I N S T I T U T   D E

S T A T I S T I Q U E

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N

P A P E R

0211

BOOTSTRAP CONFIDENCE BANDS FOR REGRESSION CURVES AND THEIR DERIVATIVES

G. CLAESKENS and I. VAN KEILEGOM

http://www.stat.ucl.ac.be

Bootstrap Confidence Bands for Regression Curves and their Derivatives

Gerda Claeskens[1] Ingrid Van Keilegom[2]

Texas A&M University *Université catholique de Louvain*

Abstract

Confidence bands for regression curves and their first p derivatives are obtained via local p-th order polynomial estimation. The method allows for multiparameter local likelihood estimation as well as other unbiased estimating equations. As an alternative to the confidence bands obtained by asymptotic distribution theory, we also study smoothed bootstrap confidence bands. Simulations illustrate the finite sample properties of the methodology.

Key Words: Confidence band; Lack-of-fit test; Local estimation equations; Local polynomial estimation; Multiparameter local likelihood; One-step bootstrap, Smoothed bootstrap.

AMS 2000 subject classifications. Primary 62G15; secondary 62E20, 62G09.

Short title. Bootstrap confidence bands

[1]Department of Statistics, Texas A&M University, College Station, TX 77843, USA. E-mail: gerda@stat.tamu.edu

[2]Institut de Statistique, Université catholique de Louvain, Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium. The research of Van Keilegom was supported by 'Projet d'Actions de Recherche Concertées', No. 98/03–217, and by an IAP research network from the Belgian government.

1 Introduction

Local polynomial estimation of a regression curve has been studied for a variety of applications and models, ranging from the classical regression setting (Cleveland, 1979), where the response is modelled as its mean plus additive error, to local quasi-likelihood (Fan, Heckman and Wand, 1995), local multiparameter likelihood (Aerts and Claeskens, 1997), local pseudo-likelihood (Claeskens and Aerts, 2000), and local estimating equations (Carroll, Ruppert and Welsh, 1998).

In the previous work the theoretical focus has mainly been on obtaining consistency and asymptotic normality of the local polynomial estimators, hereby providing the necessary ingredients to construct pointwise confidence intervals for $\theta(x)$, the regression function of interest evaluated at x. This, however, is not sufficient to get an idea about the variability of the estimator of the whole curve, neither can it be used to correctly answer questions about the curve's shape. We go one step further. First, in Section 2, we consider a likelihood model $f(y;\theta(x))$ for the conditional density of the response Y given that the covariate X equals x, where the form of f is known, but $\theta(x)$ is unspecified, and we obtain the strong uniform consistency of local polynomial estimators for the regression curve $\theta(x)$ and for its derivatives up to order p, say, the degree of the polynomial chosen for estimation, hereby we extend the results of Zhao (1994) who focuses on local constant estimation in a likelihood setting.

Further, we use this result to obtain the limit distribution of the maximal deviation of $\sup_{x\in B}|\hat{\theta}(x)-\theta(x)|$ where the set B is compact. Following the original idea of Bickel and Rosenblatt (1973), this leads to the construction of confidence bands, but importantly, not only for the regression curve of interest, but also for its derivatives up to order p.

The method of proof largely follows the construction of Härdle (1989), who derives a maximal deviation result for local constant M-smoothers. An important difference though is that in local polynomial estimation we deal with a set of estimating equations instead of just a single one. Also some assumptions of earlier work will be relaxed. In a classical regression setting, Xia (1998) obtains a confidence band for a regression curve, using local linear estimators, hereby explicitly including a nonparametric bias estimator. Explicit bias correction has earlier been introduced by Eubank and Speckman (1993) for local constant kernel estimation in regression. We avoid bias estimation by a slight undersmoothing as compared to curve estimation. Undersmoothing is also advocated by Neumann and Polzehl

(1998), following earlier results of Hall (1991a, 1992) where it is shown that undersmoothing is more efficient than explicit bias correction, when the goal is to minimize the coverage error of the confidence band. Other approaches to construct confidence bands are investigated by Knafl, Sacks and Ylvisaker (1985), Hall and Titterington (1988) and Sun and Loader (1994) who assume bounds on the derivatives of the curves.

Most earlier methods proposed in the literature restrict to asymptotic confidence bands and since the convergence of normal extremes is known to be slow, see for example Hall (1979, 1991b), all these bands perform relatively poor for small sample sizes. The motivation of the present paper is to propose a solution to this problem by using a bootstrap approach. In particular, we apply the smoothed bootstrap (see Silverman and Young, 1987) to construct novel bootstrap based confidence bands, which unlike the asymptotic bands described above, avoid application of asymptotic distributions. The bootstrap confidence bands are available for the regression curve as well as for its derivatives. We show that the bootstrap works in asymptotically obtaining the correct nominal levels for the simultaneous confidence bands; a simulation study shows numerically the advantages of the bootstrap approach. Neumann and Polzehl (1998) construct a wild bootstrap in a simple regression setting of the form $Y = \mu(x) + \varepsilon$, without first obtaining results for asymptotic confidence bands. Their method is not immediately generalized to the (likelihood) models which are the object of our paper.

One of the main motivations to construct simultaneous confidence bands is to be able to answer graphical questions about the curves. For example, if a $100(1-\alpha)\%$ simultaneous confidence band for $\theta(.)$ over the set B does not contain any linear function, this is evidence against the null hypothesis that $\theta(\cdot)$ is linear in B. A lack of fit test is therefore an immediate application of the proposed confidence bands. A graphical representation of this confidence band indicates where possibly the null hypothesis is rejected. Moreover, confidence bands give an idea about the global variability of the estimator.

Although the above results are established for one-parameter likelihood models, we also study the extension to more than one parameter. In addition, the results are extended to the situation where estimating equations, different from the full local maximum likelihood equations, are considered.

The paper is organized as follows. In Section 2 we describe the model and the estimators and state the main results. Extensions to models with more than one parameter and to other estimation equations are given in Section 3. Section 4 deals with a simulation study, and technical proofs and regularity conditions are placed in Section 5.

2 Main Results for One-Parameter Models

In Section 2.1 we first define the estimators, obtain a strong uniform consistency rate, and study the order of the remainder term in a one-step approximation to the estimators. The construction of confidence bands based on a maximal deviation result is explained in Section 2.2 and an alternative bootstrap approach is derived in Section 2.3.

All results will be presented for one-parameter likelihood models, with a one-dimensional covariate. Extensions to other estimation schemes and models are discussed in Section 3. Let us introduce some notation and definitions.

Suppose we employ a likelihood model $f(y;\theta(x))$ for the conditional density of the response Y given that $X = x$, where the form of f is known, but the dependence of Y on the covariate x, via the regression curve $\theta(x)$, is unspecified. Assuming this function to be sufficiently smooth, in the sense that it possesses at least p continuous derivatives, locally $\theta(u)$ can be well approximated by

$$\theta(x,u) = \sum_{j=0}^{p} \theta_j(x)(u-x)^j,$$

where $\theta_j(x) = \theta^{(j)}(x)/j!$, for $j = 0, \ldots, p$. This is the idea behind local polynomial estimation (Cleveland, 1979) and has since been studied in various modelling frameworks by many authors, including Fan (1992, 1993), Ruppert and Wand (1994), Aerts and Claeskens (1997) and Carroll, Ruppert and Welsh (1998). For more references refer to Fan and Gijbels (1996).

For independent and identically distributed vectors (X_i, Y_i), $i = 1, \ldots, n$, the local log likelihood at $\boldsymbol{\theta}(x) = (\theta_0(x), \ldots, \theta_p(x))^t$ is defined as

$$\mathcal{L}_n\{\boldsymbol{\theta}(x)\} = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)\log f(Y_i; \theta(x, X_i))$$

where $K_h(\cdot) = K(\cdot/h_n)/h_n$ for a kernel function K and bandwidth sequence h_n, tending to zero as n tends to infinity. Staniswalis (1989) introduced local constant likelihood estimation, that is, taking $p = 0$, see also Zhao (1994).

For all regularity and smoothness conditions and assumptions on the kernel and bandwidth sequence, we refer to Section 5.

2.1 Strong Uniform Consistency

In a first theorem we obtain the strong uniform consistency of the local polynomial likelihood estimators, the proof of which is summarized in Section 5. Under slightly stronger assumptions, a similar result for local constant estimators is derived in Zhao (1994).

Theorem 2.1 *Assume conditions (H1a), (R0)-(R2). Then, there exists for all x in an interval B, a sequence of solutions $\{\hat{\boldsymbol{\theta}}(x)\}$ to the likelihood equations ($j = 0, \ldots, p$)*

$$\frac{\partial}{\partial \theta_j} \mathcal{L}_n(\boldsymbol{\theta}) = 0$$

such that for each $j = 0, \ldots, p$:

$$\sup_x |\hat{\theta}_j(x) - \theta_j(x)| = O(h_n^{-j}\{\log n/(nh_n)\}^{1/2} + h_n^{2([(p-j)/2]+1)}) \text{ a.s.},$$

where $[a]$ denotes the integer part of a.

For asymptotic normality of these estimators, we refer to Aerts and Claeskens (1997). Before giving a convenient matrix representation of the estimator's variance, we introduce some definitions. $\boldsymbol{N}_p$ and $\boldsymbol{T}_p$ are matrices of dimension $(p+1) \times (p+1)$ of which the $(i+1, j+1)$th entry equals $\nu_{i+j}(K) = \int u^{i+j} K(u) du$ and $\int u^{i+j} K^2(u) du$ respectively ($i, j = 0, \ldots, p$). The matrix $\boldsymbol{M}_{jp}(u)$ is obtained by replacing in $\boldsymbol{N}_p$ the $(j+1)$th ($j = 0, \ldots, p$) column by $(1, u, \ldots, u^p)^t$, and for $|\boldsymbol{N}_p| \neq 0$, we define the modified kernel function $K_{jp}(u) = K(u)|\boldsymbol{M}_{jp}(u)|/|\boldsymbol{N}_p|$. Note that from Fan, Heckman and Wand (1995, Proof of Theorem 1) it follows that $\int K_{jp}^2(u) du = (\boldsymbol{N}_p^{-1} \boldsymbol{T}_p \boldsymbol{N}_p^{-1})_{j+1,j+1}$.

Aerts and Claeskens (1997), in their Theorem 2, obtain that the asymptotic variance of the local polynomial estimator is given by $\boldsymbol{V}(\boldsymbol{\theta}(x)) = f_X^{-1}(x) I^{-1}(\theta(x))(\boldsymbol{N}_p^{-1} \boldsymbol{T}_p \boldsymbol{N}_p^{-1})$, where the local Fisher information number

$$I(\theta(x)) = -E_x\left\{\frac{\partial^2}{\partial \theta^2} \log f(Y; \theta(x))\right\} = E_x\left\{\frac{\partial}{\partial \theta} \log f(Y; \theta(x))\right\}^2.$$

Throughout, E_x denotes expectation conditional on $X = x$. Note that the last equality in the above definition holds by Bartlett's identities in a full likelihood model.

For the practical construction of confidence bands, an estimator of $\boldsymbol{V}(\boldsymbol{\theta}(x))$ is required. The proposed estimator is constructed by application of the delta-method, leading to the so-called sandwich covariance estimator; see Carroll, Ruppert and Welsh (1998) for an application in the setting of local estimating equations. For the current local likelihood estimators the variance estimator reads as follows. Define the column vector $\boldsymbol{X}_i = (1, \ldots, (X_i - x)^p)^t$ and the matrix $\boldsymbol{H}_n = \text{Diag}(1, \ldots, h_n^p)$. Let $\boldsymbol{B}_n(x)$ be the matrix of dimension $(p+1) \times (p+1)$

consisting of the second partial derivatives of $\mathcal{L}_n(x)$, rescaled with the appropriate power of the bandwidth, that is,

$$\boldsymbol{B}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)\frac{\partial^2}{\partial\theta^2}\log f(Y_i;\hat\theta(x,X_i))(\boldsymbol{H}_n^{-1}\boldsymbol{X}_i)(\boldsymbol{H}_n^{-1}\boldsymbol{X}_i)^t,$$

where $\hat\theta(x,u) = \sum_{j=0}^{p}\hat\theta_j(x)(u-x)^j$. As part of the variance estimator, we also define the matrix

$$\boldsymbol{K}_n(x) = \frac{1}{n}\sum_{i=1}^{n} h_n K_h^2(X_i - x)\left\{\frac{\partial}{\partial\theta}\log f(Y_i;\hat\theta(x,X_i))\right\}^2(\boldsymbol{H}_n^{-1}\boldsymbol{X}_i)(\boldsymbol{H}_n^{-1}\boldsymbol{X}_i)^t.$$

We define $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}}(x)) = \boldsymbol{B}_n^{-1}(x)\boldsymbol{K}_n(x)\boldsymbol{B}_n^{-1}(x)$ as an estimator of the variance $\boldsymbol{V}(\hat{\boldsymbol{\theta}}(x))$. In practice we might replace $\hat\theta(x,X_i)$ by $\hat\theta(X_i)$. Note that for local likelihood estimators alternative variance estimators can be proposed and used. For example, the entry in the first row and first column of the matrix $-\boldsymbol{B}_n(x)$ is an estimator of $f_X(x)I(\theta(x))$, suggesting the variance estimator $(-\boldsymbol{B}_n(x)_{1,1})^{-1}\boldsymbol{N}_p^{-1}\boldsymbol{T}_p\boldsymbol{N}_p^{-1}$. For specific likelihood models, $I(\theta(x))$ might be calculated exactly. The next corollary reports on a consistency property of the variance estimator $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}}(x))$ and obtains a bound on the error of a one-step approximation to the local polynomial estimator $\hat{\boldsymbol{\theta}}(x)$. To make the exposition more transparent, define $\boldsymbol{A}_n(x)$ to be the vector of first partial derivatives of $\mathcal{L}_n(x)$ with respect to $\boldsymbol{\theta}$, that is,

$$\boldsymbol{A}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)\frac{\partial}{\partial\theta}\log f(Y_i;\theta(x,X_i))\boldsymbol{X}_i \tag{2.1}$$

and define the matrix $\boldsymbol{J}(x) = f_X(x)I(\theta(x))\boldsymbol{N}_p(x)$. For future use, denote $g(x) = f_X(x)I(\theta(x))$.

Corollary 2.1 *Assume conditions (H1), (R0)–(R2).*

(i) For $j,k = 0,\ldots,p$, $\sup_x|\hat V_{jk}(\boldsymbol{\theta}(x)) - V_{jk}(\boldsymbol{\theta}(x))| = o_P\{(nh_n\log n)^{-1/2}\}$.

(ii) The following representation holds:

$$\boldsymbol{H}_n(\hat{\boldsymbol{\theta}}(x) - \boldsymbol{\theta}(x)) = \boldsymbol{J}(x)^{-1}\boldsymbol{H}_n^{-1}\boldsymbol{A}_n(x) + \boldsymbol{R}_n(x),$$

where, for $j = 0,\ldots,p$, $\sup_x|R_{nj}(x)| = o_P\{(nh_n\log n)^{-1/2}\}$.

Part *(ii)* of this corollary provides the basis for the construction of one-step estimators, hereby ignoring the remainder term in the expansion. In a local quasi-likelihood setting, Fan and Chen (1999) study one-step estimators based on a similar expansion replacing $\boldsymbol{J}(x)$ by its empirical version $-\boldsymbol{B}_n(x)$. These one-step estimators are particularly useful in otherwise heavily computational intensive bootstrap methods. A similar one-step approximation has been used in a bootstrap approach by Claeskens and Aerts (2000). We will come back to this representation for the construction of bootstrap confidence bands.

2.2 Asymptotic Confidence Bands

In this section we extend the asymptotic distribution theory for local polynomial estimators in a likelihood framework with the intend of constructing confidence bands based on asymptotic distribution theory for extremes of Gaussian processes, dating back to Bickel and Rosenblatt (1973). In Section 2.3 we develop a bootstrap method to achieve the same goal.

The construction of a confidence band for the components of $\boldsymbol{\theta}(x)$ goes along a series of 5 steps, quite similar to those used in Härdle (1989) and Johnston (1982). A major difference between our approach and the one used in these two articles, is that we do not impose certain technical assumptions on f, that depend on a sequence $\{a_n\}$ that tends to infinity. See the appendix for a more detailed discussion on this issue. Another important difference with the results in Härdle (1989) is that we do not have a single estimating equation, but instead have a set of $p+1$ equations, one for each component $\theta_j(x)$, $j = 0, \ldots, p$. A welcome consequence of working with the vector $\boldsymbol{\theta}(x)$ is that it enables us to obtain confidence bands not only for the curve $\theta(\cdot)$, but also for its derivatives $\theta^{(j)}(\cdot)$ up to order p. This feature of local polynomial approximation for confidence band construction has, at least to our knowledge, not been explored before, not even in simpler regression settings. Another difference with earlier work is that we do not assume the boundedness of our estimating equations, that is, the score vector does not need to be bounded as a function of the response y, which would be an important restriction as it for example excludes application to normal regression models.

Define for $j = 0, \ldots, p$,

$$Y_{nj}(x) = (nh_n)^{1/2} h_n^{-j} \{I(\theta(x)) f_X(x)\}^{-1/2} A_{nj}(x).$$

Further, define the Rosenblatt transformation (Rosenblatt, 1952)

$$M(x, y) = (F_X(x), F(y; \theta(x))),$$

transforming (X, Y) into $(F_X(X), F(Y; \theta(X)))$, which is uniformly distributed on the unit square $[0, 1]^2$. In the next lemmas we establish asymptotically equivalent expressions for $Y_{nj}(x)$, which will eventually lead to the construction of confidence bands.

Lemma 2.1 *Assume conditions (H),(R0)-(R3). Let*

$$Y_{nj1}(x) = h_n^{1/2} g(x)^{-1/2} \int\int K_h(z-x) \left(\frac{z-x}{h_n}\right)^j \frac{\partial}{\partial\theta} \log f(y;\theta(x,z))\, dV_n(M(z,y)),$$

where $\{V_n\}$ is a sequence of 4-sided tied-down Wiener processes, defined in Lemma 5.1. Then, for $j = 0, \ldots, p$,

$$\sup_x |Y_{nj}(x) - Y_{nj1}(x)| = o_P((\log n)^{-1/2}).$$

Lemma 2.2 *Assume conditions (H),(R0)-(R3). Let*

$$Y_{nj2}(x) = h_n^{1/2} g(x)^{-1/2} \int\int K_h(z-x) \left(\frac{z-x}{h_n}\right)^j \frac{\partial}{\partial\theta} \log f(y;\theta(z))\, dV_n(M(z,y)).$$

Then, for $j = 0, \ldots, p$,

$$\sup_x |Y_{nj1}(x) - Y_{nj2}(x)| = O_P(h_n^{1/2}).$$

Lemma 2.3 *Assume conditions (H),(R0)-(R3). Let*

$$Y_{nj3}(x) = h_n^{1/2} g(x)^{-1/2} \int\int K_h(z-x) \left(\frac{z-x}{h_n}\right)^j \frac{\partial}{\partial\theta} \log f(y;\theta(z))\, dW_n(M(z,y)),$$

where $\{W_n\}$ is a sequence of standard bivariate Wiener processes satisfying $V_n(u,v) = W_n(u,v) - uW_n(1,v) - vW_n(u,1) + uvW_n(1,1)$. Then, for $j = 0, \ldots, p$,

$$\sup_x |Y_{nj2}(x) - Y_{nj3}(x)| = O_P(h_n^{1/2}).$$

Lemma 2.4 *Assume conditions (H),(R0)-(R3). Let*

$$Y_{nj4}(x) = h_n^{1/2} g(x)^{-1/2} \int\int K_h(z-x) \left(\frac{z-x}{h_n}\right)^j g(z)^{1/2}\, dW(z),$$

where W is the Wiener process on the support of X. Then, for $j = 0, \ldots, p$, $Y_{nj3}(x)$ and $Y_{nj4}(x)$ have the same distribution.

Lemma 2.5 *Assume conditions (H),(R0)-(R3). Let*

$$Y_{nj5}(x) = h_n^{1/2} \int K_h(z-x) \left(\frac{z-x}{h_n}\right)^j dW(z).$$

Then, for $j = 0, \ldots, p$,

$$\sup_x |Y_{nj4}(x) - Y_{nj5}(x)| = O_P(h_n^{1/2}).$$

The following theorem gives a maximal absolute deviation result for the local polynomial estimators of the curve $\theta(\cdot)$ and of its derivatives up to order p. First, we need some definitions. Let $\boldsymbol{Q}_p$ be a matrix of dimension $(p+1)\times(p+1)$ of which the $(i+1,j+1)$th entry equals $\int u^{i+j}\{K'(u)\}^2 du - \frac{1}{2}\{i(i-1)+j(j-1)\}\int u^{i+j-2}K^2(u)du$, $(i,j=0,\ldots,p)$. Further, for $j=0,\ldots,p$, define the kernel dependent constant $C_j = (\boldsymbol{N}_p^{-1}\boldsymbol{Q}_p\boldsymbol{N}_p^{-1})_{j+1,j+1}/\int K_{jp}^2$.

Theorem 2.2 *Assume conditions (H),(R0)-(R3), and define the random variable*

$$Z_{nj}(x) = (nh_n^{2j+1})^{1/2}(\hat{\theta}_j(x)-\theta_j(x))\left\{\hat{V}_{jj}(\hat{\boldsymbol{\theta}}(x))\right\}^{-1/2},$$

and the sequence $z_{nj} = (-2\log h_n)^{1/2} + (-2\log h_n)^{-1/2}\log(C_j^{1/2}/(2\pi))$, *then, for* $j=0,\ldots,p$,

$$P\{(-2\log h_n)^{1/2}(\sup_{x\in B}|Z_{nj}(x)|-z_{nj})<z\}\to\exp(-2\exp(-z)).$$

This theorem leads immediately to the definition of the confidence bands for the components of $\boldsymbol{\theta}(x)$.

Corollary 2.2 *Assume conditions (H),(R0)-(R3). A* $(1-\alpha)100\%$ *confidence band for* $\theta^{(j)}(\cdot)$ $(j=0,\ldots,p)$ *over region* B, *is given by the collection of all curves* ϑ_j *belonging to the set*

$$\left\{\vartheta_j : \sup_{x\in B}[|j!\hat{\theta}_j(x)-\vartheta_j(x)|\{\hat{V}_{jj}(\hat{\boldsymbol{\theta}}(x))\}^{-1/2}] \le L_{\alpha j}\right\},$$

where for $j=0,\ldots,p$,

$$L_{\alpha j} = j!(nh_n^{2j+1})^{-1/2}\{(-2\log h_n)^{1/2} + (-2\log h_n)^{-1/2}\{x_\alpha + \log(C_j^{1/2}/(2\pi))\}\},$$

and $x_\alpha = -\log\{-0.5\log(1-\alpha)\}$.

This approach contrasts the explicit bias corrected confidence bands of Eubank and Speckman (1993) and Xia (1998), where, due to forcing the bandwidth to be optimal for estimation, a bias correction term needs to be added to the curve estimator in order to be able to obtain correctly centered confidence bands, that is, bands for the regression curves themselves, not for their respective expected values of the estimated curves. Explicit "undersmoothing" for the construction of a confidence band in heteroscedastic regression models is proposed by Neumann and Polzehl (1998), following results earlier obtained by Hall (1991a, 1992) in the context of density estimation. There it is shown that undersmoothing outperforms bias correction, one explanation for this might be the large variability of bias estimators.

It is only to be expected that for the construction of confidence bands, where coverage error is the relevant criterion, a different bandwidth is to be used, as compared to curve estimation where a criterion such as the mean (integrated) squared error determines the amount of smoothing. Our theoretical investigation shows that a bandwidth of order $o((n \log n)^{-1/(4[(p-j)/2]+2j+5)})$, which goes to zero faster than the optimal rate for curve estimation by a factor $o((\log n)^{-1/(2p+3)})$ for $p-j$ odd, and $o((\log n)^{-1/(2p+5)})$ for $p-j$ even, is sufficient for the results to hold. We do not prove this rate to be optimal in any sense.

In our numerical simulation work we use the bandwidth sequence which minimizes coverage error, by performing a grid search. An interesting topic of further research is a determination of a data-driven 'optimal' bandwidth based on a theoretical study of the coverage error. For practical use in data analysis, a data-driven bandwidth can be obtained minimizing a bootstrap coverage error criterion. The procedure reads as follows. Resample the data a large number, say B times, and, for a given bandwidth, construct a confidence band for each of the bootstrap data sets by the method discussed above. The coverage probability of these bands is computed as the proportion of bands which include the estimated curve $\hat{\theta}_j(\cdot)$. We may now repeat this for a number of bandwidths and obtain that bandwidth for which the coverage probability is equal to the desired nominal coverage level.

2.3 Bootstrap confidence bands

In this section we propose a bootstrap procedure and construct a bootstrap confidence band for the unknown $\theta(x)$ and its derivatives up to order p, which serves as an alternative to the bands constructed in Corollary 2.2. The latter bands are based on the asymptotic results of Theorem 2.2, and as will be seen in the simulations, the convergence to this asymptotic distribution is quite slow. For relatively small sample sizes the bootstrap band is therefore a useful alternative. The bands we propose in this section can easily be adopted to the classical nonparametric regression context, where these bands and in particular those of the derivative curves have, to our knowledge, never been proposed.

We generate bootstrap resamples by using a smoothed bootstrap procedure. Let $g_n = h_n\hat{\sigma}_Y/\hat{\sigma}_X$ and

$$\hat{f}(x,y) = \frac{1}{nh_n g_n}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}, \frac{Y_i - y}{g_n}\right),$$

be the bivariate density estimator of (X,Y), where $\hat{\sigma}_X$ and $\hat{\sigma}_Y$ are the sample standard deviation of X and Y. The bootstrap resamples $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$ are n independent pairs, where for each i, $(X_i^*, Y_i^*) \sim \hat{f}$. We refer to Silverman and Young (1987) for more details about the smoothed bootstrap. We chose to use in the definition of $\hat{f}$ the same bandwidth h_n as before, since simulations indicate that this choice works well in practice. It is possible however, to work with different bandwidths for the bootstrap and for the construction of the confidence bands. The reason for generating bootstrap data from $\hat{f}$, rather than from the bivariate empirical distribution of the observations (X_i, Y_i) $(i = 1, \ldots, n)$, is that the asymptotic theory requires the bootstrap distribution to be smooth. Using these bootstrap data, we construct one-step Fisher-scoring estimators in the bootstrap world, hereby avoiding any iterative calculation methods to obtain the bootstrap estimators. We propose the following bootstrap analogue of the variance estimator $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}}(x))$:

$$\hat{\boldsymbol{V}}^*(\hat{\boldsymbol{\theta}}(x)) = \boldsymbol{B}_n^{-1}(x)\boldsymbol{K}_n^*(x)\boldsymbol{B}_n^{-1}(x),$$

where

$$\boldsymbol{K}_n^*(x) = \frac{1}{n}\sum_{i=1}^n h_n K_h^2(X_i^* - x)\left\{\frac{\partial}{\partial\theta}\log f(Y_i^*; \hat{\theta}(x, X_i^*))\right\}^2 (\boldsymbol{H}_n^{-1}\boldsymbol{X}_i^*)(\boldsymbol{H}_n^{-1}\boldsymbol{X}_i^*)^t,$$

where $\boldsymbol{X}_i^* = (1, \ldots, (X_i^* - x)^p)^t$. In a similar fashion we obtain the bootstrap local score vector $\boldsymbol{A}_n^*(x)$,

$$\begin{aligned}\boldsymbol{A}_n^*(x) = &\frac{1}{n}\sum_{i=1}^n K_h(X_i^* - x)\frac{\partial}{\partial\theta}\log f(Y_i^*; \hat{\theta}(x, X_i^*))\boldsymbol{X}_i^* \\ &-\frac{1}{n}\sum_{i=1}^n E^*[K_h(X_i^* - x)\frac{\partial}{\partial\theta}\log f(Y_i^*; \hat{\theta}(x, X_i^*))\boldsymbol{X}_i^*],\end{aligned}$$

where E^* denotes expectation, conditionally on the data (X_i, Y_i) $(i = 1, \ldots, n)$. For future use, let $\hat{f}_X(x) = n^{-1}\sum_{i=1}^n K_h(X_i - x)$ and $\hat{F}(y; \theta(x)) = \hat{f}_X^{-1}(x)\int_{-\infty}^y \hat{f}(x, t)\, dt$.

We can now define $\hat{\boldsymbol{\theta}}^*(x)$, the bootstrap estimator of $\boldsymbol{\theta}(x)$:

$$\boldsymbol{H}_n(\hat{\boldsymbol{\theta}}^*(x) - \hat{\boldsymbol{\theta}}(x)) = -\boldsymbol{B}_n^{-1}(x)\boldsymbol{H}_n^{-1}\boldsymbol{A}_n^*(x),$$

which we call a one-step estimator, since the above equation is the bootstrap analogue of the one-term expansion of $\hat{\boldsymbol{\theta}}(x)$ given in Corollary 2.1 (without the remainder term). See also Aerts and Claeskens (2000), for a similar bootstrap estimator of $\boldsymbol{\theta}(x)$.

In the next theorem we establish the bootstrap analogue of Theorem 2.2. The proof is given in Section 5.

Theorem 2.3 *Assume conditions (H),(R0)-(R4), and define the random variables*

$$Z_{nj}^{(1)*}(x) = (nh_n^{2j+1})^{1/2}(\hat{\theta}_j^*(x) - \hat{\theta}_j(x))\{\hat{V}_{jj}(\hat{\boldsymbol{\theta}}(x))\}^{-1/2}$$
$$Z_{nj}^{(2)*}(x) = (nh_n^{2j+1})^{1/2}(\hat{\theta}_j^*(x) - \hat{\theta}_j(x))\{\hat{V}_{jj}^*(\hat{\boldsymbol{\theta}}(x))\}^{-1/2},$$

then, for $j = 0, \ldots, p$ and $k = 1, 2$,

$$P^*\{(-2\log h_n)^{1/2}(\sup_{x \in B}|Z_{nj}^{(k)*}(x)| - z_{nj}) < z\} \to \exp(-2\exp(-z)) \qquad a.s.,$$

where z_{nj} is defined in Theorem 2.2 and where P^* denotes probability, conditionally on the data (X_i, Y_i) $(i = 1, \ldots, n)$.

This result together with Theorem 2.2 shows that $\sup_x |Z_{nj}(x)|$ and $\sup_x |Z_{nj}^{(k)*}(x)|$ $(k = 1, 2)$ have the same limiting distribution. From this, we directly obtain two bootstrap confidence bands for the components of $\boldsymbol{\theta}(x)$.

Corollary 2.3 *Assume conditions (H),(R0)-(R4). A $(1-\alpha)100\%$ confidence band for $\theta^{(j)}(\cdot)$ $(j = 0, \ldots, p)$ over region B, is given by the collection of all curves ϑ_j belonging to the set*

$$\Big\{\vartheta_j : \sup_{x \in B}[|j!\hat{\theta}_j(x) - \vartheta_j(x)|\{\hat{V}_{jj}(\hat{\boldsymbol{\theta}}(x))\}^{-1/2}] \leq L_{\alpha j}^{(k)*}\},$$

$(k = 1, 2)$, where the bound $L_{\alpha j}^{(1)*}$ satisfies

$$P^*(j!\sup_{x \in B}[|\hat{\theta}_j^*(x) - \hat{\theta}_j(x)|\{\hat{V}_{jj}(\hat{\boldsymbol{\theta}}(x))\}^{-1/2}] \leq L_{\alpha j}^{(1)*}) = 1 - \alpha,$$

and $L_{\alpha j}^{(2)*}$ satisfies

$$P^*(j!\sup_{x \in B}[|\hat{\theta}_j^*(x) - \hat{\theta}_j(x)|\{\hat{V}_{jj}^*(\hat{\boldsymbol{\theta}}(x))\}^{-1/2}] \leq L_{\alpha j}^{(2)*}) = 1 - \alpha.$$

For selecting an appropriate bandwidth in practice, we propose the following method, which is based on a double bootstrap procedure and which is similar in nature to the procedure for selecting a bandwidth for asymptotic confidence bands described at the end of Section 2.2. Let

$$I_j(h_n) = P(\hat{\ell}_{j,h}(x) \leq \theta^{(j)}(x) \leq \hat{u}_{j,h}(x) \text{ for all } x \in B)$$

be the (unknown) coverage probability of a confidence band for $\theta^{(j)}(x)$ using bandwidth h_n, where $\hat{\ell}_{j,h}(x)$ and $\hat{u}_{j,h}(x)$ are the lower and upper bounds given in Corollary 2.3. To estimate $I_j(h_n)$, generate for each of M_1 bootstrap samples a set of M_2 (say) second-level bootstrap resamples and construct lower bounds $\hat{\ell}^{*(m_1)}_{j,h}(x)$ and upperbounds $\hat{u}^{*(m_1)}_{j,h}(x)$ using these second-level resamples ($m_1 = 1, \ldots, M_1$). Define now

$$\begin{aligned}\hat{I}_j(h_n) &= P^*(\hat{\ell}^*_{j,h}(x) \leq \hat{\theta}^{(j)}(x) \leq \hat{u}^*_{j,h}(x) \text{ for all } x \in B)\\ &= \frac{1}{M_1}\sum_{m_1=1}^{M_1} I(\hat{\ell}^{*(m_1)}_{j,h}(x) \leq \hat{\theta}^{(j)}(x) \leq \hat{u}^{*(m_1)}_{j,h}(x) \text{ for all } x \in B).\end{aligned}$$

Next, calculate $\hat{I}_j(h_n)$ for a grid of h_n-values and obtain the value of h_n for which $\hat{I}_j(h_n) = 1-\alpha$ by using e.g. interpolation.

3 Applications and Extensions

One important class of likelihood models are the generalized linear models, where the conditional density of the response Y given the covariate x belongs to a one-parameter exponential family,

$$f(y; \theta(x)) = \exp\{y\theta(x) - b(\theta(x)) + c(y)\}$$

for known functions b and c, see, e.g., McCullagh and Nelder (1989). Well-known members of this family are the known variance normal regression model, and the inverse Gaussian and gamma distributions. The results in Section 2 immediately apply to these generalized linear models.

3.1 Multiparameter Likelihood

Although the class of one-parameter models is already quite large, it does not include the commonly used Gaussian heteroscedastic regression model $Y_i = \mu(X_i) + \sigma(X_i)\varepsilon_i$ where the ε_i are standard normal random variables. In the two-parameter model $f(y_i; \theta_1(x_i), \theta_2(x_i))$ both curves, e.g. $\theta_1(x) = \mu(x)$ and $\theta_2(x) = \sigma^2(x)$ can be estimated, simultaneously, by local polynomial estimators, see Aerts and Claeskens (1997). There it is advised to take both polynomials of equal degree p, which we will also assume here. Note that a link function can be included at this stage. The two-parameter normal model is actually a special example since $\mu(x)$ and $\sigma^2(x)$ are orthogonal, implying that a confidence band for $\mu(x)$ may be constructed as in the one-parameter case, provided an estimator for the variance $\sigma^2(x)$ is available. We might use the estimator resulting from the above described simultaneous estimation procedure or use any other estimator such as for example the variance estimator of Ruppert, Wand, Holst and Hössjer (1997). Other interesting multiparameter models include the generalized extreme value and Pareto distributions, see, respectively, Davison and Ramesh (2000) and Beirlant and Goegebeur (2000) for application of local polynomial estimation.

Without loss of generality we restrict attention to two parameter models. For local pth order estimation in a two-parameter model, there now is a $2(p+1)$ dimensional vector $\boldsymbol{\theta}(x) = (\boldsymbol{\theta}_1(x)^t, \boldsymbol{\theta}_2(x)^t)^t$, where, for $r = 1, 2$, $\theta_{rj}(x) = \theta_r^{(j)}(x)/j!$.

Almost sure consistency and a maximal deviation result are obtained in this multiparameter setting, for which we introduce the following notation. For $r = 1, 2$, the $(p+1)$ dimensional vectors $\boldsymbol{A}_{nr}$ are defined similarly as $\boldsymbol{A}_n$, but now taking partial derivatives,

$$\boldsymbol{A}_{nr}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)\frac{\partial}{\partial\theta_r}\log f(Y_i; \theta(x, X_i))\boldsymbol{X}_i.$$

The 2×2 dimensional local Fisher information matrix $\boldsymbol{I}(\theta_1(x), \theta_2(x))$ has as (r, s)th entry $E_x[-(\partial^2/\partial\theta_r\partial\theta_s)\log f(Y; \theta_1(x), \theta_2(x))]$. Further, the earlier defined matrix $\boldsymbol{J}(x)$ is now equal to $\boldsymbol{J}(x) = f_X(x)\boldsymbol{I}(\theta_1(x), \theta_2(x)) \otimes \boldsymbol{N}_p(x)$, where $\otimes$ denotes the Kronecker product.

Crucial for the derivation is the asymptotic behavior of the random variables

$$Y_{nrj}(x) = (nh_n)^{1/2}h_n^{-j}\left(\boldsymbol{I}^{-1}(\theta_1(x), \theta_2(x))\right)_{rr}^{1/2} f_X^{-1/2}(x)A_{nrj}(x),$$

where $r = 1, 2$ and $j = 0, \ldots, p$. A general theorem is presented in Section 3.3.

3.2 Other Estimating Equations

When the functional form of the likelihood of the data is not known, or when we do not wish to use a full likelihood approach, other estimation schemes are available. We first give some examples, before deriving the results.

Our first example is pseudolikelihood estimation (Arnold and Strauss, 1991), where for a multivariate response vector, the joint density of the data is replaced by a product of conditional densities which do not necessarily represent a joint density. The motivation behind this estimation technique is to avoid the calculation of a complicated normalizing constant, which frequently arises in exponential family models (Arnold, Castillo and Sarabia, 1992). Let A represent the set of all $2^m - 1$ vectors $\mathbf{a}$ of length m=dim($\boldsymbol{Y}$), consisting solely of zeros and ones, with each vector having at least one non-zero entry, and $\{\gamma_{\mathbf{a}} \mid \mathbf{a} \in A\}$

is a set of $2^m - 1$ given real numbers, not all zero. Denote by $\boldsymbol{Y}^{(\mathbf{a})}$ the subvector of $\boldsymbol{Y}$ corresponding to the non-zero components of $\mathbf{a}$ with associated (marginal) density function $f^{(\mathbf{a})}(\boldsymbol{y}^{(\mathbf{a})}; \theta_1(x), \theta_2(x))$. In a two-parameter model the logarithm of the local pseudolikelihood is defined as

$$\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x) \sum_{\mathbf{a}\in A} \gamma_{\mathbf{a}} \log f^{(\mathbf{a})}(\boldsymbol{Y}_i^{(\mathbf{a})}; \theta_1(x, X_i), \theta_2(x, X_i)).$$

We refer to Claeskens and Aerts (2000) for more about local polynomial estimation in these models, also for the situation that the functions $f^{(\mathbf{a})}$ are possibly misspecified.

In cases where the likelihood of the data is not available, a quasi-likelihood function can specify the relationship between mean and variance of the response (McCullagh and Nelder 1989, Wedderburn 1974). For example, let $V(\mu(x))$ represent the conditional variance of $\boldsymbol{Y}$ given x, then the quasi-likelihood function Q is such that $(\partial/\partial\mu)Q(y;\mu) = (y-\mu)/V(\mu)$. Fan, Heckman and Wand (1995) introduced local polynomial estimation in such models by defining the estimators to be the maximizers with respect to components of $\boldsymbol{\theta}(x)$, of the following function

$$\sum_{i=1}^{n} Q(Y_i; g^{-1}(\theta(x, X_i))K_h(X_i - x).$$

In the above, g is a known link function, such that $\hat{\theta}(x) = g(\hat{\mu}(x))$. For one-parameter exponential family models, using the canonical link function, quasi-likelihood and likelihood estimation coincide.

All of the above examples, including M-estimation (Huber, 1967), fit in the general estimating equations framework. Without loss of generality, assume that there are two parameter functions $\boldsymbol{\theta}(x) = (\theta_1(x), \theta_2(x))$ for which $\psi_1(Y; \theta_1(x), \theta_2(x))$ and $\psi_2(Y; \theta_1(x), \theta_2(x))$ are two unbiased estimating equations in the sense that $E_x\{\psi_r(Y; \boldsymbol{\theta}(x))\} = 0$, for $r = 1, 2$. Local polynomial estimators, both of degree p, are solutions to the following set of equations

$$\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)\psi_r(Y_i; \theta_1(x, X_i), \theta_2(x, X_i))\boldsymbol{X}_i = 0, \quad r = 1, 2.$$

We refer to Carroll, Ruppert and Welsh (1998) for more details and examples about local estimating equations. This estimating framework is particularly useful for a multivariate response vector. It is straightforward to extend the theorems to models with more than two parameter functions.

3.3 Main Result

In order to formulate our main theorem, define the 2×2 matrices $\boldsymbol{I}$ and $\boldsymbol{K}$ with (r, s)th component

$$I_{rs}(\boldsymbol{\theta}(x)) = E_x\{-\frac{\partial \psi_r}{\partial \theta_s}(Y; \boldsymbol{\theta}(x))\}, \quad \text{and} \quad K_{rs}(\boldsymbol{\theta}(x)) = E_x\{\psi_r(Y; \boldsymbol{\theta}(x))\psi_s(Y; \boldsymbol{\theta}(x))\}.$$

Note that for local likelihood estimation where $\psi_r = (\partial/\partial\theta_r)\log f$, by Bartlett's identities, the matrices $\boldsymbol{I}$ and $\boldsymbol{K}$ coincide.

Further, let

$$V_{r;jj}(\boldsymbol{\theta}(x)) = \left(\boldsymbol{I}^{-1}(\boldsymbol{\theta}(x))\boldsymbol{K}(\boldsymbol{\theta}(x))\boldsymbol{I}^{-1}(\boldsymbol{\theta}(x))\right)_{rr} f_X^{-1}(x) \int K_{jp}^2$$

be the (j, j)-th component $(j = 0, \dots, p)$ of the asymptotic variance matrix associated with local polynomial estimation of $\theta_r(x)$ $(r = 1, 2)$. As its estimator we define

$$\hat{V}_{r;jj}(\hat{\boldsymbol{\theta}}(x)) = \left(\boldsymbol{B}_n^{-1}(x)\boldsymbol{K}_n(x)\boldsymbol{B}_n^{-1}(x)\right)_{\tilde{r}\tilde{j}},$$

where $\boldsymbol{B}_n$ and $\boldsymbol{K}_n$ are partitioned matrices with $(p+1)\times(p+1)$ dimensional submatrices $\boldsymbol{B}_{n;rs}$ and $\boldsymbol{K}_{n;rs}$ respectively $(r, s = 1, 2)$,

$$\begin{aligned}
\boldsymbol{B}_{n,rs} &= \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)\frac{\partial}{\partial \theta_s}\psi_r(Y_i; \hat{\boldsymbol{\theta}}(x, X_i))(\boldsymbol{H}_n^{-1}\boldsymbol{X}_i)(\boldsymbol{H}_n^{-1}\boldsymbol{X}_i)^t \\
\boldsymbol{K}_{n,rs} &= \frac{1}{n}\sum_{i=1}^{n} h_n K_h^2(X_i - x)\psi_r(Y_i; \hat{\boldsymbol{\theta}}(x, X_i))\psi_s(Y_i; \hat{\boldsymbol{\theta}}(x, X_i))(\boldsymbol{H}_n^{-1}\boldsymbol{X}_i)(\boldsymbol{H}_n^{-1}\boldsymbol{X}_i)^t.
\end{aligned}$$

The subscript notation $\tilde{r}, \tilde{j}$ denotes that we take the jth diagonal entry of the (r, r)th submatrix.

We now present the construction of confidence bands for multiparameter curves using local estimating equations.

Theorem 3.1 *Assume conditions (H), (R0) and (R1')–(R3'). A* $(1-\alpha)100\%$ *confidence band for the* j*th derivative* $(j = 0, \dots, p)$ *of* $\theta_r(\cdot)$ $(r = 1, 2)$ *over region* B*, is the collection of all curves* ϑ_{rj} *belonging to the set*

$$\{\vartheta_{rj} : \sup_{x\in B}[|j!\hat{\theta}_{rj}(x) - \vartheta_{rj}(x)|\{\hat{V}_{r;jj}(\hat{\boldsymbol{\theta}}(x))\}^{-1/2}] \le L_{\alpha j}\},$$

where $L_{\alpha j}$ *is as in Corollary 2.2.*
Bootstrap confidence bands are constructed analogously as in Corollary 2.3.

3.4 Application to lack of fit testing

Checking whether a curve $\vartheta(\cdot)$ belongs to a simultaneous confidence band for $\theta^{(j)}(\cdot)$ corresponds to testing the null hypothesis $H_0 : \theta^{(j)}(x) = \vartheta(x)$ for all $x \in B$ versus the alternative hypothesis $H_a : \theta^{(j)}(x) \neq \vartheta(x)$ for some $x \in B$. It is readily obtained that this coincides to comparing the value of the test statistic

$$\sup_{x\in B} \left\{ \hat{V}_{jj}(\hat{\boldsymbol{\theta}}(x))^{-1/2} |j!\hat{\theta}_j(x) - \vartheta(x)| \right\}$$

with the critical value $L_{\alpha j}$, defined in Corollary 2.2. Rejection takes place for values bigger than the critical value. For a similar hypothesis test, Fan and Zhang (2000) showed that a composite null hypothesis $H_0 : \theta^{(j)}(x) = \vartheta(x, \nu)$, with the parameter vector ν unspecified, can be dealt with in a similar way, substituting ν by a root-n consistent estimator.

The power of this test will suffer from the same convergence problems as discussed before, therefore, for small or moderately sized samples, we advise to use a bootstrap version of the test instead. We refer to Aerts and Claeskens (2001), where a procedure to generate bootstrap data under the null model is proposed for (generalized) estimating equations models. Note that in certain situations, the model under the null hypothesis is completely specified, and hence in these cases, data can be generated directly from the true model, instead of making use of asymptotic or bootstrap results. This is for example the case for local likelihood models containing a single parameter $\theta(x)$, which is fully specified under H_0.

4 Simulations

In this section we apply the methods discussed above in a simulation study. For comparison purposes we also include the bias-corrected confidence bands of Xia (1998) with plug-in bandwidth choice. We construct confidence bands based on local linear estimators ($p = 1$), and use the Epanechnikov kernel function. For the confidence bands based on the asymptotic distribution theory, as well as for the two bootstrap methods, we perform a grid search to find the bandwidth which minimizes simulated coverage error.

We generate data from the normal regression model

$$Y_i = \mu(X_i) + \sigma(X_i)\varepsilon_i,$$

where the independent regression variables X_i have a uniform distribution on the unit interval, $\mu(x) = x(1-x)$ and the error terms ε_i are independent standard Gaussian random variables.

In the first setting we take $\sigma(x) = 0.1$. Table 1 presents the simulation results for three sample sizes $n = 50, 100$ and 200, and two nominal values for coverage probabilities: 90% and 95%. Simulated coverage probabilities together with the calculated area of the confidence band are obtained for the curve $\mu(\cdot)$ as well as for its first derivative $\mu'(\cdot)$. Note that the method by Xia (1998) does not provide a band for the latter curve. 500 simulation runs are carried out and for each simulation, 500 bootstrap samples are generated.

From Table 1 we observe that for the asymptotic method coverage probabilities improve with increasing sample size. Results for the derivative curve ($j = 1$) are better than for the curve itself, although the bands are wider. For this particular setting, Xia's method, focusing on the curve itself, obtains a larger coverage probability than the asymptotic method, though still significantly smaller than the nominal coverage. Both bootstrap methods arrive at nominal coverage. It is observed that the size of the bands decreases with increasing sample size. Also the effect of slow convergence for the asymptotic methods illustrates itself clearly in showing smaller coverage errors for the larger sample sizes. For $j = 0$, bootstrap method 2 has slightly narrower bands than bootstrap method 1; this reverses for $j = 1$ where bootstrap 1 is preferred.

In simulation setting 2, data are generated according to a heteroscedastic model where $\sigma(x) = 0.1 + 0.06x$. Table 2 presents the results when a global variance estimator is used. Even though heteroscedasticity is not explicitly accounted for in estimating σ, a local Fisher information number is calculated. Both bootstrap methods perform very well, clearly outperforming Xia's and the asymptotic method. Note that for $j = 0$, the area of bootstrap 2's bands is on average smaller or very comparable to that following Xia's approach, while at the same time the bootstrap achieves the correct coverage probability.

To obtain the results presented in Table 3 we explicitly take the heteroscedasticity into account by locally estimating the variance function. This additional difficulty reflects in somewhat lower coverage probabilities, especially for bootstrap 2. Also in this setting bootstrap 1 gives slightly wider confidence bands for the curve ($j = 0$), while bootstrap 2 has somewhat wider bands for the derivative curve ($j = 1$). Note that Xia's bands are not available here.

Overall, both bootstrap methods perform very well in achieving nearly perfect simulated coverage probabilities, while not sacrificing much on the width of the bands.

Nominal			$j=0$		$j=1$	
Cov. (%)	n	Method	Cov.Prob.	Area	Cov.Prob.	Area
90	50	Xia	0.738	0.171	—	—
		Asympt.	0.566	0.120	0.744	1.164
		Bootst. 1	0.900	0.217	0.900	1.360
		Bootst. 2	0.900	0.209	0.900	1.412
	100	Xia	0.758	0.121	—	—
		Asympt.	0.670	0.104	0.844	1.002
		Bootst. 1	0.900	0.144	0.900	0.865
		Bootst. 2	0.900	0.128	0.900	0.877
	200	Xia	0.760	0.086	—	—
		Asympt.	0.762	0.083	0.864	0.948
		Bootst. 1	0.900	0.097	0.900	0.733
		Bootst. 2	0.870	0.087	0.900	0.955
95	50	Xia	0.840	0.191	—	—
		Asympt.	0.688	0.133	0.826	1.167
		Bootst. 1	0.950	0.266	0.950	1.593
		Bootst. 2	0.950	0.243	0.950	2.197
	100	Xia	0.842	0.136	—	—
		Asympt.	0.784	0.110	0.920	1.013
		Bootst. 1	0.950	0.173	0.950	0.981
		Bootst. 2	0.950	0.143	0.950	0.919
	200	Xia	0.850	0.096	—	—
		Asympt.	0.862	0.096	0.934	1.100
		Bootst. 1	0.950	0.110	0.950	0.755
		Bootst. 2	0.940	0.102	0.950	1.128

Table 1: *Simulated coverage probabilities and areas of nominal 90% and 95% confidence bands.* $j = 0$ *and* $j = 1$ *denote the results for, respectively, the curves* μ *and* μ'*, by local linear estimation following Xia (1998), asymptotic distribution theory, and the two bootstrap methods as discussed in Section 2. Data are generated from a homoscedastic model.*

Nominal			$j=0$		$j=1$	
Cov. (%)	n	Method	Cov.Prob.	Area	Cov.Prob.	Area
90	50	Xia	0.722	0.233	—	—
		Asympt.	0.586	0.154	0.770	1.118
		Bootst. 1	0.900	0.276	0.900	1.473
		Bootst. 2	0.900	0.230	0.900	1.595
	100	Xia	0.742	0.156	—	—
		Asympt.	0.730	0.120	0.860	1.888
		Bootst. 1	0.900	0.167	0.900	0.987
		Bootst. 2	0.900	0.160	0.900	1.279
	200	Xia	0.766	0.112	—	—
		Asympt.	0.812	0.098	0.900	1.126
		Bootst. 1	0.900	0.125	0.900	0.740
		Bootst. 2	0.900	0.110	0.900	1.069
95	50	Xia	0.794	0.262	—	—
		Asympt.	0.724	0.171	0.868	1.326
		Bootst. 1	0.950	0.338	0.950	1.649
		Bootst. 2	0.950	0.268	0.950	1.598
	100	Xia	0.838	0.175	—	—
		Asympt.	0.854	0.142	0.934	1.510
		Bootst. 1	0.950	0.198	0.950	1.143
		Bootst. 2	0.950	0.170	0.950	1.550
	200	Xia	0.870	0.125	—	—
		Asympt.	0.902	0.109	0.946	1.340
		Bootst. 1	0.950	0.137	0.950	0.851
		Bootst. 2	0.948	0.125	0.950	1.551

Table 2: *Simulated coverage probabilities and areas of nominal 90% and 95% confidence bands. $j = 0$ and $j = 1$ denote the results for, respectively, the curves μ and μ', by local linear estimation following Xia (1998), asymptotic distribution theory, and the two bootstrap methods as discussed in Section 2. Data are generated from a heteroscedastic model, global variance estimation.*

Nominal			$j = 0$		$j = 1$	
Cov. (%)	n	Method	Cov.Prob.	Area	Cov.Prob.	Area
90	50	Asympt.	0.588	0.149	0.770	1.267
		Bootst. 1	0.798	0.281	0.900	2.192
		Bootst. 2	0.832	0.246	0.842	2.519
	100	Asympt.	0.730	0.120	0.862	1.853
		Bootst. 1	0.872	0.201	0.900	1.431
		Bootst. 2	0.836	0.140	0.868	2.358
	200	Asympt.	0.814	0.099	0.898	0.903
		Bootst. 1	0.900	0.141	0.900	0.930
		Bootst. 2	0.846	0.103	0.880	1.021
95	50	Asympt.	0.718	0.167	0.870	1.202
		Bootst. 1	0.864	0.335	0.950	2.840
		Bootst. 2	0.946	0.430	0.936	2.530
	100	Asympt.	0.854	0.142	0.936	1.759
		Bootst. 1	0.936	0.196	0.950	1.477
		Bootst. 2	0.914	0.161	0.942	1.779
	200	Asympt.	0.904	0.111	0.948	1.660
		Bootst. 1	0.950	0.142	0.950	1.029
		Bootst. 2	0.922	0.114	0.936	1.321

Table 3: *Simulated coverage probabilities and areas of nominal 90% and 95% confidence bands. $j = 0$ and $j = 1$ denote the results for, respectively, the curves μ and μ', by local linear estimation using asymptotic distribution theory, and the two bootstrap methods as discussed in Section 2. Data are generated from a heteroscedastic model, local variance estimation.*

5 Regularity conditions and proofs

Conditions

(H)(H1a) The bandwidth sequence h_n tends to zero as $n \to \infty$, in such a way that $nh_n/\log n \to \infty$ and $h_n \geq (\log n/n)^{1-2/\lambda}$ for λ as in condition (R2).

(H1b) $(\log n)^3/(nh_n) \to 0$ and for $j = 0, \ldots, p$, $nh_n^{4[(p-j)/2]+2j+5} \log n \to 0$.

(H2) $n^{-1}h_n^{-(1+b)}(\log n)^{5+b} = O(1)$ for some $b \geq 1$.

(R0) The kernel K is a symmetric, continuously differentiable pdf on $[-1, 1]$ taking on the value zero at the boundaries.
The design density f_X is differentiable on $B = [b_1, b_2]$, the derivative is continuous, and $\inf_{x \in B} f_X(x) > 0$.
The function $\theta(x)$ has $2([p/2] + 1)$ continuous derivatives on B.

(R1) For every y, third partial derivatives of $f(y, \theta)$ with respect to θ exist and are continuous in θ. The Fisher information $I(\theta(x))$, possesses a continuous derivative and $\inf_{x \in B} I(\theta(x)) > 0$.

(R2) There exists a neighborhood $N(\theta(x))$ such that

$$\max_{k=1,2} \sup_{x \in B} \left\| \sup_{\theta \in N(\theta(x))} \left| \frac{\partial^k}{\partial \theta^k} \log f(Y; \theta) \right| \right\|_{\lambda, x} < \infty$$

for some $\lambda \in (2, \infty]$, where $\| \cdot \|_{\lambda, x}$ is the L^λ-norm, conditional on $X = x$. Further,

$$\sup_{x \in B} E_x \left[\sup_{\theta \in N(\theta(x))} \left| \frac{\partial^3}{\partial \theta^3} \log f(Y; \theta) \right| \right] < \infty.$$

(R3) For some $a < b/\{2(1 + b)\}$, with b as in (H2),

$$\max_{k=1,2} \sup_{x \in B} \sup_{\theta_1, \theta_2 \in N(\theta(x))} \int [F(y; \theta_1)(1 - F(y; \theta_1))]^a \left\{ \left| \frac{\partial}{\partial y} \left[\frac{\partial}{\partial \theta} \log f(y; \theta_2) \right]^k \right| \right.$$
$$\left. + \left| \frac{\partial^2}{\partial y \partial \theta} \left[\frac{\partial}{\partial \theta} \log f(y; \theta_2) \right]^k \right| \right\} dy < \infty$$

(R4) For some $\delta > 0$,

$$\max_{k=2,3} \left\| \sup_{x \in B} \sup_{\theta \in N(\theta(x))} \left| \frac{\partial^{(k)}}{\partial \theta^{(k)}} \log f(Y^*; \theta) \right| \right\|_{2+\delta, x}^* = O(1) \text{ a.s.},$$

where for any $\lambda > 0$, $\| \cdot \|_{\lambda, x}^*$ stands for the L_λ-norm of Y^*, conditional on $X^* = x$. Further,

$$\sup_{x \in B} \sup_{|\theta - \theta(x)| \leq h_n} \left| E_x^* \left[\frac{\partial}{\partial \theta} \log f(Y^*; \theta) \right] \right| = o(h_n^{1/2}(\log n)^{-1/2}) \text{ a.s.},$$

$$\sup_{x\in B}\sup_{\theta\in N(\theta(x))}\left|E_x^*\left[\frac{\partial}{\partial\theta}\log f(Y^*;\theta)\right]^2 - E_x\left[\frac{\partial}{\partial\theta}\log f(Y;\theta)\right]^2\right| = o((\log n)^{-1/2}) \text{ a.s.},$$

and

$$\max_{k=1,2}\sup_{x\in B}\sup_{\theta_1,\theta_2\in N(\theta(x))}\int[\hat{F}(y;\theta_1)(1-\hat{F}(y;\theta_1))]^a\left\{\left|\frac{\partial}{\partial y}\left[\frac{\partial}{\partial\theta}\log f(y;\theta_2)\right]^k\right|\right.$$
$$\left.+\left|\frac{\partial^2}{\partial y\partial\theta}\left[\frac{\partial}{\partial\theta}\log f(y;\theta_2)\right]^k\right|\right\}dy = O(1) \text{ a.s.},$$

where a is as in condition (R3).

Note that condition (R4) is formulated in terms of the distribution of the bootstrap data. It is straightforward to verify this condition for specific classes of densities. For instance, (R4) is satisfied when $f(y;\theta(x))$ equals the normal density and $\theta(x)$ is the conditional mean or variance, and also when all partial derivatives of order at most three of $\log f(y;\theta(x))$ with respect to y and θ are uniformly bounded in y and θ.

The second bandwidth condition in (H1b) reduces to $nh_n^{2p+3}\log n \to 0$ for $p-j$ odd and to $nh_n^{2p+5}\log n \to 0$ for $p-j$ even.

Also note that for the multiparameter case assumption (R0) holds for all θ_r, for the Fisher information matrix in (R1), $\inf_{x\in B}\boldsymbol{I}(\theta(x))$ is positive definite and that in (R2) – (R4) partial derivatives are with respect to the components θ_r, where $r=1,2$.

For the general estimating equations situation, we replace (R1) by

(R1') For every y, second partial derivatives of $\psi_r(y;\boldsymbol{\theta})$ with respect to θ_s $(r,s=1,2)$ exist and are continuous in θ. The matrices $\boldsymbol{I}$ and $\boldsymbol{K}$ possess a continuous derivative at $\theta(x)$ and $\inf_{x\in B}\boldsymbol{I}(\theta(x))$ is positive definite.

Conditions (R2') – (R4') are (R2) – (R4) where ψ_r replaces the score function $(\partial/\partial\theta)\log f$, for $r=1,2$.

Proof of Theorem 2.1. This proof goes along the same lines as the proof of Theorem 2.1 of Zhao (1994). The major difference is that because of the local polynomial estimation, we have to deal with a vector parameter $\theta(x)=(\theta_0(x),\ldots,\theta_p(x))^t$. Using the conditions of the theorem, and similar to Lemma 1 of Zhao (1994), we obtain that

$$\sup_{x\in B}\sup_{\theta\in N(\theta(x))}\sqrt{\frac{nh_n}{\log n}}\left|\frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{X_i-x}{h_n}\right)\frac{\partial}{\partial\theta}\log f(Y_i;\theta)\left(\frac{X_i-x}{h_n}\right)^j - \right. \tag{5.1}$$
$$\left.\frac{1}{nh_n}\sum_{i=1}^n E\left[K\left(\frac{X_i-x}{h_n}\right)\frac{\partial}{\partial\theta}\log f(Y_i;\theta)\left(\frac{X_i-x}{h_n}\right)^j\right]\right| = O(1) \quad \text{a.s.}$$

For a $(p+1)$ dimensional sequence ε_n, define $\theta_\varepsilon(x,u) = \sum_{j=0}^{p}(\theta_j(x)+\varepsilon_{nj})(u-x)^j$, and let

$$A_{nj,\varepsilon}(x) = n^{-1}\sum_{i=1}^{n} K_h(X_i - x)(X_i - x)^j \frac{\partial}{\partial\theta}\log f(Y_i;\theta_\varepsilon(x, X_i)).$$

Then it follows from (5.1) that, with $j = 0,\ldots,p$,

$$\sup_{x\in B}|A_{nj,\varepsilon}(x) - E[A_{nj,\varepsilon}(x)]| = \ell_{n,j} = O\{h_n^j(\log n/(nh_n))^{1/2}\} \quad \text{a.s.}$$

Further, let $\tilde{p} = 2([p/2]+1)$, $w_j(h_n) = h_n^{\tilde{p}+j}\nu_{\tilde{p}+j}(K)\sup_{x\in B}\theta^{(\tilde{p})}/\tilde{p}! = O(h_n^{\tilde{p}+j})$ and $\tilde{w}_j(h_n) = \sup_{|u-x|\le h_n}\sum_{\ell=0}^{p}(u-x)^\ell\nu_j(K) = O(1)$. By a Taylor series expansion of $\frac{\partial}{\partial\theta}\log f(Y_i;\theta_\varepsilon(x,X_i))$ around the true parameter value, and after taking expectations, we obtain by conditions (R1) and (R2), that there exists a constant C such that $E[A_{nj,\varepsilon}(x)] \le -\frac{1}{2}C\varepsilon_{nj}\tilde{w}_j(h_n)$, where we took the sequence $\varepsilon_{nj} = \max\{2w_j(h_n)/\tilde{w}_j(h_n), 4\ell_{n,j}/(C\tilde{w}_j(h_n))\}$. In a similar way, $E[A_{nj,-\varepsilon}(x)] \ge \frac{1}{2}C\varepsilon_{nj}\tilde{w}_j(h_n)$. The proof now proceeds along the same lines as in Zhao (1994), using a continuity argument in the $(p+1)$ dimensional parameter space. □

Proof of Corollary 2.1. Let $\boldsymbol{B}_n(x)$ be the matrix,

$$\boldsymbol{B}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)\frac{\partial^2}{\partial\theta^2}\log f(Y_i;\theta(x,X_i))(\boldsymbol{H}_n^{-1}\boldsymbol{X}_i)(\boldsymbol{H}_n^{-1}\boldsymbol{X}_i)^t$$

and further, define

$$\boldsymbol{C}_n(x) = \frac{1}{2}\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)\frac{\partial^3}{\partial\theta^3}\log f(Y_i;\eta(x,X_i))(\hat{\boldsymbol{\theta}}(x)-\boldsymbol{\theta}(x))^t\boldsymbol{X}_i^t\boldsymbol{X}_i(\hat{\boldsymbol{\theta}}(x)-\boldsymbol{\theta}(x))(\boldsymbol{H}_n^{-1}\boldsymbol{X}_i),$$

where $\eta(x,X_i)$ is in between $\theta(x,X_i)$ and $\hat{\theta}(x,X_i)$. By a Taylor series expansion it is readily obtained that

$$\boldsymbol{H}_n(\hat{\boldsymbol{\theta}}(x)-\boldsymbol{\theta}(x)) = -\boldsymbol{B}_n^{-1}(x)\{\boldsymbol{H}_n^{-1}\boldsymbol{A}_n(x)+\boldsymbol{C}_n(x)\} = \{\boldsymbol{J}(x)\}^{-1}\boldsymbol{H}_n^{-1}\boldsymbol{A}_n(x)+\boldsymbol{R}_n(x),$$

where

$$\begin{aligned}\boldsymbol{R}_n(x) \;=\;& -\boldsymbol{B}_n^{-1}(x)\boldsymbol{J}^{-1}(x)\{\boldsymbol{J}(x)+\boldsymbol{B}_n(x)\}\boldsymbol{H}_n^{-1}\boldsymbol{A}_n(x)\\ & +\{\boldsymbol{B}_n^{-1}(x)\boldsymbol{J}^{-1}(x)\boldsymbol{B}_n(x)-\boldsymbol{J}^{-1}(x)\}\boldsymbol{H}_n^{-1}\boldsymbol{A}_n(x)-\boldsymbol{B}_n^{-1}(x)\boldsymbol{C}_n(x).\end{aligned}$$

Via a Taylor expansion of $\frac{\partial}{\partial\theta}\log f(Y_i;\theta(x,X_i))$ about $\hat{\theta}(x,X_i)$, taking expectations and using the symmetry of the kernel, we obtain that, under the previous set of conditions,

$$\sup_{x\in B}|E\{A_{nj}(x)\}| = O(h_n^{2([p/2]+1)+j}),$$

where $[a]$ denotes the integer part of a. This, together with equation (5.1), implies that

$$\sup_{x\in B} |A_{nj}(x)| = O_P\{h_n^j(\log n/nh_n)^{1/2} + h_n^{2([p/2]+1)+j}\}.$$

Similar techniques yield that

$$\sup_{x\in B} |B_{njk}(x) + J_{jk}(x)| = O_P\{(\log n/nh_n)^{1/2} + h_n^{2([p/2]+1)}\},$$

and that for n sufficiently large, $\inf_{x\in B} \mathrm{Det}(\boldsymbol{B}_n(x)) > 0$. Using the result of Theorem 2.1, and condition (H1b) on the bandwidth sequence, it now follows that, for $k = 0, \ldots, p$,

$$\sup_{x\in B} |R_{nk}(x)| = O_P(\{(\log n/nh_n)^{1/2} + h_n^{2([p/2]+1)}\}^2) = o_P\{(nh_n \log n)^{-1/2}\}.$$

$\square$

For the proof of Lemma 2.1, we first need to show the result below. Note that although the techniques in this paper are (in nature) quite similar to the ones used in Johnston (1982) and Härdle (1989), the latter papers do not make use of the result below. As a consequence of this, the integration over y in $Y_{nj1}(x)$ has to be restricted to $[-a_n, a_n]$ in their proofs, where a_n tends to infinity at a certain rate (while we can work with the full range $(-\infty, +\infty)$). Since their technique has the disadvantage that it leads to a number of regularity conditions on $f(y;\theta)$ that depend on the sequence $\{a_n\}$, we prefer to use a different method.

In regard to the lemma below, we like to point out that Lemma 2.1 in Härdle (1989), which is similar to the result below but does not have the denominator in (5.2), is only valid if the variables (X, Y) in that theorem follow a uniform distribution on $[0, 1]^2$.

Lemma 5.1 *Let* $(U_1, V_1), \ldots, (U_n, V_n), \ldots$ *be independent random vectors uniformly distributed on* $[0, 1]^2$*, let* $0 < r < 1$ *and* $0 < a < (1-r)/2$*. A sequence of 4-sided tight-down Wiener processes* $V_n(u, v)$ *can then be constructed such that*

$$\sup_{0\le u,v\le 1} \left| \frac{Z_n^*(u,v) - V_n(u,v)}{[u(1-u)v(1-v)]^a} \right| = o(n^{-r/2}(\log n)^{2r}) \qquad a.s., \tag{5.2}$$

where $Z_n^*(u, v) = Z_n(u, v) - vZ_{n1}(u) - uZ_{n2}(v)$,

$$Z_n(u,v) = n^{1/2}[n^{-1}\sum_{i=1}^n I(U_i \le u, V_i \le v) - uv],$$

$$Z_{n1}(u) = n^{1/2}[n^{-1}\sum_{i=1}^n I(U_i \le u) - u],$$

and similarly for $Z_{n2}(v)$.

Proof. We will restrict attention to $(u, v) \in A = [0, 1/2]^2$. The proof for the three other quadrants of $[0, 1]^2$ is similar. Hence, it suffices to consider, where $s = a/(1-r)$,

$$\sup_A \left| \frac{Z_n^*(u,v) - V_n(u,v)}{u^a v^a} \right| \qquad (5.3)$$
$$\leq \sup_A |Z_n^*(u,v) - V_n(u,v)|^r \sup_A \left| \frac{Z_n^*(u,v) - V_n(u,v)}{u^s v^s} \right|^{1-r}.$$

Since $V_n(u, v) = B_n(u, v) - vB_n(u, 1) - uB_n(1, v)$, where $\{B_n\}$ is a sequence of Brownian bridges on $[0, 1]^2$, we can write

$$Z_n^*(u,v) - V_n(u,v)$$
$$= [Z_n(u,v) - B_n(u,v)] - v[Z_{n1}(u) - B_n(u,1)] - u[Z_{n2}(v) - B_n(1,v)]$$

and this is $O(n^{-1/2}(\log n)^2)$ a.s. uniformly in u and v by the Theorem in Tusnády (1977). It follows from Einmahl et al. (1988) that the process $Z_n^*(u, v)/(u^s v^s)$ $((u, v) \in A)$ converges weakly to $V_n(u, v)/(u^s v^s)$. This, together with the Skorohod-Dudley-Wichura theorem (see Shorack and Wellner (1986), p. 47) yields that the second factor on the right hand side of (5.3) is $o(1)$ a.s., from which the result follows. □

Proof of Lemma 2.1. Define

$$L_x(z, y) = \frac{\partial}{\partial\theta} \log f(y; \theta(x, z)),$$

and let $Z_n(x, y) = n^{1/2}(F_n(x, y) - F(x, y))$ be the empirical process of (X, Y). Then, for $j = 0, \ldots, p$,

$$g(x)^{1/2} Y_{nj}(x) = n^{-1/2} h_n^{1/2} \sum_{i=1}^n K_h(X_i - x) \left(\frac{X_i - x}{h_n}\right)^j L_x(X_i, Y_i)$$
$$= h_n^{1/2} \int\int K_h(z - x) \left(\frac{z - x}{h_n}\right)^j L_x(z, y)\, dZ_n(z, y) + (nh_n)^{1/2} h_n^{-j} EA_{nj}(x).$$

Since $\sup_x |EA_{nj}(x)| = O(h_n^{2([p/2]+1)+j})$ (see the proof of Corollary 2.1), it follows that under the given conditions on the bandwidth, the second term above is $o((\log n)^{-1/2})$. Using the Rosenblatt transformation $M(x, y) = (F_X(x), F(y; \theta(x)))$ and integration by parts, the first term above can be written as (where we use the notation $q_u = F_X^{-1}(u)$)

$$h_n^{1/2} \int\int K_h(q_u - x) \left(\frac{q_u - x}{h_n}\right)^j L_x(M^{-1}(u, v))\, dZ_n(M^{-1}(u, v)), \qquad (5.4)$$

where $M^{-1}(u,v) = (F_X^{-1}(u), F^{-1}(v;\theta(F_X^{-1}(u))))$. Straightforward calculations show that in the above integral, $Z_n(M^{-1}(u,v))$ can be replaced by $\tilde{Z}_n(u,v)$, the empirical process of $(F_X(X_i), F(Y_i;\theta(X_i)))$ $(i=1,\ldots,n)$, which is uniformly distributed on $[0,1]^2$. Hence, using similar notations as in the statement of Lemma 5.1, (5.4) can be written as

$$h_n^{1/2} \int\int K_h(q_u - x) \left(\frac{q_u - x}{h_n}\right)^j L_x(M^{-1}(u,v))\, d\tilde{Z}_n^*(u,v) \tag{5.5}$$
$$+h_n^{1/2} \int\int K_h(q_u - x) \left(\frac{q_u - x}{h_n}\right)^j L_x(M^{-1}(u,v))\, d[v\tilde{Z}_{n1}(u) + u\tilde{Z}_{n2}(v)].$$

Using integration by parts, the first term of (5.5) can be written as

$$h_n^{1/2} \int\int \tilde{Z}_n^*(u,v)\, d\left[K_h(q_u - x)\left(\frac{q_u - x}{h_n}\right)^j L_x(M^{-1}(u,v))\right]$$
$$-h_n^{1/2} \int \tilde{Z}_n^*(u,1)\, d\left[K_h(q_u - x)\left(\frac{q_u - x}{h_n}\right)^j L_x(M^{-1}(u,1))\right]$$
$$+h_n^{1/2} \int \tilde{Z}_n^*(u,0)\, d\left[K_h(q_u - x)\left(\frac{q_u - x}{h_n}\right)^j L_x(M^{-1}(u,0))\right].$$

In a similar way, $g(x)^{1/2} Y_{nj1}(x)$ can be decomposed into three terms. In what follows, we consider the difference between the first term of each of both decompositions. The derivations for the second and third terms are similar, but in fact easier since only one integral is involved. Since $(F_X(X), F(Y;\theta(X)))$ is uniformly distributed on $[0,1]^2$, it follows from Lemma 5.1 that

$$\sup_{0\le u,v\le 1} \left|\frac{\tilde{Z}_n^*(u,v) - V_n(u,v)}{[u(1-u)v(1-v)]^a}\right| = o(n^{-r/2}(\log n)^{2r})$$

a.s., where $r = 1/(1+b)$ ($b > 0$ as in condition (H2)) and $0 < a < (1-r)/2 = b/\{2(1+b)\}$. Hence,

$$h_n^{1/2} \left|\int\int [\tilde{Z}_n^*(u,v) - V_n(u,v)]\, d\left[K_h(q_u - x)\left(\frac{q_u - x}{h_n}\right)^j L_x(M^{-1}(u,v))\right]\right| \tag{5.6}$$
$$\le \int\int [v(1-v)]^a \left|d\left[K_h(q_u - x)\left(\frac{q_u - x}{h_n}\right)^j L_x(M^{-1}(u,v))\right]\right|$$
$$\times o(n^{-r/2} h_n^{1/2} (\log n)^{2r}).$$

The integral in the above expression can be written as

$$\int\int [F(y;\theta(z))(1 - F(y;\theta(z)))]^a \left|d\left[K_h(z - x)\left(\frac{z - x}{h_n}\right)^j L_x(z,y)\right]\right|$$

$$= h_n^{-1} \int \int [F(y;\theta(x+uh_n))(1-F(y;\theta(x+uh_n)))]^a \Big| [K'(u)u^j + jK(u)u^{j-1}]$$
$$\times \frac{\partial}{\partial y} L_x(x+uh_n, y) + h_n K(u)u^j \frac{\partial^2}{\partial z \partial y} L_x(z = x+uh_n, y) \Big| \, dy \, du$$

and this is $O(h_n^{-1})$ by condition (R3). It now follows that (5.6) is $o(n^{-r/2} h_n^{-1/2} (\log n)^{2r}) = o((\log n)^{-1/2})$ a.s. Consider now the second term of (5.5) :

$$h_n^{1/2} \int K_h(z-x) \left(\frac{z-x}{h_n}\right)^j \left[\int \frac{\partial}{\partial \theta} \log f(y;\theta(x,z)) \, dF(y;\theta(z))\right] d\tilde{Z}_{n1}(F_X(z)) \qquad (5.7)$$
$$+ h_n^{1/2} \int K_h(z-x) \left(\frac{z-x}{h_n}\right)^j f_X(z) \left[\int \frac{\partial}{\partial \theta} \log f(y;\theta(x,z)) d\tilde{Z}_{n2}(F(y;\theta(z)))\right] dz.$$

We start with the second term of (5.7). Consider

$$\left| \int \frac{\partial}{\partial \theta} \log f(y;\theta(x,z)) d\tilde{Z}_{n2}(F(y;\theta(z))) \right|$$
$$\le \left\{ \sup_{0 \le v \le 1} \frac{|B_{n2}(v)|}{(v(1-v))^a} + o(1) \right\} \int [F(y;\theta(z))(1-F(y;\theta(z)))]^a \left| \frac{\partial^2}{\partial y \partial \theta} \log f(y;\theta(x,z)) \right| dy,$$

where $\{B_{n2}\}$ is a sequence of Brownian bridges on $[0,1]$. Hence, by condition (R3), the second term of (5.7) is $O_P(h_n^{1/2})$. Using the notation $x_u = x + uh_n$, the first term is bounded by

$$h_n^{-1/2} \int K(u)u^j \left| \int \frac{\partial}{\partial \theta} \log f(y;\theta(x,x_u)) f(y;\theta(x_u)) \, dy \right| \left| d\tilde{Z}_{n1}(F_X(x_u)) \right|.$$

Since by definition of $\theta(x_u)$,

$$E_{x_u} \left[\frac{\partial}{\partial \theta} \log f(Y;\theta(x_u))\right] = 0,$$

for any x and u, the integral between absolute values equals

$$-uh_n \frac{\partial}{\partial z} \theta(z = \tilde{x}_u, x_u) E_{x_u} \left[\frac{\partial^2}{\partial \theta^2} \log f(Y;\theta(\tilde{x}_u, x_u))\right],$$

for some $\tilde{x}_u$ between x and x_u. Hence, by condition (R2), also this term is $O_P(h_n^{1/2})$. □

Proof of Lemma 2.2. Using integration by parts, we can write $g(x)^{1/2}[Y_{nj1}(x) - Y_{nj2}(x)]$ as the sum of three terms in a similar way as in the proof of Lemma 2.1. It suffices to consider the first term of this sum, as the two others are similar :

$$h_n^{1/2}\int\int V_n(M(z,y))\,d\left[K_h(z-x)\left(\frac{z-x}{h_n}\right)^j\left\{\frac{\partial}{\partial\theta}\log f(y;\theta(x,z))-\frac{\partial}{\partial\theta}\log f(y;\theta(z))\right\}\right]$$

$$=h_n^{-1/2}\int\int V_n(M(x_u,y))\,d\left[K(u)u^j\left\{\frac{\partial}{\partial\theta}\log f(y;\theta(x,x_u))-\frac{\partial}{\partial\theta}\log f(y;\theta(x_u))\right\}\right]$$

$$=h_n^{-1/2}\int\int V_n(M(x_u,y))\{K'(u)u^j+jK(u)u^{j-1}\}\left\{\frac{\partial^2}{\partial y\partial\theta}\log f(y;\theta(x,x_u))\right.$$

$$\left.-\frac{\partial^2}{\partial y\partial\theta}\log f(y;\theta(x_u))\right\}dy\,du$$

$$+h_n^{1/2}\int\int V_n(M(x_u,y))K(u)u^j\left\{\frac{\partial^3}{\partial y\partial\theta^2}\log f(y;\theta(x,x_u))\frac{\partial}{\partial z}\theta(x,z=x_u)\right.$$

$$\left.-\frac{\partial^3}{\partial y\partial\theta^2}\log f(y;\theta(x_u))\theta'(x_u)\right\}dy\,du,$$

where $x_u=x+uh_n$. From the mean value theorem it follows that the first term above is bounded by

$$Kh_n^{1/2}\sup_{0\le u,v\le 1}\frac{|V_n(u,v)|}{(v(1-v))^a}\int\int[F(y;\theta(x_u))(1-F(y;\theta(x_u)))]^a\left|\frac{\partial^3}{\partial y\partial\theta^2}\log f(y;\theta_{xu})\right|dy\,du$$

for some $K>0$ and some θ_{xu} between $\theta(x,x_u)$ and $\theta(x_u)$, and this is $O_P(h_n^{1/2})$ by condition (R3). In a similar way it follows that the second term above is $O_P(h_n^{1/2})$. $\square$

Proof of Lemma 2.3. The difference $Y_{nj2}(x)-Y_{nj3}(x)$ can be decomposed into two components, by writing $W_n(u,v)-V_n(u,v)=[W_n(u,v)-B_n(u,v)]-[V_n(u,v)-B_n(u,v)]$, where $\{B_n\}$ is a sequence of Brownian bridges on $[0,1]$ satisfying $B_n(u,v)=W_n(u,v)-uvW_n(1,1)$. Since $V_n(u,v)-B_n(u,v)=vB_n(u,1)+uB_n(1,v)$, the proof for the second component parallels completely the derivation for the second term of (5.5) in the proof of Lemma 2.1, and hence this term is $O_P(h_n^{1/2})$. For the first component, note that $W_n(u,v)-B_n(u,v)=uvW_n(1,1)$, and hence

$$\int\int K_h(z-x)\left(\frac{z-x}{h_n}\right)^j\frac{\partial}{\partial\theta}\log f(y;\theta(z))\,d[W_n(M(z,y))-B_n(M(z,y))]$$

$$=W_n(1,1)\int K_h(z-x)\left(\frac{z-x}{h_n}\right)^j\left[\int\frac{\partial}{\partial\theta}\log f(y;\theta(z))\,dF(y;\theta(z))\right]dF_X(z)=0$$

by definition of $\theta(z)$. $\square$

Proof of Lemma 2.4. The proof is similar to that of Lemma 3.7 in Härdle (1989). First note that for any $j = 0, \ldots, p$, $Y_{nj3}(x)$ and $Y_{nj4}(x)$ are zero mean Gaussian processes. It therefore suffices to show that they have the same covariance function. Since any functions h_1 and h_2, that are defined on an interval $[a, b]$ and for which $h_i(a) = h_i(b) = 0$ $(i = 1, 2)$, satisfy $\int_a^b \int_a^b (x_1 \wedge x_2)\, dh_1(x_1)\, dh_2(x_2) = \int_a^b h_1(x) h_2(x)\, dx$, straightforward, but lengthy, calculations show that

$$
\begin{aligned}
&\mathrm{Cov}(Y_{nj3}(x_1), Y_{nj3}(x_2)) \\
&\quad = h_n[g(x_1)g(x_2)]^{-1/2} \int \int \left[\frac{\partial}{\partial \theta} \log f(y; \theta(z, z))\right]^2 dy\, K_h(z - x_1) K_h(z - x_2) f(y; \theta(z))\, dz \\
&\quad = h_n[g(x_1)g(x_2)]^{-1/2} \int g(z)\, K_h(z - x_1) K_h(z - x_2)\, dz \\
&\quad = \mathrm{Cov}(Y_{nj4}(x_1), Y_{nj4}(x_2)).
\end{aligned}
$$

□

Proof of Lemma 2.5. The proof parallels that of Lemma 3.5 in Härdle (1989). We have

$$
\begin{aligned}
&Y_{nj4}(x) - Y_{nj5}(x) \\
&\quad = h_n^{1/2} \int \left[\left(\frac{g(z)}{g(x)}\right)^{1/2} - 1\right] K_h(z - x) \left(\frac{z - x}{h_n}\right)^j dW(z) \\
&\quad = h_n^{-1/2} \int \left[\left(\frac{g(x + uh_n)}{g(x)}\right)^{1/2} - 1\right] K(u) u^j\, dW(x + uh_n) \\
&\quad = -h_n^{-1/2} \int W(x + uh_n) \left[\left(\frac{g(x + uh_n)}{g(x)}\right)^{1/2} - 1\right] \{K'(u)u^j + jK(u)u^{j-1}\}\, du \\
&\qquad -\frac{1}{2} h_n^{1/2} \int W(x + uh_n) \left(\frac{g(x + uh_n)}{g(x)}\right)^{-1/2} \left(\frac{g'(x + uh_n)}{g(x)}\right) K(u) u^j\, du. \qquad (5.8)
\end{aligned}
$$

Using the conditions on $I(\theta(x))$ and $f_X(x)$ and the fact that $\sup_x |W(x)| = O_P(1)$, it easily follows that (5.8) is $O_P(h_n^{1/2})$. □

Proof of Theorem 2.2. For $j = 0, \ldots, p$, let

$$
\begin{aligned}
r_j(x) &= \mathrm{Cov}\Big(\sum_{k=0}^p (\boldsymbol{N}_p^{-1})_{j+1,k+1} Y_{nk6}(x), \sum_{k=0}^p (\boldsymbol{N}_p^{-1})_{j+1,k+1} Y_{nk6}(0)\Big) \\
&= \sum_{k=0}^p \sum_{\ell=0}^p (\boldsymbol{N}_p^{-1})_{j+1,k+1} (\boldsymbol{N}_p^{-1})_{j+1,\ell+1} \int K(x + u) K(u) (x + u)^k u^\ell du.
\end{aligned}
$$

A Taylor series expansion about zero, yields that $r_j(0) = \int K_{jp}^2$, and that by the assumptions on the kernel, $r_j'(0) = 0$, and $r_j''(0) = -C_j \int K_{jp}^2$. The result now follows using Corollary 2.1, Lemmas 2.1–2.5 and Corollary A.1 of Bickel and Rosenblatt (1973). $\square$

Proof of Theorem 2.3. We only give the proof for $Z_{nj}^{(2)*}(x)$, since the proof for $Z_{nj}^{(1)*}(x)$ is very similar. Since by Corollary 2.1,

$$Z_{nj}(x) = \sum_{k=0}^{p} g(x)^{1/2} Y_{nk}(x) (\boldsymbol{J}(x)^{-1})_{jk} \hat{V}_{jj}(\hat{\boldsymbol{\theta}}(x))^{-1/2} + o((\log n)^{-1/2}) \text{ a.s.}$$

and

$$Z_{nj}^{(2)*}(x) = -\sum_{k=0}^{p} g(x)^{1/2} Y_{nk}^*(x) (\boldsymbol{B}_n(x)^{-1})_{jk} \hat{V}_{jj}^*(\hat{\boldsymbol{\theta}}(x))^{-1/2},$$

where

$$Y_{nk}^*(x) = (nh_n)^{1/2} h_n^{-k} g(x)^{-1/2} A_{nk}^*(x),$$

it suffices, by Slutsky's theorem, to prove that

$$\sup_x |Z_{nj}^*(x)| - \sup_x |Z_{nj}(x)| = o_P^*((\log n)^{-1/2}) \text{ a.s.}$$

To accomplish this, we will show that for all $j, k = 0, \ldots, p$,

$$\sup_x |g(x)^{1/2}[Y_{nk}^*(x) - Y_{nk}(x)]| = o_P^*((\log n)^{-1/2}) \text{ a.s.}, \tag{5.9}$$

$$\sup_x |\hat{V}_{jj}^*(\hat{\boldsymbol{\theta}}(x))^{-1/2} - \hat{V}_{jj}(\hat{\boldsymbol{\theta}}(x))^{-1/2}| = o_P^*((\log n)^{-1/2}) \text{ a.s.}, \tag{5.10}$$

$$\sup_x |(\boldsymbol{B}_n(x)^{-1})_{jk} + (\boldsymbol{J}(x)^{-1})_{jk}| = o((\log n)^{-1/2}) \text{ a.s.} \tag{5.11}$$

From the proof of Corollary 2.1, together with bandwidth condition (H1b), it follows that (5.11) holds true. For showing (5.9), write

$$\begin{aligned} & g(x)^{1/2}[Y_{nk}^*(x) - Y_{nk}(x)] \\ & \quad = g(x)^{1/2}[Y_{nk}^*(x) - \tilde{Y}_{nk}^*(x)] + g(x)^{1/2}[\tilde{Y}_{nk}^*(x) - Y_{nk}(x)] \\ & \quad = T_{nk1}(x) + T_{nk2}(x), \end{aligned}$$

where $\tilde{Y}_{nk}^*(x) = (nh_n)^{1/2} h_n^{-k} g(x)^{-1/2} \tilde{A}_{nk}^*(x)$ and $\tilde{A}_{nk}^*(x)$ is obtained by replacing $\hat{\theta}(x, X_i^*)$ $(i = 1, \ldots, n)$ in the expression of $A_{nk}^*(x)$ by $\theta(x, X_i^*)$. We start with $T_{nk1}(x)$. Write

$$\frac{\partial}{\partial \theta} \log f(Y_i^*, \hat{\theta}(x, X_i^*)) - \frac{\partial}{\partial \theta} \log f(Y_i^*, \theta(x, X_i^*))$$

$$= \frac{\partial^2}{\partial\theta^2} \log f(Y_i^*, \theta(x, X_i^*))[\hat\theta(x, X_i^*) - \theta(x, X_i^*)]$$
$$+ \frac{1}{2}\frac{\partial^3}{\partial\theta^3} \log f(Y_i^*, \eta_i^*)[\hat\theta(x, X_i^*) - \theta(x, X_i^*)]^2,$$

where η_i^* is in between $\theta(x, X_i^*)$ and $\hat\theta(x, X_i^*)$. Hence, $T_{nk1}(x)$ can be decomposed into two terms, say $T_{nk11}(x)$ and $T_{nk12}(x)$. From Theorem 2.1 it follows that $T_{nk12}(x) = O_P^*\{\log n/(nh_n)^{1/2} + (nh_n^{8[p/2]+9})^{1/2}\} = o_P^*((\log n)^{-1/2})$ uniformly in x. In order to show that $\sup_x |T_{nk11}(x)| = o_P^*((\log n)^{-1/2})$, let

$$Z_{nijk}^*(x) = d_{njk} K_h(X_i^* - x) \frac{\partial^2}{\partial\theta^2} \log f(Y_i^*, \theta(x, X_i^*))(X_i^* - x)^{j+k},$$

where $d_{njk} = b_{nk} c_{nj}$ and

$$b_{nk} = n^{-1} h_n^{-k} (nh_n)^{1/2} \log n$$
$$c_{nj} = h_n^{-j} (\frac{\log n}{nh_n})^{1/2} + h_n^{2([(p-j)/2]+1)}.$$

Then,

$$T_{nk11}(x) = \sum_{j=0}^{p} \frac{\hat\theta_j(x) - \theta_j(x)}{c_{nj}} (\log n)^{-1} \sum_{i=1}^{n} \{Z_{nijk}^*(x) - E^*[Z_{nijk}^*(x)]\}.$$

In what follows we will show that

$$\sup_x |\sum_{i=1}^{n} \{Z_{nijk}^*(x) - E^*[Z_{nijk}^*(x)]\}| = O_{P^*}(1), \qquad (5.12)$$

for all j and k, from which it follows that $\sup_x |T_{nk11}(x)| = O_{P^*}((\log n)^{-1}) = o_{P^*}((\log n)^{-1/2})$. In order to show (5.12), we establish the weak convergence of the process $\sum_i [Z_{nijk}^*(x) - E^* Z_{nijk}^*(x)]$ $(x \in B)$ by making use of Theorem 2.11.9 in van der Vaart and Wellner (1996). We start with calculating the bracketing number $N_{[]}(\varepsilon, B, L_2^n)$, which is the minimal number of sets N_ε in a partition $B = \cup_{j=1}^{N_\varepsilon} B_{\varepsilon j}$ such that for every partitioning set $B_{\varepsilon j}$,

$$\sum_{i=1}^{n} E^* \sup_{x,x' \in B_{\varepsilon j}} |Z_{nijk}^*(x) - Z_{nijk}^*(x')|^2 \le \varepsilon^2. \qquad (5.13)$$

Partition the space B into $O(\varepsilon^{-2})$ equally spaced intervals $[x_\ell, x_{\ell+1}]$ of length $K\varepsilon^2$ for some $K > 0$. We need to consider two cases. If $\varepsilon^2 \le h_n$, then for $x_\ell \le x, x' \le x_{\ell+1}$, it follows from (R4) that the left hand side of (5.13) is bounded by $K' n h_n d_{njk}^2 h_n^{-4} \varepsilon^2 h_n \le \varepsilon^2$ for some $K' > 0$. If $\varepsilon^2 \ge h_n$, then similar arguments show that the bound is now given by $K'' n \varepsilon^2 d_{njk}^2 h_n^{-2} \le \varepsilon^2$.

This shows that the bracketing number is $O(\varepsilon^{-2})$ and hence the third displayed condition in the above mentioned theorem is satisfied. Next, we show the first displayed condition, which states that $nE^*\{\sup_x |Z^*_{nijk}(x)| I[\sup_x |Z^*_{nijk}(x)| > \eta]\} \to 0$ a.s., for all $\eta > 0$. This follows easily from condition (R4) together with the fact that for any distribution F for which $\int x^2\, dF(x) < \infty$, $\int_y^{+\infty} x\, dF(x) \le y^{-1}$ for y large enough. To complete the proof of the weak convergence of $\sum_{i=1}^n Z^*_{nijk}(x)$, we still need to show the convergence of the marginals. Fix $x \in B$. It is easily shown that Liapunov's ratio (for some $\delta > 0$)

$$\frac{\sum_{i=1}^n E^*|Z^*_{nijk}(x) - E^* Z^*_{nijk}(x)|^{2+\delta}}{(\sum_{i=1}^n \mathrm{Var}^* Z^*_{nijk}(x))^{(2+\delta)/2}}$$

is $O((nh_n)^{-\delta/2}) = o(1)$ provided assumption (R4) holds. This shows that (5.12) is satisfied. It remains to consider the term $T_{nk2}(x)$. Since Lemma 2.1 entails that $Y_{nk}(x)$ is asymptotically equivalent to $Y_{nk1}(x)$, it suffices to show that

$$\tilde{T}_{nk2}(x) = g(x)^{1/2}[\tilde{Y}^*_{nk}(x) - Y_{nk1}(x)]$$

is $o^*_P((\log n)^{-1/2})$ uniformly in x. The proof for this parallels that of Lemma 2.1. Let $\hat{F}_X(x) = \int_{-\infty}^x \hat{f}_X(t)\, dt$. Then, by using the Rosenblatt transformation $(\hat{F}_X(x), \hat{F}(y;\theta(x))$ we can decompose $\tilde{T}_{nk2}(x)$ in the same way as is done in Lemma 2.1, except that $F_X(x)$ respectively $F(y;\theta(x))$ are replaced by $\hat{F}_X(x)$ respectively $\hat{F}(y;\theta(x))$. It hence follows from condition (R4) that $\sup_x |\tilde{T}_{nk2}(x)| = o^*_P((\log n)^{-1/2})$.

It remains to prove (5.10). It suffices to show that for all $j, k = 0, \ldots, p$,

$$\sup_x |K^*_{njk}(x) - K_{njk}(x)| = o^*_P((\log n)^{-1/2}) \text{ a.s.}$$

Write

$$K^*_{njk}(x) - K_{njk}(x) = [K^*_{njk}(x) - \tilde{K}^*_{njk}(x)] + [\tilde{K}^*_{njk}(x) - \tilde{K}_{njk}(x)] + [\tilde{K}_{njk}(x) - K_{njk}(x)],$$

where $\tilde{K}^*_{njk}(x)$ respectively $\tilde{K}_{njk}(x)$ is obtained by replacing $\hat{\theta}$ by θ in $K^*_{njk}(x)$ respectively $K_{njk}(x)$. A similar derivation as for the term $T_{nk12}(x)$ above shows that the first and third term are $o^*_P((nh_n \log n)^{-1/2})$ a.s. uniformly in x. Hence, it suffices to consider

$$\begin{aligned}
&\tilde{K}^*_{njk}(x) - \tilde{K}_{njk}(x) \\
&\quad = [\tilde{K}^*_{njk}(x) - E^*\tilde{K}^*_{njk}(x) - \tilde{K}_{njk}(x) + E\tilde{K}_{njk}(x)] + [E^*\tilde{K}^*_{njk}(x) - E\tilde{K}_{njk}(x)] \\
&\quad = A(x) + B(x).
\end{aligned}$$

Using a similar derivation as in the proof of Lemma 2.1, it is easy to see that

$$\sup_x |A(x)| = O_P^*((nh_n)^{-1/2}(h_n^{-1/2} + (\log n)^{-1/2})) = o_P^*((\log n)^{-1/2}) \text{ a.s.}$$

Finally, using the notation $x_u = x + uh_n$,

$$\begin{aligned} B(x) &= \int\int K^2(u)u^{j+k}\left[\frac{\partial}{\partial\theta}\log f(t;\theta(x,x_u))\right]^2 (\hat{f}(x_u,t) - f(x_u,t))\,du\,dt \\ &= \int K^2(u)u^{j+k}\int\left[\frac{\partial}{\partial\theta}\log f(t;\theta(x,x_u))\right]^2 \hat{f}(t;\theta(x_u))\,dt(\hat{f}_X(x_u) - f_X(x_u))\,du \\ &\quad + \int K^2(u)u^{j+k}\int\left[\frac{\partial}{\partial\theta}\log f(t;\theta(x,x_u))\right]^2 (\hat{f}(t;\theta(x_u)) - f(t;\theta(x_u)))\,dt\,f_X(x_u)\,du, \end{aligned}$$

where $\hat{f}(y;\theta(x)) = \hat{f}(x,y)/\hat{f}_X(x)$. From condition (R4) and the rate of convergence of $\hat{f}_X(x)$ it now follows that the first term above is $O((nh_n)^{-1/2}(\log n)^{1/2})$ a.s., while the second is $o((\log n)^{-1/2})$ a.s. This finishes the proof. □

Proof of Theorem 3.1. For $r = 1, 2$ and $j = 0, \ldots, p$, define

$$Y_{nrj}(x) \quad = \quad (nh_n)^{1/2}h_n^{-j}\left(\boldsymbol{I}^{-1}(\boldsymbol{\theta}(x))\boldsymbol{K}(\boldsymbol{\theta}(x))\boldsymbol{I}^{-1}(\boldsymbol{\theta}(x))\right)_{rr}^{1/2} f_X^{-1/2}(x)A_{nrj}(x),$$

where A_{nrj} is defined similarly as in (3.1), now replacing the rth score component by $\psi_r(Y_i;\boldsymbol{\theta}(x,X_i))$. It is readily obtained that $\mathrm{Var}(\boldsymbol{H}_n^{-1}\boldsymbol{A}_n) = (nh_n)^{-1}f_X(x)\boldsymbol{K}(\boldsymbol{\theta}(x)) \otimes \boldsymbol{T}_p + o(nh_n)^{-1}$. The proof now continues along the same lines as the proof of Theorem 2.2. □

References

Aerts, M. and Claeskens, G. (1997). Local polynomial estimators in multiparameter likelihood models. *J. Amer. Statist. Assoc.*, **92**, 1536–1545.

Aerts, M. and Claeskens, G. (2001). Bootstrap tests for misspecified models, with application to clustered binary data. *Comput. Statist. Data Anal.*, **36**, 383-401.

Arnold, B.C., Castillo, E. and Sarabia, J.-M. (1992). *Conditionally Specified Distributions.* Lecture Notes in Statistics 73, Springer-Verlag, New York.

Arnold, B.C. and Strauss, D. (1991). Pseudolikelihood estimation: some examples. *Sankhyā B*, **53**, 233–243.

Beirlant, J. and Goegebeur, Y. (2000). Local polynomial maximum likelihood estimation for Pareto-type distributions. Unpublished manuscript, UCS, Katholieke Universiteit Leuven.

Bickel, P.J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.*, **1**, 1071–1095.

Carroll, R.J., Ruppert, D. and Welsh, A.H. (1998). Local estimating equations. *J. Amer. Statist. Assoc.*, **93**, 214–227.

Claeskens, G. and Aerts, M. (2000). Bootstrapping local polynomial estimators in likelihood-based models. *J. Statist. Plann. Inference*, **86**, 63–80.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *J. Amer. Statist. Assoc.*, **74**, 829–836.

Davison, A.C. and Ramesh, N.I. (2000). Local likelihood smoothing of sample extremes. *J. R. Statist. Soc. B*, **62**, 191–208.

Einmahl, J.H.J., Ruymgaart, F.H. and Wellner, J.A. (1988). A characterization of weak convergence of weighted multivariate empirical processes. *Acta Sci. Math.*, **52**, 191–205.

Eubank, R.L. and Speckman, P.L. (1993). Confidence bands in nonparametric regression. *J. Amer. Statist. Assoc.*, **88**, 1287–1301.

Fan, J. (1992). Design adaptive nonparametric regression. *J. Amer. Statist. Assoc.*, **87**, 998–1545.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.*, **21**, 196–216.

Fan, J. and Chen, J. (1999). One-step local quasi-likelihood estimation. *J. R. Statist. Soc. B*, **61**, 927–943.

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

Fan, J., Heckman, N.E. and Wand, M.P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.*, **90**, 141–150.

Fan, J. and Zhang, W. (2000). Simultaneous confidence bands and hypothesis testing in varying coefficient models. *Scand. J. Statist.*, **27**, 715–731.

Hall, P. (1979). On the rate of convergence of normal extremes. *J. Appl. Prob.*, **16**, 433-439.

Hall, P. (1991a). Edgeworth expansions for nonparametric density estimators, with applications. *Statistics*, **22**, 215–232.

Hall, P. (1991b). On convergence rates of suprema. *Probab. Th. Rel. Fields*, **89**, 447–455.

Hall, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *Ann. Statist.*, **20**, 675–694.

Hall, P. and Titterington, D.M. (1988). On confidence bands in nonparametric density estimation and regression. *J. Multiv. Anal.*, **27**, 228–254.

Härdle, W. (1989). Asymptotic maximal deviation of M-smoothers. *J. Multiv. Anal.*, **29**, 163–179.

Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the 5th Berkeley Symposium*, **1**, 221–233.

Johnston, G.J. (1982). Probabilities of maximal deviations for nonparametric regression function estimates. *J. Multiv. Anal.*, **12**, 402–414.

Knafl, G., Sacks, J. and Ylvisaker, D. (1985). Confidence bands for regression functions. *J. Amer. Statist. Assoc.*, **80**, 683–691.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models.* Second Edition, Chapman & Hall, London.

Neumann, N.H. and Polzehl, J. (1998). Simultaneous bootstrap confidence bands in nonparametric regression. *Nonpar. Statist.*, **9**, 307–333.

Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Statist.*, **23**, 470–472.

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

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

Shorack, G.R. and Wellner, J.A. (1986). *Empirical processes with applications to statistics.* Wiley, New York.

Silverman, B.W. and Young, G.A. (1987). The bootstrap : to smooth or not to smooth? *Biometrika*, **74**, 469-479.

Staniswalis, J.G. (1989). The kernel estimate of a regression function in likelihood-based models. *J. Amer. Statist. Assoc.*, **84**, 276–283.

Sun, J. and Loader, C.R. (1994). Simultaneous confidence bands for linear regression and smoothing. *Ann. Statist.*, **22**, 1328–1345.

Tusnády, G. (1977). A remark on the approximation of the sample df in the multidimensional case. *Period. Mathem. Hungar.*, **8**, 53–55.

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak convergence and empirical processes.* Springer-Verlag, New York.

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439–447.

Xia, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *J. R. Statist. Soc. B*, **60**, 797–811.

Zhao, P.-L. (1994). Asymptotics of kernel estimators based on local maximum likelihood. *Nonpar. Statist.*, **4**, 79–90.