МНОГОМЕРНО-МАТРИЧНАЯ РЕГРЕССИЯ НА ГЛАВНЫЕ КОМПОНЕНТЫ

Муха В. С.
Кафедра информационных технологий автоматизированных систем, Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: mukha@bsuir.by

Приведена теорема о многомерно-матричном методе главных компонент. Разработана многомерно-матричная регрессия на главные компоненты, указан пример ее применения.

Введение

Метод главных компонент находит широкое применение для выявления скрытых закономерностей и экономизации алгоритмов обработки информации в хемометрике, биометрике, эконометрике, энвайрометрике и других областях знаний [1]. В настоящее время этот метод разработан лишь для случайных векторов [2]. Актуальным является его обобщение на случайные объекты более сложной структуры – обычные и многомерные случайные матрицы. Актуальным является также поиск и разработка новых направлений и областей применения этого метода. В данной работе приводится обобщение метода главных компонент на многомерные матрицы и разрабатывается многомерно-матричная линейная регрессия на главные компоненты, позволяющая использовать метод главных компонент в многомерных задачах прогнозирования.

I. Многомерно-матричный метод главных компонент

Определение. $2q$-мерную матрицу $P = P_{(q,0,q)}$ назовем $(q,0,q)$-ортогональной, если выполняется равенство

$$^{0,q}(PP^T) =^{0,q} (P^TP) = E(0,q), \quad T = B_{2q,q}, \quad (1)$$

где $E(0,q)$ – $(0,q)$-единичная матрица, $T = B_{2q,q}$ – подстановка транспонирования типа «вперед» [3].

Из этого определения следует, что $P^T = {}^{0,q}P^{-1}$, т.е. матрица P^T является $(0,q)$-обратной к P. Очевидно, что при $q = 1$ матрица P в (1) является обычной (двухмерной) ортогональной матрицей [4].

Теорема 1 *Если $\tau = (\tau_l)$, $l = (l_1, l_2, ..., l_q)$ – q-мерная гиперпрямоугольная случайная матрица со средним значением $\nu_\tau = E(\tau)$, то ее дисперсионная матрица $\mu_{\tau^2} = E(^{0,0}(\overset{o}{\tau}\overset{o}{\tau}))$, $\overset{o}{\tau} = \tau - \nu_\tau$, E – символ математического ожидания [3], может быть представлена в виде*

$$\mu_{\tau^2} =^{0,q} (^{0,q}(P\Lambda)P^T), \quad (2)$$

где Λ – $2q$-мерная матрица, составленная из $(q,0,0)$-собственных чисел α_c матрицы μ_{τ^2} (элементов матрицы $\alpha = \alpha_{(q,0,0)} = (\alpha_c)$, $c = (c_1, c_2, ..., c_q)$) в соответствии с выражением

$$\Lambda = (\lambda_{c,c'}) = \left(\left\{ \begin{array}{ll} \alpha_c, & c = c', \\ 0, & c \neq c', \end{array} \right\} \right),$$

а P – $2q$-мерная $(q,0,q)$-ортогональная матрица, составленная из нормированных фундаментальных $(q,0,0)$-собственных матриц $(y_c)_{c'}$ матрицы μ_{τ^2} по формуле

$$P = P_{(q,0,q)=(p_{c,c'})=((y_c)_{c'})}, \quad (3)$$

$$c = (c_1, c_2, ..., c_q), \quad c' = (c'_1, c'_2, ..., c'_q).$$

Случайная q-мерная матрица $\overset{o}{\xi} = (\xi - \nu_\xi) = (\overset{o}{\xi}_l)$, $\nu_\xi = E(\xi)$, $l = (l_1, l_2, ..., l_q)$, получающаяся из q-мерной случайной матрицы $\overset{o}{\tau}$ преобразованием

$$\overset{o}{\xi} =^{0,q} (P^T \overset{o}{\tau})$$

с ортогональной $2q$-мерной матрицей P из (1), называется матрицей главных компонент случайной q-мерной матрицы $\overset{o}{\tau}$, а ее элемент $\overset{o}{\xi}_l$ – l-й главной компонентой случайной матрицы $\overset{o}{\tau}$. Дисперсионная матрица матрицы главных компонент $\overset{o}{\xi}$ равна Λ:

$$\mu_{\xi^2} = E(^{0,0}(\overset{o}{\xi}, \overset{o}{\xi})) = \Lambda. \quad (4)$$

Справедливо следующее представление случайной матрицы $\overset{o}{\tau}$ посредством случайной матрицы $\overset{o}{\xi}$ главных компонент:

$$\overset{o}{\tau} =^{0,q} (P \overset{o}{\xi}). \quad (5)$$

Приведенную теорему назовем вероятностным многомерно-матричным методом главных компонент.

Отметим, что в силу равенства (4) главные компоненты $\overset{o}{\xi}_l$ случайной матрицы $\overset{o}{\tau}$ не коррелированы, а дисперсии главных компонент равны собственным числам дисперсионной матрицы μ_{τ^2}.

II. Многомерно-матричная регрессия на главные компоненты

Пусть η и $\tau - p$- и q-мерные случайные матрицы соответственно, $\mu_{\eta\tau} = E(^{0,0}(\overset{o}{\eta}\overset{o}{\tau}))$ – взаимная ковариационная матрица матриц η и τ. Как известно [5], оптимальная линейная регрессия η на τ имеет вид:

$$\hat{\eta} = \nu_\eta +^{0,q} (^{0,q}(\mu_{\eta\tau}\,^{0,q}(\mu_{\tau^2})^{-1})(\tau - \nu_\tau)). \qquad (6)$$

Пусть $\overset{o}{\xi}$ – матрица главных компонент матрицы $\overset{o}{\tau}$. Обращая матрицу μ_{τ^2} (2), получим

$$^{0,q}(\mu_{\tau^2})^{-1} =^{0,q} (^{0,q}(^{0,q}(P^T)^{-1}\,^{0,q}\Lambda^{-1})\,^{0,q}P^{-1}) =$$

$$=^{0,q} (^{0,q}(P\,^{0,q}\Lambda^{-1})\,^{0,q}P^{-1}). \qquad (7)$$

Подставляя (7), (5) в (6), получим следующее выражение

$$\hat{\eta} = \nu_\eta +^{0,q} (^{0,q}(^{0,q}(\mu_{\eta\tau}P)\,^{0,q}\Lambda^{-1})(\xi - \nu_\xi)), \qquad (8)$$

которое будем называть оптимальной линейной многомерно-матричной регрессией на главные компоненты.

Заменяя в выражениях (6) и (8) генеральные моменты эмпирическими, мы получим соответствующие эмпирические регрессии для случайных матриц η и ξ. В частности, если $(q+1)$-мерная матрица

$$X = (x_{i,j}) = ((x_i)_j), \quad i = (i_1, i_2, ..., i_q), \quad j = \overline{1,n},$$

представляет собой выборку объема n из распределения случайной матрицы τ, т. е. сечение $(x_i)_j = x_j$ ориентации j матрицы X представляет собой выборочное значение случайной матрицы τ, а $(p+1)$-мерная матрица

$$Y = (y_{i,j}) = ((y_i)_j), \quad i = (i_1, i_2, ..., i_p), \quad j = \overline{1,n},$$

– выборкой объема n из распределения случайной матрицы η, полученной при выборке X, то эмпирические моменты рассчитываются по формулам:

$$\hat{\nu}_\tau = \frac{1}{n}\sum_{j=1}^{n} x_j, \quad \hat{\nu}_\eta = \frac{1}{n}\sum_{j=1}^{n} y_j,$$

$$\hat{\mu}_{\tau^2} = \frac{1}{n}\sum_{j=1}^{n} {}^{0,0}(\overset{o}{x}_j\overset{o}{x}_j) = \frac{1}{n}\,{}^{0,1}(\overset{o}{X}\overset{o}{X}\;{}^{B_{q+1,1}}), \qquad (9)$$

$$\hat{\mu}_{\eta\tau} = \frac{1}{n}\sum_{j=1}^{n} {}^{0,0}(\overset{o}{y}_j\overset{o}{x}_j) = \frac{1}{n}\,{}^{0,1}(\overset{o}{Y}\overset{o}{X}\;{}^{B_{q+1,1}}),$$

где

$$\overset{o}{x}_j = x_j - \hat{\nu}_\tau, \quad \overset{o}{y}_j = y_j - \hat{\nu}_\eta,$$

$$\overset{o}{X} = (\overset{o}{x}_j) = ((\overset{o}{x}_i)_j), \quad \overset{o}{Y} = (\overset{o}{y}_j) = ((\overset{o}{y}_i)_j).$$

Идея линейной регрессии на главные компоненты описана в [1] для частного случая множественной эмпирической регрессии.

Эмпирический алгоритм (6) обычной линейной многомерно-матричной регрессии предполагает обращение матрицы $\hat{\mu}_{\tau^2}$ (9). Ввиду случайности измерений $\overset{o}{x}_j$ в выборке $\overset{o}{X}$ могут быть измерения, близкие к линейно зависимым. В этом случае матрица $\hat{\mu}_{\tau^2}$ (9) окажется близкой к вырожденной. Алгоритм (6) окажется неустойчивым, т. е. будет сопровождаться накоплением ошибок. В противоположность к сказанному алгоритм (8) линейной многомерно-матричной регрессии на главные компоненты не содержит такого обращения и является более устойчивым. Кроме того, в этом алгоритме мы можем не учитывать ряд главных компонент и тем самым сократить объем вычислений.

Практической задачей, в которой целесообразно воспользоваться алгоритмом регрессии на главные компоненты, является прогнозирование количественных характеристик погоды [6]. В этой задаче матрицы η и τ двухмерные ($p = q = 2$). Практический интерес представляет совместное прогнозирование двух или более характеристик (например, температуры и атмосферного давления) на 14 дней вперед по их измерениям за такой же прошлый период. В случае двух характеристик матрица $\hat{\mu}_{\tau^2}$ будет иметь размер $(2 \times 112 \times 2 \times 112)$. Обращение этой матрицы эквивалентно обращению обычной матрицы размером (224×224). При таких объемах данных следует считаться с возможной неустойчивостью алгоритма обычной эмпирической регрессии. В связи с этим был реализован и проверен на тех же данных алгоритм эмпирической регрессии на главные компоненты. При расчете главных компонент использовалась программа eig.m Matlab. Алгоритм, в котором удерживались все главные компоненты, оказался более устойчивым и показал вдвое большее быстродействие по сравнению с имеющейся реализацией алгоритма обычной эмпирической регрессии. Быстродействие алгоритма (8) можно еще несколько увеличить за счет отбрасывания части незначимых главных компонент.

III. Список литературы

1. Эсбенсен, К. Анализ многомерных данных. Пер. с англ. Под ред. О.Е. Родионовой / К. Эсбенсен. – Черноголовка: ИПХФ РАН, 2005. – 160 с.
2. Рао, С. Р. Линейные статистические методы и их применение / С. Р. Рао. – М.: Мир, 1968. – 548 с.
3. Муха, В. С. Анализ многомерных данных / В. С. Муха. – Минск: Технопринт, 2004. – 368 с.
4. Хорн, Р. Матричный анализ / Р. Хорн, Ч. Джонсон. – М.: Мир, 1989. – 656 с.
5. Муха, В.С. Наилучшая полиномиальная многомерно-матричная регрессия / В. С. Муха // Кибернетика и системный анализ. – № 3, 2007. – С. 138 – 143.
6. Муха, В.С. Статистическое векторное прогнозирование количественных характеристик погоды / В.С. Муха // Информационные системы и технологии (IST'2004). Материалы Международной конференции (Минск, 8 – 10 ноября 2004 г.). – Часть 2. – Мн.: Академия управления при президенте РБ, 2004. – С. 195 – 200.