

УДК 519.25

Г.М. Кошкин, И.Ю. Глухова

СХОДИМОСТЬ В СРЕДНЕКВАДРАТИЧЕСКОМ НЕПАРАМЕТРИЧЕСКОГО АЛГОРИТМА ИДЕНТИФИКАЦИИ ARX-ПРОЦЕССА

Для ядерных оценок функции, определяющей ARX-процесс, находятся главные части их среднеквадратических ошибок (СКО). Проводится сравнение параметрических и непараметрических алгоритмов идентификации с помощью статистического моделирования.

Ключевые слова: *условное среднее, среднеквадратическая ошибка, ARX-процесс, непараметрическая идентификация.*

Рассматривается скалярная последовательность $(Y_t)_{t=-1,0,1,\dots}$, генерируемая ARX(m,p,d)-процессом

$$Y_t = \Psi(Y_{t-i_1}, \dots, Y_{t-i_m}, X_{t-j_1}^1, \dots, X_{t-j_r}^1, \dots, X_{t-j_1}^p, \dots, X_{t-j_k}^p) + \xi_t = \Psi(Y_{t,m}, X_{t,s}) + \xi_t, \quad (1)$$

где $Y_{t,m} = (Y_{t-i_1}, \dots, Y_{t-i_m})$, $X_{t,s} = (X_{t-j_1}^1, \dots, X_{t-j_r}^1, \dots, X_{t-j_1}^p, \dots, X_{t-j_k}^p)$ – экзогенные переменные, $s = r + \dots + k$, $d = \max(r, \dots, k)$, $1 \leq i_1 < \dots < i_m < n$, $0 \leq j_1 < \dots < j_r < n, \dots$, $0 \leq j_1 < \dots < j_k < n$ – известные подпоследовательности натурального ряда чисел, (ξ_t) – последовательность независимых одинаково распределенных (с положительной на R^1 плотностью распределения) случайных величин с нулевым математическим ожиданием, конечной дисперсией, нулевым третьим и конечным четвертым моментами, $\Psi(Y_{t,m}, X_{t,s})$ – неизвестная непериодическая ограниченная функция. Модели типа (1) находят широкое применение при анализе экономических систем и финансовых временных рядов.

В работе предполагается, что выполняются условия, при которых последовательность $(Y_t)_{t=-1,0,1,\dots}$ удовлетворяет условию сильного перемешивания с коэффициентом с.п. [1,2]

$$\alpha(\tau) \approx e^{-\delta\tau}, \delta > 0, \tau \rightarrow \infty.$$

Факт, что (Y_t) удовлетворяет условию с.п. с коэффициентом с.п. $\alpha(\tau)$, будем обозначать $(Y_t) \in S(\alpha)$.

Пусть $Y_1, Y_2, \dots, Y_n$ – наблюдения, порожденные процессом (1). В качестве модели структуры Ψ в (1) возьмем условное математическое ожидание $b(y, x) = E(Y_t | Y_{t,m} = y, X_{t,s} = x) = E(Y_t | y, x)$, $(y, x) \in R^{m+s}$.

Согласно [3], $a_g(y, x) = \int q^g f(q, y, x) dq$, $g = 0, 1$, – базовые функционалы, где $f(q, y, x)$ – неизвестная плотность распределения случайного вектора $(Y_t, Y_{t,m}, X_{t,s})$ в стационарном режиме. Так как $a_0(y, x) = \int f(q, y, x) dq = p(y, x)$, где $p(y, x)$ – плот-

ность распределения вектора $(Y_{t,m}, X_{t,s})$, то условное математическое ожидание можно представить в виде

$$b(y, x) = \frac{a_1(y, x)}{a_0(y, x)} = \frac{a_1(y, x)}{p(y, x)} = \int Y_t f(Y_t | y, x) dY_t.$$

В качестве непараметрических оценок функционалов $a_g(y, x)$ в точке (y, x) возьмём

$$a_{gn}(y, x) = \frac{1}{n-Q} \sum_{i=Q+1}^n Y_i^g \frac{1}{\prod_{j=1}^m h_j} K_m \left(\frac{y - Y_{i,m}}{h^y} \right) \frac{1}{\prod_{j=1}^r h_{1j} \dots \prod_{j=1}^k h_{pj}} K_s \left(\frac{x - X_{i,s}}{h^x} \right),$$

где $Q = \max(i_m, \max(j_r, \dots, j_k))$, $h^y = (h_1, \dots, h_m)$, $h^x = (h_1^x, \dots, h_p^x)$, $h_1^x = (h_{11}, \dots, h_{1r})$, $\dots$, $h_p^x = (h_{p1}, \dots, h_{pk})$ – соответствующие параметры размытости, положительные числа, а K_m и K_s – m -мерное и s -мерное ядра.

Таким образом, ядерная оценка подстановки условного функционала $b(y, x)$ в точке (y, x) , а следовательно, и функции Ψ в (1) задаются отношением вида

$$b_n(y, x) = \Psi_n(y, x) = \frac{\sum_{i=Q+1}^n Y_i \frac{1}{\prod_{j=1}^m h_j} K_m \left(\frac{y - Y_{i,m}}{h^y} \right) \frac{1}{\prod_{j=1}^r h_{1j} \dots \prod_{j=1}^k h_{pj}} K_s \left(\frac{x - X_{i,s}}{h^x} \right)}{\sum_{i=Q+1}^n \frac{1}{\prod_{j=1}^m h_j} K_m \left(\frac{y - Y_{i,m}}{h^y} \right) \frac{1}{\prod_{j=1}^r h_{1j} \dots \prod_{j=1}^k h_{pj}} K_s \left(\frac{x - X_{i,s}}{h^x} \right)}. \quad (2)$$

Задача идентификации модели (1) является частным случаем оценивания функции вида (ср. [4])

$$H(A) = H(a_0, a_1) = \frac{a_1}{a_0}, \quad a_0 > 0,$$

где $A = (a_0, a_1)$, $a_g = a_g(u) = \int q^g f(q, u) dq$, $g = 0, 1$, $f(q, u) = f(z)$, $z \in R^{m+s+1}$ – плотность распределения случайного вектора $(Y_t, U_t) = (Y_t, Y_{t,m}, X_{t,s}) = Z_t$ в стационарном режиме.

В данной работе исследуется сходимость в среднеквадратическом оценки (2) к функции Ψ , определяющей ARX-процесс.

1. Сходимость четвертых моментов оценок базовых функционалов

Введем обозначения: $f_{l(i+1)(i+j+1)(i+j+k+1)}(z, v, u, w)$ – $4(m+s+1)$ -мерная плотность распределения выборочных величин

$$\begin{aligned} & Z_1, Z_{(i+1)}, Z_{(i+j+1)}, Z_{(i+j+k+1)}, a_{l(i+1)(i+j+1)(i+j+k+1), p}^+(z, y, z', y') = \\ & = \int_{R^4} |v u v' u'|^p f_{l(i+1)(i+j+1)(i+j+k+1)}(v, z, v, y, v', z', v', y') dv dv' dv' dv', \\ & Q+1 \leq i, j, k < n, \quad i+j+k \leq n-1; \end{aligned}$$

$$a_{1(1+j)(1+j+k),p}(z, y, z') = \int_{R^3} |v \cup v'|^p f_{1(1+j)(1+j+k)}(v, z, v, y, v', z') dv dv',$$

$$a_{1(i+1),p}^+(z, z') = \int_{R^2} |v \cup v'|^p f_{1(1+j)}(v, z, v, z') dv dv, \quad a_p^+(z) = \int |v|^p f(v, z) dv,$$

$$M_4(a_{gn}) = E[a_{gn}(z) - a_g(z)]^4, \quad S_{gn} = a_{gn}(z) - Ea_{gn}(z), \quad g = 0, 1, \quad L = m + s.$$

Определение 1. Функция $K(u)$ принадлежит классу одномерных нормированных ядер $K(\cdot) \in A_v$, если $\int |K(u)| du < \infty$, $\int K(u) du = 1$, $\int |u^v K(u)| du < \infty$, $T_j = \int u^j K(u) du = 0$, $j = 1, \dots, v-1$, $T_v \neq 0$, $K(u) = K(-u)$.

При исследовании сходимости, для простоты, для каждой переменной будем использовать один и тот же параметр размытости, а в качестве многомерных ядер соответствующей размерности – произведение одномерных ядер.

Лемма 1 (порядок сходимости четвертых центральных моментов оценок $a_{gn}(z)$ для с.п. наблюдений). Пусть

- 1) $(Z_j) \in S(\alpha)$ и $\int_0^\infty \tau^2 [\alpha(\tau)]^{\frac{\delta}{2+\delta}} d\tau < \infty$ для некоторого $0 < \delta < 2$;
- 2) $\sup_{u \in R^1} |K(u)| < \infty$, $\int |K(u)| du < \infty$;
- 3) $\lim_{n \rightarrow \infty} (h_n + 1/(nh_n^L)) = 0$;
- 4) $\sup_{u \in R^L} |a_{g\beta}^+(u)| < \infty$, $\beta = 0, 4$;
- 5) $\sup_{u \in R^L} a_{1(i+1)(i+j+1)(i+j+k+1),g}^+(u, u, u, u) < \infty$, $\sup_{u \in R^L} a_{1(1+j)(1+j+k),g(2+\delta)}^+(u, u, u) < \infty$,
 $\sup_{u \in R^L} a_{1(i+1),g(2+\delta)}^+(u, u) < \infty$.

Тогда при $n \rightarrow \infty$

$$E(S_{gn})^4 = O\left(\frac{1}{n^2 h_n^{2L}}\right).$$

Лемма 2 (порядок сходимости четвертых моментов отклонений $M_4(a_{gn})$ для с. п. наблюдений). Если $K(\cdot) \in A_v$, выполняются условия леммы 3 из [4] при $r = 0$, а также приведенной выше леммы 1, то при $n \rightarrow \infty$

$$M_4(a_{gn}) = O\left(\frac{1}{n^2 h_n^{2L}} + h_n^{4v}\right).$$

2. Главная часть СКО оценки подстановки функции Ψ

При нахождении СКО оценки $\Psi_n(y, x)$ используются лемма 1 и лемма 2. Обо-

значим $\omega_{gv}(u) = \frac{T_v}{v!} \sum_{j=1}^L \frac{\partial^v a_g(u)}{\partial u_j^v}$, $H_g = \partial H(A) / \partial(a_g)$. Отметим, что ниже, в теоре-

ме, используется $H_0 = -a_1 / a_0^2$, $H_0 = 1 / a_0$.

Определение 2. Функция $H(\cdot): R^L \rightarrow R^1$ принадлежит классу $N_v(z)$ ($H(\cdot) \in N_v(z)$), если она и все ее частные производные (до v -го порядка включительно) непрерывны в точке z . Функция $H(\cdot) \in N_v(R)$, если указанные свойства $H(\cdot)$ выполняются для всех $z \in R^L$.

Теорема (СКО оценки $H(A_n)$). Пусть при $g, p = 0, 1$

$$1) (Z_i) \in S(\alpha) \text{ и } \int_0^\infty \tau^2 [\alpha(\tau)]^{\frac{\delta}{2+\delta}} d\tau < \infty \text{ для некоторого } 0 < \delta < 2;$$

$$2) a_{g+p}(\cdot) \in N_0(R), a_{g(2+\delta)}^+(\cdot) \in N_0(u); \sup_{u \in R^L} a_{g+p}^+(u) < \infty, \sup_{u \in R^L} a_{g\beta}^+(u) < \infty, \beta = 0, 4;$$

$$3) K(\cdot) \in A_v, \sup_{u \in R^1} |K(u)| < \infty;$$

$$4) a_0(u) > 0, a_g(\cdot) \in N_v(R), \sup_{u \in R^L} |a_g(u)| < \infty; \sup_{u \in R^L} \left| \frac{\partial^v a_g(u)}{\partial u_l \dots \partial u_q} \right| < \infty, l, \dots, q = \overline{1, L};$$

5) для монотонно невозрастающей последовательности (h_n) имеет место

$$(d_n) = \left(h_n + \frac{1}{nh_n^L} \right) \downarrow 0.$$

$$6) \sup_{u \in R^L} a_{1(i+1)(i+j+1)(i+j+k+1),g}^+(u, u, u, u) < \infty, \sup_{u \in R^L} a_{1(i+1)(i+j+1),g(2+\delta)}^+(u, u, u) < \infty, \\ \sup_{u \in R^L} a_{1(i+1),g(2+\delta)}^+(u, u) < \infty, \sup_{u, u' \in R^L} a_{1(i+1),g+p}^+(u, u') < \infty \text{ для любых } i, j, k \geq 1;$$

7) для всевозможных значений $Y_1, \dots, Y_n, X_1^1, \dots, X_n^1, \dots, X_1^p, \dots, X_n^p$ последовательность $\{ |H(A_n)| \}$ мажорируется числовой последовательностью $(C_0 d_n^{-\gamma})$, C_0 – некоторая постоянная, $0 \leq \gamma \leq 1/4$.

Тогда

$$E[H(A_n) - H(A)]^2 = \sum_{g,p=0}^1 H_g H_p \left[\frac{1}{nh_n^L} a_{g+p}(x) \left(\int K^2(u) du \right)^L + \omega_{gv}(z) \omega_{pv}(z) h_n^{2v} \right] + \\ + O \left(\left[h_n^{2v} + \frac{1}{nh_n^L} \right]^{\frac{3}{2}} \right).$$

Доказательства лемм и теоремы, которые ввиду громоздкости не приводятся, базируются на результатах статей [2–5] и книги [6].

3. Сравнительный анализ параметрических и непараметрических алгоритмов

При статистическом эксперименте в качестве модельных процессов (1) возьмем

$$M(1): Y_n = 0, 2Y_{n-1} + 0, 11X_n^1 + 0, 15X_{n-1}^1 + 0, 3X_n^2 + 0, 2X_{n-2}^3 + \xi_n,$$

$$M(2): Y_n = 0, 01Y_{n-1} \cdot Y_{n-2} + 0, 2X_n^1 + 0, 03X_{n-1}^2 \cdot X_{n-1}^3 + 0, 7(X_n^3) + \xi_n,$$

$$M(3): Y_n = e^{0, 1Y_{n-1} + 0, 2X_n^1 + 0, 1X_{n-1}^1 + 0, 01X_{n-1}^2 + 0, 03X_n^3} + \xi_n.$$

Здесь переменные X^1 , X^2 , и X^3 принимают значения из соответствующих интервалов $[2; 2,5]$, $[5; 6]$, $[8; 10]$ равномерных законов распределения, а случайные величины ξ_n распределены по нормальному закону с нулевым математическим ожиданием и дисперсией, рассчитываемой для каждой модели по формуле

$$\sigma^2 = \frac{Y_{\max} - Y_{\min}}{6} \cdot a,$$

где степень зашумленности a моделей принимает значения 0,01; 0,05; 0,1; 0,15; 0,2; 0,5. Моделирование проводится по объемам наблюдений 50, 100, 200 и 500.

Для наглядности сгенерированные 200 наблюдений для моделей $M(1) - M(3)$ при отсутствии помехи представлены соответственно на рис. 1.

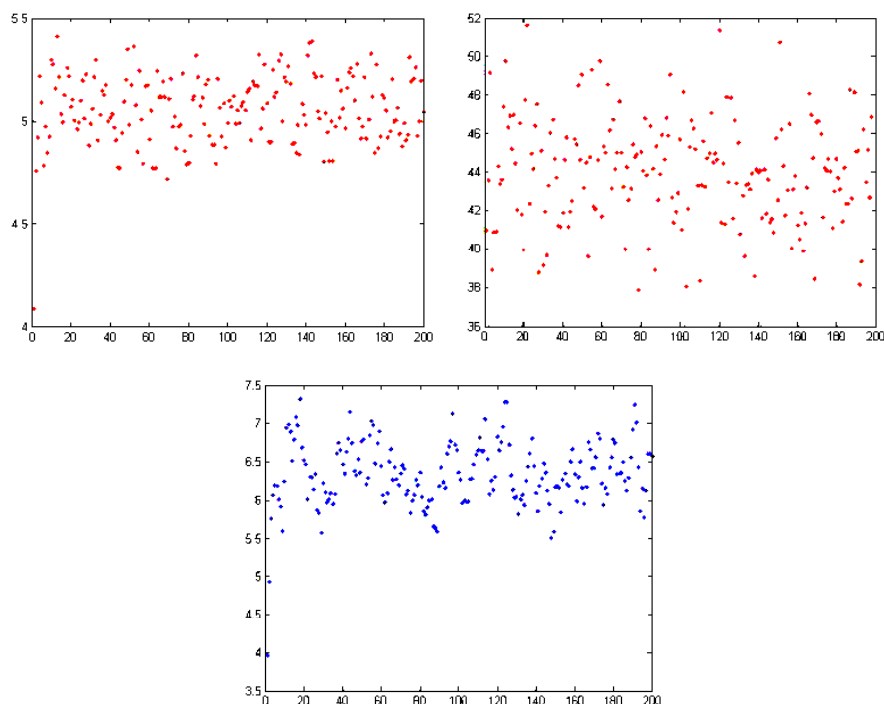


Рис. 1. Реализации 200 наблюдений для моделей $M(1)$, $M(2)$ и $M(3)$

Алгоритмы идентификации функции Ψ в (1) получены методом наименьших квадратов (МНК), итерационным взвешенным методом наименьших квадратов (ВМНК) и с помощью непараметрического подхода. Веса на текущей итерации для ВМНК определяются при помощи биквадратической функции от вектора остатков, рассчитанных на предыдущей итерации. Использование такого алгоритма позволяет задать меньшие значения весов для наблюдений, имеющих большее отклонение от регрессионной модели по отношению к остальным.

Оценки МНК и ВМНК вычисляются с помощью встроенных функций MATLAB. Моделирование непараметрических алгоритмов (2) проводится также на базе MATLAB. В качестве ядра $K(u)$ используется плотность стандартного га-

усовского закона. Параметры размытости определяются двумя способами. Следуя Сильверману [7], параметры размытости вычисляются по следующим формулам:

$$h_{i,Silverman} = 1,06\sigma_i n^{-1/4+L}, \quad i = \overline{1, L},$$

где L – размерность функции Ψ , σ_i^2 – выборочная дисперсия наблюдений i -й переменной. Второй метод нахождения параметров размытости при нахождении оценки $\overline{Y}_n$ основан на использовании эмпирического критерия:

$$h_{i,Empiric} = C_o \sigma_i n^{-1/4+L}, \quad i = \overline{1, L},$$

$$C_o = \arg \min_{0 < C < \infty} \left| Y_{n-1} - \sum_{i=Q}^{n-1} Y_i K_L \left(\frac{U_n - U_i}{h} \right) \right| / \left| \sum_{i=Q}^{n-1} K_L \left(\frac{U_n - U_i}{h} \right) \right|,$$

где $h = (h^y, h^x)$.

В таблице приведены для моделей М(1) – М(3) значения ошибок идентификации

$$A_{n,j} = \frac{1}{n-Q} \sum_{i=Q+1}^n \left| \frac{Y_i - \overline{Y}_i}{Y_i} \right| \cdot 100\%, \quad j = 1, 2, 3,$$

для объёма наблюдений $n = 50$. Для других объемов наблюдений соответствующие результаты представлены на рис. 1–3. Заметим, что все результаты моделирования для определенного объёма выборки усреднялись по 20 различным выборкам такого же объёма.

Усредненные ошибки идентификации для 50 наблюдений

Модель	Степень a зашумленности	МНК, %	ВМНК, %	Непараметрический $h_{i,Silverman}$, %	Непараметрический $h_{i,Empiric}$, %
М(1)	0,01	0,007	0,007	1,164	0,989
	0,05	0,036	0,034	1,176	0,992
	0,1	0,073	0,068	1,166	0,995
	0,15	0,110	0,102	1,173	0,999
	0,2	0,147	0,137	1,195	1,006
	0,5	0,372	0,354	1,283	1,110
М(2)	0,01	1,428	1,309	2,189	1,941
	0,05	1,427	1,309	2,180	1,935
	0,1	1,426	1,312	2,197	1,929
	0,15	1,428	1,317	2,203	1,930
	0,2	1,430	1,327	2,218	1,939
	0,5	1,476	1,416	2,306	2,039
М(3)	0,01	3,615	3,178	4,298	4,145
	0,05	3,849	3,349	4,283	4,101
	0,1	4,174	3,569	4,268	4,093
	0,15	4,821	3,797	4,294	4,115
	0,2	5,382	4,033	4,349	4,084
	0,5	7,351	5,610	4,948	4,717

Из таблицы видно, что качество идентификации для всех моделей и всех методов падает с ростом степени зашумленности.

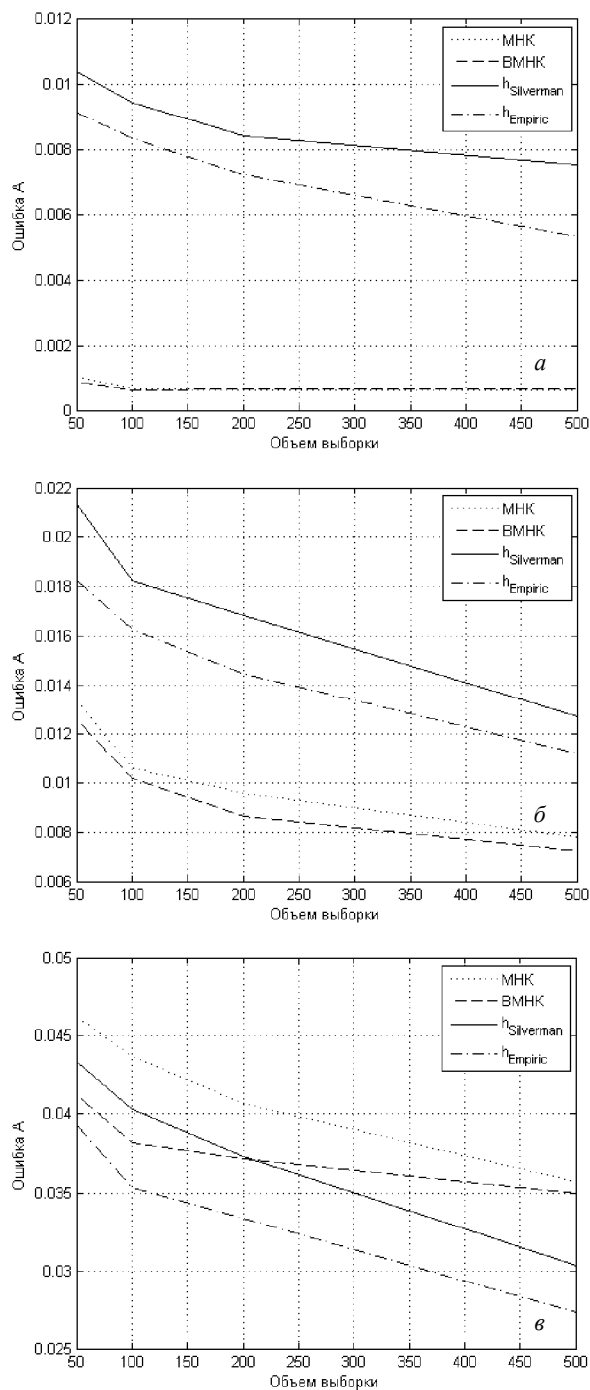


Рис. 2. Зависимость качества идентификации от объема наблюдений для степени зашумленности 0,15: *a* – модель М(1); *б* – модель М(2); *в* – модель М(3)

Анализ рисунков позволяет сделать следующие выводы:

- для нелинейной модели $M(3)$ непараметрические алгоритмы, ввиду их адаптивности, имеют преимущества перед параметрическими;
- прослеживается тенденция уменьшения ошибок идентификации для всех моделей с ростом объема наблюдений.

Заключение

Работоспособность предложенных алгоритмов проверена для различных функций Ψ , определяющих ARX-процесс (1), с помощью статистического моделирования. Показано, что непараметрические алгоритмы идентификации могут иметь преимущество перед параметрическими для нелинейных моделей. При этом непараметрические алгоритмы, в отличие от параметрических, требуют лишь информацию общего характера о структуре исследуемого объекта.

ЛИТЕРАТУРА

1. Masry E., Tjostheim D. Nonparametric estimation and identification of nonlinear ARCH time series // *Econometric Theory*. 1995. V. 11. No. 2. С. 258–289.
2. Китаева А.В., Кошкин Г.М. Полурекуррентная непараметрическая идентификация в широком смысле нелинейной гетероскедастической авторегрессии // *Автоматика и телемеханика*. 2010. № 2. С. 92–111.
3. Кошкин Г.М., Глухова И.Ю. Непараметрическая идентификация нелинейных ARX-процессов // *Вестник Томского государственного университета. Управление, вычислительная техника и информатика*. 2012. № 3 (20). С. 55–61.
4. Китаева А.В., Кошкин Г.М. Рекуррентное непараметрическое оценивание функций от функционалов многомерной плотности и их производных // *Автоматика и телемеханика*. 2009. № 3. С. 48–67.
5. Кошкин Г.М. Моменты отклонений оценки подстановки и ее кусочно-гладких аппроксимаций // *Сибирский математический журнал*. 1999. Т. 40. № 3. С. 605–618.
6. Кошкин Г.М., Пивен И.Г. Непараметрическая идентификация стохастических систем. Хабаровск: РАН, Дальневосточное отделение, 2009. 336 с.
7. Silverman B.W. *Density Estimation for Statistics and Data Analysis*. London, New York: Chapman and Hall/CRC, 1986. 174 p.

Кошкин Геннадий Михайлович

Глухова Ирина Юрьевна

Томский государственный университет

E-mail: kgm@mail.tsu.ru; win32_86@mail.ru

Поступила в редакцию 4 июля 2012 г.

Koshkin Gennady M., Glukhova Irina Yu. (Tomsk State University). Mean square convergence of nonparametric identification algorithm of ARX-process.

Keywords: kernel estimation, conditional mean, mean square error, ARX-process, nonparametric identification.

The principal parts of mean square errors for kernel estimates of the functions defining ARX-process are found. Parametric and nonparametric identification algorithms are compared.