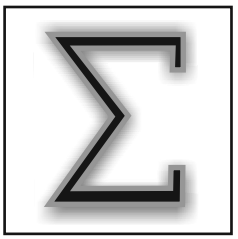
# Математика и информатика



*УДК 519.2*

## СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ AR-ВРЕМЕННЫХ РЯДОВ ПРИ НАЛИЧИИ КЛАССИФИКАЦИИ НАБЛЮДЕНИЙ

***А. В. РУДАКОВСКАЯ*** [1], ***Ю. С. ХАРИН*** [2]

[1] *Белорусский государственный университет, пр. Независимости, 4, 220030, г. Минск, Республика Беларусь*
[2] *Учреждение БГУ «Научно-исследовательский институт прикладных проблем математики и информатики», пр. Независимости, 4, 220030, г. Минск, Республика Беларусь*

Рассмотрена модель авторегрессионного временного ряда при наличии специального типа искажений гипотетической модели – классификации наблюдений. Вместо истинных значений процесса авторегрессии в каждый момент времени регистрируется лишь номер класса (числового промежутка), в который попало значение. Таким образом, задача заключается в оценивании параметров скрытой авторегрессионной модели по наблюдаемой реализации искаженного (классифицированного) временного ряда. Найдены одномерные и многомерные распределения вероятностей классифицированного временного ряда. Задача оценивания параметров решается при помощи построения частотных статистик. По наблюдаемой реализации классифицированного временного ряда строятся частотные статистики – состоятельные оценки одномерных и многомерных распределений вероятностей. Составляя соответствующую систему нелинейных уравнений и решая ее, получаем статистические оценки параметров. В качестве практического примера рассматривается модель $\mathrm{AR}(1)$ с классификацией на 2 числовых промежутка. Найден вид нелинейной системы для построения оценок, имеющей единственное решение. Представлены результаты численных экспериментов, которые иллюстрируют состоятельность построенных оценок.

***Ключевые слова:*** авторегрессионный временной ряд; неполные данные; классификация; статистические оценки; смещение; вариация.

**Образец цитирования:**

Рудаковская А. В., Харин Ю. С. Статистическое оценивание параметров AR-временных рядов при наличии классификации наблюдений // Вестн. БГУ. Сер. 1, Физика. Математика. Информатика. 2016. № 1. С. 84–89.

**For citation:**

Rudakouskaya H. V., Kharin Y. S. Statistical estimation of parameters of autoregressive time series observed under classification. *Vestnik BGU. Ser. 1, Fiz. Mat. Inform.* 2016. No. 1. P. 84–89 (in Russ.).

**Авторы:**

***Анна Вячеславовна Рудаковская*** – магистрант кафедры математического моделирования и анализа данных факультета прикладной математики и информатики.
***Юрий Семенович Харин*** – член-корреспондент НАН Беларуси, доктор физико-математических наук, профессор, директор.

**Authors:**

***Hanna Rudakouskaya***, master's degree student at the department of mathematical modeling and data analysis, school of applied mathematics and computer science.
*hanna.rudakouskaya@gmail.com*
***Yuriy Kharin***, corresponding member of the National academy of sciences of Belarus, doctor habilitatus of physics and mathematics, full professor; director.

# STATISTICAL ESTIMATION OF PARAMETERS OF AUTOREGRESSIVE TIME SERIES OBSERVED UNDER CLASSIFICATION

*H. V. RUDAKOUSKAYA*[a], *Y. S. KHARIN*[b]

[a]*Belarusian State University, Nezavisimosti avenue, 4, 220030, Minsk, Republic of Belarus*
[b]*Research Institute for Applied Problems of Mathematics and Informatics, Belarusian State University, Nezavisimosti avenue, 4, 220030, Minsk, Republic of Belarus*

The model of autoregressive time series under the special type of hypothetic model distortion – the distortion of classification type – is considered. Instead of the true values of the autoregression process we register only a number of the class (interval on the real number line) at every moment of observation. The problem is to estimate parameters of the hidden autoregression model having known the classified time series values only. Univariate and multivariate probability distributions of the classified time series are found. The problem of estimating the parameters is solved using the frequency statistics calculated on the values of observed time series which are the consistent estimates of univariate and multivariate probabilities. Estimators of the parameters are the solution of the appropriate system of nonlinear equations composed with estimates of probabilities and their analytical forms. We study the model of the $\mathrm{AR}(1)$ time series classified by 2 intervals as a practical example. An explicit form of the nonlinear system for estimation is found. The system is one-value solvable. The results of numerical experiments illustrate the consistency of the constructed estimators.

***Key words:*** autoregressive time series; incomplete data; classification; statistical estimator; bias; mean squared error.

## Введение и математическая модель

Модель *авторегрессионных временных рядов* для реальных процессов часто встречается в математической статистике [1, 2]. Однако обычно на практике временной ряд имеет некоторые искажения. Одним из типов искажений авторегрессионных временных рядов является *классификация наблюдений* [3]: регистрация в каждый момент времени вместо истинного значения временного ряда лишь некоторого классифицированного значения – номера класса (интервала), в который попало исходное наблюдение. Такой тип искажений изучался применительно к процессам скользящего среднего [4]; исследовалась модель целочисленной авторегрессии 1-го порядка [5]; в [6] изучалась модель авторегрессионного временного ряда в случае регистрации знака отсчетов; множественная регрессия при наличии классификации наблюдений рассматривалась в [7]. В настоящей работе исследуется модель авторегрессии $p$-го порядка $\mathrm{AR}(p)$ при наличии классификации наблюдений.

Пусть на вероятностном пространстве $(\Omega, \mathcal{F}, P)$ определен авторегрессионный временной ряд порядка $p$ $(\mathrm{AR}(p))$ [1]

$$x_t = \theta_1^0 x_{t-1} + \theta_2^0 x_{t-2} + \ldots + \theta_p^0 x_{t-p} + \xi_t,\ t \in \mathbb{Z}, \tag{1}$$

где $p \in \mathbb{N}$ – порядок авторегрессии; $\theta^0 = \left(\theta_i^0\right) \in \mathbb{R}^p$ – вектор коэффициентов авторегрессии; $\theta_p^0 \neq 0$; $\{\xi_t\}$ – дискретный «белый шум» (последовательность независимых одинаково распределенных гауссовских случайных величин, $\mathcal{L}\{\xi_t\} = \mathcal{N}_1\left(0, \sigma^2\right)$); коэффициенты $\theta^0$ удовлетворяют условию стационарности модели [1].

Пусть задано борелевское разбиение числовой прямой на $2 \le L < +\infty$ числовых промежутков:

$$\mathbb{R} = \bigcup_{i=0}^{L-1} A_i, \tag{2}$$

$$A_i = \left(a_i;\ a_{i+1}\right],\ -\infty = a_0 < a_1 < \ldots < a_{L-1} < a_L = +\infty.$$

Вместо исходного временного ряда $\{x_t\}$ наблюдается классифицированный временной ряд $\{y_t\}$, в котором

$$y_t = \mathrm{class}(x_t) :: = \sum_{i=0}^{L-1} iI\{x_t \in A_i\} \in B,\ B = \{0, 1, \ldots, L-1\}, \tag{3}$$

где $I\{\cdot\}$ – индикатор события, указанного в фигурных скобках; $y_t$ – номер класса, в который попало значение $x_t$.

Пусть наблюдается реализация $Y = \{y_1, \ldots, y_T\}$ классифицированного временного ряда $y_t$ длительностью $T$ при известных интервалах классификации (2). Порядок авторегрессии $p$ предполагается известным. Задача состоит в построении статистической оценки составного вектора параметров $\theta^0 = (\theta_1^0, \theta_2^0, \ldots, \theta_p^0)$, $\sigma$:

$$(\hat{\theta}, \hat{\sigma}) = (\hat{\theta}(Y), \hat{\sigma}(Y)). \tag{4}$$

## Вероятностное распределение наблюдаемого дискретного временного ряда

Для построения оценки (4) по неполным данным – наблюдаемой реализации дискретного временного ряда $Y = \{y_1, \ldots, y_T\}$ – нам понадобятся одномерные и многомерные распределения вероятностей $\{y_t\}$.

Обозначим: $\Phi(\cdot)$, $\varphi(\cdot)$ – функция распределения и плотность распределения вероятностей стандартного нормального закона $\mathcal{N}_1(0, 1)$ соответственно.

**Теорема 1.** *Если скрытый авторегрессионный процесс $x_t$ соответствует модели (1), то наблюдаемый процесс $y_t$ имеет следующее дискретное распределение вероятностей на множестве $B$:*

$$p_i(\theta, \sigma) = P\{y_t = i\} = P\{x_t \in (a_i; a_{i+1}]\} = \Phi\left(\frac{a_{i+1}\sqrt{1 - \sum_{j=1}^{p}\rho(j)\theta_j}}{\sigma}\right) - \Phi\left(\frac{a_i\sqrt{1 - \sum_{j=1}^{p}\rho(j)\theta_j}}{\sigma}\right), \tag{5}$$

*где $\rho(\cdot)$ – автокорреляционная функция исходного (неискаженного) временного ряда $\mathrm{AR}(p)$; $i \in B$.*

Доказательство. Доказательство теоремы непосредственно следует из описания модели (1)–(3) и вероятностных свойств авторегрессионных временных рядов [1, с. 70–80; 8, с. 431–449].

**Теорема 2.** *Если скрытый авторегрессионный процесс $x_t$ соответствует модели $\mathrm{AR}(1)$, то двумерное распределение вероятностей биграмм $(y_t, y_{t+1})$ наблюдаемого процесса $y_t$ имеет вид*

$$p_{ij}(\theta, \sigma) = P\{y_t = i, y_{t+1} = j\} = \frac{\sqrt{1-\theta^2}}{\sigma}\int_{a_i}^{a_{i+1}}\varphi\left(\frac{\sqrt{1-\theta^2}\,x}{\sigma}\right)\left(\Phi\left(\frac{a_{j+1} - \theta x}{\sigma}\right) - \Phi\left(\frac{a_j - \theta x}{\sigma}\right)\right)dx,\ i,\ j \in B. \tag{6}$$

Доказательство. По определению

$$p_{ij}(\theta, \sigma) = P\{y_t = i, y_{t+1} = j\} = P\{x_t \in (a_i; a_{i+1}], x_{t+1} \in (a_j; a_{j+1}]\} =$$
$$= \int_{a_j}^{a_{j+1}}\int_{a_i}^{a_{i+1}} n_2(x|\mathrm{O}_2, \Sigma)dx_1dx_2 = \int_{a_i}^{a_{i+1}} n_1(x_1|0, \sigma_{11})dx_1\int_{a_j}^{a_{j+1}} p(x_2|x_1)dx_2. \tag{7}$$

Из свойств авторегрессионных временных рядов [1, 8] следует, что

$$\sigma_{11} = \sigma_{22} = D\{x_t\} = \frac{\sigma^2}{1-\theta^2}; \tag{8}$$

$$\sigma_{12} = \sigma_{21} = \mathrm{cov}\{x_{t+1}, x_t\} = \theta\sigma_{11} = \theta\sigma_{22}. \tag{9}$$

Согласно [9, с. 33]

$$E\{x_2|x_1\} = E\{x_2\} - \frac{\sigma_{12}}{\sigma_{11}}(x_1 - E\{x_1\}) = \frac{\sigma_{12}}{\sigma_{11}}x_1 = \frac{\theta\sigma_{22}}{\sigma_{11}}x_1 = \theta x_1; \tag{10}$$

$$D\{x_2|x_1\} = \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{22}} = \sigma^2. \tag{11}$$

Подставляя выражения (8)–(11) в (7), получаем

$$p_{ij}(\theta, \sigma) = \int_{a_i}^{a_{i+1}} n_1\left(x_1 \middle| 0, \frac{\sigma^2}{1-\theta^2}\right) dx_1 \int_{a_j}^{a_{j+1}} n_1\left(x_2 \middle| \theta x_1, \sigma^2\right) dx_2 =$$

$$= \frac{\sqrt{1-\theta^2}}{\sigma} \int_{a_i}^{a_{i+1}} \varphi\left(\frac{\sqrt{1-\theta^2}\,x}{\sigma}\right)\left(\Phi\left(\frac{a_{j+1}-\theta x}{\sigma}\right) - \Phi\left(\frac{a_j - \theta x}{\sigma}\right)\right) dx, \tag{12}$$

таким образом, (12) совпадает с (6).

**Следствие.** *Если скрытый авторегрессионный процесс $x_t$ соответствует модели $\mathrm{AR}(1)$ с классификацией наблюдений на $L = 2$ класса $\left(A_0 = (-\infty, a],\ A_1 = (a, +\infty)\right)$, то вероятность биграммы $(0, 1)$ $(\{y_t = 0,\ y_{t+1} = 1\})$ наблюдаемого процесса $y_t$ имеет вид*

$$p_{01}(\theta, \sigma) = P\{y_t = 0,\ y_{t+1} = 1\} = P\{x_t \in (-\infty;\ a],\ x_{t+1} \in (a;\ +\infty)\} =$$

$$= \frac{\sqrt{1-\theta^2}}{\sigma} \int_{-\infty}^{a} \varphi\left(\frac{\sqrt{1-\theta^2}\,x}{\sigma}\right)\left(1 - \Phi\left(\frac{a-\theta x}{\sigma}\right)\right) dx. \tag{13}$$

**Теорема 3.** *Если скрытый авторегрессионный процесс $x_t$ соответствует модели (1), то распределение вероятностей k-грамм, $k \geq 2$, наблюдаемого процесса $y_t$ имеет вид*

$$p_{i_1 i_2 \ldots i_k}(\theta, \sigma) = \int_{a_{i_k}}^{a_{i_k+1}} \ldots \int_{a_{i_2}}^{a_{i_2+1}} \int_{a_{i_1}}^{a_{i_1+1}} n_k\left(x \middle| \mathbb{O}_k,\ \Sigma\right) dx_1 dx_2 \ldots dx_k, \tag{14}$$

*где $n_k(\cdot)$ – k-мерная плотность нормального распределения; $\Sigma = (\sigma_{ij})$, $\sigma_{ij} = \sigma(i-j)$, $\sigma(\cdot)$ – автоковариационная функция неискаженного временного ряда $x_t$, $i_1, i_2, \ldots, i_k \in B$.*

## Метод оценивания параметров модели на основе частотных статистик

Для построения оценки предлагается использовать тот факт, что по наблюдаемой реализации $Y = \{y_1,\ y_2,\ \ldots,\ y_T\}$ можно построить состоятельные оценки одномерных и многомерных распределений вероятностей (5), (6), (13), (14) дискретного временного ряда $\{y_t\}$:

$$\hat{p}_i = \frac{1}{T}\sum_{t=1}^{T} I\{y_t = i\},\ i = 1, \ldots, L-1; \tag{15}$$

$$\hat{p}_{ij} = \frac{1}{T-1}\sum_{t=1}^{T-1} I\{y_t = i,\ y_{t+1} = j\}; \tag{16}$$

$$\hat{p}_{i_1 i_2 \ldots i_k} = \frac{1}{T-1}\sum_{t=1}^{T-1} I\{y_t = i_1,\ y_{t+1} = i_2, \ldots, y_{t+k-1} = i_k\}.$$

С другой стороны, одномерные и многомерные распределения вероятностей (5), (6), (13), (14) для $\{y_t\}$ найдены в теоремах 1–3 и следствии. Тогда относительно параметров $\theta$, $\sigma$ с помощью (15), (16) может быть построена система нелинейных уравнений вида

$$\begin{cases} p_i(\theta, \sigma) = \hat{p}_i,\ i \in B, \\ p_{ij}(\theta, \sigma) = \hat{p}_{ij},\ i,\ j \in B, \\ p_{ijk}(\theta, \sigma) = \hat{p}_{ijk},\ i,\ j,\ k \in B, \\ \ldots. \end{cases} \tag{17}$$

Выбираем из системы (17) $p + 1$ уравнений с тем, чтобы гарантировать однозначную разрешимость системы. Решая систему, получаем оценки $\hat{\theta}$, $\hat{\sigma}$. Состоятельность построенных оценок следует из состоятельности оценок $\hat{p}_i$, $\hat{p}_{ij}$, …, $\hat{p}_{i_1 i_2 \dots i_k}$ и теоремы о функциональном преобразовании сходящихся последовательностей [10].

Проиллюстрируем применение этого метода для $\mathrm{AR}(1)$-временного ряда с классификацией наблюдений при $L = 2$:

$$x_t = \theta^0 x_{t-1} + \xi_t;\ \mathcal{L}\{\xi_t\} = N_1(0,\ \sigma^2);\ -1 < \theta^0 < 1; \tag{18}$$

$$y_t = \mathrm{class}(x_t) = I\{x_t \in A_1\},\ A_0 = (-\infty,\ a],\ A_1 = (a,\ +\infty), \tag{19}$$

или, что эквивалентно,

$$y_t = \begin{cases} 0, & x_t \le a, \\ 1, & x_t > a. \end{cases} \tag{20}$$

Частный случай (18)–(20) представляет значительный интерес для рассмотрения, поскольку процессы авторегрессии 1-го порядка имеют большое практическое значение [1], а классификация всего по двум интегралам дает предельный случай неполноты информации об исходном (неискаженном) процессе $\{x_t\}$.

Согласно представленному выше общему методу оценивания на основе частотных статистик в частном случае (18)–(20) из (17) имеем систему двух уравнений относительно $\theta$, $\sigma$:

$$\begin{cases} P_0(\theta,\ \sigma) = \hat{p}_0, \\ P_{01}(\theta,\ \sigma) = \hat{p}_{01}. \end{cases} \tag{21}$$

С использованием теоремы 1 и следствия из теоремы 2 (21) примет вид

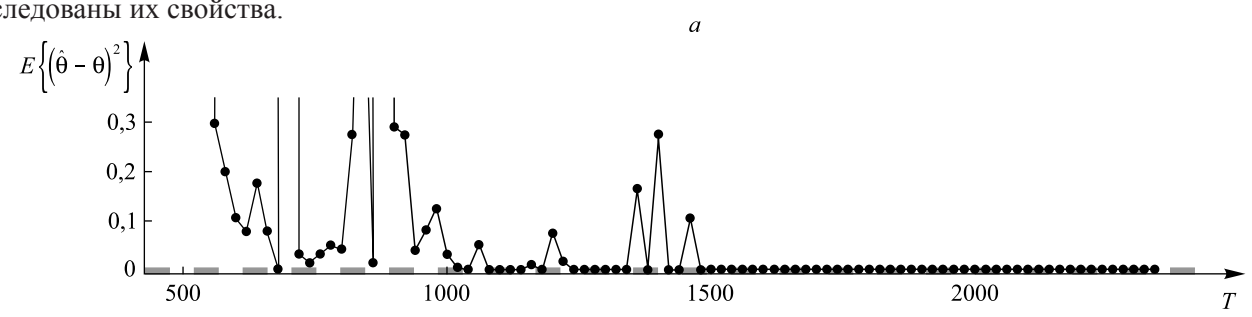
$$\begin{cases} \sigma\Phi^{-1}(\hat{p}_0) = a\sqrt{1-\theta^2}, \\ \dfrac{\sqrt{1-\theta^2}}{\sigma} \displaystyle\int_{-\infty}^{a} \varphi\left(\frac{\sqrt{1-\theta^2}\,x}{\sigma}\right)\left(1 - \Phi\left(\frac{a-\theta x}{\sigma}\right)\right) dx = \hat{p}_{01} \end{cases} \tag{22}$$

и имеет единственное решение при $a \neq 0$.

## Численные результаты

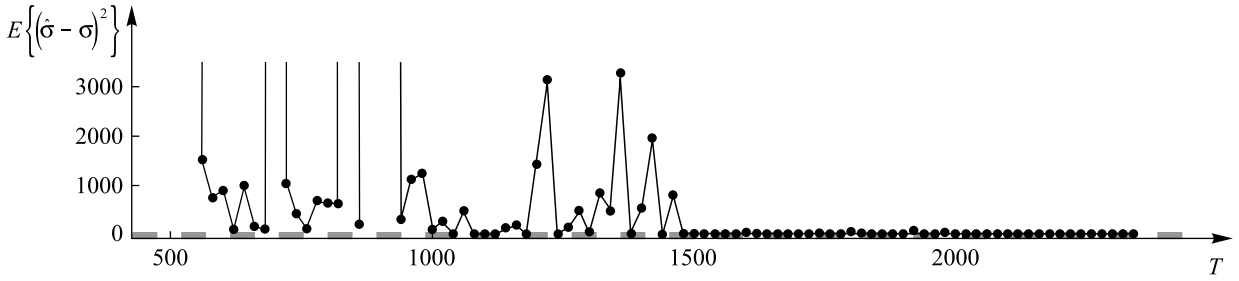
Результаты численных экспериментов по методу Монте-Карло для оценок параметров в случае (18)–(20) представлены на рисунке. Эксперименты проводились для длин реализаций временного ряда $T = 500, 520, \dots, 2340$. Для каждой длины реализации $T$ проводилось $K = 100$ повторений эксперимента, состоящих в имитации $\{x_t\}$, $\{y_t\}$ согласно (18)–(20) и вычислении оценок параметров $\hat{\theta}$, $\hat{\sigma}$ согласно (22). Численные результаты иллюстрируют состоятельность построенных оценок.

Таким образом, в рассмотренной модели авторегрессионных временных рядов при наличии искажения типа классификации найдено распределение вероятностей для наблюдаемого дискретного временного ряда. При помощи частотных оценок одномерных и многомерных распределений вероятностей дискретного временного ряда построены оценки параметров скрытой авторегрессионной модели и исследованы их свойства.




Зависимость среднеквадратической ошибки оценивания параметров от длины временного ряда: $a - \theta$ (окончание см. на с. 89)

*б*




Окончание (начало см. на с. 88): *б* – σ

## БИБЛИОГРАФИЧЕСКИЙ СПИСОК (*REFERENCES*)

1. *Бокс Дж., Дженкинс Г.* Анализ временных рядов, прогноз и управление : в 2 т. М., 1974. Т. 1.

2. *Weber R.* A course of 8 lectures to Cambridge M. Phil in Statistics students, course notes [Electronic resource]. Cambridge, 1999. URL: http://www.statslab.cam.ac.uk/~rrw1/timeseries (date of access: 10.11.2015).

3. *Харин Ю. С.* Оптимальность и робастность в статистическом прогнозировании. Минск, 2008.

4. *Dosla S.* Estimation of parameters of a clipped MA(1) process // Commun. in Stat. – Theory and Methods. 2010. Vol. 40. P. 2437–2454 [Dosla S. Estimation of parameters of a clipped MA(1) process. *Commun. in Stat. – Theory and Methods*. 2010. Vol. 40. P. 2437–2454 (in Engl.)].

5. *Yao J. F., Kachour M.* First-order rounded integer-valued autoregressive (RINAR(1)) process // J. of Time Ser. Anal. 2009. Vol. 30. P. 417–448 [Yao J. F., Kachour M. First-order rounded integer-valued autoregressive (RINAR(1)) process. *J. of Time Ser. Anal*. 2009. Vol. 30. P. 417–448 (in Engl.)].

6. *Kedem B.* Spectral analysis and discrimination by zero-crossings // Proc. IEEE. 1986. № 11 (74). P. 1477–1493 [Kedem B. Spectral analysis and discrimination by zero-crossings. *Proc. IEEE.* 1986. No. 11 (74). P. 1477–1493 (in Engl.)].

7. *Ageeva H., Kharin Y.* ML estimation of multiple regression parameters under classification of the dependent variable // Lith. Math. J. 2015. Vol. 55, № 1. P. 48–60 [Ageeva H., Kharin Y. ML estimation of multiple regression parameters under classification of the dependent variable. *Lith. Math. J.* 2015. Vol. 55, No. 1. P. 48–60 (in Engl.)].

8. *Суслов В. И., Ибрагимов Н. М., Талышева Л. П., Цыплаков А. А.* Эконометрия : учеб. пособие. Новосибирск, 2005.

9. *Андерсон Т.* Введение в многомерный статистический анализ. М., 1963.

10. *Боровков А. А.* Математическая статистика. М., 1984.

*Статья поступила в редколлегию 01.10.2015.*
*Received by editorial board 01.10.2015.*