Estimation of observation-error variance in errors-in-variables regression

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

[1] Department of Mathematics and Statistics, The University of Melbourne, VIC 3010, Australia. Research supported by a fellowship from the Australian Research Council.

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

[3] Department of Statistics, University of California at Davis, Davis, CA 95616, USA.

Abstract: Assessing the variability of an estimator is a key component of the process of statistical inference. In nonparametric regression, estimating observation-error variance is the principal ingredient needed to estimate the variance of the regression mean. Although there is an extensive literature on variance estimation in nonparametric regression, the techniques developed in conventional settings generally cannot be applied to the problem of regression with errors in variables, where the explanatory variables are not observable directly. In this paper we introduce methods for estimating observation-error variance in errors-in-variables regression. We consider cases where the variance is modelled either nonparametrically or parametrically. The performance of our methods is assessed both numerically and theoretically. We also suggest a fully data-driven bandwidth selection procedure, a problem which is notoriously difficult in errors-in-variables contexts.

Keywords: Bandwidth, kernel estimation, nonparametric curve estimation, nonparametric regression, parametric model, statistical smoothing, variance estimation.

Short title. Variance estimation.

1 Introduction

In the standard measurement error-free setting, determining the variance of an estimator of a nonparametric regression mean consists in estimating a function $\tau > 0$ from data (X, Y) that are generated by the regression model $Y = g(X) + \tau(X)^{1/2}\,\varepsilon$, where ε and X are independent variables, ε has zero mean and unit variance, and, apart from smoothness assumptions, g is completely unspecified. The quantity $\tau(X)^{1/2}\,\varepsilon$ is generally referred to as observation error, and τ is the observation-error variance. A variety of methods have been developed for treating this problem, but simple techniques that enjoy good theoretical properties are generally founded either on differencing values of Y that correspond to nearby values of X, or on residual-based approaches.

In this paper we consider estimation of τ in the more complex, nonparametric errors-in-variables setting. Here the data $(W_1, Y_1), \ldots, (W_n, Y_n)$ are generated by the model

$$W = X + U\,, \quad Y = g(X) + \tau(X)^{1/2}\,\varepsilon\,, \tag{1.1}$$

where U, ε and X are independent random variables, g and τ are smooth, unknown functions, $\tau > 0$, $E(\varepsilon) = 0$, $E(\varepsilon^2) = 1$ and the distribution of U is known.

In this context, since values of X are not observable then the popular variance-estimation methods discussed above cannot simply be modified to provide consistent estimators of τ. We shall develop alternative approaches based on deconvolution techniques and describe their performance. Their properties will be discussed in cases where τ is estimated nonparametrically, as a function, and also when τ is assumed to have a parametric form. In both settings we shall give convergence rates, and in parametric cases we shall show that our estimators are root-n consistent, provided the variance function is a sufficiently smooth functional of the unknown parameters. Our numerical work will attest to the good performance of the suggested new methodology.

Estimation of the observation-error variance is an important tool for statistical inference. In the errors-in-variables context, as in the measurement error-free case, knowledge of τ is essential if we are to assess the variability of nonparametric estimators of g such as, for example, the deconvolution-kernel estimator of Fan and Truong (1993). Indeed, the asymptotic variance of this estimator depends on the densities f_U and f_X of U and X, respectively, and on g and τ. Since f_U is known, and a variety of methods for estimating f_X and g are readily available, then the only missing ingredient is τ. Thus, estimation of τ in (1.1) is central to characterising empirically the accuracy of estimators of g.

Properties of nonparametric estimators of τ follow relatively easily from known results in the problem of estimating g, whereas their counterparts in the case of parametric models are more difficult to determine. These differences, apparent for both methodology and theory, arise because of the nonstandard way in which, in the parametric case, we must combine an infinite-parameter model for g with a finite-parameter model for τ.

In the measurement error-free case, the variance estimation problem has been addressed by many authors; see, for example, the work of Rice (1984), Gasser *et al.* (1986), Müller and Stadtmüller (1987), Buckley *et al.* (1988), Hall and Marron (1990), Hall *et al.* (1990), Müller and Stadtmüller (1992), Seifert *et al.* (1993), Neumann (1994), Müller and Zhao (1995), Ruppert *et al.* (1997), Dette *et al.* (1998), Fan and Yao (1998), Lavergne and Vuong (1998), Müller *et al.* (2003), Munk *et al.* (2005), Sheehy *et al.* (2005) and Tong and Wang (2005). The nonparametric estimators of g, τ and f_X that we use can be expressed in such a way that, when U in (1.1) is identically zero, they collapse to standard kernel estimators of those functions. In the measurement error-free setting our nonparametric estimator of variance also reverts to techniques that have been employed before.

There is a substantial literature on estimation of the regression mean, g, in errors-in-variables problems. The book by Carroll *et al.* (2006) is an excellent entry point to this work. Earlier contributions to nonparametric or semiparametric methodology include those of Carroll *et al.* (1996, 1999), Kim and Gleser (2000), Lin and Carroll (2000), Stefanski (2000), Devanarayan and Stefanski (2002), Linton and Whang (2002), Carroll and Hall (2004), Schennach (2004b), Delaigle *et al.* (2006), Huang *et al.* (2006) and Delaigle and Meister (2007).

Parametric errors-in-variables regression has also received considerable attention in the literature. References include Stefanski and Carroll (1987), Hsiao (1989), Stefanski (1989), Gleser (1990), Nakamura (1990), Cook and Stefanski (1994), Carroll *et al.* (1996), Cheng and Schneeweiss (1998), Taupin (2001), Li (2002) and Schennach (2004a). See Fuller (1987) and Carroll *et al.* (2006) for a more extensive list of references.

Sections 2 and 3 will introduce our nonparametric and parametric estimators, respectively. Their theoretical properties will be outlined in section 4. In preparation for an account of numerical properties in section 6, section 5 will discuss bandwidth choice. The methods proposed there will be used throughout our applications to simulated data. Finally, the appendix will give theoretical arguments behind the results stated in section 4.

2 Nonparametric estimators of τ

2.1 Main estimation procedure

Known results in the problem of estimating g imply simple sufficient conditions for identifiability of τ. Indeed, if (a) the distribution of ϵ has finite fourth moment and zero mean, (b) the characteristic function of the distribution of U does not vanish except at isolated points, and (c) g and τ satisfy Hölder smoothness conditions, then the function g defined by $g(x) = E(Y|X = x)$ is identifiable, because it is consistently estimated in the model at (1.1) using, for example, the methodology suggested by Fan and Truong (1993). Likewise, if (a)–(c) hold then $m(x) = E(Y^2|X = x)$ is identifiable in the model where $W = X + U$ and $Y^2 = g(X) + 2\, g(X)\, \tau(X)^{1/2}\, \varepsilon$, because it can be consistently estimated using the same technique. Moreover, using the second identity in (1.1) we see that we can write $m = g^2 + \tau$. Combining these properties we see that τ is identifiable from data generated by (1.1), provided that conditions (a)–(c) above hold.

Suppose we have a dataset $\mathcal{D} = \{(W_1, Y_1), \ldots, (W_n, Y_n)\}$ on (W, Y), generated by the model (1.1). As implied by the identifiability arguments in the previous paragraph, to construct a nonparametric estimator of the variance function τ we can first construct nonparametric estimators $\widehat{g}$ and $\widehat{m}$ of the regression curves $g(x) = E(Y|X = x)$ and $m(x) = E(Y^2|X = x)$. Then we can take our nonparametric estimator of τ to be

$$\widetilde{\tau} = \max\big(\widehat{\tau}, 0\big), \text{ where } \widehat{\tau} = \widehat{m} - \widehat{g}^2. \tag{2.1}$$

Estimation of g and m are two nonparametric errors-in-variables (or deconvolution) regression problems: in both cases the goal is to estimate a function $E(V|X)$ from data on (W, V), where $W = X + U$ is a contaminated version of X. Several nonparametric estimators have been developed in the literature, but one of the most recent methods is the local polynomial deconvolution estimator of Delaigle, Fan and Carroll (2009). To define their estimator, let K be a symmetric kernel function integrating to 1 and with compactly supported Fourier transform ϕ_K. Also, let ϕ_U denote the characteristic function corresponding to the density f_U of U, and $h > 0$ be a bandwidth. The qth order local polynomial deconvolution estimator of $E(V|X = x)$, with $q \geq 0$ an integer, is defined by

$$\widehat{E}(V|X = x) = (1, 0, \dots, 0)\, \mathbf{S}_n^{-1} \mathbf{T}_n, \tag{2.2}$$

where $\mathbf{S}_n = \{S_{n,j+k}(x)\}_{0 \leq j,k \leq q}$ and $\mathbf{T}_n = \{T_{n,0}(x), \dots, T_{n,q}(x)\}^T$, and with, for $k = 0, 1, \dots, 2q$,

$$S_{n,k}(x) = \frac{1}{nh} \sum_{j=1}^{n} L_k\left(\frac{x - W_j}{h}\right) \ , \quad T_{n,k}(x) = \frac{1}{nh} \sum_{j=1}^{n} V_j \, L_k\left(\frac{x - W_j}{h}\right), \tag{2.3}$$

$$L_k(u) = i^{-k} \frac{1}{2\pi} \int e^{-itx} \phi_K^{(k)}(t) / \phi_U(-t/h) \, dt. \tag{2.4}$$

Replacing V_j in (2.3) by Y_j and Y_j^2, respectively, (2.2) provides the local polynomial deconvolution estimators $\widehat{g}$ and $\widehat{m}$ of g and m, respectively.

Note that a version of $\widehat{\tau}$ in cases where X is observed directly, without measurement error, was given by Yao and Tong (1994).

2.2 Correcting for negativity

The estimator $\widetilde{\tau}$ is simple and straightforward, but in cases where $\widehat{m} - \widehat{g}^2$ takes negative values, $\widetilde{\tau}$ projects them to zero, which could sometimes be viewed as an unattractive property. In those cases, an alternative, "smoother" way to correct for negativity is to use the estimator

$$\bar{\tau} = E\big\{ \max\big(\widehat{m}^\dagger - \widehat{g}^{\dagger 2}, 0\big) \,\big|\, \mathcal{D} \big\}, \tag{2.5}$$

where $\widehat{m}^\dagger$ and $\widehat{g}^\dagger$ denote the versions of $\widehat{m}$ and $\widehat{g}$, respectively, computed from a resample of size n, drawn by sampling randomly, with replacement, from $\mathcal{D}$. The estimator $\bar{\tau}$, which can be thought of as being motivated by Breiman's (1996) bagging method, is a little more complicated, but when the true τ is bounded away from zero it less often takes the default value zero. Indeed, it vanishes if and only if, for all possible resamples drawn from $\mathcal{D}$, $\widehat{m}^\dagger \leq \widehat{g}^{\dagger 2}$. Note that we introduce the alternative estimator $\bar{\tau}$ only as a way to correct for negativity more smoothly than simply truncating to zero. See also Figure 4 in section 6 for an application. In particular, there is no asymptotic gain to be expected from $\bar{\tau}$ compared to $\widetilde{\tau}$. We shall show in section 4 that the two estimators are first-order equivalent.

3 Parametric estimator of τ

3.1 Background

Although we treat g from a nonparametric viewpoint, we may wish to use a parametric model, say $\tau = \tau(\cdot \,|\, \theta)$, for the variance, as is sometimes done in the measurement error-free case. For example, the homoscedastic context where $\tau(x) \equiv \theta$ is a constant, is commonly assumed in nonparametric regression. Log-linear variance and polynomial models are also in use in the measurement error-free case. See, for example, Müller and Zhao (1995), who survey literature on the topic, and Fan and Gijbels (1996, p. 146). Other work using polynomial (including linear) and log-linear models for the variance function includes that of Hasbrouck (1986), Finkenstädt *et al.* (2002) and Meyer (2005). Linear models are often fitted in response to either empirical evidence or physical considerations that indicate that measurement error variance is increasing or decreasing as a function of the explanatory variable. Sometimes quadratic models are used to reflect evidence that the rate of increase or decrease is varying. The method we present below, valid in the case of models with measurement error, is not restricted to these situations and can be used in general parametric contexts.

3.2 Estimator

Let $\widehat{\tau} = \widehat{m} - \widehat{g}^2$ be the nonparametric estimator of τ defined at (2.1), using qth order local polynomials of m and g, and let $\theta \in \mathbb{R}^p$ be the parameter of interest. Our estimator of θ relies on the following main idea: estimate θ so as to make the parametric estimator of τ sufficiently close to its nonparametric version $\widehat{\tau}$. Below, we give the definition of our estimator which results from this idea. The details leading to our definition are deferred to section 3.3. Also, to simplify the presentation we assume throughout that the distribution of U is symmetric, and so ϕ_U is real-valued.

Let $\widehat{d} = \det(\mathbf{S}_n)$, with $\mathbf{S}_n$ defined below equation (2.2), and put $\widehat{r}_1 = \widehat{g}\,\widehat{d}$ and $\widehat{r}_2 = \widehat{m}\,\widehat{d}$. We suggest choosing $\theta = \widehat{\theta}$ to solve the equation $S(\theta) = 0$, where both sides are p-vectors and

$$S(\theta) = \int \big\{(\widetilde{r_2 d})(x) - \widetilde{r}_1{}^2(x) - \tau(x \,|\, \theta)\, \widetilde{d}^2(x)\big\}\, \dot{\tau}(x \,|\, \theta)\, \omega(x)\, dx\,, \tag{3.1}$$

$\dot{\tau}(x \,|\, \theta) = (\partial/\partial\theta)\, \tau(x \,|\, \theta)$ is a p-vector, ω denotes a nonnegative, compactly supported weight function, and $\widetilde{d}^2, \widetilde{r_2 d}$ and $\widetilde{r}_1{}^2$ denote the diagonal-free versions of $\widehat{d}^2, \widehat{r}_2\widehat{d}$ and $\widehat{r}_1^2$. Here we mean that $\widehat{d}^2, \widehat{r}_2\widehat{d}$ and $\widehat{r}_1^2$ each comprise terms of the type $\sum_{i_1,\ldots,i_k=1}^n$ for some $k > 0$, where the summands involve the products $L_{j_1}\big\{(x - W_{i_1})/h\big\} \ldots L_{j_k}\big\{(x - W_{i_k})/h\big\}$ for some $j_1, \ldots, j_k$ between 0 and q, and their diagonal-free versions are those where these sums are replaced by $\sum_{i_1 \neq i_2 \neq \ldots \neq i_k}$.

Example 1 (Formula when $\widehat{\tau}$ is based on local constant estimators of m and g)**.** When nonparametric estimators of g and m are taken to be local constant, that is when $q = 0$, we have

$$\widehat{d}(x) = \frac{1}{nh} \sum_{j=1}^n L_0\Big(\frac{x - W_j}{h}\Big), \tag{3.2}$$

$$\widehat{r}_1(x) = \frac{1}{nh} \sum_{j=1}^n Y_j L_0\Big(\frac{x - W_j}{h}\Big), \quad \widehat{r}_2(x) = \frac{1}{nh} \sum_{j=1}^n Y_j^2 L_0\Big(\frac{x - W_j}{h}\Big) \tag{3.3}$$

and $S(\theta) = 0$ can be written as:

$$\sum\sum_{j_1 \neq j_2} \int \big\{Y_{j_1}^2 - Y_{j_1} Y_{j_2} - \tau(x \,|\, \theta)\big\}\, L_0\Big(\frac{x - W_{j_1}}{h}\Big)\, L_0\Big(\frac{x - W_{j_2}}{h}\Big)\, \dot{\tau}(x \,|\, \theta)\, \omega(x)\, dx = 0\,.$$

Example 2 (Estimator when $\tau(x \,|\, \theta)$ is a polynomial)**.** In this case, $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$ is a p-vector, $\tau(x \,|\, \theta) = \theta_1 + \theta_2\, x + \ldots + \theta_p\, x^{p-1}$, and the estimator takes a particularly simple form. Since $\dot{\tau}(x \,|\, \theta) = (1, x, \ldots, x^{p-1})^T$, the equation $S(\theta) = 0$ has the form $M\theta - V = 0$, where $M = (M_{i,j})_{1 \leq i,j \leq p}$ is a $p \times p$ matrix with components equal to $M_{ij} = \int \widetilde{d}^2(x) x^{i+j-2}\, \omega(x)\, dx$, and $V = (V_1, \ldots, V_p)^T$ is a p-vector whose components equal $V_j = \int \big\{(\widetilde{r_2 d})(x) - \widetilde{r}_1{}^2(x)\big\} x^{j-1}\, \omega(x)\, dx$. Thus, as long as M is invertible, we can write our estimator in the familiar form $\widehat{\theta} = M^{-1}V$. Note that, although the formula does not depend explicitly on the order q of the local polynomial estimators of m and g, the estimator $\widehat{\theta}$ depends on q through $\widetilde{r}_2$, $\widetilde{d}$ and $\widetilde{r}_1$.

As in the nonparametric case, once we have obtained the estimator $\widehat{\theta}$, we need to correct for negativity of the variance estimator $\tau(x;\widehat{\theta})$. As in section 2, we can do that in at least two ways. The first, simplest way is to take $\max\{0,\tau(x;\widehat{\theta})\}$. The drawback of this approach is that it projects negative values to zero in a rather abrupt way. An alternative and smoother way of correcting for negativity is to use the resampling procedure of section 2.2, that is to take

$$\bar{\tau}(x;\widehat{\theta}) = E\big[\max\big\{\tau(x;\widehat{\theta}^\dagger),0\big\}\,\big|\,\mathcal{D}\big], \tag{3.4}$$

where $\widehat{\theta}^\dagger$ denotes the version of $\widehat{\theta}$ computed from a resample of size n, drawn by sampling randomly, with replacement, from $\mathcal{D}$. Although we will not study theoretical properties of this estimator, it can be proved, as in the nonparametric case, that it is first-order equivalent to the estimator $\tau(x;\widehat{\theta})$. For a numerical comparison of the two ways to correct for negativity, see Figure 4 in section 6.

3.3 Motivation of the estimator

To understand the motivation for our estimator, first consider estimating θ by the vector which minimises the following least-squares criterion:

$$A_1(\theta) = \int \{\widehat{\tau}(x) - \tau(x\,|\,\theta)\}^2\, v_1(x)\, dx\,, \tag{3.5}$$

where v_1 is a weight function. In its most general form at (3.5), the least-squares distance $A_1(\theta)$ is simple to understand, but it involves the ratio of random variables, which is not particularly attractive. To overcome this problem, take $v_1 = \widehat{d}^4\, v_2$ for a function v_2. Then, recalling that $\widehat{\tau} = \widehat{m} - \widehat{g}^2$, where $\widehat{g} = \widehat{r}_1/\widehat{d}$ and $\widehat{m} = \widehat{r}_2/\widehat{d}$, (3.5) becomes

$$A_2(\theta) = \int \Big\{\widehat{r}_2(x)\,\widehat{d}(x) - \widehat{r}_1^2(x) - \tau(x\,|\,\theta)\,\widehat{d}^2(x)\Big\}^2\, v_2(x)\, dx\,, \tag{3.6}$$

which no longer involves a ratio. Next we take the diagonal-free versions of $\widehat{r}_2\,\widehat{d}$, $\widehat{r}_1^2$ and $\widehat{d}^2$ (it can be proved, employing arguments similar to those we use in our proofs, that this improves the theoretical properties of the resulting parametric estimator), so that (3.6) becomes

$$A_2(\theta) = \int \Big\{(\widetilde{r_2 d})(x) - \widetilde{r_1}^2(x) - \tau(x\,|\,\theta)\,\widetilde{d^2}(x)\Big\}^2\, v_2(x)\, dx\,.$$

To find the value of θ that minimizes $A_2(\theta)$, it remains to differentiate $A_2(\theta)$ with respect to the vector θ. Proceeding that way, we get

$$\int \Big\{ \widetilde{(r_2 d)}(x) - \widetilde{r_1}^2(x) - \tau(x \,|\, \theta)\, \widetilde{d}^2(x) \Big\}\, \widetilde{d}^2(x)\; \dot{\tau}(x \,|\, \theta)\, v_2(x)\, dx = 0\,.$$

Defining $\omega = v_2\, \widetilde{d}^2$ we deduce that θ solves $S(\theta) = 0$, where $S(\theta)$ is given by (3.1).

To appreciate why removing diagonal terms can improve performance it is instructive to consider a much simpler problem, where we wish to estimate $\psi \equiv E\{f(X)\}$ (with X denoting a random variable with density f), using data $X_1, \ldots, X_n$ drawn from the distribution with density f. One approach would be to construct a conventional kernel density estimator, $\widehat{f}$, evaluate it at X_i, and average this quantity over $i = 1, \ldots, n$. It is readily seen that the diagonal terms contribute an amount $\psi' \equiv K(0)/nh$ to this estimator, where K denotes the kernel function and h is the bandwidth. Of course, ψ' bears no relationship to the value of ψ, and if this term is removed then the performance of the estimator is improved. The same phenomenon is observed in a number of other problems, including the one treated in our paper: to first order, diagonal terms contribute only to bias, and their removal improves performance. In the case of our problem we obtain root-n consistency if the diagonal terms are dropped, but not otherwise.

4 Theoretical properties

4.1 Properties of $\widetilde{\tau}$, defined in (2.1)

Properties of our nonparametric estimator $\widetilde{\tau}$ at (2.1), using qth order local polynomial estimators of m and g, follow easily from the results of Delaigle, Fan and Carroll (2009). As usual in deconvolution problems, the asymptotic behaviour of the estimator depends on the type of error that contaminates the data. Generally a distinction is made between ordinary smooth and supersmooth errors. The latter are such that the characteristic function ϕ_U decreases exponentially fast in the tails, and for these errors it is well known that estimators converge at slow logarithmic rates. For the sake of brevity we give only properties of our estimator in the ordinary smooth case, where ϕ_U decreases polynomially fast in the tails. That is, we assume that the error density f_U is such that ϕ_U satisfies

$$d_0\,(1+|t|)^{-\alpha} \le |\phi_U(t)| \le d_1\,(1+|t|)^{-\alpha} \text{ for all } t \in \mathbb{R}, \tag{4.1}$$

for constants $d_1 \ge d_0 > 0$ and $\alpha > 1/2$. Properties of our estimator in the supersmooth case can be derived easily from Delaigle, Fan and Carroll (2009).

We assume the same regularity conditions as Delaigle, Fan and Carroll (2009). More precisely, let $\tau_2(x) = \text{var}(Y^2|X=x)$ and $\phi_X(t) = E(e^{itX})$. We assume that the following conditions are satisfied:

Condition A:

(A1) $\phi_U(t) \neq 0$ for all t;

(A2) $h \to 0$ and $nh \to \infty$ as $n \to \infty$;

(A3) $\int |\phi_X| < \infty$ and f_X is twice differentiable and $\|f_X^{(j)}\|_\infty < \infty$ for $j = 0, 1, 2$;

(A4) τ and τ_2 are bounded; m and g are $q+3$ times differentiable such that, for $j = 0, \ldots, q+3$, $||m^{(j)}||_\infty < \infty$ and $||g^{(j)}||_\infty < \infty$; and, for some $\eta > 0$, $E\big\{|Y_i - g(x)|^{2+\eta}|X=u\big\}$ and $E\big\{|Y_i^2 - m(x)|^{2+\eta}|X=u\big\}$ are bounded for all u;

(A5) K is a real and symmetric kernel such that $\int K(x)\,dx = 1$ and has finite moments of order $2q+3$; for $k = 0, \ldots, 2q+1$, $||\phi_K^{(k)}||_\infty < \infty$ and $\int \big\{|t|^\alpha + |t|^{\alpha-1}\big\}\,|\phi_K^{(k)}(t)|\,dt < \infty$; and, for $0 \le k, k' \le 2q$, $\int |t|^{2\alpha}|\phi_K^{(k)}(t)| \cdot |\phi_K^{(k')}(t)|\,dt < \infty$ and $\phi_K^{(k)}$ is not identically zero.

Asymptotic properties of our estimator are given in the next theorem. The proof is omitted since it follows from Delaigle, Fan and Carroll (2009).

Theorem 4.1. *Assume that the errors satisfy* (4.1). *Under Condition A, for each x for which $f_X(x) > 0$,*

(i) if q is even and $h =$ const. $n^{-1/(2\alpha+2q+5)}$, *then*

$$\widetilde{\tau}(x) = \tau(x) + O_p\big(n^{-(q+2)/(2\alpha+2q+5)}\big)\,; \tag{4.2}$$

(ii) if q is odd and $h =$ const. $n^{-1/(2\alpha+2q+3)}$, *then*

$$\widetilde{\tau}(x) = \tau(x) + O_p\big(n^{-(q+1)/(2\alpha+2q+3)}\big)\,. \tag{4.3}$$

Although the asymptotic rates given by the theorem improve as we increase q, in practice increasing q implies an increase in the variance of the estimator, and the versions of the local polynomial estimator that work the best are the local constant and the local linear ones. In our numerical work we use the local linear version of the estimator.

Note that Theorem 4.1 describes the behaviour of the estimator at (2.1) in the case where the design density is continuous; in this context, the rates are the same (i.e. $n^{-2/(2\alpha+5)}$) whether we use the local constant estimator ($q = 0$, corresponding to the estimator of Fan and Truong, 1993) or the local linear estimator ($q = 1$) to estimate m and g. In the case where f_X is compactly supported and is not continuous at the boundary of its support, these rates deteriorate to $\widetilde{\tau}(x) = \tau(x)+O_p(n^{-1/(2\alpha+3)})$ in the local constant case and remain $\widetilde{\tau}(x) = \tau(x) + O_p(n^{-2/(2\alpha+5)})$ in the local linear case.

Remark 1. *As already noted in Delaigle, Fan and Carroll (2009), as usual in nonparametric smoothing, many variants of these theoretical results exist. For example, in the local constant case we could use high order kernels, or even the infinite order sinc kernel. When f_X, m and g, and their relevant derivatives, are continuous on the whole real line, the sinc kernel has the advantage that it adapts automatically to the smoothness of the curves, in the sense that it produces an estimator with bias determined by the level of smoothness of the curves rather than by the kernel. See e.g. Diggle and Hall (1993) and Comte and Taupin (2007). However, when the curves have boundary points, the sinc kernel loses its theoretical advantages. In practice, the sinc kernel tends to suffer from problems such as the Gibbs phenomenon.*

4.2 Properties of $\bar{\tau}$, defined in (2.5)

Under sufficient assumptions, it can be proved that the estimators $\widetilde{\tau}$ and $\bar{\tau}$ are first-order equivalent. We give the conditions and state the result for the case where m and g are estimated by a local constant estimator ($q = 0$). The arguments can be extended to the more general version of the estimator where $\widehat{\tau}$ is based on qth order local polynomial estimators with $q \geq 1$. Assume that:

$\bar{\tau}$, in (2.5), is computed as $\bar{\tau} = B^{-1} \sum_b \max(\widehat{m}_b^\dagger - \widehat{g}_b^{\dagger 2}, 0)$, where $\widehat{m}_b^\dagger$ and $\widehat{g}_b^\dagger$ are both computed from $\mathcal{D}_b^\dagger$ for $1 \le b \le B$; $\mathcal{D}_1^\dagger, \ldots, \mathcal{D}_B^\dagger$ are resamples of size n drawn by sampling randomly, with replacement, from $\mathcal{D}$; the B resamples are independent, conditional on $\mathcal{D}$; and $B = B(n)$ diverges with n at such a rate that, for all sufficiently large n, $B \le n^{C_2}$ where $0 < C_2 < \infty$. (4.4)

Assume too that

$$|\phi_U'(t)|/|\phi_U(t)| \le C_1 \, (1 + |t|)^{-1} \, , \tag{4.5}$$

where C_1 denotes an arbitrarily large positive constant; this condition generalizes condition $A_{m,l}$(i) of Fan (1991a). Under these assumptions, the following theorem holds. A proof is given in appendix A.2.

Theorem 4.2. *Under the conditions of Theorem 4.1, if* (4.4) *and* (4.5) *hold, then at each point x for which $f_X(x)\,\tau(x) > 0$, we have*

$$\widetilde{\tau}(x) - \bar{\tau}(x) = o_p\big(n^{-2/(2\alpha+5)}\big) \, . \tag{4.6}$$

Theorem 4.2 shows that the estimators $\widetilde{\tau}$ and $\bar{\tau}$ are first-order equivalent, since the rate at (4.6) is faster than that rate at (4.2). It can also be proved that in cases where $f_X(x) > 0$ but $\tau(x) = 0$, $\bar{\tau}(x)$ generally has higher asymptotic bias than $\widetilde{\tau}(x)$, although smaller asymptotic variance. In this setting the distributions of $\widetilde{\tau}(x)$ and $\bar{\tau}(x)$ are not asymptotically normal.

Remark 2. *Condition* (4.4) *implies that B is no more than polynomially large as a function of n. This restriction is imposed to ensure that very unusual resamples, for example resamples that consist only of a single data value, arise only with particularly small probability. It protects against aberrations that would affect first-order properties of $\bar{\tau}$ when, for example, the resample is such that the denominator of $\bar{\tau}$ gets too close to zero. The condition on B could be avoided by introducing a ridge parameter in the denominator of $\bar{\tau}$.*

4.3 Theoretical properties of the parametric estimator

Under sufficient regularity conditions, the parametric estimator introduced in section 3 has the standard parametric root-n convergence rates, despite the fact that some quantities involved can be estimated only nonparametrically. As in the previous section, due to the complexity of the arguments in the general local polynomial case, we state the conditions and results in the local constant case, where $q = 0$.

Let $\| \cdot \|$ denote the usual Euclidean metric on p-variate space, let θ_0 be the true value of θ, write f_X for the density of the design variable X in (1.1), and define the $p \times p$ matrix

$$M_0 = \int f_X^2(x)\, \dot\tau(x \,|\, \theta_0)\, \dot\tau(x \,|\, \theta_0)^{\mathrm{T}}\, \omega(x)\, dx\,. \tag{4.7}$$

We assume the following conditions:

Condition B:

(B1) the weight function ω, in (3.1), is bounded, nonnegative and vanishes outside a compact set;

(B2) K is bounded and symmetric, ϕ_K is compactly supported, $\int (1+|x|^\kappa)\, |K(x)|\, dx < \infty$ where κ is a positive integer, $\int K = 1$ and $\int x^j\, K(x)\, dx = 0$ for $1 \le j < \kappa$;

(B3) the variance model $\tau(x \,|\, \theta)$ has $d_1 + 2 \ge 4$ derivatives with respect to θ, where each derivative is bounded uniformly in x in the support of ω, and in θ such that $\|\theta - \theta_0\| \le C$, for some $C > 0$;

(B4) for an integer $d_2 \ge 1$, each of the functions f_X, g, $\tau(\cdot \,|\, \theta_0)\, \omega$ and $\dot\tau(x \,|\, \theta_0)\, \omega$ has $\max(d_2, \kappa)$ derivatives, uniformly bounded on compacts; and each of the functions f_X, $f_X\, g$, $f_X\, m$, $\tau(\cdot \,|\, \theta_0)\, \omega$ and $\dot\tau(x \,|\, \theta_0)\, \omega$ has $\max(d_2, \kappa)$ absolutely integrable derivatives, where integration is over the whole real line;

(B5) $|\phi_U(t)| \ge \text{const.}\, (1 + |t|)^{-\alpha}$ for all real t, where $0 < \alpha < d_2 - \frac{1}{2}$, and $\sup f_U < \infty$;

(B6) $E\{g^4(X)\} + E\{\tau^2(X)\} < \infty$ and $E(\varepsilon^4) < \infty$;

(B7) the $p \times p$ matrix M_0, in (4.7), is nonsingular;

(B8) for κ as in (B2), and ϵ_n denoting a positive sequence such that $n^{1/2} \epsilon_n \to \infty$ as $n \to \infty$, the following properties hold: $h = h(n) \to 0$ as $n \to \infty$, $h^\kappa = o(n^{-1/2})$, $n^{-1}\, h^{-2(1+2\alpha)} \to 0$ as $n \to \infty$, and $\epsilon_n^{d_1}\, h^{-(1+2\alpha)} \to 0$ as $n \to \infty$.

Note that, for each $\alpha > 0$ and each sequence ϵ_n, we may choose κ (the order of the kernel, K; see (B2)) and d_1 and d_2 (which determine the smoothness of the model and of the weight function ω; see (B3) and (B4)) so large that (B8) holds for bandwidths h that enjoy a variety of different orders of magnitude, and such that the condition $\alpha < d_2 - \frac{1}{2}$ in (B5) obtains. Under these assumptions, the next theorem shows that our parametric estimator has the usual $\sqrt{n}$ rate of convergence. Its proof is given in appendix A.3.

Theorem 4.3. *Assume that Condition B holds. Then: (i) With probability converging to 1 as $n \to \infty$, there exists at least one solution $\widehat{\theta}$ of the equation $S(\theta) = 0$ satisfying $\|\widehat{\theta} - \theta_0\| \leq \epsilon_n$, where $S(\theta)$ is as at (3.1). (ii) There exists a positive semi-definite, symmetric matrix Σ such that, for any such solution, $n^{1/2}\,(\widehat{\theta} - \theta_0)$ is asymptotically normal* $\mathrm{N}(0, \Sigma)$.

Under stronger conditions than those imposed in the theorem it can be proved that, with probability converging to 1, a solution of $S(\theta) = 0$ exists and is unique. However, even with the assumptions in Theorem 4.3, any of the solutions identified there has the same first-order properties as any other, and so none is preferable to any other in a first-order sense.

The covariance matrix Σ is identified in Step 7 of the proof in section A.3. In the particular case where the variance function τ is a polynomial, where (with probability 1) the equation $S(\theta) = 0$ has a unique solution, part (i) of the theorem is not relevant. Part (ii), where $\widehat{\theta}$ is taken to be the uniquely defined estimator, holds under conditions B if assumption (B3) is dropped and if the constraint $\epsilon_n^{d_1}\, h^{-(1+2\alpha)} \to 0$ is removed from (B8).

5 Bandwidth selection

As for any smoothing method, the success of our estimators relies heavily on using an appropriate smoothing parameter. Data-driven bandwidth selection in errors-in-variables regression is particularly challenging, and the approach suggested here is based on bootstrap methods and the simulation-extrapolation algorithm (Cook and Stefanski,1994; Stefanski and Cook, 1995). It has points of contact with a method developed by Delaigle and Hall (2008a) in a different setting. The main similarity is that we borrow the SIMEX method, but there are more than a few dissimilarities because, in the current problem, we are estimating a variance function rather than a regression mean.

We develop two new simulation-extrapolation type bandwidth selectors, based on estimating the mean integrated squared error, denoted by MISE, and the mean squared error, or MSE, of estimators at higher levels of errors. Given the difficulty of developing bandwidth procedures in the errors-in-variables context, our new bandwidth selectors are of independent interest. They can be applied to other errors-in-variables problems.

5.1 Bandwidths for the estimators in section 2

Let h_{opt} denote the bandwidth that minimizes weighted mean integrated squared error, MISE $= E\int(\widehat{\tau}-\tau)^2\,w$, where w is a weight function. Estimating h_{opt} by directly attempting to estimate the MISE would be very difficult, so we develop an alternative approach. The idea is to create samples which contain higher levels of errors, develop estimators of bandwidths associated with two corresponding variance estimation problems, and then, using the relation that exists among the various levels of errors, deduce an estimator of h_{opt}. Higher-level versions of the variance problem are created as follows:

1. Generate a sample $U_1^\star,\ldots,U_n^\star$ from the error density f_U, and construct the sample $W_1^\star,\ldots,W_n^\star$ where $W_j^\star = W_j + U_j^\star$ for $j=1,\ldots,n$.

2. Generate a sample $U_1^{\star\star},\ldots,U_n^{\star\star}$ from f_U, and construct the sample $W_1^{\star\star},\ldots,W_n^{\star\star}$ where $W_j^{\star\star} = W_j^\star + U_j^{\star\star}$ for $j=1,\ldots,n$.

3. Define the variance functions $\tau^\star = m^\star - (g^\star)^2$ and $\tau^{\star\star} = m^{\star\star} - (g^{\star\star})^2$ corresponding to the new data, where $g^\star(x) = E(Y\,|\,W=x)$, $m^\star(x) = E(Y^2\,|\,W=x)$, $g^{\star\star}(x) = E(Y\,|\,W^\star=x)$ and $m^{\star\star}(x) = E(Y^2\,|\,W^\star=x)$. Let $\widehat{\tau}^\star$ and $\widehat{\tau}^{\star\star}$ denote the deconvolution estimators of $\tau^\star$ and $\tau^{\star\star}$ from the contaminated data $(W_j^\star, Y_j)$ and $(W_j^{\star\star}, Y_j)$, respectively, and let $h^\star_{\text{opt}}$ and $h^{\star\star}_{\text{opt}}$ be the bandwidths that minimize MISE$^\star = E\int(\widehat{\tau}^\star-\tau^\star)^2\,w^\star$ and MISE$^{\star\star} = E\int(\widehat{\tau}^{\star\star}-\tau^{\star\star})^2\,w^{\star\star}$, respectively, where $w^\star$ and $w^{\star\star}$ are weight functions.

Unlike the original problem, in these two problems with higher levels of errors the "measurement error-free data," (W,Y) and $(W^\star,Y)$ respectively, are available, and thus we can construct standard measurement error-free, difference-based estimators

$\widehat{\tau}_D^\star$ and $\widehat{\tau}_D^{\star\star}$ of $\tau^\star$ and $\tau^{\star\star}$; see section 6.1 for details, and see Rice (1984), Buckley *et al.* (1988), Hall *et al.* (1990), Müller and Stadtmüller (1992) and Seifert *et al.* (1993) for discussion of that method. Being based on a conventional regression problem with no errors in variables, these estimators converge to the correct values at a much faster rate than do $\widehat{\tau}^\star$ and $\widehat{\tau}^{\star\star}$, and so can be used, to first order, to represent the "truth" in a model for the more difficult, errors-in-variables regression problem for which $\widehat{\tau}^\star$ and $\widehat{\tau}^{\star\star}$ were computed. With this in mind we estimate $\text{MISE}^\star$ and $\text{MISE}^{\star\star}$ by $\widehat{\text{ISE}}^\star(h) = \int\{\widehat{\tau}_D^\star(x) - \widehat{\tau}^\star(x;h)\}^2\, w^\star(x)\, dx$ and $\widehat{\text{ISE}}^{\star\star}(h) = \int\{\widehat{\tau}_D^{\star\star}(x) - \widehat{\tau}^{\star\star}(x;h)\}^2\, w^{\star\star}(x)\, dx$. To avoid too strong dependence of the particular resamples generated, we repeat steps 1 and 2 B times, to generate B resamples; we calculate $\widehat{\text{ISE}}^\star$ and $\widehat{\text{ISE}}^{\star\star}$ for each of the B samples, to obtain $\widehat{\text{ISE}}_b^\star$ and $\widehat{\text{ISE}}_b^{\star\star}$, $b = 1, \ldots, B$; and we take $\widehat{\text{MISE}}^\star = B^{-1}\sum_b \text{ISE}_b^\star$ and $\widehat{\text{MISE}}^{\star\star} = B^{-1}\sum_b \text{ISE}_b^{\star\star}$.

From there, to obtain an estimator of h_{opt}, the idea, which we borrow from the simulation-extrapolation algorithm, is that $W^{\star\star}$ measures $W^\star$ in the same way that $W^\star$ measures W and W measures X, so that we can expect the relation between $h_{\text{opt}}^{\star\star}$ and $h_{\text{opt}}^\star$ to be similar to the relation between $h_{\text{opt}}^\star$ and h_{opt}, that is $h_{\text{opt}}^{\star\star}/h_{\text{opt}}^\star \approx h_{\text{opt}}^\star/h_{\text{opt}}$. Motivated by these ideas, we propose estimating h_{opt} by $\widehat{h}_{\text{opt}} = (\widehat{h}_{\text{opt}}^\star)^2/\widehat{h}_{\text{opt}}^{\star\star}$. This last step relies on the fact that $\widehat{h}_{\text{opt}}^\star$ and $\widehat{h}_{\text{opt}}^{\star\star}$ are asymptotic to constant multiples of the order of the optimal bandwidth, and the ratio $(\widehat{h}_{\text{opt}}^\star)^2/\widehat{h}_{\text{opt}}^{\star\star}$ is also asymptotic to that order. See also Remark 3. Rigorous theoretical justification can be obtained using arguments similar to Delaigle and Hall (2008a,b). Practical implementation is illustrated in section 6.

Remark 3. *(Justification of bandwidth-choice rule). Note that both $\widehat{h}_{\text{opt}}^\star$ and $\widehat{h}_{\text{opt}}^{\star\star}$ are selected to minimise mean integrated squared errors in simulated errors-in-variables problems. Since, by construction, the latter problems share the same values of α and q as the original one, they enjoy the same rates of convergence, $n^{-1/(2\alpha+2q+3)}$ if q is odd and $n^{-1/(2\alpha+2q+5)}$ if q is even, of the optimal bandwidth. Therefore, the ratio $(\widehat{h}_{\text{opt}}^\star)^2/\widehat{h}_{\text{opt}}^{\star\star}$ also has this rate. This was established by Delaigle and Hall (2008a,b) to be the case in a related setting, and indeed the property is at the heart of the widely used SIMEX method for solving deconvolution problems. The constant multiplier will* *generally not be the optimal one, and in fact, obtaining the optimal constant seems to be an especially challenging empirical problem, perhaps without a practicable solution. However, the constant determined by the ratio $(\widehat{h}^{\star}_{\text{opt}})^2/\widehat{h}^{\star\star}_{\text{opt}}$ seems to be satisfactory in many settings.*

5.2 Bandwidths for the estimator in section 3

Using ideas similar to those in the previous section, we suggest choosing the bandwidth required to calculate $\widehat{\theta}$ as follows. For $b = 1, \ldots, B$, the steps are as follows:
1–2: same as in section 5.1.
3. Define the variance functions $\tau^{\star}$ and $\tau^{\star\star}$ as in section 5.1. Let $\tau(\cdot \,|\, \widehat{\theta}^{\star})$ and $\tau(\cdot \,|\, \widehat{\theta}^{\star\star})$ denote the parametric deconvolution estimators of $\tau^{\star}$ and $\tau^{\star\star}$ from the contaminated data $(W^{\star}_j, Y_j)$ and $(W^{\star\star}_j, Y_j)$, respectively.
4. Let $\tau(\cdot \,|\, \widehat{\theta}^{\star}_D)$ and $\tau(\cdot \,|\, \widehat{\theta}^{\star\star}_D)$ denote measurement error-free, difference-based parametric estimators of $\tau^{\star}$ and $\tau^{\star\star}$, based on the data (W_j, Y_j) and $(W^{\star}_j, Y_j)$, respectively.
5. Find the bandwidths $\widehat{h}^{\star}_{\text{opt}}$ and $\widehat{h}^{\star\star}_{\text{opt}}$ that minimise $B^{-1}\sum_b \text{ISE}^{\star}_b$ and $B^{-1}\sum_b \text{ISE}^{\star\star}_b$, where $\text{ISE}^{\star} = \int \{\tau(\cdot \,|\, \widehat{\theta}^{\star}) - \tau(\cdot \,|\, \widehat{\theta}^{\star}_D)\}^2 \, w^{\star}$ and $\text{ISE}^{\star\star} = \int \{\tau(\cdot \,|\, \widehat{\theta}^{\star\star}) - \tau(\cdot \,|\, \widehat{\theta}^{\star\star}_D)\}^2 \, w^{\star\star}$, respectively, where $w^{\star}$ and $w^{\star\star}$ are weight functions. Take $\widehat{h}_{\text{opt}} = (\widehat{h}^{\star}_{\text{opt}})^2/\widehat{h}^{\star\star}_{\text{opt}}$.

6 Numerical properties

6.1 Details of implementation

For all methods, every nonparametric estimator used anywhere in the estimation procedure (to calculate the bandwidth and to calculate the estimator itself, and for our nonparametric estimator as well as for the nonparametric difference-based estimator) was a local-linear estimator (that is, we took $q = 1$ everywhere). For the bandwidth selectors of section 5, we took $w^{\star} = w^{\star\star} = 1_{[q^W_{0.025}, q^W_{0.975}]}$, with q^T_{α} denoting the αth empirical quantile of a variate T and $1_{[a,b]}$ the indicator function of the interval $[a, b]$. For the method of section 5.1, we used the nonparametric difference-based estimator with cross-validation bandwidth, constructed from the data $(W_{[i]}, D_i)_{1 \le i \le n-1}$ where $D_i = 0.5\,(Y_{[i]} - Y_{[i+1]})^2$, with $[i]$ denoting the index of the ith order statistic of W.

We used the same approach for the * data. For the bandwidth selector of section 5.2 we used the parametric version of this difference-based estimator. To speed up calculations, all nonparametric estimators used to calculate bandwidths were computed after binning the data.

For the kernel K in our nonparametric procedures we used the one suggested by Delaigle, Fan and Carroll (2009), that is, we took the kernel with Fourier transform $\phi_K(t) = (1-t^2)^8\, 1_{[-1,1]}(t)$. See Delaigle and Hall (2006) for discussion of kernels in deconvolution problems. For the parametric method of section 3.2 we took $\omega(x) = \sigma^{-1}L\{(x-\mu)/\sigma\}\,/\widehat{d}(x)$, with L the biweight kernel $L(x) = 15/16(1-x^2)^2 \cdot 1_{[-1,1]}(x)$, $\mu = (q^W_{0.01} + q^W_{0.99})/2$, $\sigma = (q^W_{0.99} - q^W_{0.01})/2$ and $\widehat{d}$ as in section 3.2.

6.2 Simulation settings

We applied our nonparametric estimators to several regression models. In each case we generated 200 samples from model (1.1), using one the following variance functions (listed in increasing order of complexity):

- $\tau_1(x) = 1$;
- $\tau_2(x) = \max(0.9x + 0.6, 0)$;
- $\tau_3(x) = \max(1.5x + 0.1, 0)$;
- $\tau_4(x) = (1.5x + 0.1)^2$;
- $\tau_5(x) = 2x^2 - 2x + 0.75$,

which we combined with one of the following regression curves (also listed in increasing level of complexity):

- $g_1(x) = 0.75$;
- $g_2(x) = 1/\big(1 + \exp\{-5(x-1/2)\}\big)$;
- $g_3(x) = 1/\big(1 + \exp\{-10(x-1/2)\}\big)$;
- $g_4(x) = 1/\big(1 + \exp\{-5(x-1/2)\}^2\big)$;
- $g_5(x) = 0.45\sin(2\pi x) + 0.5$.

In each case we took $\varepsilon \sim \mathrm{N}(0,1)$ and $X \sim \mathrm{N}(0.5, 0.1)$ or $X \sim U[0,1]$. Finally, we took U to be Laplace.

We calculated our estimators for each generated sample. To illustrate the importance of taking the error into account we also calculated naive estimators, that is, estimators that pretend there is no error in the data. We also calculated ideal estimators, that is, estimators which use the non contaminated observations X_i. Of course, these estimators are not available in practice, but they illustrate the impact that measurement errors can have on the quality of estimators. To summarize, in our numerical work we calculated the following estimators:

(1) Our nonparametric local linear estimator, which we denote by NPE;

(2) Our parametric estimator, which we denote by PE;

(3) The local linear difference-based nonparametric estimator based on the data $(W_{[i]}, D_i)_{1\le i\le n-1}$, which we denote by naive NPE;

(4) The naive parametric difference-based estimator, which is the parametric version of the naive difference-based method and which we denote by naive DBPE. This method is often used in practice because it has good theoretical properties and does not need a bandwidth;

(5) The naive version of our parametric estimator, obtained by using the data (W_i, Y_i) but setting $U \equiv 0$ everywhere else in the formulae of our estimator. We refer to this as the naive PE;

(6) The ideal parametric difference-based estimator, which is the same as the naive DBPE, except that we use the measurement error-free data $(X_{[i]}, D_i)_{1\le i\le n-1}$. We refer to this as the ideal DBPE;

(7) The ideal version of our parametric estimator, obtained by using the data (X_i, Y_i) and setting $U \equiv 0$ everywhere in the formulae of our estimator. We refer to this as the ideal PE.

Note that, although the DBPE is widely used in the measurement error-free case, partly because of its simplicity and also because it does not need a bandwidth, our results showed that in a high proportion of regression models the measurement error-free version of our estimator (i.e. ideal PE) worked better than its difference-based counterpart (ideal DBPE). Similarly, we found that the naive DBPE often gave better results than the naive PE. This complicates the comparison between our estimator and the naive methods, as in practice we would not know which of the naive DBPE and the naive PE is the best estimator. Thus, comparing our method in each case

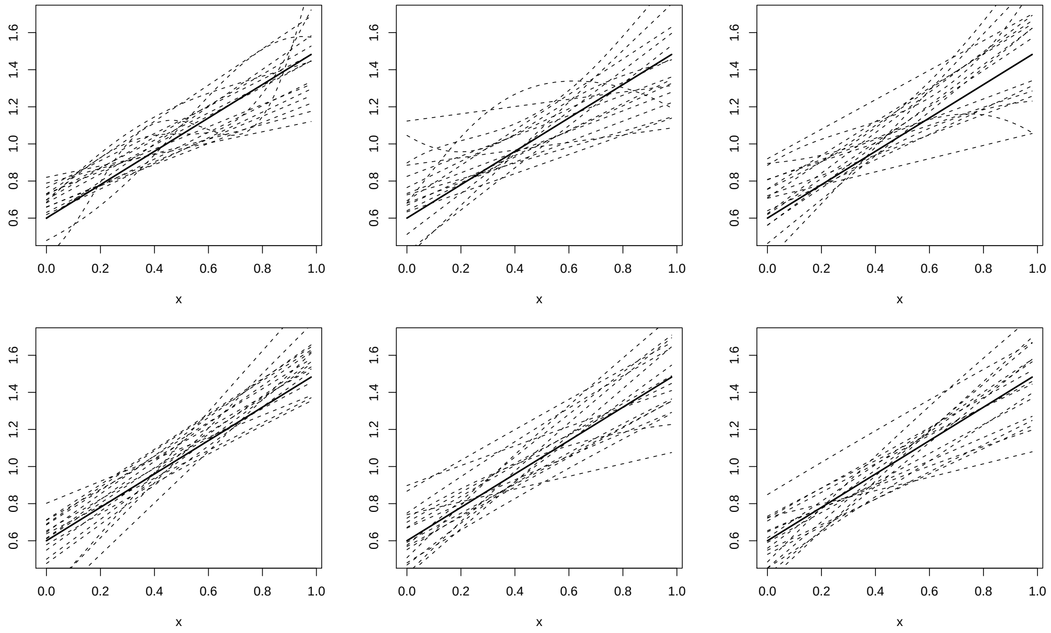

Figure 1: Estimation of τ_2 when $g = g_4$, $X \sim \mathrm{N}(0.5, 0.1)$. Top: naive NPE, bottom: our NPE. Left: $(n, \mathrm{NSR}) = (500, 0.2)$, middle: $(n, \mathrm{NSR}) = (250, 0.2)$, right: $(n, \mathrm{NSR}) = (250, 0.1)$.

with the best of the two naive methods systematically biases the comparison in favour of the naive estimators.

The performance of estimators, $\widehat{\tau}$ say, was calculated via integrated squared error, ISE $= \int_0^1 (\widehat{\tau} - \tau)^2$, except in the constant case $\tau = 1$ where we used squared error. In the figures we show the estimated curves corresponding to the quantiles $q_{0.1}, q_{0.15}, q_{0.2}, \ldots, q_{0.9}$ of the 200 calculated integrated squared errors. The true function τ is always represented by the thick solid curve.

6.3 Simulation results

In each figure, the goal is to illustrate one (or more) of the properties of the various estimators. Note that the findings discussed here were also supported by box plots. To keep this section to a reasonable length, we discuss these only briefly in the text, in cases where the graphs are not clear enough to compare the performance of the methods. A summary of the important properties is given at the end of this section.

Figure 1 illustrates the improvement one can get by taking the error into account

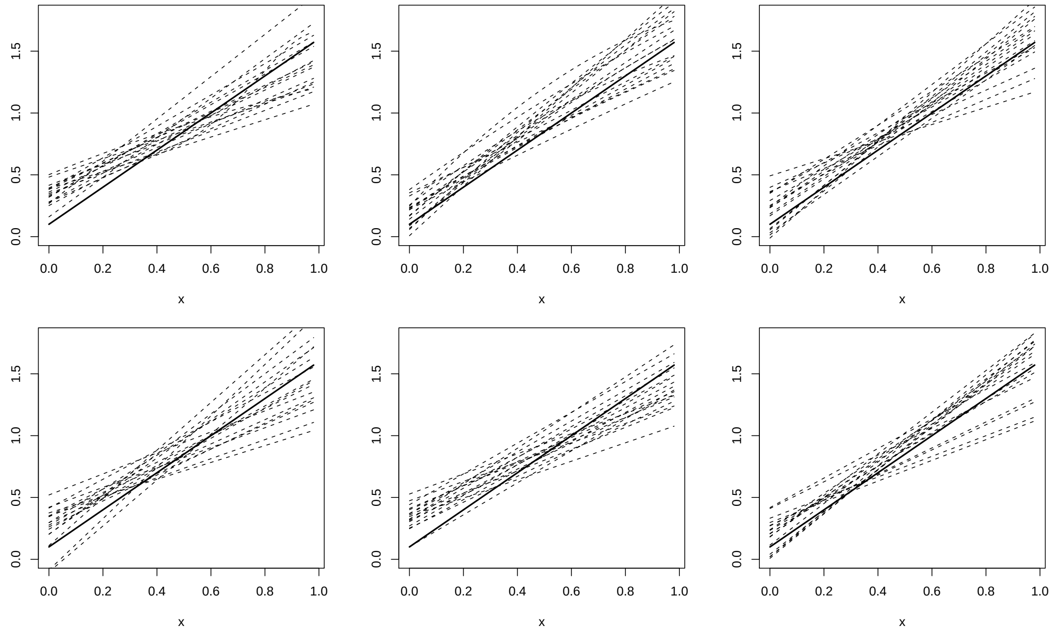

Figure 2: Quantile curves for the estimation of τ_3 when $g = g_5$, $X \sim \mathrm{N}(0.5, 0.1)$, $n = 250$ and NSR $= 0.2$, using the naive NPE (top left), our NPE (top centre), our PE (top right), the naive DBPE (bottom left), the naive PE (bottom centre) or the ideal PE (bottom right).

when calculating the nonparametric estimators. We compare our NPE and the naive NPE, by showing the quantile curves for estimating the variance function τ_2, when the regression curve is g_4, $X \sim \mathrm{N}(0.5, 0.1)$, $n = 250$ or 500, and the noise to signal ratio NSR $\equiv \mathrm{var}(U)/\mathrm{var}(X)$ is equal to 10% or 20%. The graphs show a clear superiority of our estimator compared to the naive one. They also demonstrate that the estimator improves as the sample size increases and the NSR decreases.

Figure 2 shows the quantile curves when estimating τ_3, when $g = g_5$, $X \sim \mathrm{N}(0.5, 0.1)$, $n = 250$ and NSR $= 0.2$. Here, the goal is to compare all the estimators. We see that the results improve when using our NPE compared to the naive NPE, but also that our parametric estimator (PE) improves the NPE. The graphs also show that the naive parametric estimators (naive DBPE and naive PE) are either much more biased or much more variable than our PE. The quantile curves for the ideal PE shown here demonstrate that, in this case, the impact of measurement errors on the quality of our PE is not very severe (although it may not be clear from the

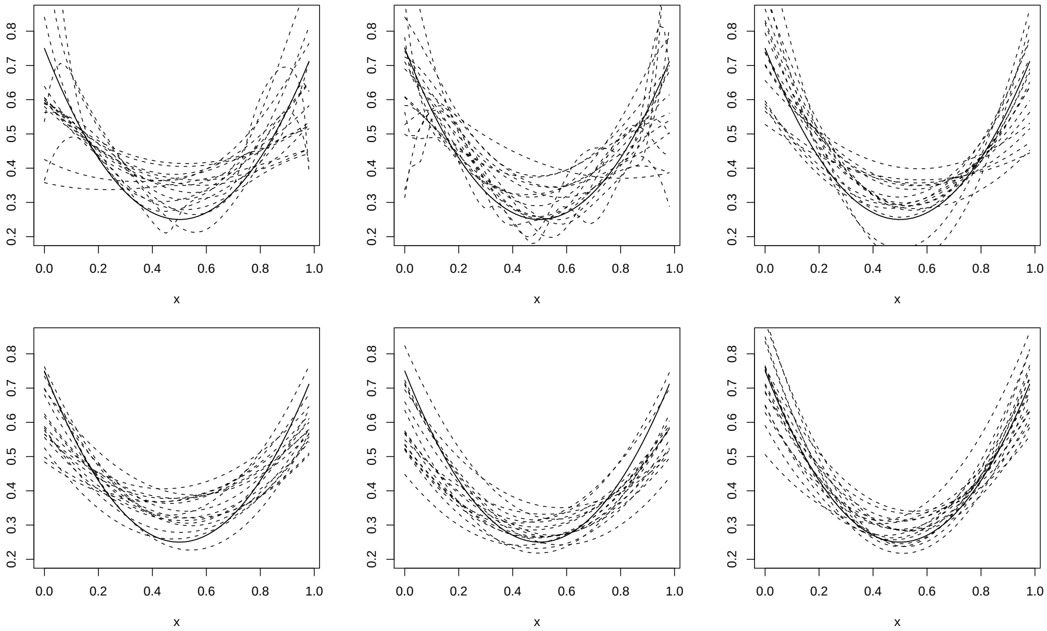

Figure 3: Quantile curves for estimation of τ_5, when $g = g_1$, $X \sim U[0, 1]$ and $n = 500$, using our NPE when NSR = 20% (top left); our in the case when NSR = 0.1, using our NPE (top center), our PE (top right), the naive PE (bottom left), the naive DBPE (bottom center) and the ideal PE (bottom right).

graph, the ideal estimator did give smaller ISEs than our PE).

Figure 3 illustrates the same properties as Figure 2, but this time for the case where the variance curve is the quadratic curve τ_5, and $g = g_1$. This case is quite difficult because of the valley in the shape of the variance curve, and estimators have a tendency to overestimate the valley. The estimators did not work very well for $n = 250$, and we show the results for $n = 500$ and NSR = 10%. In this case, the naive NPE worked so poorly that, instead of showing its quantile curves, we show those for our NPE when NSR = 20%. As above, we see that a smaller NSR implies a better estimator, our PE substantially improves our NPE, and ignoring the error (that is, using the naive estimators) results in estimators that are much more biased. In this difficult case, the impact of the measurement errors is very noticeable: the ideal PE is significantly better than our PE.

Figure 4 shows results for estimating τ_4 parametrically, when $g = g_4$, $X \sim$ N$(0.5, 0.1)$ and NSR = 0.2, for sample sizes $n = 250$ and $n = 500$. In this case

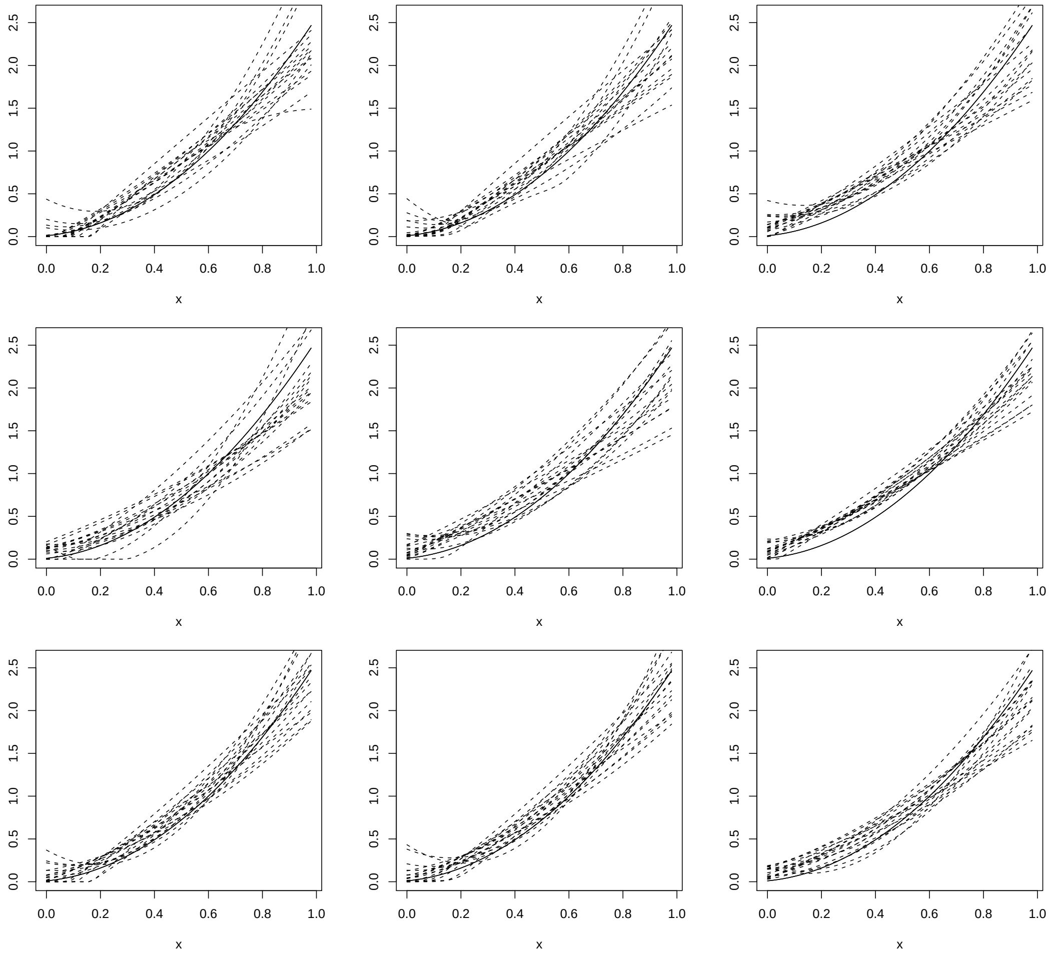

Figure 4: Estimation of τ_4 when $g = g_4$, $X \sim \text{N}(0.5, 0.1)$ and NSR = 0.2. Quantiles curves for the our PE when $n = 250$ (top left) or when $n = 500$ (bottom left), our resampling corrected PE when $n = 250$ (top center) or $n = 500$ (bottom center), the naive PE for $n = 250$ (top right) or $n = 500$ (middle right), the naive DBPE when $n = 250$ (middle left) or $n = 500$ (bottom right), or the resampling corrected DBPE when $n = 250$ (middle center).

the variance function takes values close to zero for x close to zero, and, as a result, the estimators of $\tau(x)$ often took negative values when x was close to zero. To correct for this problem we considered the two approaches discussed at the end of section 3.2. That is, we either truncated the estimator to zero or used (3.4), where the expectation was computed as the average of values computed from B resamples, as in (4.4). In the figure we show the results of both approaches. When using the second approach, we took $B = 100$ resamples. We can see that, overall, both approaches

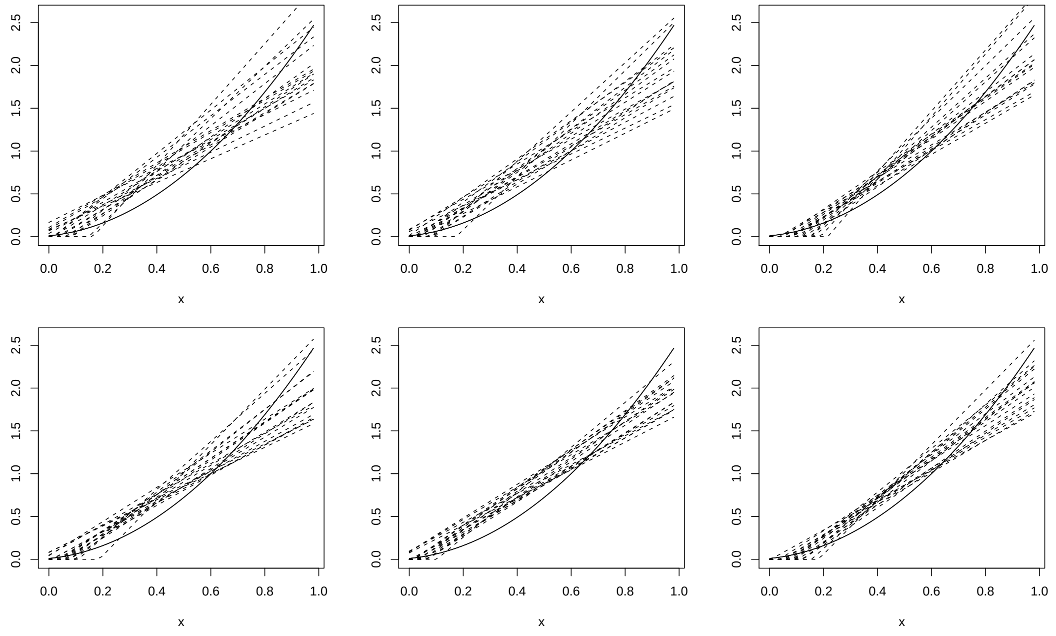

Figure 5: Estimation of τ_4 when $g = g_4$,, $X \sim \text{N}(0.5, 0.1)$ with NSR $= 0.2$, $n = 250$ (top) or $n = 500$ (bottom), and pretending that the variance is linear. Quantile curves for the naive DBPE (left), the naive PE (middle) and our PE (right).

to correcting for negativity gave similar results, but the resampling method did this correction in a smoother way. As usual, the figure also illustrates the improvement of our estimator as the sample size increases, and its superiority to the two naive parametric approaches (although the larger bias incurred by the naive estimators is more easily seen for the larger sample size, $n = 500$).

In Figure 5, we continue to consider parametric estimation of τ_4 when $g = g_4$, $X \sim \text{N}(0.5, 0.1)$ and NSR $= 0.2$, but this time we wrongly assume that τ is a linear curve. Our goal is to see whether, even when the variance model is misspecified, using an error-corrected estimator can improve on the naive estimators. In particular we want to see whether the line fitted by our PE will be closer to that fitted by the naive methods. Here, to correct for negativity, we simply truncated the fitted lines to zero. The plots of the quantile curves and the box plots (not shown here) both show that in this case, too, taking the error into account can bring significant improvement over the naive estimators, whose fitted lines are more biased than for our estimator.

Finally, in Figure 6 we show boxplots for estimating the constant variance τ_1,

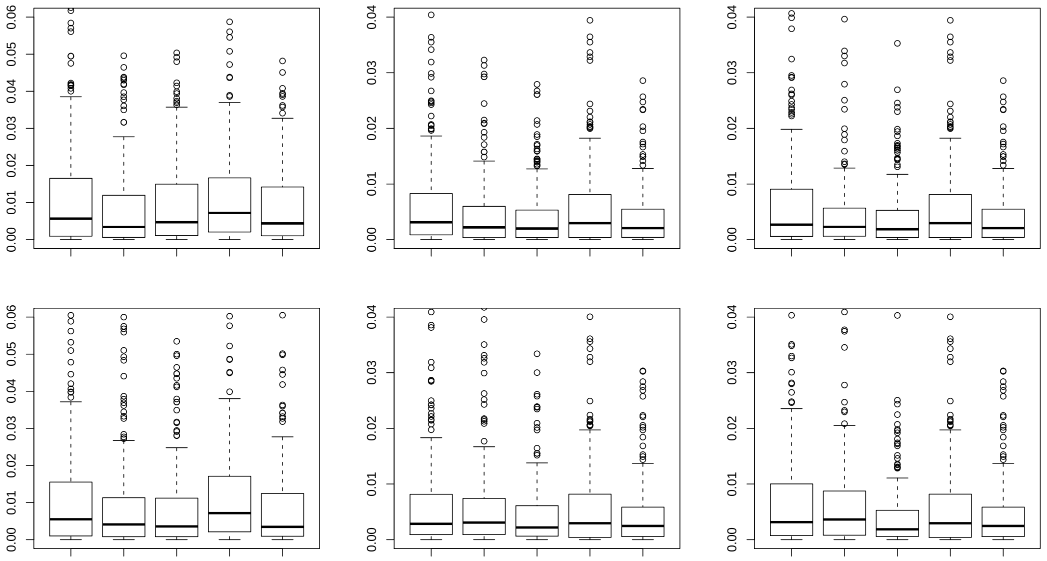

Figure 6: Boxplot for the estimation of τ_1 when $g = g_2$ (top) or $g = g_3$ (bottom), and when $(n, \text{NSR}) = (250, 10\%)$ (left), $(n, \text{NSR}) = (500, 10\%)$ (center) and $(n, \text{NSR}) = (500, 20\%)$ (right). In each graph, the first boxplot is for the naive DBPE, the second is for the naive PE, the third is for our PE, the fourth is for the ideal DBPE and the fifth is for the ideal PE.

when $g = g_2$ or g_3, for various sample sizes and NSR. In most cases (except for $(n, \text{NSR}) = (250, 10\%)$ when $g = g_2$), our PE worked better than the naive estimators, and we can see that it even worked better than the ideal DBPE. As already mentioned several times, in other cases it is the ideal DBPE that worked better than the ideal PE, and this makes the comparison of our method with the naive estimators difficult. For example in this case, if we were to compare our PE with the naive DBPE, we would find a dramatic improvement, but if we were to compare it with the naive PE, we would find that our estimator improves the naive one by a much smaller amount.

Of course, we could not present the results of all our simulations, and above we only discussed partial results. In our complete set of simulations, we also found that our estimators systematically improved as sample size increased and/or the NSR decreased. Further, we found that our parametric method almost always improved substantially at least one of the two naive methods, and usually improved both. Depending on the case, it was either the naive DBPE that we beat by a significant amount, or the naive PE. Thus, the comparison between our estimator and the naive approach is not easy. Since in practice we would not know which of the two naive methods we should use, to be fair, we should almost choose randomly one of the two naive approaches.

7 Conclusion

We have considered an important, but particularly difficult and unexplored, problem of variance estimation in the context on nonparametric errors-in-variables regression. We have proposed nonparametric and parametric variance estimators and have derived their asymptotic and finite-sample properties. We have also proposed a new bandwidth selector that is of independent interest, since it can be used in more general errors-in-variables contexts, where constructing a good data-driven bandwidth is particularly challenging.

A Appendix

A.1 Summary

This appendix contains the proofs of Theorems 4.2 and 4.3. The proofs are given in the case where $\widehat{m}$ and $\widehat{g}$ are local constant estimators ($q = 0$). In the proofs we shall use the notation $r_1 = f_X\, g$ and $r_2 = f_X\, m$.

A.2 Proof of Theorem 4.2

Let $\mathcal{D}$ be as in section 2.2, let $\mathcal{D}^\dagger$ be a resample drawn from $\mathcal{D}$, let $\widehat{f}_X$ have the definition of $\widehat{d}$ in the special case at (3.2), write $\widehat{f}_X^\dagger$ and $\widehat{r}_j^\dagger$, $j = 1, 2$, for the versions of $\widehat{f}_X$ and $\widehat{r}_j$, respectively, when the latter are computed from $\mathcal{D}^\dagger$ rather than $\mathcal{D}$, put $\Delta^\dagger = \widehat{f}_X^\dagger - \widehat{f}_X$ and $\Delta_j^\dagger = \widehat{r}_j^\dagger - \widehat{r}_j$ for $j = 1, 2$, and let $\ell = \ell(n)$ denote a sequence of positive constants. Here and below, all estimators are understood to be evaluated at x. The first step is to prove that, for all integers $p \geq 1$,

$$P(|\Delta^\dagger| > \ell) = O\Big\{\big(n\, h^{2\alpha+1}\, \ell^2\big)^{-p}\Big\}\,. \tag{A.1}$$

By Rosenthal's inequality,

$$E\big(|\Delta^\dagger|^{2p}\,\big|\,\mathcal{D}\big) \leq \frac{D_1}{(nh)^{2p}} \left[\left\{\sum_{j=1}^n \left|L_0\Big(\frac{x-W_j}{h}\Big)\right|^2\right\}^p + \sum_{j=1}^n \left|L_0\Big(\frac{x-W_j}{h}\Big)\right|^{2p}\right],$$

where $D_1, D_2, \ldots$ will denote generic positive constants not depending on n, and D_1 depends only on p. Hence,

$$\begin{aligned} E\big(|\Delta^\dagger|^{2p}\big) &\leq \frac{D_2}{(nh)^{2p}} \sum_{r=0}^p \left\{\sum_{j=1}^n E\left|L_0\Big(\frac{x-W_j}{h}\Big)\right|^2\right\}^r \sum_{j=1}^n E\left|L_0\Big(\frac{x-W_j}{h}\Big)\right|^{2(p-r)} \\ &\leq D_3 \sum_{r=0}^{p-1} \big(nh^{2\alpha+1}\big)^{-r} (nh)^{2(r-p)}\, nh^{1-2(p-r)\alpha} + D_3\, \big(nh^{2\alpha+1}\big)^{-p} \\ &= D_3 \sum_{r=0}^{p-1} (nh)^{r+1} \big(nh^{\alpha+1}\big)^{-2p} + D_3\, \big(nh^{2\alpha+1}\big)^{-p} \\ &\leq p\, D_3 \big(n\, h^{2\alpha+1}\big)^{-p}. \end{aligned} \tag{A.2}$$

To obtain the second inequality in the sequence leading to (A.2) we used the fact that $(nh)^{-2} \sum_j E|L_0\{(x-W_j)/h\}|^2 = O\{(nh^{2\alpha+1})^{-1}\}$, this being an upper bound to the variance of $\widehat{f}_X$, and the property that for $s \geq 2$, $\sum_j E|L_0\{(x-W_j)/h\}|^s = O(nh^{1-s\alpha})$, the latter identity being a consequence of the bound

$$|L_0(u)| \leq D_4\, h^{-\alpha}\, (1+|u|)^{-1}, \tag{A.3}$$

which we shall shortly derive. Result (A.1) follows on combining (A.2) and Markov's inequality.

Let $\Delta^\dagger_{[b]}$ denote the version of $\Delta^\dagger$, defined in the first paragraph of the present section, when the dataset $\mathcal{D}$ is replaced by the bth resample, $\mathcal{D}^\dagger_b$, drawn from $\mathcal{D}$. (We introduce square brackets around the subscript in $\Delta^\dagger_{[b]}$ so as not to confuse $\Delta^\dagger_{[b]}$ with $\Delta^\dagger_j$, defined in the previous paragraph.) Since (A.1) holds for all $p > 0$, and (4.4) implies that $B = O(n^{D_5})$ for some $D_5 > 0$, then provided $\ell^2 = n^\epsilon\, (n\, h^{2\alpha+1})^{-1}$ for some $\epsilon > 0$, we have, for a positive number p that can be taken arbitrarily large:

$$P\Big(\max_{1\leq b\leq B} |\Delta^\dagger_{[b]}| > \ell\Big) = O\big(n^{D_5 - p\epsilon}\big) = O\big(n^{-D_6}\big) \quad \text{for all} \quad D_6 > 0. \tag{A.4}$$

Therefore, by Taylor expansion,

$$\widehat{m}^\dagger = \frac{\widehat{r}_2^\dagger}{\widehat{f}_X^\dagger} = \frac{\widehat{r}_2 + \Delta_2^\dagger}{\widehat{f}_X + \Delta^\dagger} = \frac{\widehat{r}_2 + \Delta_2^\dagger}{\widehat{f}_X} \left\{1 - \frac{\Delta^\dagger}{\widehat{f}_X} + O_p\big(\ell^2\big)\right\},$$

$$\widehat{g}^{\dagger 2} = \frac{\widehat{r}_1^2 + 2\,\widehat{r}_1\,\Delta_1^\dagger + \Delta_1^{\dagger 2}}{\widehat{f}_X^2} \left\{1 - \frac{2\,\Delta^\dagger}{\widehat{f}_X} + O_p(\ell^2)\right\},$$

where, here and below in this paragraph, the remainders $O_p(\cdot)$ are of the stated order uniformly in all B of the resampled datasets $\mathcal{D}^\dagger = \mathcal{D}_b^\dagger$. (To see that the order of the remainder terms is as stated, note that, since B is of polynomial size in n, then (A.4) implies that the probability that either $|\Delta_2^\dagger|$ or $|\Delta^\dagger|$ is larger than ℓ for one or more of the resamples $\mathcal{D}^\dagger$, equals $O(n^{-D})$ for all $D > 0$. Therefore the remainders, which represent quadratic terms in the two respective Taylor expansion and so equal $O_p(\ell^2)$, are of that size uniformly in all B simulated values of the resample $\mathcal{D}^\dagger$.) Hence, since

$$\widehat{r}_j(x) = r_j(x) + O_p(\ell_0) \quad \text{and} \quad \widehat{f}_X(x) = f_X(x) + O_p(\ell_0) \tag{A.5}$$

where $\ell_0^2 = (n\,h^{2\alpha+1})^{-1} < \ell^2$, and since $f_X(x) > 0$, then

$$\widehat{\tau}^\dagger = \widehat{m}^\dagger - \widehat{g}^{\dagger 2} = \tilde{\tau} + \Delta_3^\dagger\,, \tag{A.6}$$

uniformly in $\mathcal{D}^\dagger = \mathcal{D}_b^\dagger$ for $1 \le b \le B$, where

$$\Delta_3^\dagger = \Delta_4^\dagger + O_p\Big\{\big(|\Delta_1^\dagger| + |\Delta_2^\dagger| + |\Delta_1^\dagger|^2\big)\,\ell + \ell^2\Big\}\,,$$
$$\Delta_4^\dagger = \frac{1}{f_X}\left(\Delta_2^\dagger - \frac{r_2}{f_X}\,\Delta^\dagger\right) - \frac{2\,r_1}{f_X^2}\left(\Delta_1^\dagger - \frac{r_1}{f_X}\,\Delta^\dagger\right).$$

(To derive (A.5) note that the second identity there is conventional, and follows for example from arguments of Delaigle *et al.* (2009), who show that the identity gives the exact rate of convergence of $\widehat{f}_X(x)$ to $f_X(x)$. The first identity is proved in the same way (and again gives the exact convergence rate), since $\widehat{r}_j$ has the same construction as $\widehat{f}_X$ except that a weight Y_j is incorporated into the series. See Example 1 in section 3.2, where (3.2) gives a formula for $\widehat{d}(x)$, which is identical to $\widehat{f}_X(x)$ in that setting, and (3.3) gives formulae for $\widehat{r}_1(x)$ and $\widehat{r}_2(x)$.)

Write $\Delta_{jb}^\dagger$ for the version of $\Delta_j^\dagger$ when $\mathcal{D}^\dagger = \mathcal{D}_b^\dagger$, and let $\mathcal{E}$ denote the event that $|\tilde{\tau} - \tau| \le \frac{1}{2}\,\tau$, i.e. that $\frac{1}{2}\,\tau(x) < \tilde{\tau}(x) < \frac{3}{2}\,\tau(x)$ where, by assumption, $\tau(x) > 0$. Since $\tilde{\tau} \to \tau$ in probability then $P(\mathcal{E}) \to 1$. If $\mathcal{E}$ holds then by (A.6),

$$|\bar{\tau} - \tilde{\tau}| = \left|\frac{1}{B}\sum_{b=1}^{B}\{\max(\widehat{\tau}_b^\dagger, 0) - \tilde{\tau}\}\right| = \left|\frac{1}{B}\sum_{b=1}^{B}\{\max(\tilde{\tau} + \Delta_{3b}^\dagger, 0) - \tilde{\tau}\}\right|$$

$$\leq \left|\frac{1}{B}\sum_{b=1}^{B}\Delta_{3b}^{\dagger}\right| + \frac{4}{B}\sum_{b=1}^{B}|\Delta_{3b}^{\dagger}|\, I\big(|\Delta_{3b}^{\dagger}| > \tfrac{1}{2}\,\tau\big)$$

$$\leq \left|\frac{1}{B}\sum_{b=1}^{B}\Delta_{4b}^{\dagger}\right| + O_p\bigg\{\ell^2 + \frac{\ell}{B}\sum_{b=1}^{B}\big(|\Delta_{1b}^{\dagger}| + |\Delta_{2b}^{\dagger}| + |\Delta_{1b}^{\dagger}|^2\big)\bigg\}$$
$$+ \frac{4}{B}\sum_{b=1}^{B}|\Delta_{3b}^{\dagger}|\, I\big(|\Delta_{3b}^{\dagger}| > \tfrac{1}{2}\,\tau\big)\,. \qquad \text{(A.7)}$$

Conditional on $\mathcal{D}$, $\sum_b \Delta_{4b}^{\dagger}$ is a sum of independent random variables with zero mean (that is, $E(\Delta_{4b}^{\dagger}\,|\,\mathcal{D}) = 0$ for each b), and from this property it can be proved that $B^{-1}\sum_b \Delta_{4b}^{\dagger} = O_p(B^{-1/2}\,\ell_0) = o_p(\ell_0)$, since we assumed that $B = B(n) \to \infty$. Similarly, $B^{-1}\sum_b (|\Delta_{1b}^{\dagger}|+|\Delta_{2b}^{\dagger}|+|\Delta_{1b}^{\dagger}|^2) = O_p(\ell_0)$ and $B^{-1}\sum_b |\Delta_{3b}^{\dagger}|\, I(|\Delta_{3b}^{\dagger}| > \frac{1}{2}\,\tau) = o_p(\ell_0)$. In relation to the last of these results, note that if $s > 0$ is fixed then

$$\frac{1}{B}\sum_{b=1}^{B}|\Delta_{4b}^{\dagger}|\, I(|\Delta_{4b}^{\dagger}| > s) = O_p\Big[E\{|\Delta_{41}^{\dagger}|\, I(|\Delta_{41}^{\dagger}| > s)\}\Big] = O_p\big\{E|\Delta_{41}^{\dagger}|^2\big\} = O_p\big(\ell_0^2\big)\,.$$

Combining these results with (A.7), and noting that $\ell^2 = o_p(\ell_0)$ provided that ϵ, in the definition of ℓ, is chosen sufficiently small, we deduce that $|\bar{\tau} - \tilde{\tau}| = o_p(\ell_0)$. In view of our choice of h (see Theorem 6.1) the latter result is equivalent to (6.9).

It remains to prove (A.3). From Conditions (A5) and (4.1) it follows that, uniformly in x,

$$2\pi\,|L_0(x)| \leq C\int_{-1}^{1}\big|\phi_K(t)\big|\,(1 + |t/h|)^{\alpha}\,dt \leq D_7\,h^{-\alpha}\,. \qquad \text{(A.8)}$$

Conditions (A5), (4.1) and (4.5), and an integration by parts, imply that

$$|2\pi x\,L_0(x)| \leq \int\left|\frac{\phi_K'(t)\,\phi_U(t/h) - h^{-1}\,\phi_K(t)\,\phi_U'(t/h)}{\phi_U(t/h)^2}\right|dt$$
$$\leq D_8\int_0^1\left\{\big|\phi_U(t/h)\big|^{-1} + \frac{|\phi_U'(t/h)|}{h\,|\phi_U(t/h)|^2}\right\}dt$$
$$\leq D_9\,h\int_0^{1/h}\left\{(1+t)^{\alpha} + \frac{(1+t)^{\alpha}}{h\,(1+t)}\right\}dt \leq D_{10}\,h^{-\alpha}\,. \qquad \text{(A.9)}$$

Result (A.3) follows from (A.8) and (A.9).

A.3 Proof of Theorem 4.3

Step 1: Approximation to $S(\theta)$.

The goal of this step is to develop an approximation to $S(\theta)$ which is simpler than

$S(\theta)$ to analyse. In the second part of this step, we illustrate our approximation in the particular case where the variance is a polynomial. Throughout the proof we shall use the notation

$$L(x \,|\, w_1, w_2) = L_0\Big(\frac{x - w_1}{h}\Big)\, L_0\Big(\frac{x - w_2}{h}\Big)\,.$$

Also, to avoid too complicated notations in this proof, we redefine $S(\theta)$ to be $n(n-1)h^2 S(\theta)$. This has no impact on the derivation of the results, as $S(\theta) = 0$ is equivalent to $n(n-1)h^2 S(\theta) = 0$. Remember, too, that we are giving the proof for the case $q = 0$ (see example 1 in section 3.2). With this in mind, and since conditions (B1) and (B3) hold, we can write:

$$\begin{aligned} S(\theta) = \sum\sum_{j_1 \neq j_2} \int \bigg\{ Y_{j_1}^2 - Y_{j_1} Y_{j_2} - \sum_{k=0}^{d_1} \tau^{(k|0)}(x \,|\, \theta, \theta_0) \bigg\}\, L(x \,|\, W_{j_1}, W_{j_2}) \\ \times \bigg\{ \sum_{k=1}^{d_1} \tau^{(k|1)}(x \,|\, \theta, \theta_0) \bigg\}\, \omega(x)\, dx \\ + \Omega_1\, \|\theta - \theta_0\|^{d_1 + 1} \sum\sum_{j_1 \neq j_2} \big(1 + Y_{j_1}^2\big) \int |L(x \,|\, W_{j_1}, W_{j_2})|\, \omega(x)\, dx\,, \end{aligned}$$

where

$$\begin{aligned} \sum_{k \geq 0} \tau^{(k|0)}(x \,|\, \theta, \theta_0) &= \tau(x \,|\, \theta_0) + (\theta - \theta_0)^{\mathrm{T}}\, \dot\tau(x \,|\, \theta_0) \\ &\qquad + \tfrac{1}{2}\, (\theta - \theta_0)^{\mathrm{T}}\, \ddot\tau(x \,|\, \theta_0)\, (\theta - \theta_0) + \ldots\,, \\ \sum_{k \geq 1} \tau^{(k|1)}(x \,|\, \theta, \theta_0) &= \dot\tau(x \,|\, \theta_0) + \ddot\tau(x \,|\, \theta_0)\, (\theta - \theta_0) + \ldots \end{aligned}$$

denote Taylor expansions of $\tau(x \,|\, \theta)$ and $\dot\tau(x \,|\, \theta)$, respectively, in terms that are of sizes $\|\theta - \theta_0\|^k$; $\ddot\tau(x \,|\, \theta_0)$ is the $p \times p$ matrix of second derivatives of $\tau(x \,|\, \theta)$ with respect to θ; and, for $\ell = 1$ and 2, Ω_ℓ is a random variable satisfying $P(|\Omega_\ell| \leq C) = 1$, with C denoting a constant depending only on the bounds to the $d_1 + 2$ derivatives of $\tau(x \,|\, \theta)$ with respect to θ. (Recall from (B3) that those derivatives are bounded uniformly in the compact set on which ω is supported.) Therefore,

$$\begin{aligned} S(\theta) = S_0(\theta) + \sum_{k=2}^{d_1} \|\theta - \theta_0\|^k \sum\sum_{j_1 \neq j_2} \int \Big[\big\{ Y_{j_1}^2 - Y_{j_1} Y_{j_2} - \tau(x \,|\, \theta_0) \big\}\, a_{1k}(x) \\ + a_{2k}(x) \Big]\, L(x \,|\, W_{j_1}, W_{j_2})\, \omega(x)\, dx \end{aligned}$$

$$+\,\Omega_2\,\|\theta-\theta_0\|^{d_1+1}\sum\sum_{j_1\neq j_2}\left(1+Y_{j_1}^2\right)\int|L(x\,|\,W_{j_1},W_{j_2})|\,\omega(x)\,dx\,, \tag{A.10}$$

where

$$\begin{aligned}S_0(\theta)=&\sum\sum_{j_1\neq j_2}\int\left\{Y_{j_1}^2-Y_{j_1}\,Y_{j_2}-\tau(x\,|\,\theta_0)\right\}L(x\,|\,W_{j_1},W_{j_2})\,\dot\tau(x\,|\,\theta_0)\,\omega(x)\,dx\\&+\Bigg[\sum\sum_{j_1\neq j_2}\int\left\{Y_{j_1}^2-Y_{j_1}\,Y_{j_2}-\tau(x\,|\,\theta_0)\right\}L(x\,|\,W_{j_1},W_{j_2})\,\ddot\tau(x\,|\,\theta_0)\,\omega(x)\,dx\\&\quad-\sum\sum_{j_1\neq j_2}\int L(x\,|\,W_{j_1},W_{j_2})\,\dot\tau(x\,|\,\theta_0)\,\dot\tau(x\,|\,\theta_0)^{\mathrm{T}}\,\omega(x)\,dx\Bigg](\theta-\theta_0)\,,\end{aligned} \tag{A.11}$$

and the vector-valued functions a_{1k} and a_{2k} are uniformly bounded and have, respectively, d_1+1-k and d_1+2-k bounded derivatives on the real line.

To understand the above calculations in a simple context, take the particular case where the variance is polynomial, that is $\tau(x\,|\,\theta)=\theta_1+\theta_2\,x+\ldots+\theta_p\,x^{p-1}$ and $\dot\tau(x\,|\,\theta)=\dot\tau(x\,|\,\theta_0)=(1,x,\ldots,x^{p-1})^T$. There, we find

$$\tau(x\,|\,\theta)=\theta_1+\theta_2\,x+\ldots+\theta_p\,x^{p-1}=\tau(x\,|\,\theta_0)+(\theta-\theta_0)^{\mathrm{T}}\dot\tau(x\,|\,\theta_0).$$

Therefore, it follows from the definition of $S(\theta)$ at (3.1), where, as indicated above, we redefine $S(\theta)$ to be $n(n-1)h^2S(\theta)$, that in that case,

$$\begin{aligned}S(\theta)=&\sum\sum_{j_1\neq j_2}\int\left\{Y_{j_1}^2-Y_{j_1}\,Y_{j_2}-\tau(x\,|\,\theta_0)-(\theta-\theta_0)^{\mathrm{T}}\dot\tau(x\,|\,\theta_0)\right\}\\&\qquad\qquad\times L(x\,|\,W_{j_1},W_{j_2})\,\dot\tau(x\,|\,\theta_0)\,\omega(x)\,dx\\=&\sum\sum_{j_1\neq j_2}\int\left\{Y_{j_1}^2-Y_{j_1}\,Y_{j_2}-\tau(x\,|\,\theta_0)\right\}L(x\,|\,W_{j_1},W_{j_2})\,\dot\tau(x\,|\,\theta_0)\,\omega(x)\,dx\\&-\sum\sum_{j_1\neq j_2}\int L(x\,|\,W_{j_1},W_{j_2})\,\dot\tau(x\,|\,\theta_0)\dot\tau(x\,|\,\theta_0)^{\mathrm{T}}\,\omega(x)\,dx\,\times(\theta-\theta_0).\end{aligned}$$

Thus, in this case, $S(\theta)$ is exactly equal to $S_0(\theta)$.

Step 2: Approximation to $S(\theta)-S_0(\theta)$.

The calculations at the end of Step 1 show that in the polynomial case we have exactly

$S(\theta) = S_0(\theta)$. However, in more general cases, $S_0(\theta)$ is only an approximation to $S(\theta)$, and the goal of this step is to assess the magnitude of $S(\theta) - S_0(\theta)$. The approximation is given at (A.24) below.

As a prelude to deriving it, let a denote a uniformly bounded function with support equal to that of ω, and define

$$S_1 = \sum\sum_{j_1 \neq j_2} \int a(x)\, L(x \mid W_{j_1}, W_{j_2})\, dx\,, \tag{A.12}$$

$$S_2 = \sum\sum_{j_1 \neq j_2} Y_{j_1}^2 \int a(x)\, L(x \mid W_{j_1}, W_{j_2})\, dx\,, \tag{A.13}$$

$$S_3 = \sum\sum_{j_1 \neq j_2} Y_{j_1}\, Y_{j_1} \int a(x)\, L(x \mid W_{j_1}, W_{j_2})\, dx\,. \tag{A.14}$$

We shall develop bounds for $E(S_\ell^2)$ for $\ell = 1$, 2 and 3, giving details of the arguments only in the relatively complex case $\ell = 3$. Now, each S_ℓ can be decomposed into "quadratic," and "linear (projection)" components. (In the case $\ell = 3$ see (A.15) and (A.16), below, for quadratic and linear components, respectively, and when $\ell = 2$ see (A.20) for the quadratic component, and (A.21) and (A.22) for the two linear components.) We bound the quadratic and linear components separately, noting that the method in the case of quadratic components will be used again in Step 4.

Given a random variable R with finite mean, let $(1 - E)\, R$ denote $R - E(R)$ and put $(1 - E)\, S_3 = S_{31} + 2\, S_{32}$, where $2\, S_{32}$ equals the linear projection of $(1 - E)\, S_3$ and S_{31} is defined by differencing:

$$\begin{aligned} S_{31} = \sum\sum_{j_1 \neq j_2} \int a(x) \bigg[& Y_{j_1}\, Y_{j_2}\, L_0\Big(\frac{x - W_{j_1}}{h}\Big)\, L_0\Big(\frac{x - W_{j_2}}{h}\Big) \\ & - \Big\{ Y_{j_1}\, L_0\Big(\frac{x - W_{j_1}}{h}\Big) + Y_{j_2}\, L_0\Big(\frac{x - W_{j_2}}{h}\Big) \Big\}\, E\Big\{ Y\, L_0\Big(\frac{x - W}{h}\Big) \Big\} \\ & + \Big\{ E\, Y\, L_0\Big(\frac{x - W}{h}\Big) \Big\}^2 \bigg]\, dx\,, \end{aligned} \tag{A.15}$$

$$S_{32} = (n - 1) \sum_{j=1}^{n} \int a(x)\, E\Big\{ Y\, L_0\Big(\frac{x - W}{h}\Big) \Big\}\, (1 - E)\, \Big\{ Y_j\, L_0\Big(\frac{x - W_j}{h}\Big) \Big\}\, dx\,. \tag{A.16}$$

Since $Y_j = g(X_j) + \tau^{1/2}(X_j)\, \varepsilon_j$ and $W_j = X_j + U_j$ then, recalling that $m = g^2 + \tau$ and noting, from the definition of L_0, that $\int L_0^2 \leq \text{const.}\, h^{-2\alpha}$ for $0 < h \leq 1$, where α is as in (B5), we have:

$$\begin{aligned}
n^{-2}\, E\big(S_{31}^2\big) &\leq 4\, E\bigg\{Y_1\, Y_2 \int a(x)\, L_0\bigg(\frac{x-W_1}{h}\bigg)\, L_0\bigg(\frac{x-W_2}{h}\bigg)\, dx\bigg\}^2 \\
&= E\Bigg(m(X_1)\, m(X_2)\, E\bigg[\bigg\{\int a(x)\, L_0\bigg(\frac{x-W_1}{h}\bigg)\, L_0\bigg(\frac{x-W_2}{h}\bigg)\, dx\bigg\}^2 \,\bigg|\, X_1, X_2\bigg]\Bigg) \\
&\leq h^2\, \big(\sup a^2\big)\, E\big\{m^2(X)\big\}\, \bigg(\int L_0^2\bigg)^2 \leq \text{const.}\, h^{2-4\alpha}\,. \qquad \text{(A.17)}
\end{aligned}$$

Also, since the function $|f_X\, g|$ is bounded (see (B4)) then $|E[g(X_1)\, K\{(x-X_1)/h\}]| \leq$ const. h, whence it follows that

$$\begin{aligned}
n^{-3}\, E\big(S_{32}^2\big) &\leq E\Bigg(m(X_2)\, \bigg[\int a(x)\, E\bigg\{g(X_1)\, K\bigg(\frac{x-X_1}{h}\bigg)\bigg\}\, L_0\bigg(\frac{x-W_2}{h}\bigg)\bigg\}\, dx\bigg]^2\Bigg) \\
&\leq \text{const.}\, h^2\, E\bigg\{m(X_2)\, \bigg(\int a^2\bigg) \int L_0^2\bigg(\frac{x-W_2}{h}\bigg)\, dx\bigg\} \\
&\leq \text{const.}\, h^3\, \big(1 + h^{-2\alpha}\big)\,. \qquad \text{(A.18)}
\end{aligned}$$

Therefore, in the case $\ell = 3$,

$$(nh)^{-4}\, \text{var}(S_\ell) = O\big(n^{-2}\, h^{-2(1+2\alpha)} + n^{-1}\, h^{-(1+2\alpha)}\big)\,. \qquad \text{(A.19)}$$

Write $S_2 = S_{21} + S_{22} + S_{23}$ where

$$\begin{aligned}
S_{21} = \sum\sum_{j_1 \neq j_2} \int a(x)\, \bigg[&Y_{j_1}^2\, L_0\bigg(\frac{x-W_{j_1}}{h}\bigg)\, L_0\bigg(\frac{x-W_{j_2}}{h}\bigg) \\
&- Y_{j_1}^2\, L_0\bigg(\frac{x-W_{j_1}}{h}\bigg)\, E\bigg\{L_0\bigg(\frac{x-W_1}{h}\bigg)\bigg\} \\
&- L_0\bigg(\frac{x-W_{j_2}}{h}\bigg)\, E\bigg\{Y_1^2\, L_0\bigg(\frac{x-W_1}{h}\bigg)\bigg\} \\
&+ E\bigg\{Y_1^2\, L_0\bigg(\frac{x-W_1}{h}\bigg)\bigg\}\, E\bigg\{L_0\bigg(\frac{x-W_1}{h}\bigg)\bigg\}\bigg]\, dx\,, \qquad \text{(A.20)}
\end{aligned}$$

$$S_{22} = (n-1) \sum_{j=1}^n \int a(x)\, E\bigg\{L_0\bigg(\frac{x-W_1}{h}\bigg)\bigg\}\, (1-E)\, \bigg\{Y_j^2\, L_0\bigg(\frac{x-W_j}{h}\bigg)\bigg\}\, dx\,, \qquad \text{(A.21)}$$

$$S_{23} = (n-1) \sum_{j=1}^n \int a(x)\, E\bigg\{Y_1^2\, L_0\bigg(\frac{x-W_1}{h}\bigg)\bigg\}\, (1-E)\, \bigg\{L_0\bigg(\frac{x-W_j}{h}\bigg)\bigg\}\, dx\,. \qquad \text{(A.22)}$$

Provided that $E(\varepsilon^4) < \infty$ and $E\{m^2(X)\} < \infty$ (see (B6)) the arguments leading to (A.17) and (A.18) give: $n^{-2}\, E(S_{21}^2) = O(h^{2-4\alpha})$ and $n^{-3}\, \max_{j=2,3} E(S_{2j}^2) = O(h^{3-2\alpha})$. This leads quickly to (A.19) in the case $\ell = 2$, and a similar approach implies that result when $\ell = 1$. Note too that, for $\ell = 1, 2, 3$, $E(S_\ell) = O\big\{(nh)^2\big\}$, and therefore $(nh)^{-2}\, E(S_\ell) = O(1)$. Combining this result with the versions of (A.19) for $\ell = 1, 2, 3$ we deduce that:

the coefficient of Ω_2 in (A.10), multiplied by $(nh)^{-2}$, equals $O_p\{\|\theta - \theta_0\|^2\, (n^{-1}\, h^{-(1+2\alpha)} + n^{-1/2}\, h^{-(1+2\alpha)/2} + 1)\}$, uniformly in θ satisfying $\|\theta - \theta_0\| \le C$, where $C > 0$ is as in (B3). (A.23)

Observe too that

$$\begin{aligned} &\sum\sum_{j_1 \ne j_2} \big(1 + Y_{j_1}^2\big) \int |L(x \,|\, W_{j_1}, W_{j_2})|\, \omega(x)\, dx \\ &\quad \le \text{const.} \sum\sum_{j_1 \ne j_2} \big(1 + Y_{j_1}^2\big) \left\{ \int L_0^2\Big(\frac{x - W_{j_1}}{h}\Big)\, dx \right\}^{1/2} \left\{ \int L_0^2\Big(\frac{x - W_{j_2}}{h}\Big)\, dx \right\}^{1/2} \\ &\quad \le \text{const.} \sum\sum_{j_1 \ne j_2} \big(1 + Y_{j_1}^2\big)\, h\, \big(1 + h^{-2\alpha}\big) = O_p\big(n^2\, h^{1-2\alpha}\big)\,. \end{aligned}$$

Therefore the coefficient of Ω_3 in (A.10), multiplied by $(nh)^{-2}$, equals $O_p(\|\theta - \theta_0\|^{d_1+1}\, h^{-(1+2\alpha)})$, uniformly in θ satisfying $\|\theta - \theta_0\| \le C$. Combining this result with (A.10) and (A.23) we deduce that:

$S(\theta) = S_0(\theta) + (nh)^2\, \Delta(\theta)$ where, uniformly in θ satisfying $\|\theta - \theta_0\| \le C$, $\Delta(\theta) = O_p(\|\theta - \theta_0\|^2\, \lambda_n + \|\theta - \theta_0\|^{d_1+1}\, h^{-(1+2\alpha)})$, $C > 0$ is as in (B3), and $\lambda_n = n^{-1}\, h^{-(1+2\alpha)} + n^{-1/2}\, h^{-(1+2\alpha)/2} + 1$. (A.24)

As mentioned earlier, in the particular case where the variance is polynomial we have $S(\theta) = S_0(\theta)$ and therefore $\Delta(\theta) = 0$.

Step 3: Solving the equation $S(\theta) = 0$.

Here we show that the equation $S(\theta) = 0$ can be written in a simpler form, specifically (A.30) below, provided that the bandwidth satisfies (A.29).

Recalling the definition of $S_0(\theta)$ at (A.11) we deduce that

$$(nh)^{-2}\, S_0(\theta) = V - (M - N)\, (\theta - \theta_0)\,, \tag{A.25}$$

where M and N are $p \times p$ matrices, V is a p-vector,

$$\begin{aligned}
M &= \frac{1}{(nh)^2} \sum\sum_{j_1 \neq j_2} \int L(x \,|\, W_{j_1}, W_{j_2})\, \dot\tau(x \,|\, \theta_0)\, \dot\tau(x \,|\, \theta_0)^{\mathrm{T}}\, \omega(x)\, dx\,, \\
N &= \frac{1}{(nh)^2} \sum\sum_{j_1 \neq j_2} \int \big\{Y_{j_1}^2 - Y_{j_1}\, Y_{j_2} - \tau(x \,|\, \theta_0)\big\}\, L(x \,|\, W_{j_1}, W_{j_2})\, \ddot\tau(x \,|\, \theta_0)\, \omega(x)\, dx\,, \\
V &= \frac{1}{(nh)^2} \sum\sum_{j_1 \neq j_2} \int \big\{Y_{j_1}^2 - Y_{j_1}\, Y_{j_2} - \tau(x \,|\, \theta_0)\big\}\, L(x \,|\, W_{j_1}, W_{j_2})\, \dot\tau(x \,|\, \theta_0)\, \omega(x)\, dx\,.
\end{aligned}$$

We shall show in three stages, respectively in Steps 4–6 below, that $n^{1/2}\,(V - EV)$ is asymptotically normally distributed with zero mean and finite variance. Similar arguments can be used to prove that

$$M - E(M) = o_p(1)\,, \quad N - E(N) = o_p(1)\,. \tag{A.26}$$

Define

$$\begin{aligned}
\phi_0(x) = \int K(u)\, f_X(x - hu)\, du\,, \quad \phi_1(x) = \int K(u)\, (g f_X)(x - hu)\, du\,, \\
\phi_2(x) = \int K(u)\, (m f_X)(x - hu)\, du\,.
\end{aligned} \tag{A.27}$$

Since the kernel K is of order κ (see (B2)) then, in view of the smoothness assumptions (B4), $\phi_0(x) = f_X(x) + O(h^\kappa)$, and so

$$\begin{aligned}
\big(1 - n^{-1}\big)^{-1} E(M) &= \int \phi_0^2(x)\, \dot\tau(x \,|\, \theta_0)\, \dot\tau(x \,|\, \theta_0)^{\mathrm{T}}\, \omega(x)\, dx \\
&= \int f_X^2(x)\, \dot\tau(x \,|\, \theta_0)\, \dot\tau(x \,|\, \theta_0)^{\mathrm{T}}\, \omega(x)\, dx + O\big(h^\kappa\big)\,.
\end{aligned}$$

Similarly, since $\phi_1(x) = (f_X\, g)(x) + O(h^\kappa)$ and $\phi_2(x) = (f_X\, m)(x) + O(h^\kappa)$, we have:

$$\begin{aligned}
\big(1 - n^{-1}\big)^{-1} E(N) &= \int \big\{\phi_2(x)\, \phi_0(x) - \phi_1^2(x) - \tau(x \,|\, \theta_0)\, \phi_0^2(x)\big\}\, \ddot\tau(x \,|\, \theta_0)\, \omega(x)\, dx \\
&= O\big(h^\kappa\big)\,,
\end{aligned}$$

and $(1 - n^{-1})^{-1}\, E(V) = O(h^\kappa)$. Combining the results from (A.26) down, and assuming that $h = h(n)$ converges to zero sufficiently fast to ensure that $h^\kappa = o(n^{-1/2})$ (see (B8)), we deduce that $M = M_0 + o_p(1)$, where M_0 is as at (4.7), $N = o_p(1)$ and $E(V) = o(n^{-1/2})$. Hence, by (A.25), the equation $S_0(\theta) = 0$ can be written as:

$$V - EV - \{M_0 + o_p(1)\}\, (\theta - \theta_0) = o_p\big(n^{-1/2}\big)\,, \tag{A.28}$$

uniformly in θ satisfying $\|\theta - \theta_0\| \le C$. From (A.24) and (A.28) we deduce that if ϵ_n denotes a sequence decreasing to zero, then, provided that

$$\begin{aligned}&h = h(n) \text{ converges to zero sufficiently fast to ensure that } h^{\kappa} = o(n^{-1/2}),\\ &\text{but so slowly that } \epsilon_n\,(n^{-1}\,h^{-(1+2\alpha)} + n^{-1/2}\,h^{-(1+2\alpha)/2}) + \epsilon_n^{d_1}\,h^{-(1+2\alpha)} \to 0,\end{aligned} \tag{A.29}$$

the following is true:

$$\begin{aligned}&\text{the equation } S(\theta) = 0 \text{ can be written as } V - EV - \{M_0 + o_p(1)\}\,(\theta - \theta_0) =\\ &o_p(n^{-1/2}), \text{ uniformly in } \theta \text{ satisfying } \|\theta - \theta_0\| \le \epsilon_n.\end{aligned} \tag{A.30}$$

Step 4. Decomposing $n^{1/2}\,(V - EV)$ into its projection plus a negligible remainder.

Put $V - E(V) = V_1 + V_2$ where, defining

$$V(j_1, j_2) = \int \bigl\{Y_{j_1}^2 - Y_{j_1}\,Y_{j_2} - \tau(x\,|\,\theta_0)\bigr\}\,L(x\,|\,W_{j_1}, W_{j_2})\,\dot{\tau}(x\,|\,\theta_0)\,\omega(x)\,dx,$$

and writing $\mathcal{F}_j$ for the sigma-field generated by (U_j, X_j, Y_j), we have:

$$\begin{aligned}V_1 &= \frac{1}{(nh)^2}\sum\sum_{j_1 \ne j_2}\Bigl[V(j_1, j_2)\\ &\qquad - E\{V(j_1, j_2)\,|\,\mathcal{F}_{j_1}\} - E\{V(j_1, j_2)\,|\,\mathcal{F}_{j_2}\} + E\{V(j_1, j_2)\}\Bigr],\\ V_2 &= \frac{n-1}{(nh)^2}\sum_{j=1}^{n}\Bigl[E\{V(j, j')\,|\,\mathcal{F}_j\} + E\{V(j', j)\,|\,\mathcal{F}_j\} - 2\,E\{V(1,2)\}\Bigr],\end{aligned} \tag{A.31}$$

and j' is taken to be any integer not equal to j. We show in the present step that

$$E\bigl(V_1^2\bigr) = o\bigl(n^{-1}\bigr)\,. \tag{A.32}$$

It follows that $V_1 = o_p(n^{-1/2})$, and thence that

$$V - E(V) = V_2 + o_p\bigl(n^{-1/2}\bigr)\,. \tag{A.33}$$

To derive (A.32), note that

$$\begin{aligned}\bigl(1 - n^{-1}\bigr)^{-1}\,E\bigl(V_1^2\bigr) &= \bigl(n^2h^4\bigr)^{-1}\,E\Bigl[V(1,2)\\ &\qquad - E\{V(1,2)\,|\,\mathcal{F}_1\} - E\{V(1,2)\,|\,\mathcal{F}_2\} + E\{V(1,2)\}\Bigr]^2\\ &\le 4\,\bigl(n^2h^4\bigr)^{-1}\,E\bigl\{V(1,2)^2\bigr\}.\end{aligned}$$

At this point we recall the arguments used to bound the quadratic components in expansions of S_1, S_2 and S_2 during Step 2. The quantities S_ℓ are defined at (A.12)–(A.14), the quadratic components of S_2 and S_3 are given at (A.20) and (A.15) respectively, and the arguments used to bound the mean squares of those components can be employed here to prove that $(n^2h^4)^{-1}\, E\{V(1,2)^2\} = O(n^{-2}\, h^{-2(1+2\alpha)})$. (Compare (A.17), which implies that $(nh)^{-4}\, E(S_{31}^2) = O(n^{-2}\, h^{-2(1+2\alpha)})$, and note that an almost identical argument gives $(nh)^{-4}\, E(S_{21}^2) = O(n^{-2}\, h^{-2(1+2\alpha)})$; see the paragraph below (A.19).) Therefore, provided that

$$h = h(n) \text{ converges to zero so slowly that } n^{-1}\, h^{-2(1+2\alpha)} \to 0\,, \tag{A.34}$$

(A.32) and hence (A.33) hold.

Assumption (B8) is included in the intersection of (A.29) and (A.34). Therefore, combining (A.30) (which is implied by (A.29)) and (A.33) (which follows from (A.34)) we deduce that if (B8) holds then:

the equation $S(\theta) = 0$ can be written as $V_2 - \{M_0 + o_p(1)\}\,(\theta - \theta_0) = o_p(n^{-1/2})$, uniformly in θ satisfying $\|\theta - \theta_0\| \le \epsilon_n$. (A.35)

Step 5. Asymptotic variance of $n^{1/2}\, V_2$.

Recall the definitions of ϕ_0, ϕ_1 and ϕ_2 at (A.27), and that V_2 is given by (A.31). In this notation,

$$\begin{aligned}
E\{L(x\,|\,W_j, W_{j'})\,|\,\mathcal{F}_j\} &= L_0\biggl(\frac{x - W_j}{h}\biggr)\, E\biggl\{K\biggl(\frac{x - X}{h}\biggr)\biggr\} \\
&= h\,\phi_0(x)\, L_0\biggl(\frac{x - W_j}{h}\biggr)\,, \\
E\bigl\{Y_j^2\, L(x\,|\,W_j, W_{j'})\,|\,\mathcal{F}_j\bigr\} &= Y_j^2\, E\{L(x\,|\,W_j, W_{j'})\,|\,\mathcal{F}_j\} \\
&= h\, Y_j^2\, \phi_0(x)\, L_0\biggl(\frac{x - W_j}{h}\biggr)\,, \\
E\bigl\{Y_{j'}^2\, L(x\,|\,W_j, W_{j'})\,|\,\mathcal{F}_j\bigr\} &= L_0\biggl(\frac{x - W_j}{h}\biggr)\, E\biggl\{m(X)\, K\biggl(\frac{x - X}{h}\biggr)\biggr\} \\
&= h\,\phi_2(x)\, L_0\biggl(\frac{x - W_j}{h}\biggr)\,, \\
E\bigl\{Y_j\, Y_{j'}\, L(x\,|\,W_j, W_{j'})\,|\,\mathcal{F}_j\bigr\} &= Y_j\, L_0\biggl(\frac{x - W_j}{h}\biggr)\, E\biggl\{g(X)\, K\biggl(\frac{x - X}{h}\biggr)\biggr\}
\end{aligned}$$

$$= h\,\phi_1(x)\,Y_j\,L_0\biggl(\frac{x-W_j}{h}\biggr)\,,$$

whence

$$\begin{aligned} Q_j &\equiv E\{V(j,j')\,|\,\mathcal{F}_j\} + E\{V(j',j)\,|\,\mathcal{F}_j\} \\ &= h \int \Bigl\{\phi_2(x) + Y_j^2\,\phi_0(x) - 2\,Y_j\,\phi_1(x) - 2\,\tau(x\,|\,\theta_0)\,\phi_0(x)\Bigr\} \\ &\qquad\qquad\qquad \times L_0\biggl(\frac{x-W_j}{h}\biggr)\,\dot\tau(x\,|\,\theta_0)\,\omega(x)\,dx\,. \end{aligned} \tag{A.36}$$

Note too that, by (B4):

> the functions ϕ_0, ϕ_1 and ϕ_2 are absolutely integrable, where the integrals are bounded uniformly in h, and $\phi_0 = f_X + o(1)$, $\phi_1 = f_X\,g + o(1)$ and $\phi_2 = f_X\,m + o(1)$ as $h \to 0$; and moreover, these properties continue to hold if ϕ_0, ϕ_1 and ϕ_2, and the functions on the right-hand sides of each of the equations, are replaced by their jth derivatives, for $1 \le j \le d_2$, where d_2 is as in (B4). (A.37)

(Here we have used the fact that $\int |K| < \infty$; see (B2).) Therefore,

$$\begin{aligned} E(Q_j) &= h\,E\biggl[\int \Bigl\{\phi_2(x) + m(X)\,\phi_0(x) - 2\,g(X)\,\phi_1(x) - 2\,\tau(x\,|\,\theta_0)\,\phi_0(x)\Bigr\} \\ &\qquad\qquad\qquad \times K\biggl(\frac{x-X}{h}\biggr)\,\dot\tau(x\,|\,\theta_0)\,\omega(x)\,dx\biggr] \\ &= 2\,h^2 \int \Bigl\{\phi_0(x)\,\phi_2(x) - \phi_1^2(x) - \tau(x\,|\,\theta_0)\,\phi_0^2(x)\Bigr\}\,\dot\tau(x\,|\,\theta_0)\,\omega(x)\,dx \\ &= 2\,h^2 \int \bigl\{m(x) - g^2(x) - \tau(x\,|\,\theta_0)\bigr\}\,f_X^2(x)\,\dot\tau(x\,|\,\theta_0)\,\omega(x)\,dx + o\bigl(h^2\bigr) \\ &= o\bigl(h^2\bigr)\,, \end{aligned} \tag{A.38}$$

since $\tau = m + g^2$.

The definition of V_2 at (A.31) implies that

$$\bigl(1 - n^{-1}\bigr)^{-1}\,V_2 = \frac{1}{nh^2}\,\sum_{j=1}^n\,(Q_j - EQ_j)\,. \tag{A.39}$$

Below, we use this formula to develop an approximation to $E(V_2 V_2^{\mathrm{T}})$. By (A.38),

$$\begin{aligned} n\,E\bigl(V_2 V_2^{\mathrm{T}}\bigr) + o(1) &= h^{-4}\,E\bigl(Q_1 Q_1^{\mathrm{T}}\bigr) \\ &= h^{-2}\,E\biggl(\biggl[\int \Bigl\{\phi_2(x) + Y^2\,\phi_0(x) - 2\,Y\,\phi_1(x) - 2\,\tau(x\,|\,\theta_0)\,\phi_0(x)\Bigr\} \end{aligned}$$

$$
\begin{aligned}
&\qquad\qquad \times L_0\bigg(\frac{x-W}{h}\bigg)\,\dot\tau(x\,|\,\theta_0)\,\omega(x)\,dx\bigg] \\
&\quad \times \bigg[\int \Big\{\phi_2(x) + Y^2\,\phi_0(x) - 2\,Y\,\phi_1(x) - 2\,\tau(x\,|\,\theta_0)\,\phi_0(x)\Big\} \\
&\qquad\qquad \times L_0\bigg(\frac{x-W}{h}\bigg)\,\dot\tau(x\,|\,\theta_0)\,\omega(x)\,dx\bigg]^{\mathrm{T}}\Bigg) \\
&= h^{-2}\iint E\bigg\{\psi(x_1,x_2,X)\,L_0\bigg(\frac{x_1-W}{h}\bigg)\,L_0\bigg(\frac{x_2-W}{h}\bigg)\bigg\}\,dx_1\,dx_2 \\
&= h^{-2}\iiiint \psi(x_1,x_2,x)\,L_0\bigg(\frac{x_1-x-u}{h}\bigg)\,L_0\bigg(\frac{x_2-x-u}{h}\bigg) \\
&\qquad\qquad \times f_X(x)\,f_U(u)\,dx_1\,dx_2\,dx\,du \\
&= \iiiint \psi(w+hv_1,w+hv_2,x)\,L_0(v_1)\,L_0(v_2) \\
&\qquad\qquad \times f_X(x)\,f_U(w-x)\,dv_1\,dv_2\,dx\,dw\,, \qquad \text{(A.40)}
\end{aligned}
$$

where the $p \times p$ matrix of functions ψ is given by

$$
\begin{aligned}
\psi(x_1,x_2,x) &= E\bigg\{\Big(\Big[\phi_2(x_1) + \{g(x)+\tau^{1/2}(x)\,\varepsilon\}^2\;\phi_0(x_1) \\
&\qquad - 2\,\{g(x)+\tau^{1/2}(x)\,\varepsilon\}\,\phi_1(x_1) - 2\,\tau(x)\,\phi_0(x_1)\Big]\,\dot\tau(x\,|\,\theta_0)\,\omega(x)\Big) \\
&\quad \times \Big(\Big[\phi_2(x_2) + \{g(x)+\tau^{1/2}(x)\,\varepsilon\}^2\;\phi_0(x_2) \\
&\qquad - 2\,\{g(x)+\tau^{1/2}(x)\,\varepsilon\}\,\phi_1(x_2) - 2\,\tau(x)\,\phi_0(x_2)\Big]\,\dot\tau(x\,|\,\theta_0)\,\omega(x)\Big)^{\mathrm{T}}\bigg\} \\
&= \Bigg(\!\!\Bigg(\sum_{\ell=1}^{\ell_0} \psi_{\ell 1}(x_1)\,\psi_{\ell 2}(x_2)\,\psi_{\ell 3}(x)\Bigg)\!\!\Bigg)\,, \qquad \text{(A.41)}
\end{aligned}
$$

the notation $(\!(\rho)\!)$ refers to a $p \times p$ matrix for which a general component has the same form as ρ, and the quantities $\psi_{\ell k}$ are functions. (To obtain the last line in (A.40) we changed variable as follows: $x_j = x + u + hv_j$ for $j = 1, 2$, and $u = w - x$.)

To establish that each component of the $p \times p$ matrix represented by the four-fold integral on the right-hand side of (A.40) is uniformly bounded, we replace $\psi(x_1, x_2, x)$ there by any one of the components $\psi_{\ell 1}(x_1)\,\psi_{\ell 2}(x_2)\,\psi_{\ell 3}(x)$ at (A.41). For notational simplicity we write the latter product as $\psi_1(x_1)\,\psi_2(x_2)\,\psi_3(x)$, and note that the respective component of the matrix of integrals at (A.40) then becomes:

$$
I(\psi_1,\psi_2,\psi_3) = \iint \psi_3(x)\,f_X(x)\,f_U(w-x)\,\bigg\{\int \psi_1(w+hv_1)\,L_0(v_1)\,dv_1\bigg\}
$$

$$\times \left\{ \int \psi_2(w + hv_2)\, L_0(v_2)\, dv_2 \right\} dx\, dw\,, \tag{A.42}$$

the absolute value of which is bounded by:

$$\begin{aligned}
&\iint |\psi_3(x)|\, f_X(x)\, f_U(w-x) \left| \int \psi_1(w + hv_1)\, L_0(v_1)\, dv_1 \right| \\
&\qquad\qquad\qquad\qquad \times \left| \int \psi_2(w + hv_2)\, L_0(v_2)\, dv_2 \right| dx\, dw \\
&\quad = \int \chi(w) \left| \int \psi_1(w + hv)\, L_0(v)\, dv \right| \left| \int \psi_2(w + hv)\, L_0(v)\, dv \right| dw \\
&\quad \leq (\sup \chi) \prod_{k=1}^{2} \left[\int \left\{ \int \psi_k(w + hv)\, L_0(v)\, dv \right\}^2 dw \right]^{1/2},
\end{aligned}$$

where $\chi(w) = \int |\psi_3(x)|\, f_X(x)\, f_U(w-x)\, dx$. It follows from (B3)–(B4) that $\sup \chi$ is bounded, uniformly in n (note that χ depends on $h = h(n)$) and in all forms of $\psi_3 = \psi_{\ell 3}$ in the representation (A.41).

By Plancherel's identity,

$$2\pi \int \left\{ \int \psi_k(w + hv)\, L_0(v)\, dv \right\}^2 dw = \int \left| \xi_k^{\mathrm{Ft}}(t) \right|^2 dt\,,$$

where ξ_k^{Ft} denotes the Fourier transform of ξ_k and $\xi_k(w) = \int \psi_k(w + hv)\, L_0(v)\, dv$. Also,

$$\begin{aligned}
\xi_k^{\mathrm{Ft}}(t) &= \iint \exp(itw)\, \psi_k(w + hv)\, L_0(v)\, dv\, dw \\
&= \iint \exp\{it\,(w + hv) - ihtv\}\, \psi_k(w + hv)\, L_0(v)\, dv\, dw \\
&= \int \exp(itx)\, \psi_k(x)\, dx \;\cdot \int \exp(-ihtv)\, L_0(v)\, dv \\
&= \psi_k^{\mathrm{Ft}}(t)\, L_0^{\mathrm{Ft}}(-ht) = \psi_k^{\mathrm{Ft}}(t)\, \phi_U^{-1}(t)\, \phi_K(ht)\,,
\end{aligned}$$

where ψ_k^{Ft} denotes the Fourier transform of ψ_k. Combining the bounds from (A.42) down we deduce that

$$|I(\psi_1, \psi_2, \psi_3)| \leq (\sup \chi) \left(\sup |\phi_K| \right) (2\pi)^{-1} \prod_{k=1}^{2} \left\{ \int \left| \psi_k^{\mathrm{Ft}}(t)\, \phi_U^{-1}(t) \right|^2 dt \right\}^{1/2}. \tag{A.43}$$

If each component of each of f_X, g, $\tau(\cdot \mid \theta_0)\, \omega$ and $\dot\tau(\cdot \mid \theta_0)\, \omega$ has d_2 absolutely integrable derivatives (see (B4)) then the same is true of each function $\psi_{\ell k}$ appearing in (A.41); and it remains true in the limit, as $n \to \infty$ (meaning here that $h \to 0$), in the sense that the integrals of the absolute values of each of the first d_2 derivatives of each component of each of f_X, g, $\tau(\cdot \,|\, \theta_0)\,\omega$ and $\dot\tau(\cdot \,|\, \theta_0)\,\omega$ are bounded as h decreases. (The functions ψ_0, ψ_1 and ψ_2 at (A.27) depend on h, but they and their derivatives also satisfy (A.37) as $n \to \infty$.) In consequence, the respective characteristic functions $\psi_{\ell k}^{\mathrm{Ft}}$ of $\psi_{\ell k}$ all satisfy

$$\big|\psi_{\ell k}^{\mathrm{Ft}}\big| \leq \text{const.}\, (1 + |t|)^{-d_2} \,. \tag{A.44}$$

If $d_2 > \alpha + \frac{1}{2}$, which is ensured by (B5), then it follows from (A.44) and the inequality $|\phi_U(t)| \geq \text{const.}\, (1+|t|)^{-\alpha}$ (this too is guaranteed by (B5)) that $\int |\psi_{\ell k}^{\mathrm{Ft}}(t)\, \phi_U^{-1}(t)|^2\, dt$ is bounded uniformly in k, ℓ and h. Hence, by (A.43), $|I(\psi_{\ell 1}, \psi_{\ell 2}, \psi_{\ell 3})|$ is bounded uniformly in $1 \leq \ell \leq \ell_0$, where ℓ_0 is as in (A.41), and so, by (A.40) and (A.42),

each component of $h^{-4}\, E(Q_1 Q_1^{\mathrm{T}})$, or equivalently of $n\, E(V_2 V_2^{\mathrm{T}})$, is bounded as $n \to \infty$. (A.45)

Calculating the limit of $n\, E(V_2 V_2^{\mathrm{T}})$, as $n \to \infty$, requires only minor modification of the argument above, as follows. For $k = 1, 2$, replace $\int \psi_{\ell k}(w + h v_k) L_0(v_k)\, dv_k$ by the limit of that quantity as $n \to \infty$, which, for almost all w, is given by

$$\begin{aligned}\lim_{h \to 0} \frac{1}{2\pi} \int e^{-itw} \psi_{\ell k}^{\mathrm{Ft}}(t) \{L_0(v_k/h)/h\}^{\mathrm{Ft}}\, dt &= \lim_{h\to 0} \frac{1}{2\pi} \int e^{-itw} \psi_{\ell k}^{\mathrm{Ft}}(t) \phi_K(ht)/\phi_U(t)\, dt \\ &= \frac{1}{2\pi} \int e^{-itw} \bar\psi_{\ell k}^{\mathrm{Ft}}(t)/\phi_U(t)\, dt,\end{aligned}$$

where $\bar\psi_{\ell k}(t) = \lim_{h \to 0} \psi_{\ell k}(t)$. Write $\bar\psi$ for the version of ψ, at (A.41), when each $\int \psi_{\ell k}(w + h v_k) L_0(v_k)\, dv_k$ in (A.42) is replaced by its limit. Then, arguing as above, we deduce that the limit (as $n \to \infty$) of $I(\psi_1, \psi_2, \psi_3)$ is finite and therefore,

$$\begin{aligned} n\, E(V_2 V_2^{\mathrm{T}}) &= h^{-4}\, E\big(Q_1 Q_1^{\mathrm{T}}\big) + o(1) \\ &\to \Sigma_1 \equiv \iint \bar\psi(w, w, x)\, f_X(x)\, f_U(w - x)\, dx\, dw \,, \end{aligned} \tag{A.46}$$

where $\bar\psi(w, w, x) = E(SS^T)$, with

$$\begin{aligned} S = \Big[&\frac{1}{2\pi} \int e^{-itw} (f_X\, m)^{\mathrm{Ft}}(t)/\phi_U(t)\, dt + \{g(x) + \tau^{1/2}(x \,|\, \theta_0)\, \varepsilon\}^2\, \frac{1}{2\pi} \int e^{-itw} \phi_X(t)/\phi_U(t)\, dt \\ &- 2\, \{g(x) + \tau^{1/2}(x \,|\, \theta_0)\, \varepsilon\}\, \frac{1}{2\pi} \int e^{-itw} (f_X\, g)^{\mathrm{Ft}}(t)/\phi_U(t)\, dt \end{aligned}$$

$$- 2\,\tau(x\,|\,\theta_0)\,\frac{1}{2\pi}\int e^{-itw}\phi_X(t)/\phi_U(t)\,dt\Big]\,\dot\tau(x\,|\,\theta_0)\,\omega(x). \tag{A.47}$$

Step 6. Central limit theorem for $n^{1/2}\,V_2$.

In view of the representation (A.39) of V_2, and of the property that $n^{-1}\,E\{h^{-2}\,\sum_j\,(Q_j - EQ_j)\}^2$ converges to a finite limit as $n \to \infty$ (see Step 5), it suffices to establish the version of Lindeberg's condition here, i.e. to show that, for each $\epsilon > 0$,

$$E\big\{\big\|h^{-2}\,Q_1\big\|^2\,I\big(\big\|h^{-2}\,Q_1\big\| > n^{1/2}\,\epsilon\big)\big\} \to 0 \tag{A.48}$$

as $n \to \infty$. We shall prove that (A.48) holds if h satisfies (B8).

Using the representation (A.36) of Q_j we deduce that $\|Q_1\| \le C_1\,(1+Y_1^2)\,h^{1-\alpha}$, where $C_1 > 0$ is a constant. Therefore the left-hand side of (A.48) is bounded above by

$$E\big\{\big\|h^{-2}\,Q_1\big\|^2\,I\big(1+Y_1^2 > C_2\,n^{1/2}\,h^{1+\alpha}\big)\big\}\,,$$

where $C_2 = \epsilon/C_1$. If h satisfies (B8) then $n^{1/2}\,h^{1+2\alpha} \to \infty$, implying that $n^{1/2}\,h^{1+\alpha} \to \infty$. Therefore it suffices to prove that

$E[\|h^{-2}\,Q_1\|^2\,I\{m(X_1) > c$ or $\epsilon^2 > c\}]$ can be made arbitrarily small, uniformly in n, by choosing the constant $c > 0$ sufficiently large. (A.49)

Since Y and U, in the model at (1.1), are independent random variables then this can be done using the method in Step 5. Specifically, in all stages in the derivation of bounds to the components of the $p \times p$ matrix $E(Q_1 Q_1^{\mathrm{T}})$, multiply the argument of the expectation throughout by the random variable

$$J = I\{m(X_1) > c\} + I(\epsilon^2 > c)\,,$$

so that we bound instead the components of $E(Q_1 Q_1^{\mathrm{T}} J)$. In the string of identities leading to (A.40), multiply the arguments of the expectations by J, leading to a version of (A.41) in which each component of the $p \times p$ matrix of functions ψ has the same general form, representable as a sum of products of three functions of the individual variables x_1, x_2 and x as at (A.41). The argument leading to (A.43) produces the same bound as before, except that now the factor $(\sup\chi)\,(\sup|\phi_K|)\,(2\pi)^{-1}$ on the right-hand side of (A.43) can be replaced by a positive number $a(c)$ that decreases to zero as c increases. Therefore (A.45) continues to hold, except that $h^{-4}\, E(Q_1 Q_1^{\mathrm{T}})$ is replaced by $h^{-4}\, E(Q_1 Q_1^{\mathrm{T}} J)$, and the bound is multiplied by $a(c)$. In particular, our bound for each component of $h^{-4}\, E(Q_1 Q_1^{\mathrm{T}} J)$ is multiplied by the factor $a(c)$. This leads quickly to (A.49).

Step 7: Conclusion.

Condition (B8) prescribes the range of values θ in which we search for a solution of the equation $S(\theta) = 0$. Provided we confine attention to that region, (A.35) and (A.45) imply that with probability converging to 1 a solution exists in the range. This establishes part (i) of the theorem. Result (A.35) connects the vector V_2 directly to a solution $\widehat{\theta}$ of the equation $S(\theta) = 0$, and through this linkage, and the asymptotic normality derived in Step 6, part (ii) of the theorem follows directly. The limiting covariance matrix Σ is identified by (A.35), (A.46) and (A.47), and is given by $\Sigma = M_0^{-1}\, \Sigma_1\, M_0^{-1}$, where M_0 is as at (4.7) (see also (B7)) and Σ_1 is as in (A.46).

References

Buckley, M.J., Eagleson, G.K. and Silverman, B.W. (1988). The estimation of residual variance in nonparametric regression. *Biometrika* **75**, 189–199.

Breiman, L. (1996). Bagging predictors. *Machine Learning* **24**, 123–140.

Carroll, R.J. and Hall, P. (2004). Low order approximations in deconvolution and regression with errors in variables. *J. Roy. Statist. Soc.* Ser. B **66** 31–46.

Carroll, R.J., Kuchenhoff, H., Lombard, F. and Stefanski, L.A. (1996). Asymptotics for the SIMEX estimator in nonlinear measurement error models. *J. Amer. Statist. Assoc.* **91**, 242–250.

Carroll, R.J., Maca, J.D. and Ruppert, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika* **86**, 541–554.

Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C.M. (2006). *Measurement Error in Nonlinear Models*, 2nd Edn. Chapman and Hall CRC Press, Boca Raton.

Cheng, C.-L. and Schneeweiss, H. (1998). Polynomial regression with errors in the variables. *J. Roy. Statist. Soc.* Ser. B **60**, 189–199

Comte, F. and Taupin, M.-L. (2007). Nonparametric estimation of the regression function in an errors-in-variables model. *Statistica Sinica* **17**, 1065–1090.

Cook, J.R. and Stefanski, L.A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *J. Amer. Statist. Assoc.* **89**, 1314–1328.

Delaigle, A. (2008). An alternative view of the deconvolution problem. *Statistica Sinica* **18**, 1025–1045.

Delaigle, A., Fan, J. and Carroll, R.J (2009). Design-adaptive local polynomial estimator for the errors-in-variables problem. *J. Amer. Statist. Assoc.* **104**, 348–359.

Delaigle, A. and Hall, P. (2006). On the optimal kernel choice for deconvolution. *Statist. Probab. Lett.* **76**, 1594–1602.

Delaigle, A. and Hall, P. (2008a). Using SIMEX for smoothing-parameter choice in errors-in-variables problems. *J. Amer. Statist. Assoc.* **103**, 280–287.

Delaigle, A. and Hall, P. (2008b). Using SIMEX for smoothing-parameter choice in errors-in-variables problems: supplementary material, downloadable from `http://www.ms.unimelb.edu.au/~aurored`

Delaigle, A., Hall, P. and Meister, A. (2008). On Deconvolution with repeated measurements. *Ann. Statist.* **36**, 665–685.

Delaigle, A., Hall, P. and Qiu, P. (2006). Nonparametric methods for solving the Berkson errors-in-variables problem. *J. Roy. Statist. Soc.* Ser. B **68**, 201–220.

Delaigle, A., and Meister, A. (2007). Nonparametric regression estimation in the heteroscedastic errors-in-variables problem. *J. Amer. Statist. Assoc.* **102**, 1416–1426.

Dette, H., Munk, A. and Wagner, T. (1998). Estimating the variance in nonparametric regression — what is a reasonable choice? *J. Roy. Statist. Soc.* Ser. B **60**, 751–764.

Devanarayan, V. and Stefanski, L.A. (2002). Empirical simulation extrapolation for measurement error models with replicate measurements. *Statist. Probab. Lett.* **59**, 219-225.

Diggle, P. and Hall, P. (1993). A Fourier approach to nonparametric deconvolution of a density estimate. *J. R. Statist. Soc.* Ser. B **55**, 523–531

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

Fan, J. (1991a). Asymptotic normality for deconvolution kernel density estimators. *Sankhya A* **53**, 97–110.

Fan, J. (1991b). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* **19**, 1257–1272.

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

Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85**, 645–660.

Finkenstädt, B.F., Bjørnstad, O.N., Grenfell, B.T. (2002). A stochastic model for extinction and recurrence of epidemics: estimation and inference for measles outbreaks. *Biostatistics* **3**, 493–510.

Gasser, T., Sroka, L. and Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73**, 625–633.

Gleser, L.J. (1990). Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models. Statistical analysis of measurement error models with applications. *Contemp. Math.* **112**, 99–114

Hall, P., Kay, J.W. and Titterington, D.M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* **77**, 521–528.

Hall, P. and Marron, J.S. (1990). On variance estimation in nonparametric regression. Biometrika **77**, 415–419.

Hall, P. and Meister, A. (2007). A ridge-parameter approach to deconvolution. *Ann. Statist.*, **35**, 1535–1558.

Hasbrouck, J. (1986). A note on linear heteroscedasticity models. *Economics Letters*

22, 349–351.

Hsiao, C. (1989) Consistent estimation for some nonlinear errors-in-variables models. *J. Econometrics* **41**, 159–185

Huang, X.Z., Stefanski, L.A. and Davidian, M. (2006). Latent-model robustness in structural measurement error models. *Biometrika* **93** 53–64.

Kim, J. and Gleser, L.J. (2000). SIMEX approaches to measurement error in ROC studies. *Comm. Statist. Theory Meth.* **29**, 2473–2491.

Komlós, J., Major, P. and Tusnády, G. (1975). An approximation of partial sums of independent RV's and the sample DF. I. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **32**, 111–131.

Lavergne, P. and Vuong, Q.H. (1998). An integral estimator of residual variance and a measure of explanatory power of covariates in nonparametric regression. *J. Nonparam. Statist.* **9**, 363–380.

Levine, M. (2006). Bandwidth selection for a class of difference-based variance estimators in the nonparametric regression: a possible approach. *Comput. Statist. Data Anal.* **50**, 3405–3431.

Li, T. (2002). Robust and consistent estimation of nonlinear errors-in-variables models. *J. Econometrics* **110**, 1–26.

Lin, X.H. And Carroll, R.J. (2000). Nonparametric function estimation for clustered data when the predictor is measured without/with error. *J. Amer. Statist. Assoc.* **95**, 520–534.

Linton, O. and Whang, Y.J. (2002). Nonparametric estimation with aggregated data. *Econometric Theory* **18**, 420–468.

Meyer, K. (2005). A Statistical Method for MOS Transistor Mismatch Analysis and its Application during Semiconductor Process Development. *Australian Journal of Experimental Agriculture* **45**, 847–858.

Müller, H.-G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* **15**, 610–635.

Müller, H.-G. and Stadtmüller, U. (1992). On variance estimation with quadratic forms. *J. Statist. Plann. Inference* **35**, 213–231.

Müller, H.-G. and Zhao, P.-L. (1995). On a semiparametric variance function model and a test for heteroscedasticity. *Ann. Statist.* **23**, 946–967.

Müller, U.U., Schick, A. and Wefelmeyer, W. (2003). Estimating the error variance in nonparametric regression by a covariate-matched U-statistic. *Statistics* **37**, 179–188.

Munk, A., Bissantz, N., Wagner, T. and Freitag, G. (2005). On difference-based variance estimation in nonparametric regression when the covariate is high dimensional. *J. Roy. Statist. Soc.* Ser. B **67**, 19–41.

Nakamura, T. (1990). Corrected score functions for errors-in-variables models: methodology and application to generalized linear models. *Biometrika* **77**, 127–137.

Neumann, M.H. (1994). Fully data-driven nonparametric variance estimators. *Scand. J. Statist.* **25**, 189–212.

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

Ruppert, D., Wand, M.P., Holst, U. and Hossjer, O. (1997). Local polynomial variance-function estimation. *Technometrics* **39**, 262–273.

Schennach, S.M. (2004a). Estimation of nonlinear models with measurement error. *Econometrica* **72**, 33–75.

Schennach, S.M. (2004b). Nonparametric regression in the presence of measurement error. *Econometric Theory* **20**, 1046–1093.

Seifert, B., Gasser, T. and Wolf, A. (1993). Nonparametric-estimation of residual variance revisited. *Biometrika* **80**, 373–383.

Spokoiny, V. (2002). Variance estimation for high-dimensional regression models. *J. Multivar. Anal.* **82**, 111–133.

Sheehy, A., Gasser, T. and Rousson, V. (2005). Nonparametric measures of variance explained. *J. Nonparam. Statist.* **17**, 765–776.

Stefanski, L.A. (1989). Unbiased estimation of a nonlinear function of a normal mean with application to measurement error models. *Comm. Statist. A* **18**, 4335–4358.

Stefanski, L.A. (2000). Measurement error models. *J. Amer. Statist. Assoc.* **95**, 1353–1358.

Stefanski, L.A. and Carroll, R.J. (1987). Conditional scores and optimal scores in generalized linear measurement error models. *Biometrika* **74**, 703–716 . Stefanski, L.A. and Cook, J.R. (1995). Simulation-extrapolation: The measurement error jackknife. *J. Amer. Statist. Assoc.* **90**, 1247–1256.

Taupin, M.L. (2001). Semi-parametric estimation in the nonlinear structural errors-in-variables model. *Ann. Statist.* **29**, 66–93.

Tong, T. and Wang, Y. (2005). Estimating residual variance in nonparametric regression using least squares. *Biometrika* **92**, 821–830.

Yao, Q. and Tong, H. (1994). Quantifying the Influence of Initial Values on Non-Linear Prediction. *J. Roy. Statist. Soc.* Ser. B **56**, 701–725.