УДК 621.391

Ю.М. Коршунов

ПОЛУЧЕНИЕ МНОГОМЕРНОЙ СТАТИСТИЧЕСКОЙ ВЫБОРКИ С ЗАДАННЫМИ КОРРЕЛЯЦИОННЫМИ СВОЙСТВАМИ

Дается описание алгоритма, позволяющего по известной корреляционной матрице получить многомерную статистическую выборку, корреляционные свойства которой близки к свойствам, заданным корреляционной матрицей.

Введение. В задачах имитационного моделирования очень важным является исследование работы моделируемого объекта при разнообразных внешних условиях, в том числе при воздействии на объект случайных факторов с заданными статистическими свойствами. В математической статистике разработаны способы моделирования одномерных случайных величин с различными законами распределения. Эти методы распространяются и на многомерные статистические выборки при отсутствии корреляционных связей между переменными. В настоящей работе рассматривается метод получения многомерной статистической выборки с заданными корреляционными свойствами.

Числовые характеристики многомерной статистической выборки. Многомерная статистическая выборка X представляет собой матрицу размером $n \times m$, строки которой представляют собой n анализируемых объектов, а столбцы описывают m признаков, характеризующих свойства этих объектов [1]. Так, элемент матрицы x_{ij} является ***признаком***, описывающим j-е свойство i-го объекта. В дальнейшем статистическую выборку будем называть ***матрицей наблюдений***. Приведем компактную запись этой матрицы:

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad x_i = \left(x_{i1} \cdots x_{im}\right), \quad i = \overline{1,n}.$$

В качестве одной из числовых характеристик полученной статистической выборки используется вектор математических ожиданий признаков

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \quad j = \overline{1,m},$$

позволяющий получить ***центрированную*** статистическую выборку X^o с элементами

$$x_{ij}^o = x_{ij} - \bar{x}_j, \quad i = \overline{1,n}, \quad j = \overline{1,m}.$$

Другой числовой характеристикой выборки является ***ковариационная матрица*** Q размером $m \times m$

$$Q = \begin{bmatrix} s_{11} \cdots s_{1m} \\ \cdots \ \cdots \ \cdots \\ s_{m1} \cdots s_{mm} \end{bmatrix},$$

диагональные члены которой s_{jj} представляют собой дисперсии признаков

$$s_{jj} = sg_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2, \ j = \overline{1,m},$$

а остальные члены $s_{jk}, j \neq k$, представляют собой ковариации между признаками

$$s_{jk} = cvr_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)\left(x_{jk} - \bar{x}_k\right), \ j,k = \overline{1,m}.$$

Решение поставленной задачи основано на анализе ***нормированной*** статистической выборки Z, которая получается путем замены элементов x_{ij} исходной выборки X на элементы

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{sg_j}, \ i = \overline{1,n}, \ j = \overline{1,m},$$

где $sg_j = \sqrt{sg_j^2}$ – среднее квадратическое отклонение j – го признака.

В нормированной статистической выборке Z математические ожидания каждого признака $\bar{z}_j$ равны нулю, дисперсии каждого признака $\overline{z_j^2}$ равны единице, а роль ковариационной матрицы играет ***корреляционная*** матрица R, имеющая вид

$$R = \begin{bmatrix} r_{11} \cdots r_{1m} \\ \cdots \ \cdots \ \cdots \\ r_{m1} \cdots r_{mm} \end{bmatrix},$$

где r_{jk} – ***коэффициенты корреляции***, связанные с элементами s_{jk} ковариационной матрицы соотношениями

$$r_{jk} = \frac{s_{jk}}{sg_j \cdot sg_k}, \ j,k = \overline{1,m}.$$

Диагональные элементы r_{jj} корреляционной матрицы равны единице, и матрица симметрична относительно главной диагонали, т.е. $r_{jk} = r_{kj}$.

Более компактную запись корреляционной матрицы можно представить в виде

$$R = \frac{1}{n-1} Z \cdot Z^T.$$

Модель статистической выборки, используемой для решения поставленной задачи. В основу построения модели, описывающей корреляционные свойства случайной выборки, положены идеи, лежащие в основе факторного анализа [2] и использованные применительно к нормированной статистической выборке Z.

Все признаки $z_{ij}, i = \overline{1,n}, \ j = \overline{1,m}$, нормированной статистической выборки обладают некоторыми общими свойствами, отражающими особенности рассматриваемой статистической выборки, но в то же время каждый из признаков обладает индивидуальными свойствами, отличающими данный признак от всех других. Поскольку признаки z_{ij} являются случайными числами, то для создания статистической выборки с требуемыми свойствами необходимо иметь ***модель***, позволяющую целенаправленно воздействовать на свойства признаков путем рационально построенного набора случайных чисел.

Для этого вводится $l \le m$ случайных векторов

$$p_{1i}, \cdots, p_{li}, \ i = \overline{1,n},$$

называемых ***общими факторами***, каждый из которых воздействует на все признаки и определяет тем самым ***общие свойства*** всей выборки. Совокупность рассмотренных векторов можно представить в виде матрицы P с элементами p_{ki}, которые должны быть случайными величинами с нулевыми математическими ожиданиями и единичной дисперсией, некоррелированными между собой:

$$\overline{p}_k = 0, \quad \overline{p_k^2} = 1, \ j = \overline{1,l}.$$
$$\overline{p_j p_k} = 0, \ j,k = \overline{1,l}, \ j \ne k.$$

Здесь и в дальнейшем черта сверху означает математическое ожидание признака.

Введем в рассмотрение также матрицу A размером $m \times l$, элементы которой a_{jk} назовем ***факторными нагрузками общих факторов*** p_{ji}.

С учетом полученных соотношений общие свойства признака z_{ij} найдутся из соотношения

$$\sum_{k=1}^{l} a_{jk} p_{ki}, \ i = \overline{1,n}, \ j = \overline{1,m}. \qquad \text{(a)}$$

Для учета индивидуальных свойств рассматриваемого признака z_{ij} вводится матрица ***характерных факторов*** в виде матрицы Y с элементами y_{ji}, представляющими собой случайные числа с нулевым математическим ожиданием и единичной дисперсией, некоррелированные между собой и некоррелированные с матрицей P. В качестве факторных нагрузок характерных факторов Y используется вектор $G = (g_1, \cdots, g_m)$. Специфические свойства признака z_{ij} найдутся из соотношения

$$g_j y_{ji}, \ i = \overline{1,n}, \ j = \overline{1,m}. \qquad \text{(b)}$$

Объединяя вместе соотношения (a) и (b), получаем полное описание математической модели нормирований матрицы наблюдений Z:

$$z_{ij} = \sum_{k=1}^{l} a_{jk} p_{ki} + g_j y_{ji}, \ i = \overline{1,n}, \ j = \overline{1,m}. \qquad (1)$$

Свойства факторных нагрузок a_{jk} и g_i найдутся из условия нормированности матрицы Z, в соответствии с которым должны выполняться условия

$$\overline{z}_j = 0, \ \overline{z_j^2} = 1.$$

Первое условие выполняется в силу того, что $\overline{p}_k = 0$ и $\overline{y}_j = 0$. Для проверки второго условия заметим, что

$$\overline{p_k^2} = 1, \ \overline{y_j^2} = 1, \ \overline{p_k y_j} = 0.$$

При этом для дисперсии $\overline{z_j^2}$ получаем выражение

$$\overline{z_j^2} = h_j^2 + g_j^2 = 1,$$

где

$$h_j^2 = \sum_{k=1}^{l} a_{jk}^2.$$

В этом выражении величину h_j^2 называют ***общностью***, а величину g_j^2 – ***характерностью*** матрицы Z.

Более удобной формой представления модели (1) будет матричная форма записи. Для этого вектор G следует представить в виде матрицы размером $m \times m$ с элементами g_{jk}, определенными как

$$g_{jk} = \begin{cases} 0, \text{если } k \ne j \\ g_j, \text{если } k = j \end{cases}.$$

При этом G будет диагональной матрицей вида $G = diag(g_1,\cdots,g_m)$ и матричная форма записи модели (1) будет иметь вид

$$Z = AP + GY. \qquad (2)$$

Если матрицы A и G заданы, то, вводя в модель (2) случайные матрицы P и Y, получим многомерную статистическую выборку, свойства которой будут определяться свойствами факторных нагрузок A и G.

Для решения поставленной задачи необходимо увязать свойства факторных нагрузок A и G со свойствами корреляционной матрицы, которая будет иметь вид

$$R = \frac{1}{n-1}ZZ^T = \frac{1}{n-1}(AP + GY)(P^TA^T + Y^TG^T).$$

Учитывая, что векторы P и Y не коррелированы между собой, а также что

$$\frac{1}{n-1}PP^T = I_l, \quad \frac{1}{n-1}YY^T = I_m,$$

где I_l и I_m - единичные матрицы размером l и m, корреляционную матрицу R можем представить в виде

$$R = Rh + U^2, \qquad (3)$$

где

$$Rh = AA^T, \qquad (4)$$

$$U^2 = GG^T = diag(g_1^2,\cdots,g_m^2). \qquad (5)$$

Свойства корреляционной матрицы R можно выразить через ее собственные значения $\Lambda = (\lambda_1,\cdots,\lambda_m)$ и собственные векторы $V = (v_1,\cdots,v_m)$. Численные значения λ_j определяют информативность собственного вектора v_j и их удобно располагать в порядке убывания значений λ_j. При этом собственные векторы, обладающие малой информативностью, могут не учитываться и свойства корреляционной матрицы будут определяться небольшим числом $l < m$ значений v_j.

Собственные векторы v_j не коррелированы между собой и определяются с точностью до постоянного множителя. Стандартные методы их определения дают нормированные значения собственных векторов, удовлетворяющих соотношению

$$\sum_{j=1}^{m} v_{ij}^2 = 1, \ i = \overline{1,l}.$$

Однако для выполнения условия (4) следует ввести новые обозначения a_{ij} элементов собственных векторов, пересчитав их по соотношению $a_{ij} = v_{ij} \cdot \sqrt{\lambda_i}$, чтобы имело место

$$\sum_{j=1}^{m} a_{ij}^2 = \lambda_i, \ i = \overline{1,l}.$$

Тогда матрица A с элементами a_{ij} будет удовлетворять соотношению (4). При этом матрица Rh, называемая ***редуцированной*** корреляционной матрицей, в соответствии с соотношением (3) найдется как

$$Rh = R - U^2. \qquad (6)$$

Поскольку U^2 согласно (5) является диагональной матрицей с диагональными элементами g_i^2, то различие между матрицами Rh и R будет только в диагональных элементах и редуцированную матрицу Rh можно получить из корреляционной матрицы R, заменив в ней диагональные члены $r_{ii} = 1$ на значения $rh_{ii} = g_i^2, i = \overline{1,m}.$

Однако значения g_i^2 нам неизвестны и можно лишь каким-либо образом задать их ***грубые оценки***. Качество этих грубых оценок можно проверить по соотношению (4). Если полученная матрица Rh этому соотношению не удовлетворяет, то можно попытаться изменить ее диагональные элементы и вновь проверить выполнение условия (4). Такие изменения можно проводить неоднократно, используя рекуррентную процедуру улучшения грубой оценки. Если улучшения не происходит, то это означает, что грубая оценка была выбрана неудачно.

Дадим описание одного из возможных методов решения задачи.

Рекуррентный метод улучшения грубой оценки редуцированной корреляционной матрицы. Одним из возможных методов получения грубой оценки диагональных элементов rh_{ii} редуцированной корреляционной матрицы Rh является замена диагональных элементов r_{ii} в каждой строке i корреляционной матрицы R на взятое с положительным знаком максимальное значение r_{ij} в этой строке

$$rh_{ii} = \max_j\left(\left|r_{ij}\right|\right), \ i,j = \overline{1,m}, \ j \neq i. \qquad (7)$$

Для полученной матрицы Rh находим собственные значения Λ, определяем число l общих факторов P и получаем матрицу собственных векторов A. Находим скалярное произведение AA^T и проверяем, совпадает ли оно со значением Rh. Если совпадения нет, то заменяем в матрице Rh диагональные члены на диагональные члены матрицы AA^T и повторяем всю процедуру для измененного значения Rh. Путем многократного повторения указанной процедуры

добиваемся выполнения соотношения (4). После этого можно, задавшись значением n и сформировав случайные матрицы факторов P и Y по уравнению (1), получить многомерную статистическую выборку, корреляционные свойства которой будут близки к свойствам корреляционной матрицы R, что можно проверить, определив корреляционную матрицу полученной выборки.

Если удовлетворительного решения не получилось, то следует вместо (7) поискать другой метод получения диагональных элементов редуцированной корреляционной матрицы.

Иллюстративные примеры. По рассмотренному алгоритму на кафедре АИТУ была разработана в пакете Matlab программа md_msa1.m, реализующая данный метод и позволившая на примерах проверить эффективность получаемого решения.

В приводимых примерах опущены все промежуточные выкладки и даются только исходная корреляционная матрица R и корреляционная матрица $\hat{R}$, полученная по многомерной статистической выборке, реализованной по рассмотренному алгоритму.

Пример 1

$$R = \begin{bmatrix} 1.0000 & 0.8000 & -0.4000 \\ 0.8000 & 1.0000 & -0.5600 \\ -0.4000 & -0.5600 & 1.0000 \end{bmatrix}$$

$$\hat{R} = \begin{bmatrix} 1.0000 & 0.7852 & -0.3315 \\ 0.7852 & 1.0000 & -0.5437 \\ -0.3315 & -0.5437 & 1.0000 \end{bmatrix}$$

Пример 2

$$R = \begin{bmatrix} 1.0000 & 0.3397 & -0.3227 & -0.0825 \\ 0.3397 & 1.0000 & 0.3603 & -0.1022 \\ -0.3227 & 0.3603 & 1.0000 & -0.2379 \\ -0.0825 & -0.1022 & -0.2379 & 1.0000 \end{bmatrix}$$

$$\hat{R} = \begin{bmatrix} 1.0000 & 0.3512 & -0.3740 & -0.1784 \\ 0.3512 & 1.0000 & 0.3159 & -0.1215 \\ -0.3740 & 0.3159 & 1.0000 & -0.1991 \\ -0.1784 & -0.1215 & -0.1991 & 1.0000 \end{bmatrix}$$

Сопоставление элементов матриц R и $\hat{R}$ в приведенных примерах показывает сходство этих матриц. Некоторое различие в элементах матриц вполне объяснимо случайным характером и ограниченным объемом полученной многомерной выборки.

Библиографический список

1. *Андерсон Т.В.* Введение в многомерный статистический анализ. - М.: Физматгиз, 1963.

2. *Харман Г.* Современный факторный анализ. - М.: Статистика, 1980.