УДК 517.2+519.2

МАТЕМАТИЧЕСКИЕ МОДЕЛИ МНОГОМЕРНЫХ ДАННЫХ

В.С. МУХА

Белорусский государственный университет информатики и радиоэлектроники
П. Бровки, 6, Минск, 220013, Беларусь

Поступила в редакцию 10 января 2014

Дан краткий обзор современного состояния проблемы анализа многомерных данных. Рассмотрены новые математические модели многомерных данных, разработанные в БГУИР.

Ключевые слова: многомерно-матричный анализ, ортогональные полиномы многих переменных, многоиндексные задачи линейного программирования, OLAP-системы, PARAFAC.

Проблема анализа многомерных данных

Проблема анализа многомерных данных является актуальной, поскольку большинство технических, экономических и других систем и процессов характеризуется не одной, а многими переменными (факторами). Эта проблема возникает в таких областях, как хемометрика, метеорология, охрана окружающей среды, обработка изображений, векторных случайных процессов и случайных полей, построение многомерных баз данных, построение моделей бизнес-процессов и многих других. Развитие подходов к анализу многомерных данных привело к целесообразности представления многомерных данных в виде многомерных матриц (multidimensional matrix). В последнее время наблюдается растущий интерес к данной проблеме во всем мире. В зарубежной литературе многомерные данные известны как многоканальные данные или многоканальные матрицы (multiway data, multiway matrix). Публикации, посвященные многоканальным данным, стали появляться, начиная с 2000 г., преимущественно на страницах журналов «Chemometrics and Intelligent Laboratory Systems» и «Journal of Chemometrics». Однако за рубежом до настоящего времени не разработаны сколько-нибудь эффективные теоретические подходы к анализу многомерных данных. Исследователи-прикладники наряду с формулировками задач разрабатывали и использовали собственные подходы к их решению, которые трудно систематизировать. Приходится также констатировать отсутствие осведомленности зарубежных авторов о всех достижениях в этой области, в частности, об основополагающей работе Соколова Н.П. по теории многомерных матриц.

Исследования в области анализа многомерных данных активно ведутся в БГУИР. Первые результаты в этой области, полученные в БГУИР, были опубликованы в 1987 г. в работе [1]. Совокупность полученных в БГУИР результатов в области анализа многомерных данных по состоянию на 2004 г., представляющих собой оригинальный многомерно-матричный математический аппарат для анализа многомерных данных, представлена в монографии [12]. В монографии обобщены и дополнены результаты работ [2–11].

Области применения разработанного многомерно-матричного математического аппарата постоянно расширяются. В данной статье приводится обзор новых результатов, полученных после выхода из печати работы [12]. В обзоре дается краткая характеристика актуальности решаемых задач, приводятся их постановки и результаты исследований, а с деталями рекомендуется ознакомиться по имеющимся публикациям. Используются обозначения работы [12].

Ортогональные полиномы векторной переменной

Важное место в анализе многомерных данных занимает аппроксимация многомерных данных рядами. Простейшей является аппроксимация рядом Тейлора, более совершенной – рядами по ортогональным полиномам. В работе [12] представлена многомерно-матричная теория дифференцирования и рядов Тейлора для многомерно-матричных функций многомерно-матричного аргумента, систем полиномов векторного аргумента x, ортогональных в R^n с гауссовским весом (полиномов Эрмита), на основе обобщенной формулы Родрига, а также общая теории систем полиномов векторной переменной, ортогональных с произвольным непрерывным весом, и построены ряды Фурье для скалярных функций векторной переменной и Грама–Шарлье для непрерывных векторных вероятностных распределений. Аналогичным образом может быть построена общая теория систем полиномов векторной переменной, ортогональных в некотором пространстве с дискретным весом [13].

Если Ω – некоторая область пространства R^n, l различным ее точкам $x_1, x_2, ..., x_l \in R^n$ приписаны в качестве весов положительные числа p_i, $i = \overline{1,l}$, и мера μ этой области определена формулой $\mu(\Omega) = \sum_{x_k \subseteq \Omega} p_k$, то такую меру будем называть вырожденной или дискретной. В случае дискретной меры интеграл от скалярной функции $f(x)$ по области Ω и мере $\mu(\Omega)$ определяется выражением $\int_{\Omega} f(x) d\mu = \sum_{x_k \subseteq \Omega} f(x_k) p_k$. Функцию $f(x)$ будем называть функцией с интегрируемым квадратом на Ω, если интеграл $\int_{\Omega} f^2(x) d\mu = \sum_{x_k \subseteq \Omega} f^2(x_k) p_k$ существует. Совокупность всех таких функций назовем пространством $L_2(\Omega, \mu)$.

Начальный многомерно-матричный момент i-го порядка ν_i дискретной весовой функции p_k, $k = \overline{1,l}$, определим выражением:

$$\nu_i = \int_{\Omega} {}^{0,0}x^i d\mu = \sum_{x_k \subseteq \Omega} {}^{0,0}(x_k)^i p_k, \; i = 0,1,2,...$$

Будем считать, что имеется l попарно различных точек области $\Omega \subseteq R^n$, не лежащих на гиперповерхности m-го порядка, причем $l \geq (m+n)!/(m!n!)$.

Многомерно-матричный полином $Q_m(x)$ степени m векторной переменной $x \in \Omega$ определим следующим образом:

$$Q_m(x) = \sum_{k=0}^{m} {}^{0,k}(C^*_{(m,k)} {}^{0,0}x^k) = \sum_{k=0}^{m} {}^{0,k}({}^{0,0}x^k C^*_{(k,m)}), \; m = 0,1,2,...,$$

где $C^*_{(m,k)} = (c^*_{i_1,...i_m, j_1,...j_k})$, $k = \overline{0,m}$, – $(m+k)$-мерные матрицы коэффициентов, симметричные относительно индексов двух своих мультииндексов $(i_1,..., i_m)$, $(j_1,..., j_k)$ и удовлетворяющие условиям $C^*_{(m,k)} = (C^*_{(k,m)})^{H_{m+k,k}}$, $C^*_{(k,m)} = (C^*_{(m,k)})^{B_{k+m,k}}$.

Последовательность многомерно-матричных полиномов $Q_m(x)$ назовем ортогональной в $L_2(\Omega, \mu)$, если выполняются условия

$$\sum_{k=1}^{l} {}^{0,0}(Q_m(x_k) Q_r(x_k)) p_k \begin{cases} = 0, & r = 0,1,..., m-1, \\ \neq 0 & r = m. \end{cases} \quad (1)$$

Основным многомерно-матричным полиномом степени m в $L_2(\Omega,\mu)$ будем называть полином вида

$$P_m(x)={}^{0,0}x^m+\sum_{k=0}^{m-1}{}^{0,k}(C_{(m,k)}\,{}^{0,0}x^k)=x^m+\sum_{k=0}^{m-1}{}^{0,k}({}^{0,0}x^kC_{(k,m)}),\ m=0,1,2,\ldots.$$

Основной многомерно-матричный полином $P_m(x)$ назовем ортогональным полиномом степени m в $L_2(\Omega,\mu)$, если он ортогонален к однородным полиномам $1, x, {}^{0,0}x^2,\ldots, {}^{0,0}x^{m-1}$:

$$\sum_{k=1}^{l}{}^{0,0}(P_m(x_k)\,{}^{0,0}(x_k)^r)p_k\begin{cases}=0, & r=\overline{0,m-1},\\ \neq 0, & r=m.\end{cases} \qquad (2)$$

Последовательности многомерно-матричных полиномов $P_m(x)$ и $Q_m(x)$ будем называть вполне биортонормальными в $L_2(\Omega,\mu)$, если выполняются соотношения (1), (2) и соотношение

$$\sum_{k=1}^{l}{}^{0,0}(Q_m(x_k)P_r(x_k))p_k=\begin{cases}0, & r=\overline{0,m-1},\\ D_{(m,m)}, & r=m,\end{cases}$$

где $D_{(m,m)}$ – $2m$-мерная матрица определенной структуры [12].

Теорема. Если последовательности многомерно-матричных полиномов $P_m(x)$ и $Q_m(x)$ вполне биортонормальны в $L_2(\Omega,\mu)$, то коэффициенты $C_{(m,k)}$ основной последовательности $P_m(x)$ определяются как решение многомерно-матричной системы линейных алгебраических уравнений

$$\nu_{m+p}+\sum_{k=0}^{m-1}{}^{0,k}(C_{(m,k)}\nu_{k+p})=0,\ m=0,1,\ldots,\ p=\overline{0,m-1},$$

а сопряженная последовательность $Q_m(x)$ может быть получена с помощью соотношения

$$Q_m(x)={}^{0,m}(A_{(m,m)}P_m(x))={}^{0,m}(P_m(x)A_{(m,m)}),$$

где

$$A_{(m,m)}=m!\,{}^{0,m}B^{-1}_{(m,m)}$$

$$B_{(m,m)}=\nu_{2m}+\sum_{r=0}^{m-1}{}^{0,r}(C_{(m,r)}\nu_{r+m})+\sum_{r=0}^{m-1}{}^{0,r}(\nu_{m+r}C_{(r,m)})+\sum_{r=0}^{m-1}\sum_{q=0}^{m-1}{}^{0,r}(C_{(m,r)}\,{}^{0,q}(\nu_{r+q}C_{(q,m)})).$$

Скалярная функция $v(x)$ векторной переменной $x\in R^n$ может быть представлена суммой многомерно-матричных ортогональных полиномов $Q_r(x)$:

$$v(x)\sim\sum_{r=0}^{\lambda}\frac{1}{r!}\,{}^{0,r}\left(B_rQ_r(x)\right),\ \text{где}\ B_r=\int_{\Omega}v(x)P_r(x)d\mu=\sum_{k=1}^{l}v(x_k)P_r(x_k)p_k\ .$$

Эта аппроксимация является аналогом ряда Фурье для скалярной функции $v(x)$ по многомерно-матричным полиномам, ортогональным с непрерывным весом [12].

Аппроксимация произвольного дискретного распределения $f_1, f_2,\ldots, f_l$ в Ω суммой многомерно-матричных ортогональных полиномов $Q_m(x_k)$ имеет вид:

$$f_k\sim p_k\sum_{m=0}^{\lambda}\frac{1}{m!}\,{}^{0,m}\left(D_mQ_m(x_k)\right),\ k=\overline{1,l},\ \text{где}\ D_m=\int_{\Omega}P_m(x)f(x)dx=\sum_{k=1}^{l}P_m(x_k)f_k\ .$$

Данная аппроксимация является аналогом ряда Грама–Шарлье для непрерывной весовой функции $f(x)$ по многомерно-матричным полиномам, ортогональным с некоторым непрерывным весом $\rho(x)$ [12].

Векторные многосвязные цепи Маркова

Цепи Маркова как математические модели дискретных процессов находят широкое применение в различных предметных областях. При этом в основном рассматриваются скалярные односвязные цепи Маркова. Обобщение на векторную односвязную цепь Маркова выполнено в работе [14]. Наиболее общий случай векторной многосвязной однородной цепи Маркова рассмотрен в [15].

Пусть $\bar{\xi}(t)=(\xi_i(t))$, $i=\overline{1,q}$, $t=0,1,2,...$, – дискретная векторная случайная последовательность, каждая компонента $\xi_i(t)$ которой может принимать k_i значений. Эта последовательность называется r-связной цепью Маркова, если условное распределение $\bar{\xi}(s)$ при известных значениях траектории во все моменты времени, предшествующие s, совпадает с условным распределением $\bar{\xi}(s)$ при известных значениях $\bar{\xi}(s-1)$, $\bar{\xi}(s-2)$,…, $\bar{\xi}(s-r)$, где r – фиксированное натуральное число.

Многосвязная (r-связная) цепь Маркова полностью определяется $(r+1)q$-мерной матрицей вероятностей перехода за один шаг

$$P=P(\bar{\xi}(s)=\bar{e}_{\nu}/\bar{\xi}(s-1)=\bar{e}_{i_r},\bar{\xi}(s-2)=\bar{e}_{i_{r-1}},...,\bar{\xi}(s-r)=\bar{e}_{i_1})=(p_{i_1,i_2,...i_r,\nu}),\ s\ge r,$$

где $i_1,i_2,..,i_r,\nu$ – q-мультииндексы, и rq-мерной матрицей безусловных вероятностей состояний

$$\Pi=P(\bar{\xi}(r-1)=\bar{e}_{i_r},\bar{\xi}(r-2)=\bar{e}_{i_{r-1}},...,\bar{\xi}(0)=\bar{e}_{i_1})=(\pi_{i_1,i_2,...i_r})$$

на тактах $s=0,1,...,r-1$.

Анализ цепи Маркова состоит в определении матрицы безусловных вероятностей

$$A(s)=P(\xi_1(s)=e_{1,i_1},\xi_2(s)=e_{2,i_2},...,\xi_q(s)=e_{q,i_q})=(a_i(s)),\ i=(i_1,i_2,...,i_q),\ i_\alpha=\overline{1,k_\alpha},\ \alpha=\overline{1,q},$$

на любом такте $s=0,1,2,...$.

Для тактов $s=0,...,r-1$ имеем $A(s)=(\sum_{i_1,...i_s,i_{s+2},...i_r}\pi_{i_1,i_2,...i_r})$. Для расчета $A(s)$ при $s\ge r$ рассматривается $(r+1)q$-мерная матрица вероятностей перехода за n шагов

$$P(n)=P(\bar{\xi}(s)=\bar{e}_{\nu}/\bar{\xi}(s-n)=\bar{e}_{i_r},\bar{\xi}(s-n-1)=\bar{e}_{i_{r-1}},...,\bar{\xi}(s-n-r)=\bar{e}_{i_1})=(p_{i_1,i_2,...i_r,\nu}(n)),$$
$$s\ge n+r.$$

Можно показать, что матрица $P(n)$ при $n>1$ удовлетворяет рекуррентному соотношению

$$P(n)=^{(r-1)q,q}(PP^{T_r}(n-1),\text{ где } T_r=(B_{rq,q},E_q),$$

а матрица безусловных вероятностей состояний r-связной векторной однородной цепи Маркова на любом такте $n\ge r$ определяется выражением

$$A(n)=^{0,rq}(\Pi P(n)).$$

Параллельный факторный анализ

Параллельный факторный анализ (PARAFAC) активно развивается в хемометрике. PARAFAC модель, известная в зарубежной литературе, имеет следующий вид:

$$G_a = (g_{a,i,j,k}) = (\sum_{f=1}^{n_f} a_{i,f} b_{j,f} c_{k,f}) + (v_{i,j,k}),\ i = \overline{1,n_a}\ \ j = \overline{1,n_b},\ k = \overline{1,n_c}, \qquad (3)$$

где $V = (v_{i,j,k})$ – матрица шума. PARAFAC задача состоит в том, чтобы по имеющейся трехмерной матрице измерений G_a получить оценки матриц $A = (a_{i,f})$, $B = (b_{j,f})$, $C = (c_{k,f})$.

Для решения сформулированной задачи используется так называемый альтернирующий метод наименьших квадратов (alternating least squares, ALS). Этот метод представлен в зарубежных литературных источниках в виде PARAFAC алгоритмов различной формы. Степень формализованности данных алгоритмов представляется не достаточно полной для их безошибочной реализации. В связи с этим в работе [16] PARAFAC рассмотрен с позиций многомерно-матричного подхода. Это позволило разработать PARAFAC алгоритм, обладающий большей степенью формализованности по сравнению с известными, а также получить некоторые новые результаты относительно PARAFAC модели.

Модель вида (3) возникает в задаче флуоресценции растворов веществ, когда растворы (образцы) содержат n_f веществ и анализируются с помощью спектрофлюорометра. Тогда $b_{j,f}$ интерпретируется как относительная эмиссия света единицей вещества f на длине волны излучения j, $c_{k,f}$ – как относительное поглощение света единицей вещества f на длине волны возбуждения k и $a_{i,f}$ означает концентрацию вещества f в i-м растворе. Элемент $g_{a,i,j,k}$ матрицы G_a означает интенсивность эмиссии света раствором i на длине волны j при его возбуждении на длине волны k.

Анализ PARAFAC модели (3) и испытания PARAFAC алгоритма в [16] показал, что невозможно получить однозначные значения матриц A, B, C модели (3) ввиду неоднозначности модели в указанной выше постановке задачи. В связи с этим PARAFAC в его существующем виде не может быть использован для полной идентификации модели (3). В приведенной постановке он может быть использован лишь для качественного анализа растворов, например, для определения числа веществ в растворах.

Более реалистичной и практически важной представляется задача получения оценки матрицы концентраций A по имеющимся матрицам G_a, B, C. Матрицы B, C могут быть известны, так как можно создать банк спектров различных веществ. В таком виде PARAFAC задача является регрессионной с неизвестным матричным параметром A. Решение этой задачи получено в работе [17].

Представим модель (3) в многомерно-матричной форме. Для этого сформируем трехмерную матрицу $D_a = (b_{j,f} c_{k,f}) = (d_{a,f,j,k})$, $f = \overline{1,n_f}$, $j = \overline{1,n_b}$, $k = \overline{1,n_c}$. Тогда модель (3) принимает вид

$$G_a = {}^{0,1}(AD_a) + V.$$

Оценка матрицы A этой модели определяется как решение оптимизационной задачи $F = {}^{0,3}(G_a - {}^{0,1}(AD_a))^2 \to \min\limits_{A}$ и имеет вид $\hat{A} = {}^{0,1}({}^{0,2}(G_a D_a^{\mathrm{T}})\,{}^{0,2}(D_a D_a^{\mathrm{T}})^{-1})$.

Дальнейший анализ многомерно-матричной PARAFAC модели может быть выполнен стандартными для регрессионного анализа приемами: проверка значимости параметров (концентраций) и отсеивание незначимых, проверка адекватности модели и ее замена в случае неадекватности путем включение в нее новых веществ.

Многоиндексные задачи линейного программирования

В исследовании операций известны так называемые многоиндексные задачи линейного программирования. К ним относятся задача об оптимальном использовании оборудования, транспортная задача, задача назначения, а также многие другие. С помощью многомерно-матричного подхода оказывается возможным разработать алгоритмы решения подобных

задач стандартным симплекс-методом с применением широко распространенных его программных реализаций. В частности, в работе [18] разработан такой алгоритм для транспортной задачи.

Транспортная задача формулируется следующим образом. Имеется m поставщиков некоторой продукции и n потребителей этой продукции. Требуется минимизировать общие затраты на перевозки

$$f = \sum_{i=1}^{m}\sum_{j=1}^{n}\alpha_{i,j}u_{i,j} \to \min_{u_{i,j}} \qquad (4)$$

при ограничениях

$$\sum_{i=1}^{m}u_{i,j} = b_j,\ j=\overline{1,n}, \qquad (5)$$

$$\sum_{j=1}^{n}u_{i,j} = s_i,\ i=\overline{1,m}, \qquad (6)$$

$$u_{i,j} \ge 0,$$

где $u_{i,j}$ – количество груза, перевозимого из пункта отправления i в пункт назначения j, $c_{i,j}$ – затраты на перевозку единицы груза из пункта i в пункт j, b_j – спрос на груз j-го потребителя, s_i – количество груза, имеющееся у i-го поставщика. Обычно рассматривается сбалансированная транспортная задача, когда $s_1 + s_2 + ... + s_m = b_1 + b_2 + ... + b_n$.

Необходимым шагом в разработке алгоритма решения транспортной задачи стандартным симплекс-методом является ее многомерно-матричная формализация. Для этого рассмотрим матрицы $\alpha = (\alpha_{i,j})$, $u = (u_{i,j})$, $i=\overline{1,m}$, $j=\overline{1,n}$. Тогда задачу (4) можно сформулировать в виде

$$f = {}^{0,2}(\alpha u) \to \min_{u}.$$

Для многомерно-матричной записи ограничения (5) сформируем трехмерную матрицу $c = (c_{k,i,j})$, $k=\overline{1,n}$, $i=\overline{1,m}$, $j=\overline{1,n}$, элементы которой определим формулой

$$c_{k,i,j} = \begin{cases} 1, & k = j, \\ 0, & k \ne j, \end{cases}$$

и вектор $b = (b_j)$, $j=\overline{1,n}$. В этих обозначениях ограничение (5) запишется в виде ${}^{0,2}(cu) = b$.

Для многомерно-матричной записи ограничения (6) сформируем трехмерную матрицу

$d = (d_{k,i,j})$, $k=\overline{1,m}$, $i=\overline{1,m}$, $j=\overline{1,n}$,

элементы которой определим формулой

$$d_{k,i,j} = \begin{cases} 1, & k = i, \\ 0, & k \ne i, \end{cases}$$

и вектор $s = (s_i)$, $i=\overline{1,m}$. Тогда условие (6) будет иметь вид ${}^{0,2}(d u) = s$.

В итоге транспортная задача (4)–(6) получит следующую стандартную векторно-матричную формулировку:

$${}^{0,1}(\tilde{\alpha}_{(0,0,2)}\tilde{u}_{(2,0,0)}) \to \min_{\tilde{u}_{(2,0,0)}}, \qquad (7)$$

$${}^{0,1}(\tilde{c}_{(1,0,2)}\tilde{u}_{(2,0,0)}) = b, \qquad (8)$$

$${}^{0,1}(\tilde{d}_{(1,0,2)}\tilde{u}_{(2,0,0)}) = s, \qquad (9)$$

где символом ~ обозначены матрицы, ассоциированные с соответствующими многомерными матрицами. Задача (7)–(9) может быть решена, например, с помощью функции linprog.m системы технических расчетов Matlab.

Задача учебного расписания

По проблеме компьютерного составления расписания занятий учебного заведения имеется обширная литература, в которой разрабатываются различные методы ее решения: генетические, муравьиные, графовые, эвристические и другие. Применение того или иного метода обусловлено особенностями расписаний различных учебных заведений, различиями в требованиях к расписанию, целевым функциям, а порой просто приоритетами авторов. Достаточно распространенным является подход, рассматривающий задачу расписания как задачу линейного программирования. Но даже при использовании этого подхода постановки задач различных авторов различны. Обилие публикаций, в том числе и свежих, свидетельствует о том, что задача расписания не получила полного разрешения. В связи с этим в работе [19] предложено задачу расписания рассматривать как задачу назначения с дополнительными ограничениями, и для ее решения использовать многомерно-матричный подход. Рассмотрена задача с двумя субъектами расписания – студенческими группами и аудиториями – при условии, что число студентов в группе не должно превосходить вместимости назначенной ей аудитории. В работе [20] этот подход обобщен на задачу расписания с произвольным числом субъектов.

Конструктивно расписание можно определить как p-мерную гиперпрямоугольную матрицу $u = (u_{i_1,\dots i_p})$, $i_\beta = \overline{1, n_\beta}$, с элементами, принимающими значения 0 или 1. Для хранения элемента расписания достаточно соответствующий элемент матрицы расписания положить равным единице.

Составлением расписания будем называть назначение элементов матрицы расписания u, удовлетворяющих определенным ограничениям. Оптимизацией расписания будем называть составление расписания, доставляющего минимум некоторому критерию.

Можно предложить следующий минимально необходимый упорядоченный набор координатных осей пространства расписания: преподаватель (Иванов В.В., Бокун М.П.,…), дисциплина (АМД, ООПиП,…), вид занятия (лекция, лабораторное занятие,…), группа (420601, 620601,…), аудитория (601-5, 602-5,…), день (понедельник, вторник,…), время (пара) (8.00–9.35, 9.45–11.20,…), неделя (1-я, 2-я, 3-я, 4-я).

Предположим, что каждый элемент расписания $u_{i_1,\dots i_p}$ имеет определенную стоимость $\alpha_{i_1,\dots i_p}$, и определим задачу расписания как оптимизационную:

$$f = \sum_{i_1=1}^{n_1} \cdots \sum_{i_p=1}^{n_p} \alpha_{i_1,\dots i_p} u_{i_1,\dots i_p} \to \min_{u_{i_1,\dots i_p}}$$

с ограничениями следующих типов:

1) типа «назначить», жестко закрепляющее некоторые элементы расписания;

2) типа «не назначать», исключающее некоторые элементы расписания;

3) определяющими не более одной комбинации «группа–аудитория» для преподавателя в фиксированное время «неделя–день–пара»;

4) определяющими не более одного преподавателя для группы в фиксированное время «неделя–день–пара»;

5) определяющими не более одного преподавателя для аудитории в фиксированное время «неделя–день–пара».

6) определяющими не более одной группы для аудитории в фиксированное время «неделя–день–пара».

7) определяющими не более одной аудитории для группы в фиксированное время «неделя–день–пара».

8) определяющими, что вместимость аудитории не меньше числа студентов в направляемой в нее группе.

Сформулированная задача расписания является многоиндексной бинарной целочисленной задачей линейного программирования. Эта задача в литературе не рассматривалась, и, как следствие, не разработан алгоритм ее решения.

Методика решения данной задачи состоит в ее приведении к виду классической векторно-матричной задачи линейного программирования. Реализация методики заключается в многомерно-матричной формализации задачи с последующим переходом к векторно-матричной формулировке в терминах ассоциированных матриц. В результате оказывается возможным получение классической векторно-матричной задачи линейного программирования, которая может быть решена стандартным симплекс-методом. Следует отметить, что если многомерно-матричная формализация целевой функции f достаточно очевидна, то формализация ограничений требует определенного искусства. Данная методика детально описана и реализована программно в работе [21].

Многомерно-матричные оптимальные статистические решения

В настоящее время для решения многих прикладных статистических задач весьма плодотворным представляется подход с позиций теории статистических решений, или, иначе, байесовский подход – благодаря его оптимизационному характеру, строгой обоснованности, использованию полной априорной информации. Наиболее известным является применение этого подхода в задаче получения точечных оценок параметров распределений, которые, вообще говоря, не являются случайными. В гораздо меньшей степени он применяется в задачах оценивания (прогнозирования) случайных состояний стохастических систем. Между тем решение именно таких задач является конечной целью многих исследований. В данной статье формулируется и решается задача принятия решения о многомерной случайной матрице по наблюдению другой многомерной случайной матрицы. В литературе данная задача не формулировалась и не решалась. В работе [21] рассматривались отдельные ее аспекты, работа [22] дает достаточно целостное и полное о ней представление.

Состояние некоторой системы считается p-мерной случайной матрицей $\eta = (\eta_{j_1, j_2, \ldots j_p})$, $\eta \in S$, S – пространство состояний, и известна плотность вероятности состояния $f(\eta)$. Над системой выполняется эксперимент (наблюдение), в результате чего получаем q-мерную матрицу наблюдений $\xi = (\xi_{i_1, i_2, \ldots i_q})$, $\xi \in X$, X – пространство наблюдений. Связь ξ с η определяется условной плотностью вероятности $f(\xi / \eta)$, которая считается известной. Задача состоит в определении плотности вероятности (решающей функции) $f(\Delta(\xi))$, минимизирующей средний риск $r = E(W(\eta, \Delta))$, где $W(\eta, \Delta)$ – функция потерь, $\Delta = \Delta(\xi) \in D$ – прогнозирующая функция или предиктор, D – пространство решений.

Аналогично скалярному случаю [23] можно показать, что оптимальное решение является нерандомизированным, и оптимальный предиктор $\Delta = \Delta(\xi)$ определяется из условия

$$r(\Delta) = \int_S W(\eta, \Delta) f(\eta / \xi) d\xi \to \min_{\Delta}. \qquad (10)$$

Уравнение Эйлера для минимизации функционала (10) имеет вид $dr(\Delta)/d\Delta = 0$. При квадратичной функции потерь $W(\eta, \Delta) =^{0,p} (\eta - \Delta)^2$ оптимальный предиктор определяется как апостериорное среднее: $\Delta^* = E(\eta / \xi)$.

Апостериорное среднее как оптимальный предиктор может быть сравнительно легко получено для матриц ξ и η с совместным гауссовским распределением. В этом случае оптимальный предиктор является линейным по наблюдению. Он определяется следующей теоремой.

Теорема. Пусть $\xi = (\xi_{i_1,i_2,\dots i_q})$ – q-мерная случайная матрица, $\eta = (\eta_{j_1,j_2,\dots j_p})$ – p-мерная случайная матрица, ξ и η имеют совместное гауссовское распределение с математическими ожиданиями $A_\xi = E(\xi)$, $A_\eta = E(\eta)$, дисперсионными матрицами $D_\xi = E({}^{0,0}(\xi - A_\xi)^2)$, $D_\eta = E({}^{0,0}(\eta - A_\eta)^2)$ и взаимной ковариационной матрицей $R_{\xi,\eta} = E({}^{0,0}((\xi - A_\xi)(\eta - A_\eta)))$. Тогда условные распределения этих матриц также гауссовские с математическими ожиданиями

$A_{\xi/\eta} = A_\xi + {}^{0,p}({}^{0,p}(R_{\xi,\eta}\ {}^{0,p}D_\eta^{-1})(\eta - A_\eta))$, $A_{\eta/\xi} = A_\eta + {}^{0,q}({}^{0,q}(R_{\eta,\xi}\ {}^{0,q}D_\xi^{-1})(\xi - A_\xi))$,

и дисперсионными матрицами

$D_{\xi/\eta} = D_\xi - {}^{0,p}({}^{0,p}(R_{\xi,\eta}\ {}^{0,p}D_\eta^{-1})R_{\eta,\xi})$, $D_{\eta/\xi} = D_\eta - {}^{0,q}({}^{0,q}(R_{\eta,\xi}\ {}^{0,q}D_\xi^{-1})R_{\xi,\eta})$,

причем $R_{\eta,\xi} = R_{\xi,\eta}^T$, где $T = B_{q+p,q}$.

В случае не гауссовского совместного распределения матриц η и ξ нет оснований рассчитывать на получение линейного оптимального предиктора, как это было выше. В этом случае можно при постановке задачи постулировать полиномиальный характер предиктора и отыскивать коэффициенты этого предиктора из условия минимума среднего риска.

Пусть $f_{\xi,\eta}(x,y)$ – совместная плотность вероятности переменных ξ, η, $f_\xi(x)$, $f_\eta(y)$ – маргинальные плотности вероятностей, $f_{\eta/\xi}(y/x)$, $f_{\xi/\eta}(x/y)$ – условные плотности вероятностей. Пусть предиктор является полиномом m-й степени,

$$\eta = \Delta(\xi) = \sum_{k=0}^{m} {}^{0,kq}(C_{(p,kq)}\xi^k) = \sum_{k=0}^{m} {}^{0,kq}(\xi^k C_{(kq,p)}),\ m = 0,1,2,\dots, \qquad (11)$$

где $C_{(p,kq)}$ – $(p+kq)$-мерные матрицы коэффициентов,

$$C_{(p,kq)} = (c_{i_{(p)},\vec{j}_{(k)}}),\ i_{(p)} = (i_1,i_2,\dots,i_p),\ \vec{j}_{(k)} = (j_{(q),0}, j_{(q),1},\dots, j_{(q),k}),$$

симметричные относительно q-мультииндексов $j_{(q),0}$, $j_{(q),1}$, …, $j_{(q),k}$ и удовлетворяющие условиям

$C_{(p,kq)} = (C_{(kq,p)})^{H_{p+kq,kq}}$, $C_{(kq,p)} = (C_{(p,kq)})^{B_{p+kq,kq}}$.

Сформулируем задачу минимизации среднего риска

$$r = E(W(\eta,\Delta)) \to \min_{C_{(p,0)},C_{(p,q)},\dots C_{(p,mq)}}. \qquad (12)$$

Функцию вида (11), коэффициенты которой определяются как решение оптимизационной задачи (12), назовем наилучшим полиномиальным многомерно-матричным предиктором.

Можно показать, что при квадратичной функции потерь $W(\eta,\Delta) = {}^{0,p}(\eta - \Delta)^2$ коэффициенты $C_{(p,0)}, C_{(p,q)}, \dots, C_{(p,mq)}$ оптимального полиномиального многомерно-матричного предиктора удовлетворяют следующей системе многомерно-матричных линейных алгебраических уравнений:

$$\sum_{k=0}^{m} {}^{0,kq}\left(C_{(p,kq)}\nu_{\xi^{k+\lambda}}\right) = \nu_{\eta\xi^\lambda},\ \lambda = \overline{0,m},$$

где $\nu_{\xi^{k+\lambda}} = E({}^{0,0}\xi^{k+\lambda})$, $\nu_{\eta\xi^\lambda} = E({}^{0,0}(\eta\ {}^{0,0}\xi^\lambda))$ – вероятностные моменты.

В работе [22] путем решения данной системы уравнений получены коэффициенты постоянного ($m=0$), линейного ($m=1$) и квадратичного ($m=2$) предикторов.

Полученные байесовские оптимальные теоретические предикторы путем замены теоретических моментов эмпирическими превращаются в эмпирические предикторы, полученные в работе [24] на основе критерия наименьших квадратов. Ввиду сходимости эмпирических моментов к теоретическим по вероятности мы сразу можем утверждать, что полученные таким путем эмпирические предикторы сходятся по вероятности к оптимальным в смысле минимума среднего риска теоретическим предикторам.

Конечномерные моменты стационарных случайных процессов и их оценки

Понятие семейства конечномерных распределений случайного процесса является основным в теории случайных процессов. В то же время понятие конечномерных моментов случайного процесса отсутствует в литературе. Между тем конечномерные моменты находят практическое применение. Например, конечномерные моменты первого и второго порядков используются в алгоритмах линейного статистического прогнозирования векторных случайных процессов. Традиционно эти моменты определяются не непосредственно, а с помощью функции математического ожидания и ковариационной функции случайного процесса. Возникает также потребность в использовании конечномерных моментов порядков выше второго. Так, в алгоритме квадратичного статистического прогнозирования векторной случайной последовательности [25] используются конечномерные моменты 3-го и 4-го порядков. Естественным представляется желание получить эти конечномерные моменты с помощью моментных функций 3-го и 4-го порядков. Однако на этом пути возникают непреодолимые трудности, связанные с громоздкостью и плохой формализуемостью условий симметрии моментных функций случайных процессов. Работа же с конечномерными моментами непосредственно по их определению снимает все препятствия на пути получения их оценок. Понятие конечномерных моментов многомерно-матричного случайного процесса введено в работе [26]. Там же получены их статистические оценки по отдельной реализации и по набору реализаций стационарного многомерно-матричного случайного процесса.

Многомерно-матричным (p-мерно-матричным) случайным процессом $\xi(\omega,t)$ назовем организованную в виде p-мерной матрицы совокупность действительных функций $\xi_{i_{(p)}}(\omega,t)$, $i_{(p)}=(i_1,i_2,...,i_p)$ – p-мультииндекс, $i_\alpha=\overline{1,n_\alpha}$, $\alpha=\overline{1,p}$, которые при каждом фиксированном значении переменной $t\in R_+$ являются измеримыми функциями $\omega\in\Omega$ на вероятностном пространстве $\{\Omega,F,P\}$.

Для p-мерно-матричного случайного процесса будем применять обозначение

$$\xi(t)=(\xi_{i_{(p)}}(t)),\ i_{(p)}=(i_1,i_2,...,i_p). \qquad (13)$$

Сечение p-мерно-матричного случайного процесса $\xi(t)$ в момент времени $t_j\in R_+$ представляет собой p-мерную случайную матрицу

$$\xi(t_j)=\xi_j=(\xi_{i_{(p)}}(t_j)),\ i_{(p)}=(i_1,i_2,...,i_p).$$

Зафиксировав s моментов времени $t_1,t_2,...,t_s$, получим s случайных p-мерных матриц (сечений) $\xi_1,\xi_2,...,\xi_s$.

Конечномерным (s-мерным) распределением p-мерно-матричного случайного процесса $\xi(t)$ (13) назовем совместное распределение его сечений $\xi_1,\xi_2,...,\xi_s$ (совокупностей случайных матриц) при любом s и любых $t_1,t_2,...,t_s$.

Начальной моментной функцией k-го порядка p-мерно-матричного случайного процесса $\xi(t)$ (13) назовем kp-мерную матрицу $\nu_\xi^{(k)}(t_1,t_2,...,t_k)$, являющуюся математическим ожиданием $(0,0)$-свернутого произведения его k сечений,

$$\nu_{\xi}^{(k)}(t_1,t_2,...,t_k)=E(^{0,0}(\xi(t_1)\cdots\xi(t_k)))=(\nu_{\xi,i_{(p)1},...i_{(p)k}}^{(k)}(t_1,t_2,...,t_k)), \quad (14)$$

где усреднение $E(\cdot)$ выполняется по k-мерному распределению случайного процесса.

Центральная моментная функция k-го порядка $\mu_{\xi}^{(k)}(t_1,t_2,...,t_k)$ p-мерно-матричного случайного процесса $\xi(t)$ определяется аналогично (14), но относительно центрированного случайного процесса $\overset{\circ}{\xi}(t)=\xi(t)-\nu_{\xi}^{(1)}(t)$.

Конечномерные моменты случайного процесса – это моменты конечномерных распределений случайного процесса. При определении конечномерных моментов рассматриваются не отдельные сечения случайного процесса, а наборы его сечений. Набор из s последовательных сечений случайного процесса $\xi(t)$ (13) в моменты времени $t_1,...,t_s$ будем называть s-набором и обозначать $\bar{\xi}_s(\bar{t}_s)$, где $\bar{t}_s=(t_1,...,t_s)$ – вектор. Для p-мерно-матричного случайного процесса $\xi(t)$ (13) набор из s его сечений образует $(p+1)$-мерную случайную матрицу, где дополнительный индекс пробегает значения от 1 до s. Ограничимся случаем двух наборов сечений случайного процесса $\xi(t)$ (13): s-набором $\bar{\xi}_s(\bar{t}_s)$,

$$\bar{\xi}_s(\bar{t}_s)=(\bar{\xi}_{i_{(p)},j})=(\xi_{i_{(p)}}(t_j)),\ j=\overline{1,s},\ \bar{t}_s=(t_1,...,t_s), \quad (15)$$

и r-набором $\bar{\eta}_r(\bar{u}_r)$,

$$\bar{\eta}_r(\bar{u}_r)=(\bar{\eta}_{i_{(p)},l})=(\xi_{i_{(p)}}(u_l)),\ l=\overline{1,r},\ \bar{u}_r=(u_1,...,u_r), \quad (16)$$

в которых $i_{(p)}=(i_1,i_2,...,i_p)$, $i_\alpha=\overline{1,n_\alpha}$, $\alpha=\overline{1,p}$.

Начальным конечномерным (s-мерным) моментом k-го порядка p-мерно-матричного случайного процесса $\xi(t)$ (13) называется $k(p+1)$-мерная матрица $\nabla_{\bar{\xi}_s}^{(k)}(\bar{t}_s)$, являющаяся математическим ожиданием k-й $(0,0)$-свернутой степени s-набора сечений $\bar{\xi}_s(\bar{t}_s)$ (15):

$$\nabla_{\bar{\xi}_s}^{(k)}(\bar{t}_s)=E(^{0,0}\bar{\xi}_s^k(\bar{t}_s))=(E(^{0,0}(\xi_{i_{(p)1}}(t_{j_1})\cdots\xi_{i_{(p)k}}(t_{j_k}))))=(\nabla_{\xi,i_{(p)1},j_1,...i_{(p)k},j_k}^{(k)}), \quad (17)$$

где $i_{(p)1},...,i_{(p)k}$ – p-мультииндексы, $j_1,...,j_k$ – индексы, принимающие значения $1,2,...,s$, и усреднение $E(\cdot)$ выполняется по s-мерному распределению случайного процесса.

Центральный s-мерный момент k-го порядка p-мерно-матричного случайного процесса $\xi(t)$ (13) определяется аналогично (17), но относительно центрированного s-набора сечений $\overset{\circ}{\bar{\xi}}_s(\bar{t}_s)=\bar{\xi}_s(\bar{t}_s)-\nabla_{\bar{\xi}_s}^{(1)}(\bar{t}_s)$.

Смешанным начальным $(s+r)$-мерным моментом $(k+q)$-го порядка p-мерно-матричного случайного процесса $\xi(t)$ (13) назовем $(k+q)(p+1)$-мерную матрицу $\nabla_{\bar{\xi}_s,\bar{\eta}_r}^{(k+q)}(\bar{t}_s,\bar{u}_r)$, являющуюся математическим ожиданием $(0,0)$-свернутого произведения k-й степени матрицы $\bar{\xi}_s(\bar{t}_s)$ (15) и q-й степени матрицы $\bar{\eta}_r(\bar{u}_r)$ (16):

$$\nabla_{\bar{\xi}_s,\bar{\eta}_r}^{(k+q)}(\bar{t}_s,\bar{u}_r)=E(^{0,0}(^{0,0}\bar{\xi}_s^k(\bar{t}_s)^{0,0}\bar{\eta}_r^q(\bar{u}_r)))=(\nabla_{\xi,\eta,i_{(p)1},j_1,...i_{(p)k},j_k,l_{(p)1},m_1,...l_{(p)q},m_q}^{(k+q)})=$$

$$=(E(^{0,0}(\xi_{i_{(p)1}}(t_{j_1})\cdots\xi_{i_{(p)k}}(t_{j_k})\xi_{l_{(p)1}}(t_{m_1})\cdots\xi_{l_{(p)k}}(t_{m_q})))), \quad (18)$$

где $j_1,...,j_k$, – индексы, принимающие значения $1,2,...,s$, $m_1,...,m_q$ – индексы, принимающие значения $1,2,...,r$, и усреднение $E(\cdot)$ выполняется по $(s+r)$-мерному распределению случайного процесса $\xi(t)$.

Смешанный центральный $(s+r)$-мерный момент $(k+q)$-го порядка p-мерно-матричного случайного процесса $\xi(t)$ (13) определяется аналогично (18), но относительно центрированных наборов сечений $\overset{\circ}{\xi}_s(\bar{t}_s)=\xi_s(\bar{t}_s)-\nu^{(1)}_{\xi_s}(\bar{t}_s)$ и $\overset{\circ}{\eta}_r(\bar{u}_r)=\eta_r(\bar{u}_r)-\nu^{(1)}_{\eta_r}(\bar{u}_r)$.

Преимуществом конечномерных моментов случайных процессов по сравнению с моментными функциями является то, что для стационарных случайных процессов они, как и конечномерные распределения, инвариантны к сдвигу по оси времени. Это преимущество позволяет сравнительно легко получать их оценки и, как следствие, использовать в практических приложениях.

Пусть известна реализация $x(jT)$, $j=\overline{1,n}$, T – интервал дискретизации, p-мерно-матричной стационарной случайной последовательности $\xi(jT)$, для которой существуют s-мерные начальные и центральные моменты k-го порядка $\nu^{(k)}_{\xi_s}(\bar{t}_s)$, $\mu^{(k)}_{\xi_s}(\bar{t}_s)$, и требуется по этой реализации найти оценки $\nu^{(k)}_{\xi_s}(\bar{t}_s)$, $\mu^{(k)}_{\xi_s}(\bar{t}_s)$ этих моментов. В работе [26] предложены оценки вида

$$\nu^{(k)}_{\xi_s}=\frac{1}{n-s+1}\sum_{j=1}^{n-s+1}(x_s(\overline{jT}))^k, \tag{19}$$

$$\mu^{(k)}_{\xi_s}=\frac{1}{n-s+1}\sum_{j=1}^{n-s+1}(\overset{*}{x}_s(\overline{jT}))^k, \tag{20}$$

где $x_s(\overline{jT})=\big(x(jT),...,x((j+s-1)T)\big)$ – набор из s сечений реализации с начальным сечением в момент времени jT, а $\overset{*}{x}_s(\overline{jT})=x_s(\overline{jT})-\nu^{(1)}_{\xi_s}$. Аналогичные оценки предложены для смешанных конечномерных моментов $\nu^{(k+q)}_{\xi_s,\eta_r}$, $\mu^{(k+q)}_{\xi_s,\eta_r}$ в случае примыкающих друг к другу наборов сечений ξ_s и η_r (15), (16):

$$\nu^{(k+q)}_{\xi_s,\eta_r}=\frac{1}{n-s-r+1}\sum_{j=1}^{n-s-r+1}(x_s(\overline{jT}))^k(x_r(\overline{(j+s)T}))^q, \tag{21}$$

$$\mu^{(k+q)}_{\xi_s,\eta_r}=\frac{1}{n-s-r+1}\sum_{j=1}^{n-s-r+1}(\overset{*}{x}_s(\overline{jT}))^k(\overset{*}{x}_r(\overline{(j+s)T}))^q, \tag{22}$$

где $\overset{*}{x}_s(\overline{jT})=x_s(\overline{jT})-\nu^{(1)}_{\xi_s}$, $\overset{*}{x}_r(\overline{(j+s)T})=x_r(\overline{(j+s)T})-\nu^{(1)}_{\xi_r}$.

На практике встречаются случаи, когда вместо одной достаточно длинной реализации стационарной случайной последовательности имеется m более коротких реализаций длиной l_i, $i=\overline{1,m}$, полученных в различные промежутки времени, и требуется по ним получить оценки конечномерных моментов случайной последовательности [27]. Для этого случая предлагаются следующие оценки:

$$\nu^{(k)}_{\xi_s}=\frac{1}{u-m(s-1)}\sum_{\nu=1}^{m}\sum_{j=1}^{l_\nu-s+1}(x_{\nu,s}(\overline{jT}))^k, \tag{23}$$

$$\mu^{(k)}_{\xi_s}=\frac{1}{u-m(s-1)}\sum_{\nu=1}^{m}\sum_{j=1}^{l_\nu-s+1}(\overset{*}{x}_{\nu,s}(\overline{jT}))^k, \tag{24}$$

$$\nu^{(k+q)}_{\xi_s,\eta_r}=\frac{1}{u-m(s+r-1)}\sum_{\nu=1}^{m}\sum_{j=1}^{l_\nu-s-r+1}(x_{\nu,s}(\overline{jT}))^k(x_{\nu,r}(\overline{(j+s)T}))^q, \tag{25}$$

$$\mu^{(k+q)}_{\xi_s,\eta_r}=\frac{1}{u-m(s+r-1)}\sum_{\nu=1}^{m}\sum_{j=1}^{l_\nu-s-r+1}(\overset{*}{x}_{\nu,s}(\overline{jT}))^k(\overset{*}{x}_{\nu,r}(\overline{(j+s)T}))^q, \tag{26}$$

где $u=\sum_{i=1}^{m} l_i$ – суммарная длина всех реализаций, $x_{\nu,s}(\overline{jT})=\left(x_\nu(jT),...,x_\nu((j+s-1)T)\right)$ – набор из s сечений ν-й реализации с начальным сечением в момент времени jT, $\overset{*}{x}_{\nu,s}(\overline{jT})=x_{\nu,s}(\overline{jT})-\nabla^{(1)}_{\xi_s}(\overline{jT})$.

В выражениях (19)–(26) все степени и произведения понимаются как $(0,0)$ - свернутые. При выполнении определенных условий эргодичности и $u\to\infty$ оценки (19)–(26) состоятельны.

Предложенные оценки были использованы без потери точности и вычислительной сложности по сравнению с классическими оценками при программной реализации алгоритма линейного статистического прогнозирования векторной случайной последовательности. Они позволили также выполнить программную реализацию алгоритма квадратичного статистического прогнозирования векторной случайной последовательности [25], что не представлялось возможным выполнить классическим подходом.

Многомерно-матричный метод главных компонент

Метод главных компонент находит применение для выявления скрытых закономерностей и экономизации алгоритмов обработки информации в хемометрике, биометрике, эконометрике, энвайрометрике и других областях знаний. В настоящее время этот метод разработан лишь для случайных векторов. Актуальным является его обобщение на случайные объекты более сложной структуры – обычные и многомерные случайные матрицы. Такое обобщение выполнено в работе [25], что позволило расширить области применения метода главных компонент.

Для обобщения метода главных компонент на случай многомерных матриц необходимо иметь решение задачи о собственных числах и собственных матрицах многомерной матрицы. Эта последняя задача получила дальнейшее развитие в работе [28] в направлениях применимости к гиперпрямоугольным матрицам, повышения уровня формализованности и компьютерной реализуемости.

Собственные числа $\alpha=\alpha_{(\mu,\lambda,0)}$ и фундаментальные собственные матрицы $Y=Y_{(\mu,\lambda,0)}$ многомерной матрицы $B_{(\mu,\lambda,\mu)}$ определяются как решения многомерно-матричного уравнения

$${}^{\lambda,\mu}((B_{(\mu,\lambda,\mu)}-\alpha_{(\mu,\lambda,0)}E_{(\mu,\lambda,\mu)})Y_{(\mu,\lambda,0)})=0.$$

$2p$-мерную матрицу $P=P_{(p,0,p)}$ назовем $(p,0,p)$-ортогональной, если ${}^{0,p}(PP^{B_{2p,p}})=E(0,p)$.

Теорема (вероятностный многомерно-матричный метод главных компонент). Если $\eta=(\eta_l)$, $l=(l_1,l_2,...,l_p)$ – p-мерная гиперпрямоугольная случайная матрица со средним значением $a_\eta=E(\eta)$, то ее дисперсионная матрица $R_\eta=E({}^{0,0}(\mathring{\eta}\mathring{\eta}))$, $\mathring{\eta}=\eta-a_\eta$, может быть представлена в виде

$$R_\eta={}^{0,p}({}^{0,p}(P\Lambda)P^{B_{2p,p}}),$$

где Λ – $2p$-мерная матрица, составленная из $(p,0,0)$-собственных чисел α_c матрицы R_η (элементов матрицы $\alpha=\alpha_{(p,0,0)}=(\alpha_c)$, $c=(c_1,c_2,...,c_p)$) по формуле

$$\Lambda=(\lambda_{c,c'})=\begin{pmatrix}\alpha_c, & c=c',\\ 0, & c\neq c',\end{pmatrix}$$

P – $2p$-мерная $(p,0,p)$-ортогональная матрица, составленная из нормированных фундаментальных $(p,0,0)$-собственных матриц $(y_c)_{c'}$ матрицы R_η по формуле

$$P = P_{(p,0,p)} = (p_{c,c'}) = ((y_c)_{c'}),\ c = (c_1, c_2, ..., c_p),\ c' = (c'_1, c'_2, ..., c'_p).$$

Случайная p-мерная матрица $\xi = (\xi_l)$, $l = (l_1, l_2, ..., l_p)$, получающаяся из p-мерной случайной матрицы η преобразованием $\xi = {}^{0,p}(P^{B_{2p,p}} \overset{\circ}{\eta})$, называется матрицей главных компонент случайной p-мерной матрицы. Дисперсионная матрица матрицы главных компонент ξ равна Λ:

$$R_\xi = E({}^{0,0}(\xi\xi)) = \Lambda.$$

Случайная матрица η представляется посредством случайной матрицы ξ главных компонент в виде:

$$\eta = a_\eta + {}^{0,p}(P\xi).$$

Очевидно, что представленную теорему можно применить к оценке $\hat{R}_\eta$ дисперсионной матрицы R_η p-мерной случайной матрицы η. В этом случае можно говорить о статистическом многомерно-матричным методе главных компонент.

Многомерная модель метеорологических данных и OLAP-системы

В настоящее время серьезное внимание уделяется концепциям построения и средствам реализации информационных систем, ориентированных на аналитическую обработку данных. Аналитическая обработка предполагает быстрый доступ к большим объемам данных и выполнение большого объема вычислений, причем используемые данные чаще всего должны быть упорядочены во времени. Это накладывает определенные ограничения на способы хранения данных, т.е. на базы данных. Наиболее распространенными в настоящее время являются реляционные базы данных. Несмотря на то, что существование реляционных баз данных, а также их использование, в том числе и для аналитической обработки данных, не подвергается сомнению, осуществляется поиск других подходов к организации баз данных. Одним из них является многомерный подход. Этот подход впервые был представлен для широкого обсуждения основоположником теории реляционных баз данных Э. Коддом, а также была определена категория OLAP-систем (Online Analytical Processing). Из 12 правил оценки OLAP-систем, представленных Э. Коддом, важнейшим является многомерность модели данных. В работе [29] изложены теоретические положения, лежащие в основе многомерной модели данных. Под многомерной моделью данных в OLAP-системе будем понимать организацию данных в виде многомерной матрицы, или, иначе, в виде гиперпрямоугольника. OLAP-системы, использующие многомерную модель данных, получили название MOLAP-систем, в отличие от ROLAP-систем, использующих реляционную модель данных. Для организации гиперпрямоугольника данных в MOLAP-системе необходимо определить оси координат (индексы) p-мерного пространства и данные, располагаемые в узлах сетки гиперпрямоугольника. Значения индексов будем называть также метками на осях координат. Оси выбираются таким образом, чтобы значение мультииндекса однозначно определяло данные. Данные чаще всего представляют собой строки, аналогичные строкам реляционной таблицы данных, так что мультииндекс является составным ключом этой таблицы. Выбор осей осуществляется для каждой подсистемы OLAP-системы на интуитивном уровне и не представляет трудностей. При описанном способе построения модели данных не нужно выполнять нормализацию таблиц данных, что существенно упрощает процесс проектирования многомерной базы данных по сравнению с реляционной.

Физически многомерная модель данных может состоять их двух текстовых файлов, один из которых определяет оси с их наименованиями, а другой – данные, соответствующие этим осям. В работе [29] рассмотрены также вопросы манипулирования данными в MOLAP-

системе, разработана многомерная модель данных в MOLAP-системе, предназначенной для аналитической обработки метеорологических данных. В частности, разработаны структуры файла осей и файла данных многомерной модели метеорологических данных. В качестве осей выбраны метеостанция, год, месяц, день и час, так как они однозначно определяют данные. Реализованное на основе многомерной модели данных программное средство статистического прогнозирования количественных характеристик погоды подтвердило свою работоспособность и успешно эксплуатируется в режиме накопления статистических данных о точности прогнозирования [30].

Заключение

Достоверность представленных в обзоре результатов подтверждена как математическими доказательствами, так и компьютерными расчетами, статистическим компьютерным моделированием, разработкой полноценно функционирующих программных средств. Наличие достаточно большого количества приложений, получивших достоверное, эффективное и завершенное разрешение в рамках разработанного многомерно-матричного математического аппарата, является подтверждением его целесообразности, достоверности и эффективности.

MATHEMATICAL MODELS OF MULTIDIMENSIONAL DATA

V.S. MUKHA

Abstract

The review of modern state of the problem of the multidimensional data analysis is given. The new mathematical models of the multidimensional data are considered.

Список литературы

1. *Муха В.С.* // Автоматика и вычислительная техника. 1987. Вып. 16. С. 65–71.
2. *Муха В.С.* // Изв. АНБ. Сер. физ.-мат. наук. 1990. № 4. С. 42–47.
3. *Муха В.С.* // Автоматика и вычислительная техника. 1991. Вып. 20. С. 128–133.
4. *Муха В.С.* // Весці АНБ. Сер. фіз.-мат. навук. 1993. № 2. С. 37–44.
5. *Муха В.С.* // Автоматика и вычислительная техника. 1993. Вып. 21. С. 65–72.
6. *Муха В.С.* // Весці АНБ. Сер. фіз.-мат. навук. 1993. № 4. С. 39–43.
7. *Муха В.С.* // Весці АНБ. Сер. фіз.-мат. навук. 1997. № 2. С. 46–53.
8. *Муха В.С.* // Электромагнитные волны и электронные системы. 1998. Т 3. № 4. С. 18–22.
9. *Муха В.С.* // Весці НАН Беларусі. Сер. фіз.-мат. навук. 2001. № 2. С. 64–68.
10. *Муха В.С.* // Автоматика и телемеханика. 2001. № 4. С. 80–90.
11. *Муха В.С.* // Докл. БГУИР. 2004. № 1 (5). С. 38–49.
12. *Муха В.С.* Анализ многомерных данных. Минск, 2004.
13. *Муха В.С.* // Весці НАН Беларусі. Сер. фіз.-мат. навук. № 1. 2004. С. 69–73.
14. *Корчиц К.С., Муха В.С.* // Докл. БГУИР. 2003. Т. 1. № 3. С. 102–105.
15. *Муха В.С.* // Проблемы управления и информатики. 2005. № 5. С. 103–109.
16. *Муха В.С.* // Проблемы управления и информатики. 2006. № 5. С. 100–108.
17. *Муха В.С.* // Весці НАН Беларусі. Сер. фіз.-мат. навук. № 3. 2007. С.131–134.
18. *Муха В.С., Маркушевский А.М.* // Докл. БГУИР. 2010. № 3 (49). С. 87–91.
19. *Муха В.С., Гончаренок Е.О.* // Докл. БГУИР. 2010. № 7 (53). С. 59–65.
20. *Муха В.С.* // Междунар. научн.-техн. журнал «Проблемы управления и информатики». 2012. № 6. С. 125–136.
21. *Муха В.С.* // Кибернетика и системный анализ. 2007. № 3. С. 138–143.
22. *Муха В.С.* // Весці НАН Беларусі. Сер. фіз.-мат. навук. № 3. 2010. С. 17–24.
23. *Муха В.С.* Статистические методы обработки данных. Минск, 2009.
24. *Муха В.С.* // Весці НАН Беларусі. Сер. фіз.-мат. навук. 2007. № 1. С. 45–51.
25. *Муха В.С., Козячий А.Н.* // Докл. БГУИР. 2011. № 3 (57). С. 25–28.
26. *Муха В.С.* // Весці НАН Беларусі. Сер. фіз.-мат. навук. 2013. № 1. С. 64–70.

27. *Муха В.С., Трофимович А.Ф.* // Докл. БГУИР. 2009. № 1 (39). С. 93–99.
28. *Муха В.С.* // Весці НАН Беларусі. Сер. фіз.-мат. навук.№ 3. 2013. С. 31–37.
29. *Муха, В.С., Козячий А.Н.* // Докл. БГУИР. 2010. № 1 (47). С. 100–105.
30. *Муха В.С., Трофимович А.Ф.* // Докл. БГУИР. 2011. № 8 (62). С. 14–21.

СВЕДЕНИЯ ОБ АВТОРЕ

Муха Владимир Степанович (1945 г.р.), д.т.н., профессор. В 1967 г. окончил МРТИ. В 1974 г. защитил кандидатскую, в 1994 г. – докторскую диссертацию. С 1999 г. по 2013 г. работал заведующим кафедрой ИТАС БГУИР. С 2013 г. – профессор кафедры ИТАС. Автор и соавтор 169 публикаций, среди которых1 монография и 2 учебных пособия с грифом Министерства образования Республики Беларусь. Область научных интересов – системы обработки информации и управления: теория, методы, алгоритмы, программные средства (многомерно-матричный анализ, статистические методы оценивания, распознавания образов).