Statistica Sinica **10**(2000), 789-817

LOCAL POLYNOMIAL REGRESSION WITH SELECTION BIASED DATA

Colin O. Wu

The Johns Hapkins University

Abstract: Let Y and $\mathbf{X}$ be real- and R^d-valued random variables. We consider the estimation of the nonparametric regression function $m(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ when $s \geq 1$ independent selection-biased samples of $(Y, \mathbf{X})$ are observed. This sampling scheme, which arises naturally in biological and epidemiological studies and many other fields, includes stratified samples, length-biased samples and other weighted distributions. A class of local polynomial estimators of $m(\mathbf{x})$ is derived by smoothing Vardi's nonparametric maximum likelihood estimator of the underlying distribution function. Large sample properties, such as asymptotic distributions and asymptotic mean squared risks, are derived explicitly. Unlike local polynomial regression with i.i.d. direct samples, we show here that kernel choices are important and optimal kernel functions may be asymmetric and discontinuous when the weight functions of the biased samples have jumps. A cross-validation criterion is proposed for the selection of data-driven bandwidths. Through a simple comparison, we show that our estimators are superior to other intuitive estimators of $m(\mathbf{x})$.

Key words and phrases: Cross-validation, local polynomials, nonparametric maximum likelihood estimator, optimal kernel and bandwidths, selection-biased sample.

1. Introduction

Let $(Y, \mathbf{X})$, $Y \in R$, $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})^T \in R^d$, $d \geq 1$, be a pair of random variables such that

$$Y = m(\mathbf{X}) + \epsilon, \tag{1.1}$$

where $m(\mathbf{x}) = E[Y|\mathbf{X} = \mathbf{x}]$ and ϵ satisfies $E(\epsilon) = 0$ and $\mathrm{Var}(\epsilon) = \sigma^2(\mathbf{X})$. Let F be the underlying joint distribution function of $(Y, \mathbf{X})$, and f be the corresponding joint density with respect to the Lebesgue measure. A selection-biased sample of $(Y, \mathbf{X})$ consists of $s \geq 1$ independent random samples so that the observations of the ith sample $\{(Y_{ij}, \mathbf{X}_{ij});\ j = 1, \ldots, n_i\}$ with $Y_{ij} \in R$ and $\mathbf{X}_{ij} = (X_{ij}^{(1)}, \ldots, X_{ij}^{(d)})^T$ have joint distribution G_i and density g_i. Here, g_i and G_i depend on the underlying distribution F through

$$g_i(y, \mathbf{x}) = \frac{w_i(y, \mathbf{x})}{W_i} f(y, \mathbf{x}), \tag{1.2}$$

where w_i is a known non-negative weight function and

$$W_i = \int\int w_i(y, \mathbf{x}) f(y, \mathbf{x}) dy d\mathbf{x}.$$

This sampling scheme arises frequently in biological and epidemiological studies, economics, survey sampling and many other fields. For example, it may be necessary in certain cases to oversample the regions where the dependent variable Y takes extreme values which could be missed by simple random samples. Stratified sampling is a common strategy to ensure observations in all regions of interest. Other useful special cases include the weighted distributions considered by Patil, Rao and Zelen (1988) and Patil and Taillie (1989), the length-biased samples considered by Vardi (1982), among others.

Theory and methods for parametric and nonparametric estimations based on $\{(Y_{ij}, \mathbf{X}_{ij}); i = 1, \ldots, s,\ j = 1, \ldots, n_i\}$ have been extensively studied in the literature. Under the framework of linear models with stratified dependent variables, Jewell (1985) and Jewell and Quesenberry (1986) considered estimation of the coefficient β in $m(\mathbf{x}) = \mathbf{x}^T\beta$, when W_i are known and $w_i(y, \mathbf{x}) = 1$ if y is from the ith stratum and 0 otherwise. Bickel and Ritov (1991) generalized these methods and studied the large sample properties of their generalized estimation procedures. Nonparametric maximum likelihood estimator of the underlying distribution function F was originally developed by Vardi (1982, 1985) and further systematically studied by Gill, Vardi and Wellner (1988). By smoothing the nonparametric maximum likelihood estimator, Jones (1991) proposed a kernel density estimator with length-biased data, and Ahmad (1995) investigated a Nadaraya-Watson type kernel regression estimator with one-sample selection-biased data. Density estimation with multi-sample biased data has been considered by Wu (1997a, b). Recently, kernel regression with size-biased data has been studied by Sköld (1999).

In this article, we propose a class of local polynomial estimators for estimating $m(\mathbf{x})$ nonparametrically with multi-sample biased data $\{(Y_{ij}, \mathbf{X}_{ij});\ i = 1, \ldots, s,\ j = 1, \ldots, n_i\}$ and investigate their statistical properties. Explicit expressions of the asymptotic distributions and the asymptotic mean squared risks, including mean squared errors and the mean integrated squared errors, are derived for general local polynomials with a single covariate ($d = 1$) and local linear estimators with multiple covariates ($d > 1$). Because local linear fittings and low-dimensional nonparametric regressions are most practical in real applications (cf. Ruppert and Wand (1994) and Fan, Gasser, Gijbels, Brockmann and Engel (1997)), the asymptotic results here provide useful insights for inferences and reliability of our estimators. Our theoretical developments can be extended, at least in principle, to higher order polynomials with multivariate covariate $\mathbf{X}$.

Such an extension would be at the expense of excessive tedious computations and complex notations, hence will not be addressed in this article.

Unlike local polynomial regressions with i.i.d. direct data, our asymptotic results show that the choices of kernels may significantly influence the statistical properties of the estimators when estimating $m(\mathbf{x})$ at a point, and the Epanechnikov kernel may not be universally optimal (Fan and Gijbels (1996) and Fan *et al.*, (1997)). In fact, the optimal kernels may be asymmetric and discontinuous when the weight functions of the biased samples have jumps. Through a simple comparison, we also show that our estimators are asymptotically superior to another class of intuitive estimators, namely linear combinations of local polynomials constructed from a separate sample.

We present our local polynomial estimators in Section 2, develop their asymptotic distributions and asymptotic risk representations in Section 3, and establish optimal bandwidth and kernel choices and a cross-validation procedure for selecting data-driven bandwidths in Section 4. Section 5 gives a simple comparison with linear combination type estimators. Proofs of the main technical results are deferred to the appendices. A complete account of the proofs and the results of a Monte Carlo simulation can be found in Wu (1999).

2. Estimation Methods

2.1. Preliminary and nonparametric maximum likelihood estimates

Integrating (1.2) with respect to $(y, \mathbf{x})$, G_i has the following expression:

$$G_i(y,\mathbf{x}) = \int_{-\infty}^{y} \left[\int_{-\infty}^{x^{(1)}} \cdots \int_{-\infty}^{x^{(d)}} \frac{w_i(z,\mathbf{t})}{W_i} f(z,\mathbf{t}) d\mathbf{t} \right] dz,$$

where $\mathbf{x} = (x^{(1)}, \ldots, x^{(d)})^T$. Averaging G_i through all s samples, $\overline{G}_N(y,\mathbf{x}) = \sum_{i=1}^{s} \lambda_{n_i} G_i(y,\mathbf{x})$ can be written as

$$\overline{G}_N(y,\mathbf{x}) = \int_{-\infty}^{y} \left[\int_{-\infty}^{x^{(1)}} \cdots \int_{-\infty}^{x^{(d)}} \left(\sum_{i=1}^{s} \frac{\lambda_{n_i} w_i(z,\mathbf{t})}{W_i} \right) f(z,\mathbf{t}) d\mathbf{t} \right] dz,$$

where $\lambda_{n_i} = n_i/N$ and $N = n_1 + \cdots + n_s$. Differentiating the right hand side of the above equation with respect to $(y, \mathbf{x})$, f and F are given by

$$f(y,\mathbf{x}) = \sum_{i=1}^{s} \left\{ \left[\sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(y,\mathbf{x})}{W_r} \right]^{-1} \lambda_{n_i} g_i(y,\mathbf{x}) \right\}$$

and

$$F(y,\mathbf{x}) = \int_{-\infty}^{y} \int_{-\infty}^{x^{(1)}} \cdots \int_{-\infty}^{x^{(d)}} \left[\sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(z,\mathbf{t})}{W_r} \right]^{-1} d\overline{G}_N(z,\mathbf{t}).$$

In general, if there are no restrictions on w_i and W_i, f may not be identifiable from $(g_1, \ldots, g_s)$ in the sense that there may not be a one-to-one correspondence between f and $(g_1, \ldots, g_s)$. For example, $f(y, \mathbf{x})$ cannot be estimated from the data unless $\sum_{i=1}^{s}[\lambda_{n_i} w_i(y, \mathbf{x}) \ / W_i] > 0$. Let $\mathcal{S}$ be the support of f on R^{d+1}, that is, $\mathcal{S}$ is the smallest closed set such that the integral of $f(y, \mathbf{x})$ is one. The following *support* and *graph connectedness* conditions are shown by Vardi (1985) to be necessary and sufficient for F to be identifiable nonparametrically.

A1: (*Support*) $\mathcal{S} \subset [(y, \mathbf{x}) : \ w_i(y, \mathbf{x}) > 0 \text{ for some } i = 1, \ldots, s]$.

A2: (*Graph Connectedness*) For any $1 \leq i \leq s$ and $1 \leq j \leq s$, there exist $i_1, \ldots, i_k$ having values between 1 and s for $1 \leq k \leq s-2$, such that

$$\int\int 1_{[w_i(y,\mathbf{x})>0]} 1_{[w_{i_1}(y,\mathbf{x})>0]} f(y, \mathbf{x}) dy d\mathbf{x} > 0,$$
$$\int\int 1_{[w_{i_k}(y,\mathbf{x})>0]} 1_{[w_j(y,\mathbf{x})>0]} f(y, \mathbf{x}) dy d\mathbf{x} > 0,$$

and

$$\int\int 1_{[w_{i_l}(y,\mathbf{x})>0]} 1_{[w_{i_{l+1}}(y,\mathbf{x})>0]} f(y, \mathbf{x}) dy d\mathbf{x} > 0, \qquad \text{for all } l = 1, \ldots, k-1.$$

Condition A2 is mainly used to guarantee that W_i can be estimated from the data. When the W_i are known, A1 alone ensures the identifiability of F, which can be naturally estimated by

$$F_N(y, \mathbf{x}) = D_N^{-1} N^{-1} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \left\{ \left[\sum_{r=1}^{s} \frac{\lambda_{n_r} w_r(Y_{ij}, \mathbf{X}_{ij})}{V_r} \right]^{-1} 1_{\left[Y_{ij} \leq y, X_{ij}^{(1)} \leq x^{(1)}, \ldots, X_{ij}^{(d)} \leq x^{(d)}\right]} \right\},$$

where $\mathbf{X}_{ij} = (X_{ij}^{(1)}, \ldots, X_{ij}^{(d)})^T$, $V_r = W_r / W_s$ and

$$D_N = N^{-1} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \left[\sum_{r=1}^{s} \frac{\lambda_{n_r} w_r \left(Y_{ij}, \mathbf{X}_{ij}\right)}{V_r} \right]^{-1}. \tag{2.1}$$

Note that D_N does not necessarily equal one. The jumps of F_N are given by

$$N^{-1} J_N \left(Y_{ij}, \mathbf{X}_{ij}\right) = \left[N D_N \sum_{r=1}^{s} \frac{\lambda_{n_r} w_r \left(Y_{ij}, \mathbf{X}_{ij}\right)}{V_r} \right]^{-1} \tag{2.2}$$

at each observation point $(Y_{ij}, \mathbf{X}_{ij})$, and zero elsewhere.

In practice, the W_i are generally unknown and have to be estimated from the data. Under conditions A1 and A2, it is shown in Vardi (1985) that, when

N is sufficiently large, the equations

$$\widehat{V}_{n_i}^{-1} N^{-1} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \left\{ w_i\left(Y_{ij}, \mathbf{X}_{ij}\right) \left[\sum_{r=1}^{s} \frac{\lambda_{n_r} w_r\left(Y_{ij}, \mathbf{X}_{ij}\right)}{\widehat{V}_{n_r}} \right]^{-1} \right\} = 1, \quad i = 1, \ldots, s-1, \tag{2.3}$$

almost surely have a unique solution $(\widehat{V}_{n_1}, \ldots, \widehat{V}_{n_{s-1}})$ which forms a natural estimator of $(V_1, \ldots, V_{s-1})$. A necessary and sufficient condition, also referred as the *strongly connected* condition, for the existence and uniqueness of $(\widehat{V}_{n_1}, \ldots, \widehat{V}_{n_{s-1}})$ is given in Theorem 1.1 of Gill, Vardi and Wellner (1988). Consequently, D_N and W_i can be estimated by

$$\widehat{D}_N = N^{-1} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \left[\sum_{r=1}^{s} \frac{\lambda_{n_r} w_r\left(Y_{ij}, \mathbf{X}_{ij}\right)}{\widehat{V}_{n_r}} \right]^{-1} \quad \text{and} \quad \widehat{W}_{n_i} = \frac{\widehat{V}_{n_i}}{\widehat{D}_N} \tag{2.4}$$

for $i = 1, \ldots, s-1$, and $\widehat{W}_{n_s} = \widehat{D}_N^{-1}$. The nonparametric maximum likelihood estimator proposed by Vardi (1982, 1985) is

$$\widehat{F}_N(y, \mathbf{x}) = \widehat{D}_N^{-1} N^{-1} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \left\{ \left[\sum_{r=1}^{s} \frac{\lambda_{n_r} w_r\left(Y_{ij}, \mathbf{X}_{ij}\right)}{\widehat{V}_{n_r}} \right]^{-1} 1_{\left[Y_{ij} \le y, X_{ij}^{(1)} \le x^{(1)}, \ldots, X_{ij}^{(d)} \le x^{(d)}\right]} \right\},$$

which has jumps

$$N^{-1} \widehat{J}_N\left(Y_{ij}, \mathbf{X}_{ij}\right) = \left[N \widehat{D}_N \sum_{r=1}^{s} \frac{\lambda_{n_r} w_r\left(Y_{ij}, \mathbf{X}_{ij}\right)}{\widehat{V}_{n_r}} \right]^{-1} \tag{2.5}$$

at each observation point $(Y_{ij}, \mathbf{X}_{ij})$, and zero elsewhere.

2.2. Local polynomials with univariate covariate

When the data are from i.i.d. direct samples, linear smoothing methods, such as kernel estimators, smoothing splines and local polynomials, have been the subject of intense investigation for many years. Theory and applications of kernel estimators and smoothing splines can be found, for example, in Stone (1982, 1984), Rice (1984), Eubank (1988), Härdle (1990) and Wahba (1990), among others. Compared with kernel estimators, local polynomial fittings have the advantages of being adaptive to both random and fixed designs, and can adjust boundary biases automatically; see, for example, Fan and Gijbels (1996), Ruppert and Wand (1994) and Cheng, Fan and Marron (1997).

When the W_i are unknown a class of local polynomial fittings, based on (2.5) and $\{(Y_{ij}, X_{ij})\}$ with $X_{ij} \in R$ and order $p \ge 1$, can be obtained by minimizing

$$\ell_N(x) = \frac{1}{N} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \left\{ \left[Y_{ij} - \sum_{k=0}^{p} \Big((X_{ij} - x)^k b_k(x) \Big) \right]^2 \widehat{J}_N(Y_{ij}, X_{ij}) K_h(X_{ij} - x) \right\}, \tag{2.6}$$

with respect to $(b_0(x), \ldots, b_p(x))$, where h is a positive bandwidth, $K_h(u) = h^{-1}K(h^{-1}u)$, and $K(\cdot)$ is a kernel function on R which satisfies $\int K(u)du = 1$. Let

$$\mathbf{Y} = (Y_{11}, \ldots, Y_{1n_1}, \ldots, Y_{s1}, \ldots, Y_{sn_s})^T,$$

$$\mathcal{Z}^T(\mathbf{x}) = \begin{pmatrix} 1 & \cdots & 1 & \cdots & 1 & \cdots & 1 \\ (X_{11}-x) & \cdots & (X_{1n_1}-x) & \cdots & (X_{s1}-x) & \cdots & (X_{sn_s}-x) \\ \vdots & & \vdots & & \vdots & & \vdots \\ (X_{11}-x)^p & \cdots & (X_{1n_1}-x)^p & \cdots & (X_{s1}-x)^p & \cdots & (X_{sn_s}-x)^p \end{pmatrix}$$

and

$$\mathcal{T}_h(x) = \operatorname{diag}\left(\frac{1}{N}\widehat{J}_N(Y_{11}, X_{11})K_h(X_{11}-x), \ldots, \frac{1}{N}\widehat{J}_N(Y_{sn_s}, X_{sn_s})K_h(X_{sn_s}-x)\right).$$

Setting the derivatives of (2.6) with respect to $b_0(x), \ldots, b_p(x)$ to 0, the maximizer $(\widehat{b}_0(x), \ldots, \widehat{b}_p(x))$ of (2.6) satisfies the normal equation

$$\left(\mathcal{Z}^T(x)\mathcal{T}_h(x)\mathcal{Z}(x)\right)\left(\widehat{b}_0(x), \ldots, \widehat{b}_p(x)\right)^T = \mathcal{Z}^T(x)\mathcal{T}_h(x)\mathbf{Y}. \tag{2.7}$$

Assume that $(\mathcal{Z}^T(x)\mathcal{T}_h(x)\mathcal{Z}(x))$ is invertible. The unique solution of (2.7) is given by

$$\left(\widehat{b}_0(x), \ldots, \widehat{b}_p(x)\right)^T = \left(\mathcal{Z}^T(x)\mathcal{T}_h(x)\mathcal{Z}(x)\right)^{-1}\left(\mathcal{Z}^T(x)\mathcal{T}_h(x)\mathbf{Y}\right). \tag{2.8}$$

We define $\widehat{b}_k(x)$ to be the nonparametric maximum likelihood (NPML) local polynomial estimator of $(k!)^{-1}m^{(k)}(x)$.

When the W_i are known, the corresponding NPML local polynomial estimators of $(k!)^{-1}m^{(k)}(x)$ can be constructed by substituting $\widehat{J}_N(Y_{ij}, X_{ij})$ of (2.6) with $J_N(Y_{ij}, X_{ij})$ defined in (2.2).

Remark 2.1. Through a local constant fitting with $p = 0$, nonparametric maximum likelihood kernel estimators of $m(x)$ and its derivatives can be easily constructed based on (2.2) or (2.5). For example, when the W_i are unknown,

$$\widehat{m}(x) = \frac{\sum_{i=1}^{s}\sum_{j=1}^{n_i}\left\{Y_{ij}\widehat{J}_N(Y_{ij}, X_{ij})K[(X_{ij}-x)/h]\right\}}{\sum_{i=1}^{s}\sum_{j=1}^{n_i}\left\{\widehat{J}_N(Y_{ij}, X_{ij})K[(X_{ij}-x)/h]\right\}} \tag{2.9}$$

is a Nadaraya-Watson type NPML kernel estimator of $m(x)$. Similarly, when the W_i are known, a NPML kernel estimator of $m(x)$ can be constructed by substituting $\widehat{J}_N(Y_{ij}, X_{ij})$ of (2.9) with $J_N(Y_{ij}, X_{ij})$. To save space, theoretical development is limited to the case of local polynomial fittings ($p \geq 1$).

2.3. Local linear fittings with multivariate covariates

The formulation of Section 2.2 can be extended to multivariate covariate $\mathbf{X}$ by adding multivariate Taylor expansion terms in (2.6). However, because of the sparsity of data in high dimensions, local linear fittings are the most practical approach for nonparametric estimation of $m(\mathbf{x})$. Here a class of multivariate local linear fittings can be obtained by minimizing

$$L_N(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{s}\sum_{j=1}^{n_i}\left\{\left[Y_{ij} - b_0(\mathbf{x}) - (\mathbf{X}_{ij}-\mathbf{x})^T\mathbf{b}_1(\mathbf{x})\right]^2 \widehat{J}_N(Y_{ij},\mathbf{X}_{ij})K_{\mathbf{H}}(\mathbf{X}_{ij}-\mathbf{x})\right\} \tag{2.10}$$

with respect to $b_0(\mathbf{x})$ and $\mathbf{b}_1(\mathbf{x}) = (b_{11}(\mathbf{x}),\ldots,b_{1d}(\mathbf{x}))^T$, where $\mathbf{H}$ is a $d\times d$ symmetric positive definite matrix, $K_{\mathbf{H}}(\mathbf{u}) = |\mathbf{H}|^{-1}K(\mathbf{H}^{-1}\mathbf{u})$, and $K(\mathbf{u})$ is a kernel function which maps R^d to R and satisfies $\int K(\mathbf{u})d\mathbf{u} = 1$. Let

$$\mathbf{Z}^T(\mathbf{x}) = \begin{pmatrix} 1 & \cdots & 1 & \cdots & 1 & \cdots & 1 \\ (\mathbf{X}_{11}-\mathbf{x}) & \cdots & (\mathbf{X}_{1n_1}-\mathbf{x}) & \cdots & (\mathbf{X}_{s1}-\mathbf{x}) & \cdots & (\mathbf{X}_{sn_s}-\mathbf{x}) \end{pmatrix}$$

and

$$\mathbf{T}_{\mathbf{H}}(\mathbf{x}) = \operatorname{diag}\left(\frac{1}{N}\widehat{J}_N(Y_{11},\mathbf{X}_{11})K_{\mathbf{H}}(\mathbf{X}_{11}-\mathbf{x}),\ldots,\frac{1}{N}\widehat{J}_N(Y_{sn_s},\mathbf{X}_{sn_s})K_{\mathbf{H}}(\mathbf{X}_{sn_s}-\mathbf{x})\right).$$

The same derivations as in Section 2.2 show that, if $(\mathbf{Z}^T(\mathbf{x})\mathbf{T}_{\mathbf{H}}(\mathbf{x})\mathbf{Z}(\mathbf{x}))$ is invertible, the unique minimizer of (2.10) is given by

$$\begin{pmatrix} \widehat{b}_0(\mathbf{x}) \\ \widehat{\mathbf{b}}_1(\mathbf{x}) \end{pmatrix} = \left(\mathbf{Z}^T(\mathbf{x})\mathbf{T}_{\mathbf{H}}(\mathbf{x})\mathbf{Z}(\mathbf{x})\right)^{-1}\left(\mathbf{Z}^T(\mathbf{x})\mathbf{T}_{\mathbf{H}}(\mathbf{x})\mathbf{Y}\right). \tag{2.11}$$

Then $\widehat{b}_0(\mathbf{x})$ is the NPML local linear estimator of $m(\mathbf{x})$, and the components of $\widehat{\mathbf{b}}_1(\mathbf{x})$ are estimators of the corresponding partial derivatives of $m(\mathbf{x})$.

2.4. An example of stratified sampling

The standard stratified sample considered by Jewell (1985) and Jewell and Quesenberry (1986) involves s independent samples and each has n_i, $i = 1,\ldots,s$, i.i.d. observations $(Y_{ij},\mathbf{X}_{ij})$, $j = 1,\ldots,n_i$, such that Y_{ij} are observed in the ith stratum $[a_{i-1},a_i)$, where $a_0,\ldots,a_s$ are constants such that $-\infty = a_0 \le a_1 \le \cdots \le a_{s-1} \le a_s = +\infty$. In order to make $m(\mathbf{x})$ identifiable, Jewell (1985) and Jewell and Quesenberry (1986) assumed that $P_i = P(a_{i-1} \le Y < a_i)$ is known, $i = 1,\ldots,s$. Thus, condition A1 is satisfied and $w_i(y,\mathbf{x}) = 1_{[a_{i-1}\le y<a_i]}$, $W_i = P_i$, $V_r = P_r/P_s$ and $D_N = P_s^{-1}$. The jumps of F_N are given by $N^{-1}J_N(Y_{ij},\mathbf{X}_{ij}) = N^{-1}\lambda_{n_i}P_i^{-1}$ at each $(Y_{ij},\mathbf{X}_{ij})$ and zero elsewhere. Substituting these jumps into

(2.10), the NPML local linear estimators $\widehat{b}_0(\mathbf{x})$ and $\widehat{\mathbf{b}}_1(\mathbf{x})$) can be computed by (2.11) with

$$\mathbf{T}_{\mathbf{H}}(\mathbf{x}) = \operatorname{diag}\left(\frac{\lambda_{n_1}}{NP_1}K_{\mathbf{H}}(\mathbf{X}_{11}-\mathbf{x}),\ldots,\frac{\lambda_{n_s}}{NP_s}K_{\mathbf{H}}(\mathbf{X}_{sn_s}-\mathbf{x})\right).$$

When the P_j are unknown, it is easy to verify that condition A2 fails for this sampling scheme, so that $m(\mathbf{x})$ cannot be estimated nonparametrically. However, if there is also an extra i.i.d. sample $\{(Y_{s+1,j},\mathbf{X}_{s+1,j}); j=1,\ldots,n_{s+1}\}$, such that $Y_{s+1,j}$ can be observed in the whole real line R, then both A1 and A2 are satisfied and P_i can be estimated by the empirical distribution of the $(s+1)$th sample:

$$\widehat{P}_i = \frac{1}{n_{s+1}}\sum_{j=1}^{n_{s+1}} 1_{[a_{i-1}\le Y_{s+1,j}<a_i]}.$$

The unique solution of (2.3) is then given by $\widehat{V}_{n_i}=\widehat{P}_i$ for $i=1,\ldots,s$, and $\widehat{V}_{n_{s+1}}=1$. Substituting $(\widehat{V}_{n_1},\ldots,\widehat{V}_{n_{s+1}})$ into (2.5), the NPML local linear estimators $\widehat{b}_0(\mathbf{x})$ and $\widehat{\mathbf{b}}_1(\mathbf{x})$ are given by (2.11) with s replaced by $s+1$ and

$$\mathbf{T}_{\mathbf{H}}(\mathbf{x}) = \operatorname{diag}\left(\frac{\widehat{P}_1}{\lambda_{n_1}+\lambda_{n_{s+1}}\widehat{P}_1}K_{\mathbf{H}}(\mathbf{X}_{11}-\mathbf{x}),\ldots,\frac{\widehat{P}_{s+1}}{\lambda_{n_{s+1}}+\lambda_{n_{s+1}}\widehat{P}_{s+1}} \times K_{\mathbf{H}}(\mathbf{X}_{s+1,n_{s+1}}-\mathbf{x})\right).$$

When the sample size is small, $\widehat{P}_i$ may not be available because there may not be enough $Y_{s+1,j}$ falling within the interval $[a_{i-1},a_i)$. When the sample size is sufficiently large, $\widehat{b}_0(\mathbf{x})$ and $\widehat{\mathbf{b}}_1(\mathbf{x})$ exist and are unique almost surely.

3. Asymptotic Properties

3.1. Asymptotic properties for univariate local polynomials

The asymptotic behavior of the $\widehat{b}_k(x)$ depends on whether x is an interior point of $\operatorname{supp}(f_X)$, the support of $f_X(\cdot)$, or x is near the boundary. We define x to be an interior point if $x\in\operatorname{supp}(f_X)$ and $|x-x_b|>ch$ for some constant $c>0$ and any x_b on the boundary of $\operatorname{supp}(f_X)$. Let x_0 be an interior point. We consider the asymptotic mean squared risks and the asymptotic distributions of $\widehat{b}_k(x_0)$.

In addition to A1 and A2, we assume the following conditions throughout the section.

A3: (a) The underlying regression function $m(x)$ is at least $p+1$ times differentiable and its $(p+1)$th derivative is continuous and bounded in a neighborhood of x_0.

(b) The marginal density $f_X(x)$ is continuous in a neighborhood of x_0. There exists a constant $\varepsilon > 0$ such that the $[2(p+1)+\varepsilon]$th moment of X and the $(2+\varepsilon)$th moment of Y are finite.

(c) There are constants $0 < \lambda_i < 1$ such that $\sum_{i=1}^{s} \lambda_i = 1$ and $\lambda_{n_i} \to \lambda_i$ as $n \to \infty$ for all $i = 1, \ldots, s$.

(d) The kernel $K(u)$ is non-negative and compactly supported on the real line. It satisfies $\int K(u)du = 1$, $\int u^k K(u)du = \mu_k$ and $\int u^k K^2(u)du < \nu_k$ for some constants μ_k and ν_k with $k = 0, \ldots, 2p+2$ such that $\mu_k \equiv 0$ if k is odd.

(e) The bandwidth h satisfies $h \to 0$ and $Nh \to \infty$, as $N \to \infty$.

(f) For each $y \in R$, $w_i(y, x)$, $i = 1, \ldots, s$, are piecewise continuous functions of x with only countably many jumps.

Ideally, we would like to measure the adequacy of $\widehat{b}_k(x_0)$ by the second moment of $[\widehat{b}_k(x_0) - (k!)^{-1}m^{(k)}(x_0)]$. However, a minor technical inconvenience that arises from minimizing (2.6) is that the moments of $\widehat{b}_k(x_0)$ may not exist. The usual approach in the literature for measuring the large sample risks of local polynomial estimators with i.i.d. direct data is to first express their conditional mean squared errors in a closed form, then evaluate the asymptotic approximations of the conditional mean squared errors; see, for example, Ruppert and Wand (1994) and Fan and Gijbels (1996). Because the jumps defined in (2.5) are random, it is difficult to obtain the closed form of the conditional moments of $\widehat{b}_k(x_0)$. As a useful alternative, we consider a modified mean squared error of $\widehat{b}_k(x_0)$.

Let $\widehat{\mathbf{b}}(x_0) = (\widehat{b}_0(x_0), \ldots, \widehat{b}_k(x_0))^T$, $H = \operatorname{diag}(1, h, \ldots, h^p)$, $\mathbf{1}_{(p+1)\times(p+1)}$ be the $(p+1) \times (p+1)$ matrix whose every single entry is 1,

$$\mathbf{m}(x_0) = \left(m(x_0), m^{(1)}(x_0), \ldots, (p!)^{-1}m^{(p)}(x_0)\right)^T,$$

and

$$\mathbf{S}_0 = \begin{pmatrix} \mu_0 & \mu_1 & \cdots & \mu_p \\ \vdots & \vdots & \vdots & \vdots \\ \mu_p & \mu_{p+1} & \cdots & \mu_{2p} \end{pmatrix}.$$

It can be shown by straightforward calculation that (see Wu (1999) for further details)

$$\mathcal{Z}^T(x_0)\mathcal{T}_h(x_0)\mathcal{Z}(x_0) = f_X(x_0)H\mathbf{S}_0H + o_p\left(\mathbf{1}_{(p+1)\times(p+1)}\right).$$

If $\mathbf{S}$ is invertible,

$$\left[\mathbf{I} + o_p\left(\mathbf{1}_{(p+1)\times(p+1)}\right)\right]\left(\widehat{\mathbf{b}}(x_0) - \mathbf{m}(x_0)\right) = \mathcal{R}(x_0), \tag{3.1}$$

where $\mathbf{I}$ is the identity matrix and

$$\mathcal{R}(x_0) = (f_X(x_0))^{-1} H^{-1} \mathbf{S}_0^{-1} H^{-1} \left\{ \mathcal{Z}^T(x_0) \mathcal{T}_h(x_0) [\mathbf{Y} - \mathcal{Z}(x_0) \mathbf{m}(x_0)] \right\}. \quad (3.2)$$

Let e_{k+1} be the $(p+1) \times 1$ vector with 1 in its $(k+1)$th position, 0 elsewhere. Because $\widehat{b}_k(x_0) = e_{k+1}^T \widehat{\mathbf{b}}(x_0)$ and $(k!)^{-1} m^{(k)}(x_0) = e_{k+1}^T \mathbf{m}(x_0)$, (3.1) implies that it is appropriate to study the asymptotic properties of $e_{k+1}^T \mathcal{R}(x_0)$. Thus, we define the modified mean squared error of $\widehat{b}_k(x_0)$ to be

$$\mathrm{MSE}\left(\widehat{b}_k(x_0)\right) = E\left\{ \left[e_{k+1}^T \mathcal{R}(x_0) \right]^2 \right\}. \quad (3.3)$$

For the global risk of $\widehat{b}_k$ over an interior region of $\mathrm{supp}(f_X)$, we define the modified mean integrated squared error of $\widehat{b}_k$ to be

$$\mathrm{MISE}\left(\widehat{b}_k\right) = \int E\left\{ \left[e_{k+1}^T \mathcal{R}(x) \right]^2 \right\} \pi(x) dx, \quad (3.4)$$

where $\pi(x)$ is a known bounded non-negative weight function whose support is a subset of $\mathrm{supp}(f_X)$.

Let $J(y,t) = [\sum_{r=1}^s (\lambda_r w_r(y,t)/W_r)]^{-1}$, $\delta(x_0-) = \lim_{t \uparrow x_0} \delta(t)$, $\delta(x_0+) = \lim_{t \downarrow x_0} \delta(t)$,

$$\delta(t) = \int (y - m(t))^2 J(y,t) f(y|t) dy$$

and $\mathbf{V}(x_0)$ be the $(p+1) \times (p+1)$ matrix, such that, for $l, q = 0, \ldots, p$, the $(l+1, q+1)$th element of $\mathbf{V}(x_0)$ is

$$V_{l+1,q+1}(x_0) = \delta(x_0-) \int_{-\infty}^0 u^{l+q} K^2(u) du + \delta(x_0+) \int_0^\infty u^{l+q} K^2(u) du.$$

Note that $\delta(x_0-) = \delta(x_0+)$ if $J(y,t)$ is continuous at x_0. But, in general, $\delta(x_0-)$ may not equal $\delta(x_0+)$.

Theorem 3.1. *Suppose that $f_X(x_0) > 0$ and Assumptions* A1, A2 *and* A3 *are satisfied.*

(a) *If $h \geq cN^{(r-1)/(1+k_1+k_2)}$ for some $c > 0$, $0 < r < 1$ and any $k_1, k_2 = 0, \ldots, p$,*

$$\begin{aligned} &\mathrm{cov}\left(e_{k_1+1}^T \mathcal{R}(x_0), e_{k_2+1}^T \mathcal{R}(x_0) \right) \\ &= \left[\frac{1}{N h^{1+k_1+k_2} f_X(x_0)} \right] e_{k_1+1}^T \mathbf{S}_0^{-1} \mathbf{V}(x_0) \mathbf{S}_0^{-1} e_{k_2+1} \left(1 + o(1)\right). \end{aligned} \quad (3.5)$$

In particular, if $h \geq cN^{(1-r)/(1+2k)}$ for some $c > 0$, $0 < r < 1$ and any $k = 0, \ldots, p$,

$$\mathrm{var}\left(e_{k+1}^T \mathcal{R}(x_0) \right) = \left[\frac{1}{N h^{1+2k} f_X(x_0)} \right] e_{k+1}^T \mathbf{S}_0^{-1} \mathbf{V}(x_0) \mathbf{S}_0^{-1} e_{k+1} (1 + o(1)). \quad (3.6)$$

(b) *If $p-k$ is odd and $h \le cN^{-r/2(p+1-k)}$ for some $c>0$ and $0<r<1$,*

$$E\left[e_{k+1}^T \mathcal{R}(x_0)\right] = h^{p+1-k} e_{k+1}^T [(p+1)!]^{-1} m^{(p+1)}(x_0) \mathbf{S}_0^{-1} \mathbf{c}_p (1+o(1)), \quad (3.7)$$

where $\mathbf{c}_p = (\mu_{p+1}, \ldots, \mu_{2p+1})^T$.

(c) *If $p-k$ is even, $h \le cN^{-r/2(p+2-k)}$ for some $c>0$, $0<r<1$ and, in addition to* A3(a) *and A3(b), $m^{(p+2)}(x)$ exists and is continuous in a neighborhood of x_0 and $f_X(\cdot)$ is continuously differentiable in a neighborhood of x_0, then*

$$E\left[e_{k+1}^T \mathcal{R}(x_0)\right] = h^{p+2-k}\left[\frac{m^{(p+2)}(x_0)}{(p+2)!} + \frac{f_X'(x_0) m^{(p+1)}(x_0)}{f_X(x_0)(p+1)!}\right] \times e_{k+1}^T \mathbf{S}_0^{-1} \mathbf{c}_{p+1} (1+o(1)). \quad (3.8)$$

Proof. See Appendix A.

Remark 3.1. One sees that (3.7) and (3.8) describe the modified asymptotic bias and (3.6) describes the modified asymptotic variance of $\widehat{b}_k(x_0)$. Because (3.7) and (3.8) do not depend on the weight functions of (1.2), they are the same as the asymptotic conditional bias of local polynomial estimators with i.i.d. direct samples (Fan and Gijbels (1996, Chapter 3)). Because (3.6) depends on $\mathbf{V}(x_0)$, which is influenced by the specific choices of the weight functions of (1.2), the asymptotic variance of $\widehat{b}_k(x_0)$ is different from that with i.i.d. direct samples. The effects of the biased sampling schemes only appear in the constant terms of (3.6).

The next theorem presents some straightforward consequences of Theorem 3.1. The proof of the theorem, which essentially checks Lyapounov's conditions for $\widehat{b}_k(x_0)$, is given in Appendix A.3 of Wu (1999).

Theorem 3.2. *Suppose the assumptions of Theorem* 3.1 *are satisfied.*

(a) *If $p-k$ is odd, $h \ge cN^{(1-r)/(1+2k)}$ and $h \le c_* N^{-r_*/[2(p+2-k)]}$ for some positive constants c, c_*, r and r_* such that $r<1$ and $r_*<1$, then, when N is sufficiently large,*

$$\mathrm{MSE}\left(\widehat{b}_k(x_0)\right) = \left\{E\left[e_{k+1}^T \mathcal{R}(x_0)\right]\right\}^2 + \mathrm{var}\left[e_{k+1}^T \mathcal{R}(x_0)\right] \quad (3.9)$$

with $\mathrm{var}[e_{k+1}^T \mathcal{R}(x_0)]$ and $E[e_{k+1}^T \mathcal{R}(x_0)]$ given by (3.6) *and* (3.7)*, respectively. In addition, when $h = O(N^{-1/(2p+3)})$,*

$$\frac{\widehat{b}_k(x_0) - (k!)^{-1} m^{(k)}(x_0) - E[e_{k+1}^T \mathcal{R}(x_0)]}{\left\{\mathrm{var}[e_{k+1}^T \mathcal{R}(x_0)]\right\}^{1/2}} \to \mathcal{N}(0,1), \quad (3.10)$$

in distribution as $N \to \infty$.

(b) *If* $p-k$ *is even and* $m^{(p+2)}(x)$ *exists and is continuous in a neighborhood of* x_0*, then* (3.9) *holds with* $E[e_{k+1}^T\mathcal{R}(x_0)]$ *given by* (3.8)*. Furthermore, when* $h = O(N^{-1/(2p+5)})$*,* (3.10) *holds with* $E[e_{k+1}^T\mathcal{R}(x_0)]$ *given by* (3.8) *and* $\mathrm{var}[e_{k+1}^T\mathcal{R}(x_0)]$ *given by* (3.6)*.*

(c) *When* N *is sufficiently large and the support of* $\pi(\cdot)$ *is a proper subset of* $\mathrm{supp}(f_X)$ *the asymptotic representation of* (3.4) *is*

$$\mathrm{MISE}\left(\widehat{b}_k\right) = \int \left\{ \left[E\left(e_{k+1}^T\mathcal{R}(x)\right)\right]^2 + \mathrm{var}\left[e_{k+1}^T\mathcal{R}(x)\right] \right\} \pi(x)dx, \qquad (3.11)$$

where $\mathrm{var}[e_{k+1}^T\mathcal{R}(x)]$ *is given by* (3.6) *and* $E[e_{k+1}^T\mathcal{R}(x)]$ *is given by* (3.7) *or* (3.8) *if the conditions of Theorem* 3.1(b) *or Theorem* 3.1(c)*, respectively, are satisfied.*

3.2. Asymptotic properties for multivariate local linear estimators

Let $\mathcal{E}_{\mathbf{x}_0,\mathbf{H}} = \{\mathbf{t} : \mathbf{H}^{-1}(\mathbf{t}-\mathbf{x}_0) \in \mathrm{supp}(K)\}$ be the support of $K_{\mathbf{H}}(\cdot - \mathbf{x}_0)$. Extending the definition of Section 3.1, $\mathbf{x}_0$ is an interior point of $\mathrm{supp}(f_{\mathbf{X}})$ if $\mathcal{E}_{\mathbf{x}_0,\mathbf{H}}$ is a proper subset of $\mathrm{supp}(f_{\mathbf{X}})$. We establish the local and global asymptotic mean squared risks of the NPML local linear estimator $\widehat{b}_0(\mathbf{x}_0)$ of (2.11) at an interior point $\mathbf{x}_0$.

As a modification of A3, we assume the following conditions throughout the section.

A4: (a) The underlying regression function $m(\mathbf{x})$ is twice differentiable with respect to $\mathbf{x}$ and its second partial derivatives are continuous in a neighborhood of $\mathbf{x}_0$.

(b) The marginal density $f_{\mathbf{X}}(\mathbf{x})$ is continuous in a neighborhood of $\mathbf{x}_0$. There exists a constant $\epsilon > 0$ so that the $(4+\epsilon)$th moments of $X^{(l)}$, $l = 1, \ldots, d$, and the $(2+\epsilon)$th moment of Y are finite.

(c) There are constants $0 < \lambda_i < 1$ such that $\sum_{i=1}^s \lambda_i = 1$ and $\lambda_{n_i} \to \lambda_i$ as $n \to \infty$ for all $i = 1, \ldots, s$.

(d) The kernel $K(\mathbf{u})$ is non-negative, compactly supported on R^d, and satisfies $\int K(\mathbf{u})d\mathbf{u} = 1$. Moreover

$$\int u^{(l)}u^{(r)}K(\mathbf{u})d\mathbf{u} = \begin{cases} 0, & \text{if } l \neq r, \\ \mu_2(K), & \text{if } l = r, \end{cases}$$

for $l, r = 1, \ldots, d$ and some $\mu_2(K) > 0$, and $\int u^{(l_1)} \cdots u^{(l_d)}K(\mathbf{u})d\mathbf{u} = 0$ for all non-negative integers $l_1, \ldots, l_d$ such that $\sum_{j=1}^d l_j$ is odd.

(e) The bandwidth matrix $\mathbf{H}$ is symmetric and positive definite, such that $N^{-1}|\mathbf{H}|^{-1}$ and every entry of $\mathbf{H}$ tend to 0 as $N \to \infty$. In addition, there is a fixed constant L such that the condition number of $\mathbf{H}$ (the ratio of its largest to smallest eigenvalues) is at most L for all N.

(f) For each $y \in R$, $w_i(y, \mathbf{x})$, $i = 1, \ldots, s$, are piecewise continuous functions of $\mathbf{x}$ and have only countably many jumps in R^d.

Assumptions A4(d) and A4(e) are the same as conditions (A1) and (A3), respectively, of Ruppert and Wand (1994). Similar to the univariate case, the moments of $\widehat{b}_0(\mathbf{x}_0)$ may not exist in general and modifications of the mean squared errors and the mean integrated squared errors of $\widehat{b}_0(\mathbf{x}_0)$ have to be considered. Let

$$m'(\mathbf{x}_0) = \left(\frac{\partial m(\mathbf{x})}{\partial x^{(1)}}\Big|_{\mathbf{x}_0}, \ldots, \frac{\partial m(\mathbf{x})}{\partial x^{(d)}}\Big|_{\mathbf{x}_0} \right)^T.$$

Then, by (2.11), we have

$$\begin{aligned} &\left(\mathbf{Z}^T(\mathbf{x}_0)\mathbf{T}_{\mathbf{H}}(\mathbf{x}_0)\mathbf{Z}(\mathbf{x}_0)\right) \begin{pmatrix} \widehat{b}_0(\mathbf{x}_0) - m(\mathbf{x}_0) \\ \widehat{\mathbf{b}}_1(\mathbf{x}_0) - m'(\mathbf{x}_0) \end{pmatrix} \\ &= \mathbf{Z}^T(\mathbf{x}_0)\mathbf{T}_{\mathbf{H}}(\mathbf{x}_0) \left(\mathbf{Y} - \mathbf{Z}(\mathbf{x}_0) \begin{pmatrix} m(\mathbf{x}_0) \\ m'(\mathbf{x}_0) \end{pmatrix} \right). \end{aligned} \tag{3.12}$$

Furthermore, we can show by similar calculations as in the derivation of (3.1) that, when N is sufficiently large,

$$\mathbf{Z}^T(\mathbf{x}_0)\mathbf{T}_{\mathbf{H}}(\mathbf{x}_0)\mathbf{Z}(\mathbf{x}_0) = f(\mathbf{x}_0) \begin{pmatrix} 1 + o_p(1) & o_p\left(\mathbf{1}_{1\times d}\right) \\ o_p\left(\mathbf{1}_{d\times 1}\right) & \mu_2(K)\mathbf{H}^2 + O_p\left(\mathbf{H}^2\right) \end{pmatrix}. \tag{3.13}$$

Multiplying $[\mathbf{Z}^T(\mathbf{x}_0)\mathbf{T}_{\mathbf{H}}(\mathbf{x}_0)\mathbf{Z}(\mathbf{x}_0]^{-1}$ by the right side of (3.12), (3.13) implies that

$$(1 + o_p(1)) \left(\widehat{b}_0(\mathbf{x}_0) - m(\mathbf{x}_0)\right) = \mathcal{R}_0(\mathbf{x}_0), \tag{3.14}$$

where

$$\mathcal{R}_0(\mathbf{x}_0) = \frac{1}{f_{\mathbf{X}}(\mathbf{x}_0)N|\mathbf{H}|} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \left[\widehat{J}_N(Y_{ij}, \mathbf{X}_{ij}) \Delta_{ij}^*(\mathbf{x}_0) K\left(\mathbf{H}^{-1}(\mathbf{X}_{ij} - \mathbf{x}_0)\right) \right]$$

and

$$\Delta_{ij}^*(\mathbf{x}_0) = Y_{ij} - m(\mathbf{x}_0) - (\mathbf{X}_{ij} - \mathbf{x}_0)^T m'(\mathbf{x}_0).$$

Thus, we measure the local and global asymptotic risks of $\widehat{b}_0(\mathbf{x}_0)$ by the modified mean squared error

$$\text{MSE}\left(\widehat{b}_0(\mathbf{x}_0)\right) = E\left[\mathcal{R}_0^2(\mathbf{x}_0)\right] \tag{3.15}$$

and the modified mean integrated squared error

$$\text{MISE}\left(\widehat{b}_0\right) = \int E\left[\mathcal{R}_0^2(\mathbf{x})\right] \pi(\mathbf{x}) d\mathbf{x}, \tag{3.16}$$

respectively, where $\pi(\mathbf{x})$ is a known bounded non-negative weight function on R^d whose support is a subset of $\text{supp}(f_{\mathbf{X}})$.

Theorem 3.3. *Suppose* $f_{\mathbf{X}}(\mathbf{x}_0) > 0$ *and Assumptions* A1, A2 *and* A4 *are satisfied.*

(a) *When* N *is sufficiently large,*

$$E\left[\mathcal{R}_0(\mathbf{x}_0)\right] = \frac{1}{2}\mu_2(K)\text{tr}\left\{\mathcal{H}_m(\mathbf{x}_0)\mathbf{H}^2\right\}(1+o(1)) \tag{3.17}$$

and

$$\text{var}\left[\mathcal{R}_0(\mathbf{x}_0)\right] = (N|\mathbf{H}|f_{\mathbf{X}}(\mathbf{x}_0))^{-1}\left[\int \delta^*(\mathbf{x}_0+\mathbf{H}\mathbf{u})K^2(\mathbf{u})d\mathbf{u}\right](1+o(1)), \tag{3.18}$$

where $tr\{\mathbf{A}\}$ *is the trace of a symmetric matrix* $\mathbf{A}$, $\mathcal{H}_m(\mathbf{x}_0)$ *is the Hessian matrix of* $m(\mathbf{x})$ *at* $\mathbf{x}_0$, *and*

$$\delta^*(\mathbf{x}_0+\mathbf{H}\mathbf{u}) = \frac{1}{4}\int J(y,\mathbf{x}_0+\mathbf{H}\mathbf{u})\left[(\mathbf{H}\mathbf{u})^T\mathcal{H}_m(\mathbf{x}_0)(\mathbf{H}\mathbf{u})\right]^2 f(y|\mathbf{x}_0+\mathbf{H}\mathbf{u})dy.$$

(b) *The asymptotic mean squared error of* $\widehat{b}_0(\mathbf{x}_0)$ *is obtained by substituting* (3.17) *and* (3.18) *into* (3.15). *If* $(N|\mathbf{H}|)^{1/2}tr\{\mathcal{H}_m(\mathbf{x}_0)\mathbf{H}^2\} \le c$ *for some* $c > 0$ *and sufficiently large* N, *then*

$$\frac{\widehat{b}_0(\mathbf{x}_0) - m(\mathbf{x}_0) - E[\mathcal{R}_0(\mathbf{x}_0)]}{\{\text{var}[\mathcal{R}_0(\mathbf{x}_0)]\}^{1/2}} \to \mathcal{N}(0,1) \quad \textit{in distribution as } N\to\infty. \tag{3.19}$$

(c) *When* $\text{supp}(\pi)$ *is a proper subset of* $\text{supp}(f_{\mathbf{X}})$, *the asymptotic mean integrated squared error of* $\widehat{b}_0$ *is obtained by substituting* (3.17) *and* (3.18) *into* (3.16).

Proof. See Appendix B.

Remark 3.2. Similar to local linear fittings with i.i.d. direct data (cf. Ruppert and Wand (1994)), the leading terms of (3.17) and (3.18) are measures of the asymptotic bias and variance of $\widehat{b}_0(\mathbf{x}_0)$, respectively. Because $\mathcal{H}_m(\mathbf{x}_0)$ measures the curvature of $m(\cdot)$ at $\mathbf{x}_0$ in a particular direction, and the corresponding entry of $\mathbf{H}$ provides the amount of smoothing for that direction, more smoothing and larger curvature lead to a larger $\text{tr}\{\mathcal{H}_m(\mathbf{x}_0)\mathbf{H}^2\}$ value and an increase in the asymptotic bias of $\widehat{b}_0(\mathbf{x}_0)$. It is also interesting to see that, because (3.17) does not depend on the weight functions $w_i(y,\mathbf{x})$, the asymptotic bias of $\widehat{b}_0(\mathbf{x}_0)$ is the same as the asymptotic conditional bias of $\widehat{b}_0(\mathbf{x}_0)$ for the i.i.d. direct data given in equation (2.3) of Ruppert and Wand (1994). On the other hand, the leading constant term of (3.18) depends on $w_i(y,\mathbf{x})$. Because $\delta^*(\mathbf{t})$ may not be a continuous function of $\mathbf{t}$, (3.18) does not necessarily depend on $\int K^2(\mathbf{u})d\mathbf{u}$.

Remark 3.3. For the special case of a diagonal bandwidth matrix $\mathbf{H} = \text{diag}(h_1, \ldots, h_d)$ with $h_r > 0$ for $r = 1, \ldots, d$, (3.17) and (3.18) reduce to

$$E\left[\mathcal{R}_0(\mathbf{x}_0)\right] = \frac{1}{2}\mu_2(K)\left\{\sum_{r=1}^{d}\left[h_r^2 \left.\frac{\partial^2 m(\mathbf{x})}{\partial (x^{(r)})^2}\right|_{\mathbf{x}_0}\right]\right\}(1 + o(1)) \tag{3.20}$$

and

$$\text{var}\left[\mathcal{R}_0(\mathbf{x}_0)\right] = (Nh_1 \cdots h_d f_{\mathbf{X}}(\mathbf{x}_0))^{-1}\left[\int \delta^*(\mathbf{x}_0 + \mathbf{H}\mathbf{u})K^2(\mathbf{u})d\mathbf{u}\right](1 + o(1)), \tag{3.21}$$

respectively. The asymptotic normality (3.19) also holds if

$$(Nh_1 \cdots h_d)^{1/2}\sum_{r=1}^{d}\left[h_r^2 \left.\frac{\partial^2 m(\mathbf{x})}{\partial (x^{(r)})^2}\right|_{\mathbf{x}_0}\right] \leq C,$$

for some $C > 0$ and sufficiently large N. Although simple bandwidth structures, such as diagonal matrices, often drastically simplify the computation of $\widehat{b}_0(\mathbf{x}_0)$ and the derivation of its theoretical properties, it is still beneficial in some circumstances to use a full bandwidth matrix (e.g. Wand and Jones (1993)). Specific theoretical advantages of using full bandwidth matrices deserve substantial attention but are out of the scope of this article.

3.3. Asymptotic properties for estimates near the boundary

In practice, the ranges of the covariates are usually finite so that $\text{supp}(f_{\mathbf{X}})$ is contained in a compact set with finite end points. For i.i.d. direct data, a major advantage of local polynomial estimators over the classical kernel estimates is that they can automatically adjust for the boundary bias without resorting to special boundary correction techniques, such as boundary kernels or reflection methods (e.g. Cheng, Fan and Marron (1997)). We show here that this automatic boundary adjustment property also holds in the current biased sampling context.

We first consider the case of local polynomials with univariate covariate X, and assume that the support of $f_X(\cdot)$ is $[a, b]$ with $f_X(x) > 0$ for all $x \in [a, b]$. Then, x_a and x_b are left and right boundary points, respectively, if $x_a = a + hc$ and $x_b = b - hc$ for some $c \geq 0$. Define $\mu_k(a, c) = \int_{-c}^{\infty} u^k K(u)du$, $\mu_k(b, c) = \int_{-\infty}^{c} u^k K(u)du$, $\nu_k(a, c) = \int_{-c}^{\infty} u^k K^2(u)du$ and $\nu_k(b, c) = \int_{-\infty}^{c} u^k K^2(u)du$, for $k = 0, \ldots, 2p + 1$. Let $\mathbf{c}_p(a, c) = (\mu_{p+1}(a, c), \ldots, \mu_{2p+1}(a, c))^T$, $\mathbf{c}_p(b, c) = (\mu_{p+1}(b, c), \ldots, \mu_{2p+1}(b, c))^T$, $\mathbf{S}_0(a, c)$ be the $(p+1)\times(p+1)$ matrix whose (r, l)th element is $\mu_{r+l-2}(a, c)$, and $\mathbf{S}_0(b, c)$ be the $(p + 1) \times (p + 1)$ matrix whose (r, l)th element is $\mu_{r+l-2}(b, c)$. Similar to the approximation of (3.1), we get (see Lemma A.2 of Wu (1999) for details)

$$\left[\mathbf{I} + o_p\left(\mathbf{1}_{(p+1)\times(p+1)}\right)\right]\left(\widehat{b}(x_a) - m(x_a)\right) = \mathcal{R}_a(x_a), \tag{3.22}$$

where

$$\mathcal{R}_a(x_a) = (f_X(a))^{-1} H^{-1} (\mathbf{S}_0(a,c))^{-1} H^{-1} \left\{ \mathcal{Z}^T(x_a) \mathcal{T}_h(x_a) [\mathbf{Y} - \mathcal{Z}(x_a) m(x_a)] \right\}. \tag{3.23}$$

Furthermore, (3.22) also holds for $\widehat{b}(x_b) - m(x_b)$ with a of (3.22) and (3.23) substituted by b.

The next theorem gives the asymptotic means and variances of $e_{k+1}^T \mathcal{R}_a(x_a)$ and $e_{k+1}^T \mathcal{R}_b(x_b)$.

Theorem 3.4. *Suppose that N is sufficiently large and assumptions* A1, A2 *and* A3 *are satisfied for x_a and x_b.*

(a) *The mean of $e_{k+1}^T \mathcal{R}_a(x_a)$ is given by* (3.7) *with $(m^{(p+1)}(x_0), \mathbf{S}_0, \mathbf{c}_p)$ substituted by $(m^{p+1)}(a), \mathbf{S}_0(a,c), \mathbf{c}_p(a,c))$, while the variance of $e_{k+1}^T \mathcal{R}_a(x_a)$ is given by* (3.6) *with $(f_X(x_0), \mathbf{S}_0, \mathbf{V}(x_0))$ substituted by $(f_X(a), \mathbf{S}_0(a,c), \mathbf{V}(a,c))$, where $\mathbf{V}(a,c)$ is the $(p+1) \times (p+1)$ matrix whose $(l+1, q+1)$th element is $\delta(a+)\nu_{l+q}(a,c)$.*

(b) *The mean and variance of $e_{k+1}^T \mathcal{R}_b(x_b)$ are the same as that of $e_{k+1}^T \mathcal{R}_a(x_a)$ with $(a, a+)$ substituted by $(b, b-)$, respectively.*

Proof. The proof follows the same steps as in Appendix A by considering integrals with appropriate boundary points.

We now consider the boundary properties of the multivariate local linear estimators. Suppose that $\text{supp}(f_{\mathbf{X}})$ is compact, $\mathbf{a}$ is on the boundary of $\text{supp}(f_{\mathbf{X}})$ and $f_{\mathbf{X}}(\mathbf{a}) > 0$. Then $\mathbf{x_a}$ is a boundary point if $\mathbf{x_a} = \mathbf{a} + \mathbf{Hc}$ for some fixed point $\mathbf{c}$ in the support of $K(\cdot)$. By (2.11), it is easy to see that (3.12) holds with $\mathbf{x}_0$ substituted by $\mathbf{x_a}$. Let

$$\mathcal{D}_{\mathbf{x_a},\mathbf{H}} = \{\mathbf{c} : \ (\mathbf{x_a} + \mathbf{Hc}) \in \text{supp}\,(f_{\mathbf{X}})\} \bigcap \text{supp}(K), \qquad \mathbf{H}^* = \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{H} \end{pmatrix},$$

$$\mathbf{P_{x_a}} = \begin{pmatrix} P_{\mathbf{x_a},11} \, P_{\mathbf{x_a},12} \\ P_{\mathbf{x_a},21} \, P_{\mathbf{x_a},22} \end{pmatrix} = \int_{\mathcal{D}_{\mathbf{x_a},\mathbf{H}}} \begin{pmatrix} 1 \\ \mathbf{u} \end{pmatrix} \begin{pmatrix} 1 \ \mathbf{u}^T \end{pmatrix} K(\mathbf{u}) d\mathbf{u}$$

and

$$\mathbf{Q_{x_a}} = \begin{pmatrix} Q_{\mathbf{x_a},11} \, Q_{\mathbf{x_a},12} \\ Q_{\mathbf{x_a},21} \, Q_{\mathbf{x_a},22} \end{pmatrix} = \int_{\mathcal{D}_{\mathbf{x_a},\mathbf{H}}} \begin{pmatrix} 1 \\ \mathbf{u} \end{pmatrix} \begin{pmatrix} 1 \ \mathbf{u}^T \end{pmatrix} \delta^* (\mathbf{x_a} + \mathbf{Hu}) K^2(\mathbf{u}) d\mathbf{u}.$$

Then, similar to (3.13), we can show that

$$(1 + o_p(1)) \left[\widehat{b}_0(\mathbf{x_a}) - m(\mathbf{x_a}) \right] = \mathcal{R}_0(\mathbf{x_a}), \tag{3.24}$$

where

$$\mathcal{R}_0(\mathbf{x_a}) = (f_{\mathbf{X}}(\mathbf{a}))^{-1} e_1^T \left(\mathbf{H}^* \mathbf{P}_{\mathbf{x_a}} \mathbf{H}^*\right)^{-1} \mathbf{Z}^T(\mathbf{x_a}) \mathbf{T}_{\mathbf{H}}(\mathbf{x_a}) \left[\mathbf{Y} - \mathbf{Z}(\mathbf{x_a}) \begin{pmatrix} m(\mathbf{x_a}) \\ m'(\mathbf{x_a}) \end{pmatrix}\right].$$

The assertions of the next theorem can be derived by a similar method (see Wu (1999), for details) as in the proof of Theorem 3.4.

Theorem 3.5. *Suppose that N is sufficiently large, and assumptions* A1, A2 *and* A4 *are satisfied for $\mathbf{x_a}$. The expectation of $\mathcal{R}_0(\mathbf{x_a})$ is*

$$E\left[\mathcal{R}_0(\mathbf{x_a})\right] = e_1^T \mathbf{P}_{\mathbf{x_a}}^{-1} \left[\int_{\mathcal{D}_{\mathbf{x_a},\mathbf{H}}} \begin{pmatrix} 1 \\ \mathbf{u} \end{pmatrix} K(\mathbf{u}) \mathbf{u}^T \left(\mathbf{H} \mathcal{H}_m(\mathbf{x_a}) \mathbf{H}\right) \mathbf{u} d\mathbf{u}\right] (1 + o(1)), \tag{3.25}$$

where $\mathcal{H}_m(\mathbf{x_a})$ is the Hessian matrix of $m(\mathbf{x})$ at $\mathbf{x_a}$. The variance of $\mathcal{R}_0(\mathbf{x_a})$ is

$$\operatorname{var}\left[\mathcal{R}_0(\mathbf{x_a})\right] = (N|\mathbf{H}| f_{\mathbf{X}}(\mathbf{a}))^{-1} e_1^T \mathbf{P}_{\mathbf{x_a}}^{-1} \mathbf{Q}_{\mathbf{x_a}} \mathbf{P}_{\mathbf{x_a}}^{-1} e_1 \left(1 + o(1)\right). \tag{3.26}$$

Remark 3.4. Theorems 3.4 and 3.5 imply that, for both univariate and multivariate covariates, the asymptotic bias and variance of $\widehat{b}_k$ at interior points differ from their counterparts at boundary points only in constant terms, and their corresponding convergence rates are the same. Thus, because the $\pi(\cdot)$ of (3.4) or (3.16) is bounded, the conclusions of Theorem 3.2(c) or Theorem 3.3(c) also hold if the support of $\pi(\cdot)$ includes boundary points of $f_{\mathbf{X}}(\cdot)$.

4. Bandwidth and Kernel Choices

We only present the case for univariate covariates. It is certainly of both theoretical and practical interest to investigate the optimal bandwidth matrix and kernel choices for the general case of multivariate local linear estimators. However, because the bandwidth matrix $\mathbf{H}$ is only required to be symmetric and positive definite, the problem of minimizing the mean squared risks in Theorem 3.3 with respect to $(\mathbf{H}, K(\cdot))$ requires more sophisticated optimization techniques than the ones used in this section.

4.1. Ideal bandwidths and kernels

Because odd order local polynomial fits are preferred to even order fits (cf. Fan and Gijbels (1996, pp.76-83)), we consider only minimizing (3.9) with respect to $(h, K(\cdot))$. Setting the partial derivative of (3.9) with respect to h to zero, straightforward algebra shows that, for a given kernel satisfying condition A3(d), the ideal local bandwidth which minimizes $\mathrm{MSE}(\widehat{b}_k(x_0)$ is given by

$$h_{opt}(x_0) = N^{-1/(2p+3)} \left\{\frac{(1+2k)(f_X(x_0))^{-1} e_{k+1}^T \mathbf{S}_0^{-1} \mathbf{V}(x_0) \mathbf{S}_0^{-1} e_{k+1}}{2(p+1-k)[((p+1)!)^{-1} m^{(p+1)}(x_0) e_{k+1}^T \mathbf{S}_0^{-1} \mathbf{c}_p]^2}\right\}^{1/(2p+3)}. \tag{4.1}$$

Similarly, the ideal global bandwidth can be obtained by minimizing the corresponding mean integrated squared error and is given by

$$h_{opt}=N^{-1/(2p+3)}\left\{\frac{(1+2k)\int[(f_X(x))^{-1}e_{k+1}^T\mathbf{S}_0^{-1}\mathbf{V}(x)\mathbf{S}^{-1}e_{k+1}]\pi(x)dx}{2(p+1-k)\int[((p+1)!)^{-1}m^{(p+1)}(x)e_{k+1}^T\mathbf{S}^{-1}\mathbf{c}_p]^2\pi(x)dx}\right\}^{1/(2p+3)}. \tag{4.2}$$

Substituting $h_{opt}(x_0)$ and h_{opt} back to (3.9) and (3.11), the corresponding mean squared risks are

$$\text{MSE}\left(\widehat{b}_k(x_0);h_{opt}(x_0)\right)=N^{-\frac{2(p+1-k)}{2p+3}}\left(\frac{2p+3}{2k+1}\right)C_{1,k}(K)$$

and

$$\text{MISE}\left(\widehat{b}_k;h_{opt}\right)=N^{-\frac{2(p+1-k)}{2p+3}}\left(\frac{2p+3}{2k+1}\right)C_{2,k}(K),$$

where

$$\begin{aligned}C_{1,k}(K)=&\left\{((p+1)!)^{-1}m^{(p+1)}(x_0)e_{k+1}^T\mathbf{S}_0^{-1}\mathbf{c}_p\right\}^{\frac{2+4k}{2p+3}} \\ &\times\left\{\left(\frac{1+2k}{2(p+1-k)}\right)(f_X(x_0))^{-1}e_{k+1}^T\mathbf{S}_0^{-1}\mathbf{V}(x_0)\mathbf{S}_0^{-1}e_{k+1}\right\}^{\frac{2(p+1-k)}{2p+3}}\end{aligned} \tag{4.3}$$

and

$$\begin{aligned}C_{2,k}(K)=&\left\{\int[((p+1)!)^{-1}m^{(p+1)}(x)e_{k+1}^T\mathbf{S}_0^{-1}\mathbf{c}_p]^2\pi(x)dx\right\}^{\frac{1+2k}{2p+3}} \\ &\times\left\{\left(\frac{1+2k}{2(p+1-k)}\right)\int[(f_X(x))^{-1}e_{k+1}^T\mathbf{S}_0^{-1}\mathbf{V}(x)\mathbf{S}_0^{-1}e_{k+1}]\pi(x)dx\right\}^{\frac{2(p+1-k)}{2p+3}}.\end{aligned} \tag{4.4}$$

The ideal local and global kernels $K_{opt}(\cdot)$ are solutions of

$$C_{1,k}(K_{opt})=\min_K C_{1,k}(K) \tag{4.5}$$

and

$$C_{2,k}(K_{opt})=\min_K C_{2,k}(K), \tag{4.6}$$

respectively, subject to the constraint that $K(\cdot)$ satisfies condition A3(d). Note here that, by a rescaled kernel argument (see, for example, Marron and Nolan (1989), or Wu (1997a, Lemma 3.1)), the optimal kernels in (4.5) and (4.6) hold not only for the optimal bandwidths given in (4.1) and (4.2), but also for all other bandwidth choices.

For the special case of estimating $m(x_0)$ with a local linear estimator, i.e., $p = 1$ and $k = 0$, we have $e_1^T \mathbf{S}_0^{-1} \mathbf{c}_p = \mu_2$ and $e_1^T \mathbf{S}_0^{-1} \mathbf{V}(x_0) \mathbf{S}_0^{-1} e_1 = V_{1,1}(x_0)$. Then (4.5) is equivalent to

$$\min_K \left\{ \left[\int u^2 K(u) du \right] \left[\delta(x_0-) \int_{-\infty}^{0} K^2(u) du + \delta(x_0+) \int_0^{\infty} K^2(u) du \right]^2 \right\}, \quad (4.7)$$

subject to the constraint that $K(\cdot)$ satisfies A3(d). This problem has been solved by Wu (1997a, Theorem 3.1) and leads to an optimal kernel of the form

$$K_{opt,\beta}(u) = \frac{\Gamma_\beta(u)}{\mu(\beta)}, \quad (4.8)$$

where $\beta = \delta(x_0+)/\delta(x_0-)$, $\mu(\beta) = \int \Gamma_\beta(u) du$,

$$\Gamma_\beta(u) = \begin{cases} 1 - (u+\theta)^2, & \text{if } -1-\theta \le u < 0, \\ \beta \left[1 - (u+\theta)^2 \right], & \text{if } 0 \le u \le 1-\theta, \\ 0, & \text{otherwise,} \end{cases}$$

and θ satisfies the equation $\int_{-1-\theta}^{1-\theta} u \Gamma_\beta(u) du = 0$. Similarly, when $p = 1$ and $k = 0$, it is straightforward to derive from (4.4) that (4.6) is equivalent to

$$\min_K \left\{ \left[\int u^2 K(u) du \right] \left[\int K^2(u) du \right]^2 \right\}, \quad (4.9)$$

subject to the constraint that $K(\cdot)$ satisfies A3(d). The solution of (4.9) (cf. Wu (1997a, Theorem 3.1)) indicates that the Epanechnikov kernel (see Härdle (1990))

$$K_{opt,E}(u) = \frac{3}{4}(1-u^2) 1_{[|u| \le 1]} \quad (4.10)$$

where $1_{[\cdot]}$ is an indicator function, is globally optimal in the sense that it minimizes MISE$(\widehat{b}_0(x_0);\ h_{opt}(x_0))$. The solutions of (4.5) and (4.6) are still not available for the general case of $p \ge 2$ and $k \ge 1$.

Remark 4.1. As noted in Wu (1997a), (4.8) reduces to (4.10) when $\delta(x)$ is continuous at x_0, that is, $\beta = 1$. Contrary to local polynomial fittings with i.i.d. direct samples where the Epanechnikov kernel provides the universal optimal weighting scheme (Fan et al. (1997)), when $\delta(x_0-) \ne \delta(x_0+)$, the optimal kernel $K_{opt,\beta}(u)$ established in (4.8) is discontinuous and asymmetric at zero. A large β value would contribute to a large jump of $K_{opt,\beta}(u)$ at $u = 0$.

Remark 4.2. Because β depends on $m(x)$ and $f(y|x)$, the optimal kernel $K_{opt,\beta}(u)$ given in (4.8) can not be directly implemented in practice. In the context of density estimation with selection-biased data, Wu (1997a) suggested an

adaptive procedure to construct asymptotically optimal kernel estimators based on sample splitting. Here a similar sampling splitting adaptive procedure can also be developed for $\widehat{b}_0(x_0)$.

4.2. Cross-validation bandwidths

The ideal bandwidths derived in (4.1) and (4.2) depend on unknown components such as the regression curve and the underlying density. Hence, for practical purposes, it is desirable to develop a procedure to select appropriate bandwidths based on the data at hand. For local polynomial regression with i.i.d. direct data, two popular bandwidth selection methods are cross-validation and "plug-in" type bandwidth selectors. Theoretical properties and practical performances of these bandwidth procedures can be found in Chiu (1991), Gasser, Kneip and Köhler (1991), Hall, Sheather, Jones and Marron (1991), Sheather and Jones (1991), Fan and Gijbels (1995) and Ruppert, Sheather and Wand (1995), among others. For density estimation with biased data, Wu (1997b) investigated the asymptotic and finite sample properties of a cross-validation bandwidth procedure. Although it is known that, in many situations, "plug-in" type procedures outperform cross-validation both asymptotically and in practice (e.g. Hall, Sheather, Jones and Marron (1991)), substantial further development is needed to establish a similar "plug-in" bandwidth selector in the current context. We present here only a "leave-one-out" cross-validation global bandwidth procedure.

The basic idea is to construct a cross-validation criterion so that a bandwidth which minimizes the cross-validation score also minimizes an approximation of the mean integrated squared error. Let $\widehat{b}_0^{-(i,j)}(x)$ be the "leave-one-out" version of $\widehat{b}_0(x)$ given by (2.8), using the data with (Y_{ij}, X_{ij}) left out. Because $\widehat{b}_0^{-(i,j)}(x)$ is also a local polynomial estimator of $m(x)$, we define our cross-validation score to be

$$\mathrm{CV}(h) = N^{-1} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \left\{ \left[Y_{ij} - \widehat{b}_0^{-(i,j)}(X_{ij}) \right]^2 \pi(X_{ij}) \frac{\widehat{W}_{n_i}}{w_i(Y_{ij}, X_{ij})} \right\}, \tag{4.11}$$

where $\pi(\cdot)$ is the non-negative weight function of (3.4). Intuitively, $\mathrm{CV}(h)$ measures the weighted average prediction error of $\widehat{b}_0^{-(i,j)}(x)$. Thus, our cross-validation global bandwidth h_{cv} is defined to be a minimizer of (4.11).

Remark 4.3. To give a heuristic justification of (4.11) we can show that, by straightforward algebra and the fact that $\widehat{W}_{n_i} = W_i + o_p(N^{-\delta})$ for any $0 < \delta \leq 1/2$ (e.g. Proposition 2.2 of Gill, Vardi and Wellner (1988) or Lemma 6.1 of Wu (1997a)),

$$\mathrm{CV}(h) \approx N^{-1} \sum_{i=1}^{s} \sum_{j=1}^{n_i} \left\{ [Y_{ij} - m(X_{ij})]^2 \frac{\pi(X_{ij}) W_i}{w_i(Y_{ij}, X_{ij})} \right\} \tag{4.12}$$

$$+N^{-1}\sum_{i=1}^{s}\sum_{j=1}^{n_i}\left\{\left[\widehat{b}_0^{-(i,j)}(X_{ij})-m(X_{ij})\right]^2\frac{\pi(X_{ij})W_i}{w_i(Y_{ij},X_{ij})}\right\}$$
$$-2N^{-1}\sum_{i=1}^{s}\sum_{j=1}^{n_i}\left\{(Y_{ij}-m(X_{ij}))\left(\widehat{b}_0^{-(i,j)}(X_{ij})-m(X_{ij})\right)\frac{\pi(X_{ij})W_i}{w_i(Y_{ij},X_{ij})}\right\}.$$

The first term of the right side of (4.12) does not involve any estimator, and hence does not depend on h. By the definition of $\widehat{b}_0^{-(i,j)}(x)$, we can verify that the second and the third terms of the right side of (4.12) converge to MISE$(\widehat{b}_0)$ and zero, respectively. Thus, by minimizing CV(h), h_{cv} also approximately minimizes MISE$(\widehat{b}_0)$. However, further theoretical properties of h_{cv} and some related computational issues require substantial development.

Remark 4.4. Extending (4.11) to the multivariate case, a cross-validation bandwidth matrix $\mathbf{H}_{cv}$ can be defined to be a minimizer of a cross-validation score CV$(\mathbf{H})$ similar to (4.11). A similar approximation to the one given in (4.12) can also be established. But, because $\mathbf{H}$ is generally not a diagonal matrix, the computation of $\mathbf{H}_{cv}$ involves more sophisticated optimization algorithms than used in the univariate case.

5. Comparison with Other Smoothing Methods

At least for many special circumstances, $m(\mathbf{x})$ and its derivatives can also be estimated by other, perhaps even computationally simpler, smoothing methods. Suppose that the underlying joint density of (1.2) is identifiable from each biased sample $\{(Y_{ij},X_{ij});\ j=1,\ldots,n_i\}$. A natural alternative to the approach of Section 2 is to first construct local polynomial estimators based on each separate sample, and then estimate the regression curve and its derivatives by some linear combinations of these local polynomials. In this section, we consider the case of a univariate covariate and compare the NPML local polynomial estimators (2.8) with such linear combination estimators.

Assume that $w_i(y,x)>0$, $i=1,\ldots,s$, for all $y\in R$ and $x\in R$ in the support $\mathcal{S}$. Then, by A1 and A2, $f(y,x)$ is identifiable from each sample separately. Based on the ith sample, $m(x)$ and its derivatives can be estimated by minimizing

$$\ell^{(i)}(x)=\frac{1}{n_i}\sum_{j=1}^{n_i}\left\{\left[Y_{ij}-\sum_{k=0}^{p}\Big((X_{ij}-x)^k b_k(x)\Big)\right]^2 K_h(X_{ij}-x)\frac{\widetilde{W}_i}{w_i(Y_{ij},X_{ij})}\right\} \tag{5.1}$$

with respect to $b_k(x)$, $k=0,\ldots p$, where

$$\widetilde{W}_i=\left\{n_i^{-1}\sum_{j=1}^{n_i}w_i^{-1}(Y_{ij},X_{ij})\right\}^{-1} \tag{5.2}$$

is an estimator of W_i. Let $T_{ij}(x) = \widetilde{W}_i[n_i w_i(Y_{ij}, X_{ij})]^{-1} K_h(X_{ij} - x)$, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})^T$, $\mathcal{T}_{h,i}(x) = \text{diag}\,(T_{i1}(x), \ldots, T_{in_i}(x))$ and

$$\mathcal{Z}_i^T(x) = \begin{pmatrix} 1 & \cdots & 1 \\ (X_{i1} - x) & \cdots & (X_{in_i} - x) \\ \vdots & \vdots & \vdots \\ (X_{i1} - x)^p & \cdots & (X_{in_i} - x)^p \end{pmatrix}.$$

When $\mathcal{Z}_i^T(x)\mathcal{T}_{h,i}(x)\mathcal{Z}_i(x)$ is invertible, the pth order local polynomial estimator of $\mathbf{m}(x) = (m(x), m^{(1)}(x), \ldots, (p!)^{-1}m^{(p)}(x))^T$ based on the ith sample is

$$\left(\widetilde{b}_{0,i}(x), \ldots, \widetilde{b}_{p,i}(x)\right)^T = \left(\mathcal{Z}_i^T(x)\mathcal{T}_{h,i}(x)\mathcal{Z}_i(x)\right)^{-1}\left(\mathcal{Z}_i^T(x)\mathcal{T}_{h,i}(x)\mathbf{Y}_i\right). \tag{5.3}$$

Let p_i, $i = 1, \ldots, s$, be non-negative weights such that $\sum_{i=1}^s p_i = 1$. A linear combination estimator of $\mathbf{m}(x)$ based on p_i and $(\widetilde{b}_{0,i}(x), \ldots, \widetilde{b}_{p,i}(x))^T$ is

$$\widetilde{\mathbf{b}}(x) = \left(\widetilde{b}_0(x), \ldots, \widetilde{b}_p(x)\right)^T = \sum_{i=1}^s \left\{p_i \left(\widetilde{b}_{0,i}(x), \ldots, \widetilde{b}_{p,i}(x)\right)^T\right\}. \tag{5.4}$$

Because $\widetilde{W}_i = W_i + o_p(N^{-\delta})$ for any $0 < \delta < 1/2$, it can be shown by similar calculations as (3.1) and (3.2) that, for an interior point x_0,

$$\left[\mathbf{I} + o_p\left(\mathbf{1}_{(p+1)\times(p+1)}\right)\right]\left(\widetilde{\mathbf{b}}(x_0) - \mathbf{m}(x_0)\right) = \widetilde{\mathcal{R}}(x_0), \tag{5.5}$$

where

$$\widetilde{\mathcal{R}}(x_0) = (f_X(x_0))^{-1} H^{-1}\mathbf{S}_0^{-1}H^{-1}\sum_{i=1}^s \left\{p_i\left[\mathcal{Z}_i^T(x_0)\mathcal{T}_{h,i}(x_0)\left(\mathbf{Y}_i - \mathcal{Z}_i(x_0)\mathbf{m}(x_0)\right)\right]\right\}. \tag{5.6}$$

Thus, with the same modification as at (3.3) and (3.4), the mean squared error of $\widetilde{b}_k(x_0)$ and the mean integrated squared error of $\widetilde{b}_k$ can be defined by

$$\text{MSE}\left(\widetilde{b}_k(x_0)\right) = E\left\{\left[e_{k+1}^T\widetilde{\mathcal{R}}(x_0)\right]^2\right\} = \left\{E\left[e_{k+1}^T\widetilde{\mathcal{R}}(x_0)\right]\right\}^2 + \text{var}\left[e_{k+1}^T\widetilde{\mathcal{R}}(x_0)\right] \tag{5.7}$$

and

$$\text{MISE}\left(\widetilde{b}_k\right) = \int E\left\{\left[e_{k+1}^T\widetilde{\mathcal{R}}(x)\right]^2\right\}\pi(x)dx. \tag{5.8}$$

Again, as in Section 4.1, odd order fits are preferred to even order fits. Thus, our discussion is limited to the case that $p-k$ is odd. By similar derivations as in the proof of Theorem 3.1, the asymptotic expressions of the mean and covariance of $\widetilde{\mathcal{R}}(x_0)$ are summarized in the following theorem, whose proof is given in Wu (1999, Appendix C).

Theorem 5.1. *Suppose that $f_X(x_0) > 0$, Assumptions* A1, A2 *and* A3 *are satisfied, $p-k$ is odd and N is sufficiently large. When $h \le c_1 N^{-r/[2(p+1-k)]}$ for some constants $c_1 > 0$ and $0 < r < 1$, $E[e_{k+1}^T \widetilde{\mathcal{R}}(x_0)]$ is given by the right side of* (3.7). *When $h \ge c_2 N^{(r-1)/(1+k_1+k_2)}$ for some constants $c_2 > 0$ and $0 < r < 1$, the covariance between $e_{k_1+1}^T \widetilde{\mathcal{R}}(x_0)$ and $e_{k_2+1}^T \widetilde{\mathcal{R}}(x_0)$ is*

$$\begin{aligned} &\operatorname{cov}\left[e_{k_1+1}^T \widetilde{\mathcal{R}}(x_0), e_{k_2+1}^T \widetilde{\mathcal{R}}(x_0)\right] \\ &= \left[\frac{1}{f_X(x_0) N h^{1+k_1+k_2}}\right] e_{k_1+1}^T \mathbf{S}_0^{-1} \widetilde{\mathbf{V}}_0(x_0) \mathbf{S}_0^{-1} e_{k_2+1}(1+o(1)), \end{aligned} \tag{5.9}$$

where $\widetilde{\mathbf{V}}_0(x_0)$ is the $(p+1)\times(p+1)$ matrix whose $(l+1, q+1)$th element is

$$\widetilde{V}_{l+1,q+1}(x_0) = \widetilde{\delta}(x_0-) \int_{-\infty}^{0} u^{l+q} K^2(u) du + \widetilde{\delta}(x_0+) \int_{0}^{\infty} u^{l+q} K^2(u) du$$

for $0 \le l \le p$ and $0 \le q \le p$, with

$$\widetilde{\delta}(t) = \int \left\{ [y - m(t)]^2 \left[\sum_{i=1}^{s} \left(\frac{p_i^2 W_i}{\lambda_i w_i(y,t)} \right) \right] f(y|t) \right\} dy. \tag{5.10}$$

Remark 5.1. A direct consequence of (5.9) is that, when $h \ge c_2 N^{(r-1)/(1+2k)}$ for $c_2 > 0$ and $0 < r < 1$, the variance of $e_{k+1}^T \widetilde{\mathcal{R}}(x_0)$ is

$$\operatorname{var}\left[e_{k+1}^T \widetilde{\mathcal{R}}(x_0)\right] = \left[\frac{1}{f_X(x_0) N h^{1+2k}}\right] e_{k+1}^T \mathbf{S}_0^{-1} \widetilde{\mathbf{V}}_0(x_0) \mathbf{S}_0^{-1} e_{k+1}(1+o(1)). \tag{5.11}$$

Thus, when $c_2 N^{(r-1)/(1+2k)} \le h \le c_1 N^{-r/[2(p+1-k)]}$ for $c_1 > 0$, $c_2 > 0$ and $0 < r < 1$, explicit expressions for $\operatorname{MSE}(\widetilde{b}_k(x_0))$ and $\operatorname{MISE}(\widetilde{b}_k)$ follow immediately from (5.7), (5.8), (3.7) and (5.11).

Remark 5.2. Comparing Theorem 5.1 with Theorem 3.1, we see that $\widehat{\mathbf{b}}(x_0)$ and $\widetilde{\mathbf{b}}(x_0)$ have the same asymptotic bias but different asymptotic variances. For the special case of $p = 1$ and $k = 0$, we can deduce from (3.6) and (5.11) that

$$\operatorname{var}\left[e_1^T \mathcal{R}(x_0)\right] = \left(\frac{1}{f_X(x_0) N h}\right) \left[\delta(x_0-) \int_{-\infty}^{0} K^2(u) du + \delta(x_0+) \int_{0}^{\infty} K^2(u) du\right]$$

and

$$\operatorname{var}\left[e_1^T \widetilde{\mathcal{R}}(x_0)\right] = \left(\frac{1}{f_X(x_0) N h}\right) \left[\widetilde{\delta}(x_0-) \int_{-\infty}^{0} K^2(u) du + \widetilde{\delta}(x_0+) \int_{0}^{\infty} K^2(u) du\right].$$

By Jensen's inequality,

$$\sum_{i=1}^{s} \left(\frac{p_i^2 W_i}{\lambda_i w_i(y,t)} \right) \ge \left(\sum_{i=1}^{s} \frac{\lambda_i w_i(y,t)}{W_i} \right)^{-1} = J(y,t) \tag{5.12}$$

for all $y \in R$, $t \in R$ and $p_i \geq 0$ satisfying $\sum_{i=1}^{s} p_i = 1$. When $s > 1$, the equality sign of (5.12) holds if and only if $p_i = \lambda_i w_i(y,t) J(y,t) W_i^{-1}$. Then, (5.10) and (5.12) imply that $\delta(t) \leq \widetilde{\delta}(t)$ for all $t \in R$. Consequently, $\text{var}[e_1^T \mathcal{R}(x_0)] \leq \text{var}[e_1^T \widetilde{\mathcal{R}}(x_0)]$ for any choice of p_i, and equality holds if and only if $p_i = \lambda_i w_i(y,t) J(y,t) W_i^{-1}$. However, the two variances are identical when $s = 1$ (e.g., Ahmad (1995)). Comparison between (3.6) and (5.11) has yet to be developed for general values of k and p.

Acknowledgements

This research was supported in part by the National Institute on Drug Abuse grant R01 DA10184-01. The author is grateful to Dr. Chin-Tsang Chiang and Ms. Vivian W.-S. Yuan for computing the numerical results. The author would also like to thank an associate editor and two referees for their insightful suggestions and comments that have greatly strengthened both the presentation and the technical aspects of this article.

Appendix A

Proof of Theorem 3.1. We first compute $E[e_{k+1}^T \mathcal{R}(x_0)]$. Let $\Delta_{ij}(x_0) = Y_{ij} - \sum_{k=0}^{p} \left[(k!)^{-1} m^{(k)}(x_0) (X_{ij} - x_0)^k \right]$. By the definitions of $\mathbf{Y}$, $\mathcal{Z}(x_0)$, $\mathcal{T}_h(x_0)$ and $\mathbf{m}(x_0)$, we have $\mathbf{Y} - \mathcal{Z}(x_0)\mathbf{m}(x_0) = (\Delta_{11}(x_0), \ldots, \Delta_{sn_s}(x_0))^T$ and, for $k = 0, \ldots, p$, the $(k+1)$th element of $\mathcal{Z}^T(x_0)\mathcal{T}_h(x_0)[\mathbf{Y} - \mathcal{Z}(x_0)\mathbf{m}(x_0)]$ is

$$\begin{aligned} & e_{k+1}^T \mathcal{Z}^T(x_0) \mathcal{T}_h(x_0) \left[\mathbf{Y} - \mathcal{Z}(x_0)\mathbf{m}(x_0) \right] \\ &= \sum_{i=1}^{s} \sum_{j=1}^{n_i} \left\{ \Delta_{ij}(x_0) (X_{ij} - x_0)^k N^{-1} \widehat{J}_N(Y_{ij}, X_{ij}) K_h(X_{ij} - x_0) \right\}. \end{aligned} \tag{A.1}$$

Because $\widehat{J}_N(Y_{ij}, X_{ij})$ involves random jumps at each observation point, we consider the asymptotic expectation and variance of

$$\mathbf{A}(x_0) = (A_0(x_0), \ldots, A_p(x_0))^T, \tag{A.2}$$

where $A_k(x_0) = \sum_{i=1}^{s} \sum_{j=1}^{n_i} \left\{ \Delta_{ij} (X_{ij} - x_0)^k N^{-1} J(Y_{ij}, X_{ij}) K_h(X_{ij} - x_0) \right\}$. Direct computation with change of variables shows that

$$\begin{aligned} E\left[A_k(x_0)\right] = & \frac{1}{N} \sum_{i=1}^{s} n_i \iint \left\{ \left[y - \sum_{r=0}^{p} (r!)^{-1} m^{(r)}(x_0)(t - x_0)^r \right] (t - x_0)^k \right. \\ & \left. \times J(y,t) K_h(t - x_0) \left(\frac{w_i(y,t)}{W_i} \right) f(y,t) \right\} dy dt \\ = & \left\{ \int \left[m(t) - \sum_{r=0}^{p} (r!)^{-1} m^{(r)}(x_0)(t - x_0)^r \right] (t - x_0)^k K_h(t - x_0) f_X(t) dt \right\} \end{aligned}$$

$$
\begin{aligned}
&\times(1+o(1))\\
=&\left\{\int\left[[(p+1)!]^{-1}m^{(p+1)}(x_0)(t-x_0)^{p+1}\right.\right.\\
&\left.\left.+o\left(|t-x_0|^{p+1}\right)\right](t-x_0)^k K_h(t-x_0)f_X(t)dt\right\}(1+o(1))\\
=&\left[\frac{h^{p+1+k}}{(p+1)!}m^{(p+1)}(x_0)f_X(x_0)\mu_{p+1+k}\right](1+o(1)). \qquad \text{(A.3)}
\end{aligned}
$$

Note that (A.3) holds for both odd and even $p-k$. Thus, when $p-k$ is odd,

$$
\begin{aligned}
&(f_X(x_0))^{-1}e_{k+1}^T H^{-1}\mathbf{S}_0^{-1}H^{-1}E\left[\mathbf{A}(x_0)\right]\\
&=\left(\frac{h^{p+1-k}}{(p+1)!}\right)m^{(p+1)}(x_0)e_{k+1}^T\mathbf{S}_0^{-1}\mathbf{c}_p(1+o(1)). \qquad \text{(A.4)}
\end{aligned}
$$

Using Lemma 6.1 of Wu (1997b), A3(b) and similar calculation as in (A.3), we can show that

$$
\begin{aligned}
&E\left|\sum_{i=1}^{s}\sum_{j=1}^{n_i}\left\{\left[\Delta_{ij}(x_0)\left(X_{ij}-x_0\right)^k N^{-1}K_h(X_{ij}-x_0)\right]\left(\widehat{J}_N(Y_{ij},X_{ij})-J(Y_{ij},X_{ij})\right)\right\}\right|\\
&\le E\left\{\left|\left(\sup_{(x,y)}\left|\widehat{J}_N(y,x)-J(y,x)\right|\right)\sum_{i=1}^{s}\sum_{j=1}^{n_i}\left[\Delta_{ij}(x_0)\left(X_{ij}-x_0\right)^k N^{-1}K_h(X_{ij}-x_0)\right]\right|\right\}\\
&=o\left(h^{p+1+k}\right). \qquad \text{(A.5)}
\end{aligned}
$$

Then (3.7) is a direct consequence of (3.2), (A.1), (A.2), (A.4) and (A.5).

Under the conditions of Theorem 3.1(c), direct computation and Taylor expansions of $m(t)$ and $f_X(t)$ show that

$$
\begin{aligned}
E\left[A_k(x_0)\right]=&\left\{\frac{h^{p+1+k}}{(p+1)!}m^{(p+1)}(x_0)f_X(x_0)\mu_{p+1+k}+h^{p+2+k}\mu_{p+2+k}\right.\\
&\left.\times\left[\frac{m^{(p+1)}(x_0)}{(p+2)!}f_X(x_0)+\frac{m^{(p+1)}(x_0)}{(p+1)!}f_X'(x_0)\right]\right\}(1+o(1)). \qquad \text{(A.6)}
\end{aligned}
$$

However, when $p-k$ is even, assumption A3(d) implies that the $(k+1)$th element of $\mathbf{S}_0^{-1}\mathbf{c}_p$ is 0 (e.g., Fan and Gijbels (1996, Section 3.7)), hence

$$
(f_X(x_0))^{-1}e_{k+1}^T H^{-1}\mathbf{S}_0^{-1}H^{-1}\left[\frac{h^{p+1+k}}{(p+1)!}m^{(p+1)}(x_0)f_X(x_0)\mathbf{c}_p\right]=0. \qquad \text{(A.7)}
$$

Then (3.8) follows from (3.2), (A.5), (A.6) and (A.7).

To prove (3.5), we consider the following expansion:

$$
A_{k_1}(x_0)A_{k_2}(x_0)=I(x_0)+II(x_0), \qquad \text{(A.8)}
$$

where

$$I(x_0)=\left(\frac{1}{Nh}\right)^2\sum_{i=1}^{s}\sum_{j=1}^{n_i}\left\{(\Delta_{ij}(x_0))^2(J(Y_{ij},X_{ij}))^2(X_{ij}-x_0)^{k_1+k_2}K^2\left(\frac{X_{ij}-x_0}{h}\right)\right\},$$

$$II(x_0)=\left(\frac{1}{Nh}\right)^2\sum_{(i_1,j_1)\neq(i_2,j_2)}\left\{\Delta_{i_1j_1}(x_0)J(Y_{i_1j_1},X_{i_1j_1})(X_{i_1j_1}-x_0)^{k_1}K\left(\frac{X_{i_1j_1}-x_0}{h}\right)\right.$$
$$\left.\times\Delta_{i_2j_2}(x_0)J(Y_{i_2j_2},X_{i_2j_2})\,(X_{i_2j_2}-x_0)^{k_2}\,K\left(\frac{X_{i_2j_2}-x_0}{h}\right)\right\}.$$

By the definitions of $\delta(t)$ and $V_{k_1+1,k_2+1}(x_0)$, we get

$$\begin{aligned}E\left[I(x_0)\right] &= \left(\frac{1}{Nh^2}\right)\iint\left\{\left[y-\sum_{k=0}^{p}\left(\frac{m^{(k)}(x_0)}{k!}(t-x_0)^k\right)\right]^2\right. \\ &\qquad\left.\times(t-x_0)^{k_1+k_2}J(y,t)K^2\left(\frac{t-x_0}{h}\right)f(y,t)\right\}dydt \\ &= \left(\frac{1}{Nh}\right)\left\{\int\delta(t)(t-x_0)^{k_1+k_2}K^2\left(\frac{t-x_0}{h}\right)f_X(t)dt\right\}(1+o(1)) \\ &= \left(\frac{f_X(x_0)}{Nh}\right)h^{k_1+k_2}V_{k_1+1,k_2+1}(x_0)(1+o(1)).\end{aligned} \tag{A.9}$$

Because the observations are independent, it is straightforward to verify that

$$E\left[II(x_0)\right]=\left(\frac{N(N-1)}{N^2}\right)\left\{E\left[A_{k_1}(x_0)\right]E\left[A_{k_2}(x_0)\right]\right\}. \tag{A.10}$$

Combining (A.8), (A.9) and (A.10), we have

$$\operatorname{cov}\left(A_{k_1}(x_0),A_{k_2}(x_0)\right)=\left(\frac{f_X(x_0)}{Nh}\right)h^{k_1+k_2}V_{k_1+1,k_2+1}(x_0)(1+o(1)). \tag{A.11}$$

The covariance of $(f(x_0))^{-1}e_{k_1+1}^TH^{-1}\mathbf{S}_0^{-1}H^{-1}\mathbf{A}(x_0)$ and $(f(x_0))^{-1}e_{k_2+1}^TH^{-1}\mathbf{S}_0^{-1}$ $H^{-1}\mathbf{A}(x_0)$ is then computed using (A.11) and is approximated by the right side of (3.5). Furthermore, by Lemma 6.1 of Wu (1997b) and A3(b),

$$\begin{aligned}&E\left|(Nh)^{-2}\left\{\sum_{i=1}^{s}\sum_{j=1}^{n_i}\left[\Delta_{ij}(x_0)(X_{ij}-x_0)^{k_1}\right.\right.\right. \\ &\quad\left.\times\left(\widehat{J}_N(Y_{ij},X_{ij})-J(Y_{ij},X_{ij})\right)K_h(X_{ij}-x_0)\right] \\ &\quad\left.\left.\times\sum_{i=1}^{s}\sum_{j=1}^{n_i}\left[\Delta_{ij}(x_0)(X_{ij}-x_0)^{k_2}\left(\widehat{J}_N(Y_{ij},X_{ij})-J(Y_{ij},X_{ij})\right)K_h(X_{ij}-x_0)\right]\right\}\right| \\ &\le E\left\{\left(\sup_{(x,y)}\left|\widehat{J}_N(y,x)-J(y,x)\right|\right)^2\right.\end{aligned}$$

$$\times \left[\sum_{i=1}^{s}\sum_{j=1}^{n_i} \Delta_{ij}(x_0)(X_{ij}-x_0)^{k_1} K_h(X_{ij}-x_0)\right]$$
$$\times \left[\sum_{i=1}^{s}\sum_{j=1}^{n_i} \Delta_{ij}(x_0)(X_{ij}-x_0)^{k_2} K_h(X_{ij}-x_0)\right]\Bigg\} = O\left(N^{-1}\right). \quad \text{(A.12)}$$

Then (A.1) and the conditions of Theorem 3.1(a) imply that (3.5) holds for the covariance between $e_{k_1+1}^T \mathcal{R}(x_0)$ and $e_{k_2+1}^T \mathcal{R}(x_0)$. This completes the proof of Theorem 3.1.

Appendix B

Proof of Theorem 3.3. Substituting $\widehat{J}_N(Y_{ij}, \mathbf{X}_{ij})$ of $\mathcal{R}_0(\mathbf{x}_0)$ with $J(Y_{ij}, \mathbf{X}_{ij})$, we first consider the asymptotic mean and variance of

$$\mathcal{R}_0^*(\mathbf{x}_0) = \frac{1}{f_{\mathbf{X}}(\mathbf{x}_0)N|\mathbf{H}|}\sum_{i=1}^{s}\sum_{j=1}^{n_i}\left[J(Y_{ij}, \mathbf{X}_{ij})\Delta_{ij}^*(\mathbf{x}_0) K\left(\mathbf{H}^{-1}(\mathbf{X}_{ij}-\mathbf{x}_0)\right)\right].$$

By direct integration and change of variables, we get

$$\begin{aligned}
E\left[\mathcal{R}_0^*(\mathbf{x}_0)\right] =& \frac{1}{f_{\mathbf{X}}(\mathbf{x}_0)N|\mathbf{H}|}\sum_{i=1}^{s}\sum_{j=1}^{n_i}\iint J(y,\mathbf{t})\frac{w_i(y,\mathbf{t})}{W_i}\left[y-m(\mathbf{x}_0)-(\mathbf{t}-\mathbf{x}_0)^T m'(\mathbf{x}_0)\right] \\
&\times K\left(\mathbf{H}^{-1}(\mathbf{t}-\mathbf{x}_0)\right) f(y,\mathbf{t})dyd\mathbf{t} \\
=& \frac{1}{f_{\mathbf{X}}(\mathbf{x}_0)|\mathbf{H}|}\int\left\{\int\left[y-m(\mathbf{x}_0)-(\mathbf{t}-\mathbf{x}_0)^T m'(\mathbf{x}_0)\right] f(y|\mathbf{t})dy\right\} \\
&\times K\left(\mathbf{H}^{-1}(\mathbf{t}-\mathbf{x}_0)\right) f_{\mathbf{X}}(\mathbf{t})d\mathbf{t} \\
=& \frac{1}{2f_{\mathbf{X}}(\mathbf{x}_0)}\int K(\mathbf{u})(\mathbf{H}\mathbf{u})^T\mathcal{H}_m(\mathbf{x}_0)(\mathbf{H}\mathbf{u})f_{\mathbf{X}}(\mathbf{x}_0+\mathbf{H}\mathbf{u})d\mathbf{u}(1+o(1)) \\
=& \frac{1}{2}\mu_2(K)\mathrm{tr}\left\{\mathcal{H}_m(\mathbf{x}_0)\mathbf{H}^2\right\}(1+o(1)).
\end{aligned}$$

For the variance of $\mathcal{R}_0^*(\mathbf{x}_0)$ we consider $\mathcal{R}_0^*(\mathbf{x}_0) = I^*(\mathbf{x}_0) + II^*(\mathbf{x}_0)$, where

$$I^*(\mathbf{x}_0) = (N|\mathbf{H}|f_{\mathbf{X}}(\mathbf{x}_0))^{-2}\sum_{i=1}^{s}\sum_{j=1}^{n_i}\left\{J^2(Y_{ij},\mathbf{X}_{ij})\left(\Delta_{ij}^*(\mathbf{x}_0)\right)^2 K^2\left(\mathbf{H}^{-1}(\mathbf{X}_{ij}-\mathbf{x}_0)\right)\right\},$$

$$\begin{aligned}
II^*(\mathbf{x}_0) =& (N|\mathbf{H}|f_{\mathbf{X}}(\mathbf{x}_0))^{-2}\sum_{(i_1,j_1)\neq(i_2,j_2)}\Big\{J(Y_{i_1j_1},\mathbf{X}_{i_1j_1})\Delta_{i_1j_1}^*(\mathbf{x}_0) \\
&\times K\left(\mathbf{H}^{-1}\left(\mathbf{X}_{i_1j_1}-\mathbf{x}_0\right)\right) J(Y_{i_2j_2},\mathbf{X}_{i_2j_2})\Delta_{i_2j_2}^*(\mathbf{x}_0) K\left(\mathbf{H}^{-1}\left(\mathbf{X}_{i_2j_2}-\mathbf{x}_0\right)\right)\Big\}.
\end{aligned}$$

Because $E[II^*(\mathbf{x}_0)] = [E(\mathcal{R}_0^*(\mathbf{x}_0))]^2(1+o(1))$ and

$$E\left[I^*(\mathbf{x}_0)\right] = (N|\mathbf{H}|f_{\mathbf{X}}(\mathbf{x}_0))^{-2}\iint J^2(y,\mathbf{t})\frac{w_i(y,\mathbf{t})}{W_i}\left[y-m(\mathbf{x}_0)-(\mathbf{t}-\mathbf{x}_0)^T m'(\mathbf{x}_0)\right]^2$$

$$\begin{aligned}
&\times K^2\left(\mathbf{H}^{-1}(\mathbf{t}-\mathbf{x}_0)\right) f(y,\mathbf{t})dyd\mathbf{t}\\
=&\frac{(f_{\mathbf{X}}(\mathbf{x}_0))^{-2}}{N|\mathbf{H}|^2}\int \delta^*(\mathbf{t})K^2\left(\mathbf{H}^{-1}(\mathbf{t}-\mathbf{x}_0)\right) f_{\mathbf{X}}(\mathbf{t})d\mathbf{t}\\
=&(N|\mathbf{H}|f_{\mathbf{X}}(\mathbf{x}_0))^{-1}\left[\int \delta^*(\mathbf{x}_0+\mathbf{H}\mathbf{u})K^2(\mathbf{u})d\mathbf{u}\right](1+o(1)),
\end{aligned}$$

the variance of $\mathcal{R}_0^*(\mathbf{x}_0)$ is given by the right side of (3.18).

Note that the conclusions of (A.5) and (A.12) also hold for the current multivariate context. Then, (3.17) and (3.18) hold because $\mathcal{R}_0^*(\mathbf{x}_0)$ and $\mathcal{R}_0(\mathbf{x}_0)$ have the same asymptotic expectation and variance. The assertions in (b) and (c) can be shown by the same arguments as in the proof of Theorem 3.2.

References

Ahmad, I. A. (1995). On multivariate kernel estimation for samples from weighted distributions. *Statist. Probab. Lett.* **22**, 121-129.

Bickel, P. J. and Ritov, Y. (1991). Large sample theory of estimation in biased sampling regression models I. *Ann. Statist.* **19**, 797-816.

Cheng, M.-Y., Fan, J. and Marron, J. S. (1997). On automatic boundary corrections. *Ann. Statist.* **25**, 1691-1708.

Chiu, S. T. (1991). Bandwidth selection for kernel density estimation. *Ann. Statist.* **19**, 1528-1546.

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

Fan, J., Gasser, T., Gijbels, I., Brockmann, M. and Engel, J. (1997). Local polynomial regression: optimal kernels and asymptotic minimax efficiency. *Ann. Inst. Statist. Math.* **49**, 79-99.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc. Ser. B* **57**, 371-394.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman & Hall, London

Gasser, T., Kneip, A. and Köhler, W. (1991). A flexible and fast method for automatic smoothing. *J. Amer. Statist. Assoc.* **86**, 643-652.

Gill, R. D., Vardi, Y. and Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.* **16**, 1069-1112.

Hall, P., Sheather, S. J., Jones, M. C. and Marron, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* **78**, 263-271.

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

Jewell, N. (1985). Regression from stratified samples of dependent variable. *Biometrika* **72**, 11-21.

Jewell, N. and Quesenberry, C. P. (1986). Regression analysis based on stratified samples. *Biometrika* **73**, 605-614.

Jones, M. C. (1991). Kernel density estimation for length biased data. *Biometrika* **78**, 511-519.

Marron, J. S. and Nolan, D. (1989). Canonical kernels for density estimation. *Statist. Probab. Lett.* **7**, 195-199.

Patil, G. P., Rao, C. R. and Zelen, M. (1988). Weighted distributions. In *Encyclopedia of Statistical Sciences* **9** (Edited by S. Kotz and N. L. Johnson), 565-571. Wiley, New York.

Patil, G. P. and Taillie, C. (1989). Probing encountered data, meta analysis and weighted distribution methods. In *Statistical Data Analysis and Inferences* (Edited by Y. Dodge), 317-346. North-Holland.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12**, 1215-1230.

Ruppert, D., Sheather, S. J. and Wand, M. P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90**, 1257-1270.

Ruppert, D. and Wand, M. P. (1994). Multivariate weighted least squares regression. *Ann. Statist.* **22**, 1346-1370.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *J. Roy. Statist. Soc. Ser. B* **53**, 683-690.

Sköld, M. (1999). Kernel regression in the presence of size-bias. *J. Nonparametr. Statist.* To appear.

Stone, C. J. (1982). Optimal global rates of convergence of nonparametric regression. *Ann. Statist.* **10**, 1040-1053.

Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12**, 1285-1297.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *Ann. Statist.* **10**, 616-620.

Vardi, Y. (1985). Empirical distributions in selection bias models (with discussions). *Ann. Statist.* **13**, 178-205.

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

Wand, M. P. and Jones, M. C. (1993). Comparison of smoothing parametrizations in bivariate kernel density estimation. *J. Amer. Statist. Assoc.* **88**, 520-528.

Wu, C. O. (1997a). The effects of kernel choices in density estimation with biased data. *Statist. Probab. Lett.* **34**, 373-383.

Wu, C. O. (1997b). A cross-validation bandwidth choice for kernel density estimates with selection biased data. *J. Multivariate Anal.* **61**, 38-60.

Wu, C. O. (1999). Large sample properties for local polynomial regression with selection biased data. Technical Report #604, Department of Mathematical Sciences, The Johns Hopkins University.

Department of Mathematical Sciences, G.W.C. Whiting School of Engineering, The Johns Hopkins University, 34th & Charles Streets, Baltimore, MD 21218-2682, U.S.A.

E-mail: colin@mts.jhu.edu

(Received May 1997; accepted October 1999)