УДК 519.2

Ю.С. ХАРИН, И.А. БОДЯГИН

ОБ ОПТИМАЛЬНОМ ПРОГНОЗИРОВАНИИ АВТОРЕГРЕССИОННЫХ ВРЕМЕННЫХ РЯДОВ ПРИ НАЛИЧИИ ИНТЕРВАЛЬНОГО ЦЕНЗУРИРОВАНИЯ

Белорусский государственный университет

(Поступила в редакцию ...)

Введение. Задача статистического прогнозирования будущих значений временного ряда по имеющимся наблюдениям возникает во многих приложениях: в медицине, экономике, метеорологии, технике, астрономии [1]. Для описания временных рядов с зависимыми наблюдениями и прогнозирования будущих значений широко применяется модель авторегрессии [1, 2]. Однако на практике значения временного ряда часто наблюдаются с искажениями различных типов: выбросы, пропуски, гетероскедастичность [3, 4], цензурирование [5] и др.; обзор типов искажений и их математические описания представлены в [3]. Цензурирование временного ряда заключается в том, что часть наблюдений ряда известна точно, а об остальных наблюдениях известно лишь, что они принадлежат некоторым числовым интервалам. Такая ситуация может возникать из-за наличия у приборов конечных пределов измерения, высокой стоимости проведения точных измерений, разладки оборудования и других причин.

В литературе имеется большое количество работ, посвященных задачам статистического оценивания модельных параметров при наличии цензурирования [5, 6, 7]. Однако статистическое прогнозирование будущих значений цензурированных временных рядов остается мало изученным и актуальным направлением исследований.

1. Математическая модель. Пусть временной ряд x_t описывается моделью AP(p) авторегрессии порядка $p \in \mathrm{N}$ (см., например, [1]):

$$x_t = \sum_{i=1}^{p} \theta_i x_{t-i} + u_t, t \in Z, \tag{1}$$

где $\{\theta_i\}_{i=1}^{p}$ — коэффициенты авторегрессии такие, что все корни порождающего характеристического многочлена $z^p - \sum_{j=1}^{p} \theta_j z^{p-j}$ лежат внутри единичного круга; $\{u_t\}$ — независимые в совокупности одинаково распределенные случайные величины, имеющие нормальный закон распределения вероятностей: $L\{u_t\} = N(0, \sigma^2)$.

Пусть вместо значений временного ряда наблюдаются случайные события:

$$A_t^* = \{x_t \in A_t\}, t \in \{1, \ldots, T\}, \tag{2}$$

где $\{A_t\}$ — заданные борелевские множества, $T > p$ — длительность наблюдения. В данной работе рассматриваются два случая: 1) A_t состоит из одного элемента $(A_t = \{x_t\})$, тогда значение x_t известно точно; 2) A_t является числовым интервалом $(A_t = [a_t, b_t), \quad a_t < b_t)$, тогда имеет место интервальное цензурирование значения x_t, а интервал $[a_t, b_t)$ называется интервалом цензурирования. Статистическое прогнозирование будущего значения $x_{T+1} \in \mathrm{R}$ заключается в вычислении оценки $\hat{x}_{T+1} \in \mathrm{R}$ на основе имеющейся информации о наступлении событий $A_1^*, \ldots, A_T^*$. Иногда возможна ситуация, когда относительно будущего значения x_{T+1} также известна некоторая дополнительная информация $A_{T+1}^* = \{x_{T+1} \in A_{T+1}\}$. Если же дополнительная информация о значении x_{T+1} отсутствует, то будем полагать, что $A_{T+1} = (-\infty, +\infty)$.

Прогнозирующая статистика для x_{T+1} в будущий момент времени $t = T+1$ является числовой функцией наблюдаемых событий:

$$\hat{x}_{T+1} = f\left(A^*_{T+1}, A^*_T, \ldots, A^*_1\right). \quad (3)$$

Погрешность прогнозирования будем характеризовать условным риском прогнозирования

$$r_T(f) = \mathrm{E}\left\{\left(x_{T+1} - \hat{x}_{T+1}\right)^2 \middle| A^*_{T+1}, A^*_T, \ldots, A^*_1\right\} \geq 0, \quad (4)$$

т.е. среднеквадратической ошибкой прогнозирования.

Рассмотрим задачу построения оптимальной прогнозирующей статистики $f_0(\cdot)$, минимизирующей условный риск (4), в случае известных параметров модели (1), (2):

$$r_T(f_0) = \inf_{f(\cdot)} r_T(f). \quad (5)$$

2. Общие результаты для модели АР(*p*).

Теорема 1**.** *Если временной ряд x_t наблюдается при наличии цензурирования общего вида (2), то среди всех прогнозирующих статистик вида (3) оптимальная по критерию минимума риска (5) прогнозирующая статистика определяется условным математическим ожиданием:*

$$\hat{x}_{T+1} = f_0\left(A^*_{T+1}, \ldots, A^*_1\right) = \mathrm{E}\left\{x_{T+1} \middle| A^*_{T+1}, A^*_T, \ldots, A^*_1\right\},\ r_T(f_0) = \mathrm{D}\left\{x_{T+1} \middle| A^*_{T+1}, A^*_T, \ldots, A^*_1\right\}. \quad (6)$$

Доказательство. Преобразуем условный риск (4):

$$r_T(f) = \mathrm{E}\left\{\left(x_{T+1} - \hat{x}_{T+1}\right)^2 \middle| A^*_{T+1}, \ldots, A^*_1\right\} = \mathrm{E}\left\{\left(x_{T+1} - f\left(A^*_{T+1}, \ldots, A^*_1\right)\right)^2 \middle| A^*_{T+1}, \ldots, A^*_1\right\} =$$

$$= \mathrm{E}\left\{\left(\left(x_{T+1} - \mathrm{E}\left\{x_{T+1} \middle| A^*_{T+1}, \ldots, A^*_1\right\}\right) + \left(\mathrm{E}\left\{x_{T+1} \middle| A^*_{T+1}, \ldots, A^*_1\right\} - f\left(A^*_{T+1}, \ldots, A^*_1\right)\right)\right)^2 \middle| A^*_{T+1}, \ldots, A^*_1\right\}.$$

Заметим, что второе слагаемое зависит только от $A^*_{T+1}, \ldots, A^*_1$:

$$r_T(f) = \mathrm{E}\left\{\left(x_{T+1} - \mathrm{E}\left\{x_{T+1} \middle| A^*_{T+1}, \ldots, A^*_1\right\}\right)^2 \middle| A^*_{T+1}, \ldots, A^*_1\right\} + \left(\mathrm{E}\left\{x_{T+1} \middle| A^*_{T+1}, \ldots, A^*_1\right\} - f\left(A^*_{T+1}, \ldots, A^*_1\right)\right)^2 +$$

$$+ 2\,\mathrm{E}\left\{\left(x_{T+1} - \mathrm{E}\left\{x_{T+1} \middle| A^*_{T+1}, \ldots, A^*_1\right\}\right) A^*_{T+1}, \ldots, A^*_1\right\}\left(\mathrm{E}\left\{x_{T+1} \middle| A^*_{T+1}, \ldots, A^*_1\right\} - f\left(A^*_{T+1}, \ldots, A^*_1\right)\right) =$$

$$= \mathrm{D}\left\{x_{T+1} \middle| A^*_{T+1}, \ldots, A^*_1\right\} + \left(\mathrm{E}\left\{x_{T+1} \middle| A^*_{T+1}, \ldots, A^*_1\right\} - f\left(A^*_{T+1}, \ldots, A^*_1\right)\right)^2 \to \min_{f(\cdot)},$$

Из этого представления следует, что (6) есть решение задачи (5). ■

Следствие 1. *Если $A_{T+1} = (-\infty, +\infty)$, то оптимальная прогнозирующая статистика* $\hat{x}_{T+1} = \mathrm{E}\left\{x_{T+1} \middle| A^*_T, A^*_{T-1}, \ldots, A^*_1\right\}$, а ее риск $r(T) = \mathrm{D}\left\{x_{T+1} \middle| A^*_T, A^*_{T-1}, \ldots, A^*_1\right\}$.

Доказательство. Явно следует из свойств математического ожидания и теоремы 1. ■

Это следствие является обобщением известного результата [2] в ситуации, когда $A_1, \ldots, A_T$ – одноточечные множества, т.е. цензурирование отсутствует.

Рассмотрим случай, когда цензурированы только последние q, $0 \leq q \leq T$, значений временного ряда, а остальные $T - q$ наблюдений известны точно.

Лемма 1. *Пусть наблюдаются значения $x_1, \ldots, x_{T-q}$ и случайные события $A^*_{T-q+1}, \ldots, A^*_T$. Тогда справедливо следующее выражение для условной плотности распределения вероятностей*

$$p\left(x_T \middle| A^*_T, \ldots, A^*_{T-q+1}, x_{T-q}, \ldots, x_1\right) = \frac{\int_{A_{T-1}} \ldots \int_{A_{T-q+1}} 1_{A_T}(x_T)\, p\left(x_T, \ldots, x_{T-q+1}, \middle| x_{T-q}, \ldots, x_1\right) dx_{T-q+1} \ldots dx_{T-1}}{\int_{A_T} \ldots \int_{A_{T-q+1}} p\left(x_T, \ldots, x_{T-q+1}, \middle| x_{T-q}, \ldots, x_1\right) dx_{T-q+1} \ldots dx_T},$$

где $1_A(x) = \{1,\ x \in A;0,$ иначе$\}$ – индикаторная функция множества A.

Доказательство. Преобразуем условную плотность распределения вероятностей:

$$p\left(x_T \middle| A_T^*,\ldots,A_{T-q+1}^*,x_{T-q},\ldots,x_1\right)=\frac{p\left(x_T,A_T^*,\ldots,A_{T-q+1}^*,x_{T-q},\ldots,x_1\right)}{p\left(A_T^*,\ldots,A_{T-q+1}^*,x_{T-q},\ldots,x_1\right)}=$$

$$=\frac{p\left(x_T,A_T^*,\ldots,A_{T-q+1}^*\middle|x_{T-q},\ldots,x_1\right)}{P\left(A_T^*,\ldots,A_{T-q+1}^*\middle|x_{T-q},\ldots,x_1\right)}=\frac{\int\limits_{A_{T-1}}\ldots\int\limits_{A_{T-q+1}}1_{A_T}(x_T)p\left(x_T,\ldots,x_{T-q+1},\middle|x_{T-q},\ldots,x_1\right)dx_{T-q+1}\ldots dx_{T-1}}{\int\limits_{A_T}\ldots\int\limits_{A_{T-q+1}}p\left(x_T,\ldots,x_{T-q+1},\middle|x_{T-q},\ldots,x_1\right)dx_{T-q+1}\ldots dx_T}. \quad \blacksquare$$

Обозначим: $\mu(t,m)=\theta_1 x_{t-1}+\ldots+\theta_m x_{t-m}=\sum\limits_{i=1}^{m}\theta_i x_{t-i}$, $t,m\in\mathrm{N}$.

Теорема 2. *Пусть в рамках модели (1), (2) наблюдаются значения* $x_1,\ldots,x_{T-q}$ *и случайные события* $A_{T-q+1}^*=\left\{x_{T-q+1}\in\left[a_{T-q+1},b_{T-q+1}\right)\right\}$, ..., $A_T^*=\left\{x_T\in\left[a_T,b_T\right)\right\}$. *Тогда оптимальная прогнозирующая статистика (6) имеет вид:*

$$\hat{x}_{T+1}=\frac{\int\limits_{a_T}^{b_T}\ldots\int\limits_{a_{T-q+1}}^{b_{T-q+1}}\mu(T+1,p)p(x_T,\ldots,x_{T-q+1}|x_{T-q},\ldots,x_1)dx_{T-q+1}\ldots dx_T}{\int\limits_{a_T}^{b_T}\ldots\int\limits_{a_{T-q+1}}^{b_{T-q+1}}p(x_T,\ldots,x_{T-q+1}|x_{T-q},\ldots,x_1)dx_{T-q+1}\ldots dx_T}. \qquad (7)$$

Доказательство. Оценку (6) в силу (1) можно представить следующим образом:

$$\hat{x}_{T+1}=\mathrm{E}\left\{x_{T+1}\middle|A_T^*,\ldots,A_{T-q+1}^*,x_{T-q},\ldots,x_1\right\}=\mathrm{E}\left\{\sum_{i=1}^{p}\theta_i x_{T+1-i}+u_{T+1}\middle|A_T^*,\ldots,A_{T-q+1}^*,x_{T-q},\ldots,x_1\right\}=$$

$$=\mathrm{E}\left\{\mu(T+1,p)\middle|A_T^*,\ldots,A_{T-q+1}^*,x_{T-q},\ldots,x_1\right\},$$

так как случайная величина u_{T+1} не зависит от $A_T^*,\ldots,A_{T-q+1}^*,x_{T-q},\ldots,x_1$ и $\mathrm{E}\{u_{T+1}\}=0$. Вычислив полученное математическое ожидание и воспользовавшись леммой 1, получим требуемое равенство (7). $\blacksquare$

Введём обозначения:

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}},\ \Phi(x)=\int\limits_{-\infty}^{x}\varphi(t)dt\ —$$

соответственно плотность и функция распределения вероятностей стандартного нормального закона N(0,1);

$$\mathrm{n}\left(x\middle|\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}=\frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)—$$

плотность распределения вероятностей нормального закона с параметрами μ и σ^2;

$$\Psi(x,y,m,s,u,v)=\frac{u\,\varphi\left(\frac{x-m}{s}\right)-v\,\varphi\left(\frac{y-m}{s}\right)}{\Phi\left(\frac{y-m}{s}\right)-\Phi\left(\frac{x-m}{s}\right)},\ x,y,m,s,u,v\in\mathrm{R}.$$

Из [9] имеем равенства:

$$\int\limits_a^b\mathrm{n}\left(x\middle|\mu,\sigma^2\right)dx=\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right), \qquad (8)$$

$$\int\limits_a^b x\,\mathrm{n}\left(x\middle|\mu,\sigma^2\right)dx=\sigma\left(\varphi\left(\frac{a-\mu}{\sigma}\right)-\varphi\left(\frac{b-\mu}{\sigma}\right)\right)+\mu\left(\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)\right), \qquad (9)$$

$$\int\limits_a^b x^2\,\mathrm{n}\left(x\middle|\mu,\sigma^2\right)dx=\sigma\left((a+\mu)\varphi\left(\frac{a-\mu}{\sigma}\right)-(b+\mu)\varphi\left(\frac{b-\mu}{\sigma}\right)\right)+\left(\sigma^2+\mu^2\right)\left(\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)\right). \qquad (10)$$

Следствие 2. *Пусть в рамках модели (1), (2) наблюдаются значения* $x_1,\ldots,x_{T-1}$ *и случайное событие* $A_T^* = \{x_T \in [a_T, b_T)\}$ $(q=1)$. *Тогда оптимальная прогнозирующая статистика (4) имеет вид:* $\hat{x}_{T+1} = \theta_1\mu(T,p) + \sum_{i=2}^{p}\theta_i x_{T-i+1} + \theta_1\sigma\Psi(a_T, b_T, \mu(T,p), \sigma, 1, 1)$.

Доказательство. Известно [1], что для модели АР(p):

$$p(x_T | x_{T-1},\ldots,x_1) = \mathrm{n}(x_T | \mu(T,p), \sigma^2). \tag{11}$$

Из теоремы 2 при $q=1$ имеем:

$$\hat{x}_{T+1} = \frac{\int_{a_T}^{b_T}\mu(T+1,p)p(x_T|x_{T-1},\ldots,x_1)dx_T}{\int_{a_T}^{b_T}p(x_T|x_{T-1},\ldots,x_1)dx_T} = \sum_{i=2}^{p}\theta_i x_{T-i+1} + \frac{\theta_1\int_{a_T}^{b_T}x_T\mathrm{n}(x_T|\mu(T,p),\sigma^2)dx_T}{\int_{a_T}^{b_T}\mathrm{n}(x_T|\mu(T,p),\sigma^2)dx_T},$$

воспользовавшись (8), (9), получим требуемое соотношение. ■

Следствие 3. *Пусть в рамках модели (1), (2) наблюдаются все значения* $x_1,\ldots,x_T$ *и случайное событие* $A_{T+1}^* = \{x_{T+1} \in [a_{T+1}, b_{T+1})\}$. *Тогда оптимальная прогнозирующая статистика (4) имеет вид:*

$$\hat{x}_{T+1} = \mu(T+1,p) + \sigma\Psi(a_{T+1}, b_{T+1}, \mu(T+1,p), \sigma, 1, 1). \tag{12}$$

Доказательство. Воспользовавшись теоремой 1, леммой 1 и (11), получим:

$$\hat{x}_{T+1} = \frac{\int_{a_{T+1}}^{b_{T+1}}x_{T+1}p(x_{T+1}|x_T,\ldots,x_1)dx_{T+1}}{\int_{a_{T+1}}^{b_{T+1}}p(x_{T+1}|x_T,\ldots,x_1)dx_{T+1}} = \frac{\int_{a_{T+1}}^{b_{T+1}}x_{T+1}\mathrm{n}(x_{T+1}|\mu(T+1,p),\sigma^2)dx_{T+1}}{\int_{a_{T+1}}^{b_{T+1}}\mathrm{n}(x_{T+1}|\mu(T+1,p),\sigma^2)dx_{T+1}}.$$

Применяя теперь (8) и (9), получим требуемое соотношение (12). ■

Известно [1], что в случае «полных данных» и отсутствия дополнительной информации об x_{T+1} оптимальная прогнозирующая статистика имеет вид $\hat{x}_{T+1} = \mu(T+1,p)$. Таким образом, второе слагаемое в (12) можно интерпретировать, как поправку на дополнительную информацию о том, что $x_{T+1} \in [a_{T+1}, b_{T+1})$.

Если среди последних q значений временного ряда имеются не только цензурированные наблюдения, но и k $(1 \le k \le q)$ известных наблюдений $x_{l_1},\ldots,x_{l_k}$ $(T-q+1 \le l_1 < \ldots < l_k \le T)$, то оптимальная прогнозирующая статистика может быть получена из (7) предельным переходом $b_{l_1} \to a_{l_1},\ldots,b_{l_k} \to a_{l_k}$.

3. Случай авторегрессии первого порядка. Рассмотрим частный случай модели (1) — модель авторегрессии первого порядка p=1 (АР(1)):

$$x_t = \theta x_{t-1} + u_t, t \in \mathrm{Z}, \tag{13}$$

причем предполагается, что параметры модели θ и σ известны. Известно [1], что временной ряд, описываемый моделью (13), является марковским временным рядом (МВР).

Лемма 2. *Для МВР* x_t *при любых* $T > q \ge 1$ *справедливо соотношение:*

$$p(x_T,\ldots,x_{T-q+1}|x_{T-q},\ldots,x_1) = p(x_T,\ldots,x_{T-q+1}|x_{T-q}), x_1,\ldots,x_T \in \mathrm{R}^1.$$

Доказательство. Воспользуемся известными равенствами для МВР:

$$p(x_T,\ldots,x_1) = p(x_1)\prod_{t=2}^{T}p(x_t|x_{t-1}),\ p(x_T,\ldots,x_{T-q}) = p(x_{T-q})\prod_{t=T-q+1}^{T}p(x_t|x_{t-1}).$$

Преобразуем условную плотность распределения вероятностей:

$$p\left(x_T,\ldots,x_{T-q+1}\middle|x_{T-q},\ldots,x_1\right)=\left(p(x_1)\prod_{t=2}^{T}p\left(x_t\middle|x_{t-1}\right)\right)\left(p(x_1)\prod_{t=2}^{T-q}p\left(x_t\middle|x_{t-1}\right)\right)^{-1}=\prod_{t=T-q+1}^{T}p\left(x_t\middle|x_{t-1}\right)=$$

$$=\left(p\left(x_{T-q}\right)\prod_{t=T-q+1}^{T}p\left(x_t\middle|x_{t-1}\right)\right)\left(p(x_{T-q})\right)^{-1}=p\left(x_T,\ldots,x_{T-q+1}\middle|x_{T-q}\right).\quad\blacksquare$$

Лемма 3. *Если x_t — произвольный МВР, то условная плотность распределения вероятностей наблюдения x_T при условии, что зарегистрированы случайные события $A_T^*,\ldots,A_{T-q+1}^*$ и наблюдения $x_{T-q},\ldots,x_1$, не зависит от «дальней предыстории» $\{x_{T-q-1},\ldots,x_1\}$:*

$$p\left(x_T\middle|A_T^*,\ldots,A_{T-q+1}^*,x_{T-q},\ldots,x_1\right)=p\left(x_T\middle|A_T^*,\ldots,A_{T-q+1}^*,x_{T-q}\right).$$

Доказательство. Согласно лемме 1:

$$p\left(x_T\middle|A_T^*,\ldots,A_{T-q+1}^*,x_{T-q},\ldots,x_1\right)=\frac{\int\limits_{A_{T-1}}\ldots\int\limits_{A_{T-q+1}}1_{A_T}(x_T)p\left(x_T,\ldots,x_{T-q+1},\middle|x_{T-q},\ldots,x_1\right)dx_{T-q+1}\ldots dx_{T-1}}{\int\limits_{A_T}\ldots\int\limits_{A_{T-q+1}}p\left(x_T,\ldots,x_{T-q+1},\middle|x_{T-q},\ldots,x_1\right)dx_{T-q+1}\ldots dx_T}.$$

Далее, применив лемму 2, получим требуемое соотношение. $\blacksquare$

Примем обозначения:

$$I_1(x_{T-m};l,m)=\int_{a_T}^{b_T}\ldots\int_{a_{T-m+1}}^{b_{T-m+1}}x_T^l\,p(x_T,\ldots,x_{T-m+1}\mid x_{T-m})dx_{T-m+1}\ldots dx_T\ ,$$

$$I_2(l,m)=\mathrm{E}\{I_1(x_{T-m};l,m)\},\ m\in\mathrm{N}\ ,\ l\in\{0,1,2\}.$$

Теорема 3. *Пусть для модели (13) наблюдаются значения $x_1,\ldots,x_{T-q}$ и случайные события A_{T-q+1}^*, ..., A_T^*, $0<q<T$. Тогда оптимальная прогнозирующая статистика и её условный риск имеют вид:*

$$\hat{x}_{T+1}=f_0\left(A_T^*,\ldots,A_{T-q+1}^*,x_{T-q}\right)=\theta\frac{I_1(x_{T-q};1,q)}{I_1(x_{T-q};0,q)},\tag{14}$$

$$r_T(f_0)=\sigma^2+\theta^2\left(\frac{I_1(x_{T-q};2,q)}{I_1(x_{T-q};0,q)}-\left(\frac{I_1(x_{T-q};1,q)}{I_1(x_{T-q};0,q)}\right)^2\right)\geq\sigma^2.\tag{15}$$

Доказательство. Оптимальная прогнозирующая статистика (14) находится по теореме 2, с учетом (11) и (13). Вычислим условный риск прогнозирования. Согласно теореме 1, учитывая независимость u_{T+1} от $A_T^*,\ldots,A_{T-q+1}^*,x_{T-q}$ и $\mathrm{E}\{u_{T+1}\}=0$, получим:

$$r_T(f_0)=\mathrm{D}\left\{x_{T+1}\middle|A_T^*,\ldots,A_{T-q+1}^*,x_{T-q}\right\}=\mathrm{E}\left\{\left(x_{T+1}-\hat{x}_{T+1}\right)^2\middle|A_T^*,\ldots,A_{T-q+1}^*,x_{T-q}\right\}=$$

$$=\mathrm{E}\left\{\left(u_{T+1}+\theta\left(x_T-\frac{I_1(x_{T-q};1,q)}{I_1(x_{T-q};0,q)}\right)\right)^2\middle|A_T^*,\ldots,A_{T-q+1}^*,x_{T-q}\right\}=$$

$$=\sigma^2+\theta^2\left(\mathrm{E}\left\{x_T^2\middle|A_T^*,\ldots,A_{T-q+1}^*,x_{T-q}\right\}+\left(\frac{I_1(x_{T-q};1,q)}{I_1(x_{T-q};0,q)}\right)^2-2\frac{I_1(x_{T-q};1,q)}{I_1(x_{T-q};0,q)}\mathrm{E}\left\{x_T\middle|A_T^*,\ldots,A_{T-q+1}^*,x_{T-q}\right\}\right).$$

Согласно лемме 2 и введенным обозначениям $\mathrm{E}\left\{x_T\middle|A_T^*,\ldots,A_{T-q+1}^*,x_{T-q}\right\}=\frac{I_1(x_{T-q};1,q)}{I_1(x_{T-q};0,q)}$,

$\mathrm{E}\left\{x_T^2\middle|A_T^*,\ldots,A_{T-q+1}^*,x_{T-q}\right\}=\frac{I_1(x_{T-q};2,q)}{I_1(x_{T-q};0,q)}$, откуда следует (15). $\blacksquare$

Исследуем ряд частных случав результатов теоремы 3.

Следствие 4. *Пусть выполнены условия теоремы 3 и* $a_T \to -\infty$, ..., $a_{T-q+1} \to -\infty$, $b_T \to +\infty$, ..., $b_{T-q+1} \to +\infty$, *тогда оптимальная прогнозирующая статистика и её условный риск имеют вид* $\hat{x}_{T+1} = f_0\left(A_T^*, \ldots, A_{T-q+1}^*, x_{T-q}\right) = \theta^{q+1} x_{T-q}$, $r_T(f_0) = \sigma^2 \sum_{i=0}^{q} \theta^{2i}$.

Доказательство. Учитывая условия следствия и условие нормировки многомерной плотности распределения вероятностей, имеем: $I_1(x_{T-q};0,q) = 1$. Тогда в силу теоремы 3 и марковости x_t находим

$$\hat{x}_{T+1} = \theta I_1(x_{T-q};1,q) = \theta \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} x_T\, p\left(x_T, \ldots, x_{T-q+1} \middle| x_{T-q}\right) dx_{T-q+1} \ldots dx_T =$$

$$= \theta \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} x_T\, p\left(x_T \middle| x_{T-1}\right) \ldots p\left(x_{T-q+1} \middle| x_{T-q}\right) dx_{T-q+1} \ldots dx_T,$$

откуда, учитывая (11) и свойства нормального закона распределения вероятностей [10], получаем требуемое выражение для оптимальной прогнозирующей статистики:

$$\hat{x}_{T+1} = \theta \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} x_T \mathrm{n}\left(x_T \middle| \theta x_{T-1}, \sigma^2\right) \ldots \mathrm{n}\left(x_{T-q+1} \middle| \theta x_{T-q}, \sigma^2\right) dx_{T-q+1} \ldots dx_T =$$

$$= \theta^2 \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} x_{T-1} \mathrm{n}\left(x_{T-1} \middle| \theta x_{T-2}, \sigma^2\right) \ldots \mathrm{n}\left(x_{T-q+1} \middle| \theta x_{T-q}, \sigma^2\right) dx_{T-q+1} \ldots dx_{T-1} = \ldots = \theta^{q+1} x_{T-q}.$$

Вычислим условный риск, для этого вычислим значение следующего выражения:

$$I_1(x_{T-q};2,q) = \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} \left(x_T - \theta x_{T-1} + \theta x_{T-1}\right)^2 \mathrm{n}\left(x_T \middle| \theta x_{T-1}, \sigma^2\right) \ldots \mathrm{n}\left(x_{T-q+1} \middle| \theta x_{T-q}, \sigma^2\right) dx_{T-q+1} \ldots dx_T =$$

$$= \sigma^2 + \theta^2 \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} x_{\mathrm{T\text{-}1}}^2 \mathrm{n}\left(x_{T-1} \middle| \theta x_{T-2}, \sigma^2\right) \ldots \mathrm{n}\left(x_{T-q+1} \middle| \theta x_{T-q}, \sigma^2\right) dx_{T-q+1} \ldots dx_{T-1} = \ldots = \sigma^2 \sum_{i=0}^{q-1} \theta^{2i} + \theta^{2q} x_{T-q}^2.$$

Подставляя найденное значение в (15), получим доказываемое выражение риска. ■

Условия следствия 4 означают, что в пределе в моменты времени $T, \ldots, T-q+1$ наблюдения $x_T, \ldots, x_{T-q+1}$ пропущены. Такая ситуация является ранее изученной и для неё получены результаты (например, в [8]), которые совпадают с результатами следствия 4.

Следствие 5. *Пусть выполнены условия теоремы 3 и* $a_T \to b_T$, ..., $a_{T-q+1} \to b_{T-q+1}$, *тогда оптимальная прогнозирующая статистика и её условный риск имеют вид:*

$$\hat{x}_{T+1} = f_0\left(x_T\right) = \theta x_T,\ r_T(f_0) = \sigma^2.$$

Доказательство. Воспользуемся предельными соотношениями, справедливыми для непрерывно-дифференцируемой функции $p(x)$:

$$\lim_{\tau \to 0} \frac{\int_a^{a+\tau} x^k p(x) dx}{\int_a^{a+\tau} p(x) dx} = \lim_{\tau \to 0} \frac{\left(\int_a^{a+\tau} x^k p(x) dx\right)'_\tau}{\left(\int_a^{a+\tau} p(x) dx\right)'_\tau} = \lim_{\tau \to 0} \frac{(a+\tau)^k p(x+\tau)}{p(x+\tau)} = a^k, k \in \{1,2\}, \qquad (16)$$

для получения которых используются правило Лопиталя и формула Лейбница. Отметим, что в условиях следствия $a_T \to x_T$, т.к. $x_T \in [a_T, b_T)$ и $a_T \to b_T$. Применяя (16) к (14), (15), получим требуемые соотношения для оптимальной прогнозирующей статистики и ее риска. ■

Условия следствия 5 означают, что в пределе известны значения $x_T = a_T = b_T$, …, $x_{T-q+1} = a_{T-q+1} = b_{T-q+1}$, то есть имеет место случай «полных данных». Для этого случая ранее получены результаты (например, в [1]), которые совпадают с результатами следствия 5.

Следствие 6. *Пусть наблюдаются случайные события* A_1^*, ..., A_T^* $(q = T)$, *т.е. все T наблюдений цензурированы. Тогда оптимальная прогнозирующая статистика и ее условный риск имеют вид:*

$$\hat{x}_{T+1} = f_0\left(A_T^*,\ldots,A_1^*\right) = \theta\frac{I_2(1,T)}{I_2(0,T)},\ r_T(f_0) = \sigma^2 + \theta^2\left(\frac{I_2(2,T)}{I_2(0,T)} - \left(\frac{I_2(1,T)}{I_2(0,T)}\right)^2\right) \geq \sigma^2.$$

Доказательство. Проводится аналогично доказательству теоремы 3. ■

Для случая $q=1$ исследуем зависимость условного риска прогнозирования от длины интервала цензурирования и проведем сравнение оптимальной прогнозирующей статистики $f_0(\cdot)$ с прогнозирующими статистиками, часто используемыми на практике. В этом случае последнее значение x_T временного ряда цензурировано интервалом $[a_T, b_T)$, а предпоследнее значение x_{T-1} известно точно. Поскольку в данном случае результат зависит только от одного интервала цензурирования $A_T = [a_T, b_T)$, то для упрощения обозначений вместо a_T и b_T будем писать a и b.

Теорема 4. *Пусть для модели (13) наблюдаются значение x_{T-1} и случайное событие $A_T^* = \{x_T \in [a,b)\}$, тогда оптимальная прогнозирующая статистика и её условный риск имеют вид:*

$$\hat{x}_{T+1} = f_0\left(A_T^*, x_{T-1}\right) = \theta^2 x_{T-1} + \theta\sigma\Psi\left(a,b,\theta x_{T-1},\sigma,1,1\right), \tag{17}$$

$$r_T(f_0) = (1+\theta^2)\sigma^2 - \left(\theta\sigma\Psi\left(a,b,\theta x_{T-1},\sigma,1,1\right)\right)^2 + \theta^2\sigma\Psi\left(a,b,\theta x_{T-1},\sigma,a-\theta\ x_{T-1},b-\theta\ x_{T-1}\right). \tag{18}$$

Доказательство. Соотношения (17), (18) следуют из теоремы 3 и равенств (8) – (10). ■

Следствие 7. *Пусть выполнены условия теоремы 4, $a \to -\infty$ и $b \to +\infty$. Тогда условный риск прогнозирования для статистики (17) имеет предел $r_T(f_0) \to \sigma^2\left(1+\theta^2\right)$.*

Доказательство. Переходя к пределу в (18), получим требуемый результат. ■

Следствие 8. *В условиях теоремы 4 для условного риска прогнозирования справедливо асимптотическое разложение при $\tau = b - a \to 0$:*

$$r_T(f_0) = \sigma^2 + \theta^2\tau^2/12 - \theta^2\tau^4\left(3a^2 - 6a\theta x_{T-1} + 3\theta^2 x_{T-1}^2 + 2\sigma^2\right)/720\sigma^4 + o(\tau^4),$$

Доказательство. Учитывая дифференцируемость функции $\Psi(\cdot)$ в (18) по τ и используя для этой функции формулу Тейлора с остаточным членом в форме Пеано, получим требуемое соотношение для условного риска. ■

Из следствия 8 получаем, что безусловный риск оптимальной прогнозирующей статистики имеет следующее асимптотическое разложение при $\tau = b - a \to 0$:

$$\mathrm{E}\{r_T(f_0)\} = \sigma^2 + \theta^2\tau^2/12 - \theta^2\tau^4\left(3a^2 + 3\theta^2\sigma^2/\left(1-\theta^2\right) + 2\sigma^2\right)/720\sigma^4 + o(\tau^4). \tag{19}$$

Согласно следствию 5 в случае полных данных $(\tau = 0)$ риск прогнозирования для оптимального прогноза равен $r_0 = \sigma^2$. Для оценки чувствительности риска к длине $\tau = b - a$ интервала цензурирования, вычислим коэффициент неустойчивости риска [3]:

$$\chi = \left(r - r_0\right)/r_0.$$

Следствие 9. *В условиях теоремы 4 для коэффициента неустойчивости риска справедливо асимптотическое разложение при $\tau \to 0$:*

$$\chi_0 = \theta^2\tau^2/12\sigma^2 + \theta^2\tau^4\left(3a^2 - 6a\theta x_{T-1} + 3\theta^2 x_{T-1}^2 + 2\sigma^2\right)/720\sigma^6 + o(\tau^4).$$

Доказательство. Разложение следует из определения χ и следствия 8. ■

Сравним теперь с оптимальной прогнозирующей статистикой (17) другие возможные прогнозирующие статистики, используемые на практике [5 – 7]. Одной из возможных альтернативных прогнозирующих статистик является статистика

$$\hat{x}_{T+1} = f_1(A_T^*, x_T) = \theta\,\mathrm{E}\left\{x_T \mid A_T^*\right\} = \theta\,\mathrm{E}\{x_T \mid x_T \in [a,b)\}. \tag{20}$$

Теорема 5. *Пусть для модели (13) наблюдаются значения $x_1,\ldots,x_{T-1}$ и случайное событие $A_T^* = \{x_T \in [a,b)\}$. Тогда прогнозирующая статистика (20) имеет вид:*

$$\hat{x}_{T+1} = f_1(A_T^*, x_T) = \frac{\theta\sigma}{\sqrt{1-\theta^2}}\Psi\left(a,b,0,\frac{\sigma}{\sqrt{1-\theta^2}},1,1\right), \tag{21}$$

и её условный риск прогнозирования равен:

$$r_T(f_1) = \frac{\sigma^2}{1-\theta^2} - \frac{\theta^2\sigma^2}{1-\theta^2}\left(\Psi\left(a,b,0,\frac{\sigma}{\sqrt{1-\theta^2}},1,1\right)\right)^2 + \frac{\theta^2\sigma}{\sqrt{1-\theta^2}}\Psi\left(a,b,0,\frac{\sigma}{\sqrt{1-\theta^2}},a,b\right). \tag{22}$$

Доказательство. Прогнозирующая статистика (20) имеет вид

$$\hat{x}_{T+1} = \theta\,\mathrm{E}\{x_T \mid x_T \in (a,b)\} = \theta\left(\int_a^b x\mathrm{n}\left(x\middle|0,\tfrac{\sigma^2}{1-\theta^2}\right)dx\right)\left(\int_a^b \mathrm{n}\left(x\middle|0,\tfrac{\sigma^2}{1-\theta^2}\right)dx\right)^{-1}.$$

Далее воспользовавшись (8), (9), получим (21). Условный риск (22) находится по определению, используя (8) – (10). ■

Следствие 10. *Пусть выполнены условия теоремы 5, $a \to -\infty$ и $b \to +\infty$. Тогда условный риск прогнозирования для статистики (20) имеет предел $r_T(f_1) \to \sigma^2/\left(1-\theta^2\right)$.*

Доказательство. Переходя к пределу в (22), получим требуемый результат. ■

Таким образом, при $a \to -\infty$ и $b \to +\infty$ риск оптимальной прогнозирующей статистики (17) меньше риска прогнозирующей статистики (20) на $\sigma^2\theta^4/\left(1-\theta^2\right)$.

Следствие 11. *В условиях теоремы 5 для условного риска прогнозирования и коэффициента неустойчивости риска справедливы асимптотические разложения при $\tau = b - a \to 0$:*

$$r_T(f_1) = \sigma^2 + \theta^2\tau^2/12 - \theta^2\left(1-\theta^2\right)^2\tau^4\left(3a^2 + 2\sigma^2/(1-\theta^2)\right)/720\sigma^4 + o(\tau^4),$$

$$\chi_1 = \theta^2\tau^2/12\sigma^2 + \theta^2\left(1-\theta^2\right)^2\tau^4\left(3a^2 + 2\sigma^2/(1-\theta^2)\right)/720\sigma^6 + o(\tau^4).$$

Доказательство. Проводится аналогично доказательствам следствий 8 и 9. ■

Сравним $\mathrm{E}\{r_T(f_0)\}$ и $r_T(f_1)$ в окрестности $\tau = 0$, учитывая (1), (19):

$$r_T(f_1) - \mathrm{E}\{r_T(f_0)\} = \frac{\theta^2\tau^4\left(3a^2 + 3\theta^2\sigma^2/(1-\theta^2) + 2\sigma^2\right) - \theta^2\left(1-\theta^2\right)^2\tau^4\left(3a^2 + 2\sigma^2/(1-\theta^2)\right)}{720\sigma^4} + o(\tau^4) =$$

$$= \frac{\theta^2\tau^4}{720\sigma^4}\left(3\theta^2\sigma^2/(1-\theta^2) + 2\theta^2\sigma^2 + 3a^2\theta^2\left(2-\theta^2\right)\right) + o(\tau^4).$$

Таким образом, при τ близких к нулю, усредненный риск оптимальной прогнозирующей статистики (17) меньше риска прогнозирующей статистики (20).

Рассмотрим ещё одну возможную прогнозирующую статистику:

$$\hat{x}_{T+1} = f_2\left(A_T^*\right) = \theta\frac{a+b}{2}. \tag{23}$$

Теорема 6. *Пусть для модели (13) наблюдаются значения $x_1,\ldots,x_{T-1}$, случайное событие $A_T^* = \{x_T \in [a,b)\}$. Тогда условный риск прогнозирования для статистики (23) равен:*

$$r_T(f_2) = \frac{\sigma^2}{1-\theta^2} + \frac{\theta^2(a+b)^2}{4} - \frac{\theta^2\sigma}{\sqrt{1-\theta^2}}\Psi\left(a,b,0,\frac{\sigma}{\sqrt{1-\theta^2}},b,a\right). \tag{24}$$

Доказательство. Следует из определения условного риска прогнозирования (3) и соотношений (8) – (10). ■

Следствие 12. *Пусть выполнены условия теоремы 6 и $\tau = b - a \to 0$. Тогда для условного риска справедливо следующее асимптотическое разложение:*

$$r_T(f_2) = \sigma^2 + \theta^2\tau^2/12 + \theta^2\left(1-\theta^2\right)^2\tau^4\left(a^2 - \sigma^2/(1-\theta^2)\right)/360\sigma^4 + o(\tau^4).$$

Доказательство. Проводится аналогично доказательству следствия 8. ■

Очевидно, что при τ близких к нулю, риск прогнозирования статистики (23) значительно больше, чем риск прогнозирования для статистик (17) и (20).

4. Численные результаты. Для сравнения прогнозирующих статистик (17), (20) и (23) проведены компьютерные эксперименты. Для оценивания риска прогнозирования при каж-

дой фиксированной длине интервала цензурирования τ использовался метод Монте-Карло с числом прогонов $N = 10000$. Использованы следующие значения параметров: $p = 1$, $\theta = 0.8$, $q = 1$, $p = 1$, $T = 100$, $\tau \in \{0,0.5,\ldots,15\}$. Моделируется временной ряд длины $T+1$. Предпоследнее наблюдение сгенерированного временного ряда x_T заменяется случайным интервалом цензурирования $[a_T, b_T)$ длины τ следующим образом: генерируется случайная величина α, равномерно распределённая на отрезке $[0,1]$, и вычисляются границы интервала цензурирования: $a_T = x_T - \alpha\tau$ и $b_T = x_T + \alpha(1-\tau)$. Результаты численных экспериментов представлены на рисунке 1.

На рисунке 1(а) изображены графики зависимостей экспериментальных значений риска для всех трёх прогнозирующих статистик от τ. Как видно из рисунка, оптимальная прогнозирующая статистика (17) имеет наименьший риск, риск статистики (20) принимает почти в два раза большие значения, а риск статистики (23) возрастает очень быстро и уже при малых τ принимает достаточно большие значения.

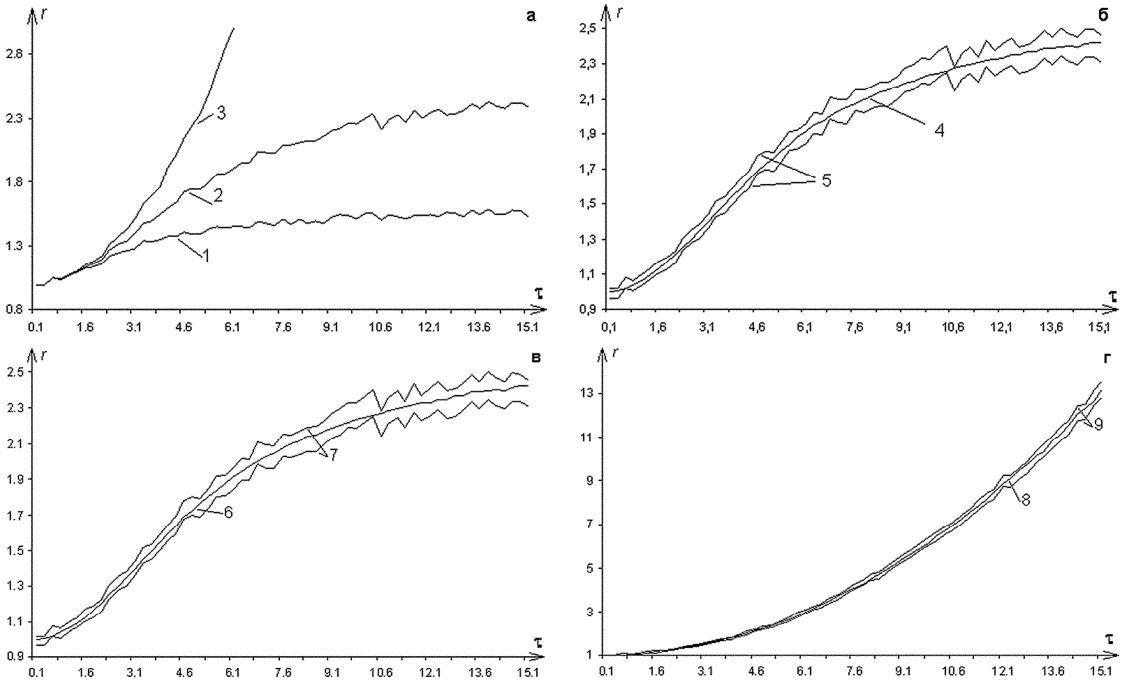

Рис.1. Результаты численных экспериментов: а) Сравнение трех прогнозирующих статистик: 1 –риск оптимальной прогнозирующей статистики (16), 2 –риск статистики (18), 3 – риск статистики (21); б) Сравнение теоретического и экспериментального значения риска оптимальной прогнозирующей статистики (16): 4 – теоретическое значение риска статистики (16), 5 – 95% доверительные границы; в) Сравнение теоретического и экспериментального значения риска прогнозирующей статистики (18): 6 – теоретическое значение риска статистики (18), 7 – 95% доверительные границы; г) Сравнение теоретического и экспериментального значения риска прогнозирующей статистики (21): 8 – теоретическое значение риска статистики (21), 9 – 95% доверительные границы

На рисунках 1(б) – 1(г) изображены усредненные теоретические значения риска прогнозирования для статистик (17), (20) и (23) в зависимости от τ, вычисленные по формулам (18), (22) и (24) соответственно, и 95%-ные доверительные границы риска. Экспериментальные и теоретические значения риска находятся в хорошем согласии.

Литература

1. Бокс Дж., Дженкинс Г. Анализ временных рядов. Прогноз и управление. М., 1974.
2. Андерсон Т. Статистический анализ временных рядов. М., 1976.
3. Харин Ю.С. Оптимальность и робастность в статистическом прогнозировании. Мн., 2008.

4. Литтл Р. Дж. А., Рубин Д.Б. Статистический анализ данных с пропусками. М., 1990.
5. Park J. W., Genton M. G., Ghosh S. K. The Canadian Journal of Statistics. 2007. Vol.35, № 1. P. 151–168.
6. Gomez G., Espinal A., Lagakos W. Statistics in medicine. 2003. № 22. P. 409–425.
7. Zeng D., Lin D.Y. Journal of Royal statistical society. Series B. 2007. № 69, part 4. P. 507–564.
8. Харин Ю.С., Гурин А.С. Искусственный интеллект. 2005. № 4. С. 292–301.
9. Градштейн И.С., Рыжик И.М. Таблицы интегралов, сумм, рядов и произведений. М., 1963.
10. Андерсон Т.В. Введение в многомерный статистический анализ данных. М., 1963.