USING SIMEX FOR SMOOTHING-PARAMETER CHOICE IN ERRORS-IN-VARIABLES PROBLEMS: technical details.

Aurore Delaigle[1,2] and Peter Hall[2,3]

Cite this work as:

Delaigle, A. and Hall, P. (2008). Using SIMEX for smoothing-parameter choice in errors-in-variables problems. *JASA*, **103**, 280-287.

and

Delaigle, A. and Hall, P. (2008). Technical details for the 2008 JASA paper: Using SIMEX for smoothing-parameter choice in errors-in-variables problems. Available at http://www.ms.unimelb.edu.au/∼aurored/links.html

WARNING:

This file is part of an old version of the paper. In particular, the text that describes the method and estimators is not the one that was published in JASA. The interest of this file are the proofs, but 1) it would have taken too much time to rewrite them using the exact notations and numberings of the published short version; (2) on the other hand, giving only the proofs without the old text describing the method would have been unreadable.

[1] Department of Mathematics, University of Bristol, Bristol BS8 1TW, UK.

[2] Department of Mathematics and Statistics, The University of Melbourne, VIC 3010, Australia.

[3] Research supported by the Australian Research Council.

1. METHODOLOGY

1.1. Model. Data (W_j, Y_j) are generated by the model,

$$Y_j = g(X_j) + V_j\,, \quad W_j = X_j + U_j\,, \tag{1.1}$$

where the variates U_j, V_j and X_j, for $1 \leq j < \infty$, are totally independent, and the sequences $U_1, U_2, \ldots$, $V_1, V_2, \ldots$ and $X_1, X_2, \ldots$ are identically distributed as U, V and X, respectively. The distribution of U is known, the distribution of V is unknown but has zero mean, and we wish to estimate the smooth function g.

1.2. Estimator. Let K denote a kernel function, let $K^{\rm Ft}(t) = \int e^{itu}\, K(u)\, du$ denote its Fourier transform, and assume that,

the function $(1+x^2)\,|K(x)|$ is integrable, $K^{\rm Ft}$ vanishes outside a compact set, and $K^{\rm Ft}(0) \neq 0$. (1.2)

Let $h > 0$ be a bandwidth, write f_U for the density of U, let $f_U^{\rm Ft}$ be the characteristic function of the distribution of U, and define

$$K_U(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itu}\, K^{\rm Ft}(t)\, f_U^{\rm Ft}(t/h)^{-1}\, dt\,. \tag{1.3}$$

A kernel estimator of g, constructed from the n data pairs (W_j, Y_j), was proposed by Fan and Truong (1993) and is given by

$$\hat{g}(x) = \frac{\sum_j Y_j\, K_U\{(x - W_j)/h\}}{\sum_j K_U\{(x - W_j)/h\}}\,, \tag{1.4}$$

where, here and in (1.8) below, each summation is over $1 \leq j \leq n$.

In order for the integrand of the integral in (1.3) not to become too large, it is generally assumed that the characteristic function $f_U^{\rm Ft}$ of the distribution of U satisfies,

$$f_U^{\rm Ft} \text{ does not vanish on the real line}\,. \tag{1.5}$$

1.3. Cross-validation criteria. If we knew the values of $X_1, \ldots, X_n$, in addition to those of $(W_1, Y_1), \ldots, (W_n, Y_n)$, we could compute a conventional cross-validation bandwidth $\hat{h}_0$: $\hat{h}_0 = \operatorname{argmin} \mathrm{CV}_0(h)$, where

$$\mathrm{CV}_0(h) = \frac{1}{n} \sum_{j=1}^{n} \{Y_j - \hat{g}_{-j}(X_j)\}^2 \, p(X_j) \,. \tag{1.6}$$

Here, $\hat{g}_{-j}$ denotes the version of $\hat{g}$, at (1.4), computed on omitting the jth data pair (W_j, Y_j) from the sample. The nonnegative function p in (1.6) denotes a weight, used to prevent CV_0 from becoming too large through attempting to estimate $g(x)$ for values of x that lie in the tails of the distribution of X. In particular, we ask that $p(x) \to 0$ as $|x| \to \infty$, and that p be an integrable function.

However, the X_j's are not observable, and so CV_0 is not a practical criterion. We develop instead two versions of CV_0 for higher levels of observation error. To this end, let $U_1^*, U_2^*, \ldots$ and $U_1^{**}, U_2^{**}, \ldots$ denote independent and identically distributed random variables, independent too of the data pairs (W_j, Y_j), and having the distribution of U. Put $W_j^* = W_j + U_j^*$ and $W_j^{**} = W_j + U_j^* + U_j^{**}$ for $1 \le j \le n$, and consider the problem of estimating g_1 from the contaminated data (W_j^*, Y_j), or of estimating g_2 from values of (W_j^{**}, Y_j), where

$$g_1(x) = E(Y \mid W = x) \,, \quad g_2(x) = E(Y \mid W^* = x) \,. \tag{1.7}$$

Appropriate estimators can be based on (1.4):

$$\hat{g}_1^*(x) = \frac{\sum_j Y_j \, K_U\{(x - W_j^*)/h\}}{\sum_j K_U\{(x - W_j^*)/h\}} \,, \quad \hat{g}_2^{**}(x) = \frac{\sum_j Y_j \, K_U\{(x - W_j^{**})/h\}}{\sum_j K_U\{(x - W_j^{**})/h\}} \,. \tag{1.8}$$

In this problem the variables W_j and W_j^* are known, and so we can use standard cross-validation to determine the bandwidths appropriate for estimating g_1 and g_2. The respective criteria are

$$\mathrm{CV}^*(h) = \frac{1}{n} \sum_{j=1}^{n} \{Y_j - \hat{g}^*_{1,-j}(W_j)\}^2 \, p(W_j) \,, \tag{1.9}$$

$$\mathrm{CV}^{**}(h) = \frac{1}{n} \sum_{j=1}^{n} \{Y_j - \hat{g}^{**}_{2,-j}(W^*_j)\}^2 \, p(W^*_j) \,, \tag{1.10}$$

where, for $k = 1$ and 2, the subscript $-j$ in $\hat{g}_{k,-j}$ indicates that we omit the jth data pair when constructing the estimator $\hat{g}_k$. The function p in (1.9) and (1.10) is identical to that in (1.6).

1.4. Using cross-validation to choose bandwidth. The cross-validation bandwidths derived from the criteria at (1.9) and (1.10) are $\hat{h}^*_1 = \mathrm{argmin}\, \mathrm{CV}^*(h)$ and $\hat{h}^{**}_2 = \mathrm{argmin}\, \mathrm{CV}^{**}(h)$. Of course, $\hat{h}^*_1$ depends on the simulated data W^*_j, and $\hat{h}^{**}_2$ depends additionally on the values of W^{**}_j. This relationship can be removed by averaging over a large number of versions of CV^* and CV^{**}, at (1.9) and (1.10), for different simulated sequences (U^*_k, U^{**}_k), thus obtaining approximations to the quantities,

$$\mathrm{CV}_1 = E(\mathrm{CV}^* \,|\, \mathcal{D}) \,, \quad \mathrm{CV}_2 = E(\mathrm{CV}^{**} \,|\, \mathcal{D}) \,, \tag{1.11}$$

where $\mathcal{D} = \{(W_1, Y_1), \ldots, (W_n, Y_n)\}$ denotes the dataset. We then define, for $j = 0, 1, 2$,

$$\hat{h}_j = \mathrm{argmin}\, \mathrm{CV}_j(h) \,. \tag{1.12}$$

We might alternatively put $\hat{h}_1 = E(\hat{h}^*_1 \,|\, \mathcal{D})$ and $\hat{h}_2 = E(\hat{h}^{**}_2 \,|\, \mathcal{D})$. These definitions lead to less computer-intensive bandwidth selectors, easier to use in a simulation study.

The relationship between $\hat{h}_0$ and $\hat{h}_1$ is similar to that between $\hat{h}_1$ and $\hat{h}_2$, and so back-extrapolation, in the SIMEX fashion, can be used to produce an approximation to $\hat{h}_0$. For example, linear back-extrapolation from the pair $(\log \hat{h}_1, \log \hat{h}_2)$ might be employed. This suggests taking the final bandwidth to be $\tilde{h}_0 = \hat{h}_1^2/\hat{h}_2$. Justification in the case of small error variance is given in section 4.4.

A more general, pragmatic approach would suggest using a "composite bandwidth,"

$$\tilde{h} = \hat{h}_1^{r_1}\, \hat{h}_2^{r_2}\,, \tag{1.13}$$

where r_1 and r_2 are chosen using experience gained in numerical studies, and subject to the constraint $r_1 + r_2 = 1$. We shall show in section 4.3 that for each such choice of r_1 and r_2, and under general constraints, the composite bandwidth is of the same size as the asymptotically optimal bandwidth, h_0 say, that minimizes asymptotic mean integrated squared error of $\hat{g}$. In order to be itself asymptotically optimal, the composite bandwidth needs adjustment only by a constant factor.

2. OVERVIEW OF THEORETICAL PROPERTIES

2.1. Relationship between cross-validation and weighted integrated squared error. We shall show that the cross-validation criteria CV_0, CV_1 and CV_2, defined at (1.6) and (1.11), can be viewed as having been constructed with the aim of finding an empirical approximation to the bandwidths that minimize certain weighted forms of asymptotic mean integrated squared error (AMISE). The weights involve the function p, which appears in each of (1.6), (1.9) and (1.10). They also incorporate, in the three respective cases, the densities f_X, f_W and f_{W^*} of X, W and W^*.

Specifically, $\hat{h}_0$, $\hat{h}_1$ and $\hat{h}_2$, defined at (1.12), can be viewed as approximations to the bandwidths h_0, h_1 and h_2 that respectively minimize

$$\text{AMISE}(\hat{g} - g) = \int \text{AMSE}\{\hat{g}(x) - g(x)\}\, f_X(x)\, p(x)\, dx\,, \tag{2.1}$$

$$\text{AMISE}(\hat{g}_1 - g_1) = \int \text{AMSE}\{\hat{g}_1^*(x) - g_1(x)\}\, f_W(x)\, p(x)\, dx\,, \tag{2.2}$$

$$\text{AMISE}(\hat{g}_2 - g_2) = \int \text{AMSE}\{\hat{g}_2^{**}(x) - g_2(x)\}\, f_{W^*}(x)\, p(x)\, dx\,. \tag{2.3}$$

Here the functions g, g_1 and g_2 are as in (1.1) and (1.7), the estimators $\hat{g}$, $\hat{g}_1$ and $\hat{g}_2$ are as in (1.4) and (1.8), and AMSE denotes asymptotic mean squared error. We shall argue in section 4.3 that $\hat{h}_j/h_j \to 1$ in probability as $n \to \infty$; see (2.12).

2.2. Mean-squared error formulae. The asymptotic bias and variance of $\hat{g}(x)$ have been derived by Fan and Truong (1993) and are given by,

$$\text{asymp-bias}\{\hat{g}(x)\} = \tfrac{1}{2}\, h^2\, \kappa_2\, q(x)\,, \text{asymp-var}\{\hat{g}(x)\} = \lambda(h)\, (nh)^{-1}\, \sigma_V^2 / f_X(x)\,, \tag{2.4}$$

respectively, where $\sigma_V^2 = \text{var}\, V$, $\kappa_2 = \int x^2\, K(x)\, dx$,

$$q = g'' + 2\, f_X'\, g'\, f_X^{-1}\,, \lambda(h) = \int K_U(v)^2\, dv = \frac{1}{2\pi} \int \left|K^{\text{Ft}}(t)\right|^2 \left|f_U^{\text{Ft}}(t/h)\right|^{-2} dt\,. \tag{2.5}$$

Of course, $\lambda(h)$ is a function of h alone, and diverges to infinity as $h \to 0$. The rate of divergence becomes faster as the distribution of U becomes smoother, or equivalently, as the distribution of U changes so that f_U^{Ft} decreases more rapidly to zero in the tails. For example, if

$$|f_U^{\text{Ft}}(t)| \sim \text{const.}\, |t|^{-\alpha} \tag{2.6}$$

as $|t|$ increases, where $\alpha > 0$ is a measure of the smoothness of the distribution of U, then

$$\lambda(h) \sim \text{const.}\, h^{-2\alpha} \tag{2.7}$$

as $h \to 0$, for a different value of the constant.

Results (2.4) and (2.1) imply that, in the case $j = 0$,

$$\text{AMISE}(\hat{g}_j - g_j) = \frac{\lambda(h)}{nh}\, A + h^4\, B_j\,, \tag{2.8}$$

where we define $g_0 = g$, $\hat{g}_0 = \hat{g}$, $q_0 = q$ (see (2.5)),

$$A = \sigma_V^2 \int p\,, \quad B_j = \tfrac{1}{4}\,\kappa_2^2 \int f_{W^{(j)}}\, p\, q_j^2\,, \tag{2.9}$$

and $W^{(j)}$ denotes X, W, W^* in the cases $j = 0, 1, 2$, respectively. Similarly it can be shown that (2.8) holds for $j = 1, 2$, where in (2.9) we take,

$$q_1 = g_1'' + 2\, f_W'\, g_1'\, f_W^{-1}\,, \quad q_2 = g_2'' + 2\, f_{W^*}'\, g_2'\, f_{W^*}^{-1}\,. \tag{2.10}$$

2.3. Properties of cross-validation bandwidths. Theorem 3.2 will give explicit conditions under which, uniformly in a range of values of h,

$$\text{CV}_j(h) = \{1 + o_p(1)\}\, \text{AMISE}(\hat{g}_j - g_j) + T\,, \tag{2.11}$$

where $T = T(n)$ denotes a quantity that does not depend on h. In a wide variety of settings, for example those where (2.6), and hence (2.7), hold, (2.11) implies that

$$\hat{h}_j / h_j \to 1 \tag{2.12}$$

in probability, where h_0, h_1 and h_2 denote the bandwidths that minimize $\text{AMISE}(\hat{g} - g)$, $\text{AMISE}(\hat{g}_1 - g_1)$ and $\text{AMISE}(\hat{g}_2 - g_2)$, respectively. (See Theorem 3.2 for details.)

To appreciate the implications of (2.12), let us assume for simplicity that the characteristic function f_U^{Ft} satisfies the polynomial decay condition (2.6).

(Other, more general constraints also lead to the conclusions we shall draw below.) Then (2.7) holds, and so (2.8) entails:

$$\mathrm{AMISE}(\hat{g}_j - g_j) \sim \left(nh^{2\alpha+1}\right)^{-1} C + h^4 B_j \,, \tag{2.13}$$

where C and, shortly, D_j will denote fixed positive constants. Results (2.12) and (2.13) imply that $\hat{h}_j$, and the bandwidth h_j that minimizes $\mathrm{AMISE}(\hat{g}_j - g_j)$, satisfy, for $j = 0, 1, 2$,

$$\hat{h}_j \sim_p h_j \sim D_j \, n^{-1/(2\alpha+5)} \,. \tag{2.14}$$

Property (2.14) implies that, no matter whether $j = 0$, 1 or 2, the bandwidth $\hat{h}_j$ that minimizes CV_j is of size $n^{-1/(2\alpha+5)}$, and that the composite bandwidth $\tilde{h}$, defined at (1.13), satisfies

$$\tilde{h} \sim_p D \, n^{-1/(2\alpha+5)} \,,$$

where $D = D_1^{r_1} \, D_2^{r_2}$, with r_1 and r_2 as in (1.13) and satisfying $r_1 + r_2 = 1$. Therefore, no matter what values of r_1 and r_2 we choose, the composite bandwidth is of the same order as the asymptotically optimal bandwidth, h_0, that minimizes $\mathrm{AMISE}(\hat{g}_0 - g_0)$.

2.4. Optimality of linear back-extrapolation in low-noise case. For the sake of simplicity we shall assume again that the characteristic function f_U^{Ft} satisfies (2.6). In this setting (2.13) holds, and the constant C there depends only on σ_V^2, on the function p and on the constant in (2.6). However, B_j depends on g_j, p and, in the respective cases $j = 0$, 1 and 2, on f_X, f_W and f_{W^*}.

In view of (2.12) and (2.13), the bandwidth h_j that minimizes $\mathrm{AMISE}(\hat{g}_j - g_j)$ satisfies,

$$\hat{h}_j \sim_p h_j \sim \left\{ \frac{(2\alpha + 1)\, C}{4\, B_j \, n} \right\}^{1/(2\alpha+5)} \tag{2.15}$$

as $n \to \infty$. Suppose we can write $U = \sigma_U T$, where the distribution of the random variable T has zero mean and unit variance, and σ_U^2 is fixed but small. (This property characterizes the "low-noise case" referred to in the heading of this section.) Then, under the regularity conditions given in Theorem 3.3, it can be shown that,

$$B_j = B_0 \left\{ 1 + j\, Q\, \sigma_U^2 + \psi_{1j}\big(\sigma_U^2\big) \right\}, \tag{2.16}$$

where the constant Q depends only on f_T, f_X and g, and in particular does not depend on σ_U^2; and, here and below, ψ (with or without subscripts) denotes a function determined by f_T, f_X and g, and satisfying $\psi(u) = o(u)$ as $u \to 0$. (A longer argument will show that, if the functions f_T, f_X and g are sufficiently smooth, $\psi(u) = O(u^2)$ as $u \to 0$.)

Properties (2.15) and (2.16) imply that,

$$\log \hat{h}_j = d \log(A/nB_0) - j\, Q\, \sigma_U^2 + \psi_{2j}\big(\sigma_U^2\big) + o_p(1)$$

for $j = 0, 1, 2$. It follows from this result, and from (2.15) for $j = 0$, that if we compute $\log \tilde{h}_0$ by linear back-extrapolation from $\log \hat{h}_1$ and $\log \hat{h}_2$ then $\tilde{h}_0$ satisfies,

$$\tilde{h}_0 \sim_p \hat{h}_0 \left\{ 1 + \psi\big(\sigma_U^2\big) \right\} \sim_p h_0 \left\{ 1 + \psi\big(\sigma_U^2\big) \right\}. \tag{2.17}$$

This result justifies linear back-extrapolation when the variance of U is small. In particular, comparing (2.15) and (2.17) we see that, if σ_U^2 is sufficiently small, although fixed, then the linearly back-extrapolated bandwidth estimator $\tilde{h}_0$ is, with probability converging to 1 as $n \to \infty$, closer to the desired bandwidth h_0 than is either $\hat{h}_1$ or $\hat{h}_2$.

3. MAIN THEOREMS

We assume throughout that the densities f_X, f_W and f_{W^*} of X, W and W^*, respectively, are well defined. First we state a result due to Fan and Truong (1993), giving the asymptotic bias and variance formulae in (2.4).

Theorem 3.1. *Assume that* (1.2) *and* (1.5) *hold, that* $\operatorname{var} V < \infty$, *that* f_W *and* g *are uniformly bounded, that* f_X *and* g *have two continuous derivatives in a neighborhood of* x, *that* $f_X(x) > 0$, *that* $h = h(n) \to 0$ *and that* $\lambda(h) = o(nh)$. *Then,* $\hat{g}(x) - g(x)$ *is asymptotically normally distributed with mean* $\frac{1}{2} h^2 \kappa_2 q(x) + o(h^2)$ *and variance* $\lambda(h)(nh)^{-1} \sigma_V^2 f_X(x)^{-1} + o\{\lambda(h)(nh)^{-1}\}$, *where* $q(x)$ *and* $\lambda(h)$ *are given by* (2.5).

Next we state a result which, under explicit regularity conditions, ensures (2.11) and (2.12). For this we assume that,

the function p is nonnegative and bounded, and its support equals a compact interval, $\mathcal{S}_p$ say. (3.1)

Taking Z to denote X, W or W^*, we ask that,

(a) f_Z has two bounded and continuous derivatives on the real line, and is bounded away from zero on $\mathcal{S}_p$; (b) $\sup f_U < \infty$ and $E|U|^\epsilon < \infty$ for some $\epsilon > 0$; (c) $E|X|^\epsilon < \infty$ for some $\epsilon > 0$; and (d) the function λ, defined at (2.5), satisfies $C_1 h^{-C_2} \le \lambda(h) \le C_3 h^{-C_4}$ for constants $0 < C_1 < C_3 < \infty$ and $0 < C_2 < C_4 < \infty$, and all $h \in (0,1]$. (3.2)

Part (a) of (3.2) is a standard assumption about the design density in a nonparametric regression problem. Indeed, in the parts of Theorem 3.2 below that correspond to $j = 0$, 1 and 2, the appropriate design density is f_X, f_W and f_{W^*}, respectively, and these are the three choices possible in (a). Parts (b) and (c) of (3.2) are mild assumptions about the distributions of U and X, respectively, and part (d) asks, in effect, that the characteristic function f_U^{Ft} not decrease to zero any faster than polynomially as $|t| \to \infty$. In particular, (3.2)(d) is implied by (2.7), which in turn follows from (2.6).

Given $\epsilon \in (0,1)$, let $\mathcal{H} = \mathcal{H}(n)$ denote a set of values of $h > 0$ with the property that, as $n \to \infty$,

(a) $\sup_{h\in\mathcal{H}} h = O(n^{-\epsilon})$, (b) $\sup_{h\in\mathcal{H}} \lambda(h)/h = O(n^{1-\epsilon})$, and (c) the bandwidths h_j that minimize $\mathrm{AMISE}(\hat{g}_j - g_j)$, are, for all sufficiently large n and for $j = 0,1,2$, in $\mathcal{H}$. (3.3)

For example, if the characteristic function f_U^{Ft} satisfies (2.6), for a constant $\alpha > 0$, then (2.7) holds and from that property it follows that a sufficient condition for part (b) of (3.3) is: $\inf_{h\in\mathcal{H}} h \geq \text{const.}\ n^{-(1-\epsilon)/(2\alpha+1)}$. Moreover, if $\kappa_2 \neq 0$ then h_0, h_1 and h_2 are each asymptotic to constant multiples of $n^{-1/(2\alpha+5)}$. Therefore, if we define $\mathcal{H}$ to be the set of values h for which $n^{-(1-\epsilon_1)/(2\alpha+1)} \leq h \leq n^{-\epsilon_2}$, where $0 < \epsilon_1 < 4/(2\alpha+5)$ and $0 < \epsilon_2 < 1/(2\alpha+5)$, then each of parts (a)–(c) of (3.3) holds, with $\epsilon \in (0, \min(\epsilon_1, \epsilon_2))$.

We assume too that,

the function g is bounded and has two continuous derivatives on the real line, $\sigma_V \neq 0$ and $E(|V|^C) < \infty$ for all $C > 0$. (3.4)

Let $\hat{h}_j$ denote the bandwidth that minimizes CV_j, and let h_j be the bandwidth that minimizes $\mathrm{AMISE}(\hat{g}_j - g_j)$, for $j = 0,1,2$. Formal definitions of CV_1 and CV_2 are given at (1.11). In practice, however, CV_1 and CV_2 would be computed as,

$$\mathrm{CV}_1 = \frac{1}{B}\sum_{b=1}^{B} \mathrm{CV}_b^*\,, \quad \mathrm{CV}_2 = \frac{1}{B}\sum_{b=1}^{B} \mathrm{CV}_b^{**}\,, \tag{3.5}$$

where CV_b^*, for $1 \leq b \leq B$, and CV_b^{**}, $1 \leq b \leq B$, denote versions of CV^* and CV^{**}, respectively, computed using independent values of U_j^* (in the definition $W_j^* = W_j + U_j^*$) and U_j^{**} (in $W_j^{**} = W_j + U_j^* + U_j^{**}$). In particular, conditional on the data $\mathcal{D} = \{(W_1, Y_1), \ldots, (W_n, Y_n)\}$, the variables $\mathrm{CV}_1^*, \ldots, \mathrm{CV}_B^*$ are independent and identically distributed, as too are $\mathrm{CV}_1^{**}, \ldots, \mathrm{CV}_B^{**}$.

In numerical practice, particularly when undertaking a simulation study, it is common to take B to be no more than polynomially large as a function of n. Formally, as $n \to \infty$,

$$B = B(n) \to \infty \quad \text{and} \quad B = O\big(n^C\big) \quad \text{for some} \quad C > 0\,. \tag{3.6}$$

This ensures that certain very pathological cases, the probabilities of which are exponentially small as functions of n, have negligibly small chance of occurring among any of the summands in the definitions at (3.5); and thereby guarantees that the denominators in the definitions of $\hat{g}^*_{1,-j}$ and $\hat{g}^{**}_{2,-j}$, in (1.8) and (1.9) respectively, are not too close to zero. For this reason we shall assume that CV_1 and CV_2 are as defined at (3.5), with B satisfying (3.6). An alternative approach would be to use a ridge-based construction to ward off problems with the denominators (see section 3.1).

Theorem 3.2. *Assume that* (1.2), (1.5), (3.1) *and* (3.3)–(3.6) *hold. If* (3.2) *is true for $Z = X$ [respectively, for $Z = W$ or W^*], then in the case $j = 0$ [respectively, in the cases $j = 1$ or $j = 2$] equation* (2.11) *holds uniformly in $h \in \mathcal{H}$, and* (2.12) *holds.*

Finally we give a formal statement of (2.16), for which we assume that:

$U = \sigma_U\, T$, where the distribution of the random variable T is held fixed, and $\sigma_U > 0$ is permitted to decrease to zero; $E(T^2) < \infty$ and $E(T) = 0$; the support, $\mathcal{S}_p$, of the bounded, nonnegative function p is a compact set; f_X has three continuous derivatives, and g has four continuous derivatives, on an open set containing $\mathcal{S}_p$; $f_X > 0$ on $\mathcal{S}_p$. (3.7)

Theorem 3.3. *If* (3.7) *holds then so too does* (2.16), *which describes behavior of B_j as $\sigma_U \to 0$.*

APPENDIX – DERIVATIONS OF THE RESULTS

A.1. Proof of Theorem 3.2.

We shall derive only the part corresponding to $j = 2$; the parts relating to $j = 0$ or 1 may be proved similarly. To simplify notation we shall give initially an expansion of $\hat{g}_2^{**}$, rather than $\hat{g}_{2,-j}^{**}$, discussing later the case of $\hat{g}_{2,-j}^{**}$.

Step (I): Approximations to $\hat{f}_{W^*}$ and $\hat{a}_2$. Define

$$\hat{f}_{W^*}(x) = \frac{1}{nh} \sum_{j=1}^{n} K_U\left(\frac{x - W_j^{**}}{h}\right), \quad \hat{a}_2(x) = \frac{1}{nh} \sum_{j=1}^{n} Y_j \, K_U\left(\frac{x - W_j^{**}}{h}\right),$$

$a_2 = f_{W^*} \, g_2$, $\Delta_{a_2} = \hat{a}_2 - a_2$ and $\Delta_{f_{W^*}} = \hat{f}_{W^*} - f_{W^*}$. In this notation, $g_2 = a_2 / f_{W^*}$ and $\hat{g}_2^{**} = \hat{a}_2 / \hat{f}_{W^*}$, and we have

$$\hat{g}_2^{**} - g_2 = \frac{\Delta_{a_2}}{f_{W^*}} - \frac{a_2 \, \Delta_{f_{W^*}}}{f_{W^*}^2} - \frac{\Delta_{a_2} \, \Delta_{f_{W^*}}}{f_{W^*}^2} + \frac{(a_2 + \Delta_{a_2}) \, \Delta_{f_{W^*}}^2}{(f_{W^*} + \Delta_{f_{W^*}}) \, f_{W^*}^2} \, . \tag{A.1}$$

Step (II): Moment bounds for $\Delta_{f_{W^*}}$ and Δ_{a_2}. Using Rosenthal's inequality it can be proved that, for each $r \geq 1$,

$$\begin{aligned} E\big\{|\Delta_{f_{W^*}} - E(\Delta_{f_{W^*}} \,|\, \mathcal{D})|^{2r} \,\big|\, \mathcal{D}\big\} &\leq \text{const.} \, \big(\hat{\sigma}_2^r + \hat{\sigma}_{2r}\big) \, , \\ E\big\{|\Delta_{a_2} - E(\Delta_{a_2} \,|\, \mathcal{D})|^{2r} \,\big|\, \mathcal{D}\big\} &\leq \text{const.} \, \big(\hat{\tau}_2^r + \hat{\tau}_{2r}\big) \, , \end{aligned} \tag{A.2}$$

where the constants depend only on r,

$$\hat{\sigma}_s = \frac{1}{(nh)^s} \sum_{j=1}^{n} E\left\{\left|K_U\left(\frac{x - W_j^{**}}{h}\right)\right|^s \;\middle|\; \mathcal{D}\right\},$$

$$\hat{\tau}_s = \frac{1}{(nh)^s} \sum_{j=1}^{n} E\left\{\left|Y_j \, K_U\left(\frac{x - W_j^{**}}{h}\right)\right|^s \;\middle|\; \mathcal{D}\right\},$$

and $\mathcal{D}$ denotes the set of data (W_j, Y_j), for $1 \leq j \leq n$. Let T have the distribution of $U_1^* + U_1^{**}$. Since, by (3.2)(b), f_U is bounded, then so too is the density f_T of T, and so,

$$\begin{aligned} E\left\{\left|K_U\left(\frac{x - W_j^{**}}{h}\right)\right|^s \;\middle|\; \mathcal{D}\right\} &= h \int |K_U(u)|^s \, f_T(x - W_j - hu) \, du \\ &\leq h \, (\sup f_T) \int |K_U(u)|^s \, du \, , \end{aligned} \tag{A.3}$$

where, here and below, unqualified integrals are over the whole real line. If $s \geq 2$,

$$(2\pi)^2 \int |K_U(u)|^s \, du \leq \Big\{ \sup_u |K_U(u)| \Big\}^{s-2} \int \bigg| \int e^{-itu} \, K^{\mathrm{Ft}}(t) \, f_U^{\mathrm{Ft}}(t/h)^{-1} \, dt \bigg|^2 du \, . \quad (A.4)$$

By Parseval's identity and the definition of $\lambda(h)$, at (2.5),

$$\int \bigg| \int e^{-itu} \, K^{\mathrm{Ft}}(t) \, f_U^{\mathrm{Ft}}(t/h)^{-1} \, dt \bigg|^2 du = (2\pi)^2 \, \lambda(h) \, , \qquad (A.5)$$

and, since K^{Ft} vanishes outside a compact set,

$$\begin{aligned} (2\pi)^2 \Big\{ \sup_u |K_U(u)| \Big\}^2 &\leq \Big\{ \int \big| K^{\mathrm{Ft}}(t) \, f_U^{\mathrm{Ft}}(t/h)^{-1} \big| \, dt \Big\}^2 \\ &\leq \text{const.} \int \big| K^{\mathrm{Ft}}(t) \, f_U^{\mathrm{Ft}}(t/h)^{-1} \big|^2 \, dt \\ &\leq \text{const.} \ \lambda(h) \, , \qquad (A.6) \end{aligned}$$

where the constant depends only on K. Together, (A.4)–(A.6) imply that,

$$\int |K_U(u)|^s \, du \leq \text{const.} \ \lambda(h)^{s/2} \, ,$$

where the constant depends only on K and s. From this result, (A.3) and the definition of $\hat{\sigma}_s$ we deduce that $\hat{\sigma}_s \leq \text{const.} \ (nh)^{-(s-1)} \, \lambda(h)^{s/2}$.

A similar bound can be derived for $\hat{\tau}_s$, on noting that, for $r_1, r_2 > 1$ such that $r_1^{-1} + r_2^{-1} = 1$,

$$\begin{aligned} 2^{-(s-1)} \, (nh)^s \, \hat{\tau}_s &\leq (nh)^s \, (\sup |g|)^s \, \hat{\sigma}_s \\ &\quad + \sum_{j=1}^n \big\{ E\big(|V_j|^{r_1 s} \mid \mathcal{D}\big) \big\}^{1/r_1} \Bigg[E\Bigg\{ \bigg| K_U\Big(\frac{x - W_j^{**}}{h}\Big) \bigg|^{r_2 s} \ \Bigg| \ \mathcal{D} \Bigg\} \Bigg]^{1/r_2} \\ &\leq \text{const.} \ nh^{1/r_2} \, \lambda(h)^{s/2} \, (1 + \hat{\xi}_s) \, , \end{aligned}$$

where the constant depends only on K, $r_1 > 1$ and s, and

$$\hat{\xi}_s = \frac{1}{n} \sum_{j=1}^n \big\{ E\big(|V_j|^{r_1 s} \mid \mathcal{D}\big) \big\}^{1/r_1} \, .$$

Since $r_1 > 1$ can be chosen arbitrarily close to 1 then, for each $\delta > 0$,

$$\hat\sigma_s \le \text{const.}\ (nh)^{-(s-1)}\, \lambda(h)^{s/2}\,, \hat\tau_s \le \text{const.}\ (nh)^{-(s-1)}\, h^{-\delta}\, (1+\hat\xi_s)\, \lambda(h)^{s/2}\,, \tag{A.7}$$

where the constants depend only on K, s and δ.

Combining (A.2) and (A.7), and recalling that, by assumption, all moments of the distribution of V are finite, we deduce that, for each $r \ge 1$, each $\delta > 0$, all $h \in \mathcal{H}$ and all real x (appearing in the suppressed arguments of the functions $\Delta_{f_{W^*}}$ and Δ_{a_2}),

$$E\big\{|\Delta_{f_{W^*}} - E(\Delta_{f_{W^*}}\,|\,\mathcal{D})|^{2r}\big\} + E\big\{|\Delta_{a_2} - E(\Delta_{a_2}\,|\,\mathcal{D})|^{2r}\big\} \le \text{const.}\ h^{-\delta}\, \big\{(nh)^{-1}\,\lambda(h)\big\}^r\,, \tag{A.8}$$

where the constants depend only on K, s and δ.

Step (III): Moderate-deviation bounds for $\Delta_{f_{W^*}}$ and Δ_{a_2} on a lattice. Result (3.3) gives,

$$\text{for all sufficiently large } n, \quad \sup_{h\in\mathcal{H}} \lambda(h) \le n\,. \tag{A.9}$$

Properties (3.2)(d) and (A.9) imply that,

$$\text{for all sufficiently large } n, \quad \sup_{h\in\mathcal{H}} h^{-1} \le (n/C_1)^{1/C_2}\,. \tag{A.10}$$

In view of (A.8) and (A.10), and by Markov's inequality, we have for each $\epsilon \in (0,1)$ and each $C > 0$, and for $m = a_2$ or f_{W^*},

$$\sup_{h\in\mathcal{H}}\ \sup_{x\in\mathbb{R}} P\Big[\,\big|\Delta_m(x) - E\{\Delta_m(x)\,|\,\mathcal{D}\}\big| > \big\{\lambda(h)/n^{1-\epsilon}\,h\big\}^{1/2}\Big] = O\big(n^{-C}\big)\,.$$

Therefore, if $\mathcal{H}_n \subseteq \mathcal{H}$ and if $\mathcal{X}_n \subseteq \mathbb{R}$ are sets containing $O(n^{C_1})$ real numbers, for $C_1 > 0$ arbitrary but fixed, we have, for each $C > 0$, and for $m = a_2$ or

f_{W^*},

$$P\Big[\big|\Delta_m(x) - E\{\Delta_m(x) \,|\, \mathcal{D}\}\big| > \big\{\lambda(h)/n^{1-\epsilon}\, h\big\}^{1/2} \text{ for some } (h,x) \in \mathcal{H}_n \times \mathcal{X}_n \Big] = O\big(n^{-C}\big)\,, \tag{A.11}$$

Step (IV): Extension of (*A*.11) *to* h *and* x *in the continuum.* Using (A.6) and the fact that K^{Ft} is compactly supported, we may deduce from (1.3) that, for each real u_1 and u_2,

$$\begin{aligned} |K_U(u_1) - K_U(u_2)| &\le \text{const.}\ |u_1 - u_2| \int \big|K^{\mathrm{Ft}}(t)\, f_U^{\mathrm{Ft}}(t/h)^{-1}\big|\, dt \\ &\le \text{const.}\ |u_1 - u_2|\ \lambda(h)^{1/2}\,. \end{aligned}$$

From this result and (A.9) we deduce that, for all n,

$$\sup_{h \in \mathcal{H}} \ \sup_{-\infty < u_1 < u_2 < \infty} |u_1 - u_2|^{-1}\, |K_U(u_1 \,|\, h) - K_U(u_2 \,|\, h)| \le \text{const.}\, n\,, \tag{A.12}$$

where we have written $K_U(u) = K_U(u \,|\, h)$ to indicate the dependence of $K_U(u)$ on h.

For each real u, and for all $h_1, h_2 \in \mathcal{H}$,

$$2\pi\, \big|K_U(u \,|\, h_1) - K_U(u \,|\, h_2)\big| \le \int \big|K_U^{\mathrm{Ft}}(t)\big|\, \big|f_U^{\mathrm{Ft}}(t/h_1)^{-1} - f_U^{\mathrm{Ft}}(t/h_2)^{-1}\big|\, dt\,; \tag{A.13}$$

$$\big|f_U^{\mathrm{Ft}}(t/h_1)^{-1} - f_U^{\mathrm{Ft}}(t/h_2)^{-1}\big| = \frac{|f_U^{\mathrm{Ft}}(t/h_1) - f_U^{\mathrm{Ft}}(t/h_2)|}{|f_U^{\mathrm{Ft}}(t/h_1)\, f_U^{\mathrm{Ft}}(t/h_2)|}\ ; \tag{A.14}$$

if $\epsilon \in (0,1]$ is as in (3.2)(b) then, for a constant $C(\epsilon) > 0$ depending only on ϵ,

$$\begin{aligned} \big|f_U^{\mathrm{Ft}}(t/h_1) - f_U^{\mathrm{Ft}}(t/h_2)\big| &\le C(\epsilon)\, E|tU|^\epsilon\, \big|h_1^{-1} - h_2^{-1}\big|^\epsilon \\ &\le C(\epsilon)\, E|tU|^\epsilon\, \frac{|h_1 - h_2|^\epsilon}{\{\min(h_1, h_2)\}^\epsilon}\ ; \end{aligned} \tag{A.15}$$

$$\left[\int |t|^\epsilon \left| K_U^{\rm Ft}(t) \right| \left\{ \left| f_U^{\rm Ft}(t/h_1)\, f_U^{\rm Ft}(t/h_2) \right| \right\}^{-1} dt \right]^2$$

$$\leq \left\{ \int \left| K_U^{\rm Ft}(t)\, f_U^{\rm Ft}(t/h_1)^{-1} \right|^2 dt \right\} \left\{ \int_{\mathcal{S}_{K^{\rm Ft}}} |t|^{2\epsilon}\, f_U^{\rm Ft}(t/h_2)^{-2}\, dt \right\}$$

$$\leq 4\pi\, \lambda(h_1) \int_0^{D_1} |t|^{2\epsilon}\, f_U^{\rm Ft}(t/h_2)^{-2}\, dt\,, \qquad (A.16)$$

where $\mathcal{S}_{K^{\rm Ft}}$ denotes the support of $K^{\rm Ft}$ and is contained in a compact interval $[-D_1, D_1]$, say. It can be proved that, if C_2, C_4 are as in (3.2)(d), with $0 < C_2 < C_4 < \infty$, then, for all sufficiently large n,

$$\sup_{h \in \mathcal{H}} \int_0^{D_1} |t|^{2\epsilon}\, f_U^{\rm Ft}(t/h)^{-2}\, dt \leq \text{const.}\, n^{C_4/C_2}\,, \qquad (A.17)$$

where the constant does not depend on n. Combining (A.9), (A.10) and (A.13)–(A.17) we deduce that, for a constant $D_2 \geq 1$,

$$\sup_{h_1, h_2 \in \mathcal{H}, h_1 \neq h_2}\ \sup_{-\infty < u < \infty} |h_1 - h_2|^{-\epsilon} \left| K_U(u \,|\, h_1) - K_U(u \,|\, h_2) \right| \leq \text{const.}\ n^{D_2}\,. \qquad (A.18)$$

Results (A.12) and (A.18) imply that,

$$\left| K_U(u_1 \,|\, h_1) - K_U(u_2 \,|\, h_2) \right| \leq \text{const.}\ n^{D_2} \left(|u_1 - u_2| + |h_1 - h_2|^\epsilon \right). \qquad (A.19)$$

In view of the conditions $E|U|^\epsilon < \infty$ and $E|X|^\epsilon < \infty$ in (3.2)(b) and (3.2)(c), respectively; and noting (A.10); we deduce that, given any compact set $\mathcal{C}$ and any $D_3 > 0$, we may choose $D_4 > 0$ so large that, with probability $1 - O(n^{-D_3})$, $|x - W_j^{**}|/h \leq n^{D_4}$ uniformly in $h \in \mathcal{H}$, $x \in \mathcal{C}$ and $1 \leq j \leq n$; and,

> this result continues to hold if the implicit supremum of $|x - W_j^{**}|/h$ is taken not just over $h \in \mathcal{H}$, $x \in \mathcal{C}$ and $1 \leq j \leq n$, but also over the B sequences $(U_1^*, U_1^{**}), \ldots, (U_n^*, U_n^{**})$.

$(A.20)$

Let $\mathcal{C}$ be the compact set noted in the previous paragraph, and let $\mathcal{X}_n$ denote a lattice in $\mathcal{C}$ of edge width $n^{-\{(1/2)+(2C_2^{-1}+D_2+D_5)/\epsilon\}}$, where $D_5 > 0$ is arbitrarily large but fixed and $\epsilon \in (0,1]$ is as in (3.2)(b). Let $\mathcal{H}_n \subseteq \mathcal{H}$ be such that no point of $\mathcal{H}$ is any further than $n^{-\{(1/2)+(2C_2^{-1}+D_2+D_5)/\epsilon+D_4\}}$ from a point in $\mathcal{H}_n$. Then, $(\#\mathcal{H}_n)+(\#\mathcal{X}_n) = O(n^{(1/2)+(2C_2^{-1}+D_2+D_5)/\epsilon+D_4})$, which is of only polynomial order in n, and so result (A.11) holds for this choice of $(\mathcal{H}_n, \mathcal{X}_n)$.

Given $(h,x) \in \mathcal{H} \times \mathcal{C}$, choose $(h',x') \in \mathcal{H}_n \times \mathcal{X}_n$ such that x' is as close as possible to x, and h' is as close as possible to h, with ties broken arbitrarily. Let $\mathcal{E}$ denote the event, having probability $1 - O(n^{-D_3})$, that $|x - W_j^{**}|/h \le n^{D_4}$ for all $h \in \mathcal{H}$, all $x \in \mathcal{C}$, all $1 \le j \le n$ and all B of the sequences $(U_1^*, U_1^{**}), \ldots, (U_n^*, U_n^{**})$. (See (A.20).) If $\mathcal{E}$ holds then, using (A.10) and noting the definitions of $\mathcal{H}_n$ and $\mathcal{X}_n$,

$$\left|\frac{x - W_j^{**}}{h} - \frac{x' - W_j^{**}}{h'}\right| \le \left|\frac{x - W_j^{**}}{h}\right| \left|1 - \frac{h}{h'}\right| + \frac{|x - x'|}{h'}$$
$$\le \text{const.} \left(n^{D_4}\, n^{1/C_2} \left|h - h'\right| + n^{1/C_2} \left|x - x'\right|\right)$$
$$\le \text{const.}\ n^{-(1/C_2)-D_2-D_5}\,.$$

Therefore, using (A.10) and (A.19),

$$\frac{1}{h}\left|K_U\left(\frac{x - W_j^{**}}{h}\right) - K_U\left(\frac{x' - W_j^{**}}{h'}\right)\right| \le \text{const.}\ n^{-D_5}\,. \qquad (A.21)$$

Using (A.6) and (A.10) we deduce that,

$$\left|h^{-1} - (h')^{-1}\right| \left\{\left|K_U\left(\frac{x - W_j^{**}}{h}\right)\right| + \left|K_U\left(\frac{x' - W_j^{**}}{h'}\right)\right|\right\}$$
$$\le \text{const.}\ \frac{|h - h'|}{\{\min(h,h')\}^2} \int \left|K^{\mathrm{Ft}}(t)\right| \left\{\left|f^{\mathrm{Ft}}(t/h)\right|^{-1} + \left|f^{\mathrm{Ft}}(t/h')\right|^{-1}\right\} dt$$
$$\le \text{const.}\ \frac{|h - h'|}{\{\min(h,h')\}^2} \left\{\lambda(h) + \lambda(h')\right\}^{1/2}$$
$$\le \text{const.}\ n^{-\{(1/2)+2C_2^{-1}+D_2+D_5\}}\, n^{2/C_2}\, n^{1/2} \le \text{const.}\ n^{-D_5}\,. \qquad (A.22)$$

Properties (A.21) and (A.22) imply that if the event $\mathcal{E}$ holds then, uniformly in $(h, x) \in \mathcal{H} \times \mathcal{C}$, in $1 \le j \le n$ and in the B sequences of values (U_k^*, U_k^{**}),

$$\left| \frac{1}{h} K_U\left(\frac{x - W_j^{**}}{h}\right) - \frac{1}{h'} K_U\left(\frac{x' - W_j^{**}}{h'}\right) \right| \le \text{const.}\ n^{-D_5}.$$

Since $D_5 > 0$ can be chosen arbitrarily large then approximation on a lattice permits us to extend (A.11) from $(h, x) \in \mathcal{H}_n \times \mathcal{X}_n$ to $(h, x) \in \mathcal{H} \times \mathcal{C}$: for each $C, \epsilon > 0$, and for $m = a_2$ or f_{W^*},

$$P\Big[\big|\Delta_m(x) - E\{\Delta_m(x) \,|\, \mathcal{D}\}\big| > \big\{\lambda(h)/n^{1-\epsilon}\, h\big\}^{1/2} \text{ for some } (h, x) \in \mathcal{H} \times \mathcal{C}\Big] = O\big(n^{-C}\big). \qquad (A.23)$$

Moreover, in view of (A.20),

> both results hold if the implicit supremum is taken not just over $h \in \mathcal{H}$ and $x \in \mathcal{C}$, but also over the different values that arise if we construct the quantities for each of $B = O(n^D)$ different sequences $(U_1^*, U_1^{**}), \ldots, (U_n^*, U_n^{**})$, for any fixed $D > 0$.
>
> $(A.24)$

Step (V): Linear approximations to $\hat{g}_2^{**} - g_2$ and $\hat{g}_{2,-j}^{**} - g_2$. We may assume, without loss of generality, that $\int K = 1$, i.e. $K^{\text{Ft}}(0) = 1$. Now,

$$\frac{1}{h} E\left\{K_U\left(\frac{x - T}{h}\right)\right\} = \int K(u)\, f_U(x - hu)\, du \equiv \kappa(x),$$

say. Put

$$\mu_{f_{W^*}}(x) = E\{\kappa(x - W)\} - f_{W^*}(x), \quad \mu_{a_2}(x) = E\{Y\, \kappa(x - W)\} - a_2(x).$$

In this notation,

$$E\{\Delta_{f_{W^*}}(x) \,|\, \mathcal{D}\} = \mu_{f_{W^*}}(x) + R_1(x), \quad E\{\Delta_{a_2}(x) \,|\, \mathcal{D}\} = \mu_{a_2}(x) + R_2(x), \qquad (A.25)$$

where

$$R_1(x) = \frac{1}{n} \sum_{j=1}^{n} \Big[\kappa(x - W_j) - E\{\kappa(x - W)\}\Big] \,,$$

$$R_2(x) = \frac{1}{n} \sum_{j=1}^{n} \Big[Y_j \, \kappa(x - W_j) - E\{Y \, \kappa(x - W)\}\Big] \,.$$

Now, $E\{R_j(x)\} = 0$ and $R_j(x) = O_p(n^{-1/2})$, for each x. The lattice approximation argument leading to (A.23)–(A.24) can be used to show that, for $j = 1, 2$, for each $C, \epsilon > 0$, and for each compact set $\mathcal{C}$,

$$P\bigg\{ \sup_{h \in \mathcal{H}} \sup_{x \in \mathcal{C}} |R_j(x)| > n^{\epsilon - (1/2)} \bigg\} = O\big(n^{-C}\big) \,. \tag{A.26}$$

Taylor expansions show that, uniformly in $h \in \mathcal{H}$ and $x \in \mathcal{C}$,

$$\mu_{f_{W^*}}(x) = \tfrac{1}{2} \, \kappa_2 \, h^2 \, f''_{W^*}(x) + o\big(h^2\big) \,, \quad \mu_{a_2}(x) = \tfrac{1}{2} \, \kappa_2 \, h^2 \, a''_2(x) + o\big(h^2\big) \,. \tag{A.27}$$

Combining (A.23) and (A.25)–(A.27), and noting that $\lambda(h)$ is bounded above zero uniformly in $h \in \mathcal{H}$, we deduce that, for $D > 0$ sufficiently large, and for each $C, \epsilon > 0$,

$$P\Big[|\Delta_{f_{W^*}}(x)| + |\Delta_{a_2}(x)| > \big\{\lambda(h)/n^{1-\epsilon} \, h\big\}^{1/2} + D \, h^2 \text{ for some } (h, x) \in \mathcal{H} \times \mathcal{C}\Big] = O\big(n^{-C}\big) \,. \tag{A.28}$$

Results (A.1) and (A.28) imply that,

$$\hat{g}_2^{**}(x) - g_2(x) = \frac{\Delta_{a_2}(x)}{f_{W^*}(x)} - \frac{a_2(x) \, \Delta_{f_{W^*}}(x)}{f_{W^*}(x)^2} + R_3(x) \,, \tag{A.29}$$

where, for $D > 0$ sufficiently large, for each $C, \epsilon > 0$, and for each compact interval $\mathcal{C}$ on which f_{W^*} is bounded away from zero,

$$P\Big[|R_3(x)| > \big\{\lambda(h)/n^{1-\epsilon} \, h\big\} + D \, h^4 \quad \text{for some} \quad (h, x) \in \mathcal{H} \times \mathcal{C}\Big] = O\big(n^{-C}\big) \,. \tag{A.30}$$

Define

$$\hat{f}_{W^*,-j}(x) = \frac{1}{nh} \sum_{k\,:\,k \neq j} K_U\Big(\frac{x - W_j^{**}}{h}\Big),$$

$$\hat{a}_{2,-j}(x) = \frac{1}{nh} \sum_{k\,:\,k \neq j} Y_j \, K_U\Big(\frac{x - W_j^{**}}{h}\Big),$$

$\Delta_{a_2,-j} = \hat{a}_{2,-j} - a_2$ and $\Delta_{f_{W^*},-j} = \hat{f}_{W^*,-j} - f_{W^*}$. An expansion analogous to (A.29), and with a bound analogous to (A.30) on the remainder, can be given for $g^{**}_{2,-j}(x) - g_2(x)$:

$$\hat{g}^{**}_{2,-j}(x) - g_2(x) = \frac{\Delta_{a_2,-j}(x)}{f_{W^*}(x)} - \frac{a_2(x)\, \Delta_{f_{W^*},-j}(x)}{f_{W^*}(x)^2} + R_{3,j}(x)\,, \qquad (A.31)$$

$$P\Big[\, |R_{3,j}(x)| > \big\{\lambda(h)/n^{1-\epsilon}\, h\big\} + D\, h^4 \quad \text{for some} \quad 1 \le j \le n \quad \text{and some} \quad (h,x) \in \mathcal{H} \times \mathcal{C} \Big] = O\big(n^{-C}\big) \quad (A.32)$$

for some $D > 0$ and all $C > 0$. (Although $\hat{a}_{2,-j}$ and $\hat{f}_{W^*,-j}$ are computed from samples of size $n-1$, rather than n, we shall construct them by dividing by n (in analogues of the definitions in Step (I)), rather than by dividing by $n-1$.) More simply than (A.32), if we define $R_4(x)$ to equal the maximum of $|\Delta_{a_2,-j}|$, $|\Delta_{f_{W^*},-j}|$, $|\Delta_{a_2}|$ and $|\Delta_{f_{W^*}}|$, where j runs over $1 \le j \le n$, then,

$$P\Big[R_4(x) > \big\{\lambda(h)/n^{1-\epsilon}\, h\big\} + D\, h^4 \quad \text{for some} \quad (h,x) \in \mathcal{H} \times \mathcal{C} \Big] = O\big(n^{-C}\big)\,. \qquad (A.33)$$

Moreover, analogues of (A.24) apply in the cases of (A.28), (A.30), (A.32) and (A.33); for example, (A.24) applies without change to (A.32) and (A.33).

Step (VI): Approximations to the cross-validation criterion. Define $V_j^* = Y_j - g_2(W_j^*) = g(X_j) + V_j - g_2(W_j^*)$, and put

$$\mathrm{CV}_1^{**} = \frac{1}{n} \sum_{j=1}^{n} \bigg\{ V_j^* - \frac{\Delta_{a_2,-j}(W_j^*)}{f_{W^*}(W_j^*)} + \frac{a_2(W_j^*)\, \Delta_{f_{W^*},-j}(W_j^*)}{f_{W^*}(W_j^*)^2} \bigg\}^2 p(W_j^*)\,,$$

$$\mathrm{CV}_2^{**} = \frac{1}{n} \sum_{j=1}^{n} \left\{ \frac{\Delta_{a_2,-j}(W_j^*)}{f_{W^*}(W_j^*)} - \frac{a_2(W_j^*)\, \Delta_{f_{W^*},-j}(W_j^*)}{f_{W^*}(W_j^*)^2} \right\}^2 p(W_j^*)\,,$$

$$\mathrm{CV}_3^{**} = \frac{1}{n} \sum_{j=1}^{n} \left\{ \frac{\Delta_{a_2}(W_j^*)}{f_{W^*}(W_j^*)} - \frac{a_2(W_j^*)\, \Delta_{f_{W^*}}(W_j^*)}{f_{W^*}(W_j^*)^2} \right\}^2 p(W_j^*)\,, \qquad (A.34)$$

$$T_1 = \frac{1}{n} \sum_{j=1}^{n} (V_j^*)^2\, p(W_j^*)\,, \qquad (A.35)$$

$$T_2 = \frac{1}{n} \sum_{j=1}^{n} V_j^* \left\{ \frac{\Delta_{a_2,-j}(W_j^*)}{f_{W^*}(W_j^*)} - \frac{a_2(W_j^*)\, \Delta_{f_{W^*},-j}(W_j^*)}{f_{W^*}(W_j^*)^2} \right\} p(W_j^*)\,. \qquad (A.36)$$

Then,

$$\mathrm{CV}_1^{**} = \mathrm{CV}_2^{**} + T_1 - 2\,T_2\,. \qquad (A.37)$$

The assumption in (3.4) that $|g|$ is bounded implies that the same is true for $|a_2|$. In view of (3.2)(a) with $Z = W^*$, f_{W^*} is bounded away from zero on the support, $\mathcal{S}_p$, of p. It follows from these properties that, uniformly in $x \in \mathcal{S}_p$,

$$\left| \frac{\Delta_{a_2}(x) - \Delta_{a_2,-j}(x)}{f_{W^*}(x)} \right| + \left| \frac{a_2(x)\, \{\Delta_{f_{W^*}}(x) - \Delta_{f_{W^*},-j}(x)\}}{f_{W^*}(x)^2} \right| \\ \le \text{const.}\ \frac{1 + |Y_j|}{nh} \left| K_U\!\left(\frac{x - W_j^{**}}{h} \right) \right| .$$

Moreover, (A.6) implies that $|K_U(u)| \le \text{const.}\ \lambda(h)^{1/2}$ for all real numbers u. Therefore,

$$\frac{1}{n} \sum_{j=1}^{n} \left\{ \left| \frac{\Delta_{a_2}(W_j^*) - \Delta_{a_2,-j}(W_j^*)}{f_{W^*}(W_j^*)} \right| \right. \\ \left. + \left| \frac{a_2(W_j^*)\, \{\Delta_{f_{W^*}}(W_j^*) - \Delta_{f_{W^*},-j}(W_j^*)\}}{f_{W^*}(W_j^*)^2} \right| \right\} p(W_j^*) \\ \le \text{const.}\ \frac{\lambda(h)^{1/2}}{n^2 h} \sum_{j=1}^{n} (1 + |Y_j|) = o_p\{\lambda(h)/nh\}\,, \qquad (A.38)$$

uniformly in $h \in \mathcal{H}$. Since the right-hand side (A.38) does not depend on the B sets of simulated data (U_k^*, U_k^{**}), then (A.38) holds uniformly in those data. This result, (A.33) and (A.24), the latter applying on this occasion to (A.32) and (A.33), imply that, uniformly in the same sense,

$$|\mathrm{CV}_2^{**} - \mathrm{CV}_3^{**}| = o_p\{\lambda(h)/nh\} . \tag{A.39}$$

Also, defining CV^{**} as at (1.10), and noting the definition of $R_{3,j}(x)$ at (A.31), we have,

$$\mathrm{CV}^{**} - \mathrm{CV}_1^{**} = 2\,T_3 + T_4 , \tag{A.40}$$

where,

$$T_3 = \frac{1}{n} \sum_{j=1}^{n} V_j^* \, R_{3,j}(W_j^*) \, p(W_j^*) , \tag{A.41}$$

$$T_4 = \frac{1}{n} \sum_{j=1}^{n} \Bigg[2 \left\{ \frac{\Delta_{a_2,-j}(W_j^*)}{f_{W^*}(W_j^*)} - \frac{a_2(W_j^*)\, \Delta_{f_{W^*},-j}(W_j^*)}{f_{W^*}(W_j^*)^2} \right\} R_{3,j}(W_j^*) + R_{3,j}(W_j^*)^2 \Bigg] \, p(W_j^*) .$$

Results (A.32), (A.33) and (A.24) (the latter applying to (A.32) and (A.33)), imply that, for each $\epsilon \in (0,1)$, all $C > 0$ and some $D > 0$,

$$P\Big[|T_4| > \big\{\lambda(h)/n^{1-\epsilon}\, h\big\}^{3/2} + D\, h^6 \quad \text{for some } h \in \mathcal{H}\Big] = O\big(n^{-C}\big) , \tag{A.42}$$

and that this result also holds if the supremum implicit in the probability statement is also taken over the B sets of simulated data (U_k^*, U_k^{**}), rather than just over $h \in \mathcal{H}$.

Combining (A.37) and (A.39)–(A.42), and recalling that T_1, T_2 and T_3 are given at (A.35), (A.36) and (A.41), respectively, we deduce that,

$$\mathrm{CV}^{**} = \mathrm{CV}_3^{**} + T_1 - 2\,T_2 + 2\,T_3 + o_p\big\{(nh)^{-1}\,\lambda(h) + h^4\big\} , \tag{A.43}$$

uniformly in $h \in \mathcal{H}$ and also in the B sets of simulated data (U_k^*, U_k^{**}).

Step (VII): Approximation to T_3. Arguing by analogy, the expansion (A.31) can be given with the precision of that at (A.1), leading to an exact formula for $R_{3,j}$, defined by (A.31). This gives the following result:

$$R_{3,j}(x)\, f_{W^*}(x)^2 = \frac{a_2(x)\, \Delta_{f_{W^*},-j}(x)^2}{f_{W^*}(x)} - \Delta_{a_2,-j}(x)\, \Delta_{f_{W^*},-j}(x) + R_{4,j}(x)\,, \tag{A.44}$$

where, for some function $0 < \theta(x) < 1$,

$$R_{4,j}(x) = \frac{\Delta_{a_2,-j}(x)\, \Delta^2_{f_{W^*},-j}(x)}{f_{W^*}(x) + \Delta_{f_{W^*},-j}(x)} - \frac{a_2(x)\, \Delta^3_{f_{W^*},-j}(x)}{\{f_{W^*}(x) + \theta(x)\Delta_{f_{W^*},-j}(x)\}^2}\,.$$

Similarly to (A.32), we have for some $D > 0$ and all $C > 0$,

$$P\Bigg[\, |R_{4,j}(x)| > \big\{\lambda(h)/n^{1-\epsilon}\, h\big\}^{3/2} + D\, h^6 \quad \text{for some} \quad 1 \le j \le n \quad \text{and some} \quad (h, x) \in \mathcal{H} \times \mathcal{C} \Bigg] = O\big(n^{-C}\big)\;, \tag{A.45}$$

and this result continues to hold if the implicit supremum is taken also over each of $B = O(n^D)$ different sequences $(U_1^*, U_1^{**}), \ldots, (U_n^*, U_n^{**})$, for any fixed $D > 0$. The quantity

$$\Delta_j(x) = \frac{a_2(x)\, \Delta_{f_{W^*},-j}(x)^2}{f_{W^*}(x)} - \Delta_{a_2,-j}(x)\, \Delta_{f_{W^*},-j}(x)\,, \tag{A.46}$$

on the right-hand side of (A.44), represents the quadratic portion of $R_{3,j}(x)\, f_{W^*}(x)^2$, and $R_{4,j}(x)$ denotes cubic and higher-order parts.

Define $\mu(h, x) = E\{\Delta_j(x)\}$, not depending on j. Standard arguments give, for each compact set $\mathcal{C}$,

$$\sup_{h\in\mathcal{H}} \sup_{x\in\mathcal{C}} \left[|\mu(h,x)| \left\{ \frac{\lambda(h)}{nh} + h^4 \right\}^{-1} \right] = O(1)\,.$$

Note too that, by definition of g_2,

$$E(V_j^* \mid W_j^* = x) = E\{g(X_j) - g_2(W_j^*) \,|\, W_j^* = x\} = 0\,, \tag{A.47}$$

for each x. Therefore, by Rosenthal's inequality, for each $r \geq 1$,

$$\sup_{h \in \mathcal{H}} \left[E \left| \frac{1}{n} \sum_{j=1}^{n} V_j^* \, \frac{\mu(W_j^*, h)}{f_{W^*}(W_j^*)^2} \, p(W_j^*) \right|^{2r} \left\{ \frac{\lambda(h)}{nh} + h^4 \right\}^{-2r} \right] = O\big(n^{-r}\big) \,. \tag{A.48}$$

This result and Markov's inequality imply that, for each subset $\mathcal{H}_n$ of $\mathcal{H}$ containing at most $O(n^{D_1})$ elements (for any fixed $D_1 > 0$), we have, for each $D_2, \epsilon > 0$,

$$P\left[\left| \frac{1}{n} \sum_{j=1}^{n} V_j^* \, \frac{\mu(W_j^*, h)}{f_{W^*}(W_j^*)^2} \, p(W_j^*) \right| > n^{\epsilon-(1/2)} \left\{ \frac{\lambda(h)}{nh} + h^4 \right\} \text{ for some } h \in \mathcal{H}_n \right] = O\big(n^{-D_2}\big) \,. \tag{A.49}$$

It can be proved from (A.19) that, for any compact set $\mathcal{C}$, there exist constants $D_1, D_2 > 0$ such that,

$$\sup_{h_1, h_2 \in \mathcal{H}, h_1 \neq h_2} \; \sup_{x \in \mathcal{C}} \frac{|\mu(h_1, x) - \mu(h_2, x)|}{|h_1 - h_2|} \leq D_1 \, n^{D_2} \,.$$

This result, and the lattice-approxiation argument leading from (A.11) to (A.23), permit us to strengthen (A.49) by replacing the supremum over $h \in \mathcal{H}_n$ by $h \in \mathcal{H}$: for each $D_2, \epsilon > 0$,

$$P\left[\left| \frac{1}{n} \sum_{j=1}^{n} V_j^* \, \frac{\mu(W_j^*, h)}{f_{W^*}(W_j^*)^2} \, p(W_j^*) \right| > n^{\epsilon-(1/2)} \left\{ \frac{\lambda(h)}{nh} + h^4 \right\} \quad \text{for some} \quad h \in \mathcal{H} \right] = O\big(n^{-D_2}\big) \,, \tag{A.50}$$

where,

> the result continues to hold if the implicit supremum is taken also over the $B = O(n^D)$ sets of simulated data (U_k^*, U_k^{**}), for any fixed $D > 0$. $\quad (A.51)$

Moment calculations lead to the following analogue of (A.48): there exists $\delta > 0$ such that, for all integers $r \geq 1$,

$$\sup_{h\in\mathcal{H}} \left[E \left| \frac{1}{n} \sum_{j=1}^{n} V_j^* \frac{\Delta_j(W_j^*) - \mu(W_j^*, h)}{f_{W^*}(W_j^*)^2} p(W_j^*) \right|^{2r} \left\{ \frac{\lambda(h)}{nh} + h^4 \right\}^{-2r} \right] = O\big(n^{-r\delta}\big) . \tag{A.52}$$

Markov's inequality, and the lattice approximation argument, lead from this result to the analogue of (A.50): for a constant $\delta > 0$, and all $C > 0$,

$$P\left[\left| \frac{1}{n} \sum_{j=1}^{n} V_j^* \frac{\Delta_j(W_j^*) - \mu(W_j^*, h)}{f_{W^*}(W_j^*)^2} p(W_j^*) \right| > n^{-\delta} \left\{ \frac{\lambda(h)}{nh} + h^4 \right\} \quad \text{for some} \quad h \in \mathcal{H} \right] = O\big(n^{-C}\big) , \tag{A.53}$$

to which (A.51) also applies.

Combining (A.50) and (A.53) we deduce that, if $\delta > 0$ is sufficiently small,

$$P\left[\left| \frac{1}{n} \sum_{j=1}^{n} V_j^* \frac{\Delta_j(W_j^*)}{f_{W^*}(W_j^*)^2} p(W_j^*) \right| > n^{-\delta} \left\{ \frac{\lambda(h)}{nh} + h^4 \right\} \quad \text{for some} \quad h \in \mathcal{H} \right] = O\big(n^{-C}\big) . \tag{A.54}$$

Properties (A.44), (A.45) and (A.54) imply that,

$$P\left[|T_3| > n^{-\delta} \left\{ \frac{\lambda(h)}{nh} + h^4 \right\} \quad \text{for some } h \in \mathcal{H} \right] = O\big(n^{-C}\big) , \tag{A.55}$$

and (A.51) also applies.

Step (VIII): Approximation to T_2, defined at (A.36). Define $\Delta_{j1} = \Delta_{a_2,-j} \times f_{W^*} - a_2 \, \Delta_{f_{W^*},-j}$ and $\mu_1(h, x) = E\{\Delta_{j1}(x)\}$, the latter not depending on j. We centre the term within braces in (A.36), writing $T_2 = T_{21} + T_{22}$ where,

$$T_{21} = \frac{1}{n} \sum_{j=1}^{n} V_j^* \frac{\Delta_{j1}(W_j^*) - \mu_1(h, W_j^*)}{f_{W^*}(W_j^*)^2} p(W_j^*) ,$$

$$T_{22} = \frac{1}{n} \sum_{j=1}^{n} V_j^* \frac{\mu_1(h, W_j^*)}{f_{W^*}(W_j^*)^2} p(W_j^*) .$$

Analogously to (A.48) it may be proved that, for all $r \geq 1$,

$$\sup_{h \in \mathcal{H}} \left[E|T_{22}|^{2r} \left\{ \frac{\lambda(h)}{nh} + h^4 \right\}^{-r} \right] = O\big(n^{-r}\big) \,,$$

whence the argument leading to (A.50) implies that, for each $D, \epsilon > 0$,

$$P\left[|T_{22}| > n^{\epsilon-(1/2)} \left\{ \frac{\lambda(h)}{nh} + h^4 \right\}^{1/2} \quad \text{for some} \quad h \in \mathcal{H} \right] = O\big(n^{-D}\big) \,. \tag{A.56}$$

Define $T_{21} = T_{211} - T_{212}$, where

$$T_{211} = \frac{1}{n} \sum_{j=1}^{n} V_j^* \, \frac{\Delta_{a_2,-j}(W_j^*) - \mu_{11}(h, W_j^*)}{f_{W^*}(W_j^*)} \, p(W_j^*) \,,$$
$$T_{212} = \frac{1}{n} \sum_{j=1}^{n} V_j^* \, \frac{a_2(W_j^*) \, \{\Delta_{f_{W^*},-j}(W_j^*) - \mu_{12}(h, W_j^*)\}}{f_{W^*}(W_j^*)^2} \, p(W_j^*) \,,$$

$\mu_{11}(x) = E\{\Delta_{a_2,-j}(x)\}$ and $\mu_{12}(x) = E\{\Delta_{f_{W^*},-j}(x)\}$. We claim that, with $T = T_{21k}$ and $k = 1, 2$, and for some $\delta > 0$ and all $D > 0$,

$$P\left[|T| > n^{-\delta} \left\{ \frac{\lambda(h)}{nh} + h^4 \right\} \quad \text{for some} \quad h \in \mathcal{H} \right] = O\big(n^{-D}\big) \,. \tag{A.57}$$

Together, this property and (A.56) imply that,

for some $\delta > 0$ and all $D > 0$, (A.57) holds with $T = T_2$. $\qquad (A.58)$

We shall treat only the case $k = 2$, showing that (A.57) holds for $T = T_{212}$. The case $k = 1$ is similar.

The quantity T_{212} is a U-statistic, although both of its linear projections equal zero. To appreciate why, note that if $\mathcal{F}_j$ and $\mathcal{G}_j$ denote the sigma-field generated by the data (and the pseudodata with superscripts $*$ or $**$) that have index j, and by all data (and pseudodata) other than those that have index j, respectively, then,

$$E\left[V_j^* \frac{a_2(W_j^*)\{\Delta_{f_{W^*},-j}(W_j^*) - \mu_{12}(h,W_j^*)\}}{f_{W^*}(W_j^*)^2}\, p(W_j^*) \,\Bigg|\, \mathcal{F}_j\right]$$
$$= \frac{V_j^*\, a_2(W_j^*)\, p(W_j^*)}{f_{W^*}(W_j^*)^2}\, E\Big\{\Delta_{f_{W^*},-j}(W_j^*) - \mu_{12}(h,W_j^*) \,\Big|\, W_j^*\Big\} = 0\,,$$
$$E\left[V_j^* \frac{a_2(W_j^*)\{\Delta_{f_{W^*},-j}(W_j^*) - \mu_{12}(h,W_j^*)\}}{f_{W^*}(W_j^*)^2}\, p(W_j^*) \,\Bigg|\, \mathcal{G}_j\right]$$
$$= E\left(E\left[V_j^* \frac{a_2(W_j^*)\{\Delta_{f_{W^*},-j}(W_j^*) - \mu_{12}(h,W_j^*)\}}{f_{W^*}(W_j^*)^2}\, p(W_j^*) \,\Bigg|\, \mathcal{G}_j, W_j^*\right] \,\Bigg|\, \mathcal{G}_j\right)$$
$$= E\left[\frac{a_2(W_j^*)\, p(W_j^*)}{f_{W^*}(W_j^*)^2}\, E(V_j^* \mid W_j^*)\, \{\Delta_{f_{W^*},-j}(W_j^*) - \mu_{12}(h,W_j^*)\} \,\Bigg|\, \mathcal{G}_j\right] = 0\,,$$

where we employed (A.47) to obtain the final identity. Using these properties, (A.4) and (A.5) it can be shown that $E(T_{212}^2) = O\{\lambda(h)/(n^2h)\}$ and

$$E\big(T_{212}^{2r}\big) = O\big[n^{-r\delta}\{\lambda(h)/(nh)\}^{2r}\big]\,,$$

where $\delta > 0$ does not depend on r. The argument leading to (A.50) now gives (A.57) with $T = T_{21k}$.

Step (IX): Approximation to CV_3^{**}*, defined at* (*A*.34)*, and completion of proof of Theorem 3.2.* This aspect of the proof amounts to a deterministic approximation to (the linear component of) summed squared error, and has been covered extensively in the literature. Therefore we shall not give details, mentioning only that the result can be obtained using a martingale-based U-statistic technique (Hall, 1984) to show that, for all integers $r \geq 1$,

$$E\{\mathrm{CV}_3^{**} - E(\mathrm{CV}_3^{**})\}^{2r} = n^{-r\delta}\left\{\frac{\lambda(h)}{nh} + h^4\right\}^{2r\delta},$$

uniformly in $h \in \mathcal{H}$ and also in the $B = O(n^D)$ different sequences $(U_1^*, U_1^{**}), \ldots,$

(U_n^*, U_n^{**}), where $\delta > 0$ does not depend on r; and then employing the lattice-approximation argument leading to (A.23)–(A.24), to prove that,

$$\mathrm{CV}_3^{**} - E(\mathrm{CV}_3^{**}) = o_p\left\{\frac{\lambda(h)}{nh} + h^4\right\},$$

uniformly in $h \in \mathcal{H}$. Simpler calculations show that,

$$E(\mathrm{CV}_3^{**}) = \mathrm{AMISE}(\hat{g}_j - g_j) + o\left\{\frac{\lambda(h)}{nh} + h^4\right\},$$

uniformly in $h \in \mathcal{H}$. Therefore, uniformly in $h \in \mathcal{H}$ and in the $B = O(n^D)$ different sequences $(U_1^*, U_1^{**}), \ldots, (U_n^*, U_n^{**})$,

$$\mathrm{CV}_3^{**} = \mathrm{AMISE}(\hat{g}_j - g_j) + o_p\left\{\frac{\lambda(h)}{nh} + h^4\right\}. \qquad (A.59)$$

Together, (A.43), (A.55), (A.58) and (A.59) imply that, uniformly in the same sense,

$$\begin{aligned}\mathrm{CV}^{**}(h) &= \mathrm{AMISE}(\hat{g}_2 - g_2)(h) + T_1 + o_p\left\{\frac{\lambda(h)}{nh} + h^4\right\} \\ &= \frac{A\,\lambda(h)}{nh} + B_2\,h^4 + T_1 + o_p\left\{\frac{\lambda(h)}{nh} + h^4\right\},\end{aligned}$$

where A and B_2 are as at (2.9), and the last line is derived using arguments in section 4.5. Therefore,

$$\mathrm{CV}_2 = \frac{1}{B}\sum_{b=1}^{B} \mathrm{CV}_b^{**} = \frac{A\,\lambda(h)}{nh} + B_2\,h^4 + T_1 + o_p\left\{\frac{\lambda(h)}{nh} + h^4\right\}. \qquad (A.60)$$

This result is equivalent to (2.11) in the case $j = 2$, and directly implies (2.12) in that instance. Therefore, (A.60) entails the part of Theorem 3.2 that relates to $j = 2$.

A.2. Proof of Theorem 3.3. Put $j = 1$ or 2, and let A and τ denote respectively a random variable independent of X, and a positive constant, chosen such that τA has the distribution of U when $j=1$, or the distribution of U^* when $j=2$. We fix the value of τ, and the distribution of A, by asking that $E(A^2)=1$. Note too that, since $E(U)=0$ (see (3.7)), then $E(A)=0$. Put $Z=X+\tau A$,

$$g_3(x)=E\{g(X)\,|\,Z=x\}\,,\quad q_3=g_3''+2\,f_Z'\,g_3'\,f_Z^{-1}\,,\quad C=\int f_Z\,p\,q_3^2\,. \tag{A.61}$$

(Thus, $C=B_j$, the latter defined at (2.9), when $j=1,2$.) To establish (2.16) it suffices to show that

$$C=B_0+\tau^2\,D+o\big(\tau^2\big) \tag{A.62}$$

as $\tau\to 0$, where B_0 is given by (2.9) with q_j there replaced by q, given at (2.5), and where the constant D does not depend on τ.

By Taylor expansion,

$$\begin{aligned}\alpha(x)\equiv\frac{d}{dx}\,E\{A\,I(X+\tau A\le x)\}&=\int a\,\frac{d^2}{da\,dx}\,P(A\le a,X\le x-\tau A)\,da\\&=\int a\,\{f_X(x)-\tau\,f_X'(x)\,a\}\,dF_A(a)+o(\tau)=-\tau\,f_X'(x)+o(\tau)\,,\\f_Z(x)=E\{f_X(x-\tau A)\}&=f_X(x)+\tfrac{1}{2}\,\tau^2\,f_X''(x)+o\big(\tau^2\big)\,.\end{aligned} \tag{A.63}$$

(Here and below the remainders are of the stated orders uniformly in $x\in\mathcal{S}_p$.) Hence,

$$E(A\,|\,X+\tau A=x)=\frac{\alpha(x)}{f_Z(x)}=-\tau\,f_X'(x)\,f_X(x)^{-1}+o(\tau)\,,$$

$$\begin{aligned}g_3(x)&=E\{g(x-\tau A)\,|\,X+\tau A=x\}\\&=E\Big\{g(x)-\tau A\,g'(x)+\tfrac{1}{2}\,(\tau A)^2\,g''(x)\;\Big|\;X+\tau A=x\Big\}+o\big(\tau^2\big)\\&=g(x)+\tfrac{1}{2}\,\tau^2\,\big\{2\,g'(x)\,f_X'(x)\,f_X(x)^{-1}+g''(x)\big\}+o\big(\tau^2\big)\\&=g(x)+\tfrac{1}{2}\,\tau^2\,q(x)+o\big(\tau^2\big)\,,\end{aligned}$$

where q is as at (2.5). A similar argument shows that $g_3^{(k)} = g^{(k)} + \frac{1}{2}\,\tau^2\, q^{(k)} + o(\tau^2)$ for $k = 0, 1, 2$. An analogous extension of (A.63) gives an expansion of $f_Z^{(k)}$ for $k = 0, 1$. This leads to the result,

$$q_3 = g_3'' + 2\,\frac{f_X' + \frac{1}{2}\,\tau^2\, f_X'''}{f_X + \frac{1}{2}\,\tau^2\, f_X''}\,\big(g' + \tfrac{1}{2}\,\tau^2\, q'\big) + o\big(\tau^2\big) = q + \tau^2\,\gamma + o\big(\tau^2\big)\,, \quad (A.64)$$

uniformly on $\mathcal{S}_p$, where γ is a function defined by the second identity in (A.64). Substituting (A.64), and formula (A.63) for f_Z (and similarly for its derivatives), into the expression for C in (A.61); and noting that p is bounded and compactly supported; we deduce (A.62).

REFERENCES

Delaigle, A. and Hall, P. (2008). Using SIMEX for smoothing-parameter choice in errors-in-variables problems. *JASA*, **103**, 280-287.

Fan, J. and Truong, Y.K. (1993). Nonparametric regression with errors in variables. *Ann. Statist.* **21**, 1900–1925.

Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *J. Multivariate Anal.* **14**, 1–16.