{L}}$ and convergence in probability by $\xrightarrow{p}$, as $n \to \infty$.

(C1) The joint distribution $F(\cdot, \cdot)$ of (X, Y) has a density $f(u, v)$, which is continuous on $N(x) \times \Re^p$.

(C2) $f(u, v)$ is twice continuously differentiable in the first argument on $N(x) \times \Re^p$.

(C3) The marginal density f_X of X is twice continuously differentiable on $N(x)$ and satisfies $f_X(x) > 0$.

(C4) The function $g(x) = G\{\mu_{\alpha_1}(x), ..., \mu_{\alpha_k}(x)\}$ is twice continuously differentiable on $N(x)$; this implies corresponding differentiability conditions for G and moment functions $\mu_{\alpha_m}(x)$.

(C5) The bandwidth sequence satisfies

$$b \to 0, \quad nb \to \infty \quad \text{and} \quad nb^5 \to d^2 \quad \text{as} \quad n \to \infty \quad \text{for} \quad \text{a} \quad d \geq 0.$$

(C6) The kernel function K satisfies

$$K \geq 0, \quad \int K = 1, \quad \int Ku = 0, \quad \int Ku^2 < \infty, \quad \int K^2 < \infty.$$

Consider now weighted averages

$$\Psi_{\lambda n} = (nb)^{-1} \sum_{i=1}^{n} \psi_\lambda (X_i, Y_i) K \left(b^{-1} (x - X_i)\right), \quad \lambda = 1, ..., m,$$

where the real valued functions ψ_λ satisfy:

(C7) All ψ_λ are bounded and continuous on $\{x\} \times \Re^p$.

(C8) The second derivatives with respect to the first argument exist for all ψ_λ and are continuous on $\{x\} \times \Re^p$.

Define

$$\zeta_\lambda(x) = \int \psi_\lambda(x, v) f(x, v)\, dv \tag{31}$$

$$\sigma_{\kappa\lambda}(x) = \int \psi_\kappa(x, v)\psi_\lambda(x, v) f(x, v) dv - \int \psi_\kappa(x, v) f(x, v) dv \int \psi_\lambda(x, v) f(x, v)\, dv, \tag{32}$$

and let $H : \Re^q \to \Re$ be a function with continuous second derivatives and $\zeta = (\zeta_1, ..., \zeta_q)$. The following auxiliary result is a direct consequence of Theorem 4.1 in Bhattacharya and Müller (1993); the proof is omitted.

Lemma 1: *Under (C1)-(C8),*

$$(nb)^{1/2} \left[H\left(\Psi_{1n}, ..., \Psi_{qn}\right) - H\left(\zeta_1, ..., \zeta_q\right)\right] \xrightarrow{\mathcal{L}} \mathcal{N}\left(B, V\right),$$

where

$$B = c_B \sum_{\lambda=1}^{q} \frac{dH}{dx_\lambda} |_\zeta \frac{d^2}{d^2 x} \zeta_\lambda(x), \; V = c_V \sum_{\kappa,\lambda=1}^{q} \frac{dH}{dx_\kappa} |_\zeta \frac{dH}{dx_\lambda} |_\zeta \sigma_{\kappa\lambda}(x).$$

Note for the following that in the case where

$$H(x_1, ..., x_q) = H(zx_1, ..., zx_q)$$

for all $z \neq 0$, a simple chain rule argument shows that $\sigma_{\kappa\lambda}$ in (33) can be replaced by

$$\tilde{\sigma}_{\kappa\lambda} = \int \psi_\kappa(x,v)\psi_\lambda(x,v)f(x,v)dv.$$

6.2. *Proof of Theorem 1*

For $\hat{\delta}$ as defined in (20), it follows from Corollary 4.3 in Bhattacharya and Müller (1993) that

$$(nb)^{1/2}\left\{\hat{\delta}_{e_1}(x) - \mu_{e_1}^{(1)}(x)\right\} \xrightarrow{p} 0,$$

where $e_1 = (1, 0, ..., 0)$. This generalizes easily to

$$(nb)^{1/2}\left\{\hat{\delta}_\alpha(x) - \mu_\alpha^{(1)}(x)\right\} \xrightarrow{p} 0 \tag{33}$$

for any multi-index α. Define the function

$$H_1(x_1, x_2, x_3) = \frac{x_1 - x_2\mu_\alpha^{(1)}(x)}{x_3}.$$

Choosing $x_q = \Psi_{qn}$, $q = 1, 2, 3$, with $\psi_1(u,v) = v^\alpha$, $\psi_2(u,v) = u - x$, $\psi_3(u,v) = 1$, we find that (33), Slutsky's Theorem and Lemma 1 imply that

$$(nb)^{1/2}\left\{\hat{\mu}_\alpha(x) - \mu_\alpha(x)\right\} \xrightarrow{\mathcal{L}} \mathcal{N}(B_\alpha, V_\alpha), \tag{34}$$

with $B_\alpha = c_B\,\mu_\alpha^{(2)}(x)$, $V_\alpha = c_V\{\mu_{2\alpha}(x) - \mu_\alpha^2(x)\}/f_X(x)$.

Furthermore, for any given constants $c_1, c_2 \neq 0$, and any multi-indices α_1, α_2 with $|\alpha_1| > 0, |\alpha_2| > 0$, generalizing (34), one obtains

$$\begin{aligned}(nb)^{1/2}\{c_1(\hat{\mu}_{\alpha_1}(x) - \mu_{\alpha_1}(x)) + c_2(\hat{\mu}_{\alpha_2}(x) - \mu_{\alpha_2}(x))\}\\ \to \mathcal{N}(B_{\alpha_1,\alpha_2}, V_{\alpha_1,\alpha_2})\end{aligned} \tag{35}$$

where

$$\begin{aligned}B_{\alpha_1,\alpha_2} &= c_B\{c_1\mu_{\alpha_1}^{(2)}(x) + c_2\mu_{\alpha_2}^{(2)}(x)\},\\ V_{\alpha_1,\alpha_2} &= c_V\{c_1^2\mu_{2\alpha_1}(x) + c_2^2\mu_{2\alpha_2}(x) + 2c_1c_2\mu_{\alpha_1+\alpha_2}(x)\\ &\quad - (c_1\mu_{\alpha_1}(x) + c_2\mu_{2\alpha_2}(x))^2\}/f_X(x).\end{aligned}$$

This follows from Lemma 1, choosing

$$H_2(x_1, x_2, x_3, x_4) = c_1\frac{x_1 - x_2\mu_{\alpha_1}^{(1)}(x)}{x_3} + c_2\frac{x_4 - x_2\mu_{\alpha_2}^{(1)}(x)}{x_3},$$

with $x_q = \Psi_{qn}, \quad q = 1, ..., 4, \quad \psi_1(u,v) = v^{\alpha_1}, \psi_2(u,v) = u - x, \quad \psi_3(u,v) = 1$, and $\psi_4(u,v) = v^{\alpha_2}$. Extending (35) to a linear combination of more than two estimators and a Taylor expansion

$$\hat{g}(x) - g(x) = \sum_{m=1}^{k} \frac{dG}{dx_m}|_\mu \{\hat{\mu}_{\alpha_m}(x) - \mu_{\alpha_m}(x)\} + o\left[\sum_{m=1}^{k} \{\hat{\mu}_{\alpha_m}(x) - \mu_{\alpha_m}(x)\}\right]$$

conclude the proof.

6.3. *Proof of (28) and (30) for estimators $\hat{v}_{MSE}$*

Let $p = 1$ and set

$$H_3(x_1, x_2, x_3) = (x_1/x_2) - (x_3^2/x_2^2), \; x_q = \Psi_{qn}, \; q = 1, 2, 3$$

with

$$\psi_1(u,v) = v^2, \; \psi_2(u,v) \equiv 1, \; \psi_3(u,v) = v.$$

Applying Lemma 1, we obtain

$$(nb)^{1/2} \left\{ \frac{\sum_{i=1}^n w_i(x) Y_i^2}{\sum_{i=1}^n w_i(x)} - \left(\frac{\sum_{i=1}^n w_i(x) Y_i}{\sum_{i=1}^n w_i(x)} \right)^2 \right\} \xrightarrow{\mathcal{L}} \mathcal{N}(\bar{B}, \bar{V}), \qquad (36)$$

with

$$\bar{B} = c_B(v_{(2)} + 2\mu_1^{(1)^2} + 2v^{(1)} \frac{f_X^{(1)}}{f_X}(x), \quad \bar{V} = V_V$$

as in (34).

We note in passing that (36) provides a result for plug-in variance function estimation based on the Nadaraya-Watson quotient type kernel estimator. Furthermore, by (33),

$$\hat{\delta}_1(x) \xrightarrow{p} \mu_1^{(1)}(x). \qquad (37)$$

Note that s_x^2 (21) can be equivalently written as

$$s_x^2 = \frac{\sum_{i=1}^n w_i(x) X_i^2}{\sum_{i=1}^n w_i(x)} - \left(\frac{\sum_{i=1}^n w_i(x) X_i}{\sum_{i=1}^n w_i(x)} \right)^2.$$

Therefore, (36) applies, with Y_i replaced by X_i, and analogous changes in the definition of $v(\cdot)$ and $\mu_\ell(\cdot)$. This leads to $\bar{B} = 2c_B$ and $\bar{V} = 0$, so that

$$(nb)^{1/2} s_x^2 \xrightarrow{p} 2c_B. \qquad (38)$$

The result now follows by combining (37)-(38) and applying Slutsky's theorem.

6.4. *Proof of Theorem 2*

Using analogous arguments as in (37) and (38), we find

$$(nb)^{1/2}\hat{\Delta}(x) \xrightarrow{p} c_B \left(\sum_{l,m=1}^{k} \frac{d^2 G}{dx_l dx_m} |_\mu \right) \delta_{\alpha_l}(x)\delta_{\alpha_m}(x).$$

The result follows from Theorem 1 and Slutsky's theorem.

Acknowledgments

Part of this research was carried out while I was visiting the Statistical Institute of Academia Sinica at Taipei, and I wish to express my sincere gratitude for the hospitality received there. Special thanks are due to Philip Cheng and Zhen Zhang for helpful feedback on a previous version and to two anonymous referees for their comments and suggestions. This research was supported in part by NSF grants DMS03-54448 and DMS05-05537.

References

1. Bhattacharya, P.K. and Müller, H.G. (1993). Asymptotics for nonparametric regression. *Sankhyā* A 55, 420-441.
2. Bjerve, O. and Doksum, K. (1993). Correlation curves as a measure of dependence conditional on a covariate. *Ann. Statist.* **21**, 890-902.
3. Chu, C.K. and Marron, J.S. (1991). Choosing a kernel regression estimator (with discussion). *Statist. Science* **6**, 404-436.
4. Claeskens, G. and van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *Ann. Statist.* **31**, 1852-1884.
5. Dette, H. and Munk, A. (1998). Testing heteroscedasticity in nonparametric regression. *J. R. Statist. Soc. B* **60**, 693-708.
6. Doksum, K., Blyth, S., Bradlow, E., Meng, X.L. and Zhao, H.Y. (1994). Correlation curves as local measures of variance explained by regression. *J. Am. Statist. Assoc.* **89**, 571-582.
7. Doksum, K. and Samarov, A. (1995). Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *Ann. Statist.* **23**, 1443-1473.
8. Eubank, R.L. and Speckman, P.L. (1993). Confidence bands in nonparametric regression. *J. Am. Statist. Assoc.* **88**, 1287-1301.
9. Eubank, R. and Thomas, W. (1993). Detecting heteroscedasticity in nonparametric regression. *J. R. Statist. Soc. B* **55**, 145-155.
10. Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling.* Chapman and Hall, London.
11. Fan, J. and Yao, Q.W. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85**, 645-660.

12. Gasser, T., Müller, H.G. and Mammitzsch, V. (1985). Kernels for nonparametric curve estimation. *J. R. Statist. Soc. B* **47**, 238-252.
13. Hall, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *Ann. Statist.* **20**, 675-694.
14. Jennen-Steinmetz, C. and Gasser, T. (1987). A unifying approach to nonparametric regression estimation. *J. Am. Statist. Assoc.* **83**, 1084-1089.
15. Jones, M.C. and Park, B.U. (1994). Versions of kernel-type regression estimators. *J. Am. Statist. Assoc.* **89**, 825-832.
16. Mammen, E. and Marron, J.S. (1997). Mass recentred kernel smoothers. *Biometrika* **84**, 765-777.
17. Müller, H.G. (1997). Density adjusted kernel estimators for random design nonparametric regression. *Statist. Prob. Letters* **36**, 161-172.
18. Müller, H.G. and Song, K.S. (1993). Identity reproducing multivariate nonparametric regression. *J. Multiv. Anal.* **46**, 237-253.
19. Müller, H.G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* **15**, 610-625.
20. Müller, H.G. and Stadtmüller, U. (1993). On variance function estimation with quadratic forms. *J. Statist. Plann. Inf.* **35**, 213-231.
21. Park, B.U., Kim, W.C. and Jones, M.C. (1997). On identity reproducing nonparametric regression estimators. *Statist. Prob. Letters* **32**, 279-290.
22. Picard, D. and Tribouley, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.* **28**, 298-335.
23. Ruppert, D., Wand, M.P., Holst, U. and Hössjer, O. (1997). Local polynomial variance function estimation. *Technometrics* **39**, 262-273.
24. Seifert, B. and Gasser, T. (1996). Finite-sample variance of local polynomials – analysis and solutions. *J. Am. Statist. Assoc.* **91**, 267-275.
25. Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing*. Chapman and Hall, London.