Scandinavian Journal of Statistics Theory and Applications

doi: 10.1111/j.1467-9469.2012.00821.x

Measuring the Discrepancy of a Parametric Model via Local Polynomial Smoothing

ANOUAR EL GHOUCH

Université catholique de Louvain

MARC G. GENTON

Texas A&M University

TAOUFIK BOUEZMARNI

Université de Sherbrooke

ABSTRACT. In the context of multivariate mean regression, we propose a new method to measure and estimate the inadequacy of a given parametric model. The measure is basically the missed fraction of variation after adjusting the best possible parametric model from a given family. The proposed approach is based on the minimum L^2-distance between the true but unknown regression curve and a given model. The estimation method is based on local polynomial averaging of residuals with a polynomial degree that increases with the dimension d of the covariate. For any $d \geq 1$ and under some weak assumptions we give a Bahadur-type representation of the estimator from which $\sqrt{n}$-consistency and asymptotic normality are derived for strongly mixing variables. We report the outcomes of a simulation study that aims at checking the finite sample properties of these techniques. We present the analysis of a dataset on ultrasonic calibration for illustration.

Key words: explanatory power, inadequacy index, model misspecification, multivariate local polynomial smoothing, strong mixing sequence, validation test

1. Introduction

In the context of regression with high dimensional predictors, it is difficult to get an efficient non-parametric estimator for the true regression function because of the sparsity of the data. For that reason and for the purpose of interpretability, simple parametric models with few covariates are usually preferred to a purely non-parametric fit. However, the selection of an appropriate parametric function to fit and make inference about the data is a challenging problem in any real data analysis. 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 or Bayesian information criteria and the use of test statistics such as Wald or likelihood ratio tests. In the first case, the selection is done by choosing the model with the smallest criterion (error) among competing models 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. For an excellent review and a detailed discussion of model 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 non-parametric techniques that allow for great flexibility. Methods such as kernel smoothing and splines have become widely used to justify

a parametric restriction. The literature related to this subject is very vast and includes the work of Cristobal Cristobal *et al.* (1987), Härdle & Mammen (1993), Hong & White (1995), Zheng (1996), Li & Wang (1998), Delgado & González Manteiga (2001), Zhang & Dette (2004) and Jun & Pinkse (2009) for consistently testing a parametric regression functional form.

These 'classical' non-parametric methods focus on the behaviour 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 and regardless of whether the given parametric model is correct or not our purpose is to construct an 'inadequacy index' based on a distance between the given parametric family and the unknown target function. This index serves as a kind of inverse coefficient of determination: it takes values in $[0, 1]$ and when its value is close to 0 the parametric fit becomes better. We propose a consistent estimator of this 'inadequacy index' and study its asymptotic properties. 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 (strictly stationary) 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. As a by-product, using these results we develop a new validation test. The main difficulty here is degeneracy of the asymptotic distribution under correct specification. To bypass this problem without sacrificing the power and the rate of convergence we adopt the concept of neighbourhood hypotheses; see Dette & 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.

The rest of the paper is organized as follows. In section 2, we introduce our inadequacy index in terms of an L^2-distance between the mean regression function and a parametric model. The estimation procedure is described in section 3. Section 4 is devoted to the asymptotic properties of the proposed estimator. In section 5, we show how the inadequacy index can be used to validate a given model via neighbourhood hypotheses testing. The performance of the proposed method is examined in section 6 via a Monte Carlo simulation study. A real data analysis on ultrasonic calibration is carried out in section 7. The proofs of the asymptotic results are collected in appendix S1 that is provided as Supporting Information on the journal website.

2. Closeness of parametric approximation: the inadequacy index

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 $\mathcal{M} = \{m(\theta, x), \theta \in \Theta\}$.

Following an original idea of Doksum & 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

$$\eta^2 = \frac{\mathrm{Var}[m(X)]}{\mathrm{Var}(Y)} = 1 - \frac{\mathrm{Var}(\epsilon)}{\mathrm{Var}(Y)}.$$

This coefficient gives the fraction of variability of Y explained by X through the true mean regression function $m(x)$. It is a direct consequence of the ANOVA decomposition $E(Y - g(X))^2 = E(m(X) - g(X))^2 + E(\epsilon^2)$, where g is any real-valued function with $E(g^2(X)) < \infty$. Now, for a given θ, letting $g(x) = m(\theta, x)$ we get $E(Y - m(\theta, X))^2 = E(m(X) - m(\theta, X))^2 + \mathrm{Var}(\epsilon)$. This is a decomposition of the parametric residual variation into unexplained variation due to model misspecification and error variation. Therefore, the coefficient

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

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 of $m(\theta, x)$. Note that the smallest the value of $\zeta^2(\theta)$ the best the model $m(\theta, x)$ is. The case $\zeta^2(\theta) = 0$ is equivalent to $m(\theta, X) = m(X)$ with probability 1. If $\mathcal{M}$ includes constants (which should always be the case) then $\zeta^2(\theta) \leq \eta^2 \leq 1$. The equality $\zeta^2(\theta) = \eta^2$ occurs if and only if $m(\theta, X) = E(Y)$ with probability 1. This is the case when the parametric model fails to capture any variability in the data.

Let θ^* be the pseudo-true parameter, i.e. $m(\theta^*, x)$ is the best approximation to the true regression function m that can be found within the parametric family $\mathcal{M} = \{m(\theta, x), \theta \in \Theta\}$. We define our parameter of interest to be

$$\zeta^2 \equiv \zeta^2_\varphi(\theta^*) = \frac{E[(m(X) - m(\theta^*, X))\varphi(X)]^2}{E[(Y - m(\theta^*, X))\varphi(X)]^2} = 1 - \frac{E[(Y - m(X))\varphi(X)]^2}{E[(Y - m(\theta^*, X))\varphi(X)]^2},$$

where we introduce the known weight function φ in order to get a more flexible and 'robust' measure. In fact, φ will typically be one in the central part of the support of X and zero near the boundary and so the non-parametric estimator of ζ^2 will be less sensitive to the boundary points, an inherent problem in kernel regression. The weight allows also the user to focus the analysis on a given range of the covariates. For example, one can measure how good is a given parametric model for small values of X compared to large values. A more clever but also more technical approach could be the use of a data driven weight as for example $\varphi(x) = I(\hat{f}(x) > a)$, where $\hat{f}$ is the kernel density estimator of X and a is a given small positive constant.

The index ζ^2 is the missed fraction of variation after adjusting the best possible parametric model from the family $\mathcal{M}$. In other words ζ^2 is the inadequacy index of the family $\mathcal{M}$. To illustrate the usefulness of the index ζ^2 and to exemplify its interpretation, we consider the following regression model

$$Y = 6 + \beta_1 X_1 + \beta_2 X_2 + 0 X_3 + \lambda(2X_3 + \sin|3\pi X_3 + \pi|) + \tau\epsilon.$$

To generate the data we use the same procedure as in our simulation study; see section 6. We calculate ζ^2 for the eight possible simple linear models (including the intercept only model) with $\{X_1, X_2, X_3\}$ as covariates. It is important to note that these linear models are note necessarily nested and that they involve different number of parameters. Observe also that when $\lambda = 0$, the true regression function reduces to $6 + \beta_1 X_1 + \beta_2 X_2$ and when $\lambda \neq 0$, no linear model can fit the data correctly. It is also interesting to note that ζ^2 coincides with the explanatory power index η^2 for the intercept only model. In other words, η^2 can be seen as a special case of ζ^2. By varying any of the parameters β_1, β_2, λ and τ, the data generating procedure changes and so does the 'reference' index η^2. For easy comparison, we report in Table 1 the values of ζ^2/η^2 under different situations. The latter values vary in $[0, 1]$ and represent the relative index of inadequacy.

Table 1. $100\times$ *the relative index of inadequacy* ζ^2/η^2

β_1	β_2	λ	τ	η^2	X_1	X_2	X_3	(X_1, X_2)	(X_1, X_3)	(X_2, X_3)	(X_1, X_2, X_3)
0.5	1.5	0	0.5	45.44	94.28	16.92	100.0	34.59×10^{-6}	94.28	16.92	36.14×10^{-6}
2	2	0	0.5	72.73	78.57	78.57	100.0	58.06×10^{-6}	78.57	78.57	60.88×10^{-6}
2	2	0	1	40.00	62.49	62.49	100.0	98.52×10^{-6}	62.49	62.49	115.4×10^{-6}
2	2	1	1	59.25	89.20	89.20	89.20	74.38	74.38	74.38	52.76

When $\lambda=0$, the linear model with only X_3 as a covariate is equivalent to an intercept only model, i.e. we lose 100% of the explanatory power of (X_1, X_2, X_3). In this case, adding X_3 to any linear model involving $\{X_1, X_2\}$ does not decrease the value of ζ^2. This is an important fact as it means that, at least theoretically, our index does not suffer from the well known overfitting or over parametrization problem. This is due to the fact that, in the definition of ζ^2, the parametric model affects both the nominator and the denominator, hence, the penalization for the number of parameters is somewhat automatic. In the first case ($\beta_2 = 3\times\beta_1 = 1.5$), the linear model with only X_2 as a covariate fits the data very well, since the loss in the explanatory power is less than 17%. This percentage reduces to almost 0% when we incorporate X_1 in the model. The situation becomes completely different in the last case (see the last row in Table 1) where any linear model leads to a loss in the explanatory power of more than 50%. Among all the candidate models, the best one, in this case, is clearly the one with all covariates.

We now give a formal justification for our claim that the proposed index is not affected by over-parametrization. Let $\hat{\theta}$ be the least square estimator of θ and $\hat{m}$ be the local constant estimator of m as defined in section 3.2. From Dette (1999), see also Biedermann & Dette (2000), under some regularity conditions, we know that, for a sufficiently large sample size n,

$$E\left[n^{-1}\sum_i(\hat{m}(X_i)-m(\hat{\theta},X_i))^2\right]\approx \mathrm{Var}(Y)(nh)^{-1}\int K^2(x)\,\mathrm{d}x+E(m(X)-\mathcal{P}_q m(X))^2,$$

and

$$E\left[n^{-1}\sum_i(Y_i-m(\hat{\theta},X_i))^2\right]\approx n^{-1}(n-q)[\mathrm{Var}(Y)+E(m(X)-m(\theta^*,X))^2],$$

where h is the bandwidth, K is the kernel and $\mathcal{P}_q m(X)$ is the orthogonal projection of $m(X)$ onto the subspace spanned by $\nabla_\theta m(X,\theta^*)$ with respect to the inner product $\langle m_1, m_2\rangle = E[m_1(X)m_2(X)]$. It follows that, if the parametric model is correct, then the ratio of the expected estimated quantities given above is approximatively

$$\frac{h^{-1}\int K^2(x)\,\mathrm{d}x}{n-q}.$$

So, increasing the number of parameters q in the parametric model should cause an increase in the value of the estimated index.

3. Estimation procedure

3.1. Problem setting

For a given $\theta\in\Theta$, put $\Delta(\theta,x)=m(x)-m(\theta,x)$ and $Y(\theta)=Y-m(\theta,X)$. Clearly, to estimate ζ^2 we need an estimator for both $T(\theta^*)\equiv T_\varphi(\theta^*) := E[\Delta^2(\theta^*,X)\varphi^2(X)]$ and $S(\theta^*)\equiv S_\varphi(\theta^*) := E[Y^2(\theta^*)\varphi^2(X)]$. We first propose an estimator for $T(\theta^*)$ and study its asymptotic

properties. Before that, in the following two subsections, we introduce some key assumptions about the data and the kernel smoothing procedure.

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 $\mathcal{F}_I^L$ $(-\infty \le I, L \le \infty)$ denote the σ-field generated by the family $\{(X_t, Y_t), I \le t \le L\}$. The stochastic process $\{(X_t, Y_t)\}$ 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 & Yao (2003) and Carrasco & 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.

3.2. The local polynomial smoother

We now explain the kernel smoothing procedure that will be used in the estimation of $T(\theta)$. Let K denote a non-negative kernel function defined on $\mathbb{R}^d$, $0<h_n\equiv h\to 0$ be a bandwidth parameter and $K_h(x)=h^{-d}K(x/h)$. By definition, $\Delta(\theta,x)=\mathbb{E}[Y(\theta)\,|\,X=x]$. If 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 non-parametric 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 non-parametric estimators of a regression function; see, for example, Fan & Gijbels (1996) for more details. In particular, the multivariate local polynomial estimator of order p, $p\in\mathbb{N}$, 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 $p\ge 0$ and $d\ge 1$, can be found in appendix S1.

The estimator given by (1) is only available when θ is known, which is not the case here. Let $\hat{\theta}$ be any consistent estimator of θ^*, i.e. $\hat{\theta}=\theta^*+o_p(1)$; details for the parametric estimation procedure will be given later. We now have a feasible estimator of $\Delta(\theta,x)$: $\hat{\Delta}(\hat{\theta},x)=\sum_{j=1}^n w_j(x)\times Y_j(\hat{\theta})=\hat{m}(x)-\hat{m}(\hat{\theta},x)$, where $\hat{m}(x)=\sum_{j=1}^n w_j(x)\,Y_j$ is the standard (non-parametric) 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 non-parametric fit $\hat{m}(x)$.

3.3. Our estimator of $T(\theta^*)$

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

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

This quantity is related to a discrete (Riemann sum) version of a test statistic that was proposed by Härdle & Mammen (1993) in the context of goodness-of-fit tests. In the present work, the primary objective is not about testing, but about constructing an estimator of ζ^2 with some 'good' properties. To this end, we consider here another estimator of $T(\theta^*)$ given by $T_n(\hat{\theta})$, with

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

It is straightforward to show that this estimator is simply the empirical version of the expression $T(\theta) = E[(2\,Y(\theta) - \Delta(\theta, X))\Delta(\theta, X)\varphi^2(X)]$. Later, see remark 1 and section 6, 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 be 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 & 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).

4. Asymptotic properties

4.1. Assumptions

Before starting with the study of the asymptotic properties of $T_n(\hat{\theta})$ we need first to introduce some notations and give a set of sufficient regularity conditions needed to obtain the results. For a d-tuple $k = (k_1, \ldots, k_d)^T \in \mathbb{N}^d$ and a d-vector $x = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$, 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 $\dot{m} := \partial m/\partial\theta$ and $\ddot{m} := \partial^2 m/\partial\theta\theta^T$ 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

$$\bar{a} > \max\left(\frac{2v}{v-2}, \frac{\delta(7+2d)-4}{\delta(1-ad)-2}\right)$$

for some $v \in (2, \delta]$.

(A7) The kernel K is a bounded non-negative function with compact support, say $[-1, 1]^{\otimes d}$ and for every k with $0 \leq |k| \leq 2p$ (p is the highest order in the local polynomial approximation) 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 assumption (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 assumption (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).

4.2. Main results

Our first result is formulated in the following lemma.

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

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

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

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 $\hat{B} = n^{-1}\sum_{i=1}^n \times \dot{m}(\hat{\theta}, X)Y(\hat{\theta})\varphi^2(X)$.

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

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

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

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 these conditions to hold, we need that $p > d/2 - 1$. In other words, to ensure the optimal root-n convergence rate, the order of the local polynomial approximation should increase as the dimension d of the covariates X increases.
- Although our estimator converges to the population parameter as the root-n convergence rate due to the average done across the samples, it does not mean that such an estimator is completely free from the curse-of-dimensionality. The inaccuracy of the first-step estimation of the unknown curve of high dimension will be passed on to the second-step estimator as it is illustrated in the simulation study; see section 4. As any non-parametric kernel based method, our estimator is sensitive to the choice of

bandwidth. In the simulation study we select the bandwidth via the cross-validation method; see Xia and Li (2002) and Li and Racine (2004). When the objective of the study is to compare two or many parametric models based on the same set of covariates, one should use the same bandwidth to calculate ζ^2 for every model, otherwise the comparison may be unfair.

- All the bandwidth restrictions given in assumption (A6), (i) and (ii) are fulfilled whenever the assumption (i') given below is satisfied:

  (i') $\delta \geq 4$, $p > d/2 - 1$ and $h_n \sim (n^{-1} \ln n)^a$ for some $\dfrac{1}{4(p+1)} < a < \dfrac{2}{d}$

- From the proofs given in appendix S1 it is 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,

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

  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 $\Delta_n(x)$.

It is also important to note that, until now, 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

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

In this definition we used the weighted version of the least squares estimator, since, as motivated in section 4.3, we are interested in assessing the quality of the parametric $m(\theta, x)$ within the support of $\varphi(x)$. From corollary 3.1 in Domowitz & White (1982) we claim that, under assumptions (A), $\hat{\theta}$ converges with probability 1 to

$$\theta^* = \arg\min_{\theta \in \Theta} E[(m(X) - m(\theta, X))\varphi(X)]^2 = \arg\min_{\theta \in \Theta} T(\theta). \tag{5}$$

Observe that $T(\theta^*)$ coincides with $\min_{\theta \in \Theta} T(\theta)$, 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

$$-2B^T = E[-2\dot{m}^T(\theta^*, X)\Delta(\theta^*, X)\varphi^2(X)] = E\left[\frac{\partial \Delta^2(\theta^*, X)}{\partial \theta}\varphi^2(X)\right] = \frac{\mathrm{d}T(\theta^*)}{\mathrm{d}\theta} = 0.$$

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*

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

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

As $\zeta^2 = T(\theta^*)/S(\theta^*)$, an obvious estimator of this index is provided by $\hat{\zeta}^2 := T_n(\hat{\theta})/S_n(\hat{\theta})$, where $T_n(\theta)$ is given by (3), $\hat{\theta}$ is given by (4) and $S_n(\theta) = n^{-1} \sum_{i=1}^{n} (Y_i - m(\theta, X_i))^2 \varphi^2(X_i)$.

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* $\xi_{i,\varphi}(\theta^*)$, $\xi_{i,\varphi}(\theta) = [(1-\zeta^2)Y^2(\theta) - \epsilon_i^2]\varphi^2(X_i)/S(\theta)$ *with* $Y_i(\theta) = Y_i - m(\theta, X_i)$, *and* $\epsilon_i = Y_i - m(X_i)$.

4.3. Asymptotic normality and variance

A direct consequence of theorem 2 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 & 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 $\sigma^2 \equiv \sigma_\varphi^2(\theta^*)$, with $\sigma_\varphi^2(\theta) := \lim_{n\to\infty} n^{-1}\mathrm{Var}(\sum_{t=1}^{n} \xi_t(\theta)) = \mathrm{Var}[\xi_1(\theta)] + 2\sum_{t>1}\mathrm{Cov}(\xi_1(\theta), \xi_t(\theta))$.

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 & 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' $\{\xi_i\}_{1\le i\le n}$ into $N := l+1$ 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 $l \to \infty$ and $l = o(n)$. Let $U_i = l^{-1}\sum_{j=i}^{i+l-1} \xi_j$ be the sample mean of the ith 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 $\hat{\sigma}^2 = lN^{-1}\sum_{i=1}^{N}(U_i - \bar{U})^2$. 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 parameters ζ^2, θ^*, m and S. To overcome this problem, we simply suggest to plugging-in $\hat{\zeta}^2$ $\hat{\theta}$, $\hat{m}$ and S_n, as defined above, into the definition of $\hat{\sigma}^2$ to get $\hat{\sigma}_n^2$ as our feasible estimator of the asymptotic variance.

5. Validation of a parametric model

A direct application of the previous results is that one can construct an asymptotically valid Wald-type confidence interval for ζ^2 that is given by $\hat{\zeta}^2 \pm \frac{\hat{\sigma}_n}{\sqrt{n}} z_{1-\alpha/2}$, where z_α is the α-quantile of the standard normal distribution. Although this confidence interval gives us valuable information about the quality of the parametric approximation, we still need a formal approach to test the goodness-of-fit hypothesis: $H_0 : m \in \mathcal{M}$ versus $H_1 : m \notin \mathcal{M}$. In terms of ζ^2, 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 asymptotic results in such a form have been noted

before by many authors; see for example Fan & Li (1996, 1999). 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-statistic defined in appendix S1. 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 neighbourhood hypothesis or 'precise' hypothesis; see Hodges & Lehmann (1954). The drawbacks of (6) over (7) were largely documented by many authors; see for example Dette & Munk (1998) and the references given therein. To cite just an argument in favour of the concept of neighbourhood testing, observe that (7) is designed to provide evidence in favour 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 neighbourhood hypothesis we refer to Dette & Munk (2003).

As noted by those authors, the main difficulty with neighbourhood 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 approach adopted, in this work that consists in studying the estimated distance $T(\theta^*)$ without restrictions on the model specification allows us to easily overcome this difficulty. In fact, by theorem 2, we directly conclude that a critical region for $H_{\pi,0}$ is provided by $\hat{\zeta}^2 < \pi + z_\alpha \frac{\hat{\sigma}}{\sqrt{n}}$. 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^2 is given by $[0, \zeta^2_{n,+}]$, with $\zeta^2_{n,+} = \hat{\zeta}^2 + \frac{\hat{\sigma}}{\sqrt{n}} z_{1-\alpha}$. So one can state that, at risk $\alpha \times 100\%$, 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.

6. 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 $\hat{\zeta}^2$ and its asymptotic properties as stated in the previous sections. The open source software R was used in this study; see R Development Core Team (2012). The simulation considers univariate and multivariate cases with both i.i.d. data and weakly dependent data using the weight function $\varphi(t) = I(0 \leq t \leq 1)$ and $N := 2000$ replications. We generate $n := 2000$ 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

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

We are interested in measuring and testing the quality 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 $\zeta^2 \nearrow 1$.

In the two-dimensional case, we choose

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

and for $d = 3$,

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

The covariates are independent of each other, independent of the error variable ϵ_t and are $Unif[-\varepsilon, 1+\varepsilon]$. We use the local linear smoother with the Epanechnikov kernel function. As a data-driven bandwidth ($\hat{h}$) selection criterion, we use the likelihood cross-validation method; see Xia & Li (2002) and also Li & 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: $\hat{\zeta}_0^2 = T_n^0(\hat{\theta})/S_n(\hat{\theta})$ and $\hat{\zeta}^2 = T_n(\hat{\theta})/S_n(\hat{\theta})$; see (2) and (3). In Table 2 we report the empirical root mean squared error (RMSE $= \sqrt{\text{MSE}}$) based on 2000 replications using the data-driven bandwidth $\hat{h}$. We also report RMSE*, the empirical root mean squared error based on a (fixed) optimal bandwidth, i.e. the one that minimize RMSE over the grid $0.01, 0.02 \ldots, 0.99$. Table 2 shows the values of d, τ, λ used to generate the data together with the corresponding values of ζ^2 in percentage. For this latter, only the case $d = 1$ is shown since the other values are somewhat similar. From Table 2, we observe that both $\hat{\zeta}^2$ and $\hat{\zeta}_0^2$ perform very well with respect to the MSE criterion, with a clear advantage of $\hat{\zeta}^2$ over $\hat{\zeta}_0^2$. In fact, we obtain almost systematically a smaller MSE when we use our estimator $T_n(\hat{\theta})$ instead of the naive one, $T_n^0(\hat{\theta})$. 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 these estimators are mainly affected by the true values of the parameter ζ^2 and by the dimensionality d, along with interaction between these two factors. For example, when $d = 1$ or 2, as ζ^2 increases the MSE of $\hat{\zeta}^2$ initially increases and then rapidly decreases. $\hat{\zeta}_0^2$ behaves similarly, but its MSE increases rapidly and then decreases slowly. As d increases, we can see that $\hat{\zeta}_0^2$ behaves more and more badly, compared to $\hat{\zeta}^2$, especially when ζ^2 becomes large. Both the absolute value of the bias and the variance increase with d.

Table 2. $100 \times RMSE^*$ *and* $100 \times RMSE$ *for* $\hat{\zeta}^2$ *and* $\hat{\zeta}_0^2$. $p = 1$, *i.e. local linear approximation*

			$100 \times \text{RMSE}^*$						$100 \times \text{RMSE}$					
$d=1$			$d=1$		$d=2$		$d=3$		$d=1$		$d=2$		$d=3$	
λ	τ	ζ^2%	$\hat{\zeta}_0^2$	$\hat{\zeta}^2$	$\hat{\zeta}_0^2$	$\hat{\zeta}^2$	$\hat{\zeta}_0^2$	$\hat{\zeta}^2$	$\hat{\zeta}_0^2$	$\hat{\zeta}^2$	$\hat{\zeta}_0^2$	$\hat{\zeta}^2$	$\hat{\zeta}_0^2$	$\hat{\zeta}^2$
0	0.5	0.0	0.30	0.28	0.36	0.30	0.42	0.31	1.04	1.70	1.92	3.23	2.29	4.26
	1	0.0	0.30	0.28	0.36	0.30	0.42	0.31	1.04	1.70	1.87	3.14	2.35	4.36
	2	0.0	0.30	0.28	0.36	0.30	0.42	0.31	1.04	1.70	1.94	3.24	2.31	4.27
0.8	0.5	53.8	4.62	4.48	5.68	4.07	17.05	4.81	9.65	5.15	19.35	5.85	30.50	8.02
	1	22.6	4.66	4.97	3.89	4.36	3.98	4.03	7.29	6.30	10.62	7.79	15.26	10.02
	2	6.8	2.71	3.05	2.03	2.59	1.96	2.38	4.09	5.38	4.66	6.94	5.71	7.25
1.5	0.5	80.4	2.45	2.28	13.58	2.14	35.21	5.01	8.30	2.57	21.59	2.71	38.67	7.86
	1	50.6	4.72	4.74	5.11	4.15	16.57	4.93	9.57	5.49	18.70	6.21	29.15	8.19
	2	20.4	4.61	4.88	3.63	4.27	3.76	4.00	6.98	6.34	9.84	7.82	14.16	10.28
2.5	0.5	91.9	1.32	0.97	17.58	1.37	42.25	9.57	6.39	1.13	19.36	1.07	39.61	8.19
	1	74.0	2.86	2.92	11.45	2.54	31.27	3.10	8.95	3.32	21.72	3.54	37.28	7.77
	2	41.6	4.90	5.11	4.81	4.53	6.60	4.24	9.22	6.05	16.56	7.04	24.96	8.41

Globally, the variance is the main component of the mean squared error (the results are not displayed here for the sake of brevity). Typically, its contribution decreases with λ and increases with τ and it is more sensitive to the alteration in τ. As expected, increasing the covariate dimensionality d causes the MSE to increase, but $\hat{\zeta}_0^2$ is 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 9.3 whereas the MSE of $\hat{\zeta}^2$ increases by only a factor of 2.2. This becomes even more striking when we consider the optimal bandwidths. To give just an example, under the same scenario as above, the MSE* of $\hat{\zeta}_0^2$ increases by a factor of 12.3 whereas the MSE* of $\hat{\zeta}^2$ increases only by a factor of 1.08. This definitely demonstrates the advantages of the proposed estimator. Regarding the usefulness of the bandwidth selection procedure, we have, globally, observed that the results obtained using $\hat{h}$ were quite close to those obtained using the 'optimal' bandwidth. However, the dimensionality has again a clear negative impact. We have also observed that the loss of efficiency due to the estimated bandwidth is larger for $\hat{\zeta}_0^2$. In fact, the average (maximum) value of $|\text{MSE} - \text{MSE}^*|$ is 0.002 (0.009) and 0.015 (0.064) for $\hat{\zeta}^2$ and $\hat{\zeta}_0^2$, respectively. This indicates a greater robustness of $\hat{\zeta}^2$ to bandwidth misspecification.

Table 3 gives the relative efficiency of the local linear to the local constant approximation, i.e. $\text{MSE}(\hat{\zeta}_{p=0}^2)/\text{MSE}(\hat{\zeta}_{p=1}^2)$, where $\hat{\zeta}_{p=1}^2 \equiv \hat{\zeta}^2$ and $\hat{\zeta}_{p=0}^2$ are the estimators of ζ^2 using the local linear and local constant approximation, respectively. For $d = 1$, the two approximations give similar results. However, as d increases, the local linear estimator becomes more and more efficient. This is in agreement with the requirement $p > d/2 - 1$; see remark 1 in section 4.1.

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 given in section 4.3. The small mean squared errors, see Table 4, demonstrate the consistent nature of the proposed method. Globally, the MSE performance is very satisfactory and better than expected.

To complete the picture, Table 5 shows the coverage probability for the upper confidence intervals for ζ^2 at nominal level 95% computed using $\hat{\zeta}^2$ and its estimated asymptotic variance. The accuracy of confidence limits was assessed by calculating the proportion of times the true value was below the confidence limit. Globally, the empirical coverage probability was often different from the expected values especially for $\tau = 2$. This is because the data become too

Table 3. *Relative efficiency of the local linear to the local constant approximation*

λ	0			0.8			1.5			2.5		
τ	0.5	1	2	0.5	1	2	0.5	1	2	0.5	1	2
$d = 1$	0.87	0.87	0.87	1.17	1.28	1.43	1.26	1.17	1.19	1.13	1.17	1.10
$d = 2$	6.09	11.03	7.35	3.94	3.28	3.58	2.22	3.86	3.40	2.72	2.29	3.43
$d = 3$	34.53	35.00	43.54	3.50	5.25	13.87	1.59	4.09	5.72	0.82	2.11	4.97

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

λ	0			0.8			1.5			2.5		
τ	0.5	1	2	0.5	1	2	0.5	1	2	0.5	1	2
$d = 1$	3.86	3.86	3.86	12.43	18.84	13.66	3.31	13.60	18.74	0.65	5.34	16.28
$d = 2$	7.12	6.97	7.10	19.97	29.56	16.82	4.21	22.02	28.71	0.65	7.32	27.03
$d = 3$	8.43	8.47	8.49	18.98	35.35	20.83	3.06	21.54	35.18	2.18	4.77	28.27

Table 5. *The empirical coverage probability for the upper confidence intervals for* ζ^2 *using data driven bandwidth* ($\hat{h}$) *and using the optimal bandwidth* (h_o). *Nominal coverage* = 95%

	λ	0			0.8			1.5			2.5		
	τ	0.5	1	2	0.5	1	2	0.5	1	2	0.5	1	2
$d=1$	$\hat{h}$	1.00	1.00	1.00	0.96	0.93	0.83	0.98	0.96	0.93	0.99	0.98	0.96
	h_o	1.00	1.00	1.00	0.94	0.95	0.95	0.95	0.94	0.96	0.96	0.96	0.95
$d=2$	$\hat{h}$	1.00	1.00	1.00	0.95	0.89	0.69	0.96	0.94	0.88	0.95	0.95	0.93
	h_o	1.00	1.00	1.00	0.97	0.95	0.95	0.99	0.97	0.96	0.97	0.95	0.96
$d=3$	h_o	0.98	0.98	0.98	0.98	0.96	0.96	1.00	0.93	0.94	0.98	0.99	0.94

Table 6. $100 \times RMSE^{(*)}$ *of* $\hat{\zeta}^2$ *and the empirical coverage probability using data driven bandwidth* ($\hat{h}$) *and using the optimal bandwidth* (h_o). *Nominal coverage* = 95%, $\rho = 0.9$ *and* $d = 1$

λ	τ	ζ^2	$100\times$RMSE*	$100\times$RMSE	Emp. Cov. h_o	Emp. Cov. $\hat{h}$
0	0.5	0.0	0.28	1.78	1.00	1.00
	1	0.0	0.28	1.78	1.00	1.00
2.5	0.5	52.6	8.73	12.92	0.95	0.95
	1	21.7	6.86	12.33	0.95	0.94

noisy in such a case. For $\tau = 0.5$ or for $\lambda = 0$, our intervals appear to be too conservative, but as d increases they become anti-conservative. For $d = 3$ the results were unsatisfactory (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 5 gives the optimal coverage probability obtained using a fixed but optimal bandwidth (the one that minimizes the coverage error). 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 $\sqrt{n}(\hat{\zeta}^2 - \zeta^2)$ 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 4.3), we use the block selection method of Patton *et al.* (2009) provided by the R package *np* of Hayfield & 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 6 for the case $\rho = 0.9$ and $d = 1$. Table 6 (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 ζ^2 and the variance of $\hat{\zeta}^2$. For example, when $d = 1$, $\lambda = 2.5$ and $\tau = 0.5$, $\zeta^2 \approx 53\%$ and $\mathrm{Var}(\hat{\zeta}^2) \approx 0.54$ for $\rho = 0.95$, while for $\rho = 0$ (i.i.d.) $\zeta^2 \approx 92\%$ and $\mathrm{Var}(\hat{\zeta}^2) \approx 0.02$.

7. Ultrasonic calibration data

In section 7, we consider a real data analysis. Our objective is to illustrate the usefulness of $\hat{\zeta}^2$ as a decision rule to find the best approximation among several candidate parametric models. The data are the result of a National Institute of Standards and Technology (NIST) study involving ultrasonic calibration. The response variable is ultrasonic response, and the predictor variable is metal distance. There are 214 observations; see http://www.nist.gov/srd/ for more details about the data. Here, we study and compare the following models:

- M1: Simple linear regression model: $\beta_1 + \beta_2 x$;
- M2: Polynomial model of degree 2: $\beta_1 + \beta_2 x + \beta_3 x^2$;
- M3: Polynomial model of degree 3;
- M4: The nonlinear model: $\exp(-\beta_1 x)/(\beta_2 + \beta_3 x)$;
- M5: The Biexponential model: $\beta_1 \exp(-\exp(\beta_2)x) + \beta_3 \exp(-\exp(\beta_4)x)$;
- M6: The Asymptotic regression model: $\beta_1 + (\beta_2 - \beta_3) \exp(-\exp(\beta_4)x)$;
- M7: The Gompertz Growth model: $\beta_1 \exp(-\beta_2 \beta_3^x)$;
- M8: The Michaelis–Menten model: $(\beta_1 x)/(\beta_2 + x)$; and
- M9: The Weibull growth curve model: $\beta_1 - \beta_2 \exp(-\exp(\beta_3)x^{\beta_4})$.

We calculate $\hat{\zeta}^2$, its estimated asymptotic standard deviation, its corresponding 95%-upper confidence limit (UCL) and the mean squared prediction error (MSPE), the AIC and the BIC of each model. All the results are shown in Table 7. It can be seen that all the linear models

Table 7. $\hat{\zeta}^2$, *its estimated asymptotic standard deviation, the* 95%*-upper confidence limit (UCL) and the mean squared prediction error (MSPE), the AIC and the BIC of each model*

	$\hat{\zeta}^2 \times 100$	$Asym.\hat{sd} \times 100$	$UCL \times 100$	*MSPE*	AIC	BIC
M7	0.00002	0.08	0.009	10.6	1120.8	1968.8
M9	0.00116	0.69	0.079	10.7	1123.7	2183.7
M5	0.00337	1.18	0.136	10.6	1123.1	2183.1
M4	2.37645	16.0	4.183	11.1	1131.2	1979.2
M6	4.67667	24.0	7.380	11.2	1132.1	1980.1
M3	25.6416	66.0	33.06	14.4	1187.8	2247.8
M2	67.8416	54.0	73.92	33.2	1364.8	2212.8
M8	81.1551	40.8	85.74	158	1697.7	2333.7
M1	93.4079	14.2	95.00	162	1702.2	2338.2

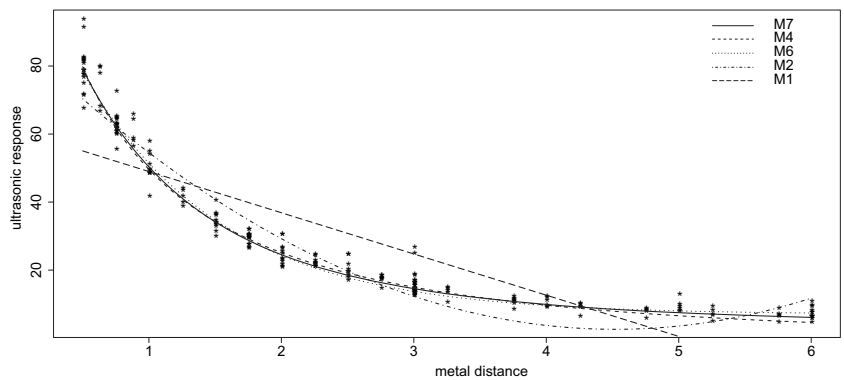

Fig. 1. Scatter plot of the ultrasonic calibration data with some fitted curves.

(M1, M2 and M3) give unsatisfactory results. M1 is the worst model with an inadequacy index of about 93%. The best model is M7 with almost zero inadequacy index. Although it should not be always the case, M7 is also the best model according to the MSPE criterion, and according also to the AIC and BIC. For M7, given that the 95%-upper limit of ζ^2 is of only 0.009%, we can definitively validate this model as being the (most) correct one. Note that the model M6 recognized as the best by NIST is ranked 4th by our inadequacy index ($\approx 2\%$). Finally, Fig. 1 shows the scatter plot of the data with some fitted curves.

Acknowledgements

We thank the associate editor who gave many suggestions that improved this article. We also thank the two anonymous referees for their valuable comments. A. El Ghouch acknowledges financial support from IAP research network P6/03 of the Belgian Government (Belgian Science Policy), and from the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique', granted by the 'Académie universitaire Louvain'.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Proofs of the asymptotic results.

References

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

Biedermann, S. & Dette, H. (2000). Testing linearity of regression models with dependent errors by kernel based methods. *TEST* **9**, 417–438.

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

Cristobal Cristobal, J. A., Roca Faraldo, P. & González Manteiga, W. (1987). A class of linear regression parameter estimators constructed by nonparametric estimation. *Ann. Statist.* **15**, 603–609.

Delgado, M. A. & González Manteiga, W. (2001). Significance testing in nonparametric regression based on the bootstrap. *Ann. Statist.* **29**, 1469–1507.

Dette, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimators. *Ann. Statist.* **27**, 1012–1040.

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

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

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

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

Fan, J. & Gijbels, I. (1996). *Local polynomial modelling and its applications, Monographs on statistics and applied probability*, Vol. **66**. Chapman & Hall, London.

Fan, J. & Yao, Q. (2003). *Nonlinear time series. Springer series in statistics. Nonparametric and parametric methods*. Springer-Verlag, New York.

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

Fan, Y. & Li, Q. (1999). Central limit theorem for degenerate U-statistics of absolutely regular processes with applications to model specification testing. *J. Nonparametr. Statist.* **10**, 245–271.

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

Härdle, W. & Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21**, 1926–1947.
Hayfield, T. & Racine, J. S. (2008). Nonparametric econometrics: The np package. *J. Statist. Softw.* **27**, 1–32.
Hodges, J. L. & Lehmann, E. L. (1954). Testing the approximative validity of statistical hypotheses. *J. Roy. Statist. Soc. Ser. B* **16**, 261–268.
Hong, Y. & White, H. (1995). Consistent specification testing via nonparametric series regression. *Econometrica* **63**, 1133–1159.
Jun, S. J. & Pinkse, J. (2009). Semiparametric tests of conditional moment restrictions under weak or partial identification. *J. Econometrics* **152**, 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*. Springer-Verlag, New York.
Lavergne, P. (1998). Selection of regressors in econometrics: parametric and nonparametric methods selection of regressors in econometrics. *Econometric Rev.* **17**, 227–273.
Li, Q. & Racine, J. (2004). Cross-validated local linear nonparametric regression. *Statist. Sinica* **14**, 485–512.
Li, Q. & Wang, S. (1998). A simple consistent bootstrap test for a parametric regression function. *J. Econometrics* **87**, 145–165.
Liu, R. & Singh, K. (1992). Moving blocks jackknife and bootstrap capture weak dependence. In *Exploring the limits of bootstrap (east lansing, mi, 1990)* (ed L. Billard), 225–248. Wiley, New York.
Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *J. Time Ser. Anal.* **17**, 571–599.
Patton, A., Politis, D. N. & White, H. (2009). Correction to 'Automatic block-length selection for the dependent bootstrap' by D. Politis & H. White [mr2041534]. *Econometric Rev.* **28**, 372–375.
R Development Core Team (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
Xia, Y. & Li, W. K. (2002). Asymptotic behavior of bandwidth selected by the cross-validation method for local polynomial fitting. *J. Multivariate Anal.* **83**, 265–287.
Zhang, C. & Dette, H. (2004). A power comparison between nonparametric regression tests. *Statist. Probab. Lett.* **66**, 289–301.
Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *J. Econometrics* **75**, 263–289.

Received December 2011, in final form October 2012

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