ИНФОРМАТИКА

2008 **июль-сентябрь** **№ 3**

УДК 519.2: 681.3

В.И. Малюгин

ДИСКРИМИНАНТНЫЙ АНАЛИЗ МНОГОМЕРНЫХ АВТОКОРРЕЛИРОВАННЫХ РЕГРЕССИОННЫХ НАБЛЮДЕНИЙ В УСЛОВИЯХ ПАРАМЕТРИЧЕСКОЙ НЕОДНОРОДНОСТИ МОДЕЛЕЙ

Рассматривается задача дискриминантного анализа моделей многомерной линейной регрессии с неоднородной структурой и автокоррелированными ошибками наблюдения. Предлагается состоятельное решающее правило классификации многомерных неоднородных автокоррелированных регрессионных наблюдений, а также итерационный алгоритм вычисления оценок параметров модели. Исследуется эффективность итерационного алгоритма на модельных данных.

Введение

Модели статистических зависимостей регрессионного и авторегрессионного типов применяются в разнообразных приложениях [1, 2]. В настоящей статье рассматривается задача дискриминантного анализа статистических зависимостей, описываемых моделью многомерной линейной регрессии со структурной параметрической неоднородностью в предположении, что случайные ошибки наблюдения регрессионных моделей являются автокоррелированными. Автокорреляция ошибок служит причиной статистической зависимости регрессионных наблюдений, что позволяет называть подобные модели моделями динамической регрессии [3]. Отличительной особенностью данной задачи от традиционных задач дискриминантного анализа случайных выборок из смесей распределений является нарушение двух традиционных для указанных задач предположений: предположения относительно одинаковой распределенности наблюдений из одного класса и предположения о независимости наблюдений. Задачи дискриминантного анализа статистических зависимостей в условиях параметрической неоднородности возникают при разработке автоматизированных систем кредитного скоринга [4], при ранжировании коммерческих банков по степени устойчивости [5] и в ряде других приложений.

Задачи дискриминантного и кластерного анализа независимых многомерных регрессионных наблюдений исследовались в [6]. В случае авторегрессионных моделей временных рядов указанные задачи рассматривались в [7, 8]. В настоящей статье в предположении, что случайные ошибки наблюдений многомерных регрессионных моделей описываются стационарной моделью векторной авторегрессии первого порядка, найдены представления для оценок неизвестных параметров модели на основе обобщенного метода наименьших квадратов и показана их состоятельность. Получено представление для подстановочного байесовского решающего правила в форме Зигерта-Котельникова для классификации многомерных зависимых регрессионных наблюдений. Для вычисления оценок параметров моделей, используемых в подстановочном решающем правиле, предлагается использовать итерационный алгоритм. Приводятся результаты экспериментального исследования данного алгоритма на модельных данных.

1. Математическая модель наблюдений и постановка задач исследования

Будем предполагать, что для произвольного момента времени t (t = 1,...,T, …) наблюдаемые объекты описываются случайным вектором наблюдений $\mathbf{y}_t \in \Re^{N+M}$, который допускает разбиение на подвекторы:

$$\mathbf{y}_t = \begin{pmatrix} \mathbf{x}_t \\ \mathbf{z}_t \end{pmatrix} \in \Re^{N+M},\ \mathbf{x}_t \in \Re^N,\ \mathbf{z}_t \in \mathbf{Z} \subset \Re^M, \quad (1)$$

где $\mathbf{x}_t = (x_{t1},...,x_{tN})'$ – вектор эндогенных переменных (признаков), характеризующих состояние анализируемого объекта; $\mathbf{z}_t = (z_{t1},...,z_{tM})'$ – вектор экзогенных переменных (факторов), описывающих внешние воздействия на состояние объектов.

В момент времени *t* наблюдаемый объект может находиться в одном из двух классов состояний. Номер состояния описывается ненаблюдаемой случайной величиной $d_t \in S = \{0,1\}$ с распределением вероятностей

$$P\{d_t = i\} = \pi_{it} > 0 (i \in S), \pi_{0t} + \pi_{1t} = 1, \quad (2)$$

параметры $\{\pi_{it}\}(i \in S)$ при этом соответствуют априорным вероятностям классов состояний в момент времени *t*.

Между подвекторами составного вектора наблюдений (1) существует статистическая зависимость, которая для класса $d_t = i$ $(i \in S)$ описывается моделью многомерной линейной регрессии вида

$$\mathbf{x}_{it} = B_i \mathbf{z}_t + \mathbf{v}_{it},\ i \in S = \{0,1\}, \quad (3)$$

где $B_i = (b_{ike})$ – $(N \times M)$-матрица коэффициентов регрессии $(B_0 \neq B_1)$; $\mathbf{v}_{it} = (v_{it1},...,v_{itN})' \in \Re^N$ – вектор случайных ошибок наблюдения, который описывается моделью векторной авторегрессии первого порядка (vector autoregressive – *VAR*(1)) вида

$$\mathbf{v}_{it} = A_i \mathbf{v}_{it-1} + \mathbf{u}_{it}, \quad (4)$$

для которой $A_i = (a_{ike})$ – $(N \times M)$-матрица коэффициентов авторегрессии; $\mathbf{v}_{i0}$ – начальное значение случайного процесса $\mathbf{v}_{it}$.

Относительно модели (2)–(4) в данном исследовании возможны следующие предположения:

П1. $\mathbf{u}_{it} = (u_{it1},...,u_{itN})' \in \Re^N$ – независимые одинаково распределенные случайные векторы с нулевым средним значением и невырожденной ковариационной $(N \times N)$-матрицей $\Sigma_i = (\sigma_{ikl}) : E\{\mathbf{u}_{it}\} = 0,\ E\{\mathbf{u}_{it}\mathbf{u}'_{i\tau}\} = \delta_{t\tau}\Sigma_i$, где $\delta_{t\tau}$ – символ Кронекера.

П2. Случайный процесс $\mathbf{u}_{it} \in \Re^N$ является гауссовским, для него используется обозначение $\{\mathbf{u}_{it}\} \sim \mathbf{N}_N(\mathbf{0}, \Sigma_i)$.

П3. $\mathbf{z}_{it} \in \mathbf{Z} \subset \Re^M$, где область $\mathbf{Z} = \{\mathbf{z} \in \Re^M : \mathbf{z}'\mathbf{z} \le c,\ 0 < c < \infty\}$ – гипершар радиуса $\sqrt{c}$, причем векторы $\mathbf{z}_{it}, \mathbf{z}_{it-1}$ линейно независимы от $\forall i,t$.

П4. Модель векторной авторегрессии (4) удовлетворяет условию стационарности, т. е. корни алгебраического уравнения $\left|I_N \lambda - A_i\right| = 0$ лежат внутри единичного круга: $\left|\lambda_j\right| < 1,\ j = 1,...,N$, где I_N – $(N \times N)$-единичная матрица.

П5. Имеет место структурная параметрическая неоднородность модели, т. е. матрицы коэффициентов регрессии, определяющие параметрическую структуру зависимостей, удовлетворяют условию $(B_0 - B_1)\mathbf{z} \neq \mathbf{0}\ \forall \mathbf{z} \in \mathbf{Z}$.

П6. Априорные вероятности классов не зависят от времени: $\pi_{it} = \pi_i\ (t = 1,...,T,\ i \in S)$ и являются заданными.

Истинные значения параметров $\{B_i\}$, $\{A_i\}$, $\{\Sigma_i\}$ модели (2)–(4) неизвестны. Имеется классифицированная обучающая выборка значений признаков $X = X_0 \bigcup X_1 (X_i = \{\mathbf{x}_{i1}, ..., \mathbf{x}_{iT_i}\}, i \in S)$ объема $T = T_0 + T_1$ из классов $\Omega_0 \cup \Omega_1$, соответствующая последовательности значений факторов $Z = Z_0 \bigcup Z_1 (Z_i = \{\mathbf{z}_{i1}, ..., \mathbf{z}_{iT_i}\}, i \in S)$, где значение вектора признаков $\mathbf{x}_{it} \in \Re^N$ определяется на основании (3) для заданного значения вектора факторов $\mathbf{z}_{it} \in \mathbf{Z}$ $(t = 1, ..., T)$.

Имеют место следующие задачи статистического анализа многомерных неоднородных наблюдений, порождаемых моделью (2)–(4):

– построение состоятельных статистических оценок параметров модели $\{B_i\}$, $\{A_i\}$, $\{\Sigma_i\}$;

– построение состоятельного в смысле вероятности ошибки подстановочного байесовского решающего правила классификации новых наблюдений $(\mathbf{x}_\tau, \mathbf{z}_\tau)$ $(\tau = T + 1, ..., T + n,\ n \ge 1)$.

2. Статистическое оценивание параметров модели

Как отмечалось ранее, наличие классифицированной обучающей выборки позволяет решать задачи оценивания параметров модели и классификации новых наблюдений последовательно. Рассмотрим задачу оценивания параметров для произвольного класса наблюдений, порождаемых моделью (3), (4). Для простоты изложения в этом разделе будем опускать индекс *i*, соответствующий номеру класса. При этом модель (3), (4) принимает вид

$$\mathbf{x}_t = B\mathbf{z}_t + \mathbf{v}_t; \tag{5}$$

$$\mathbf{v}_t = A\mathbf{v}_{t-1} + \mathbf{u}_t \tag{6}$$

и удовлетворяет предположениям П1–П4. Степень автокорреляции ошибок наблюдения $\{\mathbf{v}_t\}$ и соответственно зависимости регрессионных наблюдений $\{\mathbf{x}_t\}$ будем характеризовать показателем $\delta = \|A\| / N$, где $\|A\| = \sqrt{\sum_{kl}^{N} a_{kl}^2}$ – норма матрицы $A = (a_{ke})$.

Получим представления для статистических оценок параметров B, A, Σ модели (5), (6) по реализации временного ряда значений зависимых переменных $\{\mathbf{x}_t\}$, соответствующей заданной реализации временного ряда значений факторов $\{\mathbf{z}_t\}$ $(t = 1, ..., T)$.

Заметим, что при N=1 модель (5), (6) представляет собой модель множественной линейной регрессии с автокоррелированными ошибками, для оценивания параметров которой традиционно используется обобщенный метод наименьших квадратов (МНК) [1, 9]. Для практической реализации данного метода могут использоваться итерационные алгоритмы Кохрейна – Окартта, Хилдрета – Лу, Дарбина [9]. Ни один из этих подходов не может непосредственно применяться в рассматриваемом многомерном случае, когда N>1.

Получим вначале аналитические представления для обобщенных МНК-оценок неизвестных параметров модели (5), (6), а затем опишем алгоритм вычисления данных оценок. В соответствии с общей методологией построения обобщенных МНК-оценок осуществим преобразование переменных в модели (5), (6), которое позволит перейти к традиционной модели с некоррелированными случайными ошибками наблюдения. Обычные МНК-оценки параметров для преобразованной модели являются обобщенными МНК-оценками для исходной модели.

Лемма. Пусть для модели VAR(1) вида (6) выполняется условие стационарности П4, $\mathbf{v}_0 \in \Re^N$ – заданное начальное значение, а случайный процесс $\{\mathbf{u}_t\}$ является бесконечным в обе стороны $(t = ... - 2, -1, 0, 1, 2...)$ и удовлетворяет предположению П1, тогда модель (5), (6) допус-

кает эквивалентное представление в виде модели VAR(1) с экзогенными переменными и некоррелированными ошибками:

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{z}_t - AB\mathbf{z}_{t-1} + \mathbf{u}_t = A\mathbf{x}_{t-1} + C\mathbf{Z}_t + \mathbf{u}_t, \tag{7}$$

где $C = \left(C_0 \middle| C_1\right)$ – блочная $\left(N \times 2M\right)$-матрица $\left(C_0 = B, C_1 = -AB\right)$; $\mathbf{Z}_t' = \left(\mathbf{z}_t' \ \mathbf{z}_{t-1}'\right) \in \mathbf{Z}^2$ – составной вектор факторов.

Доказательство. Пусть L – оператор сдвига ($L\mathbf{x}_t = \mathbf{x}_{t-1}$), тогда в условиях леммы стационарный процесс VAR(1) допускает представление

$$\mathbf{v}_t = \sum_{l=0}^{\infty} A^l \mathbf{u}_{t-l} = \sum_{l=0}^{\infty} \left(AL\right)^l \mathbf{u}_t. \tag{8}$$

С учетом стационарности процесса $\left\{\mathbf{v}_t\right\}$ и свойства оператора сдвига из (8) следует

$$\mathbf{v}_t = \left(I_N - AL\right)^{-1} \mathbf{u}_t. \tag{9}$$

Учитывая представление (9) в (5), получаем $\left(I_N - AL\right)\mathbf{x}_t = \left(I_N - AL\right)B\mathbf{z}_t + \mathbf{u}_t$, что на основании свойства оператора сдвига влечет (7). ■

Следствие. *Если осуществить преобразование переменных вида* $\mathbf{y}_t = \mathbf{x}_t - A\mathbf{v}_{t-1}$*, то в модель (5), (6) допускает представление в виде модели многомерной линейной регрессии с некоррелированными ошибками:*

$$\mathbf{y}_t = B\mathbf{z}_t + \mathbf{u}_t. \tag{10}$$

Получим аналитические представления для МП-оценок параметров A, B, Σ модели (7). Введем обозначения для матриц:

$$G = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{t-1}\mathbf{x}_{t-1}', \ g = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_{t-1}'; \tag{11}$$

$$H = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{t-1}\mathbf{Z}', \ h = \frac{1}{T}\sum_{t=1}^{T} \mathbf{Z}_t\mathbf{x}'; \tag{12}$$

$$Q = \frac{1}{T}\sum_{t-1}^{T} \mathbf{Z}_t\mathbf{Z}_t', \tag{13}$$

где G, g – $\left(N \times M\right)$-матрицы; H', h – $\left(2M \times N\right)$-матрицы; Q – $\left(2M \times 2M\right)$-матрица.

Теорема 1. *Пусть временной ряд значений эндогенных переменных* $\left\{\mathbf{x}_t\right\}(t = 1,...,T)$*, соответствующий временному ряду значений факторов* $\left\{\mathbf{z_t}\right\}$*, удовлетворяет модели (7), для которой выполняются предположения П1–П4,* $\mathbf{z}_0 = \mathbf{0}$, $\mathbf{x}_0$ *– заданные начальные значения. Если матрицы* $Q, G, Q - HG^{-1}H', G - HQ^{-1}H'$ *не вырождены, то оценки* $\widehat{A}, \widehat{C}, \widehat{\Sigma}$ *параметров модели (7) по методу максимального правдоподобия (МП-оценки) определены единственным образом и допускают представления*

$$\widehat{C}' = \left(\widehat{C}_0' \middle| \widehat{C}_1'\right) = \left(Q - H'G^{-1}H\right)^{-1}\left(h - H'G^{-1}g\right); \tag{14}$$

$$\widehat{A}' = \left(G - HQ^{-1}H'\right)^{-1}\left(g - HQ^{-1}h\right); \tag{15}$$

$$\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T}\left(\mathbf{x}_t - \hat{A}\mathbf{x}_{t-1} - \hat{C}\mathbf{Z}_t\right)\left(\mathbf{x}_t - \hat{A}\mathbf{x}_{t-1} - \hat{C}\mathbf{Z}_t\right)'. \tag{16}$$

Доказательство. Как и в случае с моделью авторегрессии, включающей детерминированные экзогенные переменные [9], можно показать, что совместная плотность распределения случайных векторов $\{\mathbf{x}_t\}$, которую можно рассматривать как функцию правдоподобия для искомых параметров A, C и Σ, имеет вид

$$L_N\left(A, C, \Sigma\right) = \left(2\pi\right)^{-\frac{NT}{2}} \left|\Sigma\right|^{-\frac{T}{2}} \exp\left\{-\frac{1}{2} tr\left(\Sigma^{-1} S_T\left(A, C\right)\right)\right\}, \tag{17}$$

где

$$S_T\left(A, C\right) = \sum_{t=1}^{T}\left(\mathbf{x}_t - A\mathbf{x}_{t-1} - C\mathbf{Z}_t\right)\left(\mathbf{x}_t - A\mathbf{x}_{t-1} - C\mathbf{Z}_t\right)' \tag{18}$$

так называемая сумма квадратов отклонений модельных значений временного ряда от наблюдаемых значений $\{\mathbf{x}_t\}$.

Задача нахождения МП-оценок $\hat{A}, \hat{C}$ эквивалентна задаче минимизации суммы квадратов отклонений $S_T\left(A, C\right)$ вида (18) по матрицам A, C. Система уравнений для задачи минимизации $S_T\left(A, C\right)$ может быть записана в виде

$$\sum_{t=1}^{T}\begin{pmatrix}\mathbf{x}_{t-1}\\ \mathbf{Z}_t\end{pmatrix}\begin{pmatrix}\mathbf{x}'_{t-1} & \mathbf{Z}'_t\end{pmatrix}\begin{pmatrix}A'\\ C'\end{pmatrix} = \sum_{t=1}^{T}\begin{pmatrix}\mathbf{x}_{t-1}\\ \mathbf{Z}_t\end{pmatrix}\mathbf{x}'_t. \tag{19}$$

Из системы уравнений (19) с учетом обозначений (11)–(3) следуют представления (14), (5).

Оценка (16) для Σ получается в результате максимизации оценки логарифмической функции правдоподобия $l_N\left(\hat{A}, \hat{C}, \Sigma\right) = \ln\left(L_N\left(\hat{A}, \hat{C}, \Sigma\right)\right)$ на множестве симметричных положительно определенных матриц. ■

Следствие 1. *МНК-оценки параметров A, C совпадают с МП-оценками данных параметров и определяются соотношениями (13), (14).*

Следствие 2. *В условиях теоремы 1 МНК-оценка искомой матрицы B модели (7) выражается через элементы матрицы $\hat{C}$, удовлетворяет соотношениям $\hat{C}_0 = \hat{B}, \hat{C}_1 = -\hat{A}\hat{B}$ и является обобщенной МНК-оценкой матрицы B для модели (5), (6).*

Исследуем состоятельность оценок $\hat{A}$, $\hat{\Sigma}$ и $\hat{C} = \left(\hat{C}_0 \middle| \hat{C}_1\right)\left(\hat{C}_0 = \hat{B}, \hat{C}_1 = -\hat{A}\hat{B}\right)$ в смысле сходимости по вероятности при $T \to \infty$ к истинным значениям.

Для фиксированной последовательности $\mathbf{Z}'_t = \left(\mathbf{z}'_t\ \mathbf{z}'_{t-1}\right) \in \mathbf{Z}^2\ \left(t = 1, ..., T\right)$ обозначим $Q = Q\left(T\right)$, где

$$Q\left(T\right) = \begin{pmatrix} Q_{11}\left(T\right) & Q_{12}\left(T\right)\\ Q_{21}\left(T\right) & Q_{22}\left(T\right)\end{pmatrix}, \tag{20}$$

а $Q_{ij}\left(T\right)\left(i, j = 1, 2\right)$ согласно (12) имеют вид

$$Q_{11}\left(T\right) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_t\mathbf{z}'_t\ ,\ Q_{22}\left(T\right) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_{t-1}\mathbf{z}'_{t-1},\ Q_{21}\left(T\right) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_{t-1}\mathbf{z}'_t\ ,\ Q_{12}\left(T\right) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_t\mathbf{z}'_{t-1}.$$

Обозначим также

$$Q^0 = \lim_{T\to\infty} Q(T),\ Q_{ij}^0 = \lim_{T\to\infty} Q_{ij}(T). \tag{21}$$

Теорема 2. *Если для модели (7) выполняются предположения П1–П4, временной ряд* $\{\mathbf{x}_t\}(t=1,...,T)$ *соответствует фиксированным значениям факторов* $\{\mathbf{z}_t\}$, *а матрицы* Σ *и* Q_{11}^0 *не вырождены, то МП-оценки* $\hat{A}, \hat{C}, \hat{\Sigma}$ *вида (14)–(16) при* $T\to\infty$ *сходятся по вероятности к истинным значениям параметров* A, C, Σ.

Доказательство. В условиях теоремы модель (7) может рассматриваться как частный случай модели

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + D\mathbf{w}_t + \mathbf{u}_t,\ t=1,...,T, \tag{22}$$

при некоторых ограничениях, касающихся ее структурного компонента $D\mathbf{w}_t$. При этом последовательность $\{\mathbf{w}_t\}$ в модели (22) считается фиксированной, а относительно модели в целом делаются предположения П1, П4.

В работе [9] доказана состоятельность по вероятности МП-оценок $\hat{A}, \hat{C}, \hat{\Sigma}$ при следующих условиях:

А. Ковариационная матрица $F = \mathrm{Cov}(\mathbf{x}_t, \mathbf{x}_t)$ является положительно определенной.

Б. $\mathbf{w}_t'\mathbf{w}_t < c\,(0<c<\infty)\,\forall t = 1,...,T$.

В. Матрица $Q^* = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbf{w}_t\mathbf{w}_t'$ является невырожденной.

Покажем, что условия А–В выполняются в условиях теоремы 2 и для модели (7). Поскольку структурный компонент $C\mathbf{Z}_t$ модели (7) является фиксированным, то с учетом представления (3) для ковариационной матрицы $F = \mathrm{Cov}(\mathbf{x}_t, \mathbf{x}_t)$ случайного вектора $\mathbf{x}_t \in \Re^N$ имеем

$$F = \sum_{l=0}^{\infty} A^l \Sigma (A^l)'.$$

В работе [9] показано, что ковариационная матрица F удовлетворяет уравнению $F = \Sigma + AFA'$ и, следовательно, является положительно-определенной как сумма положительно-определенной матрицы Σ и неотрицательно определенной матрицы AFA'. Таким образом, выполняется условие А.

Справедливость условия Б следует из предположения П3. Действительно:

$$\mathbf{Z}_t'\,\mathbf{Z}_t = \left(\mathbf{z}_t'\ \mathbf{z}_{t-1}'\right)\begin{pmatrix}\mathbf{z}_t\\ \mathbf{z}_{t-1}\end{pmatrix} = \mathbf{z}_t'\,\mathbf{z}_t + \mathbf{z}_{t-1}'z_{t-1} < c\,(0<c<\infty)\,\forall t = 1,...,T.$$

Из невырожденности матрицы Q_{11}^0 следует невырожденность матрицы Q^0. Действительно, в силу (20), (21)

$$\left|Q^0\right| = \left|Q_{11}^0\right|\left|Q_{22}^0 - Q_{12}^0 Q_{11}^{0\,-1} Q_{21}^0\right|. \tag{23}$$

Очевидно, $\left|Q^0\right| \ge 0$, причем равенство может иметь место, только если $\left|Q_{11}^0\right| = 0$ либо $\left|Q_{22}^0 - Q_{12}^0 Q_{11}^{0\,-1} Q_{21}^0\right| = 0$. По условию теоремы имеем $\left|Q_{11}^0\right| \ne 0$, второй определитель в (23) отличен от нуля, поскольку векторы $\mathbf{z}_t, \mathbf{z}_{t-1}$ линейно независимы по определению. Таким образом, выполняется условие В. ■

3. Алгоритм вычисления оценок параметров модели

Формулы (14)–(16) неудобны в вычислительном отношении и «избыточно сложны» для оценивания параметров модели (5), (6): требуется оценивать $2MN + N^2 + N^2$ параметров (элементов матриц C, A и Σ), чтобы найти $MN + N^2 + N^2$ искомых параметров (элементов матриц B, A и Σ). При фиксированном объеме данных это может отразиться и на точности оценок. Поэтому приведем описание алгоритма вычисления обобщенных МНК-оценок параметров модели (5), (6), которые могут быть реализованы как автономно, так и с использованием статистических пакетов прикладных программ, в которых имеются возможности оценивания векторных авторегрессионных моделей с экзогенными переменными при выполнении традиционных предположений о некоррелированности случайных ошибок наблюдения. Предлагаемый алгоритм следует схеме итерационного алгоритма Кохрейна – Окартта [10] в случае $N = 1$, однако использует для исходных переменных в многомерной модели (5), (6) другой вариант преобразований, приводящих данную модель к виду (10).

Описание алгоритма. Для обозначения номера итерации будем использовать верхний индекс k в скобках при соответствующих параметрах и переменных. Опишем действия, которые выполняются на начальной (k=0) и последующих (k=1, 2,…) итерациях алгоритма.

Начальная итерация (k=0)

Шаг 01. Полагаем, что $A^{(0)} = O_N$ (т. е. считаем случайные ошибки $\{\mathbf{v}_t\}(t = 1,...,T)$ некоррелированными); $B^{(0)}$ – МНК-оценка матрицы B в модели

$$\mathbf{x}_t = B\mathbf{z}_t + \mathbf{v}_t,\ t = 1,...,T. \tag{24}$$

Шаг 02. Вычисляем остатки для модели (24) по формуле

$$\mathbf{v}_t^{(0)} = \mathbf{x}_t - B^{(0)}\mathbf{z}_t,\ t = 1,...,T.$$

Итерация k (k=1, 2, …)

Шаг k1. Полагаем, что $A^{(k)}$ – МНК-оценка матрицы A в модели

$$\mathbf{v}_t^{(k-1)} = A\mathbf{v}_{t-1}^{(k-1)} + \mathbf{u}_t.$$

Шаг k2. Проверяем условие остановки алгоритма. Если элементы матриц $A^{(l)} = (a_{ij}^{(l)})$ $(l = k-1, k)$ удовлетворяют условию

$$\left\| \left\| A^{(k)} \right\| - \left\| A^{(k-1)} \right\| \right\| \le \varepsilon,\ 0 < \varepsilon \ll 1,$$

то полагаем, что $\widehat{A} = A^{(k-1)}$, $\widehat{B} = B^{(k-1)}$ и работа алгоритма прекращается. В противном случае переходим к шагу k3.

Шаг k3. Осуществляем преобразование переменных по формуле

$$\mathbf{y}_t^{(k)} = \mathbf{x}_t - A^{(k)}\mathbf{v}_{t-1}^{(k-1)},\ t = 1,...,T. \tag{25}$$

Полагаем, что $B^{(k)}$ – МНК-оценка матрицы B в модели

$$\mathbf{y}_t^{(k)} = B\mathbf{z}_t + \mathbf{v}_t,\ t = 1,...,T. \tag{26}$$

Шаг k4. Вычисляем остатки для модели (26) по формуле

$$\mathbf{v}_t^{(k)} = \mathbf{x}_t - B^{(k)}\mathbf{z}_t,\ t = 1,...,T\ ,$$

полагаем, что k:= k+1, и переходим к шагу k1 на новой итерации.

С учетом специфики задачи статистического оценивания параметров в качестве характеристики сложности алгоритма целесообразно использовать общее число оцениваемых параметров модели по выборке фиксированного объема. По этой характеристике обычный МНК (число оцениваемых параметров равно общему числу элементов матриц B и Σ, т. е. $NM + N^2$) превосходит предлагаемый алгоритм (общее число элементов матриц B, A и Σ равно $NM + N^2 + N^2$) и тем более алгоритм вычисления параметров по аналитическим формулам (14)–(16) ($2MN + N^2 + N^2$ – общее число элементов матриц C, A и Σ). Однако обычный МНК не учитывает автокорреляцию случайных ошибок наблюдений (матрица A не оценивается), поэтому следует ожидать выигрыш в точности оценивания параметров при использовании специально разработанного итерационного алгоритма. Сложность рассматриваемой задачи как задачи статистического оценивания параметров для обычного МНК можно характеризовать ранее введенной величиной $\delta = \|A\| / N \geq 0$. При $\delta = 0$ выполняется традиционное для МНК предположение о некоррелированности ошибок наблюдения. Чем больше значение δ, тем сильнее степень нарушения этого предположения.

Результаты экспериментального исследования предлагаемого итерационного алгоритма для модели многомерной линейной регрессии (5), (6), а также сравнительный анализ точности оценок на основе этого алгоритма и алгоритмов Кохрейна – Окартта и Дарбина в условиях применимости последних (т. е. в случае N=1) представлены в разд. 5.

4. Решающее правило классификации неоднородных зависимых регрессионных наблюдений

Рассмотрим исходную модель наблюдений (3), (4). Для классификации новых наблюдений $(\mathbf{x}_\tau, \mathbf{z}_\tau)$ $(\tau = T+1,...,T+n,\ n \geq 1)$ будем использовать подстановочное байесовское решающее правило (ПБРП), получающееся из оптимального в смысле минимума вероятности ошибки байесовского решающего правила (БРП) в форме Зигерта–Котельникова [7] подстановкой в него состоятельных оценок неизвестных параметров.

Пусть $\varphi_N\left(u \middle| \mathbf{0}, \Sigma\right)$ – плотность N-мерного нормального распределения с нулевым математическим ожиданием и невырожденной ковариационной матрицей Σ; $U(z) = \begin{cases} 0, z < 0, \\ 1, z \geq 0 \end{cases}$ – единичная функция Хэвисайда.

Теорема 3. *Если для модели наблюдений (3), (4) выполняются предположения П1–П6 и* $\Sigma_0 \neq \Sigma_1$, *то БРП классификации случайного вектора* $\mathbf{x}_\tau$ *для заданных значений* $\mathbf{x}_{\tau-1}, \mathbf{z}_\tau$ $(\tau = T+1,...,T+n,\ n \geq 1)$ *имеет вид*

$$d_t = d\left(\mathbf{x}_\tau, \mathbf{z}_\tau\right) = U\left[\frac{\varphi_N\left(\mathbf{x}_\tau \middle| B_1\mathbf{z}_\tau + \boldsymbol{\beta}_{1\tau-1}, \Sigma_1\right)}{\varphi_N\left(\mathbf{x}_\tau \middle| B_0\mathbf{z}_\tau + \boldsymbol{\beta}_{0\tau-1}, \Sigma_0\right)} - \frac{\pi_0}{\pi_1}\right], \tag{27}$$

где $\boldsymbol{\beta}_{it-1} = A_i\left(\mathbf{x}_{\tau-1} - B_i\mathbf{z}_{\tau-1}\right),\ i \in S = \{0,1\}$.

Доказательство. Для фиксированных в момент времени τ значений $\mathbf{x}_{\tau-1}, \mathbf{z}_\tau$ $(\tau = T+1,...,T+n,\ n \geq 1)$ случайный вектор $\mathbf{x}_\tau$ имеет с вероятностью π_i плотность распределения $\varphi_N\left(\mathbf{x}_\tau \middle| B_i\mathbf{z}_\tau + \boldsymbol{\beta}_{i\tau-1}, \Sigma_i\right)$ и, следовательно, согласно [6] БРП классификации определяется соотношением (27). ■

Следствие 1. *В условиях теоремы 2 состоятельное ПБРП классификации получается подстановкой в (24) обобщенных МНК-оценок* $\widehat{A}, \widehat{C}, \widehat{\Sigma}$, *для которых справедливо представление (12)–(14).*

Доказательство. Состоятельность ПБРП в смысле сходимости риска ПБРП к риску БРП при $T_i \to \infty (i \in S)$ следует из состоятельности оценок параметров модели [3]. ■

Следствие 2. *Если модели наблюдений для различных классов различаются лишь матрицами коэффициентов регрессии, т. е.* $B_0 \neq B_1, A_0 = A_1 = A, \Sigma_0 = \Sigma_1 = \Sigma$, *то ПБРП для модели (3), (4) имеет вид*

$$d_\tau = d(\mathbf{x}_\tau, \mathbf{z}_\tau) = U\left[\lambda_\tau(\mathbf{x}_\tau, \mathbf{z}_\tau)\right], \quad \lambda_\tau = \lambda_\tau^*(\mathbf{x}_\tau, \mathbf{z}_\tau) - \alpha_{1\tau} - \alpha_{2\tau}, \tag{28}$$

где $\alpha_{1\tau} = \frac{1}{2}\left(\hat{\boldsymbol{\beta}}_{1\tau-1} - \hat{\boldsymbol{\beta}}_{0\tau-1}\right)' \hat{\Sigma}^{-1}\left(\hat{\boldsymbol{\beta}}_{1\tau-1} - \hat{\boldsymbol{\beta}}_{0\tau-1}\right)$;

$\alpha_{2\tau} = \hat{\boldsymbol{\beta}}'_{1\tau-1}\hat{\Sigma}^{-1}\hat{B}_1\mathbf{z}_\tau - \hat{\boldsymbol{\beta}}'_{0\tau}\hat{\Sigma}^{-1}\hat{B}_0\mathbf{z}_\tau$;

$\hat{\boldsymbol{\beta}}_{i\tau-1} = \widehat{A}\left(\mathbf{x}_{\tau-1} - \hat{B}_i\mathbf{z}_{\tau-1}\right),\ i \in S.$

$\lambda_\tau^*(\mathbf{x}_\tau, \mathbf{z}_\tau) = \mathbf{z}'\left(\hat{B}_1 - \hat{B}_0\right)' \hat{\Sigma}^{-1}\mathbf{x}_\tau - \frac{1}{2}\mathbf{z}'\left(\hat{B}_1 + \hat{B}_0\right)' \hat{\Sigma}^{-1}(B_1 - B_0) - \ln\frac{\pi_0}{\pi_1}$ – линейная дискриминантная функция ПБРП в случае независимых ошибок наблюдения $\{\mathbf{v}_t\}$ в модели (3).

Таким образом, если $A = 0$, т. е. регрессионные наблюдения $\{\mathbf{x}_t\}$ являются некоррелированными, то ПБРП (28) совпадает с ранее построенным в [8] решающим правилом.

Заметим, что вероятность ошибки БРП (28) в случае равновероятных классов $(\pi_0 = \pi_1 = 0{,}5)$ определяется по формуле $P = \Phi\left(-\frac{\Delta_\tau}{2}\right)$, где $\Delta_\tau = \left(\boldsymbol{\psi}_{1\tau} - \boldsymbol{\psi}_{0\tau}\right)' \Sigma^{-1}\left(\boldsymbol{\psi}_{1\tau} - \boldsymbol{\psi}_{0\tau}\right)$ – межклассовое расстояние Махаланобиса, $\boldsymbol{\psi}_{i\tau} = B_i\mathbf{z}_\tau + \boldsymbol{\beta}_{i\tau-1}$, $i \in S$; $\Phi(\cdot)$ – функция распределения стандартного нормального закона.

5. Численные эксперименты

Проведены три серии экспериментов для различных тестовых моделей, целью которых является:

1) сравнительный анализ точности оценок параметров модели (5), (6) при различных значениях δ и T на основе предлагаемого алгоритма, обычного МНК и альтернативных алгоритмов вычисления обобщенных МНК-оценок (алгоритмов Кохрейна – Окартта и Дарбина [9]) для модели множественной линейной регрессии ($N = 1$) с ошибками наблюдений, описываемыми моделью авторегрессии первого порядка AR(1), т. е. в условиях, когда возможно применение альтернативных алгоритмов;

2) иллюстрация возможности использования предлагаемого алгоритма оценивания параметров в случае, когда модель ошибок наблюдения (6) имеет более высокий порядок авторегрессии $p \geq 1$;

3) сравнительный анализ точности оценок параметров модели (5), (6), получаемых с помощью обычного МНК и предлагаемого алгоритма при различных значениях δ, соответствующих различным степеням зависимости регрессионных наблюдений.

Тестовая модель 1 определяется следующими соотношениями: $N = M = 1$, $T \in \{12,\ 102\}$, $x_{t1} = b_1 z_t + v_{t1}$, $v_{t1} = a_{11}v_{t-1,1} + u_{t1}$, $u_{t1} \sim \mathbf{N}(0, \sigma_{11}^2)$, значения z_t равномерно распределены на интервале $[-1,1]$, $a_{11} \in \{0{,}5,\ 0{,}9,\ 0{,}95\}$, $\delta = a_{11}$, $b_1 = 2$, $\sigma_{11} = 0{,}625$. Оценки параметров для рассматриваемых алгоритмов (усредненные по пяти прогонам) представлены в табл. 1 и 2. В случае алгоритма Дарбина оцениваются параметры модели вида

$$x_t = a_{11}x_{t-1} + b_1 z_t - a_{11}b_1 z_{t-1} + u_t = a_{11}x_{t-1} + b_1 z_t + dz_{t-1} + u_t$$

(частного случая модели (7) при N=1), поэтому в таблицах указываются две оценки параметра b_1: оценка коэффициента регрессии при переменной z_t и оценка, вычисляемая как $\hat{d}/\hat{a}_{11}$. В столбце МНК для параметра a_{11} приводятся МНК-оценки по смоделированной реализации случайных ошибок $\{v_{t1}\}$.

Таблица 1

Модель множественной линейной регрессии с ошибками вида AR(1), T = 12

Истинное значение a_{11}	Параметры	МНК	Алгоритм Кохрейна – Окартта	Алгоритм Дарбина	Предлагаемый алгоритм
0,5	a_{11}	0,3041	0,2278	0,2555	0,2278
	b_1	2,0746	2,0619	2,1238/1,3102	2,1050
0,9	a_{11}	0,8705	0,8502	0,8379	0,8502
	b_1	2,2027	1,9895	2,1504/1,7536	2,1129
0,95	a_{11}	0,9497	0,9307	0,9258	0,9307
	b_1	2,3407	1,9781	2,1436/1,7886	2,0958

Таблица 2

Модель множественной линейной регрессии с ошибками вида AR(1), T = 102

Истинное значение a_{11}	Параметры		Кохрейна – Окартта	Алгоритм Дарбина	Предлагаемый алгоритм
0,5	a_{11}	0,3325	0,3251	0,3331	0,3251
	b_1	2,0319	2,0031	2,0281/1,7122	2,0337
0,9	a_{11}	0,8132	0,8093	0,8083	0,8093
	b_1	2,0312	1,9755	2,0321/1,8858	2,0344
0,95	a_{11}	0,9289	0,9283	0,9290	0,9283
	b_1	2,0077	1,9724	2,0287/1,9064	2,0332

Тестовая модель 2 соответствует модели множественной линейной регрессии с ошибками наблюдения, описываемыми моделью авторегрессии AR(p): $N = M = 1$, $T = 502$, $p = 2$, $x_{t1} = b_1 z_t + v_{t1}$, $v_{t1} = a_{11}v_{t-1,1} + a_{12}v_{t-2,1} + u_{t1}$, $u_{t1} \sim \mathbf{N}(0, \sigma_{11}^2)$, значения z_t равномерно распределены на интервале $[-1,1]$, $a_{11} = 0,9$, $a_{12} = -0,7$ ($\delta = 0,57$), $b_1 = 2$, $\sigma_{11} = 0,625$. Оценки параметров на основе предлагаемого алгоритма (усредненные по пяти прогонам) представлены в табл. 3.

Таблица 3

Модель множественной линейной регрессии с ошибками вида AR(2)

Параметры	МНК	После первой итерации	После второй итерации	После третьей итерации
a_{11} a_{12}	0,8683, –0,7534	0,8671, –0,7520	0,8686, –0,7530	0,8687, –0,7530
$b_1 = 2$	2,0249	2,0076	2,0073	Не оценивается

Тестовая модель 3 определяется следующими соотношениями: $N = 2$, M=1, $T = 300$, $x_{t1} = b_1 z_t + v_{t1}$, $x_{t2} = b_2 z_t + v_{t2}$, $v_{t1} = a_{11}v_{t-1,1} + a_{12}v_{t-1,2} + u_{t1}$, $v_{t2} = a_{21}v_{t-1,1} + a_{22}v_{t-1,2} + u_{t2}$, $u_{tj} \sim \mathbf{N}(0, \sigma_j^2)$, значения z_t равномерно распределены на интервале $[-1,1]$.

Рассматривались два тестовых примера, отличающихся степенью корреляции случайных ошибок наблюдения в модели (5), (6):

1) $b_1 = 2,\ b_1 = 5,\ a_{11} = 0,4,\ a_{12} = 0,\ a_{21} = -0,2,\ a_{22} = 0,5,\ \sigma_1 = 0,8,\ \sigma_2 = 0,4,\ \delta$=0,5;

2) $b_1 = 2,\ b_1 = 5,\ a_{11} = 0,9,\ a_{12} = 0,\ a_{21} = -0,7,\ a_{22} = 0,8,\ \sigma_1 = 0,8,\ \sigma_2 = 0,4,\ \delta$=0,7.

Оценки параметров моделей для рассматриваемых случаев на основе обычного МНК, не учитывающего автокорреляцию случайных ошибок, и предлагаемого алгоритма (усредненные по пяти прогонам) представлены в табл. 4.

Таблица 4

Модель многомерной линейной регрессии с ошибками вида VAR(1)

Тестовый пример	Параметры	МНК	Предлагаемый алгоритм
1	b_1, b_2	1,9948, 5,0009	1,9981, 4,9258
2	b_1, b_2	1,8717, 5,6312	1,9906, 4,9486

В табл. 5 даны значения основных тестовых статистик, используемых для оценки качества построенных многомерных статистических моделей, включая информационные статистики Акаике (AIC) и Шварца (SC), нормированную матричную сумму квадратов остатков $\left|\hat{\Sigma}\right|$ и логарифмическую функцию правдоподобия для вычисленных оценок $l_N\left(\hat{A},\hat{C},\hat{\Sigma}\right)$ [1, 10]. Меньшие значения статистик AIC, SC, $\left|\hat{\Sigma}\right|$ и большие значения статистики $l_N\left(\hat{A},\hat{C},\hat{\Sigma}\right)$ соответствуют более качественной модели. Заметим также, что для всех моделей оценки коэффициентов регрессии являются статистически значимыми на уровне 0,05, а для моделей, построенных с помощью предлагаемого алгоритма, остатки являются гауссовским белым шумом.

Таблица 5

Характеристики качества статистических моделей

Алгоритм	Тестовые статистики			
	AIC	SC	$\left\|\hat{\Sigma}\right\|$	$l_N\left(\hat{A},\hat{C},\hat{\Sigma}\right)$
Тестовый пример 1				
МНК	4,0736	4,0983	0,2001	–609,045
Предлагаемый	3,9312	3,9556	0,1736	–585,714
Тестовый пример 2				
МНК	8,1854	8,2101	12,2190	–1225,810
Предлагаемый	3,4506	3,4754	0,1073	–513,8673

Заключение

Результаты проведенных экспериментов позволяют сделать следующие выводы:

1. В случае $N = 1$ (см. табл. 1 и 2) предлагаемый алгоритм занимает промежуточное положение по точности оценивания регрессионных коэффициентов между алгоритмами Кохрейна – Окартта и Дарбина; преимущество итерационных алгоритмов над обычным МНК проявляется в условиях коротких временных рядов и тем сильнее, чем больше степень зависимости регрессионных наблюдений.

2. Предлагаемый алгоритм может использоваться для оценивания параметров рассматриваемой модели в случае, когда модель ошибок наблюдения имеет более высокий порядок авторегрессии (см. табл. 3).

3. В случае $N > 1$ при высокой степени зависимости регрессионных наблюдений предлагаемый алгоритм имеет существенный выигрыш по сравнению с обычным МНК как по точности оценивания регрессионных коэффициентов (см. табл. 4), так и по всем основным характеристикам качества статистических моделей (см. табл. 5).

4. При практической реализации описанных выше решающих правил рекомендуется использовать оценки неизвестных параметров, полученные с помощью предлагаемого итерационного алгоритма.

Список литературы

1. Харин, Ю.С. Эконометрическое моделирование / Ю.С. Харин, В.И. Малюгин, А.Ю. Харин. – Минск: БГУ, 2003. – 313 с.

2. Прикладная статистика. Классификация и снижение размерности / С.А. Айвазян [и др.] – М.: Финансы и статистика, 1989. – 607 с.

3. Песаран, М. Динамическая регрессия: теория и алгоритмы / М. Песаран, Л. Слейтер. – М.: Финансы и статистика, 1984. – 310 с.

4. Гринь, Н.В. Исследование точности методов классификации многомерных данных в задачах кредитного скоринга / Н.В. Гринь, В.И. Малюгин // Вестник ГрГУ. Сер. 2. – 2008. – № 1. – С.77–85.

5. Малюгин, В.И. Оценка устойчивости коммерческих банков на основе эконометрических моделей с дискретными зависимыми переменными / В.И. Малюгин, Е.В. Пытляк // Банковский вестник. – 2007. – № 4 (369). – С. 30–36.

6. Малюгин, В.И. Об оптимальности классификации случайных наблюдений, различающихся уравнениями регрессии / В.И. Малюгин, Ю.С. Харин // Автоматика и телемеханика. – 1986. – № 7. – С. 35–46.

7. Харин, Ю.С. Робастность в статистическом распознавании образов. – Минск: Университетское, 1992. – 232 с.

8. Жук, Е.Е. Устойчивость в кластер-анализе многомерных наблюдений / Е.Е. Жук, Ю.С. Харин. – Минск: БГУ, 1998. – 240 с.

9. Андерсон, Т. Статистический анализ временных рядов / Т. Андерсон. – М.: Мир, 1976. – 755 с.

10. Магнус, Я.Р. Эконометрика. Начальный курс / Я.Р. Магнус, П.К. Катышев, А.А. Пересецкий. – М.: Дело, 2004. – 576 с.

Поступила 21.03.08

НИИ прикладных проблем математики и информатики
Белорусского государственного университета,
Минск, пр. Независимости, 4
e-mail: Malugin@bsu.by

V.I. Malugin

DISCRIMINANT ANALYSIS OF THE MULTIVARIATE AUTOCORRELATED REGRESSION OBSERVATIONS ON CONDITIONS OF A PARAMETRIC HETEROGENEITY OF THE MODELS

The paper is devoted to the problem of discriminant analysis of the multivariate regression models with heterogeneous structure and autocorrelated innovations. A consistent decision rule of classification of multivariate heterogeneous dependent regression observations is obtained. An iteration algorithm for the estimation of the model parameters is suggested and examined on the base of simulated data.