МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
КАЗАНСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ

Н.А. МИРСАЕВА

ИСПОЛЬЗОВАНИЕ НЕКОТОРЫХ МЕТОДОВ ВЕКТОРНО-МАТРИЧНОЙ АЛГЕБРЫ В ЗАДАЧАХ МЕТЕОРОЛОГИИ

Учебно-методическое пособие
для бакалавров направления 05.03.04 «Гидрометеорология»

КАЗАНЬ
2018

УДК 551.509
ББК 26.23

Печатается по рекомендации методической комиссии
Института экологии и природопользования
ФГАОУ ВО «Казанский (Приволжский) федеральный университет»
(протокол №7 от 28 ноября 2018 г).

заседания кафедры метеорологии, климатологии и экологии атмосферы
(протокол №4 от 15 ноября 2018 г).

Рецензенты:
кандидат географических наук, доцент КФУ **М.А. Верещагин;**
кандидат географических наук, доцент КФУ **А.А. Николаев**

Мирсаева Н.А.
Использование некоторых методов векторно-матричной алгебры в задачах метеорологии: учеб.-метод. пособие / Н.А. Мирсаева – Казань: Изд-во Казан. ун-та, 2018. – 21 с.

Данное пособие составлено в соответствии с материалом курса лекций «Статистические методы гидрометеорологического прогнозирования», читаемой бакалаврам, обучающимся по направлению 05.03.04 «Гидрометеорология». В него включены основные понятия, используемые в указанной дисциплине, а также уделяется основное внимание теоретическим основам статистических методов гидрометеорологического прогнозирования. Пособие предназначено для студентов вузов, аспирантов и преподавателей.

УДК 551.509
ББК 26.23

Введение

Метеорология, как наука об атмосфере, выводит многие свои закономерности, опираясь на результаты обобщений массовых приземных аэрологических, спутниковых и других измерений параметров ее физического состояния. Собираемая таким образом первичная информация отличается, как правило, ее значительной емкостью, что нередко приводит к необходимости ее предварительного группирования в виде некоторых компактных множеств, например, в виде векторов и матриц, служащих основанием линейной алгебры.

Ниже иллюстрируются отдельные примеры того, какие задачи из области метеорологии можно успешно решать, опираясь на фундаментальные понятия векторно-матричной алгебры.

1. Отображение метеорологических рядов и карт в виде n-мерных векторов

Если совокупность результатов регулярных измерений температуры воздуха (x_i), выполненных в стандартные сроки наблюдений ($i=\overline{1,n}$), записать в идее столбца

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{pmatrix} \qquad (1)$$

то, сообразуясь с теоретическими положениями [1], эту же совокупность можно интерпретировать, с одной стороны, как реализацию n-мерного вектора-столбца, и, одновременно, как матрицу-столбец порядка (1×n).

Если результаты тех же измерений записать в виде строки

$$x^{\mathrm{T}} = [x_1, x_2, x_3] \qquad (2)$$

то, последняя может рассматриваться как транспонированное (т) значение того же вектора, именуемого как вектор-строка или матрица-строка (порядка n×1).

Условным отображением вектора (1) может быть его запись в виде $\|x_i\|$ или в виде $\{x_i\}$ ($i=\overline{1,n}$). Любое метеорологическое поле некоторой величины (температура, геопотенциал и др.), задаваемое в

узлах регулярной географической сетки, может быть представлено так же в виде n-мерного вектора x (1) или x^T (2).

Табличные данные приложения 1 с позиций матрично-векторной алгебры [1] можно рассматривать и как выборку векторов (1), (2), и как матрицу порядка (n×m)

$$M = \begin{pmatrix} x_{11}\, x_{12} \dots x_{1m} \\ x_{21}\, x_{22} \dots x_{2m} \\ \dots \dots \dots \dots \dots \dots \\ x_{n1}\, x_{n2} \dots x_{nm} \end{pmatrix} = \|x_{ij}\|, \quad (3)$$

элементы которой (x_{ij}) представляют собою значения аномалий температуры воздуха (АТВ) на ст. Казань, университет в i-м году и в j-м месяце.

Одной из часто решаемых задач метеорологии является задача о сравнении хода во времени какой-либо метеорологической величины, зафиксированной на самой поздней части истории и в предыстории. Например, может возникнуть потребность сравнения хода АТВ (прил. 1), начиная с января 2001 г. по 2010 г. (реализация вектора x (1)) с многолетним ходом тех же январских АТВ, начиная с 1981 г., по 1990 г. (реализация вектора y(1)).

Векторы $\|x_i\|$ и $\|y_i\|$ можно сравнивать, вычисляя, например, квадрат расстояния («евклидово расстояние») между ними

$$r(x,y) = (x-y)^T(x-y) = \sum_{i=1}^{i=n}(x_i - y_i)^2, \quad (4)$$

($r(x,y) \geq 0$), либо путем оценки косинуса угла между ними

$$\cos(x,y) = x^T y[(x^T x)(y^T y)]^{-0,5} = \sum_{i=1}^{i=n} x_i\, y_i \left[\sum_{i=1}^{i=n} x_i^2 \sum_{i=1}^{i=n} y_i^2\right]^{-0,5}, \quad (5)$$

(-1,0≤cos(x,y)≤1,0).

Однако, ни один из указанных подходов (4), (5) к сравнению векторов x и y не является исчерпывающим. Самой же полной мерой сходства указанных векторов является комплексный показатель их сходства

$$\xi(x,y) = \cos(x,y) + (1 + r(x,y))^{-1}. \quad (6)$$

При этом -1,0≤ξ (x,y)≤2,0. При ξ (x,y)=2,0 сравниваемые векторы строго коллинеарны и имеют одинаковую длину, при ξ (x,y)=-1,0 векторы x и y противоположны по направлению и в общем случае имеют неодинаковую длину.

Задача. С использованием алгоритмов (4) – (6) оценить показатели сходства (аналогичности) между многолетней динамикой январских АТВ в Казани в 2001 – 2010 гг. (вектор x) и динамикой тех же АТВ в 1981 – 1990 гг. (вектор y) и в 1991 – 2000 гг. (вектор z, прил. 1).

Решение. Сравнение векторов x и y приводит к следующим результатам:

$r(\mathrm{x},\mathrm{y})=[(7{,}65-5{,}55)^2+(5{,}15-0{,}05)^2+\ldots((-4{,}25)-1{,}95)^2]=395{,}460$

$$\cos(x,y)=\frac{5{,}55\cdot 7{,}65+0{,}05\cdot 5{,}15+\ldots 1{,}95\cdot(-4{,}25)}{[(5{,}55^2+0{,}05^2+\ldots+1{,}95^2)(7{,}85^2+5{,}15^2+\ldots+(-4{,}25^2))]^{0{,}5}}=0{,}080,$$

$\xi(\mathrm{x},\mathrm{y})=\cos(x,y)+(1+r(x,y))^{-1}=0{,}083.$

Действуя аналогичным образом, найдем показатели сходства между динамикой АТВ в 2001 – 2010 гг. (вектор x) и динамикой АТВ в 1991 – 2000 гг. (вектор z):

$r(\mathrm{x},z)=286{,}980;\ \cos(x,z)=0{,}304;\ \xi(\mathrm{x},z)=0{,}307.$

Как видно, многолетняя динамика АТВ в 1991 – 2000 гг. (вектор z) в сравнении ее с той же динамикой в 1981 – 1990 гг. (вектор y) по всем показателям находилась в более лучшем согласии с динамикой АТВ в 2001 – 2009 гг. (вектор x).

Задачи для самостоятельного решения

1. Многолетние изменения АТВ в апреле и марте в 1980 – 2000 –м гг. (прил. 1), рассматриваются как реализации векторов x и y (соответственно). Оценить величины квадратов длин S(x) и S(y) указанных векторов.

2. Многолетние (2000 – 2010 гг.) изменения АТВ в ноябре и декабре (прил. 1), рассматриваются как реализации векторов x и y (соответственно). Оценить величину угла (α, град) между указанными векторами.

3. Рассматривая многолетние изменения АТВ в октябре на отрезках времени 1987 – 1992 гг., 1993 – 1998 гг., 1999 – 2004 гг., 2005 – 2010 гг., как реализации векторов κ, z, y и x (соответственно), рассчитать полные меры сходства $\xi(\mathrm{x},\mathrm{y})$, $\xi(\mathrm{x},z)$, $\xi(\mathrm{x},\mathrm{к})$. На каком из указанных 3-х отрезков времени многолетний ход АТВ ближе всего согласуется с ходом тех же АТВ в 2005 – 2010 гг.?

4. После визуального изучения многолетнего хода АТВ в 2001 – 2010 гг. в январе (вектор x), феврале (вектор y) и в марте (вектор z) постулировалось: ближе всего к динамике АТВ в январе были их изменения в марте. Подкрепляется ли постулируемое утверждение показателями сходства векторов $\xi(\mathrm{x},\mathrm{y})$, $\xi(\mathrm{x},z)$?

5. После визуального изучения графического отображения многолетнего (2001 – 2010 гг.) хода АТВ в январе (вектор x), феврале (вектор y) и в марте (вектор z) создалось впечатление, что вектор y по направленности изменений его компонентов ближе всего согласуется

с направленностью изменений компонентов вектора *z*. Проверить, насколько оказалось верным начальное суждение в отношении направленности многолетних изменений АТВ в указанных месяцах?

2. Построение линейных регрессионных моделей

В учебных дисциплинах по метеорологии можно найти множество примеров, когда изменения той или иной метеорологической величины (*y*), именуемой как результативный признак (предиктант), находятся в зависимости от изменений комплекса факториальных признаков (x_1, x_2, x_3,…, x_n), и когда эта зависимость аппроксимируется (в простейшем случае) полиномом первой степени

$$a_1 x_1 + a_2 x_2 + \ldots a_n x_n = y. \qquad (7)$$

Уравнение (7) известно, как уравнение линейной регрессии, в котором $a_i (i=\overline{1,n})$ – некоторые множители, подлежащие определению.

Регрессионная модель может быть с успехом использована, например, при изучении изменений температуры воздуха (*y*) в зависимости от широты (x_1), географической долготы (x_2) и высоты орографии (x_3) [5]. Множество примеров практического использования регрессионных моделей (7) можно найти также в курсах по синоптической метеорологии [6], долгосрочным метеорологическим прогнозам [2, 9].

Самостоятельная разработка моделей (7) предполагает практическое владение методикой определения оптимальных значений множителей $a_i (i=\overline{1,n})$, обеспечивающих минимум погрешности ее применения. Достигается это использованием «метода наименьших квадратов (МНК)» [2, 3].

Прежде, чем раскрывать суть МНК, условимся, что с самого начала под x_i и *y* в равенстве (7) будем рассматривать не сами их натуральные величины, а их отклонения от средних значений (величины аномалий), обратим внимание также на то, что равенство (7) в векторной форме может быть записано в виде произведения

$$a^T \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{pmatrix} = y, \qquad (8)$$

в котором $a^T = (a_1 a_2 \dots a_n)$ – символ транспонированного вектора множителей a_i, содержимое круглой скобы интерпретируется как *n*-мерный вектор-предиктор *x*.

Реализация МНК предполагает наличие архивной выборки: матрицы векторов-предикторов $x_{m\times n}$

$$x = \begin{pmatrix} x_1\ x_2 \dots x_n \\ x_1\ x_2 \dots x_n \\ \dots\ \dots\ \dots\ \dots\ \dots \\ x_1\ x_2 \dots x_n \end{pmatrix} \quad (9)$$

и компонентов вектора-предиктанта $Y_{(m\times 1)}$

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{pmatrix}. \quad (10)$$

Здесь m – объем архивной выборки (9, 10) или число опытов, в которых фиксировались компоненты x_1, x_2, x_3,…, x_n вектора-предиктора x_i и соответствующие им значения компонентов вектора-предиктанта (y_i).

Для разыскания неизвестных множителей $a_i(i=\overline{1,n})$ вначале записывается система линейных уравнения с n-неизвестными

$$(x^T x)\begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_n \end{pmatrix} = x^T y = Q, \quad (11)$$

а затем формируется определитель (n-го порядка) системы уравнений (11)

$$\begin{vmatrix} \sum x_1^2 & \sum x_1 x_2 & \dots & \sum x_1 x_n \\ \sum x_2 x_1 & \sum x_2^2 & \dots & \sum x_2 x_n \\ \dots & \dots & \dots & \dots \\ \sum x_n x_1 & \sum x_n x_2 & \dots & \sum x_n^2 \end{vmatrix}, \quad (12)$$

и состав вектора сумм смешанных произведений

$$Q = \begin{pmatrix} \sum x_1 Y \\ \sum x_2 Y \\ \dots\ \dots \\ \sum x_n Y \end{pmatrix}. \quad (13)$$

Разыскиваемые множители a_i определяются по методу Крамера [3]

$$a_i = \frac{\Delta a_i}{\Delta}. \quad (14)$$

Здесь $i=\overline{1,n}$, Δ - числовое значение определителя (12), Δa_i – числовые оценки определителей n-го порядка, образуемых заменой *j*-

го столбца определителя (12) столбцом с компонентами вектора *Q*. Суммирования смешанных произведений в (12) производятся в пределах архивной выборки (9) (от *i*=1 до *i*=*m*).

Реализация МНК предполагает обязательную выполнимость требования неравенства: Δ≠0.

Для разыскивания множителей a_i($i=\overline{1,n}$) может быть реализован также иной алгоритм

$$a = \begin{pmatrix} a_1 \\ a_2 \\ \dots \\ a_n \end{pmatrix} = M^{-1}q, \qquad (15)$$

в котором M^{-1} – матрица, обратная матрице ковариаций

$$M = \begin{pmatrix} \overline{x_1^2} & \overline{x_1 x_2} & \dots & \overline{x_1 x_n} \\ \overline{x_2 x_1} & \overline{x_2^2} & \dots & \overline{x_2 x_n} \\ \dots & \dots & \dots & \dots \\ \overline{x_n x_1} & \overline{x_n x_2} & \dots & \overline{x_n^2} \end{pmatrix}. \qquad (16)$$

Верхние черточки над элементами матрицы *M* являются символом осреднения (деления на объем выборки – *m*) всех элементов определителя (12).

Вектор ковариаций q

$$q = \frac{1}{m}Q \qquad (17)$$

является результатом перемножения вектора смешанных произведений *Q* (13) на скаляр 1/*m*.

Важным показателем разрешающей способности линейной регрессионной модели (1) является ее коэффициент детерминации

$$R^2_{y.x_1x_2\dots x_n} = R^2 = \frac{1}{\sigma^2(y)}[q^T M^{-1} q], \qquad (18)$$

представляющий собою отношение дисперсии $\sigma^2(y)_{x_1x_2\dots x_n}$ результативного признака, воспроизводимой учетом совместного действия учтенного комплекса факториальных признаков, к полной дисперсии предиктанта $\sigma^2(y)$. При этом $0 \le R^2 \le 1{,}0$. При $R^2 \to 0$ разрешающая способность модели (1) очень мала, при $R^2 \to 1$ – максимальна.

Задача. Имеется архивная выборка векторов-предикторов x_1, x_2 и компонентов вектора предиктанта *y* (9,10). Соответствующие данные отражены в табл. 1.

Таблица 1

Номера испытаний	Компоненты векторов		
	x_1	x_2	y
1	0,8	-0,1	2,1
2	-0,6	0,4	3,8
3	0,2	-0,2	0,2
4	1,4	0,5	0,4
5	1,8	0,6	1,0

Необходимо определить множители a_1, a_2 модели

$$a_1 x_1 + a_2 x_2 = y.$$

При этом необходимо показать:

- вид определителя 2-го порядка системы линейных уравнений (11), числовые оценки определителей Δ, Δa_1, Δa_2, векторов Q (13), q (16); сформировать матрицу ковариаций M (16), оценить дисперсию предиктанта $\sigma^2(y)$.

Оценку матрицы M^1 выполнить согласно рекомендаций [1].

Решение.

Действуя согласно вышесказанному и принимая во внимание объем выборки m=5, имеем:

- определитель системы линейных уравнений

$$\begin{vmatrix} 6{,}240 & 1{,}420 \\ 1{,}420 & 0{,}820 \end{vmatrix}$$

- матрицу ковариаций

$$M = \begin{pmatrix} 1{,}248 & 0{,}284 \\ 0{,}284 & 0{,}164 \end{pmatrix}$$

- векторы Q и q

$$Q = \begin{pmatrix} 1{,}800 \\ 2{,}070 \end{pmatrix}$$

$$q = \frac{1}{5} Q = \begin{pmatrix} 0{,}360 \\ 0{,}414 \end{pmatrix}$$

- числовые оценки определителей

Δ=3,100

Δa_1=-1,463

Δa_2=10,361

- множители

a_1=-0,472

a_2=3,342

- вектор q^T=(0,360 0,414)
- обратная ковариационная матрица M^{-1}=$\begin{pmatrix} 1{,}322 & -2{,}290 \\ -2{,}290 & 10{,}063 \end{pmatrix}$
- дисперсия предиктанта $\sigma^2(y)$=1,760.
- коэффициент детерминации R^2=0,689.

Результаты определения множителей a_1 и a_2 вторым способом:

$$a = M^{-1}q = \begin{pmatrix} -0{,}472 \\ 3{,}342 \end{pmatrix}.$$

Задачи для самостоятельного решения

1. Определить оптимальные значения параметров линейной регрессионной модели (1) (см. п. 1), в которой x_1, x_2, x_3 – аномалии температуры воздуха (АТВ) на ст. Казань, университет (прил .1) в январе, феврале и марте (соответственно), и те же АТВ наблюдавшиеся в апреле (y) в том же периоде времени. За основу решения взять МНК (Г. Крамера). Полное решение должно содержать запись: определителя системы линейных уравнений и его числа (Δ), определителей $\Delta a_i\,(i=\overline{1,3})$, вектора сумм смешанных произведений Q (13) и искомых множителей a_1, a_2, a_3.
2. Рассматривая выборки многолетних (2001 – 2010 гг.) данных об АТВ в январе (x_1), феврале (x_2) и марте (x_3) (прил. 1) как матрицу X порядка 10×3, сформировать матрицу M^{-1} и проверить выполнимость условия $MM^{-1}=M^{-1}M=I$, где I – единичная матрица.
3. Выборка многолетних (2001 – 2010 гг.) данных об АТВ, положенная в основу для решения задачи 6 (см. выше) рассматривается как матрица X порядка 10×3, а данные об АТВ в те же годы в апреле – как реализация 10-ти-мерного вектора-предиктанта Y. Вычислить множители a_1, a_2, a_3 регрессионной модели (1) (п. 2.1), решая систему линейных уравнений с использованием обратной ковариационной матрицы (M^{-1}) и вектора ковариации q (17). Полное решение задачи должно сопровождаться записями матрицы M^{-1}, вектора q и множителей a_1, a_2, a_3.
4. Многолетние (2001 – 2010 гг.) данные по АТПО в сентябре (x_1), октябре (x_2) и ноябре (x_3) рассматриваются как матрица $X_{(10\times3)}$, а в декабре как реализация 10-ти мерного вектора-предиктанта Y. С использованием МНК определить множители a_1, a_2, a_3 регрессионной модели (1) (п. 2.1) и оценить величину коэффициента детерминации R^2. К ответу приложить оценку дисперсии предиктанта $[\sigma_0^2]$, вид

обратной ковариационной матрицы M^{-1}, вектора ковариаций q(17) и вектора q^T.

5. Применительно к условиям задачи 9 записать вид матрицы ковариаций M. Оценить квадраты длин векторов x_1[S(x_1)=?], x_2[S(x_2)=?] и x_3[S(x_3)=?].

3. Построение линейных регрессионных моделей с использованием элементов структуры корреляционных матриц

Для упрощения последующих рассуждений в линейной регрессионной модели (1) (п. 2.1) значение результативного признака (предиктанта y) перепишем на новую буквенную переменную x_1, факториальные признаки x_1, x_2, x_3… – соответственно, на буквенные переменные с измененными номерами – x_2, x_3, x_4…, и будем, таким образом, разыскивать параметры a_{1i} ($i=\overline{2,n}$) в несколько видоизмененном уравнении регрессии

$$x_1 = a_{12}x_2 + a_{13}x_3 + a_{14}x_4. \quad (19)$$

В двойных числовых подстрочных индексах a_{1i} в уравнении (19) отражены номера взаимодействующих признаков. Первый из них является индикатором предиктанта, а второй – порядковым номером учитываемого предиктора.

Можно показать, что любой множитель a_{1n} можно найти как произведение [3]

$$a_{1n} = (-1)^n \frac{\sigma_1}{\sigma_n}\frac{D_{1n}}{D_{11}}, \quad (20)$$

в котором σ_1 и σ_n – символы средних квадратических отклонений признаков 1 (предиктант) и n-го (факториального), D_{1n} и D_{11} – миноры определителя

$$D = \begin{vmatrix} 1 & r_{12} & r_{13} & \dots & r_{1n} \\ r_{21} & 1 & r_{23} & \dots & r_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & r_{n3} & \dots & 1 \end{vmatrix} \quad (21)$$

корреляционной матрицы $\|r_{ij}\|$ ($i=\overline{1,n}$, $j=\overline{1,n}$).

Элементами определителя (21) являются парные коэффициенты корреляции r_{ij} между признаками i и j.

Коэффициент детерминации комплекса предсказателей x_1, x_2, x_3 в этом случае можно определить, следуя равенству

$$R^2_{1.2,3,4} = \left(1 - \frac{D}{D_{11}}\right). \quad (22)$$

Рассматриваемый метод построения линейных регрессионных моделей (22) выгодно отличается от предыдущих методов тем, что позволяет легко получить (помимо $R^2_{1.2,3\ldots n}$) важный дополнительный показатель их разрешающей способности. В данном случае речь идет о величине средней квадратической погрешности использования модели (19)

$$S_{1.2,3\ldots n} = \sigma_1\sqrt{\frac{D}{D_{11}}}. \qquad (23)$$

Здесь σ_1– величина среднего квадратического отклонения результативного признака (1), все другие обозначения прежние.

Ранее уже обращалось внимание на то, что в регрессионной модели (19) взаимодействующие признаки x_i (i=$\overline{1,n}$) представляют собою их отклонения от средних арифметических значений (аномалии). В этой связи оценка средних квадратических отклонений σ_i (i=$\overline{1,n}$) всех признаков, действие которых предусмотрено в модели (19), значительно упрощаются

$$\sigma_i = \sqrt{\sum_{j=1}^{j=m} \mathrm{x}_{ij}^2}. \qquad (24)$$

Как видно (24), суммирование квадратов x_{ij}^2 признака i осуществляется в порядке очередности их следования в архивной выборке (x_{i1}, x_{i2}, … x^2_{ij} … x^2_{im}).

Задача. Многолетние (2001 – 2010 гг.) колебания АТВ в апреле, январе, феврале и марте (прил. 1) рассматриваются как изменения результативного (x_1) и факториальных (x_2, x_3, x_4) признаков (соответственно). Сформировать корреляционную матрицу $\|r_{ij}\|$ взаимодействующих признаков, определить множители a_{12}, a_{13} и a_{14} в уравнении регрессии (19) и коэффициент детерминации $R^2_{1.2,3,4}$.

Решение. После ввода в ПК базы данных, указанных в начале задачи, и запуска стандартной программы вычислений коэффициентов корреляции (Excel) имеем следующий вид корреляционной матрицы

$$\|r_{ij}\| = \begin{pmatrix} 1 & 0{,}027 & 0{,}055 & -0{,}063 \\ 0{,}027 & 1 & 0{,}204 & 0{,}198 \\ 0{,}055 & 0{,}204 & 1 & 0{,}572 \\ -0{,}063 & 0{,}198 & 0{,}572 & 1 \end{pmatrix}$$

Далее согласно алгоритму (24) находим:

- оценки средних квадратических отклонений

$\sigma_1 = 6{,}980^{\circ}\mathrm{C}$, $\sigma_2 = 16{,}408^{\circ}\mathrm{C}$, $\sigma_3 = 11{,}488^{\circ}\mathrm{C}$, $\sigma_4 = 10{,}074^{\circ}\mathrm{C}$.

- значения определителя корреляционной матрицы и его миноров

D_{11}=0,638, D_{12}=0,018, D_{13}=-0,084, D_{14}=-0,092.

Решение завершается оценками искомых множителей согласно алгоритму (20)

$$a_{12} = \frac{\sigma_1}{\sigma_2}\frac{D_{12}}{D_{11}} = 0{,}012,$$

$$a_{13} = -\frac{\sigma_1}{\sigma_3}\frac{D_{13}}{D_{11}} = 0{,}080,$$

$$a_{14} = \frac{\sigma_1}{\sigma_4}\frac{D_{14}}{D_{11}} = 0{,}100,$$

В итоге общий вид регрессионной модели (19) определится следующим образом

$x_1 = a_{12}x_2 + a_{13}x_3 + a_{14}x_4$=$0{,}012x_2 + 0{,}080x_3 + 0{,}100x_4$.

Задачи для самостоятельного решения

1. Используя данные прил. 1 о многолетней (2001 – 2010 гг.) динамике АТВ в марте (результативный признак – x_1) в декабре, январе и феврале (факториальные признаки x_2, x_3, x_4 – соответственно) сформировать корреляционную матрицу $\|r_{ij}\|$. К какому типу матриц она относится?
2. Многолетние колебания АТВ в сентябре 2001 – 2010 гг. (прил. 1) рассматриваются как поведение результативного (x_1) признака, а колебания АТВ в те же годы в июне, июле и августе – как поведение факториальных (x_2, x_3, x_4) признаков. Построить линейную прогностическую зависимость x_1=f(x_2, x_3, x_4). При этом выполнить оценки а) коэффициента детерминации $R^2_{x_1.x_2,x_3,x_4.}$

б) величины средней квадратической погрешности применения в прогностических целях линейной регрессии.

3. Корреляционная матрица комплекса взаимодействующих факторов имеет вид

$$\|r_{ij}\| = \begin{pmatrix} 1 & 0{,}21 & 0{,}49 \\ 0{,}21 & 1 & -0{,}25 \\ 0{,}49 & -0{,}25 & 1 \end{pmatrix},$$

Среднее квадратическое отклонение результативного признака $\sigma_1 = 10{,}0$°C. Определить величины: а) факториальной дисперсии $\sigma^2_{1.2,3}$ результативного признака, б) величину средней квадратической погрешности использования линейной регрессионной модели $x_1 = a_{12}x_2 + a_{13}x_3$.

4. Коэффициент детерминации модели $x_1 = a_{12}x_2 + a_{13}x_3 + a_{14}x_4$,

в которой x_1 – температура воздуха, x_2 – широта места, x_3 – долгота, x_4 –высота орографии, равен $R^2_{1.2,3,4} = 0{,}60$. Определитель корреляционной матрицы (D) и минор D_{12} равны 0,60 и 0,40 (соответственно), σ_1=0,50, σ_2=0,25. Оценить величины: а) средней квадратической погрешности $S_{1.2,3,4}$ рассматриваемой модели и б) множителя a_{12}.

4. Множественная и частная корреляция

В метеорологии и климатологии можно найти немало примеров, когда требуется установить степень согласованности изменений между какими-либо признаками, например, X и Y и оценить меру влияния поведения одного из них (X) на другой (Y).

В этом случае, обыкновенно, прибегают к вычислениям парного (полного) коэффициента корреляции между ними

$$r_{xy} = \frac{\sum_{i=1}^{i=n}(X_i - \bar{X})(Y_i - \bar{Y})}{n\,\sigma(X)\sigma(Y)}. \qquad (25)$$

Здесь X_i и Y_i – текущие значения признаков и их средние арифметические значения $(\bar{X}, \bar{Y})$; $\sigma(X), \sigma(Y)$ – средние квадратические отклонения, n – объем выборки. При этом $-1{,}0 \le r_{xy} \le 1{,}0$.

При $|r_{xy}|=1{,}0$ поведение факториального признака X полностью объясняет поведение результативного признака Y, и в этом случае можно говорить о наличии функциональной (жесткой) связи между ними. При $|r_{xy}|<1{,}0$, связи между указанными признаками оцениваются уже как стохастические (вероятностные, случайные). В этом случае можно говорить уже о том, что вариации результативного признака Y определяются действием не только признака X, но и других (Z, K, …).

Чтобы учесть множественные связи между вариациями признаков X, Z, K и др. и результативным Y прибегают к вычислениям множественного (совокупного) коэффициента корреляции [8] $R_{y.x,z,к}$. При этом $0 \le R_{y.x,z,к} \le 1{,}0$.

Здесь в подстрочных индексах при R первая буквенная переменная относится к результативному признаку (предиктанту), и она отделена от группы факториальных признаков точкой [3, 8].

Для упрощения всех последующих записей результативному признаку присвоим порядковый номер 1, а факториальным признакам X, Z, K – номера 2,3,4 и др. (соответственно).

При этом следует иметь в виду, что квадрат коэффициента множественной корреляции (КМК) полностью совпадает с ранее рассматривавшемся коэффициентом детерминации (22).

Расчет величины КМК предполагает:

- уяснение состава комплекса факториальных признаков, с действием которых связаны (априори) изменения результативного признака;

- формирование архивной выборки векторов, подобной (9);

- формирование определителя D (21) n-го порядка, его минора D_{11} и оценку их числовых значений.

Само значение КМК при этом определяется согласно равенству

$$R_{1.2,3,4,\dots n} = \sqrt{1 - \frac{D}{D_{11}}}. \quad (26)$$

При этом должны выдерживаться обязательные требование неравенства: $0 \leq R_{1.2,3,4,\dots n} \leq 1{,}0$.

Если коэффициент парной корреляции r_{12} (25) значим, и если при этом $R^2_{1.2,3} > r^2_{12}$, то присоединение к факториальному признаку 2 признака 3 влечет за собою некоторое увеличение разрешающей способности (увеличение коэффициента детерминации и уменьшение среднего квадрата погрешности $S_{1.2,3}$) линейной модели (19) с 2-мя предсказателями против линейной модели с 1-м предсказателем.

Если же окажется, что $r^2_{12} \geq R^2_{1.2,3}$, то в отношении признака 3 можно говорить о том, что он не содержит полезной информации относительно результативного признака 1 и, следовательно, не оказывает влияния на его поведение. Эти же рассуждения далее могут быть распространены и в отношении факториальных признаков 4,5 и последующих.

Таким образом, расчеты КМК могут быть полезны

- и как мера тесноты множественных связей между результативным (1) и комплексом факториальных (2, 3…n) признаков;

- и как инструмент для разыскания наиболее информативных и «отсеивания» малоинформативных (в отношении поведения предиктанта 1) факторов.

При вычислениях парных коэффициентов корреляции r_{12} (25) между признаками 1 и 2 не учитывается зависимость каждого из них от сложного действия на них всех варьирующих факторов, что в ряде случаев может сопровождаться значительным искажением (маскировкой) истинной связи между ними. В связи с этим нередко

возникает потребность вычислений частных коэффициентов корреляции

$$r_{1к.2,3,\dots n} = (-1)^{к}\frac{D_{1к}}{\sqrt{D_{11}D_{кк}}}, \qquad (27)$$

измеряющих направленности и тесноту связи между результативным (1) и факториальным признаком (к) при выключенном влиянии на него (постоянстве) всех других составляющих факториального комплекса.

При этом -1,0≤$r_{1к.2,3,\dots n}$≤1,0. Подобно тому, как r_{yx}^2 определяет величину относительного вклада варьирующего признака X в полную дисперсию σ^2(y), квадрат частного коэффициента корреляции (27) описывает «отфильтрованную» от искажающих влияний других факторов долю вклада к-го признака в полное многообразие поведения $[\sigma_1^2]$ того же результативного признака.

Таким образом, коэффициент детерминации полного комплекса факториальных признаков (18) можно записать в виде суммы

$$R_{1.2,3,\dots,n}^2 = r_{12.3,4,\dots,n}^2 + r_{13.2,4,\dots,n}^2 + \dots r_{1n.2,3,\dots,n-1}^2 + \delta_{1.2,3,\dots,n}. \qquad (28)$$

Последнее слагаемое в правой части отражает собою «остаточную» часть полного коэффициента детерминации, которая отражает ту часть поведения результативного признака, которая связана с влиянием на него интерактивных (межфакторных) взаимодействий. При $\delta_{1.2,3,\dots,n} < 0$ взаимодействия составляющих факторного комплекса влекут за собою некоторое уменьшение определенности поведения $[\sigma_1^2]$ результативного признака, и, наоборот, если $\delta_{1.2,3,\dots,n} > 0$. Очевидно, что величина последнего слагаемого в (28) может быть определена путем замыкания рассматриваемого равенства.

Задача. Из результатов исследований [4] закономерностей географического распределения по территории Приволжского федерального округа (ПФО) средних месячных температур воздуха в январе (результативный признак – 1) в их зависимости от географической широты (2), долготы (3) и высоты рельефа (4) вытекает следующий вид корреляционной матрицы (21) при полном учете комплекса указанных факторов

$$\|r_{ij}\| = \begin{pmatrix} r_{11}\ r_{12}\ r_{13}\ r_{14} \\ r_{21}\ r_{22}\ r_{23}\ r_{24} \\ r_{31}\ r_{32}\ r_{33}\ r_{34} \\ r_{41}\ r_{42}\ r_{43}\ r_{44} \end{pmatrix} = \begin{pmatrix} 1\ -0{,}516\ -0{,}469\ -0{,}337 \\ -0{,}516\ 1\ 0{,}120\ -0{,}067 \\ -0{,}469\ 0{,}120\ 1\ 0{,}178 \\ -0{,}337\ -0{,}067\ 0{,}178\ 1 \end{pmatrix}.$$

Необходимо оценить величины частных коэффициентов корреляции: $r_{12.3,4}$, $r_{13.2,4}$, $r_{14.2,3}$. Определить: - действие какого из трех независимых факторов (широты – 2, долготы – 3 или высоты – 4?) на результативный признак в наибольшей и в наименьшей мере замаскировано межфакторными связями?

Решение. Вначале, имея в виду указанную здесь матрицу $\|r_{ij}\|$, выполним оценку ряда ее определителей. Они таковы: D=0,450, D_{11}=0,947, D_{12}=-0,468, D_{13}=0,336, D_{14}=-0,290, D_{22}=0,691, D_{33}=0,592, D_{44}=0,557.

$R^2_{1.2,3,4} = 1 - \frac{D}{D_{11}}$=0,525,

$r_{12.3,4} = \frac{D_{12}}{\sqrt{D_{11}D_{22}}}$=-0,578,

$r_{13.2,4} = -\frac{D_{13}}{\sqrt{D_{11}D_{33}}}$=-0,449,

$r_{14.2,3} = \frac{D_{14}}{\sqrt{D_{11}D_{44}}}$=-0,400.

Из сравнений r_{12} с $r_{12.3,4}$, r_{13} с $r_{13.2,4}$ и r_{14} с $r_{14.2,3}$ следует, что из 3-х факторов, оказывающих свое влияние на изменения по территории ПФО температуры воздуха в январе, в наибольшей мере искажается (интерактивными связями) действие широты и высоты, и в наименьшей мере – действие долготы.

Из сравнений приведенных абсолютных величин коэффициентов частной корреляции следует, что в распределении температуры воздуха в январе по территории ПФО решающее значение принадлежит изменениям географической широты ($|r_{12.3,4}|$)> $|r_{13.2,4}|$ и ($|r_{12.3,4}| > |r_{14.2,3}|$)).

Задачи для самостоятельного решения

1. По тем же результатам исследований [4], базирующихся на обобщении многолетних (1955 – 2009 гг.) наблюдений за температурой воздуха в феврале на 215 станциях ПФО был определен вид корреляционной матрицы

$$\|r_{ij}\| = \begin{pmatrix} 1 & -0{,}310 & -0{,}495 & -0{,}334 \\ -0{,}310 & 1 & 0{,}120 & -0{,}067 \\ -0{,}495 & 0{,}120 & 1 & 0{,}178 \\ -0{,}334 & -0{,}067 & 0{,}178 & 1 \end{pmatrix},$$

в которой индексация взаимодействующих признаков осталась прежней (см. условия предыдущей типовой задачи).

Вычислить величины: 1) КМК, 2) коэффициента детерминации $R^2_{1.2,3,4}$ и относительную величину вклада в него (в долях единицы) изменений широты (2) места [$r^2_{12.3,4}/R^2_{1.2,3,4}$=?].

2. По данным тех же результатов исследований [4] определен вид корреляционной матрицы

$$\|r_{ij}\| = \begin{pmatrix} 1 & -0{,}795 & -0{,}155 & -0{,}415 \\ -0{,}795 & 1 & 0{,}120 & -0{,}067 \\ -0{,}155 & 0{,}120 & 1 & 0{,}178 \\ -0{,}415 & -0{,}067 & 0{,}178 & 1 \end{pmatrix},$$

полученной в результате изучения условий распределения по территории ПФО температуры воздуха в июле (результативный признак) 1). Здесь индексация факториальных признаков 2,3,4 – прежняя.

Оценить величины КМК, частных коэффициентов корреляции $r_{12.3,4}$, $r_{13.2,4}$, $r_{14.2,3}$ (27) и выполнить их сравнение с полными коэффициентами корреляции r_{12}, r_{13}, r_{14} (25).

3. В ходе изучения закономерностей распределения по территории ПФО месячных сумм осадков (1) [4] в зависимости от изменений географической широты (2), долготы (3) и высот орографии (4) корреляционная матрица взаимодействующих признаков определилась следующим образом:

$$\|r_{ij}\| = \begin{pmatrix} 1 & 0{,}791 & 0{,}045 & 0{,}225 \\ 0{,}791 & 1 & 0{,}120 & -0{,}067 \\ 0{,}045 & 0{,}120 & 1 & 0{,}178 \\ 0{,}225 & -0{,}067 & 0{,}178 & 1 \end{pmatrix}.$$

Определить КМК и коэффициент детерминации комплекса действующих факторов (2,3,4). Полученные в ходе решения указанные оценки сравнить с аналогичными показателями полученными в ходе решения задания 17. Какой из двух климатических показателей (температура воздуха или атмосферные осадки?) более чувствителен по отношению к изменениям координат географического пространства? Ответ на последний вопрос должен быть аргументирован сравнениями соответствующих коэффициентов детерминации.

4. Из результатов исследований [4] закономерностей пространственного распределения по территории ПФО температуры воздуха в апреле (результативный признак – 1) в зависимости от изменений географической широты (2), долготы (3) и высоты

орографии (4) вытекает следующий вид корреляционной матрицы взаимодействующих факторов

$$\|r_{ij}\| = \begin{pmatrix} 1 & -0{,}822 & -0{,}320 & -0{,}355 \\ -0{,}822 & 1 & 0{,}120 & -0{,}067 \\ -0{,}320 & 0{,}120 & 1 & 0{,}178 \\ -0{,}355 & -0{,}067 & 0{,}178 & 1 \end{pmatrix}.$$

Оценить величину КМК ($R_{1.2,3,4}$) и маскирующего действия на поведение результативного признака межфакторных (интерактивных) взаимодействий [$\delta_{1.2,3,4}$ =? – в долях единицы]. Выявить из совокупности факториальных признаков наиболее «мощный» и «слабый».

5. Из результатов исследований [4] закономерностей пространственного распределения по территории ПФО атмосферных осадков в апреле (результативный признак – 1) в зависимости от изменений географической широты (2), долготы (3) и высоты орографии (4) вытекает следующий вид корреляционной матрицы взаимодействующих факторов

$$\|r_{ij}\| = \begin{pmatrix} 1 & 0{,}442 & -0{,}008 & 0{,}242 \\ 0{,}442 & 1 & 0{,}120 & -0{,}067 \\ -0{,}008 & 0{,}120 & 1 & 0{,}178 \\ 0{,}242 & -0{,}067 & 0{,}178 & 1 \end{pmatrix}.$$

Выполнить оценки КМК, коэффициента детерминации ($R^2_{1.2,3,4}$ – в долях единицы), маскирующего влияния интерактивных межфакторных взаимодействий на распределение атмосферных осадков по территории [$\delta_{1.2,3,4}$ =?], относительных вкладов в полное значение коэффициента детерминации широты [$r^2_{12.3,4}/R^2_{1.2,3,4}$=?], долготы [$r^2_{13.2,4}/R^2_{1.2,3,4}$=?] и высоты места [$r^2_{14.2,3}/R^2_{1.2,3,4}$=?].

Список литературы

1. *Абубакиров Н.Р.* Элементы линейной алгебры / Н.Р. Абубакиров, В.А. Халямина. – Казань: Казанский госуд. ун-т, 2004. – 52 с.
2. *Багров Н.А.* Долгосрочные метеорологические прогнозы / Н.А. Багров, К.В. Кондратович, Д.А. Педь, А.И. Угрюмов. – Л.: Гидрометеоиздат, 1985. – 248 с.
3. *Борисенков Е.П.* Алгоритмы и программы статистической обработки информации на ЭВМ / Е.П. Борисенков, М.А. Романов. – Л.: Гидрометеоиздат, 1969. – 354 с.
4. *Важнова Н.А.* Влияние неоднородностей географической среды на распределение температуры воздуха в Приволжском федеральном округе (ПФО) / Н.А. Важнова // Вестник Удмуртского университета. Биология. Науки о Земле. – 2013. Вып. 3. – С. 91 – 99.
5. *Выгодский М.Я.* Справочник по высшей математике / М.Я. Выгодский. – М.: Физматгиз, 1962. – 870 с.
6. *Дроздов О.А.* Климатология / О.А. Дроздов, В.А. Васильев, Н.В. Кобышева, А.Н. Раевский, Л.К. Смекалова. – Л.: Гидрометеоиздат, 1989. – 567 с.
7. *Зверев А.С.* Синоптическая метеорология / А.С. Зверев. – Л.: Гидрометеоиздат, 1968. – 774 с.
8. *Пановский Г.А.* Статистические методы в метеорологии / Г.А. Пановский, Г.В. Брайер. – Л.: Гидрометеоиздат, 1972. – 209 с.
9. *Чичасов Г.Н.* Технология долгосрочных прогнозов погоды /Г.Н. Чичасов. – Спб.: Гидрометеоиздат, 1991. – 304 с.

Приложение 1

Средние месячные аномалии температуры воздуха (АТВ, ˚С)
на ст. Казань, университет (норма 1961 – 1990 гг.)

ГОД	1	2	3	4	5	6	7	8	9	10	11	12
1980	-0,35	0,34	-3,51	0,88	-0,01	0,00	-2,06	-3,75	-1,02	-0,27	-0,12	4,34
1981	5,55	3,34	-1,91	-2,72	-1,01	3,40	3,54	3,35	0,08	3,83	1,68	2,74
1982	0,05	-0,76	-0,51	0,18	-0,91	-3,30	0,44	-0,25	1,48	-0,57	2,68	5,24
1983	6,75	3,24	-0,91	4,68	-0,71	-2,30	0,04	-0,85	-0,12	1,63	0,58	4,14
1984	4,05	-1,96	0,89	-0,62	3,39	0,40	1,24	-2,15	1,08	0,23	-2,02	-5,66
1985	2,05	-2,76	-0,81	-1,62	-1,11	-1,60	-1,56	2,45	-0,32	-0,17	-0,72	-0,06
1986	1,65	-4,26	1,69	3,68	-2,21	1,00	-2,16	-0,65	-1,52	-0,37	-1,62	-2,56
1987	-6,85	3,24	-1,91	-3,32	1,99	3,50	-1,46	-0,85	-1,72	-1,57	-5,32	-1,96
1988	0,05	1,84	1,59	-0,02	0,29	3,30	3,74	1,45	-0,62	1,23	-4,52	-1,26
1989	3,05	4,84	1,19	-0,72	-0,51	4,50	0,94	-0,55	0,48	1,13	0,28	1,64
1990	1,95	7,64	4,49	2,48	-2,31	-1,50	-0,36	-0,75	-1,32	-0,37	0,28	2,64
1991	2,45	1,94	0,19	3,08	2,29	3,60	-0,16	-1,55	-0,12	4,03	1,38	-0,96
1992	3,55	1,04	1,99	0,78	-1,91	-0,30	-1,86	-0,45	3,28	-1,07	0,28	2,54
1993	5,35	0,34	-0,51	-0,02	0,69	0,00	0,34	-0,45	-4,02	0,03	-6,82	1,04
1994	4,95	-5,16	-1,01	0,48	-1,61	-1,30	-2,96	-1,55	2,48	1,53	-0,62	-0,66
1995	2,55	7,34	3,49	6,58	2,79	3,70	0,14	0,35	2,38	2,63	-0,12	-3,06
1996	-1,75	-1,36	-1,31	-2,32	2,19	0,50	0,84	-0,25	-0,72	0,33	4,28	-1,66
1997	-2,55	1,94	1,79	-0,02	-1,51	2,30	-1,36	-0,85	-0,72	1,23	-1,62	-1,86
1998	2,45	-3,16	2,49	-4,02	0,59	4,10	1,84	-0,15	-0,22	1,43	-5,42	2,44
1999	3,65	3,74	-1,11	2,68	-4,21	2,20	2,14	-0,45	-0,52	3,23	-4,92	3,74
2000	5,05	5,34	1,89	4,08	-3,81	1,20	2,64	0,45	-0,72	1,23	-0,92	0,74
2001	7,65	1,94	1,89	4,48	0,59	-0,50	2,34	-0,15	1,18	-0,07	1,88	-3,26
2002	5,15	8,64	4,89	0,48	-3,21	-0,70	3,14	-2,05	1,18	0,23	0,98	-8,86
2003	2,75	-0,06	-0,31	0,88	0,89	-3,50	1,34	2,35	0,98	2,83	2,48	5,34
2004	4,15	0,74	4,19	-2,32	0,89	-0,50	1,84	2,45	-4,55	-13,05	-18,95	-24,25
2005	5,55	-0,86	-3,21	0,98	3,99	-0,10	0,24	1,15	-3,65	-10,75	-16,75	-22,65
2006	-3,25	-2,96	0,59	0,48	0,59	4,20	-1,46	1,55	-3,45	-12,45	-20,15	-19,75
2007	9,65	-3,16	3,59	1,18	2,79	-0,80	0,54	5,15	-4,55	-10,45	-21,55	-27,75
2008	1,25	4,84	5,19	3,48	-0,81	-0,60	1,44	2,35	-7,05	-9,05	-14,95	-22,15
2009	2,25	2,14	2,39	-1,22	1,29	3,20	0,64	0,55	-1,75	-10,55	-18,35	-27,65
2010	-4,25	-2,36	0,39	2,48	4,49	4,40	6,44	5,45	-3,55	-13,45	-16,15	-26,25