ISSN 0002-3574 (print)
УДК 517.514

Поступила в редакцию 23.09.2016
Received 23.09.2016

**В. С. Муха**

*Белорусский государственный университет информатики и радиоэлектроники, Минск, Беларусь*

# СИММЕТРИЧНАЯ АППРОКСИМАЦИЯ ВЕКТОРНЫХ СТАТИСТИЧЕСКИХ ДАННЫХ ЛИНЕЙНЫМИ МНОГООБРАЗИЯМИ

Рассматривается задача линейной аппроксимации векторных статистических данных. Как известно, классическая линейная функция регрессии минимизирует сумму квадратов вертикальных расстояний от системы точек до аппроксимирующей плоскости. В данной статье рассматривается иной подход к аппроксимации, когда минимизируется сумма квадратов перпендикулярных расстояний от системы точек до плоскости. Такая аппроксимация названа симметричной. Получены формулы аппроксимирующих линейных многообразий в параметрической форме. Решение задачи выполнено в векторно-матричной форме. Приведены численные примеры и их графические иллюстрации в сравнении с известными результатами из литературы и классического линейного регрессионного анализа.

*Ключевые слова:* аппроксимация, векторные статистические данные, линейные многообразия, линейный регрессионный анализ, компьютерный анализ данных.

**V. S. Mukha**

*Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus*

## SYMMETRIC APPROXIMATION OF VECTOR STATISTICAL DATA BY LINEAR VARIETIES

The problem of the approximation of vector statistical data is considered. As it is known the classical linear regression function minimizes the sum of squares of the vertical distances from the system of points to the approximating plane. The problem in this case consists of choos the parameters $c_0$, $c_1$, …, $c_{m-1}$ of linear function $x_m = c_0 + c_1x_1 + ... + c_{m-1}x_{m-1}$ from the condition

$$\sum_{i=1}^{n}(x_{m,i} - c_0 - c_1x_{1,i} - ... - c_{m-1}x_{m-1,i})^2 \to \min_{c_0,c_1,...,c_{m-1}},$$

where $x_{m,i}$, $i = \overline{1,n}$, are the observations of endogenous component $\xi_m$ of random vector $\xi = (\xi_1,...,\xi_m)$ and $n$ is the size of the sample. In this article, another approach to the approximation is considered when the sum of squares of the perpendicular distances from the system of points $X_i = (x_{1,i},...,x_{m,i}) \in E^m$, $i = \overline{1,n}$, to the plane $x_m = c_0 + c_1x_1 + ... + c_{m-1}x_{m-1}$ is minimized. Such an approximation was called as symmetric approximation of vector statistical data. We solve the problem in more general form. We look for the linear variety in $E^m$ ($(m - r)$-dimensional plane) in parametric form

$$X = C_0 + t_1C_1 + ... + t_{m-r}C_{m-r} \in E^m, \ 1 \le r \le m, \ t_0 = 1,$$

where $C_0$, $C_1$, …, $C_{m-r}$ are the unknown linear independent vectors in $E^m$, $t_1$,…, $t_{m-r}$ are the scalar real parameters, $E^m$ is the $m$-dimensional Euclidean space. This variety can be presented in form of vector-vector linear dependence

$$X_r = A + B\,X_{m-r},$$

where $X_r^T = (x_{m-r+1}, x_{m-r+2},...x_m,)$, $X_{m-r}^T = (x_1, x_2,...x_{m-r})$, $B$ is the $(r \times (m-r))$-matrix. We give the solution in vector-matrix form and the numerical examples compared with the known results from literature and the classical linear regression analysis.

*Keywords:* approximation, vector statistical data, linear varieties, linear regression analysis, computer data analysis.

**Введение.** В настоящее время теоретическое обоснование и широкое применение получил линейный регрессионный анализ в виде множественной (скалярно-векторной) линейной регрессии. Задача в этом случае состоит в выборе параметров $c_0$, $c_1$, …, $c_{m-1}$ линейной функции $x_m = c_0 + c_1x_1 + ... + c_{m-1}x_{m-1}$, доставляющих минимум сумме квадратов расстояний от наблюдений $x_{m,i}$, эндогенной компоненты $\xi_m$ случайного вектора $\xi = (\xi_1,...,\xi_m)$ до значений $c_0 + c_1x_{1,i} + ... + c_{m-1}x_{m-1,i}$, $i = \overline{1,n}$, предсказанных данной функцией:

$$\sum_{i=1}^{n}(x_{m,i} - c_0 - c_1x_{1,i} - ... - c_{m-1}x_{m-1,i})^2 \to \min_{c_0,c_1,...,c_{m-1}}. \quad (1)$$

Аппроксимирующая функция $x_m = c_0 + c_1 x_1 + ... + c_{m-1} x_{m-1}$ представляет собой гиперплоскость в $E^m$, а критерий (1) минимизирует сумму квадратов расстояний от точек $X_i = (x_{1,i}, ..., x_{m,i}) \in E^m$ до этой гиперплоскости, отсчитываемых вдоль оси $x_m$, т. е. по вертикали. Полученная при этом функция является условным эмпирическим математическим ожиданием случайной величины $\xi_m$ на случайные величины $\xi_1, \ldots, \xi_{m-1}$.

Возможен также иной подход к аппроксимации векторных данных $X_i = (x_{1,i}, ..., x_{m,i}) \in E^m$, состоящий в минимизации суммы квадратов расстояний от точек $X_i \in E^m$ до искомой гиперплоскости в $E^m$, который рассматривался в работах [1, 2], однако не получил широкого освещения в статистической литературе. В данной работе дается независимое решение этой проблемы. В отличие от [1, 2], решение получено в векторно-матричной форме.

**1. Линейные многообразия в многомерном арифметическом пространстве.** Как известно [3], в $R^m$ можно определить линейные многообразия в параметрической форме

$$X = C_0 + t_1 C_1 + ... + t_{m-r} C_{m-r} \in R^m, 1 \le r \le m, t_0 = 1, \tag{2}$$

где $C_0^T = (c_{0,1}, ..., c_{0,m})$, $C_1^T = (c_{1,1}, ..., c_{1,m})$, …, $C_{m-r}^T = (c_{m-r,1}, ..., c_{m-r,m})$ – линейно независимые векторы в $R^m$, $t_1, \ldots, t_{m-r}$ – скалярные действительные параметры, $X$ – вектор-столбец в $R^m$. Многообразие (2) называется $(m-r)$-мерной гиперплоскостью в $R^m$, а векторы $C_1, \ldots, C_{m-r}$ – направляющими векторами этой гиперплоскости.

Для выяснения содержательного смысла линейного многообразия (2) запишем его в виде

$$X = C_0 + CT, \tag{3}$$

где $C = [C_1, C_2, ..., C_{m-r}]$ – $(m \times (m-r))$-матрица, столбцами которой являются векторы-столбцы $C_1, C_2, ..., C_{m-r}$, а $T^T = [t_1, t_2, ..., t_{m-r}]$ – вектор параметров $t_1, t_2, ..., t_{m-r}$. Представим матрицу $C$ и векторы $X$, $C_0$ в виде

$$C = \begin{pmatrix} \overline{C}_{m-r} \\ \overline{C}_r \end{pmatrix}, X = \begin{pmatrix} \overline{X}_{m-r} \\ \overline{X}_r \end{pmatrix}, C_0 = \begin{pmatrix} \overline{C}_{0,m-r} \\ \overline{C}_{0,r} \end{pmatrix}, \tag{4}$$

где $\overline{C}_{m-r}$ – $((m-r) \times (m-r))$-матрица, $\overline{C}_r$ – $(r \times (m-r))$-матрица, $\overline{X}_{m-r}^T = (x_1, x_2, ..., x_{m-r})$, $\overline{X}_r^T = (x_{m-r+1}, x_{m-r+2}, ..., x_m)$, $\overline{C}_{0,m-r}^T = (c_{0,1}, c_{0,2}, ..., c_{0,m-r})$, $\overline{C}_r^T = (c_{0,m-r+1}, c_{0,m-r+2}, ...c_{0,m})$. Тогда вместо уравнения (3) мы можем записать два уравнения

$$\begin{cases} \overline{X}_{m-r} = \overline{C}_{0,m-r} + \overline{C}_{m-r} T, \\ \overline{X}_r = \overline{C}_{0,r} + \overline{C}_r T. \end{cases} \tag{5}$$

В силу линейной независимости векторов $C_1, C_2, ..., C_{m-r}$ матрица $\overline{C}_{m-r}$ не вырожденная, и мы можем из первого уравнения системы уравнений (5) найти вектор параметров $T$:

$$T = (\overline{C}_{m-r})^{-1} (\overline{X}_{m-r} - \overline{C}_{0,m-r}).$$

Подставляя это решение во второе уравнение системы (5), получим

$$\overline{X}_r = \overline{C}_{0,r} + \overline{C}_r (\overline{C}_{m-r})^{-1} (\overline{X}_{m-r} - \overline{C}_{0,m-r}). \tag{6}$$

Последнее выражение показывает, что в случае $(m-r)$-мерной гиперплоскости в $R^m$ $r$ произвольных компонент вектора $X$, принадлежащего гиперплоскости, линейно выражаются через остальные $(m-r)$ компонент этого вектора. Формула (6) представляет собой эту зависимость в явной форме для последних $r$ компонент вектора $X$. В частности, при $m - r = 0$ многообразие (2) называется нульмерной гиперплоскостью в $E^m$ с уравнением $X = C_0$ и является точкой в $R^m$. При $m - r = 1$ это одномерная гиперплоскость в $R^m$, т. е. прямая с уравнением $X = C_0 + t_1 C_1$, проходящая через точку $C_0$ в направлении вектора $C_1$. При $m - r = m - 1$ это уравнение $(m-1)$-мерной гиперплоскости в $R^m$ (гиперплоскости в $R^m$ в классическом понимании). Итак, выражение (6) является эквивалентным представлением гиперплоскости (2).

Обратно, легко заметить, что выражение (6) представляет собой векторно-векторную (многомерную) линейную среднюю квадратичную регрессию, если считать, что $\overline{C}_{m-r} = D(\eta_{m-r}) = D_{m-r}$ – дисперсионная матрица некоторого случайного вектора $\eta_{m-r}$, $\overline{C}_r = \mathrm{cov}(\eta_r, \eta_{m-r}) = R_{r,m-r}$ – ковариационная матрица случайных векторов $\eta_r$ и $\eta_{m-r}$, $\overline{C}_{0,m-r} = E(\eta_{m-r})$, $\overline{C}_{0,r} = E(\eta_r)$ – математические ожидания случайных векторов $\eta_{m-r}$ и $\eta_r$ соответственно. Это значит, что многомерная линейная средняя квадратичная регрессия вида

$$\overline{X}_r = \overline{C}_{0,r} + R_{r,m-r}\,(D_{m-r})^{-1}(\overline{X}_{m-r} - \overline{C}_{0,m-r})$$

представляет собой $(m-r)$-мерное многообразие в $R^m$, имеющее параметрическое уравнение (3) с

$$C = \begin{pmatrix} D_{m-r} \\ R_{r,m-r} \end{pmatrix},\ X = \begin{pmatrix} \overline{X}_{m-r} \\ \overline{X}_r \end{pmatrix},\ C_0 = \begin{pmatrix} \overline{C}_{m-r} \\ \overline{C}_r \end{pmatrix} = E\begin{pmatrix} \eta_{m-r} \\ \eta_r \end{pmatrix}.$$

**2. Постановка и решение задачи симметричной аппроксимации векторных статистических данных.** Пусть $\xi^T = (\xi_1, ..., \xi_m) \in E^m$ – случайный вектор в $E^m$ со средним значением $A_\xi = E(\xi)$ и положительно определенной дисперсионной матрицей $D_\xi = E((\xi - A_\xi)(\xi - A_\xi)^T)$. Поставим задачу определения векторов $C_0$, $C_1$,…, $C_{m-r}$ линейного многообразия (2), обеспечивающих минимальную вариацию квадрата расстояния $\rho^2$ от точки $\xi$ до многообразия:

$$E\left(\rho^2(\xi, C_0 + t_1 C_1 + ... + t_{m-r} C_{m-r})\right) \to \min_{C_0, C_1, ..., C_{m-r}}. \tag{7}$$

Перейдем к решению задачи. Квадрат расстояния $\rho^2$ от точки $\xi$ до линейного многообразия (2) вычисляется по формуле [3]:

$$\rho^2 = \frac{\det(\tilde{C}^T \tilde{C})}{\det(C^T C)}, \tag{8}$$

где $C = [C_1, C_2, ..., C_{m-r}]$ – $(m \times (m-r))$-матрица, столбцами которой являются векторы-столбцы $C_1, C_2, ..., C_{m-r}$, $\tilde{C} = [C_1, C_2, ..., C_{m-r}, \xi - C_0] = [C, \xi - C_0]$ – $(m \times (m-r+1))$-матрица. Рассматривая эти матрицы как блочные, получим

$$Z = C^T C = (z_{i,j}) = (C_i^T C_j),\ i, j = \overline{1, m-r},$$

$$\tilde{Z} = \tilde{C}^T \tilde{C} = (\tilde{z}_{i,j}) = \begin{pmatrix} C_i^T C_j, & i, j = \overline{1, m-r}, \\ (\xi - C_0)^T C_j, & i = m-r+1,\ j = \overline{1, m-r}, \\ C_i^T (\xi - C_0), & i = \overline{1, m-r},\ j = m-r+1, \\ (\xi - C_0)^T (\xi - C_0), & i, j = m-r+1, \end{pmatrix}.$$

Будем искать единичные ортогональные векторы $C_1, C_2, ..., C_{m-r}$. В этом случае искомыми компонентами векторов $C_1, C_2, ..., C_{m-r}$ будут направляющие косинусы, будут выполняться равенства $C_i^T C_j = 0$ при $i \neq j$, $C_i^T C_i = 1$, матрица $Z$ будет единичной, а матрица $\tilde{Z}$ в развернутой форме примет следующий вид:

$$\tilde{Z} = \begin{pmatrix} 1, & 0, & \dots, & 0, & C_1^T(\xi - C_0) \\ 0, & 1, & \dots, & 0, & C_2^T(\xi - C_0) \\ 0, & 0, & \dots, & 0, & C_3^T(\xi - C_0) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0, & 0, & \dots, & 1, & C_{m-r}^T(\xi - C_0) \\ (\xi - C_0)^T C_1, & (\xi - C_0)^T C_2, & \dots, & (\xi - C_0)^T C_{m-r}, & (\xi - C_0)^T(\xi - C_0) \end{pmatrix}.$$

Определитель матрицы $Z$ равен единице: $\det(C^T C) = 1$. Определитель матрицы $\tilde{Z}$ получим, раскрывая его по элементам последней строки матрицы $\tilde{Z}$:

$$\rho^2 = \det(\tilde{C}^T \tilde{C}) = (\xi - C_0)^T (\xi - C_0) - \sum_{i=1}^{m-r} C_i^T (\xi - C_0)(\xi - C_0)^T C_i .$$

Нас интересует математическое ожидание случайной величины $\rho^2$. Легко понять, что

$$E(\rho^2) = \operatorname{tr} \tilde{D} - \sum_{i=1}^{m-r} C_i^T \tilde{D} C_i , \tag{9}$$

где $\tilde{D} = E((\xi - C_0)(\xi - C_0)^T)$. Преобразуем матрицу $\tilde{D}$ следующим образом:

$$\tilde{D} = E((\xi - C_0)(\xi - C_0)^T) = E(((\xi - E(\xi)) + (E(\xi) - C_0))((\xi - E(\xi)) + (E(\xi) - C_0))^T) = D_\xi + M,$$

где

$$D_\xi = E((\xi - E(\xi))(\xi - E(\xi))^T),$$
$$M = (E(\xi) - C_0)(E(\xi) - C_0)^T,$$

$D_\xi$ – дисперсионная матрица случайного вектора $\xi$. С учетом данного представления матрицы $\tilde{D}$ выражение (9) принимает вид

$$E(\rho^2) = \left( \operatorname{tr} D_\xi - \sum_{i=1}^{m-r} C_i^T D_\xi C_i \right) - \left( \operatorname{tr} M - \sum_{i=1}^{m-r} C_i^T M C_i \right). \tag{10}$$

Необходимыми условиями минимума функции $E(\rho^2)$ являются следующие уравнения:

$$\frac{\partial}{\partial C_0} E(\rho^2) = 0, \frac{\partial}{\partial C_i} E(\rho^2) = 0, i = \overline{1, m-r},$$

где $\frac{\partial}{\partial C_0} E(\rho^2), \frac{\partial}{\partial C_i} E(\rho^2), i = \overline{1, m-r}$, – производные скалярной функции $E(\rho^2)$ по векторным переменным $C_0, C_1, C_2, ..., C_{m-r}$ (скалярно-векторные производные) [4, 5]. Дифференцирование $E(\rho^2)$ (10) по $C_0$ дает уравнение

$$\frac{\partial}{\partial C_0} E(\rho^2) = \frac{\partial}{\partial C_0} \left( \operatorname{tr} M - \sum_{i=1}^{m-r} C_i^T M C_i \right) = -2(E(\xi) - C_0) + \sum_{i=1}^{m-r} C_i C_i^T (E(\xi) - C_0) = 0,$$

из которого получаем

$$C_0 = E(\xi) = A_\xi. \tag{11}$$

При таком значении $C_0$ получим $M = 0$, $\operatorname{tr} M - \sum_{i=1}^{m-r} C_i^T M C_i = 0$, так что вместо дифференцирования по $C_1, C_2, ..., C_{m-r}$ функции $E(\rho^2)$ (10) мы можем выполнять дифференцирование функции

$$\phi(C_1, C_2, ..., C_{m-r}) = \left( \operatorname{tr} D_\xi - \sum_{i=1}^{m-r} C_i^T D_\xi C_i \right). \tag{12}$$

В результате дифференцирования получим уравнения

$$\frac{d}{dC_k} \phi(C_1, C_2, ..., C_{m-r}) = -2 D_\xi C_k = 0, k = \overline{1, m-r}, 1 \le r \le m. \tag{13}$$

Представим матрицу $D_\xi$ в виде $D_\xi = D_1 + \lambda I$, где $I$ – единичная матрица, $\lambda$ – действительное число. Тогда уравнения (13) примут вид

$$D_1 C_k = \lambda C_k, k = \overline{1, m-r}, 1 \le r \le m.$$

Это значит, что искомые векторы $C_k$ являются собственными векторами матрицы $D_1 = D_\xi - \lambda I$. Однако, как известно [6], матрицы $D_\xi$ и $D_\xi - \lambda I$ имеют одни и те же собственные векторы. В таком случае в качестве искомых векторов $C_k$ можно взять собственные векторы дисперсионной матрицы $D_\xi$ случайного вектора $\xi$. Матрица $D_\xi$ имеет $m$ собственных векторов. При отыскании $(m-r)$-мерного многообразия нам необходимо из $m$ собственных векторов матрицы $D_\xi$ отобрать $(m-r)$ векторов, доставляющих минимальное значение функции $\phi(C_1, C_2, ..., C_{m-r})$ (12), или, в силу положительной определенности матрицы $D_\xi$, максимальное значение функции

$$\psi(C_1, C_2, ..., C_{m-r}) = \sum_{i=1}^{m-r} C_i^T D_\xi C_i .$$

Можно показать [6], что если $C_1, C_2, ..., C_m$ – единичные собственные векторы матрицы $D_\xi$ и $\lambda_1 > \lambda_2 > ... > \lambda_m$ – соответствующие им собственные числа, то $C_1^T D_\xi C_1 > C_2^T D_\xi C_2 > ... > C_m^T D_\xi C_m$. В силу этого свойства в качестве направляющих векторов $C_1, C_2, ..., C_{m-r}$ искомых линейных многообразий следует взять $(m-r)$ собственных векторов матрицы $D_\xi$, соответствующих максимальным собственным числам. Например, при отыскании прямой линии ($r = m-1$) нужно выбрать один из $m$ собственных векторов, а именно, собственный вектор $C_1$, соответствующий максимальному собственному числу $\lambda_1$. При отыскании $(m-1)$-мерной гиперплоскости ($r$ = 1) выбираются $m-1$ собственных векторов $C_1, C_2, ..., C_{m-1}$, соответствующих собственным числам $\lambda_1 > \lambda_2 > ... > \lambda_m$. При отыскании точки собственные векторы находить не нужно: искомой точкой $C_0$ будет математическое ожидание случайного вектора $\xi$ (11).

Таким образом, мы доказали следующую теорему.

Т е о р е м а. *Пусть* $\xi^T = (\xi_1, ..., \xi_m) \in E^m$ *– случайный вектор в m-мерном евклидовом пространстве* $E^m$ *со средним значением* $C_0 = A_\xi = E(\xi)$ *и положительно определенной дисперсионной матрицей* $D_\xi = E((\xi - A_\xi)(\xi - A_\xi)^T)$, $\lambda_1 > \lambda_2 > ... > \lambda_m$ *– собственные числа матрицы* $D_\xi$, $C_1, C_2, ..., C_m$ *– соответствующие этим числам собственные векторы. Тогда линейное многообразие* (2) (*или* (3)) *обеспечивает минимальную вариацию* $E(\rho^2(\xi, X))$ *квадрата расстояния* $\rho^2$ *от точки* $\xi$ *до этого многообразия. При разбиении в* (3) *векторов X,* $C_0$ *и матрицы C на блоки в виде* (4) *полученное многообразие* (3) *может быть представлено в форме* (6).

Если теоретические моменты $A_\xi$, $D_\xi$ случайного вектора $\xi$ заменить соответствующими эмпирическими моментами

$$\hat{A}_\xi = \frac{1}{n}\sum_{i=1}^{n} X_{o,i}, \tag{14}$$

$$\hat{D}_\xi = \frac{1}{n}\sum_{i=1}^{n} (X_{o,i} - \hat{A}_\xi)(X_{o,i} - \hat{A}_\xi)^T, \tag{15}$$

где $X_{o,i}$, $i = \overline{1,n}$, – векторы-столбцы наблюдений случайного вектора $\xi$, то мы получим симметричную эмпирическую аппроксимацию векторных статистических данных линейными многообразиями.

**3. Примеры.** Для демонстрации техники предложенной аппроксимации и ее графической иллюстрации рассмотрим два примера.

П р и м е р 1. Этот пример взят из работы К. Пирсона [1]. Заданы 4 точки $X_{o,i}^T = (x_{o,1}, x_{o,2}, x_{o,3})$ в трехмерном пространстве:

$$X_{o,1}^T = (2,16,219),\ X_{o,2}^T = (2,26,261),\ X_{o,3}^T = (4,16,127),\ X_{o,4}^T = (4,26,231). \tag{16}$$

В [1] получена следующая симметричная аппроксимация этих данных плоскостью:

$$x_3 + 38,02187x_1 - 7,35823x_2 - 169,03778 = 0. \tag{17}$$

В соответствии с подходом нашей работы для получения симметричной аппроксимации необходимо по имеющимся данным найти выборочное среднее $\hat{A}_\xi$ (14) и выборочную дисперсионную матрицу $\hat{D}_\xi$ (15). Получим

$$\hat{A}_\xi = \begin{pmatrix} 3{,}0000 \\ 21{,}0000 \\ 209{,}5000 \end{pmatrix}, \hat{D}_\xi = \begin{pmatrix} 0{,}0010 & 0 & -0{,}0305 \\ 0 & 0{,}0250 & 0{,}1825 \\ -0{,}0305 & 0{,}1825 & 2{,}5027 \end{pmatrix} \cdot 10^3.$$

Далее необходимо найти собственные числа $\lambda_1 > \lambda_2 > \lambda_3$ и соответствующие им собственные векторы $C_1$, $C_2$, $C_3$ матрицы $\hat{D}_\xi$. Имеем

$$\lambda_1 = 2{,}5165 \cdot 10^3, \lambda_2 = 0{,}0121 \cdot 10^3, \lambda_3 = 0{,}0002 \cdot 10^3,$$

$$\mathrm{C}_1 = \begin{pmatrix} -0{,}0121 \\ 0{,}0730 \\ 0{,}9973 \end{pmatrix}, \mathrm{C}_2 = \begin{pmatrix} 0{,}1913 \\ 0{,}9791 \\ -0{,}0694 \end{pmatrix}, \mathrm{C}_3 = \begin{pmatrix} 0{,}9815 \\ -0{,}899 \\ 0{,}0258 \end{pmatrix}.$$

Наконец, записываем аппроксимирующие многообразия в параметрической форме. Для двухмерной плоскости

$$Y = \hat{A}_\xi + t_1 C_1 + t_2 C_2, \tag{18}$$

прямой линии

$$Y = \hat{A}_\xi + t_1 C_1 \tag{19}$$

и точки

$$Y = \hat{A}_\xi. \tag{20}$$

С помощью уравнения (6) вместо уравнения плоскости в параметрической форме (18) мы имеем эквивалентное уравнение в обычной форме:

$$x_3 + 38{,}02214 x_1 - 7{,}35822 x_2 - 169{,}04363 = 0. \tag{21}$$

Соответствующие коэффициенты уравнений (17) и (21) совпадают с точностью до четырех значащих цифр. Кратчайшее расстояние от точки $X_{o,3}^T = (4{,}16{,}127)$ до плоскости (18) (или (21)), подсчитанное по формуле (8), равняется 0,1984, в то время как расстояние от этой точки до плоскости по вертикали составляет 7,6867. Эти числа согласуются с соответствующими числами работы [1]. Сумма квадратов кратчайших расстояний от всех четырех точек до плоскости (18) (или (21)), подсчитанная по формуле (8), составляет 0,7913, а сумма квадратов расстояний от четырех точек до этой же плоскости по вертикали – 1187,7. Полученные цифры служат подтверждением правильности решения оптимизационной задачи (7).

Классическая линейная средняя квадратичная регрессия, рассчитанная по тем же данным (16), определяется выражением

$$x_3 = 209{,}5 - 30{,}5(x_1 - 3) + 7{,}3(x_2 - 21). \tag{22}$$

Сумма квадратов расстояний от четырех точек до этой плоскости по вертикали составляет 961. Это меньше, чем для предыдущей аппроксимации, что также согласуется с теорией.

На рис. 1 изображены заданные точки (16), аппроксимирующие плоскости (17), (21), (22), аппроксимирующая прямая (19) и аппроксимирующая точка (20). Как видно, аппроксимирующая плоскость (21) данной статьи совпадает с плоскостью (17), полученной К. Пирсоном. Вместе с тем эти плоскости отличаются, хотя и незначительно, от плоскости (22), определяемой классической линейной средней квадратичной регрессией.

П р и м е р 2. Моделировались реализации четырехмерного случайного вектора $\xi^T = (\xi_1, \xi_2, \xi_3, \xi_4)$ с нормальным (гауссовским) распределением, нулевым средним значением $A_\xi^T = (0, 0, 0, 0)$, дисперсиями компонент $D(\xi_1) = D(\xi_2) = D(\xi_3) = D(\xi_4) = 1$ и корреляционной матрицей

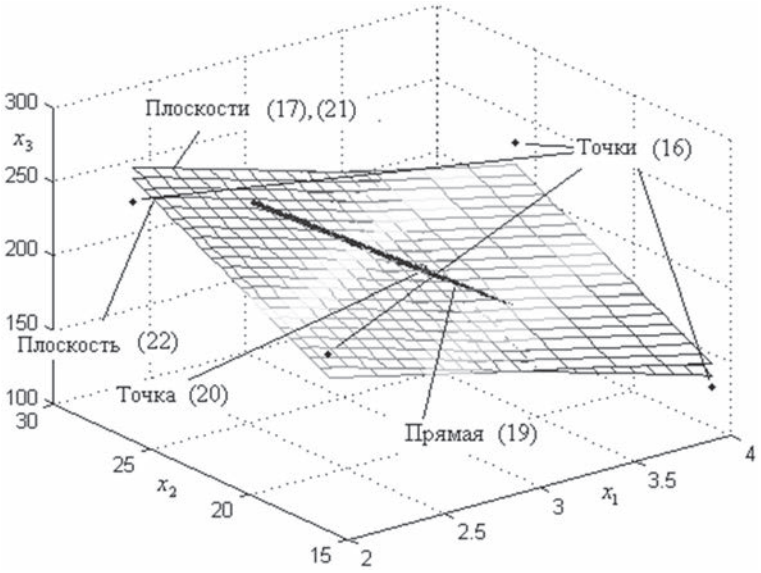


Рис. 1. Графическая иллюстрация к примеру 1
Fig. 1. Graphical illustration to example 1

$$\rho = \begin{pmatrix} 1 & 0,6 & 0,3 & 0,7 \\ 0,6 & 1 & 0,6 & 0,6 \\ 0,3 & 0,6 & 1 & 0,5 \\ 0,7 & 0,6 & 0,5 & 1 \end{pmatrix}.$$

По одной из выборок объема $n = 100$ получены следующие характеристики: выборочное среднее

$$\widehat{A}_\xi^T = C_0^T = \begin{pmatrix} 0,1698 & -0,1051 & -0,0761 & -0,0524 \end{pmatrix},$$

выборочная дисперсионная матрица

$$\widehat{D}_\xi = \begin{pmatrix} 2,5210 & 1,5246 & 0,4992 & 1,5448 \\ 1,5246 & 2,5348 & 1,4030 & 1,4540 \\ 0,4992 & 1,4030 & 3,2561 & 1,1649 \\ 1,5448 & 1,4540 & 1,1649 & 2,5803 \end{pmatrix},$$

собственные числа $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4$ выборочной дисперсионной матрицы $\widehat{D}_\xi$

$$\lambda_1 = 6,5369,\ \lambda_2 = 2,5226,\ \lambda_3 = 1,0973,\ \lambda_4 = 0,7354$$

и соответствующие этим собственным числам собственные векторы $C_1$, $C_2$, $C_3$, $C_4$:

$$C_1 = \begin{pmatrix} 0,4620 \\ 0,5336 \\ 0,4826 \\ 0,5186 \end{pmatrix},\ C_2 = \begin{pmatrix} 0,5481 \\ 0,0575 \\ -0,8088 \\ 0,2052 \end{pmatrix},\ C_3 = \begin{pmatrix} -0,1678 \\ 0,6257 \\ 0,0349 \\ 0,7610 \end{pmatrix},\ C_4 = \begin{pmatrix} -0,6767 \\ 0,5661 \\ -0,3343 \\ 0,3315 \end{pmatrix}.$$

Полученные собственные векторы позволяют записать уравнения аппроксимирующих плоскостей: нульмерной $X = \widehat{A}_\xi$, одномерной $X = A\ + t_1 C_1$, двумерной $X = \widehat{A}_\xi + t_1 C_1 + t_2 C_2$ и трехмерной $X = \widehat{A}_\xi + t_1 C_1 + t_2 C_2 + t_3 C_3$. Остановимся более подробно на двумерной аппроксимирующей плоскости $X = \widehat{A}_\xi + t_1 C_1 + t_2 C_2$, для чего получим ее эквивалентное представление в виде (6). В соответствии с формулами (4) имеем

$$C = (C_1, C_2) = \begin{pmatrix} 0{,}4620 & 0{,}5481 \\ 0{,}5336 & 0{,}0575 \\ 0{,}4826 & -0{,}8088 \\ 0{,}5186 & 0{,}2052 \end{pmatrix}, \; X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}, \; C_0 = \begin{pmatrix} 0{,}1698 \\ -0{,}1051 \\ -0{,}0761 \\ -0{,}0524 \end{pmatrix},$$

$$\overline{C}_{m-r} = \begin{pmatrix} 0{,}4620 & 0{,}5481 \\ 0{,}5336 & 0{,}0575 \end{pmatrix}, \; \overline{C}_r = \begin{pmatrix} 0{,}4826 & -0{,}8088 \\ 0{,}5186 & 0{,}2052 \end{pmatrix}, \; \overline{X}_{m-r} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \; \overline{X}_r = \begin{pmatrix} x_3 \\ x_4 \end{pmatrix},$$

$$\overline{C}_{0,m-r} = \begin{pmatrix} 0{,}1698 \\ -0{,}1051 \end{pmatrix}, \; \overline{C}_{0,r} = \begin{pmatrix} -0{,}0761 \\ -0{,}0524 \end{pmatrix}, \; \overline{C}_r (\overline{C}_{m-r})^{-1} = \begin{pmatrix} -1{,}7271 & 2{,}3999 \\ 0{,}2997 & 0{,}7123 \end{pmatrix}.$$

По формуле (6) получим

$$\begin{pmatrix} x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} -0{,}0761 \\ -0{,}0524 \end{pmatrix} + \begin{pmatrix} -1{,}7271 & 2{,}3999 \\ 0{,}2997 & 0{,}7123 \end{pmatrix} \cdot \begin{pmatrix} x_1 - 0{,}1698 \\ x_2 + 0{,}1051 \end{pmatrix}, \tag{23}$$

или

$$\begin{aligned} x_3 &= -0{,}0761 - 1{,}7271(x_1 - 0{,}1698) + 2{,}3999(x_2 + 0{,}1051), \\ x_4 &= -0{,}0524 + 0{,}2997(x_1 - 0{,}1698) + 0{,}7123(x_2 + 0{,}1051). \end{aligned} \tag{24}$$

Таким образом, двумерная аппроксимирующая плоскость $X = \hat{A}_\xi + t_1 C_1 + t_2 C_2$ эквивалентна линейной связи (23) двумерного вектора $(x_3, x_4)$ с двумерным вектором $(x_1, x_2)$.

По вектору $\hat{A}_\xi^T$ и матрице $\hat{D}_\xi$ можно также получить классическую линейную среднюю квадратичную регрессию $(\xi_3, \xi_4)$ на $(\xi_1, \xi_2)$ [7]:

$$\begin{pmatrix} x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} -0{,}0761 \\ -0{,}0524 \end{pmatrix} + \begin{pmatrix} 0{,}4992 & 1{,}4030 \\ 1{,}5448 & 1{,}4540 \end{pmatrix} \cdot \begin{pmatrix} 2{,}5210 & 1{,}5246 \\ 1{,}5246 & 2{,}5348 \end{pmatrix}^{-1} \cdot \begin{pmatrix} x_1 - 0{,}1698 \\ x_2 + 0{,}1051 \end{pmatrix} =$$

$$= \begin{pmatrix} -0{,}0761 \\ -0{,}0524 \end{pmatrix} + \begin{pmatrix} -0{,}2149 & 0{,}6827 \\ 0{,}4179 & 0{,}3223 \end{pmatrix} \cdot \begin{pmatrix} x_1 - 0{,}1698 \\ x_2 + 0{,}1051 \end{pmatrix},$$

или

$$\begin{aligned} x_3 &= -0{,}0761 - 0{,}2149(x_1 - 0{,}1698) + 0{,}6827(x_2 + 0{,}1051), \\ x_4 &= -0{,}0524 + 0{,}4179(x_1 - 0{,}1698) + 0{,}3223(x_2 + 0{,}1051). \end{aligned} \tag{25}$$

На рис. 2 изображены случайные числа $(x_1, x_2, x_4)$ и аппроксимирующие их плоскости (24), (25), полученные различными методами. Как видно, различные методы аппроксимации приводят в данном примере к существенно различным аппроксимирующим плоскостям.

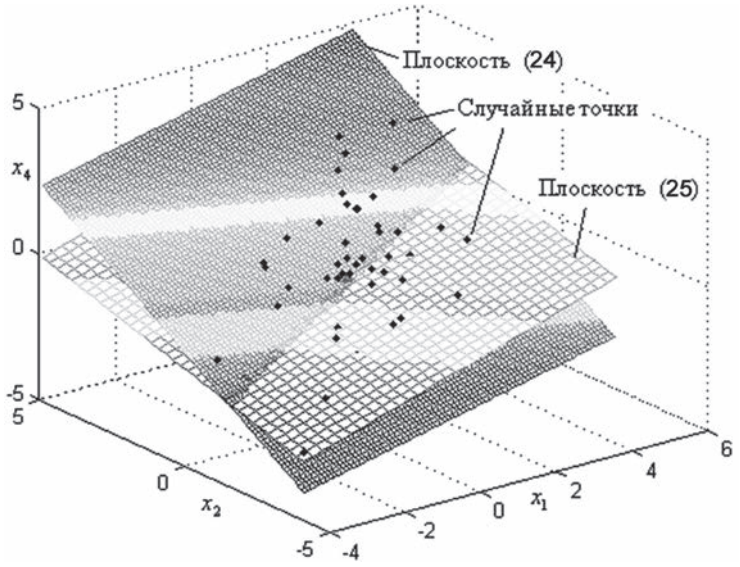


Рис. 2. Графическая иллюстрация к примеру 2

Fig. 2. Graphical illustration to example 2

**Заключение.** Предложенная в статье симметричная аппроксимация может быть выполнена как в случае, когда вектор $\xi^T = (\xi_1, ..., \xi_m)$ является случайным по природе, так и в случае, когда он представляет собой измеренные с ошибками входные и выходные переменные некоторого детерминированного объекта, и мы желаем аппроксимировать этот объект линейной зависимостью. Однако вопрос предпочтительного использования симметричной линейной аппроксимации или линейной регрессии требует отдельных исследований. Можно лишь с уверенностью сказать, что симметричную аппроксимацию следует применять тогда, когда минимизация ее критерия является очевидной целью аппроксимации, например, при отыскании линейных маршрутов, обладающих минимальным (максимальным) суммарным удалением от некоторых объектов.

## Список использованных источников

1. Pearson, K. On lines and planes of closest fit to systems of points in space / K. Pearson / Philos. Mag. – 1901. – Vol. 6, N 2. – P. 559–572.

2. Крамер, Г. Математические методы статистики / Г. Крамер. – М.: Мир, 1975. – 648 с.

3. Утешев, А. Ю. Записная книжка на виртуальном факультете [Электронный ресурс] / А. Ю. Утешев. – Режим доступа: http://www.apmath.spbu.ru/ru/staff/uteshev/index.html, свободный.

4. Амосов, А. А. Скалярно-матричное дифференцирование и его приложения к конструктивным задачам теории связи / А. А. Амосов, В. В. Колпаков // Проблемы передачи информации. – 1972. – №. 8, вып. 1. – С. 3–15.

5. Муха, В. С. Анализ многомерных данных / В. С. Муха. – Минск: Технопринт, 2004. – 366 с.

6. Вержбицкий, В. М. Численные методы (линейная алгебра и нелинейные уравнения) / В. М. Вержбицкий. – М.: Высш. шк., 2000. – 266 с.

7. Рао, С. Р. Линейные статистические методы и их применение / С. Р. Рао. – М.: Наука, 1968. – 548 с.

## References

1. Pearson K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 1901, vol. 6, no. 2, pp. 559–572. doi:10.1080/14786440109462720.

2. Cramer H. *Mathematical Methods of Statistics (PMS-9).* Princeton University Press, 1999. 575 p.

3. Uteshev A.Yu. *Notebook on the virtual faculty.* Available at: http://www.apmath.spbu.ru/ru/staff/uteshev/index.html. (in Russian)

4. Amosov A.A., Kolpakov V.V. Scalar-matrix differentiation and its applications to the constructive tasks of communication. *Problemy peredachi informatsii* [Problems of Information Transmission], 1972, no. 8, iss. 1, pp. 3–15. (in Russian)

5. Mukha V.S. *Analysis of multidimensional data*. Minsk, Technoprint, 2004. 366 p. (in Russian)

6. Verzhbitskii V.M. *Numerical methods (linear algebra and nonlinear equations).* Moscow, Vysshaia shkola Publ., 2000. 266 p. (in Russian)

7. Rao C.R. *Linear statistical inference and its applications*. 2 ed. Wiley, 1973. 648 p.

## Информация об авторе

**Муха Владимир Степанович** – доктор технических наук, прфессор, профессор кафедры информационных технологий автоматизированных систем, Белорусский государственный университет информатики и радиоэлектроники (ул. П. Бровки, 6, 220013, г. Минск, Республика Беларусь). E-mail: mukha@bsuir.by

## Information about the author

**Mukha Vladimir Stepanovich** – D. Sc. (Engineering), Professor, Professor of the Department of Automated Data Processing Systems, Belarusian State University of Informatics and Radioelectronics (6, P. Brovka Str., 220013, Minsk, Republic of Belarus). E-mail: mukha@bsuir.by

## Для цитирования

Муха, В. С. Симметричная аппроксимация векторных статистических данных линейными многообразиями / В. С. Муха // Вес. Нац. акад. навук Беларусі. Сер. фіз.-мат. навук. – 2016. – № 4. – С. 23–31.

## For citation

Mukha V.S. Symmetric approximation of vector statistical data by linear varieties. *Vestsi Natsyianal'nai akademii navuk Belarusi. Seryia fizika-matematychnykh navuk* [Proceedings of the National Academy of Sciences of Belarus. Physics and Mathematics series], 2016, no. 4, pp. 23–31. (in Russian)