On an Adaptive Goodness–of–Fit test with Finite Sample Validity for Random Design Regression Models

Brutti Pierpaolo
"Sapienza" Università di Roma.

Keywords: Goodness–of–fit; Adaptive test; Nonparametric Regression; Separation Rates; Warped Wavelets; U–statistics; Multiple Test.
AMS: 62G08, 62G10

Abstract

Given an i.i.d. sample $\{(X_i, Y_i)\}_{i\in\{1,\dots,n\}}$ from the random design regression model $Y = f(X) + \epsilon$ with $(X, Y) \in [0, 1] \times [-M, M]$, in this paper we consider the problem of testing the (simple) null hypothesis "$f = f_0$", against the alternative "$f \neq f_0$" for a fixed $f_0 \in \mathsf{L}^2([0, 1], G_X)$, where $G_X(\cdot)$ denotes the (known) marginal distribution of the design variable X. The procedure proposed is an adaptation to the regression setting of a multiple testing technique introduced by Fromont and Laurent [5], and it amounts to consider a suitable collection of unbiased estimators of the L^2–distance $\mathsf{d}_2(f, f_0) = \int [f(x) - f_0(x)]^2 \mathrm{d}G_X(x)$, rejecting the null hypothesis when at least one of them is greater than its $(1 - u_\alpha)$ quantile, with u_α calibrated to obtain a level–α test. To build these estimators, we will use the *warped wavelet* basis recently introduced by Picard and Kerkyacharian [7]. We do not assume that the errors are normally distributed, and we do not assume that X and ϵ are independent but, mainly for technical reasons, we will assume, as in most part of the current literature in learning theory, that $|f(x) - y|$ is uniformly bounded (almost everywhere). We show that our test is adaptive over a particular collection of approximation spaces linked to the classical Besov spaces.

1. Introduction

Consider the usual nonparametric regression problem with random design. In this model we observe an i.i.d. sample $\mathcal{D}_n = \{\mathbf{Z}_i = (X_i, Y_i)\}_{i\in\{1\dots n\}}$ from the distribution of a vector $\mathbf{Z} = (X, Y)$ where

$$Y = f(X) + \epsilon, \tag{1}$$

for (X, ϵ) a random vector with $\mathbb{E}(\epsilon|X) = 0$ and $\mathbb{E}(\epsilon^2|X) < \infty$ almost surely. The regression function is known to belong to a subset $\mathcal{F}$ of $\mathsf{L}^2([0, 1], G_X)$ for G_X the marginal distribution of X, which will be assumed known. Let $f_0 \in \mathcal{F}$ be fixed. In this paper we consider the problem of testing the (simple) null hypothesis "$\mathsf{H}_0 : f = f_0$" against the alternative "$\mathsf{H}_1 : f \neq f_0$". Since $f \in \mathsf{L}^2([0, 1], G_X)$, it seems natural to consider a test statistic somehow linked to an

estimator of the (weighted) L^2–distance $\mathsf{d}_2(f, f_0) = \int [f(x) - f_0(x)]^2 \mathrm{d}G_X(x)$. The approach considered in the present paper is an adaptation to the regression setting with random design of the work by Fromont and Laurent (2006) for density models, and it amounts to consider a suitable collection of unbiased estimators for $\mathsf{d}_2(f, f_0)$, rejecting the null hypothesis when at least one of them is greater than its $(1 - u_\alpha)$ quantile, with u_α calibrated to obtain a level–α test. Similar problems have been widely studied in the testing literature. See, for example, the nice review provided by Hart [6]. From a more theoretical point of view, Spokoiny [10] considers a Gaussian white noise model $\mathrm{d}X(t) = f(t)\mathrm{d}t + \epsilon \mathrm{d}W(t)$, and propose to test "$f \equiv 0$" adaptively using a wavelet based procedure. He also study the (asymptotic) properties of his approach and show that, in general, adaptation is not possible without some loss of efficiency of the order of an extra $\log\log(n)$ factor, where n is the sample size. In the same setting, Ingster (see Ingster and Suslina [8]) builds an adaptive test based on chi–square statistics, and study its asymptotic properties. The literature regarding goodness–of–fit testing in a density model is also vast (see [5] and references therein).

The pre–testing approach considered here has been initiated by Baraud, Huet and Laurent [2] for the problem of testing linear or qualitative hypotheses in the Gaussian regression model. One nice feature of their approach is that the properties of the procedures are non asymptotic.

2. A Goodness–of–Fit Test

Consider again the regression model in Equation 1. We do not assume that the errors are normally distributed, and we do not assume that X and ϵ are independent but, mainly for technical reasons, we will assume, as in most part of the current literature in learning theory (see Cucker and Smale [3]), that $|f(x) - y|$ is uniformly bounded (almost everywhere) by a positive constant M. Doing so, all the proofs will be greatly simplified without moving too far away from a realistic (although surely not minimal) set of assumptions (in particular considering the finite–sample scope of the analysis). What we propose is a goodness–of–fit test similar to the one introduced in [5]. To describe it, let $f_0(\cdot)$ be some fixed function in $\mathsf{L}^2([0,1], G_X)$ and $\alpha \in (0,1)$. Now suppose that our goal is to build a level–α test of the null hypothesis $\mathsf{H}_0 : f \equiv f_0$ against the alternative $\mathsf{H}_1 : f \neq f_0$ from the data $\{\mathbf{Z}_i\}_{i \in \{1,\ldots,n\}}$. The test is based on an estimation of

$$\|f - f_0\|^2_{\mathsf{L}^2(G_X)} = \|f\|^2_{\mathsf{L}^2(G_X)} + \|f_0\|^2_{\mathsf{L}^2(G_X)} - 2\langle f, f_0\rangle_{\mathsf{L}^2(G_X)}.$$

Since the last (linear) term $\langle f, f_0\rangle_{\mathsf{L}^2(G_X)}$ can be easily estimated by the empirical estimator $\frac{1}{n}\sum_{i=1}^n Y_i\, f_0(X_i)$, the key problem is the estimation of the first term $\|f\|^2_{\mathsf{L}^2(G_X)}$. Adapting the arguments in Laurent [9], we can consider an at most countable collection of linear subspaces of $\mathsf{L}^2([0,1], G_X)$ denoted by

$\mathcal{S} = \{\mathrm{S}_k\}_{k\in\mathrm{K}}$. For all $k \in \mathrm{K}$, let $\{e_\ell\}_{\ell\in\mathcal{I}_k}$ be some orthonormal basis of S_k. The estimator

$$\hat{\theta}_{n,k} = \frac{1}{n(n-1)} \sum_{i=2}^{n} \sum_{j=1}^{n-1} \left[\sum_{\ell\in\mathcal{I}_k} \{Y_i e_\ell(X_i)\} \cdot \{Y_j e_\ell(X_j)\} \right], \tag{2}$$

is a U–statistic of order two for $\|\Pi_{\mathrm{S}_k}(f)\|^2_{\mathsf{L}^2(G_X)}$ – where $\Pi_{\mathrm{S}_k}(\cdot)$ denotes the orthogonal projection onto S_k – with kernel

$$h_k(\mathbf{z}_1, \mathbf{z}_2) = \sum_{\ell\in\mathcal{I}_k} \{y_1 e_\ell(x_1)\} \cdot \{y_2 e_\ell(x_2)\}, \quad \mathbf{z}_i = (x_i, y_i),\ i \in \{1, 2\}.$$

Then, for any $k \in \mathrm{K}$, $\|f - f_0\|^2_{\mathsf{L}^2(G_X)}$ can be estimated by

$$\widehat{\mathsf{R}}_{n,k} = \widehat{\theta}_{n,k} + \|f_0\|^2_{\mathsf{L}^2(G_X)} - \frac{2}{n} \sum_{i=1}^{n} Y_i\ f_0(X_i). \tag{3}$$

Now that we have an estimator $\widehat{\mathsf{R}}_{n,k}$, lets denote by $r_{n,k}(u)$ its $1-u$ quantile under H_0, and consider

$$u_\alpha = \sup \left\{ u \in (0,1) : \mathsf{P}^{\otimes n}_{f_0} \left[\sup_{k\in K} \left\{ \widehat{\mathsf{R}}_{n,k} - r_{n,k}(u) \right\} > 0 \right] \leqslant \alpha \right\},$$

where $\mathsf{P}^{\otimes n}_{f_0}\{\cdot\}$ is the law of the observations $\{\mathbf{Z}_i\}_{i\in\{1,\ldots,n\}}$ under the the null hypothesis. Then introduce the test statistics R_α defined by

$$\mathsf{R}_\alpha = \sup_{k\in K} \left\{ \widehat{\mathsf{R}}_{n,k} - r_{n,k}(u_\alpha) \right\},$$

so that we reject the null whenever R_α is positive.

This method amounts to a multiple testing procedure. Indeed, for all $k \in \mathrm{K}$, we construct a level–u_α test by rejecting $\mathsf{H}_0 : f \equiv f_0$ if $\widehat{\mathsf{R}}_{n,k}$ is greater than its $(1-u_\alpha)$ quantile under H_0. After this, we are left with a collection of tests and we decide to reject H_0 if, for some of the tests in the collection, the hypothesis is rejected.

3. Power of the Test

Both the practical and theoretical performances of the proposed test, depend strongly on the orthogonal system we adopt to generate the collection of linear subspaces $\{S_k\}_{k\in\mathrm{K}}$. A basis that fit perfectly in the present framework, is the so–called *warped wavelet* basis studied by Kerkyacharian and Picard [7]. The idea is to start from a standard wavelet basis $\{\psi_(\cdot)j,k\}_{(j,k)}$ and build by composition a new system $\{\psi_{j,k}(G(\cdot))\}_{(j,k)}$, where $G(\cdot)$ is adapting to the design: it may be the distribution function of the design $G_X(\cdot)$ itself, or its

estimation when it is unknown. An appealing feature of this method is that it does not need a new algorithm to be implemented.

At this point, for each $\mathrm{J} \in \mathbb{N}$, we have a system of scaling functions $\{\phi_{\mathrm{J},k}(G)\}_k$ that we can use to generate the subspaces $\mathcal{S} = \{\mathrm{S}_\mathrm{J}\}_{\mathrm{J}\in\mathbb{N}}$ where we have slightly changed the indexing notation: from k to J.

The following theorem, describes the class of alternatives over which the test has a prescribed power.

Theorem 3.1 *Let $\{\mathbf{Z}_i = (X_i, Y_i)\}_{i\in\{1,\dots,n\}}$ be an i.i.d. sequence from the distribution of a vector $\mathbf{Z} = (X, Y)$ described structurally by the nonparametric regression model*

$$Y = f(X) + \epsilon,$$

for (X, ϵ) a random vector with $\mathbb{E}(\epsilon|X) = 0$ and $\mathbb{E}(\epsilon^2|X) < +\infty$. Assume further that $f_0(\cdot)$ and the unknown regression function $f(\cdot)$ belong to $\mathsf{L}^2([0,1], G_X)$ for $G_X(\cdot)$ the marginal distribution of X, assumed know and absolutely continuous with density $g_X(\cdot)$ bounded from below and above. Finally assume that $|f(x) - y|$ is uniformly bounded (almost everywhere) by a positive constant M.

Now let $\beta \in (0,1)$. For all $\gamma \in (0,2)$, there exist positive constants $C_1 \equiv C_1(\beta)$ and $C_2 \equiv C_2(\beta, \gamma, \tau_\infty, M, \|f_0\|_\infty)$ such that, defining

$$\mathsf{V}_{n,\mathrm{J}}(\beta) = \frac{C_1}{n}\left\{\tau_\infty \cdot \sqrt{2^\mathrm{J}} + \frac{M^2}{n}2^\mathrm{J}\right\} + \frac{C_2}{n},$$

with $\tau_\infty = \|f\|_\infty^2 + \|\sigma^2\|_\infty$, then, for every $f(\cdot)$ such that

$$\|f - f_0\|^2_{\mathsf{L}^2(G_X)} > (1+\gamma)\inf_{\mathrm{J}\in\mathcal{J}_n}\left\{\|f - \Pi_{S_\mathrm{J}}(f)\|^2_{\mathsf{L}^2(G_X)} + r_{n,\mathrm{J}}(u_\alpha) + \mathsf{V}_{n,\mathrm{J}}(\beta)\right\},$$

the following inequality holds: $\mathsf{P}_f^{\otimes n}\{\mathsf{R}_\alpha \leqslant 0\} \leqslant \beta$.

4. Uniform Separation Rates

Now that we know against what kind of alternatives our multiple testing procedure has guaranteed power, we can move on, and examine the problem of establishing uniform non–asymptotic separation rates (see Ingster and Suslina [8] and Baraud [1]) over well–suited functional classes included in $\mathsf{L}^2([0,1], G_X)$. We will start by defining for all $s > 0$, $R > 0$, and $M > 0$, the following (linear) approximation space (see the review by Devore [4]):

$$\mathcal{A}^s(R, M, G_X) = \left\{w \in \mathsf{L}^2(G_X) : \|w\|_\infty \leqslant M, \|w - \Pi_{\mathrm{S}_\mathrm{J}}(w)\|^2_{\mathsf{L}^2(G_X)} \leqslant R^2 2^{-2\mathrm{J}\,s}\right\}.$$

When $\mathrm{d}G_X(x) = \mathrm{d}x$ is the Lebesgue measure, $\mathcal{A}^s(R, M, \mathrm{d}x)$ is strictly related to a well–known Besov body. In our case, instead, it is a bit less clear how

to "visualize" the content of $\mathcal{A}^s(R, M, G_X)$ in terms of common smoothness classes. The easiest way, is to notice that, for each $w \in \mathsf{L}^2([0,1], G_X)$

$$\|w - \Pi_{S_\mathrm{J}}(w)\|^2_{\mathsf{L}^2(G_X)} = \left\|w(G_X^{-1}) - \Pi_{S_\mathrm{J}}\big(w(G_X^{-1})\big)\right\|^2_{\mathsf{L}^2(\mathrm{d}x)},$$

where

$$G_X^{-1}(x) = \inf\{t \in \mathbb{R} : G_X(t) \geqslant x\}$$

is the *quantile function* of the design distribution $G_X(\cdot)$. Consequently,

$$f \in \mathcal{A}^s(R, M, G_X) \Leftrightarrow f(G_X^{-1}) \in \mathcal{A}^s(R, M, \mathrm{d}x),$$

so that the regularity conditions that hide behind the definition of the approximation space $\mathcal{A}^s(R, M, G_X)$ could be expressed more explicitly in terms of the *warped* function $f \circ G_X^{-1}(\cdot)$, mixing the smoothness of $f(\cdot)$ with the (very regular, indeed) design $G_X(\cdot)$ (see [7]).

From here we can prove that, under suitable conditions, our procedure adapts over the approximation space $\mathcal{A}^s(R, M, G_X)$ at a rate known to be optimal for a particular scale of Besov spaces (Spokoiny [10]).

5. Bibliography

[1] Baraud Y. (2000). *Non asymptotic separation rates of testing in signal detection.* Technical report, Ecolé Normale Supérieure, Paris.

[2] Baraud Y., Huet S., and Laurent B. (2003). *Adaptive tests of linear hypotheses by model selection.* The Annals of Statistics, 31, 225-Ũ251.

[3] Cucker F. and Smale S. (2002) *On the mathematical foundations of learning.* Bull. Amer. Math. Soc. (N.S.), 39(1), 1–49.

[4] Devore R. (1998) *Nonlinear approximation.* Acta Numerica, 1–99.

[5] Fromont M. and Laurent B. (2006) *Adaptive goodness–of–fit tests in a density model.* The Annals of Statistics, 34(2), 680–720.

[6] Hart J.D. (1997). Nonparametric Smoothing and Lack–of–Fit Tests. *Springer Series in Statistics.* Springer.

[7] Kerkyacharian, G. and Picard D. (2004). *Regression in random design and warped wavelets.* Bernoulli, 10(6), 1053–1105.

[8] Ingster Y. and Suslina I. A. Nonparametric Goodness–of–Fit Testing Under Gaussian Models. Number 169 in *Lecture Notes in Statistics.* Springer.

[9] Laurent B. (2005). *Adaptive estimation of a quadratic functional of a density by model selection.* ESAIM: Probability and Statistics, 9, 1–18.

[10] Spokoiny V. G. (1996). *Adaptive hypothesis testing using wavelets.* The Annals of Statistics, 24, 24770–2498.