Канонический корреляционный анализ

Н.П. Алексеева,

СПбГУ, математико-механический факультет

2020 г.

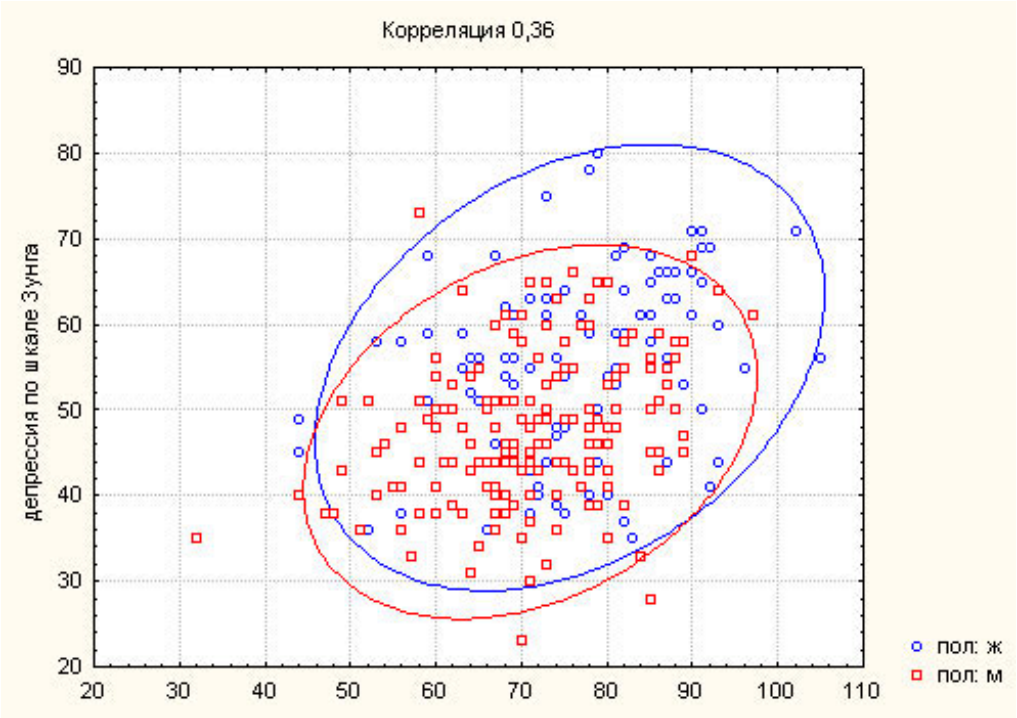
Структура анализа корреляционных зависимостей между метрическими переменными

- Коэффициенты корреляции Пирсона, Спирмена, Кендала между двумя признаками $(x_j, y_j), j = 1, \ldots, n$.
- Множественный коэффициент корреляции между переменной, представленной выборкой $\{y_j\}$, и совокупностью $\{x_{1j}, \ldots x_{pj}\}$, $j = 1, 2, \ldots, n$.
- Канонический корреляционный анализ зависимости двумя совокупностями переменных, представленных выборками $\{x_{1j}, \ldots x_{pj}\}$, $\{y_{1j}, \ldots y_{kj}\}$, $j = 1, 2, \ldots, n$.

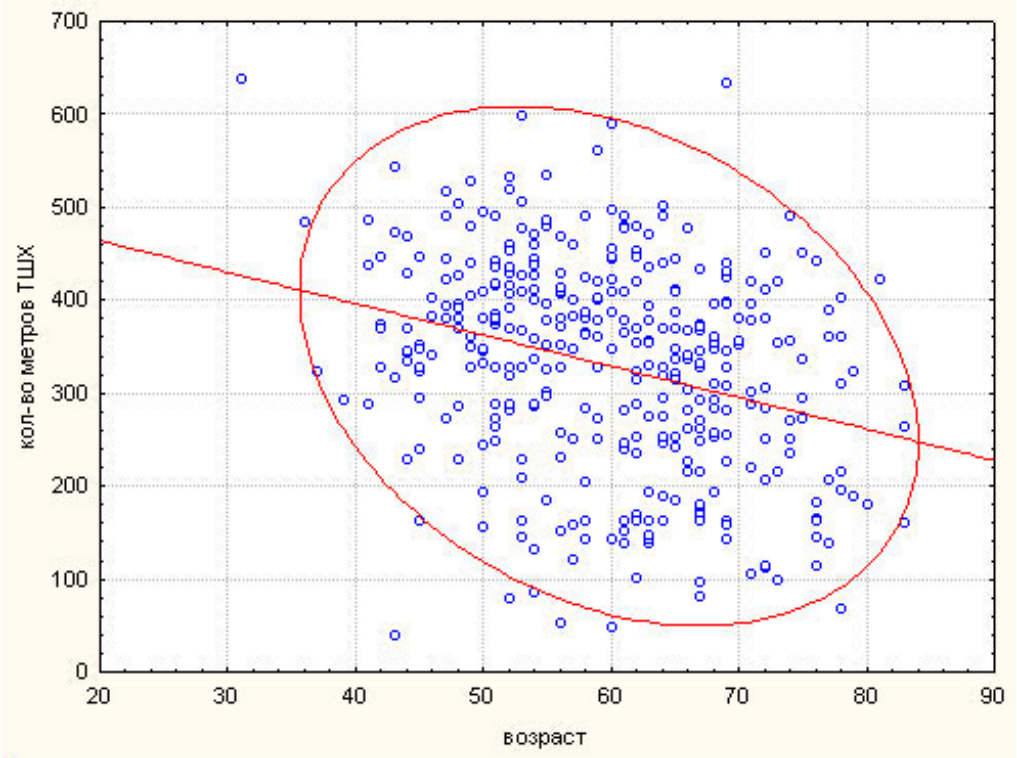
Корреляция Пирсона между двумя признаками $(x_j, y_j), j = 1, \ldots, n$.

$$r = \frac{\sum_{j=1}^{n} x_j y_j - n\bar{x}\bar{y}}{\sqrt{\sum_{j=1}^{n} x_j^2 - n\bar{x}^2}\sqrt{\sum_{j=1}^{n} y_j^2 - n\bar{y}^2}}$$

При $H_0 : \varrho = 0$ статистика $t = \frac{r}{\sqrt{1 - r^2}}\sqrt{n-2} \sim T(n-2)$.



Пример отрицательной зависимости



Корреляция Спирмена

Пусть $x'_1, \ldots, x'_n$ и $y'_1, \ldots, y'_n$ – ранги (порядковый номер элемента в упорядоченной выборке) индивидов по признакам X и Y,

k_i – количество индивидов, которых превосходит x_i,
l_i– количество индивидов, которые превосходят x_i,
$x_{i*} = k_i - l_i = 2(x'_i - \frac{n+1}{2})$.

ранг	1	2	3	4	5	6	7	8	9	10
Масса до диеты	57	66	68	70	76	78	79	80	81	92
x_{j*}	-9	-7	-5	-3	-1	1	3	5	7	9

$$k_3 = 2, l_3 = 7, x_{3*} = 2 - 7 = -5 = 2\left(3 - \tfrac{11}{2}\right) = 2 \cdot (-2.5)$$

Коэффициент корреляции Спирмена R_S

$$x_{i*} = k_i - l_i = 2(x'_i - \tfrac{n+1}{2}),\ y_{i*} = 2(y'_i - \tfrac{n+1}{2}).$$

$$Q_{xy} = \sum_{i=1}^{n} x_{i*}y_{i*},\ Q = Q_{xx} = \sum_{i=1}^{n} x_{i*}^2 = Q_{yy} = \sum_{i=1}^{n} y_{i*}^2$$

$$R_S = \frac{Q_{xy}}{Q} = \frac{Q_{xy}}{\sqrt{Q_{xx}}\sqrt{Q_{yy}}}$$

При полной положительной коррелиованности $Q_{xy} = Q$, $R_S = 1$.

$$\begin{array}{c} X \\ Y \end{array} \begin{pmatrix} 1 & 2 & \dots & n-1 & n \\ 1 & 2 & \dots & n-1 & n \end{pmatrix}$$

При полной отрицательной коррелиованности $Q_{xy} = -Q$, $R_S = -1$.

$$\begin{array}{c} X \\ Y \end{array} \begin{pmatrix} 1 & 2 & \dots & n-1 & n \\ n & n-1 & \dots & 2 & 1 \end{pmatrix}$$

Коэффициент корреляции Спирмена как выборочная корреляция между рангами

$$R_S = \frac{Q_{xy}}{\sqrt{Q_{xx}}\sqrt{Q_{yy}}} = \frac{\sum\limits_{i=1}^{n} 2(x'_i - \frac{n+1}{2})2(y'_i - \frac{n+1}{2})}{\sqrt{\sum\limits_{i=1}^{n} 2(x'_i - \frac{n+1}{2})}\sqrt{\sum\limits_{i=1}^{n} 2(y'_i - \frac{n+1}{2})}}$$

Так как $1 + 2 + \ldots + n = \frac{n(n+1)}{2}$, то среднее арифметическое последовательности $1, 2, \ldots, n$ равно $\frac{n+1}{2}$.

Значимость коэффициента корреляции Спирмена

$$H_0 : R_S = 0$$

$$T = \frac{R_S}{\sqrt{1 - R_S^2}}\sqrt{n-2} \sim T(n-2)$$

Масса до диеты	68	80	92	81	70	79	78	66	57	76
Масса после диеты	60	84	87	79	74	71	72	67	57	70

cor.test(x,y,method="spearman")

Spearman's rank correlation rho
data: x and y
S = 18, p-value = 0.00138
alternative hypothesis: true rho is not equal to 0
sample estimates:
rho
0.8909091

Коэффициент корреляция Кендала

$x_1, \ldots, x_n$ и $y_1, \ldots, y_n$

$$x_{ik} = \begin{cases} 1, & x_i < x_k, \\ 0, & x_i = x_k, \\ -1, & x_i > x_k \end{cases} \quad \text{и} \quad y_{ik} = \begin{cases} 1, & y_i < y_k, \\ 0, & y_i = y_k, \\ -1, & y_i > y_k \end{cases}$$

и введем показатель однонаправленности неравенств

$$S = \sum_{i=1}^{n} \sum_{k=i+1}^{n} x_{ik} y_{ik} \, .$$

Поскольку число всевозможных пар равно $N = C_n^2 = \frac{n(n-1)}{2}$, то доля однонаправленных пар, называемая коэффициентом корреляции Кендалла, равна $T = \frac{S}{N}$.

Пример корреляции Кендала

$$\begin{array}{c} X \\ Y \end{array} \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 4 & 1 & 3 & 2 \end{pmatrix}.$$

ik	12	13	14	15	23	24	25	34	35	45
x_{ik}	1	1	1	1	1	1	1	1	1	1
y_{ik}	-1	-1	-1	-1	-1	-1	-1	1	1	-1
$x_{ik}y_{ik}$	-1	-1	-1	-1	-1	-1	-1	1	1	-1

$S_+ = 2$ и $S_- = 8$ количество 1 и -1 соответственно

$$S = 1 \cdot S_+ + (-1) \cdot S_- = S_+ - S_- = S_+ - (N - S_+) = 2S_+ - N.$$

S_+ равен количеству элементов выборки Y, больших y_m по всем m при упорядочивании их по признаку X.

$$T = \frac{S}{N} = \frac{2S_+ - N}{N} = \frac{2S_+}{N} - 1\,.$$

Значимость корреляции Кендала

Для проверки значимости отклонения от нуля коэффициента корреляции Кендалла используется то, что при независимости признаков

$$\mathrm{T} \sim \mathcal{N}(0, \sigma), \text{ где } \sigma^2 = \frac{2(2\mathrm{n}+5)}{9\mathrm{n}(\mathrm{n}-1)}.$$

Например, $\mathrm{T}_* = -0.6$, $\mathrm{n} = 5$, $\sigma^2 = \frac{2(2\cdot 5+5)}{9\cdot 5\cdot 4} = \frac{1}{6}$, $\sigma \approx 0.4$. При альтернативе отрицательной корреляции
$\alpha_* = \mathrm{P}\{\mathrm{T} < \mathrm{T}_*\} = \mathrm{pnorm}(-0.6; 0; 0.4) = 0.07$, отсюда отрицательную зависимость нельзя объяснить случайностью при $\alpha > 0.07$.

Канонический корреляционный анализ. Постановка задачи

Два набора признаков $X_1, \ldots, X_r$ и $X_{r+1}, \ldots, X_{r+s}$.
Математическая задача здесь состоит в том, чтобы найти такие нормированные линейные комбинации

$$U = \beta_0 + \beta_1 X_1 + \ldots + \beta_r X_r \quad \text{и} \quad V = \alpha_0 + \alpha_1 X_{r+1} + \ldots + \alpha_s X_{r+s},$$

чтобы корреляция $R = \mathrm{cor}(U, V)$ между U и V была максимальной.
Случайные величины U и V, а также их коэффициенты называются каноническими.

Метод нахождения канонических коэффициентов

$$X = (X_1, \ldots, X_r, X_{r+1}, \ldots, X_{r+s})^T$$

Не умаляя общности, $EX_i = 0$, $i = 1, 2, \ldots, r+s$.
Обозначим через $U_1 = (X_1, \ldots, X_r)^T$, $U_2 = (X_{r+1}, \ldots, X_{r+s})^T$.
$\Sigma_{11} = EU_1U_1^T$, $\Sigma_{22} = EU_2U_2^T$ и $\Sigma_{12} = EU_1U_2^T$, $\Sigma_{21} = EU_2U_1^T$, $\Sigma_{12} = \Sigma_{21}$

$$\Sigma = EXX^T = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Канонические случайные величины $V_1 = L^T U_1$, $V_2 = M^T U_2$, где
$L = (l_1, \ldots, l_r)^T$, $M = (m_1, \ldots, m_s)^T$ канонические коэффициенты.
Поскольку предполагается нормированность канонических величин

$$DV_1 = EV_1^2 = EL^T U_1 U_1^T L = L^T \Sigma_{11} L = 1,$$
$$DV_2 = EV_2^2 = EM^T U_2 U_2^T M = M^T \Sigma_{22} M = 1,$$
$$\operatorname{cov}(V_1, V_2) = EV_1V_2 = EL^T U_1 U_2^T M = L^T \Sigma_{12} M = M^T \Sigma_{21} L,$$

то экстремальная задача нахождения L и M таких, что

$$\operatorname{cor}(V_1, V_2) = \operatorname{cov}(V_1, V_2) \to \max \quad \text{при} \quad DV_1 = DV_2 = 1 \qquad (1)$$

сводится к задаче:

$$L^T \Sigma_{12} M \to \max \quad \text{при} \quad L^T \Sigma_{11} L = 1 \text{ и } M^T \Sigma_{22} M = 1. \qquad (2)$$

Для решения задачи

$$\mathrm{L}^{\mathrm{T}}\Sigma_{12}\mathrm{M} \;\to\; \max \quad \text{при} \quad \mathrm{L}^{\mathrm{T}}\Sigma_{11}\mathrm{L} = 1 \text{ и } \mathrm{M}^{\mathrm{T}}\Sigma_{22}\mathrm{M} = 1.$$

введем функцию Лагранжа

$$\mathcal{L}(\mathrm{L},\mathrm{M}) = \mathrm{L}^{\mathrm{T}}\Sigma_{12}\mathrm{M} - \frac{\lambda_1}{2}\left(\mathrm{L}^{\mathrm{T}}\Sigma_{11}\mathrm{L} - 1\right) - \frac{\lambda_2}{2}\left(\mathrm{M}^{\mathrm{T}}\Sigma_{22}\mathrm{M} - 1\right)$$

и построим систему нормальных уравнений:

$$\begin{cases} \Sigma_{12}\mathrm{M} - \lambda_1\Sigma_{11}\mathrm{L} = 0 \\ \Sigma_{21}\mathrm{L} - \lambda_2\Sigma_{22}\mathrm{M} = 0 \\ \mathrm{L}^{\mathrm{T}}\Sigma_{11}\mathrm{L} = 1 \\ \mathrm{M}^{\mathrm{T}}\Sigma_{22}\mathrm{M} = 1 \end{cases}$$

Умножая первое уравнение на L^{T}, а второе на M^{T}, получаем

$$\begin{cases} \Sigma_{12}\mathrm{M} - \lambda_1\Sigma_{11}\mathrm{L} = 0 \\ \Sigma_{21}\mathrm{L} - \lambda_2\Sigma_{22}\mathrm{M} = 0 \\ \mathrm{L}^{\mathrm{T}}\Sigma_{11}\mathrm{L} = 1 \\ \mathrm{M}^{\mathrm{T}}\Sigma_{22}\mathrm{M} = 1 \end{cases} \iff \begin{cases} \mathrm{L}^{\mathrm{T}}\Sigma_{12}\mathrm{M} - \lambda_1\mathrm{L}^{\mathrm{T}}\Sigma_{11}\mathrm{L} = 0 \\ \mathrm{M}^{\mathrm{T}}\Sigma_{21}\mathrm{L} - \lambda_2\mathrm{M}^{\mathrm{T}}\Sigma_{22}\mathrm{M} = 0 \\ \mathrm{L}^{\mathrm{T}}\Sigma_{11}\mathrm{L} = 1 \\ \mathrm{M}^{\mathrm{T}}\Sigma_{22}\mathrm{M} = 1 \end{cases}$$

$$\iff \quad \lambda_1 = \lambda_2 = \mathrm{cor}(\mathrm{V}_1, \mathrm{V}_2) = \varrho\,.$$

Итак, из $\lambda_1 = \lambda_2 = \varrho$

$$\begin{cases} \Sigma_{12}\mathrm{M} - \lambda_1\Sigma_{11}\mathrm{L} = 0 \\ \Sigma_{21}\mathrm{L} - \lambda_2\Sigma_{22}\mathrm{M} = 0 \end{cases} \Longleftrightarrow \begin{cases} \Sigma_{12}\mathrm{M} - \varrho\Sigma_{11}\mathrm{L} = 0 \\ \Sigma_{21}\mathrm{L} - \varrho\Sigma_{22}\mathrm{M} = 0 \end{cases}. \quad (3)$$

Умножим первое уравнение на $\Sigma_{21}\Sigma_{11}^{-1}$, а второе на ϱ, и сложим уравнения:

$$\begin{cases} \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\mathrm{M} - \varrho\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{11}\mathrm{L} = 0 \\ \varrho\Sigma_{21}\mathrm{L} - \varrho^2\Sigma_{22}\mathrm{M} = 0 \end{cases}$$
$$\Longrightarrow \quad (\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} - \varrho^2\Sigma_{22})\mathrm{M} = 0,$$

откуда получаем, что ϱ^2 является собственным корнем, а M собственным вектором уравнения

$$|\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} - \varrho^2\Sigma_{22}| = 0.$$

Аналогично в случае другого множества величин

$$(\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \varrho^2\Sigma_{11})\mathrm{L} = 0,$$

и ϱ^2 является собственным корнем, а L собственным вектором уравнения

$$|\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - \varrho^2\Sigma_{11}| = 0.$$

Таким образом, имеем $\mathrm{m} \leq \min(\mathrm{r}, \mathrm{s})$ канонических случайных величин $\mathrm{V_{1i}} = \mathrm{L_i^T U_2}$ и $\mathrm{V_{2i}} = \mathrm{M_i^T U_2}$, упорядоченных в соответствии с корнями $\varrho_1^2 \geq \varrho_2^2 \geq \ldots \geq \varrho_\mathrm{m}^2$. Далее для удобства будем считать канонические корреляции ϱ_i положительными, иначе можно поменять знаки канонических коэффициентов M или L.

Свойства канонических случайных величин

$$\operatorname{cov}(V_{1i}, V_{1j}) = \begin{cases} 1, & \text{при } i = j, \\ 0, & \text{при } i \neq j \end{cases},$$

$$\operatorname{cov}(V_{2i}, V_{2j}) = \begin{cases} 1, & \text{при } i = j, \\ 0, & \text{при } i \neq j \end{cases}.$$

$$\operatorname{cov}(V_{1i}, V_{2j}) = \begin{cases} \varrho_i, & \text{при } i = j \leq m, \\ 0, & \text{при } i \neq j \text{ или } i = j > m \end{cases}$$

Характеристическое уравнение $(\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} - \varrho^2\Sigma_{22})\mathrm{M} = 0$
Пусть $\mathrm{A} = \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, $\mathrm{M_i}$ и $\mathrm{M_j}$ собственные векторы, соответствующие собственным числам $\varrho_i^2 \neq \varrho_j^2$.

$$\begin{cases} (\mathrm{A} - \varrho_i^2\Sigma_{22})\mathrm{M_i} = 0 \\ (\mathrm{A} - \varrho_j^2\Sigma_{22})\mathrm{M_j} = 0 \end{cases}$$

первое уравнение умножим на $-\mathrm{M_j^T}$, второе на $\mathrm{M_i^T}$, и уравнения сложим:

$$\begin{cases} -\mathrm{M_j^T A M_i} + \varrho_i^2\mathrm{M_j^T}\Sigma_{22}\mathrm{M_i} = 0 \\ \mathrm{M_i^T A M_j} - \varrho_j^2\mathrm{M_i^T}\Sigma_{22}\mathrm{M_j} = 0 \end{cases} \quad \Longrightarrow \quad (\varrho_i^2 - \varrho_j^2)\mathrm{M_i^T}\Sigma_{22}\mathrm{M_j} = 0,$$

$$\mathrm{cov}(\mathrm{V_{2i}}, \mathrm{V_{2j}}) = \mathrm{EV_{2i}V_{2j}} = \mathrm{EM_i^T U_2 U_2^T M_j} = \mathrm{M_i^T}\Sigma_{22}\mathrm{M_j} = 0.$$

Аналогично

$$\mathrm{cov}(\mathrm{V_{1i}}, \mathrm{V_{1j}}) = \mathrm{EV_{1i}V_{1j}} = \mathrm{EL_i^T U_1 U_1^T L_j} = \mathrm{L_i^T}\Sigma_{11}\mathrm{L_j} = 0.$$

$$\mathrm{cov}(\mathrm{V_{1i}}, \mathrm{V_{2j}}) = \mathrm{E}\mathrm{L_1^T}\mathrm{U_1}\mathrm{U_2^T}\mathrm{M_j} = \mathrm{L_i^T}\mathbf{\Sigma}_{12}\mathrm{M_j}$$

Рассмотрим первое уравнение $\mathbf{\Sigma}_{12}\mathrm{M_i} - \varrho_\mathrm{i}\mathbf{\Sigma}_{11}\mathrm{L_i} = 0$ из системы (3) в случае $\varrho_\mathrm{i} \neq 0$ и умножим его обе части на $\mathrm{L_i^T}$:

$$\mathbf{\Sigma}_{12}\mathrm{M_i} = \varrho_\mathrm{i}\mathbf{\Sigma}_{11}\mathrm{L_i} \iff \mathrm{L_i^T}\mathbf{\Sigma}_{12}\mathrm{M_i} = \varrho_\mathrm{i}\mathrm{L_i^T}\mathbf{\Sigma}_{11}\mathrm{L_i} = \varrho_\mathrm{i},$$

при умножении на $\mathrm{L_j^T}$ получим $\mathrm{L_j^T}\mathbf{\Sigma}_{12}\mathrm{M_i} = \varrho_\mathrm{i}\mathrm{L_j^T}\mathbf{\Sigma}_{11}\mathrm{L_i} = 0$. В случае $\varrho_\mathrm{i} = 0$ $\mathbf{\Sigma}_{12}\mathrm{M_i} = 0 \iff \mathrm{L_i^T}\mathbf{\Sigma}_{12}\mathrm{M_i} = 0$.

Проверка значимости канонических корреляций

Бартлетту принадлежит критерий, основанный на распределении логарифма s сомножителей $\prod_{j=1}^{s}(1-\varrho_j^2)$. Для определенности считаем, что $r \leq s$, иначе группы просто меняются местами. Если предположить, что k канонических корреляций не равны нулю, то статистика критерия для проверки того, что остальные равны нулю, такова:

$$-\left(n-k-1-\frac{r+s+1}{2}+\sum_{j=1}^{k}\varrho_j^2\right)\ln\left(\prod_{j=k+1}^{r}(1-\varrho_j^2)\right)\sim$$

$$\sim\chi^2((r-k)(s-k)).$$

Пример канонических корреляций

Исследуем корреляционную структуру между психологическими признаками и показателями тяжести наркотической зависимости.

Индекс тяжести	психологические характеристики
X_1 – медицинский статус	Y_1 – рискованность
X_2 – наркотический статус	Y_2 – уровень депрессии
X_3 – проблемы с законом	Y_3 – уровень тревоги
X_4 – психиатрический статус	

Канонические факторные нагрузки

	R1 = 0.352	$R_2 = 0.193$	$R_3 = 0.143$
мед.ст.	-0.641	0.298	-0.707
нарк.ст.	-0.545	0.199	0.390
проблемы с законом	0.296	-0.628	-0.227
психиатр. ст.	-0.663	-0.704	0.111
рискованность	-0.091	-0.934	0.346
депрессия	-0.735	-0.243	-0.633
тревожность	-0.961	0.182	0.209