

ДИСКРИМИНАНТНЫЙ АНАЛИЗ СТАЦИОНАРНЫХ ВРЕМЕННЫХ РЯДОВ В ПРОСТРАНСТВЕ МП-ОЦЕНОК КОЭФФИЦИЕНТОВ АВТОРЕГРЕССИИ

Г. В. Микулич

Белорусский государственный университет, пр. Независимости, 4,
220030, г. Минск, Беларусь, aragorniguga@gmail.com
Научный руководитель — Е. Е. Жук, доктор физико-математических наук,
профессор

Рассматривается проблема статистической классификации реализаций стационарных в широком смысле временных рядов, различающихся по классам значениями коэффициентов авторегрессии. Для данной проблемы приводится алгоритм дискриминантного анализа и оценка его эффективности.

Ключевые слова: авторегрессия; дискриминантный анализ; оценка максимального правдоподобия.

Пусть наблюдается выборка $\tilde{X}^n = \{X_1, \dots, X_n\}$ размера n , состоящая из наблюдений из $L \geq 2$ классов $\{\Omega_1, \dots, \Omega_L\}$. Наблюдение X_t принадлежит к классу со случайным индексом $d_t^0 \in S$, $S = \{1, 2, \dots, L\}$ ($t = \overline{1, n}$).

При фиксированном значении индекса $d_t^0 = i$ ($i \in S$) наблюдение $X_t = (x_{t1}, \dots, x_{tT_t})' \in R^{T_t}$ является реализацией длины T_t временного ряда авторегрессии порядка $p \geq 1$ (АР(p)):

$$x_t^i + \theta_{i1}^0 x_{t-1}^i + \dots + \theta_{ip}^0 x_{t-p}^i = u_t^i, l \in Z = \{0, \pm 1, \pm 2, \dots\}, \quad (1)$$

где $\theta_i^0 = (\theta_{i1}^0, \dots, \theta_{ip}^0)' \in R^p$ — вектор коэффициентов авторегрессии для i -го класса; $\{u_t^i\}_{t=-\infty}^{+\infty}$ — независимые в совокупности нормальные (гауссовские) случайные величины:

$$E\{u_t^i\} = 0, D\{u_t^i\} = E\{(u_t^i)^2\} = \sigma^2, l \in Z, i \in S. \quad (2)$$

Также классы $\{\Omega_i\}_{i \in S}$ характеризуются априорными вероятностями:

$$\pi_i^0 = P\{d_t^0 = i\} > 0, i \in S \quad (\pi_1^0 + \dots + \pi_L^0 = 1). \quad (3)$$

Таким образом, классы $\{\Omega_i\}_{i \in S}$ однозначно определяются значениями $\{\pi_i^0, \theta_i^0\}_{i \in S}$ и дисперсией σ^2 . Предполагаем относительно значений $\{\theta_i^0\}_{i \in S}$:

$$z: z^p + \sum_{j=1}^p \theta_{ij}^0 z^{p-j} = 0, |z| < 1, i \in S. \quad (4)$$

Задача дискриминантного анализа состоит в следующем [1]: при известном векторе истинной классификации $D^0 = (d_1^0, \dots, d_n^0)' \in S^n$ построить оценки $\hat{d}_{n+1}, \hat{d}_{n+2} \in S$ для индексов классов, к которым принадлежат вновь поступающие наблюдения X_{n+1}, X_{n+2} по исходной выборке $\tilde{X}^n = \{X_1, \dots, X_n\}$.

Преобразуем исходную выборку $\tilde{X}^n = \{X_1, \dots, X_n\}$ в выборку $\tilde{Y}^n = \{Y_1, \dots, Y_n\}$, где $Y_t \in R^p$ ($t = \overline{1, n}$) – оценка максимального правдоподобия для p -вектора коэффициентов авторегрессии $\theta_{d_i^0}^o \in R^p$, построенная по наблюдению $X_t \in R^{T_t}$:

$$\{Y_t, \hat{\sigma}_t\} = \arg \max_{\{\bar{\theta}, \sigma\}} \ln p(X_t; \bar{\theta}, \sigma); \quad (5)$$

при фиксированном индексе класса $d_i^0 = i$ ($i \in S$):

$$p(X_t; \theta_i^0, \sigma) = n_p(X_t^p | O_p, R_p(\theta_i^0, \sigma)) (2\pi)^{-(T_t-p)/2} \sigma^{-(T_t-p)} \times \\ \times \exp \left(-\frac{1}{2\sigma^2} \sum_{l=p+1}^{T_t} \left(x_{tl} + \theta_{il}^0 x_{t, l-1} + \dots + \theta_{ip}^0 x_{t, l-p} \right)^2 \right),$$

где O_p – нулевой p -вектор;

$$X_t = (x_{t1}, \dots, x_{t, T_t})' = ((X_t^p)', x_{t, p+1}, \dots, x_{t, T_t})' \in R^{T_t}; \quad X_t^p = (x_{t1}, \dots, x_{tp})' \in R^p;$$

$n_p(y | \mu, \Sigma) = (2\pi)^{-p/2} (\det(\Sigma))^{-1/2} \exp(-\frac{1}{2}(y - \mu)' \Sigma^{-1}(y - \mu))$ – функция плотности p -мерного ($y \in R^p$) нормального распределения $N_p(\mu, \Sigma)$ с вектором средних $\mu \in R^p$ и ковариационной ($p \times p$)-матрицей Σ ($\det(\Sigma) \neq 0$);

$$R_p(\theta_i^0, \sigma) = \left(\rho_{|k-l|}(\theta_i^0, \sigma) \right)_{k,l=1}^p = E\{X_t^p (X_t^p)' | d^0 = i\}; \quad (6)$$

невыврожденная [2] ковариационная ($p \times p$)-матрица сформирована автоковариациями $\rho_k(\theta_i^0, \sigma) = E\{x_{ty} x_{t, j+k} | d^0 = i\}$, $k = 0, 1, 2, \dots$, определяемыми из уравнений Юла-Уокера [3]:

$$\sum_{j=1}^p \theta_{ij}^0 \rho_j(\theta_i^0, \sigma) + \rho_0(\theta_i^0, \sigma) = \sigma^2; \quad (7)$$

$$\sum_{j=1}^p \theta_{ij}^0 \rho_{|k-j|}(\theta_i^0, \sigma) + \rho_k(\theta_i^0, \sigma) = 0, \quad k = 1, 2, \dots$$

При увеличении числа наблюдений:

$$T_t \rightarrow +\infty, \quad t = \overline{1, n},$$

оценки $\{Y_t\}_{t=1}^n$ могут быть заменены на среднеквадратичные:

$$Y_t = - \left(\sum_{l=p+1}^{T_t} X_{tl}^p (X_{tl}^p)' \right)^{-1} \sum_{l=p+1}^{T_t} x_{tl} X_{tl}^p,$$

$$X_{tl}^p = (x_{t,l-1}, \dots, x_{t,l-p})' \in R^p, \quad t = \overline{1, n},$$

или на оценки Юла-Уокера:

$$Y_t = -(\hat{R}_t^p)^{-1} \hat{r}_t^p, \quad \hat{R}_t^p = (\hat{\rho}_{|k-l|}^t)_{k,l=1}^p, \quad \hat{r}_t^p = (\hat{\rho}_1^t, \dots, \hat{\rho}_p^t)', \quad t = \overline{1, n},$$

где $\hat{\rho}_j^p = \sum_{s=1}^{T_t-j} x_{ts} x_{t,s+j} / (T_t - j)$ — оценка для автоковариации $\rho_j(\theta_{d_t^0}^0, \sigma)$, $j = 0, 1, \dots, p$, построенная по реализации X_t длины T_t .

Пусть в условиях модели (1)–(3) выполняется условие стационарности (4). Тогда при фиксированном векторе $D^0 = (d_1^0, \dots, d_n^0)' \in S^n$ оценки максимального правдоподобия $Y_1, \dots, Y_n$ из (5) являются состоятельными оценками соответствующих коэффициентов авторегрессии ($t = \overline{1, n}$):

$$Y_t \xrightarrow{P} \theta_{d_t^0}^0, \quad T_t \rightarrow +\infty,$$

и имеют асимптотически нормальное распределение:

$$\sqrt{T_t} (Y_t - \theta_{d_t^0}^0) \sim N_p(O_p, \sigma^2 R_p^{-1}(\theta_{d_t^0}^0, \sigma)), \quad T_t \rightarrow +\infty,$$

где ковариационные матрицы $\{R_p(\theta_i^0, \sigma)\}_{i \in S}$ из (6), (7).

Для построения классификации вновь поступающих наблюдений можно использовать следующий алгоритм дискриминантного анализа:

1. По преобразованной выборке $\tilde{Y}^n = \{Y_1, \dots, Y_n\}$ и вектору истинной классификации $D^0 = (d_1^0, \dots, d_n^0)' \in S^n$ оцениваются «центры» классов $\{\theta_i^0\}_{i \in S}$:

- 2.

$$\hat{\theta}_i = \left(\sum_{t=1}^n \delta_{d_t^0, i} \right)^{-1} \sum_{t=1}^n \delta_{d_t^0, i} Y_t, \quad i \in S, \quad (8)$$

где $\delta_{ij} = \{1, j=i; 0, j \neq i\}$ – символ Кронекера.

3. Классифицируется новое наблюдение Y (оценка, построенная по $X \in \{X_{n+1}, X_{n+2}, \dots\}$):

- 4.

$$\hat{d} = d(Y; \hat{\theta}) = \arg \min_{i \in S} |Y - \hat{\theta}_i|. \quad (9)$$

Исследуем теперь эффективность процедуры (8), (9). Алгоритм построен на следующем решающем правиле (РП):

$$d = d(Y; \theta^0) = \arg \min_{i \in S} |Y - \theta_i^0|, \quad Y \in R^p, \quad (10)$$

и на подстановочном РП $\hat{d}_t = d(Y_t; \hat{\theta}) = \arg \min_{i \in S} |Y_t - \hat{\theta}_i|$, $t = 1, n, n+1, n+2, \dots$, полученном из РП (10) подстановкой статистической оценки $\hat{\theta} = (\hat{\theta}_1', \dots, \hat{\theta}_L')'$ вместо неизвестного вектора $\theta^0 \in R^{Lp}$ коэффициентов авторегрессии $\{\theta_i^0\}_{i \in S}$. Оценим риск (вероятность ошибочной классификации) РП (10):

$$r_T = P\{d(Y; \theta^0) \neq d^0\},$$

при увеличении числа наблюдений ($T \rightarrow +\infty$). Это можно сделать, используя следующие соотношения:

$$r_T / r_T^* \rightarrow 1, \quad T \rightarrow +\infty;$$

где $U(\omega) = \{1, \text{if } \omega \geq 0; 0, \text{if } \omega < 0\}$ – единичная функция.

Таким образом, для оценки риска может использоваться значение r_T^* . В случае двух классов ($L=2$):

$$r_T^* = \pi_1^0 \Phi \left(-\sqrt{T} \frac{|\theta_1^0 - \theta_2^0|^2}{2\sigma\Delta(\theta_1^0, \theta_2^0)} \right) + \pi_2^0 \Phi \left(-\sqrt{T} \frac{|\theta_1^0 - \theta_2^0|^2}{2\sigma\Delta(\theta_2^0, \theta_1^0)} \right),$$

где $\Delta(\theta_i^0, \theta_j^0) = \sqrt{(\theta_i^0 - \theta_j^0)' R_p^{-1}(\theta_i^0, \sigma)(\theta_i^0 - \theta_j^0)}$; $\Phi(\cdot)$ – функция стандартного нормального распределения $N_1(0,1)$.

Библиографические ссылки

1. Харин Ю. С., Зуев Н. М., Жук Е. Е. Теория вероятностей, математическая и прикладная статистика: учебник. – Мн. : БГУ, 2011. – 463 с.
2. Бокс Дж., Дженкинс Г. Анализ временных рядов. Прогноз и управление. Вып. 1 М., 1974.
3. Андерсон Т. Статистический анализ временных рядов / Т. Андерсон. М. : Мир, 1976.