УДК 519.224:330.46

А.А. Новоселов

ПОСТРОЕНИЕ МНОГОМЕРНЫХ ДИСКРЕТНЫХ РАСПРЕДЕЛЕНИЙ С ЗАДАННОЙ КОРРЕЛЯЦИОННОЙ СТРУКТУРОЙ

В работе рассматриваются методы воспроизведения многомерного дискретного распределения с заданной корреляционной структурой и маргинальными распределениями. Для воспроизведения используются смеси базовых распределений и решение некоторых оптимизационных задач.

Ключевые слова: дискретное распределение, корреляция, копула, смесь

Введение

Пусть заданы нормальные распределения со средними значениями $\mu_1,\ldots,\ \mu_d$ и стандартными отклонениями $\sigma_1,\ldots,\ \sigma_d$. Для произвольной корреляционной матрицы R существует единственное многомерное нормальное распределение, обладающее такими маргинальными распределениями и корреляционной матрицей. Хорошо известный алгоритм воспроизведения соответствующего случайного вектора $X=(X_1,\ldots,\ X_d)$ основан на факторизации ковариационной матрицы.

Обозначим

$$\Lambda=\begin{pmatrix}\sigma_1 & 0 & \cdots & 0\\ 0 & \sigma_2 & \cdots & 0\\ \cdots & \cdots & \cdots & \cdots\\ 0 & 0 & \cdots & \sigma_d\end{pmatrix}$$

диагональную матрицу со стандартными отклонениями на диагонали, тогда $C=\Lambda R\Lambda$ является ковариационной матрицей распределения вектора X. Будучи неотрицательно определенной и симметричной, ковариационная матрица C может быть представлена в виде

$$C=A'A \qquad (1)$$

с некоторой матрицей A, причем последняя определяется не единственным образом. Примерами такого представления являются разложение Холецкого и ортогональное разложение.

При наличии разложения (1) вектор X воспроизводится из стандартного нормального случайного вектора Z по формуле

$$X=A'Z. \qquad (2)$$

Действительно, для Z справедливо $\boldsymbol{E}ZZ'=I$, где I — единичная матрица соответствующего размера, поэтому $\boldsymbol{E}XX'=\boldsymbol{E}(A'ZZ'A)=A'(\boldsymbol{E}ZZ')A=A'A=C$, так что X обладает требуемой ковариационной структурой.

В случае, когда компоненты X имеют фиксированные дискретные распределения, аналогичный метод оказывается неприменимым. Во-первых, заданным маргинальным распределениям и ковариационной матрице соответствуют, вообще говоря, многие многомерные дискретные распределения. Может оказаться и так, что подходящее многомерное распределение не существует.

Во-вторых, алгоритм вращения (2) не сохраняет дискретную решетку значений, на которой задано распределение.

В работах [1], [2] анонсированы методы воспроизведения двумерного дискретного распределения с заданными маргинальными распределениями и корреляцией, основанные на смесях некоторых базовых распределений и минимизации уклонения от независимого распределения. В настоящей работе предлагается обоснование этих методов.

Описание двумерного дискретного распределения

Пусть размерность $d=2$. Обозначим $K=\{1,\dots,m\}\times\{1,\dots,n\}$. Дискретное распределение вектора $X=(X_1, X_2)'$ задается на прямоугольной сетке значений $\{x_{11},\dots, x_{1m}\}\times\{x_{21},\dots, x_{2n}\}$ в виде $\boldsymbol{P}(X_1=x_{1i}; X_2=x_{2j})=r_{ij}$, $(i,j)\in K$. Обозначим

$$r=\{r_{ij}, (i,j)\in K\} \tag{3}$$

совместное распределение компонент вектора X. Здесь $P(X_1=x_{1i})=p_i$, $i=1,\dots,m$ и $P(X_2=x_{2j})=q_j$, $j=1,\dots,n$, так что векторы

$$p=(p_1,\dots,p_m),\ \ q=(q_1,\dots,q_n) \tag{4}$$

описывают маргинальные распределения компонент.

Средние значения

$$a_1 = EX_1 = \sum_{i=1}^{m} x_{1i} p_i\ ,\ \ a_2 = EX_2 = \sum_{j=1}^{n} x_{2j} q_j\ ,\ \ a_{12} = E(X_1X_2) = \sum_{i=1}^{m}\sum_{j=1}^{n} x_{1i}x_{2j}r_{ij}\ ,$$

стандартные отклонения

$$\sigma_1 = \sqrt{E(X_1 - EX_1)^2}\ ,\ \ \sigma_2 = \sqrt{E(X_2 - EX_2)^2}$$

и коэффициент корреляции

$$c = \frac{E(X_1X_2) - EX_1EX_2}{\sigma_1\sigma_2} \tag{5}$$

вычисляются, как обычно.

Если маргинальные распределения (4) известны, то выполняются соотношения

$$\sum_{i=1}^{m} r_{ij} = q_j,\ j=1,\dots,n \tag{6}$$

и

$$\sum_{j=1}^{n} r_{ij} = p_i, i=1,\dots,m\,. \tag{7}$$

Отметим, что среди $m+n$ уравнений (6), (7) имеется лишь $m+n-1$ независимых, поскольку сумма компонент любого распределения равна 1.

Если же дополнительно известен коэффициент корреляции c компонент X, то справедливо и уравнение

$$\sum_{i=1}^{m}\sum_{j=1}^{n} x_{1i}x_{2j}r_{ij} = c\sigma_1\sigma_2 + a_1a_2\,. \tag{8}$$

Обозначим $F(p,q)$ класс всех двумерных распределений (3) с маргинальными распределениями (4) (его еще называют классом Фреше), а $F_c(p,q)$ — его подкласс распределений с корреляцией c. Известно [3], что среди всех распределений в $F(p,q)$ наименьшей корреляцией $c_{min}=c_{min}(p,q)$ обладает антикомонотонное распределение $R^-(p,q)$, а наибольшей корреляцией $c_{max}=c_{max}(p,q)$ обладает комонотонное распределение $R^+(p,q)$, причем, вообще говоря, $c_{min}>-1$ и $c_{max}<1$. Класс Фреше представим в виде

$$F(p,q) = \bigcup_{c\in[c_{\min},c_{\max}]} F_c(p,q)\,. \tag{9}$$

Обозначим еще $R^0(p,q)$ независимое распределение из класса Фреше $F(p,q)$.

Пример

Пусть на сетке значений

$$\{0,1,2\}\times\{0,1\} \tag{10}$$

заданы маргинальные распределения

$$p=(1/3,1/2,1/6),\ \ q=(2/5,3/5). \tag{11}$$

Основные характеристики маргинальных распределений равны

$$a_1=\frac{5}{6}, a_2=\frac{3}{5}, \sigma_1=\frac{\sqrt{17}}{6}, \sigma_2=\frac{\sqrt{6}}{5}\ .$$

Двухпараметрическое представление класса Фреше имеет вид

$$R=\begin{pmatrix} u & v & 2/5-u-v \\ 1/3-u & 1/2-v & u+v-7/30 \end{pmatrix}. \tag{12}$$

Область допустимых значений параметров задается неравенствами

$$0\le u\le 1/3, 0\le v\le 1/2, 7/30\le u+v\le 2/5 \tag{13}$$

и представлена на рис. 1. Окружностями на нем отмечены антикомонотонное $R^-(p,q)$

$$R^-(p,q)=\begin{pmatrix} 0 & 7/30 & 1/6 \\ 1/3 & 4/15 & 0 \end{pmatrix}. \tag{14}$$

независимое $R^0(p,q)$

$$R^0(p,q)=\begin{pmatrix} 2/15 & 1/5 & 1/15 \\ 1/5 & 3/10 & 1/10 \end{pmatrix}. \tag{15}$$

и комонотонное $R^+(p,q)$

$$R^-(p,q)=\begin{pmatrix} 1/3 & 1/15 & 0 \\ 0 & 13/30 & 1/6 \end{pmatrix}. \tag{16}$$

распределения.

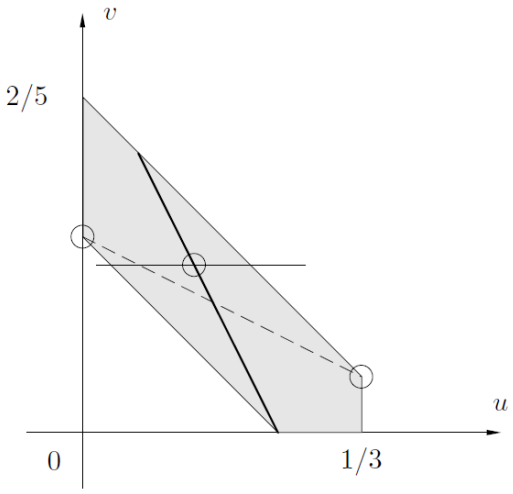

Рис. 1: Допустимая область в плоскости параметров (u, v) в примере, жирная линия соответствует некоррелированным компонентам; окружностями (слева направо) отмечены антикомонотонное, независимое и комонотонное распределения; горизонтальный отрезок изображает решения задачи 1

Класс $F_c(p,q)$ описывается в параметрическом представлении уравнением

$$v=\frac{7}{15}+c\frac{\sqrt{102}}{30}-2u\ . \tag{17}$$

Некоррелированные распределения лежат на отрезке прямой, описываемом уравнением

$$u+v=\frac{7}{15}, \frac{1}{15}\le u\le\frac{7}{30}\ .$$

Коэффициент корреляции в данном примере заключен в интервале

$$c \in \left[-\frac{7}{\sqrt{102}}, \frac{8}{\sqrt{102}} \right]. \tag{18}$$

Критерий близости к независимому распределению

Теперь сформулируем две задачи выбора из всего множества совместных распределений, удовлетворяющих условиям (6) – (8), единственного распределения, наименее уклоняющегося от независимого распределения в смысле критерия

$$f(r) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} \left(r_{ij} - p_i q_j \right)^2 \to \min_r . \tag{19}$$

Задача 1. В первой задаче целевая функция (19) минимизируется при ограничениях (6) – (8).

Функция Лагранжа этой задачи имеет вид

$$L(r, \lambda, \mu, \nu) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{n} \left(r_{ij} - p_i q_j \right)^2 + \sum_{j=1}^{n} \lambda_j \left(\sum_{i=1}^{m} r_{ij} - q_j \right) +$$
$$+ \sum_{i=1}^{m} \mu_i \left(\sum_{j=1}^{n} r_{ij} - p_i \right) + \nu \left(\sum_{i=1}^{m} \sum_{j=1}^{n} x_{1i} x_{2j} r_{ij} - c\sigma_1\sigma_2 - a_1 a_2 \right)$$

Дифференцируя функцию Лагранжа по всем переменным, и приравнивая производные к нулю, получаем уравнения

$$\partial L / \partial r_{kl} = (r_{kl} - p_k q_l) + \lambda_l + \mu_k + \nu x_{1k} x_{2l} = 0, k = 1, \ldots m, l = 1, \ldots n, \tag{20}$$

а также уравнения (6) – (8). Отметим, что ввиду соотношения $\sum_{i=1}^{m} p_i = \sum_{j=1}^{n} q_j = 1$ в системе (6) – (8) одно из уравнений, например, первое, является следствием остальных, и его можно отбросить вместе с соответствующим множителем Лагранжа λ_1. Для решения полученной системы уравнений выразим элементы матрицы r из (20) в виде функции множителей Лагранжа λ, μ, ν:

$$r_{ij} = p_i q_j - \left(\lambda_j + \mu_i + \nu x_{1i} x_{2j} \right) = 0, (i, j) \in K, \tag{21}$$

и подставим полученные выражения в уравнения (6) – (8). Имеем:

$$-m\lambda_j - \sum_{i=1}^{m} \mu_i - \nu x_{2j} \sum_{i=1}^{m} x_{1i} = 0, \; j = 2, \ldots n, \tag{22}$$

$$-\sum_{j=1}^{n} \lambda_j - n\mu_i - \nu x_{1i} \sum_{j=1}^{n} x_{2j} = 0, i = 1, \ldots m, \tag{23}$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} x_{1i} x_{2j} \left(p_i q_j - \left(\lambda_j + \mu_i + \nu x_{1i} x_{2j} \right) \right) = c\sigma_1\sigma_2 + a_1 a_2 .$$

Обозначив

$$s_1 = \sum_{i=1}^{m} x_{1i}, s_2 = \sum_{j=1}^{m} x_{2j}, s_{12} = \sum_{i=1}^{m} x_{1i}^2 \sum_{j=1}^{n} x_{2j}^2 ,$$

преобразуем последнее уравнение к виду

$$-s_1 \sum_{j=1}^{n} \lambda_j x_{2j} - s_2 \sum_{i=1}^{m} \mu_i x_{1i} - \nu s_{12} = c\sigma_1\sigma_2 , \tag{24}$$

Обозначим I_n единичную матрицу размера $n \times n$, а J_{mn}— прямоугольную матрицу размера $m \times n$, все элементы которой равны 1. Далее, обозначим $\bar{x}_2 = (x_{22}, \ldots, x_{2n})', \bar{x}_1 = (x_{11}, \ldots, x_{1m})'$. В системе уравнений (6) – (8), как уже отмечалось, первое уравнение является следствием

остальных, поэтому его можно отбросить. Кроме того, в полученной системе $m+n$ уравнений относительно $m+n+1$ неизвестных одну из неизвестных можно выбрать произвольным образом. Мы будем полагать $\lambda_1 = 0$, что соответствует отбрасыванию первого столбца в матрице системы. После этих операций система уравнений запишется в виде

$$A\gamma = b, \tag{25}$$

где неизвестные множители Лагранжа обозначены

$$\gamma = (\lambda_2, \ldots, \lambda_n, \mu_1, \ldots, \mu_n, \nu)', \tag{26}$$

а матрица системы A и вектор правых частей имеют вид

$$A = \begin{pmatrix} -mI_{n-1} & -J_{n-1,m} & -s_1\bar{x}_2 \\ -J_{m,n-1} & -nI_m & -s_2\bar{x}_1 \\ -s_1\bar{x}_2' & -s_2\bar{x}_1' & -s_{12} \end{pmatrix}, \quad b = (0, \ldots, 0, c\sigma_1\sigma_2).$$

Квадратная система линейных уравнений (25) имеет единственное решение (26), добавляя в которое значение $\lambda_1 = 0$, из (21) вычисляем искомое распределение.

Задача 2. Во второй задаче целевая функция (19) минимизируется при ограничениях (6) – (8) и

$$r_{ij} \geq 0, (i, j) \in K. \tag{27}$$

Аналитическое решение этой задачи в общем виде недоступно, однако численные методы позволяют эффективно решать ее. Приведем решения для рассмотренного примера.

Максимальному значению $c = c_{\max} = 8/\sqrt{102}$ соответствует решение

$$r = \begin{pmatrix} 1/3 & 1/15 & 0 \\ 0 & 13/30 & 1/6 \end{pmatrix},$$

которое, как нетрудно заметить, представляет собой комонотонное распределение (16).

Минимальному значению $c = c_{\min} = -7/\sqrt{102}$ соответствует решение

$$r = \begin{pmatrix} 0 & 7/30 & 1/6 \\ 1/3 & 4/15 & 0 \end{pmatrix},$$

представляющее собой антикомонотонное распределение (14).

Результаты работы позволяют эффективно строить дискретные вероятностные модели с заданной корреляционной структурой. Предложенная методика без труда обобщается на многомерный случай *d>2*.

Библиографический список

1. А.А. Новоселов. Воспроизведение дискретных распределений с заданной ковариационной структурой. Материалы II Межрегиональной конференции КГТЭИ, Красноярск, 1:229–234, 2009.
2. А.А. Новоселов. Дискретные распределения с заданной корреляцией, наименее отклоняющиеся от независимого. Труды XIII Международной конференции по эвентологической математике и смежным вопросам, Красноярск, КГТЭИ, СФУ, 2009, 126-131.
3. R.B. Nelsen. An Introduction To Copulas. Springer, 1998.

A.A. Novosyolov

BUILDING MULTIDIMENSIONAL DISCRETE DISTRIBUTIONS POSSESSING PREDEFINED CORRELATION STRUCTURE

The paper is devoted to methods of building multidimensional discrete distributions possessing predefined correlation structure and marginal distributions. The methods are based on mixing special distributions and solving some optimization problems.

Keywords: discrete distribution, correlation, copula, mixing