УДК 519.2

Е. Е. ЖУК

СТАТИСТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ БЛИЖАЙШИХ СТАЦИОНАРНЫХ ВРЕМЕННЫХ РЯДОВ В ПРОСТРАНСТВЕ КОЭФФЦИЕНТОВ АВТОРЕГРЕССИИ

Белорусский государственный университет, Минск, Беларусь,
e-mail: zhukee@mail.ru

Исследуется проблема статистического определения ближайших стационарных в широком смысле временных рядов на основе их описания авторегрессионными моделями. Предлагается использовать решающие правила в пространстве коэффициентов авторегрессии. В качестве меры эффективности принимаемых решений аналитически вычислен риск (вероятность ошибочно определить ближайшие временные ряды). Рассмотрен случай двух классов.

Ключевые слова: стационарный временной ряд, авторегрессионная модель, реализация, решающее правило, риск.

E. E. ZHUK

STATISTICAL DETERMINATION OF THE NEAREST STATIONARY TIME SERIES IN A SPACE OF AUTOREGRESSIVE COEFFICIENTS

Belarusian State University, Minsk, Belarus,e-mail: zhukee@mail.ru

The problem of statistical determination of the nearest stationary time series is considered. The decision rules in a space of autoregressive coefficients are proposed and their efficiency is analytically investigated. The case of two classes is studied.

Keywords: stationary time series, autoregressive model, realization, decision rule, risk.

1. Математическая модель и постановка задач. Как известно, при весьма общих условиях регулярности стационарный в широком смысле временной ряд (ВР) $\{x_t\}_{t\in Z}$ ($Z=\{0,\pm1,\pm2,\ldots\}$ – множество целых чисел) с отсчетами $x_t \in R$, $t\in Z$, имеющими нулевое математическое ожидание: $\mathrm{E}\{x_t\}=0$, $t\in Z$, можно представить в виде разложения Вольда [1, 2]:

$$x_t + \sum_{j=1}^{+\infty} \beta_j x_{t-j} = u_t,\ \ t\in Z; \qquad \sum_{j=1}^{+\infty} \beta_j^2 < +\infty, \tag{1}$$

где случайные величины $\{u_t\}_{t\in Z}$ (ошибки наблюдений) некоррелированы и имеют нулевые математические ожидания и одинаковую ограниченную дисперсию:

$$\mathrm{E}\{u_t\}=0,\ \ \mathrm{D}\{u_t\}=\mathrm{E}\{u_t^2\}=\sigma^2<+\infty; \tag{2}$$
$$\mathrm{E}\{u_t u_l\}=0,\ \ \forall t,l\in Z,\ l\neq t.$$

Разложение (1), (2) представляет собой так называемую модель авторегрессии бесконечного порядка [1, 2] и однозначно определяется коэффициентами авторегрессии $\beta=(\beta_j)_{j=1}^{+\infty}$. Модель (1), (2) обозначим как АР($+\infty,\beta,\sigma^2$).

С другой стороны, из (1) видно, что $\beta_j \to 0$, $j\to+\infty$, и на практике порядок авторегрессии выбирают конечным [1, 2], пренебрегая близкими к нулю значениями коэффициентов и полагая $\beta_j=0$, $j>p$, где p – порядок авторегрессии. Соотношение (1) при этом принимает вид

$$x_t + \sum_{j=1}^{p} \beta_j x_{t-j} = u_t, \quad t \in Z, \tag{3}$$

и временной ряд $\{x_t\}_{t\in Z}$ описывается моделью АР($p, \beta_{(p)}, \sigma^2$) [1, 2], где $\beta_{(p)} = (\beta_1, ..., \beta_p)' \in R^p$ – p-вектор коэффициентов авторегрессии («′» – символ транспонирования), а случайные величины $\{u_t\}_{t\in Z}$ определены в (2).

Пусть наряду с ВР $\{x_t\}_{t\in Z}$, определяемым моделью АР($p, \beta_{(p)}, \sigma^2$) из (2), (3), имеется $L \geq 2$ стационарных в широком смысле ВР $\{x_t^{(i)}\}_{t\in Z}$, $i \in S$, где $S = \{1, \ldots, L\}$ – множество номеров этих рядов. ВР $\{x_t^{(i)}\}_{t\in Z}$ задается своей моделью авторегрессии АР($p, \beta_{(p)}^{(i)}, \sigma_{(i)}^2$) типа (2), (3), где $\beta_{(p)}^{(i)} = (\beta_1^{(i)}, ..., \beta_p^{(i)})'$ и $\sigma_{(i)}^2$ – соответствующие значения коэффициентов авторегрессии и дисперсии ошибок наблюдений ($i \in S$). Порядок авторегрессии p везде считается одинаковым и выбирается исходя из соображений: $\beta_j = 0$, $\beta_j^{(i)} = 0$, $i \in S$, $j > p$. Задача заключается в нахождении для ВР $\{x_t\}_{t\in Z}$ «ближайшего» среди $L \geq 2$ ВР $\{x_t^{(i)}\}_{t\in Z}$, $i \in S$.

Если все коэффициенты авторегрессионных моделей $\beta_{(p)} = (\beta_1, ..., \beta_p)'$ и $\beta_{(p)}^{(i)} = (\beta_1^{(i)}, ..., \beta_p^{(i)})'$, $i \in S$, заданы (известны), то в качестве решения задачи можно предложить

$$D^o = \left\{ k : \left| \beta_{(p)} - \beta_{(p)}^{(k)} \right| = \min_{i\in S} \left| \beta_{(p)} - \beta_{(p)}^{(i)} \right| \right\}, \tag{4}$$

где $D^o \subseteq S$ – множество номеров тех ВР из $\{x_t^{(i)}\}_{t\in Z}$, $i \in S$, к которым ВР $\{x_t\}_{t\in Z}$ ближе всего в смысле расстояний Евклида между векторами коэффициентов авторегрессии (учтено, что могут быть совпадающие по значению расстояния).

Если в (4) нет совпадающих между собой расстояний, то множество D^o состоит из одного элемента (имеется один ближайший ВР):

$$D^o = \{d^o\}, \quad d^o = d\left(\beta_{(p)}, \left\{\beta_{(p)}^{(i)}\right\}_{i\in S}\right) = \arg\min_{i\in S} \left| \beta_{(p)} - \beta_{(p)}^{(i)} \right|, \tag{5}$$

где $d(\beta_{(p)}, \{\beta_{(p)}^{(i)}\}_{i\in S}) \in S$ – так называемое решающее правило (РП) L-средних [2–4].

По аналогии с [5] рассмотрим далее три случая априорной неопределенности.

I. Наблюдается реализация $X = \{x_t\}_{t=1}^T$ длительности T ВР $\{x_t\}_{t\in Z}$, а сами коэффициенты авторегрессии $\beta_{(p)}$ для него неизвестны. При этом коэффициенты $\left\{\beta_{(p)}^{(i)}\right\}_{i\in S}$ ВР $\left\{x_t^{(i)}\right\}_{t\in Z}$, $i \in S$, заданы.

II. Коэффициенты авторегрессии $\beta_{(p)}$ ВР $\{x_t\}_{t\in Z}$ известны, а вместо $\left\{\beta_{(p)}^{(i)}\right\}_{i\in S}$ предложены реализации $X^{(i)} = \left\{x_t^{(i)}\right\}_{t=1}^{T_i}$, $i \in S$, соответствующих им временных рядов.

III. Все временные ряды представлены своими реализациями $X = \{x_t\}_{t=1}^T$ и $X^{(i)} = \{x_t^{(i)}\}_{t=1}^{T_i}$, $i \in S$.

Для решения задачи воспользуемся подстановочным принципом [2], применив его к РП L-средних из (5).

2. Решающие правила в пространстве МНК-оценок коэффициентов авторегрессии и их риск. Построим по реализациям $X = \{x_t\}_{t=1}^T$ и $X^{(i)} = \{x_t^{(i)}\}_{t=1}^{T_i}$, $i \in S$, МНК-оценки [2] параметров $\beta_{(p)}$ и $\{\beta_{(p)}^{(i)}\}_{i\in S}$ соответственно:

$$\hat{\beta}_{(p)} = -\left(\sum_{t=p+1}^{T} X_t X_t'\right)^{-1} \sum_{t=p+1}^{T} x_t X_t ; \qquad \hat{\beta}_{(p)}^{(i)} = -\left(\sum_{t=p+1}^{T_i} X_t^{(i)} (X_t^{(i)})'\right)^{-1} \sum_{t=p+1}^{T_i} x_t^{(i)} X_t^{(i)} , \ i \in S , \qquad (6)$$

где предполагается, что $T > p$ и $T_i > p$, $i \in S$, и обозначено: $X_t = (x_{t-1},...,x_{t-p})' \in R^p$, $t = \overline{p+1,T}$; $X_t^{(i)} = (x_{t-1}^{(i)},...,x_{t-p}^{(i)})' \in R^p$, $t = \overline{p+1,T}$, $i \in S$.

В зависимости от уровня априорной неопределенности подстановочные РП, основанные на РП L-средних из (5), будут иметь вид

$$d_I = d\left(\hat{\beta}_{(p)}, \left\{\beta_{(p)}^{(i)}\right\}_{i\in S}\right), \ d_{II} = d\left(\beta_{(p)}, \left\{\hat{\beta}_{(p)}^{(i)}\right\}_{i\in S}\right), \ d_{III} = d\left(\hat{\beta}_{(p)}, \left\{\hat{\beta}_{(p)}^{(i)}\right\}_{i\in S}\right), \qquad (7)$$

где учтено, что статистические оценки $\hat{\beta}_{(p)}$ и $\left\{\hat{\beta}_{(p)}^{(i)}\right\}_{i\in S}$ имеют абсолютно непрерывные распределения вероятностей [2], поэтому вероятность совпадения расстояний равна нулю и решения в (7) выносятся однозначно (как и в (5)).

В качестве меры эффективности РП (7) по аналогии с [4, 5] определим риск как вероятность ошибочно определить пару ближайших в смысле (4) временных рядов:

$$r_I = \mathrm{P}\{d_I \notin D^o\}, \ r_{II} = \mathrm{P}\{d_{II} \notin D^o\}, \ r_{III} = \mathrm{P}\{d_{III} \notin D^o\}. \qquad (8)$$

Отметим, что если $D^o = S$, то $r_I = r_I = r_{III} = 0$ и выносимое РП (7) решение не принципиально. Если множество D^o состоит из одного элемента (один ближайший ВР), то

$$r_I = \mathrm{P}\{d_I \neq d^o\}, \ r_{II} = \mathrm{P}\{d_{II} \neq d^o\}, \ r_{III} = \mathrm{P}\{d_{III} \neq d^o\}, \ d^o = \arg\min_{i\in S}\left|\beta_{(p)} - \beta_{(p)}^{(i)}\right|. \qquad (9)$$

Чем меньше (ближе к нулю) значения риска из (8), (9), тем эффективнее выносимые при помощи РП (7) решения.

Вычислим здесь риск r_I (случаи II и III аналогичны в предположении, что реализации $X = \{x_t\}_{t=1}^{T}$ и $X^{(i)} = \{x_t^{(i)}\}_{t=1}^{T_i}$, $i \in S$, независимы в совокупности [2]). Введем обозначение:

$$n_N(y \mid \mu, \Sigma) = (2\pi)^{-\frac{N}{2}} (\det(\Sigma))^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y-\mu)'\Sigma^{-1}(y-\mu)\right), \ y \in R^N ,$$

– плотность N-мерного нормального (гауссовского) закона распределения вероятностей (закон $N_N(\mu,\Sigma)$ [2]) с вектором математического ожидания $\mu \in R^N$ и невырожденной ковариационной $(N \times N)$-матрицей Σ ($\det(\Sigma) \neq 0$).

Теорема. *Пусть* $X = \{x_t\}_{t=1}^{T}$ *– реализация длительности* T *временного ряда, определяемого моделью* АР($p, \beta_{(p)}, \sigma^2$) *из* (3)*, где случайные величины* $\{u_t\}_{t\in Z}$ *независимы в совокупности и одинаково распределены по нормальному закону* $N_1(0,\sigma^2)$ ($\sigma^2 < +\infty$)*, а корни характеристического уравнения*

$$\sum_{j=1}^{p} \beta_j z^{p-j} = 0$$

лежат внутри единичного круга ($|z| < 1$)*, тогда в условиях* (5) *риск* $r_I = r_I(T) = \mathrm{P}\{d_I \neq d^o\}$ *РП* $d_I = d\left(\hat{\beta}_{(p)}, \left\{\beta_{(p)}^{(i)}\right\}_{i\in S}\right)$ *в асимптотике растущей длительности:* $T \to +\infty$*, удовлетворяет соотношению:*

$$r_I(T)/\widetilde{r}_I(T) \to 1, \;\; T \to +\infty; \qquad (10)$$

$$\tilde{r}_I(T) = 1 - \int_{R^p} \prod_{\substack{j \in S \\ j \neq d^o}} U\left(\left|y - \beta_{(p)}^{(j)}\right| - \left|y - \beta_{(p)}^{(d^o)}\right|\right) n_p\left(y \middle| \beta_{(p)}, \frac{1}{T}\sigma^2 \Sigma_{p,p}^{-1}\right) dy,$$

где $U(z) = \{1,$ *если* $z \geq 0;\; 0,$ *если* $z < 0\}$ *– единичная функция Хэвисайда,* d^o *– истинный номер ближайшего ВР из* (5), *а* $\Sigma_{p,p} = \Sigma_{p,p}(\beta_{(p)}) = (\sigma(i-j))_{i,j=1}^{p}$ *– ковариационная* $(p \times p)$*-матрица, элементы которой определяются ковариационной функцией* $\sigma(\tau) = \mathrm{cov}\{x_t, x_{t+\tau}\} = \mathrm{E}\{x_t x_{t+\tau}\}$, $\forall\, t, \tau \in Z$*, вычисляемой по* $\beta_{(p)}$ *из уравнений Юла – Уокера* [1, 2]:

$$\sigma(\tau) + \sum_{j=1}^{p} \beta_j \sigma(\tau - j) = 0\,, \;\; \tau = \overline{1, p}\,.$$

Доказательство. В условиях теоремы, согласно известным результатам [1, 2], при $T \to +\infty$ МНК-оценка $\hat{\beta}_{(p)}$ из (6) асимптотически нормально распределена:

$$L\left\{\sqrt{T}\left(\hat{\beta}_{(p)} - \beta_{(p)}\right)\right\} \to N_p\left(0_p, \sigma^2 \Sigma_{p,p}^{-1}\right), \qquad (11)$$

где 0_p – нулевой p-вектор, а $\Sigma_{p,p} = \left(\sigma(i-j)\right)_{i,j=1}^{p}$ – ковариационная $(p \times p)$-матрица, описанная в условиях теоремы.

Для риска r_I из (9), с учетом вида соответствующего ему РП из (7):

$$d_I = d\left(\hat{\beta}_{(p)}, \left\{\beta_{(p)}^{(i)}\right\}_{i \in S}\right) = \arg\min_{i \in S} \left|\hat{\beta}_{(p)} - \beta_{(p)}^{(i)}\right|,$$

справедлива цепочка равенств:

$$r_I = \mathrm{P}\{d_I \neq d^o\} = 1 - \mathrm{P}\{d_I = d^o\} = 1 - \mathrm{P}\left\{\bigcap_{\substack{j \in S \\ j \neq d^o}} \left\{\left|\hat{\beta} - \beta^{(j)}\right| \geq \left|\hat{\beta} - \beta^{(d^o)}\right|\right\}\right\},$$

откуда с учетом (11) и получаем доказываемое асимптотическое соотношение (10).

Практическая значимость результата (10) состоит в том, что он позволяет при больших значениях длительности реализации ($T \to +\infty$) приближенно вычислить риск: $r_I \approx \tilde{r}_I$, аналитически оценив эффективность принимаемых решений. Однако простой вид величина $\widetilde{r}_I = \widetilde{r}_I(T)$ из (10) принимает лишь при $L = 2$.

3. Случай отнесения реализаций к двум временным рядам. Пусть своими коэффициентами $\beta_{(p)}^{(1)}$ и $\beta_{(p)}^{(2)}$ заданы два авторегрессионных временных ряда ($L = 2$, $S = \{1,2\}$). И к ближайшему из них в смысле коэффициентов авторегрессии необходимо отнести реализацию $X = \{x_t\}_{t=1}^{T}$ длительности $T > p$ стационарного временного ряда, описываемого моделью АР($p, \beta_{(p)}, \sigma^2$). Соответствующее РП из (7) принимает при $L = 2$ вид

$$d_I = d\left(\hat{\beta}_{(p)}, \left\{\beta_{(p)}^{(i)}\right\}_{i \in S}\right) = U\left(\left|\hat{\beta}_{(p)} - \beta_{(p)}^{(1)}\right| - \left|\hat{\beta}_{(p)} - \beta_{(p)}^{(2)}\right|\right) + 1. \qquad (12)$$

Вычислим риск РП (12).

Следствие. *Пусть в условиях теоремы* $L = 2$*, тогда величина* $\widetilde{r}_I(T)$ *из* (10) *может быть представлена в виде:*

$$\tilde{r}_I(T) = \Phi\left(-\sqrt{T}\,\frac{\left| \left|\beta_{(p)} - \beta_{(p)}^{(1)}\right|^2 - \left|\beta_{(p)} - \beta_{(p)}^{(2)}\right|^2 \right|}{2\sigma\Delta} \right), \tag{13}$$

где $\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} \exp\left(-\frac{w^2}{2}\right)dw,\ z \in R$ *– функция распределения вероятностей стандартного нормального закона* $N_1(0{,}1)$*, а величина*

$$\Delta = \sqrt{\left(\beta_{(p)}^{(1)} - \beta_{(p)}^{(2)}\right)' \Sigma_{p,p}^{-1} \left(\beta_{(p)}^{(1)} - \beta_{(p)}^{(2)}\right)} \tag{14}$$

– аналог расстояния Махаланобиса [2] *между* $\beta_{(p)}^{(1)}$ *и* $\beta_{(p)}^{(2)}$ *относительно ковариационной матрицы* $\Sigma_{p,p}$ *из* (10).

Доказательство. Из (12) с учетом (11) получаем:

$$\tilde{r}_I(T) = \begin{cases} 1 - \mathrm{P}\left\{ \left|\xi - \beta_{(p)}^{(2)}\right|^2 - \left|\xi - \beta_{(p)}^{(1)}\right|^2 \geq 0 \right\}, \text{ если } d^o = 1; \\ \mathrm{P}\left\{ \left|\xi - \beta_{(p)}^{(2)}\right|^2 - \left|\xi - \beta_{(p)}^{(1)}\right|^2 \geq 0 \right\}, \text{ если } d^o = 2, \end{cases} \tag{15}$$

где случайный p-вектор $\xi \in R^p$ распределен по нормальному закону $N_p\left(\beta_{(p)}, \sigma^2\Sigma_{p,p}^{-1}/T\right)$, а случайная величина

$$\left|\xi - \beta_{(p)}^{(2)}\right|^2 - \left|\xi - \beta_{(p)}^{(1)}\right|^2 = 2\left(\xi - \frac{\beta_{(p)}^{(1)} + \beta_{(p)}^{(2)}}{2}\right)\left(\beta_{(p)}^{(1)} - \beta_{(p)}^{(2)}\right) \in R$$

линейна по ξ, и потому является нормальной случайной величиной с математическим ожиданием $\left|\beta_{(p)} - \beta_{(p)}^{(2)}\right|^2 - \left|\beta_{(p)} - \beta_{(p)}^{(1)}\right|^2$ и дисперсией $4\sigma^2\left(\beta_{(p)}^{(1)} - \beta_{(p)}^{(2)}\right)' \Sigma_{p,p}^{-1}\left(\beta_{(p)}^{(1)} - \beta_{(p)}^{(2)}\right)/T = 4\sigma^2\Delta^2/T$. Нормируя данную случайную величину до стандартного нормального закона $N_1(0{,}1)$, из (15) получаем:

$$\tilde{r}_T = \begin{cases} \Phi\left(-\sqrt{T}\,\frac{\left|\beta_{(p)} - \beta_{(p)}^{(2)}\right|^2 - \left|\beta_{(p)} - \beta_{(p)}^{(1)}\right|^2}{2\sigma\Delta} \right), \text{ если } d^o = 1; \\ \Phi\left(\sqrt{T}\,\frac{\left|\beta_{(p)} - \beta_{(p)}^{(2)}\right|^2 - \left|\beta_{(p)} - \beta_{(p)}^{(1)}\right|^2}{2\sigma\Delta} \right), \text{ если } d^o = 2, \end{cases}$$

где учтено известное свойство функции распределения вероятностей закона $N_1(0{,}1)$: $1 - \Phi(z) = \Phi(-z)$, $z \in R$. Но в данном случае, согласно условию теоремы, $\left|\beta_{(p)} - \beta_{(p)}^{(1)}\right| \neq \left|\beta_{(p)} - \beta_{(p)}^{(2)}\right|$ и

$$d^o = \begin{cases} 1, \text{ если } \left|\beta_{(p)} - \beta_{(p)}^{(1)}\right|^2 < \left|\beta_{(p)} - \beta_{(p)}^{(2)}\right|^2; \\ 2, \text{ если } \left|\beta_{(p)} - \beta_{(p)}^{(1)}\right|^2 > \left|\beta_{(p)} - \beta_{(p)}^{(2)}\right|^2, \end{cases}$$

что приводит к (13).

Из (13) следует (напомним: $r_I(T)/\tilde{r}_I(T) \to 1,\ T \to +\infty$), что $r_I = r_I(T) \to 0$, $T \to +\infty$, т. е. с увеличением длительности подлежащей отнесению реализации эффективность принимаемых решений повышается (значение риска уменьшается). Из (13) также видно, что риск уменьшается с увеличением различия между собой расстояний $\left|\beta_{(p)} - \beta_{(p)}^{(1)}\right|$ и $\left|\beta_{(p)} - \beta_{(p)}^{(2)}\right|$ между коэффициентами авторегрессии $\beta_{(p)}$ подлежащего отнесению ВР и коэффициентами $\beta_{(p)}^{(1)}$ и $\beta_{(p)}^{(2)}$ заданных авторегрессионных моделей. При $\left|\beta_{(p)} - \beta_{(p)}^{(1)}\right| = \left|\beta_{(p)} - \beta_{(p)}^{(2)}\right|$ риск заведомо равен нулю: $r_I = r_I(T) = 0$, $\forall T$, а использование представления (13) некорректно, поскольку оно получено в предположении $\left|\beta_{(p)} - \beta_{(p)}^{(1)}\right| \neq \left|\beta_{(p)} - \beta_{(p)}^{(2)}\right|$.

Отметим также, что в [4] подобная задача соотнесения решалась в пространстве ковариационных функций.

Список использованной литературы

1. *Андерсон, Т.* Статистический анализ временных рядов: пер. с англ. / Т. Андерсон. – М.: Мир, 1976.

2. *Харин, Ю. С.* Математическая и прикладная статистика: учеб. пособие / Ю. С. Харин, Е. Е. Жук. – Минск: БГУ, 2005.

3. *Жук, Е. Е.* Непараметрическая статистическая классификация стационарных временных рядов / Е. Е. Жук // Вес. Нац. акад. навук Беларусі. Сер. фіз.-мат. навук. – 2004. – № 4. – С. 26–30.

4. *Жук, Е. Е.* Статистическое отнесение реализаций стационарных временных рядов к классам, определенным в пространстве ковариационных функций / Е. Е. Жук // Вес. Нац. акад. навук Беларусі. Сер. фіз.-мат. навук. – 2015. – № 1. – С. 47–51.

5. *Жук, Е. Е.* Статистическое определение ближайших классов по многомерным обучающим выборкам / Е. Е. Жук // Теория вероятностей, случайные процессы, математическая статистика и их приложения: сб науч. ст. – Минск: РИВШ, 2014. – С. 59–63.

Поступила в редакцию 15.12.2015