Measuring the Discrepancy of a Parametric Model via Local Polynomial Smoothing

Anouar El Ghouch[1] , **Marc G. Genton**[2] and **Taoufik Bouezmarni**[3]

[1] Institut de Statistique, Université Catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium. E-mail: Anouar.Elghouch@uclouvain.be

[2] Department of Statistics, Texas A&M University, College Station, TX 77843-3143, U.S.A. E-mail: genton@stat.tamu.edu

[3] Département de mathématiques, Université de Sherbrooke, Québec, Canada, E-mail: taoufik.bouezmarni@usherbrooke.ca.

September 27, 2010

Abstract

In the context of multivariate mean regression we propose a new estimator of the minimum L^2-distance between the true but unknown regression curve and a given parametric family. The method is based on local polynomial averaging of residuals with a polynomial degree that increases with the dimension of the covariate. Under some weak assumptions we give a Bahadur-type representation of the estimated distance from which root-n consistency and asymptotic normality are derived for strongly mixing variables. We then show how to use the proposed method to: (i) measure the explanatory power of a given set of covariates or a given parametric model; and (ii) test the goodness-of-fit hypothesis. We conclude with a simulation study that aims at checking the finite sample properties of these techniques.

KEY WORDS: Explanatory power; Goodness-of-fit; Model misspecification; Multivariate local polynomial smoothing; Strong mixing sequences; Hypothesis test.

Short title: Discrepancy of a Parametric Model

1 Introduction

In the context of regression with high dimensional predictors, it is difficult to get an efficient nonparametric estimator for the true regression function because of the sparsity of the data, the so-called curse of dimensionality. For that reason and for the purpose of interpretability, simple parametric models with few covariates are usually preferred to a purely nonparametric fit or to complex and possibly redundant parametric models. Therefore, the selection of an appropriate set of covariates to be used and the selection of an adequate parametric function to fit and make inference about the data are two of the most important and challenging problems in real data analysis. Typically, attention is focused on checking the adequacy of a particular parametric model without inquiring about the overall quality of the regressors. However, if the covariates at hand only capture a small amount of the variability of the output variable then it is obviously difficult to justify the choice of any particular parametric model. The problem becomes even more difficult with a small sample size or with a complex data structure such as in the presence of dependency among the observations. It turns out that one should consider both aspects in order to make the best decision by first selecting a subset of regressors without any parametric restrictions and then choosing an appropriate model taking into account the explanatory power of the selected covariates. Although covariate selection and model selection are of different nature, the main question in both problems is whether the variance explained by the function/covariates is relevant when compared to the variance due to errors.

During the last decades a very large amount of research related to this topic has been proposed with a variety of procedures, justifications and assumptions. The traditional literature includes the use of model selection criteria such as Akaike and Bayesian information criteria and the use of test statistics such as Wald and likelihood ratio tests. In the first case the

selection is done by choosing the model with the smallest criterion (error) among competing models/regressors while in the second case the selection is based on a test statistic that measures the departure from the null hypothesis in the direction of an alternative. As mentioned by Taylor (2009), "the ordering of statistical tests and the choice of the significance level can influence the outcome, as can the choice of the model under the null hypothesis". For an excellent review and a detailed discussion of model/covariate selection procedures and tests we refer to Lavergne (1998). To be consistent, the classical approaches impose severe restrictions on the true underling parametric structure and depend heavily on some strong assumptions about data such as normality of residuals, homoscedasticity, or the fact that the correct model belongs to the set of candidate models. Modern literature avoids these drawbacks by using nonparametric techniques that allow for great flexibility. Methods such as kernel smoothing and splines have become widely used to justify covariate and model selection. The literature related to this subject is vast and includes the work of Härdle and Mammen (1993), Hong and White (1995), Zheng (1996), Li and Wang (1998) and Jun and Pinkse (2009) for consistently testing a parametric regression functional form and Fan and Li (1996), Lavergne and Vuong (1996) and Gu et al. (2007) for testing relevant regressors.

In this context, we propose a new method based on a minimum L^2-distance between the parametric family and the unknown true regression function. This basic idea is behind many goodness-of-fit tests proposed in the literature. Commonly, these classical methods focus on the behavior of a test statistic under the null hypothesis that the given model is correct. Our approach differs from existing methods in many aspects. Rather than a testing problem, our purpose is to construct an estimator of the distance between the parametric model and the target function with some good statistical properties regardless of whether the given parametric model is correct or not. In that sense, the proposed estimator can be seen as a selection criterion and, so, be used to discriminate between several models. Under some

weak assumptions, our estimator is shown to be consistent and asymptotically unbiased. We also prove its asymptotic normality with the optimal root-n convergence rate. These results are stated under a random design and we allow for weakly dependent data which means that our method can be applied also in time series or spatial frameworks. Unlike many existing methods that treat only some particular parametric functions such as linear or polynomial functions, our approach can be applied to check the quality of any smooth parametric model without further restriction on its form and without the need of any bias correction or bootstrap procedure. The proposed distance estimator is then used to estimate an "inadequacy index" which serves as a kind of an inverse coefficient of determination (R^2): it takes values in $[0, 1]$ and when its value is close to 0 the parametric fit becomes better. Based on this result we are able to test the adequacy of a given parametric model using the asymptotic normality of that estimator. The main difficulty here is degeneracy of the test statistic under correct specification. To bypass this problem without sacrificing the power and the rate of convergence we adopt the concept of neighborhood hypotheses; see Dette and Munk (2003) for a very nice discussion. This method allows for the validation of a given parametric model which cannot be done with the classical goodness-of-fit tests. Finally we also discuss the problem of variable selection and show how the methodologies described above can be directly adapted to that.

The rest of the paper is organized as follows. In Section 2 we introduce the L^2-distance between the mean regression function and a parametric model. The nonparametric estimation procedure is also described in this section. Section 3 is devoted to the asymptotic properties of the proposed estimator. In Section 4 we define an inadequacy index and show how it can be used to validate a given model via neighborhood hypotheses testing. In that section, we also discuss the problem of covariate selection. The performance of the proposed method is examined in Section 5 via a Monte Carlo simulation study. The proofs of the asymptotic results are collected in the Appendix.

2 Estimation Procedure

2.1 Problem setting

Let (X, Y) be a random vector in $\mathbb{R}^d \times \mathbb{R}$. For a given $x \in \mathbb{R}^d$, we denote by $m(x)$ the conditional mean of Y given that $X = x$. Define $\epsilon = Y - m(X)$ and denote by f the marginal density of X. Let the function $m(\theta, x)$ be a parametric model. This function is known up to the finite parameter θ that belongs to the parameter space Θ which is assumed to be a compact subset of $\mathbb{R}^q$. We consider the function $m(\theta, x)$ as a member of the family of parametric functions $\mathcal{M} = \{m(\theta, x),\ \theta \in \Theta\}$.

The data are given by (X_i, Y_i), $i = 1, \ldots, n$ and have the same distribution as (X, Y). As dependent observations are considered in this paper, we introduce here the mixing coefficient. Let $\mathcal{F}_I^L$ $(-\infty \leq I, L \leq \infty)$ denote the σ-field generated by the family $\{(X_t, Y_t),\ I \leq t \leq L\}$. The stochastic processes $\{(X_t, Y_t)\}$ is said to be strongly mixing if the α-mixing coefficient $\alpha(t) = \sup_{A \in \mathcal{F}_{-\infty}^0, B \in \mathcal{F}_t^\infty} |P(A \cap B) - P(A)P(B)|$ converges to 0 as $t \to \infty$. This dependency structure includes numerous random sequences. Among them are the independent and m-dependent variables and, under some weak conditions, the classical linear and nonlinear ARMA and (G)ARCH time series; see, for example, Fan and Yao (2003) and Carrasco and Chen (2002) for further details. As we will see later, the dependency among observations does not have any impact on the asymptotic results, provided that the degree of the dependence, as measured by the mixing coefficient $\alpha(t)$, is weak enough such that assumption (A6) given below is satisfied.

2.2 The parameter of interest

As mentioned in the introduction, our main objective here is to assess the adequacy of the function $m(\theta, x)$ as an approximation of the true but unknown mean regression function

$m(x)$. We measure the distance between the regression function m and $m(\theta, x)$ by $T(\theta) = \mathbb{E}[(m(X) - m(\theta, X))^2]$. This is nothing but the L^2-distance, with respect to f, that separates $m(\theta, X)$ from $m(X)$. Because in many situations one is only interested in measuring the quality of $m(\theta, x)$ within a subset of the support of the design density f, it is better to work with the following weighted version

$$T_\varphi(\theta) = \mathbb{E}[\Delta^2(\theta, X)\varphi^2(X)],$$

where $\Delta(\theta, x) = m(x) - m(\theta, x)$ and $\varphi(x)$ is a known weight function. In this way only the set $\{x : \varphi(x) \neq 0\}$ matters. In practice, the weighted version may also be used to avoid highly uncertain estimation in regions with sparse or noisy data. For example, by choosing $\varphi = I_{\{f>\delta\}}$, with I being the standard indicator function and $\{f > \delta\} = \{x : f(x) > \delta\}$, for some $\delta > 0$, one may avoid troubles that arise when the density f approaches 0, which is especially problematic in the context of nonparametric inference. From now on we will only consider the weighted version T_φ of T, so all upcoming formulas will depend on φ, but for notational convenience we will suppress φ in all our notations. Let $\hat{\theta}$ denote an estimator of θ. The parametric estimation procedure will be discussed latter. For now, the only assumption that we need is that $\hat{\theta}$ is pseudo-consistent. By this, we mean that there exists a unique parameter θ^* in the interior of Θ such that $\hat{\theta}$ converges in probability to θ^*, i.e. $\hat{\theta} = \theta^* + o_p(1)$. Our parameter of interest is $T(\theta^*) = T_\varphi(\theta^*)$, i.e., the L^2-distance between $m(\cdot)$ and $m(\theta^*, \cdot)$.

2.3 The local polynomial smoother

We now explain the kernel smoothing procedure that is used in the estimation step. Let K denote a nonnegative kernel function defined on $\mathbb{R}^d$, $0 < h_n \equiv h \to 0$ be a bandwidth parameter and $K_h(x) = h^{-d}K(x/h)$. For any $\theta \in \Theta$, denote $Y(\theta) = Y - m(\theta, X)$. By definition, $\Delta(\theta, x) = \mathbb{E}[Y(\theta)|X = x]$. If a parameter θ is available and if we consider $(X_i, Y_i(\theta))$,

$i = 1, \ldots, n$, as the observed sample and $\Delta(\theta, x)$ as the objective function, then we could directly apply classical smoothing techniques to construct a valid nonparametric estimator of $\Delta(\theta, x)$. In fact, using the local averaging principle, we propose to estimate $\Delta(\theta, x)$ by

$$\hat{\Delta}(\theta, x) = \sum_{j=1}^{n} w_j(x) Y_j(\theta), \tag{1}$$

where $w_j(x)$, $j = 1, \ldots, n$, are local weight functions depending on x, on $\{X_1, \ldots, X_n\}$, on the bandwidth parameter h and on the kernel function K. The form of (1) is shared by many nonparametric estimators of a regression function; see, for example, Fan and Gijbels (1996) for more details. In particular, the multivariate local polynomial estimator of order p, $p \in \mathbb{N}$, which is based on a multivariate local polynomial approximation, i.e., a Taylor expansion, of the target function $\Delta(\theta, x)$, can be expressed as (1). In this case the weight functions, $w_j(x)$, take different forms depending on the dimension d and the value of p. For the local constant regressor, i.e., $p = 0$, $w_j(x) = K_h(X_j - x) / \sum_{i=1}^{n} K_h(X_j - x)$. For the univariate local linear estimator, i.e., $p = 1$, and $d = 1$, $w_j(x) = n^{-1} K_h(X_j - x)[s_{n,2}(x) - \frac{X_j - x}{h} s_{n,1}(x)] / [s_{n,0}(x) s_{n,2}(x) - s_{n,1}^2(x)]$, where $s_{n,k}(x) = n^{-1} \sum_{j=1}^{n} [(X_j - x)/h]^k K_h(X_j - x)$. The general expression of the local polynomial multivariate weight function $w_j(x)$, for any $p \geq 0$ and $d \geq 1$, can be found in the Appendix, see (9). The estimator given by (1) is only available when θ is known, which, of course, is not the case here. So, we use the pseudo-consistent estimator $\hat{\theta}$ to get the following feasible estimator: $\hat{\Delta}(\hat{\theta}, x) = \sum_{j=1}^{n} w_j(x) Y_j(\hat{\theta}) = \hat{m}(x) - \hat{m}(\hat{\theta}, x)$, where $\hat{m}(x) = \sum_{j=1}^{n} w_j(x) Y_j$ is the standard (nonparametric) local polynomial estimator of the mean regression function $m(x)$ and $\hat{m}(\hat{\theta}, x) = \sum_{j=1}^{n} w_j(x) m(\hat{\theta}, X_j)$ is a smooth version of the parametric estimator $m(\hat{\theta}, x)$. It is known that smoothing the parametric estimator makes it asymptotically biased exactly as the standard nonparametric fit $\hat{m}(x)$. Such a procedure is largely used in the theory of goodness-of-fit test to improve power.

2.4 Our estimator of $T(\theta^*)$

Now that we have a valid estimator of $\Delta(\theta^*, x)$, and given the fact that $T(\theta) = \mathbb{E}[\Delta^2(\theta, X)\varphi^2(X)]$ we may consider estimating $T(\theta^*)$ using the obvious statistic

$$T_n^0(\hat{\theta}) = n^{-1} \sum_{i=1}^{n} \hat{\Delta}^2(\hat{\theta}, X_i)\varphi^2(X_i). \tag{2}$$

Up to the factor $h^{d/2}$, this is a discrete (Riemann sum) version of the test statistic that was proposed by Härdle and Mammen (1993) in the context of goodness-of-fit test. They used the local constant weight function $w_j(x)$, as defined above, and under some regularity assumptions, they showed the asymptotic normality of their test statistic when $m \in \mathcal{M}$. In the present work, the primary objective is not about testing but about constructing an estimator of $T(\theta^*)$ with some "good" properties. To this end, we consider here another estimator of $T(\theta^*)$ given by $T_n(\hat{\theta})$, with

$$T_n(\theta) = n^{-1} \sum_{i=1}^{n} \left(2Y_i(\theta)\hat{\Delta}(\theta, X_i) - \hat{\Delta}^2(\theta, X_i) \right) \varphi^2(X_i). \tag{3}$$

It is straightforward to show that this estimator is simply the empirical version of the following expression:

$$T(\theta) = \mathbb{E}\left[(2Y(\theta) - \Delta(\theta, X))\, \Delta(\theta, X)\varphi^2(X) \right].$$

Later, see Remark 2 and Section 5, the advantages of T_n over T_n^0 will become clear. Another way to motivate the choice of this estimation procedure is via the influence function approach. In fact, it can bee shown that $T_n(\theta)$ is the one-step estimator of $T(\theta)$ based on its influence function. More details can be found in Doksum and Samarov (1995). This approach is widely used in parametric and semiparametric theory to construct asymptotically linear estimators with high efficiency; see Bickel et al. (1993).

3 Main Results

Before starting with the study of the asymptotic properties of the proposed estimator we need first to introduce some notations and give a set of sufficient regularity conditions needed for the results to hold. For a d-tuple $k = (k_1, \ldots, k_d) \in \mathbb{N}^d$ and a d-vector $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we write

$$x^k = x_1^{k_1} \times \cdots \times x_d^{k_d}, \ \ |k| = \sum_{l=1}^{d} k_l, \ \text{ and } \ (D^k m)(x) = \frac{\partial^k m(x)}{\partial x_1^{k_1} \cdots x_d^{k_d}}.$$

ASSUMPTIONS (A)

(A1) $x \to m(\theta, x)$ is a continuous function on $S \subset \mathbb{R}^d$ for each θ in Θ. $\theta \to m(\theta, x)$ is twice differentiable on Θ for each x in S. The functions $\dot{m} := \dfrac{\partial m}{\partial \theta}$ and $\ddot{m} := \dfrac{\partial^2 m}{\partial \theta \theta^T}$ are continuous on $\Theta \times S$.

(A2) φ has a compact support $D \subset int(S)$, where $int(S)$ is the interior of S.

(A3) The marginal density f of X is bounded, uniformly continuous, and for all $x \in D$ $f(x) > L$ for some $L > 0$. For every $l \geq 1$, the joint density of (X_1, X_{l+1}) is bounded.

(A4) The conditional density of X given Y exists and is bounded. For every $l \geq 1$, the conditional density of (X_1, X_{l+1}) given (Y_1, Y_{l+1}) exists and is bounded.

(A5) For every k with $|k| = p + 1$, $(D^k m)$ is a bounded Lipschitz function.

(A6) $\mathbb{E}|Y|^\delta < \infty$ for some $\delta > 2$, $h_n \sim (n^{-1} \ln n)^a$ for some $0 < a < d^{-1}(1 - 2/\delta)$ and $\alpha(t) = O(t^{-\bar{a}})$, with $\bar{a} > \max\left(\frac{2\nu}{\nu-2}, \frac{\delta(7+2d)-4}{\delta(1-ad)-2}\right)$ fore some $\nu \in (2, \delta]$.

(A7) The kernel K is a bounded nonnegative function with compact support, say $[-1, 1]^{\otimes d}$ and for every k with $0 \leq |k| \leq 2p$ the function $u \to u^k K(u)$ is Lipschitz .

Some comments on our assumptions are worth noting. Assumption (A1) is mainly needed to apply the mean value theorem. The compactness stipulation in (A2) is used to derive asymptotic uniform bounds. Assumptions (A3)-(A7) are largely used in the theory of kernel regression with dependent data. Those assumptions can be found, for example, in Masry (1996). The stipulations about the bandwidth and the mixing coefficient in (A6) are just a simple, i.e., stronger, version of the necessary assumptions given by Conditions (7d), (4.5) and (4.7) in Masry (1996).

Our first result is formulated in the following Lemma.

LEMMA **1** *Under assumptions (A), if* $\mathbb{E}[\varphi^2(X)] < \infty$ *and* $\mathbb{E}(|\epsilon|\varphi^2(X)) < \infty$ *then*

$$T_n(\hat{\theta}) = T_n(\theta^*) - 2B^T(\hat{\theta} - \theta^*) + o_p(||\hat{\theta} - \theta^*||),$$

where $B = \mathbb{E}\left[\dot{m}(\theta^*, X)Y(\theta^*)\varphi^2(X)\right]$ *and* $\epsilon = Y - m(X)$.

This Lemma states that, asymptotically, the only impact of using the estimator $\hat{\theta}$ instead of θ^* is to shift T_n by the term $2B^T(\hat{\theta} - \theta^*)$. This quantity vanishes whenever $m \in \mathcal{M}$ since in that case $B = 0$. Otherwise B can be easily estimated by its empirical version $\hat{B} = n^{-1}\sum_{i=1}^n \dot{m}(\hat{\theta}, X)Y(\hat{\theta})\varphi^2(X)$.

The next Lemma gives a Bahadur-type representation of the estimator $T_n(\theta)$.

LEMMA **2** *Under assumptions (A2)-(A7), if (i)* $\frac{\ln n}{n^{1/2}h^d} = o(1)$*, (ii)* $n^{1/2}h^{2(p+1)} = o(1)$*, (iii)* $\mathbb{E}|\epsilon|^\nu < \infty$*,* $\mathbb{E}|\varphi^2(X)|^\nu < \infty$ *and* $\mathbb{E}|\epsilon\varphi^2(X)|^\nu < \infty$*, and (iv) for any* $t > 1$*,* $\mathbb{E}|\epsilon_1\epsilon_t\varphi^2(X_t)|^\nu < \infty$ *and* $\mathbb{E}|\epsilon_1\epsilon_t\varphi^2(X_1)|^\nu < \infty$*, then for any* $\theta \in \Theta$*,*

$$T_n(\theta) = n^{-1}\sum_{i=1}^n \left[2Y_i(\theta)\Delta(\theta, X_i) - \Delta^2(\theta, X_i)\right]\varphi^2(X_i) + o_p(n^{-1/2}).$$

This is a very simple asymptotic representation of $T_n(\theta)$ as a sum of weakly dependent random variables whose mean is exactly $T(\theta)$. The simplicity of this representation comes from the

fact that it is free from the bandwidth parameter h and the fact that it depends only on $Y(\theta)$, $\Delta(\theta, x)$ and on the known function φ.

REMARK **1**

- *Assumptions (i) and (ii) imply that* $n^{1/2}h^d \to \infty$ *and* $h^{2(p+1)-d} \to 0$. *For this conditions to hold, we need that* $p > d/2 - 1$. *In other words, to guarantee the optimal root-n convergence rate, the order of the local polynomial approximation should increase as the dimension* d *of the covariates* X *increases. This result appears to be new in the kernel smoothing literature.*

- *All the bandwidth restrictions given in (A6), (i) and (ii) are fulfilled whenever the assumption (i') given bellow is satisfied:*
  (i') $\delta \geq 4$, $p > d/2 - 1$ *and* $h_n \sim (n^{-1} \ln n)^a$ *for some* $\frac{1}{4(p+1)} < a < \frac{2}{d}$.

- *From the proofs given in the Appendix, it easy to see that Lemma 1 remains valid if instead of the statistic* T_n *we use* T_n^0. *However this is not the case when we consider Lemma 2. In fact, without adding extra assumptions, one can only state that,*

$$T_n^0(\theta^*) = n^{-1} \sum_{i=1}^{n} \Delta^2(\theta^*, X_i)\varphi^2(X_i) + \sup_{x \in D} |\Delta(\theta^*, x)| \left\{ O_p(\left(\ln n/(nh^d)\right)^{1/2}) + O_p(h^{p+1}) \right\}.$$

  From this expression it is clear that in order to achieve a higher rate of convergence for T_n^0, *one needs to impose some restrictions on* $\Delta(\theta^*, x)$, *such as for example* $\Delta(\theta^*, x) = c_n \Delta_n(x)$, *for certain sequences* $c_n \to 0$ *and a bounded function* $\Delta_n(x)$.

Of course, all the results discussed above are only available under our restrictions on the bandwidth parameter as formulated in the assumptions given in Lemma 2. Other bandwidth choices may also be considered but this will inevitably affect the limiting distribution of the estimator. It is also important to note that no restriction was made on the parametric

estimation procedure and so one can use any available parametric method. Here, for its simplicity and desirable properties, we suggest to use the least squares technique. Thus we propose to estimate θ by

$$\hat{\theta} = \arg\min_{\theta\in\Theta} n^{-1} \sum_{i=1}^{n} Y_i^2(\theta)\varphi^2(X_i). \tag{4}$$

In this definition we used the weighted version of the least squares estimator, since, as motivated in Section 4.2, we are interested in assessing the quality of the parametric model $m(\theta, x)$ within the support of $\varphi(x)$. From Corollary 3.1 in Domowitz and White (1982) we claim that, under Assumptions (A), $\hat{\theta}$ converges with probability 1 to

$$\theta^* = \arg\min_{\theta\in\Theta} T(\theta). \tag{5}$$

Interestingly, in this particular case, the parameter of interest $T(\theta^*)$ coincides with $\min_{\theta\in\Theta} T(\theta)$. In other words, our parameter of interest becomes exactly the minimum L^2-distance between m and the parametric family $\mathcal{M}$. Moreover, since θ^* minimize $T(\theta)$ in the interior of Θ, Assumption (A1) implies that,

$$\begin{aligned} -2B^T &= \mathbb{E}\left[-2\dot{m}^T(\theta^*, X)\Delta(\theta^*, X)\varphi^2(X)\right] \\ &= \mathbb{E}\left[\frac{\partial\Delta^2(\theta^*, X)}{\partial\theta}\varphi^2(X)\right] = \frac{dT(\theta^*)}{d\theta} = 0, \end{aligned}$$

so no bias estimation or correction will be needed. This result, together with Lemma 1 and Lemma 2, leads to the following theorem.

THEOREM **1** *Under Assumptions (A), if the conditions (i)-(iv) given in Lemma 2 are satisfied, then*

$$T_n(\hat{\theta}) = n^{-1} \sum_{i=1}^{n} \left[2Y_i(\theta^*)\Delta(\theta^*, X_i) - \Delta^2(\theta^*, X_i)\right] \varphi^2(X_i) + o_p(n^{-1/2}),$$

where $\hat{\theta}$ and θ^* are given by (4) and (5), respectively.

In the next section we demonstrate some applications of this result.

4 Closeness of Parametric Approximation and Explanatory Power of a Covariate

4.1 An inadequacy index

Following an original idea of Doksum and Samarov (1995), we introduce a measure of model deficiency based on the L^2 loss function. The idea is in the spirit of the well-known Pearson's correlation ratio $\eta^2 = \dfrac{\mathbb{V}ar[m(X)]}{\mathbb{V}ar(Y)} = 1 - \dfrac{\mathbb{V}ar(\epsilon)}{\mathbb{V}ar(Y)}$. This coefficient gives the fraction of variability of Y explained by X through the true mean regression function $m(x)$. In other words this coefficient measures the ability of the covariates, in a correctly specified model, to distinguish differing outcomes. It is a direct consequence of the ANOVA decomposition $\mathbb{V}ar(Y) = \mathbb{V}ar(m(X)) + \mathbb{V}ar(\epsilon)$. This equality is just a special case of a more general one given by $\mathbb{E}(Y - g(X))^2 = \mathbb{E}(m(X) - g(X))^2 + \mathbb{E}(\epsilon^2)$, where g is any real-valued function with $\mathbb{E}(g^2(X)) < \infty$. Now, letting $g(x) = m(\theta, x)$ we get $\mathbb{E}(Y - m(\theta, X))^2 = \mathbb{E}(m(X) - m(\theta, X))^2 + \mathbb{E}(\epsilon^2)$. This is a decomposition of the parametric residual variation into unexplained variation due to model misspecification and error variation. Therefore, the coefficient

$$\zeta^2(\theta) = \frac{\mathbb{E}(m(X) - m(\theta, X))^2}{\mathbb{E}(Y - m(\theta, X))^2} = 1 - \frac{\mathbb{E}(Y - m(X))^2}{\mathbb{E}(Y - m(\theta, X))^2}$$

is the fraction of the parametric residual variation that can be completely attributed to the lack of fit in the parametric function $m(\theta, x)$ which will be described shortly as the missed fraction of variation or the inadequacy index. In particular, the index $\zeta^2 := \zeta^2(\theta^*)$ is the missed fraction of variation after adjusting the best possible parametric model from the family $\mathcal{M} = \{m(\theta, x),\ \theta \in \Theta\}$. It is interesting to note that η^2 and ζ^2 behave in an opposite way in the sense that while a large value of η^2 indicates a good explanatory power of the covariates, the smallest the value of ζ^2 the best the model is. Both of these coefficients range from 0 to

1. However, if $\mathcal{M}$ includes constants (which should always be the case) then $\zeta^2 \leq \eta^2$, with equality if $m(\theta^*, X) = \mathbb{E}(Y)$. This is the case when the parametric model fails to capture any variability in the data.

4.2 An estimator of ζ^2 and its asymptotic variance

For the same reasons as previously mentioned we prefer to work with a weighted version of ζ^2, which corresponds to $T(\theta^*)/S^2(\theta^*)$, where $S^2(\theta) = \mathbb{E}\left[(Y - m(\theta, X))^2 \varphi^2(X)\right]$. Hereafter $T(\theta^*)/S^2(\theta^*)$ will be denoted by $\zeta^2 \equiv \zeta^2(\theta^*)$. An obvious estimator of this index is provided by $\hat{\zeta}^2 := T_n(\hat{\theta})/S_n^2(\hat{\theta})$, where $T_n(\theta)$ is given by (3), $\hat{\theta}$ is given by (4) and $S_n^2(\theta) = n^{-1}\sum_{i=1}^n (Y_i - m(\theta, X_i))^2 \varphi^2(X_i)$. Based on the result of Theorem 1, the next theorem gives a very useful asymptotic expression for $\hat{\zeta}^2$.

THEOREM **2** *Under Assumptions (A), if the conditions (i)-(iv) given in Lemma 2 are satisfied, then*

$$\hat{\zeta}^2 - \zeta^2 = n^{-1}\sum_{i=1}^n \xi_i + o_p(n^{-1/2}),$$

where ξ_i is a shortcut for $\xi_i(\theta^*)$, $\xi_i(\theta) = \left[(1-\zeta^2)Y^2(\theta) - \epsilon_i^2\right]\varphi^2(X_i)/S^2(\theta)$ with $Y_i(\theta) = Y_i - m(\theta, X_i)$, and $\epsilon_i = Y_i - m(X_i)$.

A direct consequence of this theorem is the asymptotic normality of $\hat{\zeta}^2$. In fact, applying the central limit theorem to the strong mixing sequence $\{\xi_t\}$, see for example Theorem 2.21 in Fan and Yao (2003), we have that under the assumptions of Lemma 2:

$$\sqrt{n}(\hat{\zeta}^2 - \zeta^2) \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

where the asymptotic variance $\sigma^2 \equiv \sigma^2(\theta^*)$, with $\sigma^2(\theta) := \lim_{n\to\infty} n^{-1}\mathbb{V}ar\left(\sum_{t=1}^n \xi_t(\theta)\right) = \mathbb{V}ar[\xi_1(\theta)] + 2\sum_{t>1}\mathbb{C}ov(\xi_1(\theta), \xi_t(\theta))$.

To use this property in practice we need a consistent estimator for the asymptotic variance σ^2. In the case of i.i.d. data this can be done by using the classical sample variance estimator.

In the presence of correlated data, we adopt here the moving block bootstrap (MBB) procedure as proposed by Künsch (1989) and Liu and Singh (1992). This approach allows us to estimate σ^2 without making any parametric model restriction and without resort to any Monte Carlo simulation. A detailed description of this method and its merits over other competing methods can be found in the book by Lahiri (2003). To fix ideas, we start by splitting the "data" $\{\xi_i\}_{1\leq i\leq n}$ into $N := n-l+1$ blocks $\mathcal{B}_i = \{\xi_i, \ldots, \xi_{i+l-1}\}$, $i = 1, \ldots, N$, of length $l \equiv l_n \in [1, n]$. We require that $l \to 0$ and $l = o(n^{-1})$. Let $U_i = l^{-1}\sum_{j=i}^{i+l-1}\xi_j$ be the sample mean of the i-th block and $\bar{U}$ the sample mean of $\{U_1, \ldots, U_N\}$. Like in the i.i.d. case ($l = 1$), the MBB estimator of σ^2 is $\hat{\sigma}^2 = lN^{-1}\sum_{i=1}^{N}\left(U_i - \bar{U}\right)^2$. By Theorem 3.1 in Lahiri (2003), one can easily check that under the assumption of Lemma 2, $\hat{\sigma}^2$ converges in probability to σ^2. However, this estimator depends on the unknown parameter θ^*. To overcome this problem, we simply suggest to plugging-in $\hat{\theta}$ into the definition of $\hat{\sigma}^2(\theta)$ to get $\hat{\sigma}^2(\hat{\theta})$ as our feasible estimator of the asymptotic variance. Following similar techniques as those used in the Appendix, it can be shown that $\hat{\sigma}^2(\hat{\theta}) - \hat{\sigma}^2(\theta^*) = o_p(1)$ and so that $\hat{\sigma}^2(\hat{\theta})$ is consistent.

4.3 Validation of a parametric model

A first and direct application of the previous results is that one can construct an asymptotically valid Wald-type confidence interval for ζ that is given by $\hat{\zeta} \pm \frac{\hat{\sigma}}{\sqrt{n}}z_{1-\alpha/2}$, where z_α is the α-quantile of the standard normal distribution and $\hat{\sigma}^2 = \hat{\sigma}^2(\hat{\theta})$. Although this confidence interval gives us valuable information about the quality of the parametric model $m(\theta, x)$, we still need a formal approach to test the goodness-of-fit hypothesis

$$H_0 : m \in \mathcal{M} \quad \text{versus} \quad H_1 : m \notin \mathcal{M}.$$

In term of ζ, this hypothesis can be formulated as

$$H_0 : \zeta^2 = 0 \quad \text{versus} \quad H_1 : \zeta^2 > 0. \tag{6}$$

Unfortunately, $\hat{\zeta}^2$ cannot be directly used as a test statistic for (6) since under the null hypothesis ξ vanishes and so does σ^2. The asymptotics in such a form have been noted before by many authors; see for example Fan and Li (1996). This degeneracy can be handled by considering higher-order terms in the expansion of $T_n(\hat{\theta})$. In fact, under H_0, it can be shown that $T_n(\hat{\theta}) = J''_{n,1} + o_p(n^{-1}h^{-d/2})$, where $J''_{n,1}$ is a degenerate U-statistics as defined by (21) in the Appendix. This remark can be used to prove the asymptotic normality of $nh^{d/2}\hat{\zeta}^2$ under correct specification and so to get a valid test statistic for the hypothesis (6). Such an approach will lead inevitably to the curse of dimensionality as the convergence rate decreases with d. Here instead of (6) we propose to test the following hypothesis

$$H_{\pi,0} : \zeta^2 \geq \pi \quad \text{versus} \quad H_{\pi,1} : \zeta^2 < \pi, \tag{7}$$

where $\pi \in (0, 1)$ is a small constant that can be considered by the analyst as a tolerable missed fraction of variation. In the literature, (7) is known as a neighborhood hypothesis or "precise" hypothesis; see Hodges and Lehmann (1954). The drawbacks of testing the classical hypothesis (6) with the hypothesis (7) were largely documented by many authors; see for example Dette and Munk (1998) and the references given therein. To cite just an argument in favor of the concept of neighborhood testing, observe that (7) is designed to provide evidence in favor of the tested model $m(\theta, x)$ while the latter cannot be confirmed even if the p-value associated with (6) is large. For a detailed discussion of many other aspects related with neighborhood hypothesis we refer to Dette and Munk (2003).

As noted by those authors, the main difficulty with neighborhood testing is the need of the asymptotic distribution of the test statistic not only under the assumption that $m \in \mathcal{M}$, as is classically done in the literature of goodness-of-fit testing, but at any point in the model space $\mathcal{M}$. The method adopted in this work consists in studying the estimated distance $T_n(\hat{\theta})$ without restrictions on the model specification and allows us to easily overcome this difficulty.

In fact, by Theorem 2, we directly conclude that a consistent critical region for $H_{\pi,0}$ is provided by

$$\hat{\zeta}^2 < \pi + z_\alpha \frac{\hat{\sigma}}{\sqrt{n}}. \tag{8}$$

Another difficulty usually associated with this procedure is the selection of π. In our case, this is facilitated by the fact that the coefficient ζ^2 is a proportion bounded above by 1 and hence π should be as well. One can also get around this difficulty by reformulating the problem of testing (7) in terms of interval estimation. In fact, an asymptotic $100 \times (1-\alpha)\%$ upper confidence interval for ζ is given by $[0, \zeta^2_{n,+}]$, with

$$\zeta^2_{n,+} = \hat{\zeta}^2 + \frac{\hat{\sigma}}{\sqrt{n}} z_{1-\alpha}.$$

Of course, the rule given by (8) is equivalent to $\zeta^2_{n,+} \leq \pi$. In any case, given a sufficiently large data set, since $P(\zeta^2 \leq \zeta^2_{n,+}) \to 1-\alpha$, one can state, at risk $\alpha \times 100\%$, that the missed fraction of variation does not exceed $\zeta^2_{n,+}$. According to the value of the latter, the tested model can be judged as admissible or not.

REMARK **2** *If one is interested in the explanatory power of a random covariate X then, instead of ζ^2, η^2 should be used. An estimator of the weighted version of the latter, i.e., $\mathbb{V}ar(m(X)\varphi(X))/\mathbb{V}ar(Y\varphi(X))$, is given by $\hat{\eta}^2 = T_n - \bar{Y}^2/S_n^2$, where $\bar{Y} = n^{-1}\sum_i Y_i\varphi(X_i)$, $S_n^2 = n^{-1}\sum_{i=1}^n (Y_i\varphi(X_i) - \bar{Y})^2$, and $T_n = n^{-1}\sum_{i=1}^n \left(2Y_i\hat{m}(X_i) - \hat{m}^2(X_i)\right)\varphi^2(X_i)$, with $\hat{m}(x)$ being the local polynomial estimator of $m(x)$ of order p. Then T_n can be seen as a special case of $T_n(\theta)$ with $m(\theta, X) \equiv 0$. Therefore, as a direct consequence of Lemma 2, it follows that*

$$T_n = n^{-1}\sum_i (2Y_i m(X_i) - m^2(X_i))\varphi(X_i) + o_p(n^{-1/2}).$$

From this result and following a similar approach as in the proof of Theorem 2, it can be shown that, under the assumptions of Lemma 2,

$$\hat{\eta}^2 - \eta^2 = \sigma^{-2} n^{-1} \sum_{i=1}^n \left[(1-\eta^2)(Y_i\varphi(X_i) - \mu)^2 - \epsilon_i^2\varphi^2(X_i)\right] + o_p(n^{-1/2}),$$

where $\mu = \mathbb{E}(Y\varphi(X))$ *and* $\sigma^2 = \mathbb{V}ar(Y\varphi(X))$. *As we have done for* ζ^2, *this asymptotic expression can be used to provide consistent confidence intervals and tests procedures for* η^2.

5 Monte Carlo Simulations

In this section we report the results of an extensive simulation study that was designed to evaluate the finite sample performance of $\hat{\eta}^2$ and $\hat{\zeta}^2$ and their asymptotic properties as stated in the previous sections. The simulation considers univariate and multivariate cases with both i.i.d. data and weak dependent data using the weight function $\varphi(t) = I(0 \leq t \leq 1)$ and $N := 2000$ replications. We generate $n := 200$ data according to the following model

$$Y_t = m_1(X_t) + \lambda m_2(X_t) + \tau\epsilon_t,$$

where $X_t \sim Unif[-\varepsilon, 1+\varepsilon]$ and $\epsilon_t \sim \mathcal{N}(0,1)$. Here ε was chosen so that $P(0 \leq X_t \leq 1) = 0.95$. For $d = 1$, m_1 and m_2 are given by

$$m_1(x) = 6 + 2x, \;\; m_2(x) = \sin(\sqrt{(3\pi x + \pi)^2}).$$

We are interested in measuring and testing the quality of the covariate X and of the linear parametric model $m(\theta, x) = \theta_0 + \theta_1 x$. To this end we vary the values of λ and τ. The linear model $m(\theta, x)$ is correct only when $\lambda = 0$. In this case $\zeta^2 = 0$, but as λ increases, $m(\theta, x)$ becomes more and more inadequate and $\zeta^2 \nearrow 1$. On the other hand, by increasing τ, X becomes more and more irrelevant and $\eta^2 \searrow 0$. Table 1 shows the values of τ and λ used to generate the data together with the corresponding values of η^2 and ζ^2.

In the two-dimensional case, we choose

$$m_1(x) = 6 + 2x_1 + 2x_2, \;\; m_2(x) = \sin(\sqrt{(3\pi x_1 + \pi)^2 + (3\pi x_2 + \pi)^2}),$$

and for $d = 3$,

$$m_1(x) = 6 + 2x_1 + 2x_2 + 2x_3, \;\; m_2(x) = \sin(\sqrt{(3\pi x_1 + \pi)^2 + (3\pi x_2 + \pi)^2 + (3\pi x_3 + \pi)^2}).$$

The covariates are independent of each other, independent of the error variable ϵ_t and are $Unif[-\varepsilon, 1+\varepsilon]$.

Table 1: *The true values of* η^2 *and* ζ^2 *in* %

	η^2					ζ^2				
λ	0	0.8	1.5	2.5	5	0	0.8	1.5	2.5	5
τ			d=1					d=1		
0.5	95.44	95.48	95.88	96.72	98.39	0.00	53.82	80.38	91.93	97.85
1	83.95	84.09	85.33	88.04	93.84	0.00	22.56	50.60	74.00	91.93
2	56.66	56.93	59.26	64.79	79.21	0.00	6.79	20.38	41.56	74.00
τ			d=2					d=2		
0.5	96.58	96.78	97.11	97.68	98.77	0.00	55.28	81.30	92.36	97.98
1	87.59	88.24	89.36	91.31	95.25	0.00	23.60	52.07	75.12	92.36
2	63.83	65.22	67.73	72.42	83.36	0.00	7.17	21.36	43.00	75.12
τ			d=3					d=3		
0.5	97.32	97.38	97.56	97.94	98.82	0.00	56.01	81.74	92.56	98.03
1	90.09	90.29	90.92	92.25	95.46	0.00	24.14	52.81	75.67	92.56
2	69.45	69.93	71.47	74.84	84.01	0.00	7.37	21.86	43.73	75.67

To calculate each estimator we use the local linear smoother with the Epanechnikov kernel function. As a data-driven bandwidth selection criterion, we use the least squares cross-validation method; see Xia and Li (2002) and also Li and Racine (2004). In the multivariate cases, we use the product kernel and let all components of each bandwidth vector to be equal.

Our first objective is to perform a comparison between two estimators of ζ^2: $\hat{\zeta}^2 = T_n(\hat{\theta})/S_n^2(\hat{\theta})$ and $\hat{\zeta}_0^2 = T_n^0(\hat{\theta})/S_n^2(\hat{\theta})$ and two estimators of η^2: $\hat{\eta}^2 = T_n/S_n^2$ and $\hat{\eta}_0^2 = T_n^0/S_n^2$, where $T_n^0 = n^{-1}\sum_{i=1}^n \hat{m}^2(X_i)\varphi^2(X_i)$. In Table 2 we report the root mean squared error (RMSE) for all those estimators, for $d = 1, 2, 3$, $\tau = 0.5, 1, 2$ and $\lambda = 0, 0.8, 1.5, 2.5, 5$. From this table we observe that both $\hat{\eta}^2$ and $\hat{\eta}_0^2$ perform very well with respect to the MSE criterion, with a clear advantage of $\hat{\eta}^2$ over $\hat{\eta}_0^2$. This is also the case for ζ^2. In fact, we obtain almost

systematically a smaller MSE when we use our estimators $T_n(\hat{\theta})$ instead of $T_n^0(\hat{\theta})$. The only exception happened with a very small value of λ ($\zeta^2 \to 0$) where $\hat{\zeta}_0^2$ provided a slightly better result. The performances of all these estimators are clearly affected by the true values of the corresponding parameters and by the dimensionality d, along with interaction between this two factors. For example, when $d = 1$ or 2, as ζ increases the MSE of $\hat{\zeta}^2$ initially increases and then rapidly decreases. Also $\hat{\zeta}_0^2$ behaves similarly but its MSE increases rapidly and then decreases slowly. For $d = 3$, we can see that $\hat{\zeta}_0^2$ behaves badly as ζ^2 increases and approaches 1. As expected, increasing the covariate dimensionality d causes the MSE to increase but $\hat{\eta}_0^2$ and $\hat{\zeta}_0^2$ are clearly more sensitive to the curse of dimensionality. For example, for $\tau = 1$ and $\lambda = 1.5$, when d moves from 1 to 3, the MSE of $\hat{\zeta}_0^2$ increases by a factor of 6.8 whereas the MSE of $\hat{\zeta}^2$ increases by only a factor of 1.8. This becomes even more striking when we consider the optimal bandwidths (the results are not shown here). To give just an example, under the same scenario as above and using the optimal bandwidth for both estimators, the MSE of $\hat{\zeta}_0^2$ increases by a factor of 12.3 whereas the MSE of $\hat{\zeta}^2$ increases by only a factor of 0.8. This definitely demonstrates the advantages of the proposed estimators over the "classical" ones.

Table 3 gives the optimal root mean squared error (RMSE*) obtained using the optimal bandwidth (the one that minimize the MSE). For easy comparison, the table shows the RMSE obtained using the data-driven bandwidth. We also report the bias and the standard deviation (S.D.) for both $\hat{\eta}^2$ and $\hat{\zeta}^2$ when $d = 2$. The other results are not displayed here for the sake of brevity. Both the absolute value of the bias and the variance increase with d. For $d = 3$, we have also noticed that both $\hat{\eta}^2$ and $\hat{\zeta}^2$ become more biased as λ increases. Globally, the variance is the main component of the mean squared error. Typically, its contribution decreases with λ and increases with τ and it is more sensitive to the alteration in τ. Regarding the usefulness of the bandwidth selection procedure, we have observed that, in general, the resulting values for the two estimators $\hat{\eta}^2$ and $\hat{\zeta}^2$ obtained using the data driven bandwidth

were quite close to the optimal ones obtained using the optimal bandwidth. However, the dimensionality has again a negative impact. The observed averaged value of the difference in mean squared error $|MSE - MSE^*|$ was 0.004 for $\hat{\eta}^2$ and 0.021 for $\hat{\zeta}^2$ with maximum value equal to 0.015 and 0.062 for $\hat{\eta}^2$ and $\hat{\zeta}^2$, respectively. We have also observed that the loss of efficiency due to the estimated bandwidth is larger for $\hat{\eta}_0^2$ and $\hat{\zeta}_0^2$. In fact, for these estimators the average (maximum) of $|MSE - MSE^*|$ was 0.019 (0.074) and 0.048 (0.137), respectively. This indicates a greater robustness of $\hat{\eta}^2$ and $\hat{\zeta}^2$ to bandwidth misspecification.

Another objective of this simulation study is to verify the validity of the proposed testing procedures. As the estimation of the asymptotic variance plays a crucial role, we start by checking the finite sample performance of our variance estimator of $\hat{\zeta}^2$ as introduced in Section 4.2 . To estimate the asymptotic variance of $\hat{\eta}^2$ we use exactly the same procedure. Table 3 illustrates the consistent nature of the these estimators via a small mean squared error. The behavior of the latter differs for the various cases but we observe that the MSE performance is better than expected even for $d = 3$. To complete the picture, Table 5 shows the coverage probability for the lower and upper confidence intervals (LCI and UCI) for η^2 and ζ^2, respectively, of nominal level 95% computed using $\hat{\eta}^2$ and $\hat{\zeta}^2$. The accuracy of confidence limits was assessed by calculating the proportion of times the true value was above or below the confidence limit. For both estimators, the observed coverage probability was often different from the expected values especially for $\tau = 2$. For $\tau = 0.5$ or for $\lambda = 0$, our intervals appear overly conservative but as d increases they become anti-conservative. For $d = 3$ the results were unsatisfactory especially for ζ^2 (the results are not shown). This is not really surprising given that we use the same bandwidth parameter that we used to estimate our parameters. As we have seen, this bandwidth is appropriate for MSE minimization but now we need a balance between narrow confidence interval and minimum coverage error. To illustrate the benefit of our method when the bandwidth is correctly specified, Table 6 gives the optimal

coverage probability obtained using a fixed but optimal bandwidth (the one that minimize the coverage error) for $d = 3$. These results clearly demonstrate the usefulness and the good performance of the Normal approximation and the resulting confidence limits given a "good" bandwidth parameter. Theoretically, the optimal bandwidth parameter can be determined by studying how fast $\sqrt{n}(\hat{\zeta} - \zeta)$ converges to its limit using, for instance, Edgeworth expansions; see for example Hall (1992). Practically, bootstrap methods may be used to estimate the coverage error associated with a given bandwidth and so to approximate the optimal one, leading to further improvements of the confidence intervals. This is however beyond the scope of the present work but may be a topic of further research.

Finally, the entire simulation study was re-run using data with correlated errors generated according to an autoregressive $AR(\rho)$ process of order 1 with different values of the autocorrelation parameter ρ. To be more precise, we generate ϵ_t according to the model $\epsilon_t = \rho\epsilon_{t-1} + \omega_t$, with $\omega_t \sim$ i.i.d. $\mathcal{N}(0, 1)$. To choose the block length needed for the asymptotic variance estimator (see Section 4.2), we use the block selection method of Patton et al. (2009) provided by the R package *np* of Hayfield and Racine (2008). The results for the dependent case were globally similar to those obtained with i.i.d. data and so we only provide here a brief summary given in Table 7 for the case $\rho = 0.95$ and $d = 1$. This table (and other results not shown here) clearly indicate that this dependency structure has almost no effect on our estimators and the proposed confidence intervals. Nevertheless, comparing the i.i.d. case and the dependent case is difficult here because changing ρ affects the variation in Y and so it also affects η^2, ζ^2 and the variance of their corresponding estimators. For example, when $d = 1$, $\tau = 0.5$ and $\lambda = 0$, $\eta^2 \approx 0.67$ and $\mathbb{V}ar(\hat{\eta}^2) \approx 0.55$ for $\rho = 0.95$, while for $\rho = 0$ (i.i.d.) $\eta^2 \approx 0.95$ and $\mathbb{V}ar(\hat{\eta}^2) \approx 0.02$.

Table 2: $100 \times RMSE$ *for* $\hat{\eta}_0^2$, $\hat{\eta}^2$, $\hat{\zeta}_0^2$ *and* $\hat{\zeta}^2$.

		λ						λ				
		0	0.8	1.5	2.5	5		0	0.8	1.5	2.5	5
τ				d=1						d=1		
0.5	$\hat{\eta}_0$	3.702	3.057	2.302	1.331	1.825	$\hat{\zeta}_0$	1.044	9.286	8.043	6.192	3.897
	$\hat{\eta}$	1.228	1.078	0.880	0.618	0.411	$\hat{\zeta}$	1.682	5.146	2.574	1.131	0.320
1	$\hat{\eta}_0$	6.862	5.749	4.619	2.824	2.637	$\hat{\zeta}_0$	1.044	7.255	9.226	8.615	6.187
	$\hat{\eta}$	3.319	3.086	2.628	1.858	0.810	$\hat{\zeta}$	1.682	6.295	5.486	3.322	1.131
2	$\hat{\eta}_0$	9.688	8.872	7.552	5.433	3.904	$\hat{\zeta}_0$	1.044	4.392	6.953	8.990	8.603
	$\hat{\eta}$	5.765	5.747	5.312	4.483	2.635	$\hat{\zeta}$	1.682	4.085	6.342	6.051	3.322
τ				d=2						d=2		
0.5	$\hat{\eta}_0$	3.507	3.966	5.999	8.450	11.083	$\hat{\zeta}_0$	1.921	17.825	20.987	19.361	14.968
	$\hat{\eta}$	0.936	0.743	0.657	0.591	0.874	$\hat{\zeta}$	2.217	5.849	2.710	1.069	1.552
1	$\hat{\eta}_0$	6.650	6.352	7.594	10.436	14.313	$\hat{\zeta}_0$	1.873	9.962	17.225	20.657	19.334
	$\hat{\eta}$	2.578	2.276	1.908	1.417	0.727	$\hat{\zeta}$	2.127	7.788	6.207	3.541	1.074
2	$\hat{\eta}_0$	10.206	9.969	9.476	11.070	16.310	$\hat{\zeta}_0$	1.936	4.660	9.839	15.186	20.649
	$\hat{\eta}$	5.289	5.349	4.752	3.998	2.321	$\hat{\zeta}$	2.235	5.937	7.822	7.035	3.537
τ				d=3						d=3		
0.5	$\hat{\eta}_0$	3.150	6.270	10.475	16.454	56.755	$\hat{\zeta}_0$	2.294	25.516	35.952	38.987	40.211
	$\hat{\eta}$	0.725	0.819	1.640	3.282	7.043	$\hat{\zeta}$	3.248	7.165	6.897	8.187	9.923
1	$\hat{\eta}_0$	5.910	8.864	11.893	17.537	56.415	$\hat{\zeta}_0$	2.350	15.263	24.139	33.452	39.586
	$\hat{\eta}$	1.980	2.111	1.962	2.546	6.092	$\hat{\zeta}$	3.356	9.988	7.455	6.524	8.188
2	$\hat{\eta}_0$	9.452	10.900	13.533	20.256	29.625	$\hat{\zeta}_0$	2.308	5.709	14.164	20.291	33.186
	$\hat{\eta}$	4.637	5.115	5.251	4.548	5.064	$\hat{\zeta}$	3.265	5.000	10.243	8.413	6.039

Table 3: $100 \times RMSE^*$, $100 \times RMSE$, $100 \times Bias$ *and* $100 \times S.D.$ *for* $\hat{\eta}^2$ *and* $\hat{\zeta}^2$ *with* $d = 2$

		$\hat{\eta}^2$					$\hat{\zeta}^2$				
τ	λ	0	0.8	1.5	2.5	5	0	0.8	1.5	2.5	5
0.5	RMSE*	0.662	0.596	0.602	0.926	1.927	0.000	4.069	2.145	1.369	2.398
	RMSE	0.936	0.743	0.657	0.591	0.874	3.217	5.849	2.710	1.069	1.552
	Bias	-0.513	-0.282	-0.337	-0.418	-0.669	1.435	2.754	1.231	0.283	-1.103
	S.D.	0.783	0.688	0.564	0.418	0.563	2.879	5.160	2.414	1.031	1.092
1	RMSE*	2.349	2.127	1.799	1.346	1.250	0.000	4.356	4.154	2.537	1.369
	RMSE	2.578	2.276	1.908	1.417	0.727	3.127	7.788	6.207	3.541	1.074
	Bias	-0.514	-0.098	0.101	0.098	-0.268	1.401	3.298	2.904	1.655	0.290
	S.D.	2.526	2.274	1.905	1.414	0.676	2.796	7.056	5.486	3.131	1.034
2	RMSE*	5.116	4.794	4.267	3.411	1.825	0.000	2.590	4.274	4.533	2.537
	RMSE	5.289	5.349	4.752	3.998	2.321	3.235	6.937	7.822	7.035	3.537
	Bias	0.010	-0.087	0.960	1.300	0.822	1.429	0.902	3.176	3.242	1.656
	S.D.	5.289	5.348	4.654	3.780	2.171	2.902	6.878	7.148	6.244	3.125

Table 4: $100 \times RMSE$ *for the estimated asymptotic variance of* $\hat{\eta}^2$ *and* $\hat{\zeta}^2$.

		$\hat{\eta}^2$					$\hat{\zeta}^2$				
	$\tau \vert \lambda$	0	0.8	1.5	2.5	5	0	0.8	1.5	2.5	5
$d = 1$	0.5	1.940	1.364	0.691	0.211	0.028	3.856	12.327	3.307	0.646	0.049
	1	11.328	8.623	5.044	2.008	0.381	3.856	18.337	13.444	5.336	0.646
	2	14.488	13.073	10.533	7.826	3.462	3.856	13.657	18.189	16.023	5.336
$d = 2$	0.5	0.988	0.618	0.353	0.138	16.652	7.119	19.859	4.214	0.652	22.416
	1	7.115	5.061	3.014	1.325	0.239	6.970	28.357	21.812	7.320	0.653
	2	15.214	14.212	11.016	7.873	3.084	7.105	16.824	28.711	27.034	7.315
$d = 3$	0.5	0.553	0.563	13.454	9.476	15.720	8.429	18.980	3.059	2.182	4.886
	1	4.390	4.034	3.068	2.677	14.501	8.471	35.351	21.536	4.772	2.221
	2	13.721	14.352	12.178	7.598	2.679	8.486	20.825	35.183	27.299	5.315

Table 5: *Lower and upper confidence intervals (LCI and UCI) for η^2 and ζ^2, respectively, using data driven bandwidth. Nominal coverage = 95%.*

		LCI for η^2					UCI for ζ^2				
	$\tau\|\lambda$	0	0.8	1.5	2.5	5	0	0.8	1.5	2.5	5
$d=1$	0.5	97.71	96.24	96.27	97.20	99.93	98.55	96.76	98.47	98.87	99.20
	1	93.62	90.33	90.13	90.13	90.67	100.00	93.71	96.80	98.00	98.87
	2	93.14	89.67	87.93	86.33	84.47	98.90	83.90	93.40	96.33	98.00
$d=2$	0.5	98.60	97.20	97.53	99.53	100.00	98.52	94.80	95.67	94.60	80.33
	1	93.47	89.73	87.67	87.87	95.27	98.84	90.47	94.87	95.47	94.40
	2	91.87	89.87	85.60	83.33	82.31	98.79	82.67	90.00	93.87	95.47

Table 6: *Lower and upper confidence intervals (LCI and UCI) for η^2 and ζ^2, respectively using optimal bandwidth. Nominal coverage = 95%.*

	LCI for η^2					UCI for ζ^2				
$\tau\|\lambda$	0	0.8	1.5	2.5	5	0	0.8	1.5	2.5	5
0.5	95.10	95.10	96.00	98.89	99.90	100.00	98.20	99.60	97.80	100.00
1	94.60	94.40	94.40	98.80	93.89	100.00	96.20	93.30	99.10	97.80
2	92.50	95.00	95.00	96.20	92.60	100.00	96.20	93.60	93.70	99.10

Table 7: $100 \times RMSE$ *and lower and upper confidence intervals (LCI and UCI) for η^2 and ζ^2, respectively, using data driven bandwidth. Nominal coverage = 95%.*

		$\hat{\eta}^2$						$\hat{\zeta}^2$				
$\tau\|\lambda$		0	0.8	1.5	2.5	5		0	0.8	1.5	2.5	5
0.5	RMSE	1.23	1.08	0.88	0.62	0.41	RMSE	1.68	5.14	2.57	1.13	0.32
	LCI	97.71	96.23	96.26	97.20	99.93	UCI	100.00	96.76	98.46	98.86	99.20
1	RMSE	3.32	3.88	2.63	1.86	0.81	RMSE	1.69	2.52	5.48	3.32	1.13
	LCI	93.62	93.81	90.13	90.13	90.67	UCI	100.00	95.00	96.80	98.00	98.87

6 Appendix

This appendix collects proofs of the main results stated in the previous sections. Throughout, when we evaluate the order of some terms, the symbol C denotes a generic constant.

For a given $u = 0, \ldots, p$, let $N_u := \frac{(u+d-1)!}{(d-1)!u!}$ be the number of distinct $d-$tuples k with $|k| = u$. Arrange the N_u elements of $\{k,\ |k| = u\}$ in a lexicographical descending order. Let l_u denote this one-to-one mapping, i.e., $l_u(1) = (0, \ldots, 0, u), \ldots, l_u(N_u) = (u, 0, \ldots, 0)$. Now for a given x, define $\gamma_{u,h}(x)$ to be the $N_u \times 1$ vector of the lexicographical arrangement of $\{(x/h)^k, |k| = u\}$, i.e., $\gamma_{u,h}(x) = ((x/h)^{l_u(1)}, \ldots, (x/h)^{l_u(N_u)})^T$. Put $\gamma_h(X_j - x) = (\gamma_{0,h}^T(X_j - x), \ldots, \gamma_{p,h}^T(X_j - x))^T$. This is a column vector of dimension $N := \sum_{u=0}^{p} N_u$. Let $\mathcal{X}_{n,u}(x)$ be the $n \times N_u$ matrix $[\gamma_{u,h}(X_1 - x) \ldots \gamma_{u,h}(X_n - x)]^T$, $\mathcal{X}_n(x) \equiv \mathcal{X}$ be the $N \times N$ matrix $[\mathcal{X}_{n,0}(x) \ldots \mathcal{X}_{n,p}(x)]$, $\mathbf{W}$ be the $n \times n$ diagonal matrix with a diagonal given by $\{n^{-1}K_h(X_j - x),\ j = 1, \ldots, n\}$, $\mathbf{Y}$ be the $n \times 1$ vector $(Y_1, \ldots, Y_n)^T$, $\mathbf{m}(X, \theta)$ be the the $n \times 1$ vector $(m(X_1, \theta), \ldots, m(X_n, \theta))^T$, and $\mathbf{Y}(\theta) = \mathbf{Y} - \mathbf{m}(X, \theta)$.

By definition, see e.g. Masry (1996), the local multivariate polynomial estimator of $\Delta(\theta, x)$ is the first element of the $N \times 1$ vector $\mathbf{S}_n^{-1}(x)\mathcal{X}^T\mathbf{W}\mathbf{Y}(\theta)$, with $\mathbf{S}_n(x) = \mathcal{X}^T\mathbf{W}\mathcal{X}$. Thus, $\hat{\Delta}(\theta, x) = e_{N,1}^T\mathbf{S}_n^{-1}(x)\mathcal{X}^T\mathbf{W}\mathbf{Y}(\theta) = \sum_{j=1}^{n} w_j(x)Y_j(\theta)$, with

$$\begin{aligned} w_j(x) &= e_{N,1}^T\mathbf{S}_n^{-1}(x)\mathcal{X}^T\mathbf{W}e_{n,j} \\ &= n^{-1}e_{N,1}^T\mathbf{S}_n^{-1}(x)\gamma_h(X_j - x)K_h(X_j - x) \end{aligned} \tag{9}$$

where, for $r = n, N$ and $l = 1, \ldots r$, $e_{r,l}$ is the $r \times 1$ vector with the lth element being 1 and the rest of elements being zero. First note that, $\sum_{j=1}^{n} w_j(x) = 1$. Also, observe that

$$\sum_{j=1}^{n} |w_j(x)| \leq n^{-1}||e_{N,1}^T\mathbf{S}_n^{-1}(x)|| \sum_{j=1}^{n} ||\gamma_h(X_j - x)||K_h(X_j - x).$$

It is easy to check that the elements of the matrix $\mathbf{S}_n(x)$ are $s_{n,k}(x) = n^{-1}\sum_{j=1}^{n}\left(\frac{X_j - x}{h}\right)^k K_h(X_j - x)$, $0 \leq |k| \leq 2p$. On the other hand, by assumption (A7), $||\gamma_h(X_j - x)|| \leq C$. So,

$n^{-1}\sum_{j=1}^{n}||\gamma_h(X_j - x)||K_h(X_j - x) \leq Cs_{0,n}(x)$. Now, by Corollary 1 in Masry (1996), $\sup_D |s_{n,k}(x) - f(x)\mu_k| = o_p(1)$, where $\mu_k = \int u^k K(u)du$. So,

$$\sup_{x\in D}\sum_{j=1}^{n}|w_j(x)| = O_P(1). \tag{10}$$

REMARK **3** *$\mathcal{X}^T\mathbf{WY}$ corresponds to the vector $\boldsymbol{\tau}_n(x)$ as defined by equation (2.2) in Masry (1996). Also* $\Delta(\theta, x) = e_{N,1}^T\mathbf{S}_n^{-1}(x)\mathcal{X}^T\mathbf{W}[\mathbf{Y} - \mathbf{m}(\theta, X)] = \hat{m}(x) - e_{N,1}^T\mathbf{S}_n^{-1}(x)\mathcal{X}^T\mathbf{W}\mathbf{m}(\theta, X) = \hat{m}(x) - \hat{m}(\hat{\theta}, x)$.

Proof of Lemma 1

By the mean value Theorem,

$$T_n(\hat{\theta}) - T_n(\theta^*) = (\hat{\theta} - \theta^*)^T\dot{T}_n(\tilde{\theta}_n), \tag{11}$$

where $\tilde{\theta}_n = \theta^* + \eta(\hat{\theta} - \theta^*)$, for some $\eta \in (0, 1)$, and

$$\dot{T}_n(\theta) = 2n^{-1}\sum_i\left[\dot{Y}_i(\theta)\hat{\Delta}(\theta, X_i) + (Y_i(\theta) - \hat{\Delta}(\theta, X_i))\dot{\hat{\Delta}}(\theta, X_i)\right]\varphi^2(X_i), \tag{12}$$

with $\dot{\hat{\Delta}}(\theta, x) = -\sum_j w_j(x)\dot{m}(\theta, X_j)$, and $\dot{Y}_i(\theta) = -\dot{m}(\theta, X_i)$.

From (11), it is clear that Lemma 1 is equivalent to say that $\dot{T}_n(\tilde{\theta}_n) = -2B + o_p(1)$. Let $I_n = n^{-1}\sum_{i=1}^{n}(Y_i(\tilde{\theta}_n) - \hat{\Delta}(\tilde{\theta}_n, X_i))\dot{\hat{\Delta}}(\tilde{\theta}_n, X_i)\varphi^2(X_i)$. Then $\dot{T}_n(\tilde{\theta}_n) = 2n^{-1}\sum_i \dot{Y}_i(\tilde{\theta}_n)\hat{\Delta}(\tilde{\theta}_n, X_i)\varphi^2(X_i) + 2I_n$. First we will show that $I_n = o_p(1)$. Observe that,

$$I_n = I_{n,1} + I_{2,n} + I_{3,n} - I_{4,n} - I_{5,n},$$

with $I_{n,1} = n^{-1}\sum_i \epsilon_i\dot{\hat{\Delta}}(\theta^*, X_i)\varphi^2(X_i)$, $I_{n,2} = n^{-1}\sum_i \epsilon_i(\dot{\hat{\Delta}}(\tilde{\theta}_n, X_i) - \dot{\hat{\Delta}}(\theta^*, X_i))\varphi^2(X_i)$, $I_{n,3} = n^{-1}\sum_i(\Delta(\tilde{\theta}_n, X_i) - \Delta(\theta^*, X_i))\dot{\hat{\Delta}}(\tilde{\theta}_n, X_i)\varphi^2(X_i)$, $I_{n,4} = n^{-1}\sum_i(\hat{\Delta}(\tilde{\theta}_n, X_i) - \hat{\Delta}(\theta^*, X_i))\dot{\hat{\Delta}}(\tilde{\theta}_n, X_i)\varphi(^2X_i)$, and $I_{n,5} = n^{-1}\sum_i(\hat{\Delta}(\theta^*, X_i) - \Delta(\theta^*, X_i))\dot{\hat{\Delta}}(\tilde{\theta}_n, X_i)\varphi^2(X_i)$.

For some given $\epsilon > 0$, let $\mathcal{K} = \{u : |u - x| \leq \epsilon \text{ and } x \in D\}$. By choosing ϵ sufficiently small,

$\mathcal{K}$ becomes a compact subset of S.

For any x in D, any $\theta \in \Theta$ and for n sufficiently large,

$$||\dot{\hat{\Delta}}(\theta, x)|| \leq \sum_j |w_j(x)|||\dot{m}(\theta, X_j)|| \\ \leq \sup_{(\theta,u)\in\Theta\times\mathcal{K}} ||\dot{m}(\theta, u)|| \sum_j |w_j(x)|.$$

So, by (10) and assumption (A1),

$$\sup_{(\theta,x)\in\Theta\times D} ||\dot{\hat{\Delta}}(\theta, x)|| = O_p(1). \tag{13}$$

By the mean value Theorem, for any x in D and for n sufficiently large, there exists a $\tilde{\theta}_{n,x} \in \Theta$ such that $|\Delta(\tilde{\theta}_n, x) - \Delta(\theta^*, x)| = |m(\tilde{\theta}_n, x) - m(\theta^*, x)| \leq ||\hat{\theta} - \theta^*||||\dot{m}(\tilde{\theta}_{n,x}, x)||$. So, by assumption (A1)

$$\sup_{x\in D} |\Delta(\tilde{\theta}_n, x) - \Delta(\theta^*, x)| = O_p(||\hat{\theta} - \theta^*||). \tag{14}$$

Using (13) we can also check that

$$\sup_{x\in D} |\hat{\Delta}(\tilde{\theta}_n, x) - \hat{\Delta}(\theta^*, x)| = O_p(||\hat{\theta} - \theta^*||). \tag{15}$$

Similarly one can prove that,

$$\sup_{x\in D} ||\dot{\hat{\Delta}}(\tilde{\theta}_n, x) - \dot{\hat{\Delta}}(\theta^*, x)|| = O_p(||\hat{\theta} - \theta^*||). \tag{16}$$

From the definition of $\hat{\Delta}(\theta, x)$, see (1), we have that $\hat{\Delta}(\theta^*, x) - \Delta(\theta^*, x) = (\hat{m}(x) - m(x)) - \sum_{j=1}^n w_j(x)(m(\theta^*, X_j) - m(\theta^*, x))$. By Theorem 6 in Masry (1996), $\sup_{x\in D} |\hat{m}(x) - m(x)| = o_p(1)$. On the other hand, $|\sum_j w_j(x)(m(\theta^*, X_j) - m(\theta^*, x))| \leq \sup_{|u-x|\leq h_n} ||m(\theta^*, u) - m(\theta^*, x)|| \sum_j |w_j(x)|$. So using (10) and assumption (A1), we have that

$$\sup_{x\in D} |\hat{\Delta}(\theta^*, x) - \Delta(\theta^*, x)| = o_p(1). \tag{17}$$

Similarly,

$$\sup_{x\in D}||\dot{\hat{\Delta}}(\theta^*,x)+\dot{m}(\theta^*,x)|=o_p(1). \tag{18}$$

Now we have all the ingredients needed to show that, $I_{n,l}=o_p(1)$, for, $l=1,\ldots,5$. In fact, using the LLN and (13), we get from (18), (16), (14), (15) and (17) that, respectively,

$$\begin{aligned}
I_{n,1} &= n^{-1}\sum_i \epsilon_i(\dot{\hat{\Delta}}(\theta^*,X_i)+\dot{m}(\theta^*,X_i))\varphi^2(X_i) - n^{-1}\sum_i \epsilon_i\dot{m}(\theta^*,X_i)\varphi^2(X_i)\\
&= o_p(1)-o_p(1)=o_p(1),\\
|I_{n,2}| &\le \sup_{x\in D}||\dot{\hat{\Delta}}(\tilde{\theta}_n,x)-\dot{\hat{\Delta}}(\theta^*,x)||n^{-1}\sum_i|\epsilon_i|\varphi^2(X_i)\\
&= O_p(||\hat{\theta}-\theta^*||)O_p(1)=o_p(1),\\
|I_{n,3}| &\le \sup_{x\in D}|\Delta(\tilde{\theta}_n,x)-\Delta(\theta^*,x)|\sup_{(\theta,x)\in\Theta\times D}||\dot{\hat{\Delta}}(\theta,x)||n^{-1}\sum_i\varphi^2(X_i)\\
&= O_p(||\hat{\theta}-\theta^*||)O_p(1)O_p(1)=o_p(1),\\
I_{n,4} &= o_p(1),\ \text{and}\ I_{n,5}=o_p(1).
\end{aligned}$$

So, we have established that $I_n=o_p(1)$, hence, by (12)

$$\begin{aligned}
\dot{T}_n(\tilde{\theta}_n) &= 2n^{-1}\sum_i \dot{Y}_i(\tilde{\theta}_n)\hat{\Delta}(\tilde{\theta}_n,X_i)\varphi^2(X_i)+o_p(1)\\
&= -2n^{-1}\sum_i \dot{m}(\tilde{\theta}_n,X_i)\hat{\Delta}(\tilde{\theta}_n,X_i)\varphi^2(X_i)+o_p(1)\\
&= -2n^{-1}\sum_i \dot{m}(\theta^*,X_i)\Delta(\theta^*,X_i)\varphi^2(X_i)+o_p(1) \qquad (19)\\
&= -2B+o_p(1),
\end{aligned}$$

where in (19) we have used (15), $\sup_{x\in D}||\dot{m}(\tilde{\theta}_n,x)-\dot{m}(\theta^*,x)||=O_p(||\tilde{\theta}_n-\theta^*)||)$ and the fact that both $\dot{m}(\theta^*,x)$ and $\Delta(\theta^*,x)$ are bounded in D. This together with (11) concludes the proof of Lemma 1. □

Proof of Lemma 2

Substituting $\hat{\Delta}(\theta,x)$ in (3) by $\hat{\Delta}(\theta,x)-\Delta(\theta,x)+\Delta(\theta,x)$, we obtain

$$\begin{aligned} T_n(\theta) =& n^{-1}\sum_{i=1}^{n}\left[2Y_i(\theta)\Delta(\theta,X_i)-\Delta^2(\theta,X_i)\right]\varphi^2(X_i) \\ &+2n^{-1}\sum_{i=1}^{n}\left[\hat{\Delta}(\theta,X_i)-\Delta(\theta,X_i)\right]\epsilon_i\varphi^2(X_i) \\ &-n^{-1}\sum_{i=1}^{n}\left[\hat{\Delta}(\theta,X_i)-\Delta(\theta,X_i)\right]^2\varphi^2(X_i). \end{aligned} \tag{20}$$

By Theorem 6 in Masry (1996), $\sup_{x\in D}|\hat{m}(x)-m(x)| = O_p((\ln n/(nh^d))^{1/2})+O_p(h^{p+1}) = o_p(n^{-1/4})$. Similar arguments as those used in the proof of that result lead to the property $\sup_{x\in D}|\sum_{j=1}^{n}w_j(x)m(\theta^*,X_j)-m(\theta^*,x)| = O_p(h^{p+1}) = o_p(n^{-1/4})$. So,

$$n^{-1}\sum_{i=1}^{n}\left[\hat{\Delta}(\theta,X_i)-\Delta(\theta,X_i)\right]^2\varphi^2(X_i) = o_p(n^{-1/2}).$$

It remains to show that $J_n(\theta) := n^{-1}\sum_{i=1}^{n}\left[\hat{\Delta}(\theta,X_i)-\Delta(\theta,X_i)\right]\epsilon_i\varphi^2(X_i) = o_p(n^{-1/2})$. To do so, we need the following Lemma.

LEMMA **3** *Under the conditions of Lemma 2, for any* $\theta\in\Theta$

$$\begin{aligned} \hat{\Delta}(\theta,x)-\Delta(\theta,x) =& d_{n,0}(x)n^{-1}\sum_{j=1}^{n}\gamma_h(X_j-x)K_h(X_j-x)\epsilon_j \\ &+h^{p+1}d_{n,1}(\theta,x)+d_{n,2}(\theta,x)+r_n(\theta,x), \end{aligned}$$

where $d_{n,0}(x) := e_{N,1}^T\mathbb{E}(\mathbf{S}_n^{-1}(x))$ *is a deterministic* $1\times N$ *vector that satisfies* $\sup_{x\in D}||d_{n,0}(x)|| = O(1)$, $d_{n,1}(\theta,x)$ *and* $d_{n,2}(\theta,x)$ *are two deterministic scalars that satisfy* $\sup_{x\in D}|d_{n,1}(\theta,x)| = O(1)$ *and* $\sup_{x\in D}|d_{n,2}(\theta,x)| = o(h^{p+1})$. *The remainder term* r_n *satisfies* $\sup_{x\in D}|r_n(\theta,x)| = O_p(\ln n/(nh^d))+O_p(h^{p+1}(\ln n/(nh^d))^{1/2}) = o_p(n^{1/2})$.

The proof of this Lemma is omitted since it follows directly from equation (2.13) in Masry (1996) through his Theorem 2, Theorem 4, Theorem 5, Proposition 1, Corollary 2 and Corollary 3.

From Lemma 3 we can decompose $J_n(\theta)$ as $J_{n,1} + J_{n,2}(\theta) + J_{n,3}(\theta) + J_{n,4}(\theta)$, with $J_{n,1} = n^{-2}\sum_{i=1}^n\sum_{j=1}^n d_{n,0}(X_i)\gamma_h(X_j-X_i)K_h(X_j-X_i)\epsilon_i\epsilon_j\varphi^2(X_i)$, $J_{n,2}(\theta) = h^{p+1}n^{-1}\sum_{i=1}^n d_{n,1}(\theta,X_i)\epsilon_i\varphi^2(X_i)$, $J_{n,3}(\theta) = n^{-1}\sum_{i=1}^n d_{n,2}(\theta,X_i)\epsilon_i\varphi^2(X_i)$ and $J_{n,4}(\theta) = n^{-1}\sum_{i=1}^n r_n(\theta,X_i)\epsilon_i\varphi^2(X_i)$. Clearly

$$\begin{aligned}|J_{n,4}(\theta)| &\leq sup_{x\in D}|r_n(\theta,x)|n^{-1}\sum_{i=1}^n|\epsilon_i|\varphi^2(X_i)\\ &= o_p(n^{-1/2})O_p(1) = o_p(n^{-1/2}).\end{aligned}$$

Now consider $J_{n,3}(\theta)$. By Lemma 3 and Lemma 7 in Doukhan and Louhichi (1999), using assumption (A6)

$$\begin{aligned}\mathbb{E}[J_{n,3}^2(\theta)] &\leq Cn^{-1}(\mathbb{E}|d_{n,2}(\theta,X_i)\epsilon_i\varphi^2(X_i)|^\nu)^{2/\nu}\sum_{t\geq 1}\alpha^{1-2/\nu}(t)\\ &\leq Cn^{-1}h^{2(p+1)}(\mathbb{E}|\epsilon_i\varphi^2(X_i)|^\nu)^{2/\nu}\sum_{t\geq 1}\alpha^{1-2/\nu}(t)\\ &\leq Cn^{-1}h^{2(p+1)}.\end{aligned}$$

We conclude that $J_{n,3}(\theta) = O_p(n^{-1/2}h^{(p+1)}) = o_p(n^{-1/2})$. Similarly, one can check that $J_{n,2}(\theta) = o_p(n^{-1/2})$. To get the desired result, it remains to prove that $J_{n,1} = o_p(n^{-1/2})$.
Let $J_{n,1}^{'} = n^{-2}\sum_{i=1}^n d_{n,0}(X_i)\gamma_h(0)K_h(0)\epsilon_i^2\varphi^2(X_i)$. Observe that, for n sufficiently large,

$$\begin{aligned}|J_{n,1}^{'}| &\leq n^{-1}h^{-d}K(0)||\gamma_h(0)||\sup_{x\in D}||d_{n,0}(X_i)||\left(n^{-1}\sum_{i=1}^n\epsilon_i^2\varphi^2(X_i)\right)\\ &\leq Cn^{-1}h^{-d}\left(n^{-1}\sum_{i=1}^n\epsilon_i^2\varphi^2(X_i)\right)\\ &= O_p(n^{-1}h^{-d}) = o_p(n^{-1/2}).\end{aligned}$$

So, we have that $J_{n,1} = J_{n,1}^{''} + o_p(n^{-1/2})$, with

$$J_{n,1}^{''} := n^{-2}\sum_{i=1}^n\sum_{j\neq i} d_{n,0}(X_i)\gamma_h(X_j-X_i)K_h(X_j-X_i)\epsilon_i\epsilon_j\varphi^2(X_i). \tag{21}$$

Put $\xi_i = (X_i,\epsilon_i)^T$ and $\eta(\xi_i,\xi_j) = d_{n,0}(X_i)\gamma_h(X_j-X_i)K_h(X_j-X_i)\epsilon_i\epsilon_j\varphi^2(X_i)$. Let $h(\xi_i,\xi_j) = \eta(\xi_i,\xi_j) + \eta(\xi_j,\xi_i)$. With these notations, we can write $J_{n,1}^{''}$ as $J_{n,1}^{''} = \sum_{1\leq i<j\leq n} h(\xi_i,\xi_j)$.

Observe that $h(\xi_i, \xi_j)$ is symmetric and $\mathbb{E}[h(\xi_i, v)] = 0$, for any v. So by Lemma C.2(ii) in Gao and King (2006), using assumption (A6)

$$\mathbb{E}[(J_{n,1}^{''})^2] = n^{-4}\mathbb{E}[(\sum_{1\leq i<j\leq n} h(\xi_i, \xi_j))^2] \leq Cn^{-2}M_n^{2/\nu},$$

with $M_n = \max_{1<i<j\leq n} \max\left\{E|h(\xi_i, \xi_j)|^\nu, \int |h(\xi_i, \xi_j)|^\nu dP(\xi_i)dP(\xi_j)\right\}$. Under our assumptions (iii) and (iv) given in Lemma 2, using the fact that, for n sufficiently large, $|d_{n,0}(X_i)\gamma_h(X_j - X_i)K_h(X_j - X_i)| \leq C/h^d$, one can easily verify that $M_n^{2/\nu} \leq Ch^{-2d}$. So, $J_{n,1}^{''} = O_p(n^{-1}h^{-d}) = o_p(n^{-1/2})$ which concludes the proof of Lemma 2. □

Proof of Theorem 2

Using Theorem 1, and the fact that $S_n^2(\hat{\theta}) - S^2(\theta^*) = (S_n^2(\hat{\theta}) - S^2(\theta^*)) + (S_n^2(\theta^*) - S^2(\theta^*)) = o_p(||\hat{\theta} - \theta^*||) + O_p(n^{-1/2})$, the desired result follows directly from the identity

$$\frac{\hat{a}}{\hat{b}} = \frac{a}{b} + \hat{b}\left[\hat{a} - a - (\hat{b} - b)\frac{a}{b}\right].$$

Details are omitted. □

References

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993). *Efficient and adaptive estimation for semiparametric models.* Johns Hopkins Series in the Mathematical Sciences. Baltimore, MD: Johns Hopkins University Press.

Carrasco, M. and X. Chen (2002). Mixing and moment properties of various GARCH and stochastic volatility models. *Econometric Theory 18*, 17–39.

Dette, H. and A. Munk (1998). Validation of linear regression models. *Ann. Statist. 26*(2), 778–800.

Dette, H. and A. Munk (2003). Some methodological aspects of validation of models in nonparametric regression. *Statist. Neerlandica 57*(2), 207–244.

Doksum, K. and A. Samarov (1995). Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *Ann. Statist. 23*(5), 1443–1473.

Domowitz, I. and H. White (1982). Misspecified models with dependent observations. *J. Econometrics 20*(1), 35–58.

Doukhan, P. and S. Louhichi (1999). A new weak dependence condition and applications to moment inequalities. *Stochastic Process. Appl. 84*(2), 313–342.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and its Applications*, Volume 66 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall.

Fan, J. and Q. Yao (2003). *Nonlinear Time Series*. Springer Series in Statistics. Springer-Verlag. Nonparametric and Parametric Methods.

Fan, Y. and Q. Li (1996). Consistent model specification tests: omitted variables and semiparametric functional forms. *Econometrica 64*(4), 865–890.

Gao, J. and M. King (2006). Estimation and model specification testing in nonparametric and semiparametric econometric models. Mpra paper, University Library of Munich, Germany.

Gu, J., D. Li, and D. Liu (2007). Bootstrap non-parametric significance test. *J. Nonparametr. Stat. 19*(6-8), 215–230.

Hall, P. (1992). On bootstrap confidence intervals in nonparametric regression. *Ann. Statist. 20*(2), 695–711.

Härdle, W. and E. Mammen (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist. 21*(4), 1926–1947.

Hayfield, T. and J. S. Racine (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software 27*(5).

Hodges, J. L. and E. L. Lehmann (1954). Testing the approximative validity of statistical hypotheses. *Journal of the Royal Statistical Society, Series B 16*, 261–268.

Hong, Y. and H. White (1995). Consistent specification testing via nonparametric series regression. *Econometrica 63*(5), 1133–1159.

Jun, S. J. and J. Pinkse (2009). Semiparametric tests of conditional moment restrictions under weak or partial identification. *J. Econometrics 152*(1), 3–18.

Künsch, H. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist. 17*, 1217–1241.

Lahiri, S. N. (2003). *Resampling methods for dependent data*. Springer Series in Statistics. New York: Springer-Verlag.

Lavergne, P. (1998). Selection of regressors in econometrics: parametric and nonparametric methods selection of regressors in econometrics. *Econometric Reviews 17*(3), 227–273.

Lavergne, P. and Q. H. Vuong (1996). Nonparametric selection of regressors: the nonnested case. *Econometrica 64*(1), 207–219.

Li, Q. and J. Racine (2004). Cross-validated local linear nonparametric regression. *Statist. Sinica 14*(2), 485–512.

Li, Q. and S. Wang (1998). A simple consistent bootstrap test for a parametric regression function. *J. Econometrics 87*(1), 145–165.

Liu, R. and K. Singh (1992). Moving blocks jackknife and bootstrap capture weak dependence. In *Exploring the Limits of Bootstrap (East Lansing, MI, 1990)*, pp. 225–248. Wiley.

Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *J. Time Ser. Anal. 17*(6), 571–599.

Patton, A., D. N. Politis, and H. White (2009). Correction to "Automatic block-length selection for the dependent bootstrap" by D. Politis and H. White [mr2041534]. *Econometric Rev. 28*(4), 372–375.

Taylor, L. W. (2009). Penalized-r^2 criteria for model selection. *Manchester School 77*(6), 699–717.

Xia, Y. and W. K. Li (2002). Asymptotic behavior of bandwidth selected by the cross-validation method for local polynomial fitting. *J. Multivariate Anal. 83*(2), 265–287.

Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *J. Econometrics 75*(2), 263–289.