INSTITUT DE STATISTIQUE
BIOSTATISTIQUE ET
SCIENCES ACTUARIELLES
(ISBA)

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION
PAPER

2012/01

MEASURING THE DISCREPANCY OF A PARAMETRIC MODEL VIA LOCAL POLYNOMIAL SMOOTHING

EL GHOUCH, A., GENTON, G.M. and T. BOUEZMARNI

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.

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: Inadequacy index; Explanatory power ; Model misspecification; Validation test; Multivariate local polynomial smoothing; Strong mixing sequence.

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. For that reason and for the purpose of interpretability, simple parametric models with few covariates are usually preferred to a purely nonparametric 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 nonparametric 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 and Mammen (1993), Hong and White (1995), Zheng (1996), Li and Wang (1998), Delgado and González Manteiga (2001), Zhang and Dette (2004) and Jun and Pinkse (2009) for consistently testing a parametric regression functional form.

These "classical" nonparametric 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 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 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.

The rest of the paper is organized as follows. In Section 2 we introduce our inadequacy index in term 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 neighborhood 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 the Appendix.

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

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

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 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 $m(\theta, X) = \mathbb{E}(Y)$ 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 find within the parametric family $\mathcal{M} = \{m(\theta, x),\ \theta \in \Theta\}$. We define our parameter of interest to be

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

where we introduce the known weight function φ in order, e.g., to avoid highly uncertain estimation in regions with sparse or noisy data. 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 $\mathcal{M}$.

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 $T(\theta^*) \equiv T_\varphi(\theta^*) := \mathbb{E}[\Delta^2(\theta^*, X)\varphi^2(X)]$ and $S(\theta^*) \equiv S_\varphi(\theta^*) := \mathbb{E}[Y^2(\theta^*)\varphi^2(X)]$. 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 $\mathcal{F}_I^L$ $(-\infty \leq I, L \leq \infty)$ denote the σ-field generated by the family $\{(X_t, Y_t),\ I \leq t \leq L\}$. 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 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.

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 nonnegative kernel function defined on $\mathbb{R}^d$, $0 < h_n \equiv h \to 0$ be a bandwidth parameter and $K_h(x) = h^{-d}K(x/h)$. 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 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, $p \in \mathbb{N}$, 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 (8).

The estimator given by (1) is only available when θ is known, which, of course, 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 $\Delta(\theta, x)$: $\hat{\Delta}(\hat{\theta}, x) = \sum_{j=1}^n w_j(x)Y_j(\hat{\theta}) = \hat{m}(x) - \hat{m}(\hat{\theta}, x)$, 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 $\hat{m}(x)$.

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

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

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

This quantity is related to a discrete (Riemann sum) version of a test statistic that was proposed by Härdle and 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

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

It is straightforward to show that this estimator is simply the empirical version of the expression $T(\theta) = \mathbb{E}\left[(2Y(\theta) - \Delta(\theta, X))\, \Delta(\theta, X)\varphi^2(X) \right]$. 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 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).

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 for the results to hold. 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 $\dot{m} := \dfrac{\partial m}{\partial \theta}$ 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 $0 \leq |k| \leq 2p$ (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 (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).

4.2 Main results

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* $Y(\theta) = Y - m(\theta, X)$.

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

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 these 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.*

- *Although the 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 6.

- *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 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,*

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

  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* $\Delta_n(x)$.

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 and White (1982) we claim that, under Assumptions (A), $\hat{\theta}$ converges with probability 1 to

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

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

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

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.

Since $\zeta^2 = T(\theta^*)/S(\theta^*)$, 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 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 $\sigma^2 \equiv \sigma^2_\varphi(\theta^*)$, with $\sigma^2_\varphi(\theta) := \lim_{n\to\infty} n^{-1}\mathbb{V}ar\left(\sum_{t=1}^n \xi_t(\theta)\right) = \mathbb{V}ar[\xi_1(\theta)] + 2\sum_{t>1}\mathbb{C}ov(\xi_1(\theta), \xi_t(\theta))$.

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" $\{\xi_i\}_{1\le i\le n}$ 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 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 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 $\hat{\sigma}^2_n$ 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 in-

formation about the quality 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 term 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 asymptotics in such a form have been noted before by many authors; see for example Fan and Li (1996) and Fan and Li (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-statistics as defined by (20) 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 (6) over (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 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 it asymptotic properties as stated in the previous sections. 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 $n := 200$ 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 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

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

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 $Unif[-\varepsilon, 1+\varepsilon]$. 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 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: $\hat{\zeta}_0^2 = T_n^0(\hat{\theta})/S_n(\hat{\theta})$ and $\hat{\zeta}^2 = T_n(\hat{\theta})/S_n(\hat{\theta})$; see equations (2) and (3). In Table 1 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 1 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 this table 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. 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 $|MSE - MSE^*|$ 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 2 gives the Relative efficiency of the local linear to the local constant approximation, i.e. $MSE(\hat{\zeta}_{p=0}^2)/MSE(\hat{\zeta}_{p=1}^2)$, 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 3, demonstrate the consistent nature of the proposed method. Globally, the MSE performance is very satisfactory and better than expected.

To complete the picture, Table 4 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 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 4 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). Practi-

cally, 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 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 5 for the case $\rho = 0.9$ 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 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 $\mathbb{V}ar(\hat{\zeta}^2) \approx 0.54$ for $\rho = 0.95$, while for $\rho = 0$ (i.i.d.) $\zeta^2 \approx 92\%$ and $\mathbb{V}ar(\hat{\zeta}^2) \approx 0.02$.

7 Ultrasonic Calibration Data

In this section 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: $\beta_1 \frac{x}{\beta_2 + x}$, 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 6. It can be seen that all the linear models (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, Figure 1 shows the scatter plot of the data with some fitted curves.

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

$$\begin{aligned} w_j(x) &= e_{N,1}^T\mathbf{S}_n^{-1}(x)\mathcal{X}^T\mathbf{W}e_{n,j} \\ &= n^{-1}e_{N,1}^T\mathbf{S}_n^{-1}(x)\gamma_h(X_j - x)K_h(X_j - x) \end{aligned} \tag{8}$$

where, for r, 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,

$$\sup_{x \in D} \sum_{j=1}^{n} |w_j(x)| = O_P(1). \tag{9}$$

REMARK **2** *$\mathcal{X}^T\mathbf{W}\mathbf{Y}$ 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,

$$T_n(\hat{\theta}) - T_n(\theta^*) = (\hat{\theta} - \theta^*)^T \dot{T}_n(\tilde{\theta}_n), \tag{10}$$

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

$$\dot{T}_n(\theta) = 2n^{-1} \sum_i \left[\dot{Y}_i(\theta)\hat{\Delta}(\theta, X_i) + (Y_i(\theta) - \hat{\Delta}(\theta, X_i))\dot{\hat{\Delta}}(\theta, X_i) \right] \varphi^2(X_i), \tag{11}$$

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 (10), it is clear that Lemma 1 is equivalent to say that $\dot{T}_n(\tilde{\theta}_n) = -2B + o_p(1)$. Let $I_n^{-1} \sum_{i=1}^n (Y_i(\tilde{\theta}_n) - \hat{\Delta}(\tilde{\theta}_n, X_i))\dot{\hat{\Delta}}(\tilde{\theta}_n, X_i)\varphi^2(X_i)$. 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 $I_{n,1} = n^{-1} \sum_i \epsilon_i \dot{\hat{\Delta}}(\theta^*, X_i)\varphi^2(X_i)$, $I_{n,2} = n^{-1} \sum_i \epsilon_i (\dot{\hat{\Delta}}(\tilde{\theta}_n, X_i) - \dot{\hat{\Delta}}(\theta^*, X_i))\varphi^2(X_i)$, $I_{n,3} = n^{-1} \sum_i (\Delta(\tilde{\theta}_n, X_i) - \Delta(\theta^*, X_i))\dot{\hat{\Delta}}(\tilde{\theta}_n, X_i)\varphi^2(X_i)$, $I_{n,4} = n^{-1} \sum_i (\hat{\Delta}(\tilde{\theta}_n, X_i) - \hat{\Delta}(\theta^*, X_i))\dot{\hat{\Delta}}(\tilde{\theta}_n, X_i)\varphi(^2 X_i)$, 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,

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

So, by (9) and assumption (A1),

$$\sup_{(\theta,x) \in \Theta \times D} ||\dot{\hat{\Delta}}(\theta, x)|| = O_p(1). \tag{12}$$

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)

$$\sup_{x \in D} |\Delta(\tilde{\theta}_n, x) - \Delta(\theta^*, x)| = O_p(||\hat{\theta} - \theta^*||). \tag{13}$$

Using (12) we can also check that

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

Similarly one can prove that,

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

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 (9) and assumption (A1), we have that

$$\sup_{x \in D} |\hat{\Delta}(\theta^*, x) - \Delta(\theta^*, x)| = o_p(1). \tag{16}$$

Similarly,

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

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 (12), we get from (17), (15), (13), (14) and (16) that, respectively,

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

So, we have established that $I_n = o_p(1)$, hence, by (11)

$$
\begin{aligned}
\dot{T}_n(\tilde{\theta}_n) &= 2n^{-1}\sum_i \dot{Y}_i(\tilde{\theta}_n)\hat{\Delta}(\tilde{\theta}_n, X_i)\varphi^2(X_i) + o_p(1) \\
&= -2n^{-1}\sum_i \dot{m}(\tilde{\theta}_n, X_i)\hat{\Delta}(\tilde{\theta}_n, X_i)\varphi^2(X_i) + o_p(1) \\
&= -2n^{-1}\sum_i \dot{m}(\theta^*, X_i)\Delta(\theta^*, X_i)\varphi^2(X_i) + o_p(1) \qquad (18) \\
&= -2B + o_p(1),
\end{aligned}
$$

where in (18) we have used (14), $\sup_{x\in D} ||\dot{m}(\tilde{\theta}_n, x) - \dot{m}(\theta^*, x)|| = O_p(||\tilde{\theta}_n - \theta^*)||)$ and the fact that both $\dot{m}(\theta^*, x)$ and $\Delta(\theta^*, x)$ are bounded in D. This together with (10) 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

$$\begin{aligned} T_n(\theta) =& n^{-1} \sum_{i=1}^{n} \left[2Y_i(\theta)\Delta(\theta, X_i) - \Delta^2(\theta, X_i) \right] \varphi^2(X_i) \\ &+ 2n^{-1} \sum_{i=1}^{n} \left[\hat{\Delta}(\theta, X_i) - \Delta(\theta, X_i) \right] \epsilon_i \varphi^2(X_i) \\ &- n^{-1} \sum_{i=1}^{n} \left[\hat{\Delta}(\theta, X_i) - \Delta(\theta, X_i) \right]^2 \varphi^2(X_i). \end{aligned} \tag{19}$$

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

$$J_{n,1}^{''} := n^{-2}\sum_{i=1}^n\sum_{j\neq i} d_{n,0}(X_i)\gamma_h(X_j-X_i)K_h(X_j-X_i)\epsilon_i\epsilon_j\varphi^2(X_i). \tag{20}$$

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)

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

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

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

Delgado, M. A. and W. González Manteiga (2001). Significance testing in nonparametric regression based on the bootstrap. *The Annals of Statistics 29*, 1469–1507.

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

Dette, H. and A. Munk (2003). Some methodological aspects of validation of models in nonparametric regression. *Statistica 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. *The Annals of Statistics 23*(5), 1443–1473.

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

Doukhan, P. and S. Louhichi (1999). A new weak dependence condition and applications to moment inequalities. *Stochastic Processes and their Applications 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.

Fan, Y. and Q. Li (1999). Central limit theorem for degenerate U-statistics of absolutely regular processes with applications to model specification testing. *Journal of Nonparametric Statistics 10*(3), 245–271.

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

Hall, P. (1992). On bootstrap confidence intervals in nonparametric regression. *The Annals of Statistics 20*(2), 695–711.

Härdle, W. and E. Mammen (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics 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. *Journal of Econometrics 152*(1), 3–18.

Künsch, H. (1989). The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics 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.

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

Li, Q. and S. Wang (1998). A simple consistent bootstrap test for a parametric regression function. *Journal of 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. *Journal of Time Series Analysis 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 Reviews 28*(4), 372–375.

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

Zhang, C. and H. Dette (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. *Journal of Econometrics 75*(2), 263–289.

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

			$100 \times RMSE^*$						$100 \times RMSE$					
		$d=1$	$d=1$		$d=2$		$d=3$		$d=1$		$d=2$		$d=3$	
λ	τ	$\zeta^2\%$	$\hat{\zeta}_0^2$	$\hat{\zeta}^2$	$\hat{\zeta}_0^2$	$\hat{\zeta}^2$	$\hat{\zeta}_0^2$	$\hat{\zeta}^2$	$\hat{\zeta}_0^2$	$\hat{\zeta}^2$	$\hat{\zeta}_0^2$	$\hat{\zeta}^2$	$\hat{\zeta}_0^2$	$\hat{\zeta}^2$
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

Table 2: 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 3: 100×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 4: 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 5: 100×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×RMSE*	100×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

Table 6: $\hat{\zeta}^2$, 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

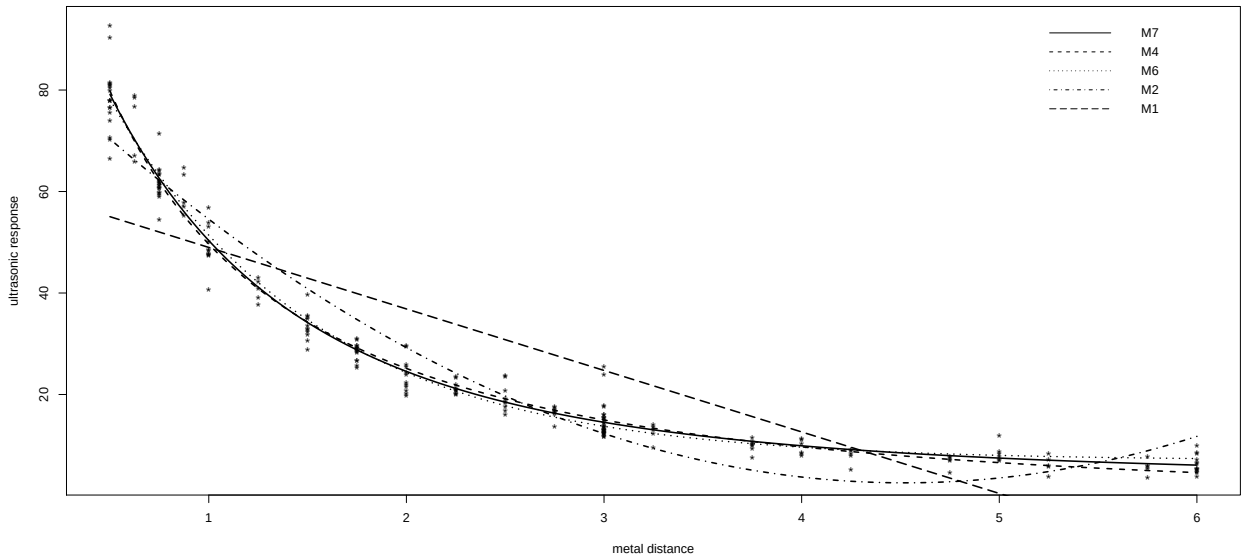

Figure 1: Scatter plot of the ultrasonic calibration data with some fitted curves.