

Mathematics

On the Regression Estimation in a Hilbert Space

Elizbar Nadaraya^{*}, Petre Babilua^{**}, Grigol Sokhadze^{**}

^{*} Academy Member, I. Javakhishvili Tbilisi State University

^{**} I. Javakhishvili Tbilisi State University

ABSTRACT. The problem of estimation of a regression curve in an infinite dimensional space is considered. Such problems arise in the statistics of random processes. Observation data in such processes is a pair whose one component is an element of an infinite dimensional space. It is shown that the finite dimensional projections of observations and the regression curve estimators constructed on their basis give an approximation of a regression function in the initial space. The method of infinite dimensional analysis and in particular the notion of the logarithmic gradient of a distribution function are used. © 2010 Bull. Georg. Natl. Acad. Sci.

Key words: regression, statistical estimators, Hilbert space.

Many problems of the theory of random process statistics lead to the necessity to estimate the relations of trajectories with the functional of this process or another by means of observation data. Hence it is interesting to investigate the procedure of regression estimation by observation data on a process-random value pair. Such problems are naturally formulated in terms of Hilbert spaces. They are well studied for the finite dimensional case. Nadaraya-Watson estimators are the best known ones for a regression function. Vast literature on this topic is available, see, for example, [1-4]. As to the infinite dimensional analogue of these estimators, these problems are in the stage of development. An interesting approach is presented in [5, 6].

In the present paper we indicate one relation of statistical estimation of a regression function with problems of infinite dimensional analysis, in particular with the notion of a logarithmic derivative of a measure. Earlier, in [7] this relation was used for the construction of a statistical estimator of a logarithmic derivative in a Hilbert space. By modifying this construction procedure we can construct estimators for a regression function in an infinite dimensional space.

Let $\{H, \mathcal{R}\}$ be a measurable Hilbert space. Here H is assumed to be real and separable. Scalar products and norms will be provided with subscripts of respective spaces. \mathcal{R} is the Borel σ -algebra of subsets. $\{\Omega, \mathcal{J}, P\}$ is a probability space with complete measure. Let $\xi: \Omega \rightarrow H$ and $\eta: \Omega \rightarrow H$ be random elements with values in H .

We say that the part (ξ, η) satisfies the integration by parts formula (briefly, $IP(\xi, \eta)$) if there exists an integrable random value $L(\xi, \eta)$ such that for any bounded smooth functional $\varphi \in C_b^\infty(H)$ the following equality (integration by parts formula) is fulfilled:

$$E(\varphi'(\xi), \eta)_H = E\varphi(\xi)L(\xi, \eta). \quad (1)$$

From this definition it is easy to see that $L(x,y)$ is a linear mapping with respect to the second argument. As an example we can consider the random elements $\xi = \zeta$ and $\eta = g(\zeta)$, where ζ is a Gaussian element in H with zero mean and kernel type correlation operator B and $g : H \rightarrow H$ is a sufficiently smooth function. Then, following [8], we can calculate

$$L(\xi, \eta) = (R^{-1}g(\zeta), \zeta)_H - Trg'(\zeta).$$

Let $h \in H$ be a fixed vector, and η be a random value. Then, by virtue of (1), for any $\varphi \in C_b^\infty(H)$ we can write

$$E(\varphi'(\xi), \eta h)_H = E\varphi(\xi)L(\xi, \eta h) \tag{2}$$

and say that the pair (ξ, η) satisfied the integration by parts formula (briefly, $IP_h(\xi, \eta)$).

If $H = \mathbb{R}^n$, then, by the well-known Malyavin's formula ([9,10]), from (2) we can derive the smoothness properties of the distribution ξ . In particular, for $\eta = 1$ there exists a distribution density of the vector ξ with respect to the Lebesgue measure and it is continuous. Also, the representation

$$p_\xi(x) = E\chi_{I(x)}L(\xi, h) \tag{3}$$

is valid, where χ is the event indicator and $I(x) = \prod_{i=1}^n [x_i, \infty)$.

If a random vector ξ and a random value η are such that $IP_h(\xi, \eta)$ and $IP_h(\xi, 1)$ take place, then a regression curve can be represented in the form

$$r(x) = E(\eta | \xi = x) = \frac{E(\chi_{I(x)}(\xi)L(\xi, \eta h))}{E(\chi_{I(x)}(\xi)L(\xi, h))}. \tag{4}$$

Using (3), the regression curve can be approximated in the Hilbert space. Assume that we have a sampling $(X_1, Y_1), (X_2, Y_2), \dots, (X_m, Y_m)$ of independent and equally distributed random pairs from (ξ, η) , where $X_k \in H$ and $Y_k \in \mathbb{R}$, $k = 1, 2, \dots, m$.

Denote by U_n an orthogonal finite dimensional projector in H . Let $H_n = U_n H$, $h_n = U_n h$, $h \in H$, $\xi_n = U_n \xi$, $x_n = U_n x$. Let further $X_1^{(n)}, X_2^{(n)}, \dots, X_m^{(n)}$ be the projections of observed values $X_k^{(n)} = U_n X_k$, $k = 1, 2, \dots, m$, and construct the pairs $(X_1^{(n)}, Y_1), (X_2^{(n)}, Y_2), \dots, (X_m^{(n)}, Y_m)$.

If we assume that $IP_h(\xi, \eta)$ and $IP_h(\xi, 1)$ are valid, then it can be easily verified that $IP_{h_n}(\xi_n, \eta)$ and $IP_{h_n}(\xi_n, 1)$ are valid, too, for all n . By virtue of formula (3), for the projection we can write the regression equation

$$r_n(x_n) = E(\eta | \xi_n = x_n) = \frac{E(\chi_{I(x_n)}(\xi_n)L(\xi_n, \eta h_n))}{E(\chi_{I(x_n)}(\xi_n)L(\xi_n, h_n))}. \tag{5}$$

Theorem 1. Assume that in a separable real Hilbert space H we have a sampling $(X_1, Y_1), (X_2, Y_2), \dots, (X_m, Y_m)$ of independent and equally distributed random pairs from (ξ, η) , where $X_k \in H$ and $Y_k \in \mathbb{R}$, $k = 1, 2, \dots, m$. If $IP_h(\xi, \eta)$ and $IP_h(\xi, 1)$ are valid, then $r_n(x_n)$ defined by formula (4) converges to $r(x) = E(\eta | \xi = x)$.

This theorem enables one to construct Nadaraya-Watson type estimators of a regression curve using the approach from [7]. Let $K(x)$ be a continuously differentiable, everywhere positive function and $\int_R K(x)dx = 1$. We will symbolically write $K \in CL$. Let $X_{ij}^{(n)}$, $j = 1, \dots, n$, be the components of the vector $X_i^{(n)}$, $i = 1, \dots, m$; x_i^j , $j = 1, \dots, n$, be the components of the vector x_i , $i = 1, \dots, m$. Analogously, h_i^j , $j = 1, \dots, n$, be the components of the vector h_i , $i = 1, \dots, m$. Taking into account the fact that the functional $L(\xi, \eta h)$ is the logarithmic derivative of the joint distribution along h , by the method of [7] we construct the regression estimator

$$\hat{r}_n^m(x_n) = \frac{E \sum_{i=1}^m Y_i \sum_{s=1}^n h_m^s K'(\lambda_m(x_n^s - X_{is}^{(n)})) \prod_{\substack{j=1 \\ j \neq s}}^n K(\lambda_m(x_n^j - X_{ij}^{(n)}))}{E \sum_{i=1}^m \sum_{s=1}^n h_m^s K'(\lambda_m(x_n^s - X_{is}^{(n)})) \prod_{\substack{j=1 \\ j \neq s}}^n K(\lambda_m(x_n^j - X_{ij}^{(n)}))}, \quad (6)$$

where $\{\lambda_m\}_{m=1}^\infty$ is some diverging numerical sequence.

After the transformation of (6) we obtain

$$\hat{r}_n^m(x_n) = \frac{\sum_{i=1}^m Y_i \prod_{j=1}^n K(\lambda_m(x_n^j - X_{ij}^{(n)}))}{\sum_{i=1}^m \prod_{j=1}^n K(\lambda_m(x_n^j - X_{ij}^{(n)}))}. \quad (7)$$

The following statement is valid.

Theorem 2. Let $IP_h(\xi, \eta)$ and $IP_h(\xi, 1)$ be valid. If $K \in CL$, $\lambda_n \rightarrow \infty$, $\frac{\lambda_n^2 \ln n}{n} \rightarrow 0$ as $n \rightarrow \infty$, then estimator (6) converges uniformly, with probability 1, to $r(x) = E[\eta | \xi = x]$.

The proof of Theorems 1 and 2 is based on the martingale property of the random sequence (6) with respect to an increasing sequence of Borel σ -algebras with respect to the variable n , while with respect to the variable m \hat{r}_n^m is a finite dimensional estimator of the regression function r_n .

მათემატიკა

რეგრესიის შეფასებისათვის ჰილბერტის სივრცეში

ე. ნადარაია*, პ. ბაბილუა**, გ. სოხაძე**

* აკადემიის წევრი, ი. ჯავახიშვილის სახ. თბილისის სახელმწიფო უნივერსიტეტი

** ი. ჯავახიშვილის სახ. თბილისის სახელმწიფო უნივერსიტეტი

ნაშრომში განხილულია რეგრესიის წირის შეფასების საკითხი უსასრულოგანზომილებიან სივრცეში. საკითხის ასეთი დასმა ხშირად გვხვდება შემთხვევითი პროცესების სტატისტიკაში. ასეთ ამოცანებში შერჩევა წყვილია, რომელთაგან ერთ-ერთი უსასრულოგანზომილებიანი სივრცის ელემენტია. ნაჩვენებია, რომ დაკვირვებათა სასრულგანზომილებიანი პროექციების საფუძველზე აგებული რეგრესიის წირის შეფასება კარგ აპროქსიმაციას აკეთებს ძირითადი სივრცის რეგრესიის ფუნქციისათვის. გამოიყენება უსასრულოგანზომილებიანი ანალიზის მეთოდები, კერძოდ, განაწილების ლოგარითმული გრადიენტის ცნება.

REFERENCES

1. *E. Nadaraya* (1989), *Nonparametric Estimation of Probability Densities and Regression Curves*. Dordrecht.
2. *J. S. Marron* (1986), *J. Multivariate Anal.*, **20**, No. 1: 91-113.
3. *L. Devroye, L. Györfi, A. Krzyżak* (1998), *J. Multivariate Anal.*, **65**, 2: 209-227.
4. *G. Collomb* (1981), *Internat. Statist. Rev.*, **49**, 1: 75-93.
5. *S. Dabo-Niang* (2004), *Appl. Math. Lett.*, **1**, 4: 381-386.
6. *S. Dabo-Niang, F. Ferraty, Ph. Vieu* (2004), *C. R. Math. Acad. Sci. Paris*, **339**, 9: 659-662.
7. *E. A. Nadaraya, G. A. Sokhadze, A. D. Shatashvili* (2009), *Kibernetika i Sistemnyi Analiz*, **5**: 106-110.
8. *Yu. L. Daletski, S. V. Fomin* (1983), *Mery i differentsial'nye uravneniya v beskonechnomernykh prostranstvakh*. Moscow, 383 p. (in Russian).
9. *P. Malliavin* (1997), *Stochastic Analysis. Grundlehren der Mathematischen Wissenschaften*. Berlin, Heidelberg, New York, 313.
10. *M. Sanz-Sole* (2003), *Lecture Notes*. **2**, Barcelona, 128 p.

Received November, 2009