# Nonparametric Algorithms of Identification and Prediction in the ARX-Models

Gennady M. Koshkin, Vadim Yu. Lukov, Iosif G. Piven
Tomsk State University
Department of Applied Mathematics and Cybernetics
Tomsk, Russia
kgm@mail.tsu.ru, lukov_vadim@rambler.ru, josef__@mail.ru

**Abstract - To identify an unknown function defining of a nonlinear ARX-process, we use kernel regression estimators. The principal parts of mean square errors for these estimators are found. The proposed algorithms are applied to the real data processing.**

***Keywords—ARX-processes, kernel estimators, regression function***

## I. INTRODUCTION

Let the random sequence $Y_t$ be generated by the model [1]

$$Y_t = \psi(Y_{t-1},...,Y_{t-k}, X_{t-1},...,X_{t-q}) + \varepsilon_t, \quad t = 1,...,n, \quad (1)$$

where $Y_t$ is an output variable at the moment $t$, $X_t$ is the random exogenous factor, which independent of $\varepsilon_t$, $\varepsilon_t$ is a sequence of independent and identically distributed random variables with zero mean and bounded variance, $\psi$ is an unknown function.

Denote

$$Y_{t,k} = (Y_{t-1},...,Y_{t-k}), \quad X_{t,q} = (X_{t-1},...,X_{t-q}). \quad (2)$$

The model (1) with $k = 1$ is given in [2]. Let the function $\psi$ be bounded, the distribution density of $\varepsilon_t$ on $R^1$ be nonnegative with $E\varepsilon_1^3 = 0$ and $E\varepsilon_1^4 < \infty$. In this case, the process (1) is the geometric Markov chain with the strong mixing coefficient [3-5]

$$\alpha(\tau) \le c_0 \rho_0^{\tau}, \quad 0 < \rho_0 < 1, \quad c_0 > 0. \quad (3)$$

Then, we can estimate the function $\psi$ by the following analog of Nadaraya-Watson estimator [6-9]

$$\psi_n(z) = \frac{\sum_{t=1}^{n} Y_i \cdot K\left(\frac{z - Z_t}{h_n}\right)}{\sum_{t=1}^{n} K\left(\frac{z - Z_t}{h_n}\right)}, \quad (4)$$

or by the piecewise-smooth approximation [9-11]

$$\tilde{\psi}_n(z) = \frac{\psi_n(z)}{(1 + \delta_n |\psi_n(z)|^{\tau})^{\rho}}, \quad (5)$$

where $K\left(\frac{y - Z_t}{h_n}\right)$ is a $(k+q)$-dimensional kernel, (the product of one-dimensional densities of the standard normal law), $h_n \in R^1$ is a bandwidth parameter, $h_n \downarrow 0$, $\delta_n \downarrow 0$, $\tau > 0$, $\rho > 0$, $\tau\rho \ge 1$, $Z_t = (Y_{t-1},...,Y_{t-k}, X_{t-1},...,X_{t-q})$.

## II. ASYMPTOTIC PROPERTIES OF ESTIMATORS

Let $(Y_t)_{t=1,...,n}$ be observations generated by process (1).

It is well known that the regression function can be used as a model for $\psi$ in (1):

$$r(y,x) = E(Y_t \mid Y_{t,k} = y, X_{t,q} = x) = E(Y_t \mid y, x) = \\ = \frac{a(y,x)}{p(y,x)} = \int Y_t f(Y_t \mid y, x) dY_t, \quad (6)$$

where $(y,x) \in R^{k+q}$, $a(y,x) = \int z f(z,y,x) dz$ is the basic functional, $p(y,x)$ is a density function of $(Y_{t,k}, X_{t,q})$, $f(z,y,x)$ is an unknown density function of $(Y_t, Y_{t,k}, X_{t,q})$ [5].

Introduce the following notation:

$$h_n^{\langle y \rangle} = \left(h_{n,1}^{\langle y \rangle},...,h_{n,k}^{\langle y \rangle}\right), \quad h_n^{\langle x \rangle} = \left(h_{n,1}^{\langle x \rangle},...,h_{n,q}^{\langle x \rangle}\right),$$

This work was supported by Russian Foundation for Basic Research, project 13-08-00744, and the TSU Competitiveness Improvement Program.

978-1-4673-9941-8/16 $31.00 © 2016 IEEE
DOI 10.1109/SMRLO.2016.109

$$K_k\left(\frac{y-Y_{i,k}}{h_n^{\langle y\rangle}}\right)=K\left(\frac{y_1-Y_{i,1}}{h_{n,1}^{\langle y\rangle}}\right)\times\ldots\times K\left(\frac{y_k-Y_{i,k}}{h_{n,k}^{\langle y\rangle}}\right),$$

$$K_q\left(\frac{x-X_{i,q}}{h_n^{\langle x\rangle}}\right)=K\left(\frac{x_1-X_{i,1}}{h_{n,1}^{\langle x\rangle}}\right)\times\ldots\times K\left(\frac{x_k-X_{i,q}}{h_{n,q}^{\langle x\rangle}}\right).$$

As nonparametric estimators for $a(y,x)$ and $r(y,x)$ at the point $(y,x)$, we take the statistics:

$$a_n(y,x)=\frac{1}{n-s}\sum_{i=s+1}^{n}Y_i\frac{1}{\prod_{j=1}^{k}h_{nj}^{\langle y\rangle}}K_k\left(\frac{y-Y_{i,k}}{h_n^{\langle y\rangle}}\right)\times \times\frac{1}{\prod_{j=1}^{q}h_{nj}^{\langle x\rangle}}K_q\left(\frac{x-X_{i,q}}{h_n^{\langle x\rangle}}\right), \quad (7)$$

$$r_n(y,x)=\psi_n=\frac{\sum_{i=s+1}^{n}Y_i\frac{1}{\prod_{j=1}^{k}h_{nj}^{\langle y\rangle}}K_k\left(\frac{y-Y_{i,k}}{h_n^{\langle y\rangle}}\right)\frac{1}{\prod_{j=1}^{q}h_{nj}^{\langle x\rangle}}K_q\left(\frac{x-X_{i,q}}{h_n^{\langle x\rangle}}\right)}{\sum_{i=s+1}^{n}\frac{1}{\prod_{j=1}^{k}h_{nj}^{\langle y\rangle}}K_k\left(\frac{y-Y_{i,k}}{h_n^{\langle y\rangle}}\right)\frac{1}{\prod_{j=1}^{q}h_{nj}^{\langle x\rangle}}K_q\left(\frac{x-X_{i,q}}{h_n^{\langle x\rangle}}\right)}, \quad (8)$$

where $s=\max(k,q)$ [1].

The quality of estimators will characterize by the MSE (mean square error):

$$u^2(a_n)=u^2(a_n(y,x))=E(a_n-a)^2=D(a_n)+b^2(a_n), \quad (9)$$

where $D(a_n)$ is the variance and $b(a_n)$ is the bias of $a_n$.

According to [12-15], asymptotical unbiasedness of $r_n(y,x)$ was proved, and also the optimal estimators and the optimal MSE were found.

## III. Real Data Processing

Consider the stock prices of the company "Gazprom" from April 2014 to April 2015. In the framework of model (1), let dollar rate, the oil and gas prices be exogenous factors.

First, the case when there are no exogenous factors, i.e. (1) is the AR-model [4,13,15], is studied.

The quality of models will characterize by the relative error of identification ($\delta$) and the average error of forecasting ($\varepsilon$):

$$\delta=\frac{1}{n}\cdot\sum_{i=1}^{n}\frac{|Y_i-\psi_n|}{Y_i}\cdot 100\%, \quad (10)$$

$$\varepsilon=\frac{1}{30}\cdot\sum_{i=n+1}^{n+30}|Y_i-\psi_n|, \quad (11)$$

Analysis of AR-models show that the best model is AR(1)-model. The results are presented in Table 1.

TABLE I. Comparison of Models AR(1)- AR(4)

| Models | AR(1) | AR(2) | AR(3) | AR(4) |
|---|---|---|---|---|
| Optimal Bandwidths | 1.3 | 1.4<br>9.5 | 1.6<br>9.7<br>10.3 | 0.6<br>18.3<br>8<br>6.4 |
| Relative Errors of Identification, % | 1.87 | 1.937 | 2.032 | 2.201 |
| Average Errors of Forecasting, Rubles | 1.822 | 1.889 | 1.874 | 1.878 |

Then, consider ARX-model [1,2] with a dependence on the dollar rate:

$$\psi_n\left(h_n^{\langle y\rangle},h_n^{\langle x\rangle}\right)=\frac{\sum_{j\geq 2}Y_j\cdot K\left(\frac{y-Y_{j-1}}{h_n^{\langle y\rangle}}\right)\cdot K\left(\frac{x-X_{j-1}}{h_n^{\langle x\rangle}}\right)}{\sum_{j\geq 2}K\left(\frac{y-Y_{j-1}}{h_n^{\langle y\rangle}}\right)\cdot K\left(\frac{x-X_{j-1}}{h_n^{\langle x\rangle}}\right)}. \quad (12)$$

The optimal bandwidths $h^{\langle y\rangle}$ and $h^{\langle x\rangle}$ were obtained by minimizing the relative error of identification

$$\varepsilon(h)=\frac{1}{30}\cdot\sum_{i=n-30}^{n}\frac{|Y_i-\psi_n|}{Y_i}\cdot 100\%. \quad (13)$$

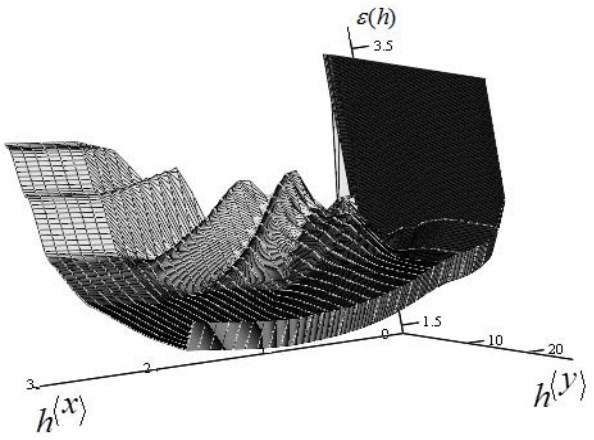


Fig. 1. Dependence of the error ε on the bandwidths $h^{\langle y\rangle}$, $h^{\langle x\rangle}$.

In this case, the minimum error is achieved at $h^{\langle y \rangle} = 18.9$, $h^{\langle x \rangle} = 2$ (see Fig. 1). The real prices and the results of nonparametric identification are given in Fig. 2.

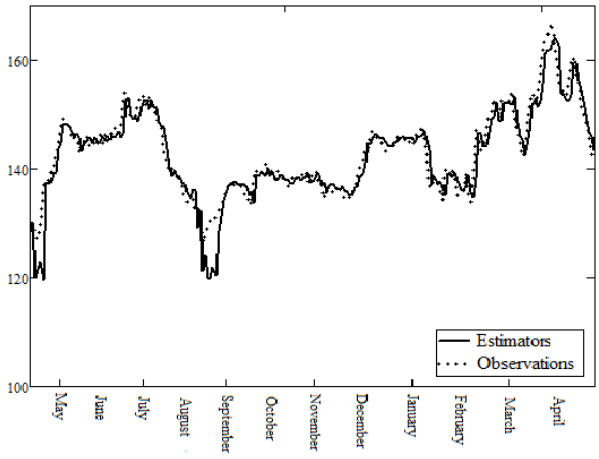


Fig. 2. Observations $X_i$ and their estimators $\psi_n(2, 18.9)$.

The large errors in August 2014 can be explained by the sharp fall of oil prices (about 4% per month in comparison with 1% for the period January-July).

Further, consider a model with a dependence on the oil prices. Here, the minimum error is achieved at $h^{\langle y \rangle} = 13.5$, $h^{\langle x \rangle} = 0.9$ (see Fig. 3).

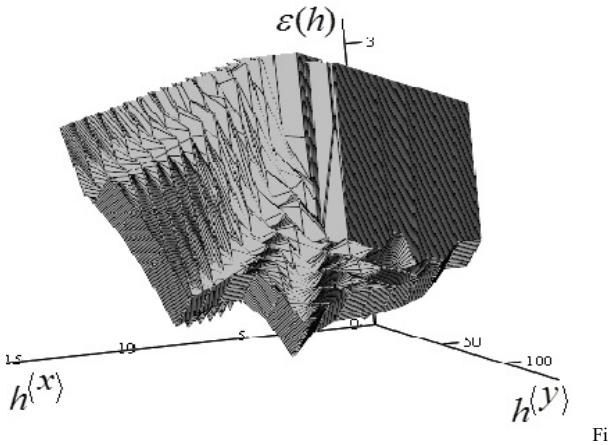


Fig. 3. Dependence of the error ε on the bandwidths $h^{\langle y \rangle}$, $h^{\langle x \rangle}$.

A comparison of the real prices and results of nonparametric identification is presented in Fig. 4.

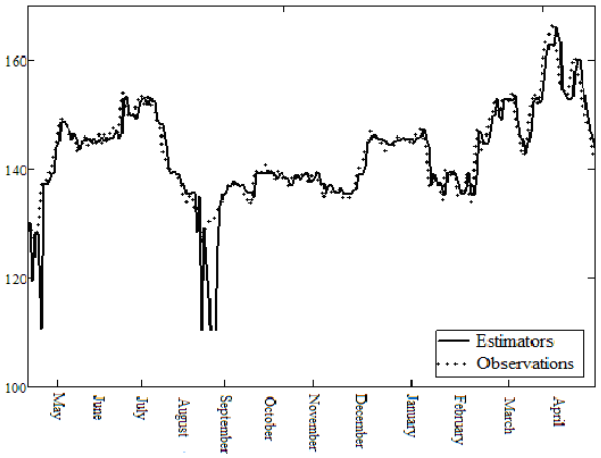


Fig. 4. Observations $X_i$ and their estimators $\psi_n(0.9, 13.5)$.

At last, consider a model with a dependence on the gas prices. It is seen, the minimum error is achieved at $h^{\langle y \rangle} = 7.3$, $h^{\langle x \rangle} = 1.1$ (see Fig. 5).

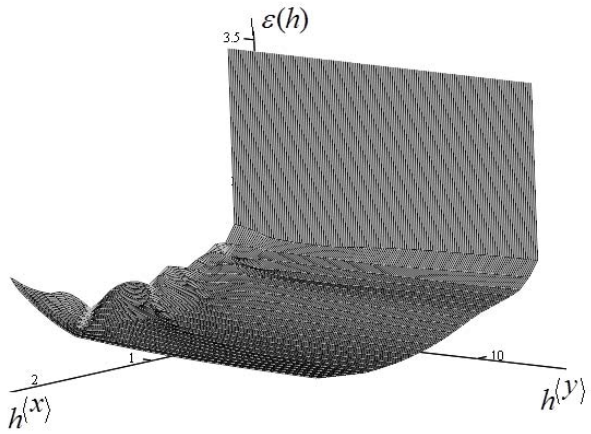


Fig. 5. Dependence of the error ε on the bandwidths $h^{\langle y \rangle}$ and $h^{\langle x \rangle}$.

A graphical comparison of the real prices and results of nonparametric identification is presented in Fig. 6.

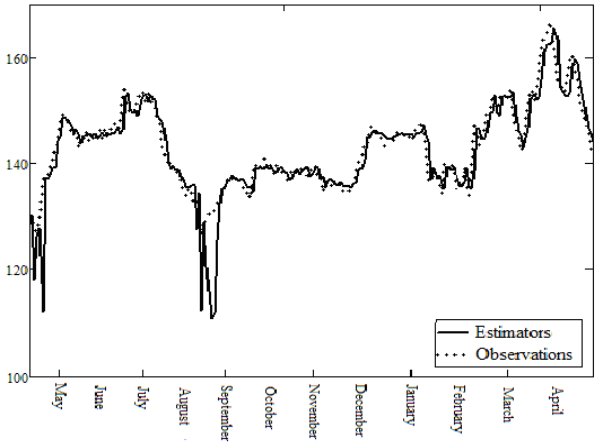


Fig. 6. Observations $X_i$ and their estimators $\psi_n$ (1.1, 7.3).

We can see that the model with the USD rate is the best both in the sense of minimizing the relative error and in the sense of minimizing the average error (results are given in Table 2).

TABLE II. COMPARISON OF MODELS

| Models | AR(1) | USD | OIL | GAS |
|---|---|---|---|---|
| Optimal Bandwidths | 1.3 | 2<br>18.9 | 0.9<br>13.5 | 1.1<br>7.3 |
| Relative Errors of Identification, % | 1.87 | 1.745 | 1.948 | 1.919 |
| Average Errors of Forecasting, Rubles | 1.84 | 1.687 | 2.102 | 1.998 |

## IV. CONCLUSION

The paper deals with the problem of estimating the unknown function defining a nonlinear ARX-process with making use of kernel regression estimators. The asymptotic properties of such estimators, namely, unbiasedness and consistency are proved. In addition, we found the main parts of the asymptotic MSEs of the estimators.

In the frameworks of AR-model and ARX-model, the stock prices of the company "Gazprom" from April 2014 to April 2015 were studied. Analysis of AR-models show that the best model is AR(1)-model. Also, ARX-model with the exogenous factor of the dollar rate is better than ARX-models with the exogenous factors of oil or gas prices.

Note that the considered algorithms can be used on the nonparametric identification and prediction of static and dynamic production functions [16,17]. The improved estimators of identification and prediction can be obtained using a priori information (see [18,19]).

## REFERENCES

[1] I. Glukhova and G. Koshkin, "Non-parametric identification of nonlinear ARX-processes," Tomsk State University. Journal of Control and Computer Science, vol. 20, 3, pp. 55-61, 2012.

[2] A. Georgiev, "Nonparametric system identification by kernel methods," IEEE Transactions on Automatic Control, vol. AC-29, pp. 356-358, 1984.

[3] P. Ango Nze, "Criteres d'ergodicite de quelques modeles a representation marcovienne," C. R. Acad. Sci. Paris. Serie I, 315, pp. 1301-1304, 1992.

[4] W. Hardle, A. Tsybakov, and L. Yang, "Nonparametric vector autoregression," Journal of Statistical Planning and Inference, vol. 68, pp. 221-245, 1998.

[5] A. Kitaeva and G. Koshkin, "Nonparametric semirecursive identification in a wide sense of strong mixing processes", Problems of Information Transmission, vol. 46, 1, pp. 22-37, 2010.

[6] E. Nadaraya, "On estimating of regression," Theory Probab. Appl., vol. 9, pp. 141–142, 1964.

[7] G. Koshkin, "Approach to the investigation of functionals of conditional distributions under statistical uncertainty," Automation and Remote Control, vol. 39, 8, pp. 1141-1151, 1978.

[8] A. Kitaeva and G. Koshkin, "Semi-recursive kernel estimation of functions of density functionals and their derivatives), IFAC Proceeding Volumes (IFAC-PapersOnline), 9 (PART 1), pp. 423-428, 2007.

[9] A. Dobrovidov, G. Koshkin, and V. Vasiliev, Non-Parametric State Space Models. Heber, UT 84032, USA. Kendrick Press, Inc. 2012.

[10] G.M. Koshkin, "Stable estimation of ratios of random functions from experimental data", Russian Physics Journal, vol. 36, 10, pp. 1008-1015, 1993.

[11] G.M. Koshkin, "Deviation moments of the substitution estimator and of its piecewise-smooth approximations", Siberian Mathematical Journal, vol. 40, 3, pp. 515-527, 1999.

[12] G.M. Koshkin, "Asymptotic properties of functions of statistics and their application to nonparametric estimation," Automation and Remote Control, vol. 51, 3, pp. 345-357, Part: 1, March 1990.

[13] V.A. Vasiliev and G.M. Koshkin, "Nonparametric identification of autoregressions," Theory of Probability and Its Applications, vol. 43, 3, pp. 507-517, 1999.

[14] A.V. Kitaeva and G.M. Koshkin, "Recurrent nonparametric estimation of functions from functionals of multidimensional density and their derivatives," Automation and Remote Control, vol. 70, 3, pp. 389-407, March 2009.

[15] A.V. Kitaeva and G.M. Koshkin, "Semi-recursive nonparametric identification in the general sense of a nonlinear heteroscedastic autoregression," Automation and Remote Control, vol. 71, 2, pp. 257-274, February 2010.

[16] A.V. Kitaeva and G.M. Koshkin, "Nonparametric identification of the production functions," Proceedings of the World Congress on Engineering 2011, WCE 2011, 1, pp. 276-280, 2011.

[17] A.V. Kitaeva and G.M. Koshkin, "Nonparametric identification of static and dynamic production functions", IAENG International Journal of Applied Mathematics, vol. 41, 3, pp. 228-234, 2011.

[18] Yu.G. Dmitriev and G.M. Koshkin, "On the use of a priori information in nonparametric regression estimation", IFAC Proceedings Series, vol. 2, pp. 223-228,1987.

[19] Yu.G. Dmitriev and G.M. Koshkin, "Using additional information in nonparametric estimation of density functionals", Automation and Remote Control, vol. 48, 10, pp. 1307-1316, October 1987.