X Congreso Galego de Estatística e Investigación de Operacións
Pontevedra, 3–4–5 de novembro de 2011

PROJECTION-BASED NONPARAMETRIC TESTING FOR FUNCTIONAL COVARIATE EFFECT

Valentin Patilea[1], César Sánchez Sellero[2] and Matthieu Saumard[3]

[1]CREST (Ensai) & IRMAR, France, patilea@ensai.fr
[2]Departamento de Estadística e Investigación Operativa,
Universidad de Santiago de Compostela, Spain, cesar.sanchez@usc.es,
[3]INSA-IRMAR, France, matthieu.saumard@insa-rennes.fr.

ABSTRACT

A regression setup is considered where one observes independent copies $(U_1, X_1), \cdots, (U_n, X_n)$ of (U, X) with U a scalar response and X a functional covariate taking values in the Hilbert space $L^2[0,1]$. The problem we address here is that of checking the null hypothesis $H_0 : E(U \mid X) = 0$, almost surely.

Our approach is an extension of the one proposed by Lavergne and Patilea (2008) and is based on the remark that H_0 is equivalent to the following statement: $\forall p \geq 1$, $E(U \mid \langle X^{(p)}, \gamma \rangle) = 0$, almost surely, for any γ a $p-$dimension vector of norm equal to 1. Here $X^{(p)}$ is the random vector of the first p coordinates of the decomposition of X in a given basis of $L^2[0,1]$ and $\langle \cdot, \cdot \rangle$ is the usual scalar product. The idea is to search for a least-favorable direction γ for H_0, where γ belongs to a finite-dimension space. Next, we will let p grow to infinity with the sample size.

These ideas are being extended to tests of a parametric funtional regression model, and to functional responses.

Keywords: Functional data; Nonparametric testing.

1. INTRODUCTION

Consider a sample of independent copies $(U_1, X_1), \cdots, (U_n, X_n)$ of (U, X) where U is a real-valued random variable and X is a random variable taking values in $L^2[0,1]$, that is the Hilbert space of the squared integrable real-valued functions defined on the unit interval.

The problem we investigate herein is the test of the hypothesis

$$H_0 : \ \mathrm{E}\,(U|X) = 0 \quad \text{almost surely (a.s.)} \tag{1}$$

against the nonparametric alternative $\mathrm{P}[\mathrm{E}\,(U|X) = 0] < 1$.

In this contribution we present the case where U is directly observed. The case where U is estimated as a residual of a parametric model, which leads to goodness-of-fit tests, will be studied in a forthcoming work.

The approach we consider is based on a dimension reduction idea used by Lavergne and Patilea (2008) in a finite dimension setup.

Other works on nonparametric testing in regression with functional covariate can be found in Cardot et al. (2003), Kokoszka et al. (2008) and Delsol et al. (2011).

2. DIMENSION REDUCTION IN NONPARAMETRIC TESTING

Let us introduce some notation. For any $p \geq 1$, let $\mathcal{S}^p = \{\gamma \in \mathbb{R}^p : \|\gamma\| = 1\}$ denote the unit hypersphere in $\mathbb{R}^p$. Let $\langle \cdot, \cdot \rangle$ denote the inner product in $L^2[0,1]$, that is for any $X_1, X_2 \in L^2[0,1]$

$$\langle X_1, X_2 \rangle = \int_0^1 X_1(t) X_2(t) dt,$$

and let $\| \cdot \|$ be the associated norm. Hereafter $\mathcal{R} = \{\rho_1, \rho_2, \cdots\}$ be an arbitrarily fixed orthonormal basis of the function space $L^2[0,1]$, that is $\langle \rho_i, \rho_j \rangle = \delta_{ij}$. Then any element of the predictor process X can be expanded into

$$X(t) = \sum_{j=1}^{\infty} x_j \rho_j(t)$$

where the random coefficients x_j are given by $x_j = \langle X, \rho_j \rangle$. We say that X is bounded if $\|X\|^2 = \sum_{j=1}^{\infty} x_j^2$ is a bounded random variable. For a fixed positive integer p, $X^{(p)} \in L^2[0,1]$ will be the projection of X on the subspace generated by the first p elements of the basis $\mathcal{R}$, that is

$$X^{(p)}(t) = \sum_{j=1}^{p} x_j \rho_j(t).$$

Let us notice that $\|X^{(p)}\|$ coincides with the Euclidean norm of the vector $(x_1, \cdots, x_p)$ in $\mathbb{R}^p$. By abuse we also identify $X^{(p)}$ with the $p-$dimension random vector $(x_1, \cdots, x_p)$. On the other hand, for any integer $p > 1$ and non random vector $\gamma = (\gamma_1, \cdots, \gamma_p) \in \mathbb{R}^p$, we consider by abuse γ an element in $L^2[0,1]$ with $(\gamma_1, \cdots, \gamma_p, 0, 0, \cdots)$ the coefficients of its expansion and hence $\langle X, \gamma \rangle = \langle X^{(p)}, \gamma \rangle = \sum_{i=1}^{p} x_j \gamma_j$. In the following we will also use $\beta = \sum_{j=1}^{\infty} b_j \rho_j(t)$ to denote a non random element of $L^2[0,1]$.

Our approach relies on the following lemma, an extension of Lemma 2.1 of Lavergne and Patilea (2008) to Hilbert space-valued conditioning random variables. The result shows that for checking nullity of a conditional expectation, it is equivalent to consider expectations conditional on X and expectations conditional on $L^2[0,1]$ projections of X on a sufficiently rich set of directions.

Lemma 1 *Let $X \in L^2[0,1]$ and $Z \in \mathbb{R}$ be random variables. Assume that X is bounded, $E|Z| < \infty$ and $E(Z) = 0$.*

(A) The following statements are equivalent:

1. $E(Z \mid X) = 0$ *a.s.*

2. $E(Z \mid \langle X, \beta \rangle) = 0$ *a.s.* $\forall \beta \in L^2[0,1]$ *with* $\|\beta\| = 1$.

3. *for any integer* $p \geq 1$, $E(Z \mid \langle X, \gamma \rangle) = 0$ *a.s.* $\forall \gamma \in \mathcal{S}^p$.

4. *for any integer* $p \geq 1$, $E(Z \mid X^{(p)}) = 0$ *a.s.*

(B) If $P[E(Z \mid X) = 0] < 1$*, then there exists a positive integer* $p_0 \geq 1$ *such that for any integer* $p \geq p_0$*, the set*

$$\{\gamma \in \mathcal{S}^p : E(Z \mid \langle X, \gamma \rangle) = 0 \text{ a.s.}\}$$

has Lebesgue measure zero on the unit hypersphere $\mathcal{S}^p$ *and is not dense.*

Our Lemma 1 readily yields new formulations of H_0.

Corollary 1 *Consider a* $L^2[0,1]-$*valued random variable* U *such that* $\mathbb{E}\|U(\theta)\| < \infty$. *Let* $\omega(\beta, t)$, $\beta \in L^2[0,1]$ *and* $t \in \mathbb{R}$*, be a function such that* $\omega(\beta, \langle X, \beta \rangle) > 0$ *for all* $\|\beta\| = 1$. *For any* $p \geq 1$*, let* $w_p(\gamma, t)$, $\gamma \in \mathbb{R}^p$ *and* $t \in \mathbb{R}$*, be a function such that* $w_p(\gamma, \langle X, \gamma \rangle) > 0$ *for all* $\|\gamma\| = 1$. *The following statements are equivalent:*

1. *The null hypothesis (1) holds true.*

2. $$\max_{\beta \in L^2[0,1],\ \|\beta\|=1} \mathbb{E}\left[U \mathbb{E}\left(U | \langle X, \beta \rangle\right) \omega(\beta, \langle X, \beta \rangle)\right] = 0. \tag{2}$$

3. *for any* $p \geq 1$ *and any set* $B_p \subset \mathcal{S}^p$ *with strictly positive Lebesgue measure in on the unit hypersphere* $\mathcal{S}^p$,
$$\max_{\gamma \in B_p} \mathbb{E}\left[U \mathbb{E}\left(U | \langle X, \gamma \rangle\right) w_p(\gamma, \langle X, \gamma \rangle)\right] = 0. \tag{3}$$

3. TESTING THE EFFECT OF A FUNCTIONAL COVARIATE

We introduce a general approach for nonparametric testing of the effect of a functional covariate X on a real-valued random variable U. For simplicity, here we assume that $\mathbb{E}(U) = 0$. Our approach is based on Corollary 1-(3) and *univariate* kernel smoothing. In this way we avoid the problem of smoothing in infinite-dimension.

To avoid handling denominators close to zero, we set the weight function $\omega(\gamma, \cdot)$ in Corollary 1 equal to the density of $\langle X, \gamma \rangle$, denoted by $f_\gamma(\cdot)$, which is assumed to exist for any γ. For any $\gamma \in \mathbb{R}^p$, let

$$Q(\gamma) = \mathbb{E}\{U\ \mathbb{E}[U \mid \langle X, \gamma \rangle] f_\gamma(\langle X, \gamma \rangle)\} = \mathbb{E}\{\mathbb{E}^2[U \mid \langle X, \gamma \rangle] f_\gamma(\langle X, \gamma \rangle)\}.$$

For any $p \geq 1$, let $B_p \subset \mathcal{S}^p$ be a set with strictly positive Lebesgue measure in $\mathcal{S}^p$. By Corollary 1, the null hypothesis (1) holds true if and only if

$$\forall p \geq 1, \quad \max_{\gamma \in B_p} Q(\gamma) = 0. \tag{4}$$

In view of equation (4), our goal is to estimate $Q(\gamma)$. With at hand a sample of (U, X), define

$$Q_n(\gamma) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} U_i U_j \frac{1}{h} K_h \left(\langle X_i - X_j, \gamma \rangle \right), \quad \gamma \in \mathcal{S}^p,$$

where $K_h(\cdot) = K(\cdot/h)$, where $K(\cdot)$ is a kernel and h a bandwidth. In the case of finite dimension covariates, the function $\gamma \mapsto Q_n(\gamma)$ is the statistic considered by Lavergne and Patilea (2008), see also Bierens (1990). For fixed p and $\gamma \in \mathcal{S}^p$, it is well-known that $Q_n(\gamma)$ has asymptotic centered normal distribution with rate $nh^{1/2}$ under H_0; see for instance Guerre and Lavergne (2005). We will show that the asymptotic normal distribution is preserved even when p grows at a suitable rate with the sample size.

For a fixed p the statistic $Q_n(\gamma)$ is expected to be close to $Q(\gamma)$ uniformly in γ. Then a natural idea would be to build a test statistic using the maximum of $Q_n(\gamma)$ with respect to γ. We will choose a direction γ as the least favorable direction for the null hypothesis H_0 obtained from a penalized criterion based on a standardized version of $Q_n(\gamma)$. Lavergne and Patilea (2008) and Bierens (1990) considered this idea using $Q_n(\gamma)$. Here we use a standardized version of $Q_n(\gamma)$, that is,

$$nh^{1/2} \frac{Q_n(\gamma)}{\widehat{v}_n(\gamma)}$$

where $\widehat{v}_n^2(\cdot)$ is a suitable estimator of the variance of $nh^{1/2} Q_n(\cdot)$.

Now, fix some $\beta_0 \in L^2[0,1]$ that could be interpreted as an initial *guess* of an unfavorable direction for H_0. Let b_{0j}, $j \geq 1$, be the coefficients in the expansion of β_0 in the basis $\mathcal{R}$. For any given $p \geq 1$ such that $\sum_{j=1}^p b_{0j}^2 > 0$, let

$$\gamma_0^{(p)} = (b_{01}, \cdots, b_{0p}) \Big/ \sqrt{\sum_{j=1}^p b_{0j}^2} \ .$$

Given $B_p \subset \mathcal{S}^p$ with strictly positive Lebesgue measure in $\mathcal{S}^p$ that contains $\gamma_0^{(p)}$, the least favorable direction γ for H_0 is defined as

$$\widehat{\gamma}_n = \arg\max_{\gamma \in B_p} \left[nh^{1/2} Q_n(\gamma)/\widehat{v}_n(\gamma) - \alpha_n \mathbb{I}_{\left\{\gamma \neq \gamma_0^{(p)}\right\}} \right], \tag{5}$$

where $\mathbb{I}_A$ is the indicator function of a set A, and α_n, $n \geq 1$ is a sequence of positive real numbers decreasing to zero at an appropriate rate that depends on the rates of h and p and that will be made explicit below. Let us notice that the maximization used to define $\widehat{\gamma}_n \in \mathcal{S}^p$ is a finite dimension optimization problem. The choice of β_0, and thus of $\gamma_0^{(p)}$, is theoretically irrelevant, it does not affect the asymptotic critical values and the consistency results. However, in practice the choice of β_0 could be related to a priori information of the practitioner on a class of alternatives, like for instance the class of functions depending only on $\langle X, \beta_0 \rangle$.

We have proven that with suitable rates of increase for α_n and p and decrease for h, the probability of the event $\{\widehat{\gamma}_n = \gamma_0^{(p)}\}$ tends to 1 under H_0. Hence $Q_n(\widehat{\gamma}_n)/\widehat{v}_n(\widehat{\gamma})$ behaves asymptotically like $Q_n(\gamma_0^{(p)})/\widehat{v}_n(\gamma_0^{(p)})$, even when p grows with the sample size. Therefore the test statistic we consider is

$$T_n = nh^{1/2}\frac{Q_n(\widehat{\gamma}_n)}{\widehat{v}_n(\widehat{\gamma}_n)}\ . \tag{6}$$

and its asymptotic distribution is standard normal.

Regarding the variance estimation, $\widehat{v}_n^2(\cdot)$, there are several possibilities to face this problem. In the case of high dimension covariates, a simple and convenient solution is given by

$$\widehat{\tau}_n^2\,(\gamma) = \frac{2}{n(n-1)h}\sum_{j\neq i} U_i^2 U_j^2 K_h^2\left(\langle X_i - X_j, \gamma\rangle\right). \tag{7}$$

A bootstrap procedure is proposed to approximate the critical values of the test statistic.

We also show how the testing procedure can be extended to the case of parametric functional models like linear model and generalized linear models (see Müller and Stadtmüller (2005)).

Simulations will also be shown where the performance of the test is evaluated. Practical advices will be provided about the parameters to be chosen in the procedure. Special attention will be paid to the basis and the direction β_0.

4. ACKNOWLEDGEMENTS

The second author acknowledges support coming from Spanish Ministry of Science through Project MTM2008-03010.

REFERENCES

Bierens, H.J. (1990) A consistent conditional moment test of functional form. Econometrica, 58, 1443-1458.

Cardot, H., Ferraty, F., Mas, A. and Sarda, P. (2003). Testing hypotheses in the functional linear model. Scandinavian Journal of Statistics, 30, 241-255.

Delsol, L., Ferraty, F. and Vieu, P. (2011). Structural test in regression on functional variables. Journal of Multivariate Analysis, 102, 422-447.

Guerre, E. and Lavergne, P. (2005). Data-driven rate-optimal specification testing in regression models. The Annals of Statistics, 33, 840-870.

Kokoszka, P., Maslova, J., Sojka, J. and Zhu, L. (2008). Testing for lack of dependence in the functional linear model. The Canadian Journal of Statistics, 36, 116.

Lavergne, P. and Patilea, V. (2008). Breaking the curse of dimesionality in nonparametric testing. Journal of Econometrics, 143, 103-122.

Müller, H.G. and Stadtmüller, U. (2005). Generalized functional linear models. The Annals of Statistics, 33, 774-805.