ОЦЕНКА КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ ВРЕМЕННЫХ РЯДОВ НА ОСНОВЕ ТРАНСПОРТНОЙ ЗАДАЧИ ЛИНЕЙНОГО ПРОГРАММИРОВАНИЯ

А.П. Котенко, М.Б. Букаренко

Самарский государственный аэрокосмический университет имени академика С.П. Королёва, г. Самара, Россия

Современные экономические системы характеризуются большим числом сложных взаимосвязей. При этом отдельные факторы являются чаще всего случайными величинами (СВ). Изучение их статистических распределений представляет важную задачу, которая многократно усложняется при необходимости учёта их взаимодействия.

При фиксированном распределении отдельных факторов полезно знать границы их коррелированности. Покажем, как в этом случае найти границы коэффициента корреляции по заданным частным распределениям.

Рассмотрим две неотрицательных дискретных случайных величины X и Y с конечными рядами распределений:

X	x_1	…	x_i	…	x_n
P_X	p_{1*}	…	p_{i*}	…	p_{n*}

,

Y	y_1	…	y_j	…	y_m
P_Y	p_{*1}	…	p_{*j}	…	p_{*m}

,

$0 \le x_1 < x_2 < \ldots < x_i < \ldots < x_n < +\infty$, $0 \le y_1 < y_2 < \ldots < y_j < \ldots < y_m < +\infty$.

Тогда $0 < MX < +\infty$, $0 < MY < +\infty$, $0 < \sigma_X < +\infty$, $0 < \sigma_Y < +\infty$. Представим корреляционное поле таблицей 1.

Таблица 1. Совместное распределение двумерной дискретной случайной величины.

(X,Y)		y_1	…	y_j	…	y_m
	$P_X \setminus P_Y$	p_{*1}	…	p_{*j}	…	p_{*m}
x_1	p_{1*}	p_{11}	…	p_{1j}	…	p_{1m}
…	…	…	…	…	…	…
x_i	p_{i*}	p_{i1}	…	p_{ij}	…	p_{im}
…	…	…	…	…	…	…
x_n	p_{n*}	p_{n1}	…	p_{nj}	…	p_{nm}

Заметим, что $p_{ij}=p_{i*}\cdot p_{*j}$ при независимости случайных величин X и Y; при этом $r_{XY} = \dfrac{\operatorname{cov}(X,Y)}{\sigma_X \cdot \sigma_Y} = 0$.

Пример. Рассмотрим два фиксированных дискретных распределения:

X	0	1	3
P_X	0,25	0,5	0,25

, $\mathrm{M}X=\frac{5}{4}$, $\sigma_X^2=\frac{19}{16}$;

Y	0	1	5	6
P_Y	0,25	0,25	0,25	0,25

, $\mathrm{M}Y=3$, $\sigma_Y^2=\frac{26}{4}$.

Нулевую корреляцию r_{XY}=0 получим для независимых СВ X и Y при совместном распределении, заданном таблицей 2.

Таблица 2. Совместное распределение двумерной случайной величины (X,Y) при отсутствии корреляции.

(X,Y)		0	1	5	6
	$P_X \setminus P_Y$	0,25	0,25	0,25	0,25
0	0,25	0,0625	0,0625	0,0625	0,0625
1	0,5	0,125	0,125	0,125	0,125
3	0,25	0,0625	0,0625	0,0625	0,0625

Поставим задачу: найти совместное распределение $P_{(X,Y)}$, доставляющее максимальное (минимальное) значение коэффициента корреляции r_{XY}. Поскольку дисперсии $\sigma_X^2>0$, $\sigma_Y^2>0$ заданы, задача сводится к определению экстремальных значений ковариации $\mathrm{cov}(X,Y)=\mathrm{M}_{P_{(X,Y)}}XY-\mathrm{M}X\cdot\mathrm{M}Y$, которая при заданных математических ожиданиях MX, MY превращается в задачу

$$\mathrm{M}_{P_{(X,Y)}}XY\to\underset{P_{(X,Y)}}{\mathrm{extr}}.$$

Решим её с помощью двух транспортных задач линейного программирования. Неизвестный оптимальный план $\overline{P}:=\left(p_{ij}\right)_{i\in\overline{1,n};\,j\in\overline{1,m}}\in[0,1]^{nm}$ удовлетворяет $n\cdot m$ условиям неотрицательности $p_{ij}\geq 0$ и системе $n+m$ линейных ограничений в виде равенств $\sum_{j=1}^{m}p_{ij}=p_{i*}$, $1\leq i\leq n$; $\sum_{i=1}^{n}p_{ij}=p_{*j}$, $1\leq j\leq m$, из которых вытекает ограниченность вероятности совместного распределения $p_{ij}\leq 1$.

В качестве целевой функции возьмём

$$F\left(P_{(X,Y)}\right):=\mathrm{M}_{P_{(X,Y)}}XY=\sum_{i=1}^{n}\sum_{j=1}^{m}p_{ij}c_{ij}\to\max(\min)$$

с удельными транспортными затратами $c_{ij}:=x_iy_j\geq 0$.

Задача максимизации корреляции $\Leftrightarrow\left(\mathrm{M}_{P_{(X,Y)}}XY\to\max\limits_{P_{(X,Y)}}\right)$.

Начальный план $\overline{P}_0$ определим методом «северо-западного угла»,

поскольку максимальное значение $r_{XY}\approx 1$ соответствует сонаправленному изменению значений носителей величин X и Y. Для СВ X и Y получим следующую замкнутую транспортную задачу максимизации (табл. 3).

Таблица 3. Совместное распределение для случая строгой положительной корреляции.

$x_i \setminus y_j$	0	1	5	6	p_{i*}
0	0,25	×	×	×	0,25
1	0	0,25	0,25	×	0,5
3	×	×	0	0,25	0,25
p_{*j}	0,25	0,25	0,25	0,25	1\1

Её начальный опорный план $\overline{P}_0^{\max}(0{,}25;0;0;0;0;0{,}25;0{,}25;0;0;0;0;0{,}25)$ – единственный оптимальный (проверяется методом потенциалов). Тогда максимальное значение ковариации $\max\limits_{P_{(X,Y)}} \operatorname{cov}(X,Y) = 2{,}25$. Поэтому максимальное значение корреляции $\max\limits_{P_{(X,Y)}} r_{XY} = 18/\sqrt{19\cdot 26} \approx 0{,}81$.

Задача минимизации корреляции $\Leftrightarrow \left(\mathrm{M}_{P_{(X,Y)}} XY \to \min\limits_{P_{(X,Y)}} \right)$.

Начальный план $\overline{P}_0$ определим методом «северо-восточного угла», поскольку минимальное значение $r_{XY}\approx -1$ соответствует противоположному изменению значений носителей величин X и Y. Для СВ X и Y получим следующую замкнутую транспортную задачу минимизации (табл. 4).

Таблица 4. Совместное распределение для случая строгой отрицательной корреляции.

$x_i \setminus y_j$	0	1	5	6	p_{i*}
0	×	×	×	0,25	0,25
1	0	0,25	0,25	0	0,5
3	0,25	×	×	×	0,25
p_{*j}	0,25	0,25	0,25	0,25	1\1

Её начальный опорный план $\overline{P}_0^{\min}(0;0;0;0{,}25;0;0{,}25;0{,}25;0;0{,}25;0;0;0)$ – единственный оптимальный (проверяется методом потенциалов). Тогда минимальное значение ковариации $\min\limits_{P_{(X,Y)}} \operatorname{cov}(X,Y) = -2{,}25$. Поэтому минимальное значение корреляции $\min\limits_{P_{(X,Y)}} r_{XY} = -18/\sqrt{19\cdot 26} \approx -0{,}81$.

Таким образом, для любых СВ X_1 и Y_1 с заданными рядами распределений

$$\operatorname{cov}(X_1, Y_1) \in \left[-\frac{18}{\sqrt{494}}, \frac{18}{\sqrt{494}} \right].$$

Это позволяет сделать вывод о возможности сильной (но не строгой) корреляционной зависимости факторов X и Y, что позволяет, к примеру, совместно рассматривать их в модели множественной регрессии.

В общем случае, для любого значения $a \in \left(\min\limits_{P_{(X,Y)}} r_{XY}, \max\limits_{P_{(X,Y)}} r_{XY} \right)$ найдётся по крайней мере одна пара СВ $X_1(a)$ и $Y_1(a)$ с теми же заданными распределениями и корреляцией:

$$r_{X_1(a)Y_1(a)} = \frac{\operatorname{cov}(X_1(a), Y_1(a))}{\sigma_X \cdot \sigma_Y} = \frac{\mathrm{M}_{P_{(X_1(a),Y_1(a))}} X_1(a) Y_1(a) - \mathrm{M}X \cdot \mathrm{M}Y}{\sigma_X \cdot \sigma_Y} = a.$$

Представим значение a как точку числового интервала:

$$a = \min_{P_{(X,Y)}} r_{XY} + \lambda \left(\max_{P_{(X,Y)}} r_{XY} - \min_{P_{(X,Y)}} r_{XY} \right), \text{ где } \lambda := \frac{a - \min\limits_{P_{(X,Y)}} r_{XY}}{\max\limits_{P_{(X,Y)}} r_{XY} - \min\limits_{P_{(X,Y)}} r_{XY}} \in (0,1).$$

Тогда в силу выпуклости множества допустимых планов рассмотренных транспортных задач допустимым будет план $\overline{P} := \overline{P}_0^{\min} + \lambda \left(\overline{P}_0^{\max} - \overline{P}_0^{\min} \right)$, для которого $F(\overline{P}) = F\left(\overline{P}_0^{\min}\right) + \lambda \left[F\left(\overline{P}_0^{\max}\right) - F\left(\overline{P}_0^{\min}\right) \right] = \mathrm{M}_{\overline{P}_0^{\min}} XY + \lambda \left[\mathrm{M}_{\overline{P}_0^{\max}} XY - \mathrm{M}_{\overline{P}_0^{\min}} XY \right]$.

Этот план обеспечит заданное значение корреляции a.

Предложенное решение может быть обобщено на случай произвольного числа взаимодействующих дискретных случайных величин.

Приведённый метод оценки коэффициента парной корреляции позволяет по частным распределениям факторов оценить возможность их сильной коррелированности без привлечения экспериментов о совместном распределении. Это позволяет, к примеру, априорно оценить возможность совместного использования факторов в модели множественной регрессии.