Дискретные распределения с заданной корреляцией, наименее уклоняющиеся от независимого

Аркадий Арсеньевич Новоселов
Сибирский федеральный университет
Красноярский государственный торгово-экономический институт
Красноярск
arcady@novosyolov.ru

Аннотация. *Воспроизведение многомерного нормального распределения с заданной корреляционной структурой является хорошо изученной задачей, и обычно осуществляется с использованием факторизации ковариационной матрицы. Для дискретных распределений такой метод неприменим. В настоящей работе описывается один метод построения двумерного дискретного распределения по заданным маргинальным распределениям и коэффициенту корреляции. Метод основан на поиске распределения, наименее уклоняющегося от независимого распределения.*

Ключевые слова. *Маргинальное распределение, класс Фреше, корреляция, дискретное многомерное распределение*

1 Введение

Пусть заданы нормальные распределения со средними значениями $\mu_1, \dots, \mu_d$ и дисперсиями $\sigma_1^2, \dots, \sigma_d^2$. Для произвольной корреляционной матрицы R существует единственное многомерное нормальное распределение, обладающее такими маргинальными распределениями и корреляционой матрицей. Хорошо известный алгоритм вопроизведения соответствующего случайного вектора $X = (X_1, \dots, X_d)'$ основан на факторизации ковариационной матрицы.

Обозначим

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

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

$$C = A'A, \tag{1}$$

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

$$X = A'Z. \tag{2}$$

Действительно, для Z справедливо $\mathbf{E}ZZ' = I$, где I — единичная матрица соответствующего размера, поэтому

$$\mathbf{E}XX' = \mathbf{E}(A'ZZ'A) = A'(\mathbf{E}ZZ')A = A'A = C,$$

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 работе [1] предложен метод воспроизведения двумерного дискретного распределения с заданными маргинальными распределениями и корреляцией, основанный на смесях некоторых базовых распределений. В настоящей работе предлагается метод решения такой же задачи, основанный на поиске распределения, наименее уклоняющегося от независимого распределения.

2 Описание двумерного дискретного распределения

Пусть $p = 2$. Обозначим

$$K = \{1, \dots, m\} \times \{1, \dots, n\}.$$

Дискретное распределение вектора $X = (X_1, X_2)'$ задается на прямоугольной сетке значений

$$\{x_{11}, \ldots, x_{1m}\} \times \{x_{21}, \ldots, x_{2n}\},$$

показанной на рис. 1, в виде

$$\mathbf{P}(X_1 = x_{1i};\ X_2 = x_{2j}) = r_{ij},\ (i,j) \in K.$$

Совместные вероятности

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

представлены на рис. 2. Здесь $\mathbf{P}(X_1 = x_{1i}) = p_i$, $i = 1, \ldots, m$ и $\mathbf{P}(X_2 = x_{2j}) = q_j$, $j = 1, \ldots, n$, так что векторы

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

описывают маргинальные распределения компонент. Средние значения

$$a_1 = \mathbf{E}X_1 = \sum_{i=1}^{m} x_{1i} p_i,\quad a_2 = \mathbf{E}X_2 = \sum_{j=1}^{n} x_{2j} q_j,$$

$$a_{12} = \mathbf{E}(X_1 X_2) = \sum_{i=1}^{m}\sum_{j=1}^{n} x_{1i} x_{2j} r_{ij},$$

стандартные отклонения

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

и коэффициент корреляции

$$\rho = \frac{\mathbf{E}(X_1 X_2) - \mathbf{E}X_1 \mathbf{E}X_2}{\sigma_1 \sigma_2}, \tag{5}$$

вычисляются, как обычно.

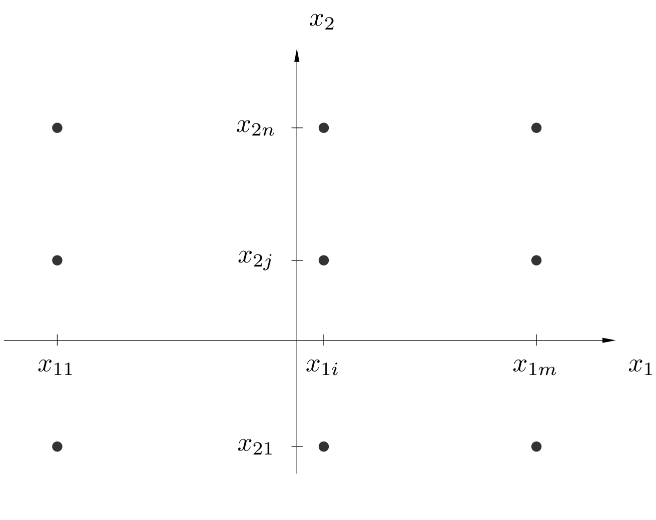

Рис. 1: Сетка значений вектора X

$$\begin{array}{cc|ccccc}
x_{21} & q_1 & r_{11} & \ldots & r_{i1} & \ldots & r_{m1} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
x_{2j} & q_j & r_{1j} & \ldots & r_{ij} & \ldots & r_{mj} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
x_{2n} & q_n & r_{1n} & \ldots & r_{in} & \ldots & r_{mn} \\
\hline
 & & p_1 & \ldots & p_i & \ldots & p_m \\
 & & x_{11} & \ldots & x_{1i} & \ldots & x_{1m}
\end{array}$$

Рис. 2: Совместное распределение компонент вектора X

Если маргинальные распределения (4) известны, то выполняются соотношения

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

и

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

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 ρ компонент X, то справедливо и уравнение

$$\sum_{i=1}^{m}\sum_{j=1}^{n} x_{1i} x_{2j} r_{ij} = \rho \sigma_1 \sigma_2 + a_1 a_2. \tag{8}$$

Обозначим $\mathcal{F}(p,q)$ класс всех двумерных распределений (3) с маргинальными распределениями (4) (его еще называют классом Фреше), а $\mathcal{F}_\rho(p,q)$ — его подкласс распределений с корреляцией ρ. Известно [2], что среди всех распределений в $\mathcal{F}(p,q)$ наименьшей корреляцией $\rho_{\min} = \rho_{\min}(p,q)$ обладает антикомонотонное распределение $R^-(p,q)$, а наибольшей корреляцией $\rho_{\max} = \rho_{\max}(p,q)$ обладает комонотонное распределение $R^+(p,q)$, причем, вообще говоря, $\rho_{\min} > -1$ и $\rho_{\max} < 1$. Класс Фреше представим в виде

$$\mathcal{F}(p,q) = \bigcup_{\rho \in [\rho_{\min}, \rho_{\max}]} \mathcal{F}_\rho(p,q). \tag{9}$$

Обозначим еще $R^\diamond(p,q)$ независимое распределение из класса Фреше $\mathcal{F}(p,q)$.

3 Примеры

В настоящем параграфе рассмотрим примеры различных двумерных случайных векторов, имеющих одинаковые маргинальные распределения и одинаковую корреляцию.

Пример 1. Пусть сетка значений случайного вектора X задана в виде

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

как показано на рис. 3. Здесь $m=3$, $n=2$.

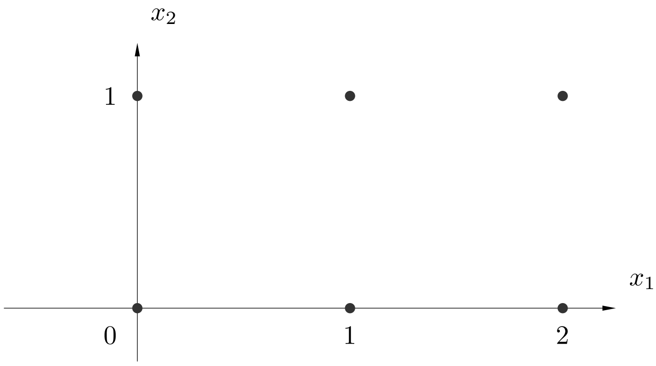

Рис. 3: Сетка значений вектора X, примеры 1, 2

Зададим маргинальные распределения компонент X_1 и X_2 посредством вероятностей

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

Средние значения и стандартные отклонения маргинальных распределений равны

$$a_1=1,\ a_2=\frac{1}{2},\ \sigma_1=\sqrt{\frac{2}{3}},\ \ \sigma_2=\frac{1}{2}.$$

Ограничения (6), (7) в данном случае задают 4 независимых условия, поэтому совместное распределение компонент X зависит от 2 произвольных параметров. Обозначим эти параметры u,v и запишем совместное распределение X в виде

$$\begin{array}{cc|ccc}
0 & 1/2 & u & v & 1/2-u-v\\
1 & 1/2 & 1/3-u & 1/3-v & u+v-1/6\\
\hline
 & & 1/3 & 1/3 & 1/3\\
 & & 0 & 1 & 2
\end{array} \tag{12}$$

Для того, чтобы (12) задавало совместное распределение, параметры u,v должны удовлетворять ограничениям

$$0\le u\le 1/3,\ \ 0\le v\le 1/3,\ \ 1/6\le u+v\le 1/2. \tag{13}$$

Область плоскости (u,v), задаваемая ограничениями (13), показана на рис. 4.

Зафиксируем произвольное значение корреляции ρ и изучим класс $\mathcal{F}_\rho(p,q)$ совместных распределений на сетке (10), обладающих маргинальными распределениями (11) и заданной корреляцией ρ. Уравнение (8) в этом случае дает дополнительное соотношение

$$v=\frac{1}{2}\left(1+\rho\sqrt{\frac{2}{3}}\right)-2u. \tag{14}$$

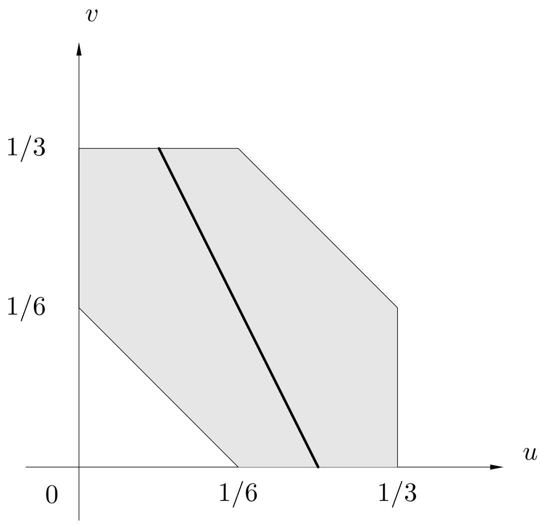

Рис. 4: Допустимая область в плоскости параметров (u,v) в примере 1, жирная линия соответствует некоррелированным компонентам

Прямая (14) пересекает допустимую область параметров (13) при значениях

$$\rho\in\left[-\sqrt{\frac{2}{3}},\sqrt{\frac{2}{3}}\right]. \tag{15}$$

Это соотношение и определяет допустимые значения коэффициента корреляции в данном примере. Комонотонному распределению соответствует значение корреляции $\rho_{\max}=\sqrt{2/3}$, а антикомонотонному — значение $\rho_{\min}=-\sqrt{2/3}$.

Пусть сначала $\rho=0$. Из (14) получаем связь между параметрами распределения

$$v=\frac{1}{2}-2u,\ \ \frac{1}{12}\le u\le\frac{1}{4}\,;$$

это множество пар значений параметров изображено на рис. 4 жирной линией. Соответствующее семейство $\mathcal{F}_0(p,q)$ совместных распределений имеет следующий вид:

$$\begin{array}{cc|ccc}
0 & 1/2 & u & 1/2-2u & u\\
1 & 1/2 & 1/3-u & 2u-1/6 & 1/3-u\\
\hline
 & & 1/3 & 1/3 & 1/3\\
 & & 0 & 1 & 2
\end{array} \tag{16}$$

где

$$\frac{1}{12}\le u\le\frac{1}{4}\,.$$

Далее приведены распределения из $\mathcal{F}_0(p,q)$, соответствующие граничным значениям $u = 1/12$ и $u = 1/4$.

0	1/2	1/12	1/3	1/12
1	1/2	1/4	0	1/4
		1/3	1/3	1/3
		0	1	2

0	1/2	1/4	0	1/4
1	1/2	1/12	1/3	1/12
		1/3	1/3	1/3
		0	1	2

Аналогично строятся семейства распределений $\mathcal{F}_\rho(p,q)$, обладающих другим фиксированным коэффициентом корреляции ρ, удовлетворяющим (15). Каждое такое семейство $\mathcal{F}_\rho(p,q)$ содержит континуум распределений, за исключением семейств $\mathcal{F}_{\rho_{\min}}(p,q)$ и $\mathcal{F}_{\rho_{\max}}(p,q)$, которые содержат по одному совместному распределению. Это антикомонотонное

0	1/2	0	1/6	1/3
1	1/2	1/3	1/6	0
		1/3	1/3	1/3
		0	1	2

и комонотонное

0	1/2	1/3	1/6	0
1	1/2	0	1/6	1/3
		1/3	1/3	1/3
		0	1	2

распределения, соответственно.

Пример 2. Теперь рассмотрим более асимметричный пример. Пусть на той же сетке значений (10) заданы маргинальные распределения

$$p = (1/3, 1/2, 1/6), \ \ q = (2/5, 3/5). \tag{17}$$

Основные характеристики маргинальных распределений равны

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

Двухпараметрическое представление класса Фреше имеет вид

0	2/5	u	v	$2/5 - u - v$
1	3/5	$1/3 - u$	$1/2 - v$	$u + v - 7/30$
		1/3	1/2	1/6
		0	1	2

(18)

Область допустимых значений параметров задается неравенствами

$$0 \le u \le 1/3, \ \ 0 \le v \le 1/2, \ \ 7/30 \le u + v \le 2/5 \tag{19}$$

и представлена на рис. 5. Окружностями на нем отмечены антикомонотонное $R^-(p,q)$

0	2/5	0	7/30	1/6
1	3/5	1/3	4/15	0
		1/3	1/2	1/6
		0	1	2

(20)

независимое $R^\diamond(p,q)$

0	2/5	2/15	1/5	1/15
1	3/5	1/5	3/10	1/10
		1/3	1/2	1/6
		0	1	2

(21)

и комонотонное $R^+(p,q)$

0	2/5	1/3	1/15	0
1	3/5	0	13/30	1/6
		1/3	1/2	1/6
		0	1	2

(22)

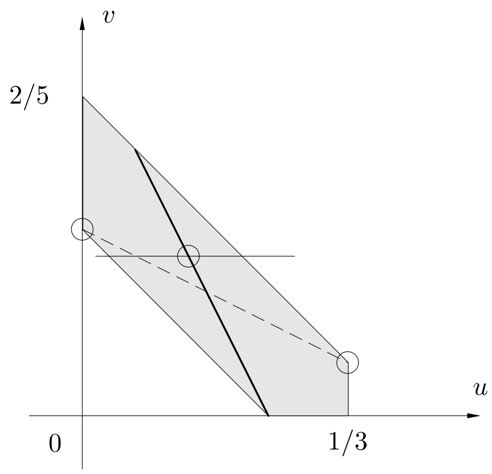

Рис. 5: Допустимая область в плоскости параметров (u, v) в примере 2, жирная линия соответствует некоррелированным компонентам; окружностями (слева направо) отмечены антикомонотонное, независимое и комонотонное распределения; горизонтальный отрезок изображает решения задачи 1 из §4 для примера 2, см. (33)

Класс $\mathcal{F}_\rho(p,q)$ описывается в параметрическом представлении уравнением

$$v = \frac{7}{15} + \rho\frac{\sqrt{102}}{30} - 2u. \tag{23}$$

Некоррелированные распределения лежат на отрезке прямой, описываемом уравнением

$$2u + v = \frac{7}{15}, \ \frac{1}{15} \le u \le \frac{7}{30}.$$

Коэффициент корреляции в данном примере заключен в интервале

$$\rho \in \left[-\frac{7}{\sqrt{102}}, \frac{8}{\sqrt{102}}\right]. \qquad (24)$$

4 Близость 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

$$f(r) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}(r_{ij} - p_i q_j)^2 \longrightarrow \min_r. \qquad (25)$$

Задача 1. В первой задаче целевая функция (25) минимизируется при ограничениях (6) — (8). Функция Лагранжа этой задачи имеет вид

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

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

$$\frac{\partial\mathcal{L}}{\partial r_{kl}} = (r_{kl} - p_k q_l) + \lambda_l + \mu_k + \nu x_{1k}x_{2l} = 0, \qquad (26)$$

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), (7) 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 (26) в виде функции множителей Лагранжа λ, μ, ν:

$$r_{ij} = p_i q_j - (\lambda_j + \mu_i + \nu x_{1i}x_{2j}),\ (i,j) \in K, \qquad (27)$$

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

$$-m\lambda_j - \sum_{i=1}^{m}\mu_i - \nu x_{2j}\sum_{i=1}^{m} x_{1i} = 0,\ j = 2,\dots,n, \qquad (28)$$

$$-\sum_{j=2}^{n}\lambda_j - n\mu_i - \nu x_{1i}\sum_{j=1}^{n} x_{2j} = 0,\ i = 1,\dots,m, \qquad (29)$$

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

Обозначив

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

преобразуем последнее уравнение к виду

$$-s_1\sum_{j=1}^{n}\lambda_j x_{2j} - s_2\sum_{i=1}^{m}\mu_i x_{1i} - \nu s_{12} = \rho\sigma_1\sigma_2. \qquad (30)$$

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

$$\bar{x}_2 = (x_{22},\dots,x_{2n})',\quad \bar{x}_1 = (x_{11},\dots,x_{1m})',$$

В системе уравнений (28) — (30), 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

$$A\gamma = b, \qquad (31)$$

где неизвестные множители Лагранжа обозначены

$$\gamma = (\lambda_2,\dots,\lambda_n,\mu_1,\dots,\mu_m,\nu)', \qquad (32)$$

а матрица системы A и вектор правых частей имеют вид

$$A = \begin{pmatrix} -mI_{n-1} & -J_{n-1,m} & -s_1\bar{x}_2 \\ -J_{m,n-1} & -nI_m & -s_2\bar{x}_1 \\ -s_1\bar{x}_2' & -s_2\bar{x}_1' & -s_{12}\end{pmatrix}$$

$$b = \Big(\ \overbrace{0\ \dots\ 0}^{n-1}\ \ \overbrace{0\ \dots\ 0}^{m}\ \ \overbrace{\rho\sigma_1\sigma_2}^{1}\ \Big)'.$$

Квадратная система линейных уравнений (31) имеет единственное решение (32), добавляя в которое значение $\lambda_1 = 0$, из (27) вычисляем искомое распределение.

Приведем решение первой задачи для нескольких значений корреляции в ситуации примера 2. Для $\rho = 0$, как и ожидалось, получаем независимое распределение (21). Для $\rho = -4/\sqrt{102}$ получаем распределение

$$\begin{array}{cc|ccc} 0 & 2/5 & 1/15 & 1/5 & 2/15 \\ 1 & 3/5 & 4/15 & 3/10 & 1/30 \\ \hline & & 1/3 & 1/2 & 1/6 \\ & & 0 & 1 & 2 \end{array}$$

Максимальному значению $\rho = \rho_{\max} = 8/\sqrt{102}$ соответствует решение

$$\begin{array}{cc|ccc} 0 & 2/5 & 4/15 & 1/5 & -1/15 \\ 1 & 3/5 & 1/15 & 3/10 & 7/30 \\ \hline & & 1/3 & 1/2 & 1/6 \\ & & 0 & 1 & 2 \end{array},$$

которое не является вероятностным распределением ввиду присутствия отрицательных элементов. Минимальному значению $\rho = \rho_{\min} = -7/\sqrt{102}$ соответствует решение

$$\begin{array}{cc|ccc} 0 & 2/5 & 1/60 & 1/5 & 11/60 \\ 1 & 3/5 & 19/60 & 3/10 & -1/60 \\ \hline & & 1/3 & 1/2 & 1/6 \\ & & 0 & 1 & 2 \end{array},$$

которое также не является вероятностным распределением ввиду присутствия отрицательных элементов. Поэтому естественно рассмотреть задачу с дополнительными ограничениями неотрицательности.

Отметим, что в данном примере нетрудно вывести явную зависимость решения задачи от ρ в терминах параметров u, v

$$u = \frac{\sqrt{102}}{60}\rho + \frac{2}{15},\ v = \frac{1}{5}. \tag{33}$$

Это множество изображено в виде горизонтального отрезка на рис. 5. Пара (u, v) лежит в допустимом множестве при

$$\rho \in \left[-\frac{6}{\sqrt{102}}, \frac{4}{\sqrt{102}}\right]. \tag{34}$$

Задача 2. Во второй задаче целевая функция (25) минимизируется при ограничениях (6) — (8) и

$$r_{ij} \geq 0,\ (i, j) \in K. \tag{35}$$

Аналитическое решение этой задачи в общем виде недоступно, однако численные методы позволяют эффективно решать ее. Приведем решения для примера 2. Отметим. что при выполнении условия (34) решения задач 1 и 2 совпадают. Поэтому интересно рассматривать решение задачи 2 при ρ, не удовлетворяющих условию (34).

Максимальному значению $\rho = \rho_{\max} = 8/\sqrt{102}$ соответствует решение

$$\begin{array}{cc|ccc} 0 & 2/5 & 1/3 & 1/15 & 0 \\ 1 & 3/5 & 0 & 13/30 & 1/6 \\ \hline & & 1/3 & 1/2 & 1/6 \\ & & 0 & 1 & 2 \end{array},$$

которое, как нетрудно заметить, представляет собой комонотонное распределение (22).

Минимальному значению $\rho = \rho_{\max} = -7/\sqrt{102}$ соответствует решение

$$\begin{array}{cc|ccc} 0 & 2/5 & 0 & 7/30 & 1/6 \\ 1 & 3/5 & 1/3 & 4/15 & 0 \\ \hline & & 1/3 & 1/2 & 1/6 \\ & & 0 & 1 & 2 \end{array},$$

представляющее собой антикомонотонное распределение (20).

Список литературы

[1] А.А. Новоселов. Воспроизведение дискретных распределений с заданной ковариационной структурой. *Материалы II Межрегиональной конференции КГТ-ЭИ, Красноярск*, 1:229–234, 2009.

[2] R.B. Nelsen. *An Introduction To Copulas.* Springer, 1998.

Arcady Novosyolov (Krasnoyarsk)

Discrete distributions with given correlation closest to the independent distribution

Abstract. *Building multidimensional normal distribution possessing predefined correlation structure is a well known problem. It is usually solved by factorization of covariance matrix. The method does not work in case of discrete distributions. The present paper describes one method of building a discrete distribution given its marginals and correlation coefficient. The method is based on minimizing the deviation from independent distribution.*

Keywords. *Marginal distribution, Frechet class, correlation, multidimensional discrete distribution.*