Testing for Lack of Fit in Functional Regression Models

Samuel Maistre[a] & Valentin Patilea[a,b,1]

[a] CREST-Ensai & IRMAR, Campus de Ker Lann, rue Blaise Pascal, BP 37203, 35172 Bruz cedex, France.

[b] Corresponding author. Email: patilea@ensai.fr

Abstract. We consider regression models with a response variable taking values in a Hilbert space, of finite or infinite dimension, and hybrid covariates. This means there are two sets of regressors, one of finite dimension and a second one functional with values in a Hilbert space. The problem we address is the test of the effect of the functional covariates. This problem occurs in many situations: testing the effect of the functional covariate in a semi-functional partial linear regression with scalar responses, significance test for functional regressors in nonparametric regression with hybrid covariates and scalar or functional responses, testing the effect of a functional covariate on a scalar or functional outcome. We propose a new test based on univariate kernel smoothing. The test statistic is asymptotically standard normal under the null hypothesis provided the smoothing parameter tends to zero at a suitable rate. The one-sided test is consistent against any fixed alternative and detects local alternatives a la Pitman approaching the null hypothesis quickly enough. In particular we show that neither the dimension of the outcome nor the dimension of the functional covariates influences the theoretical power of the test against such local alternatives.

Keywords. Regression, lack-of-fit test, functional data, $U-$statistics

1 Introduction

Let$(\mathcal{H}_1, \langle\cdot,\cdot\rangle_{\mathcal{H}_1})$ and $(\mathcal{H}_2, \langle\cdot,\cdot\rangle_{\mathcal{H}_2})$ denote two possibly different Hilbert spaces. Herein we focus on the following situations: $\mathcal{H}_1 = \mathbb{R}$, $\mathcal{H}_2 = L^2[0,1]$ and $\mathcal{H}_1 = \mathcal{H}_2 = L^2[0,1]$.

Consider $U \in \mathcal{H}_1$, $Z \in \mathbb{R}^q$ and $W \in \mathcal{H}_2$ and let (U_i, Z_i, W_i), $1 \leq i \leq n$, denote a sample of independent copies of (U, Z, W). The statistical problem we consider is the test of the hypothesis

$$\mathbb{E}[U \mid Z, W] = 0 \qquad \text{a.s.} \tag{1}$$

[1] This author gratefully acknowledges financial support from the Romanian National Authority for Scientific Research, CNCS-UEFISCDI, project PN-II-ID-PCE-2011-3-0893.

against a general alternative like $\mathbb{P}(\mathbb{E}[U \mid Z, W] = 0) < 1$. This problem occurs in many model check problems.

a) *Semiparametric functional partially linear models.* Aneiros-Pérez and Vieu (2006) proposed an extension of the partially linear model to functional data. Their model writes as

$$Y = \sum_{j=1}^{q} X_j \beta_j + m(W) + \varepsilon$$

where the response Y and the covariates X_j are real-valued random variables, W is a random variable taking values in a functional space, typically $L^2[0,1]$, and the error term satisfies $\mathbb{E}[\varepsilon \mid X_1, \cdots X_q, W] = 0$ a.s. The coefficients β_j and the function $m(\cdot)$ have to be estimated. Before estimating $m(\cdot)$ nonparametrically, one should check the significance of the variable W. Let $Z = (X_1, \cdots, X_q)$ and $U = Y - \mathbb{E}[Y \mid Z]$. Then testing the significance of W is exactly testing condition (1). In this example, the variable U is not observed and the sample $U_1, \cdots, U_n$ has to be estimated by the residuals of the linear fit of Y given $X_1, \cdots, X_q$.

b) *Variable selection in functional nonparametric regression with functional responses.* Regression models for functional responses are now widely used, see for instance Faraway (1997). Two situations were studied: finite and infinite dimension covariates; see Ramsay and Silverman (2005), Ferraty *et al.* (2011), Ferraty *et al.* (2012). Consider the hybrid case with both finite and infinite dimension covariates. An important question is the significance of the functional covariates. In a more formal way, let $Y \in \mathcal{H}_1$ be the regressor and let $Z \in \mathbb{R}^q$ and $W \in \mathbb{H}_2$ denote the covariates. Then the problem is to test

$$\mathbb{E}[Y \mid Z, W] = \mathbb{E}[Y \mid Z].$$

Let $U = Y - \mathbb{E}[Y \mid Z]$. Then the problem becomes to test condition (1). In this example also the sample of the variable U_i is not observed and has to be estimated by the residuals of the nonparametric regression of Y given Z.

c) *Testing the effect of a functional variable.* Consider the random variables $U \in \mathcal{H}_1$, $\widetilde{W} \in \mathcal{H}_2$. Without loss of generality, we suppose that U is centered. We want to check if $\mathbb{E}[U \mid \widetilde{W}] = 0$. Patilea *et al.* (2012b) proposed a test procedure based on finite dimension subspaces of $\mathcal{H}_2$. Their test statistic is somehow related to a Kolmogorov-Smirnov statistic in a finite dimension space with the dimension growing with the sample size. Here we propose an alternative route. Let $Z = \langle \widetilde{W}, \psi_1 \rangle_{\mathcal{H}_2}$ where ψ_1 is an element of an orthonormal basis of $\mathcal{H}_2$. Suppose that Z admits a density with respect to the Lebesgue measure. The basis of $\mathcal{H}_2$ could be the one given by the functional principal components which in general have to be estimated from the data. In such a case, the sample of $Z_i's$ has to estimated too. Let $W = \widetilde{W} - \langle \widetilde{W}, \psi_1 \rangle_{\mathcal{H}_2} \psi_1$. Then, testing $\mathbb{E}[U \mid \widetilde{W}] = 0$ is nothing but testing condition (1).

2 Testing the significance of functional covariates.

Let $\{\phi_1, \phi_2, \cdots\}$ be an orthonormal basis of $\mathcal{H}_2$. Let $\beta_k = \langle W, \phi_k \rangle_{\mathcal{H}_2}$ and $W_{\underline{p}} = \sum_{k=1}^{p} \beta_k \phi_k$. For a function l, let $\mathcal{F}[l]$ denote the Fourier Transform of l. Let K be a multivariate kernel defined on $\mathbb{R}^q$ such that $\mathcal{F}[K] > 0$ and $\phi(s) = \exp(-\|s\|^2)$, $\forall s \in \mathbb{R}^p$. Many kernels satisfy the positive Fourier Transform condition, for instance the gaussian, triangle, Student and logistic densities.

Our new procedure is based on the following facts. First, for any positive function $\omega(\cdot)$ and any $h > 0$ and p positive integer

$$\begin{aligned} I(h) &= \mathbb{E}\left[\langle U_1, U_2 \rangle_{\mathcal{H}_1} \omega(Z_1)\omega(Z_2) h^{-q} K((Z_1 - Z_2)/h)\phi(W_{1,\underline{p}} - W_{2,\underline{p}})\right] \\ &= \mathbb{E}\left[\langle U_1, U_2 \rangle_{\mathcal{H}_1} \omega(Z_1)\omega(Z_2) \int_{\mathbb{R}^q} e^{2\pi i t'(Z_1 - Z_2)} \mathcal{F}[K](th) dt \right. \\ &\qquad \left. \times \int_{\mathbb{R}^p} e^{2\pi i s'(W_{1,\underline{p}} - W_{2,\underline{p}})} \mathcal{F}[\phi](s) ds \right] \\ &= \int_{\mathbb{R}^q} \int_{\mathbb{R}^p} \left\| \mathbb{E}\left[\mathbb{E}[U \mid Z, W] \omega(Z) e^{-i\{t'Z + s'W_{\underline{p}}\}}\right] \right\|_{\mathcal{H}_1}^2 \mathcal{F}[K](th)\mathcal{F}[\phi](s) dt ds. \end{aligned}$$

Using the conditions $\mathcal{F}[\phi], \mathcal{F}[K] > 0$ (and $\omega > 0$), for any p we have the equivalence

$$\mathbb{E}(U \mid Z, W_{\underline{p}}) = 0 \ \ a.s. \ \ \Leftrightarrow \ \ I(h) = 0, \quad \forall h > 0.$$

Second,

$$\mathbb{E}(U \mid Z, W) = 0 \ \ a.s. \ \ \Leftrightarrow \ \ \mathbb{E}(U \mid Z, W_{\underline{p}}) = 0 \ \ a.s. \quad \forall p \in \{1, 2, \cdots\};$$

see Patilea *et al.* (2012a,b).

Then, the idea of the new approach is to build a test statistic using an approximation of $I(h)$. A convenient choice of the function $\omega(\cdot)$ will allow to simplify this task. In the examples a) and c) we could simply take $\omega(\cdot) \equiv 1$. In example b) we take $\omega(\cdot)$ equal to the density of Z. Moreover, in example c) we can smooth in one dimension and hence avoid the curse of dimensionality.

To estimate $I(h)$, we use the $U-$statistic

$$I_n(h) = \frac{1}{n(n-1)h^q} \sum_{1 \le i \ne j \le n} \left\langle \widehat{U_i \omega(Z_i)}, \ \widehat{U_j \omega(Z_j)} \right\rangle_{\mathcal{H}_1} K_{ij}(h)\ \phi_{ij},$$

where

$$K_{ij}(h) = K((Z_i - Z_j)/h), \qquad \phi_{ij} = \exp(-\|W_i - W_j\|^2_{\mathcal{H}_2}).$$

In example b) we take

$$\widehat{U_i\omega(Z_i)} = \frac{1}{n(n-1)} \sum_{k\neq i} (Y_i - Y_k) \frac{1}{g^q} L_{ik}(g),$$

where L is another kernel, $L_{ik}(g) = L((Z_i - Z_k)/g)$ and g is a bandwidth converging to zero at a suitable rate. The variance of $I_n(h)$ could be estimated by

$$v_n^2(h) = \frac{2}{n^2(n-1)^2 h^{2q}} \sum_{1\leq i\neq j\leq n} \left\langle \widehat{U_i\omega(Z_i)},\ \widehat{U_j\omega(Z_j)} \right\rangle^2_{\mathcal{H}_1} K_{ij}^2(h)\ \phi_{ij}^2.$$

Then, the test statistic is

$$T_n = \frac{I_n(h)}{v_n(h)}.$$

Under mild technical conditions we show that the test statistic is asymptotically standard normal under the null hypothesis $\mathbb{E}[U \mid Z, W] = 0$ a.s. We also show that it is consistent against fixed alternative and detect Pitman alternatives

$$H_{1n}:\ \mathbb{E}(U \mid Z, W) = r_n \delta(Z, W), \quad n \geq 1,$$

with probability tending to 1, provided that $r_n^2 n h^{q/2} \to \infty$.

References

[1] Aneiros-Pérez, G., Philippe Vieu, P. (2006), Semi-functional partial linear regression, *Statist. Prob. Lett.* 76, 1102–1110.

[2] Delsol, L., Ferraty, F., and Vieu, P. (2011), Structural test in regression on functional variables, *J. Multivariate Anal.* 102, 422–447.

[3] Faraway, J.J. (1997), Regression analysis for a functional response, *Technometrics* 39, 254–261.

[4] Ferraty, F., Laksaci, A., Tadj, A., and Vieu, P. (2011), Kernel regression with functional response, *Electron. J. Stat.* 5, 159–171.

[5] Ferraty, F., Van Keilegom, I., and Vieu, P. (2012), Regression when both response and predictor are functions, *J. Multivariate Anal.* 109, 10–28.

[6] García-Portugués, E., González-Manteiga, W., and Febrero-Bande, M. (2012), A goodness-of-fit test for the functional linear model with scalar response, arXiv:1205.6167v3 [stat.ME].

[7] Lavergne, P., Maistre, S., Patilea, V. (2012), A significance test for explanatory variables in nonparametric regression, Working Paper CREST-Ensai.

[8] Patilea, V., Sánchez-Sellero, C., and Saumard, M. (2012a), Nonparametric testing for no-effect with functional responses and functional covariates, arXiv:1209.2085 [math.ST].

[9] Patilea, V., Sánchez-Sellero, C., and Saumard, M. (2012b), Projection-based nonparametric goodness-of-fit testing with functional covariates, arXiv:1205.5578 [math.ST]

[10] Ramsay, J., and Silverman, B.W. (2005). *Functional Data Analysis* (2nd ed.). Springer-Verlag, New York.