1. Канонические корреляции

1.1. Постановка задачи

Обычные коэффициенты корреляции используются для выявления линейной зависимости между двумя признаками X и Y. Если нужно выявить зависимость между признаком X_0 и набором признаков $X_1, \ldots, X_p$, то в качестве характеристики этой зависимости рассматривается множественный коэффициент корреляции, равный коэффициенту корреляции между X_0 и его наилучшим линейным предсказанием $\widehat{X_0} = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$.

Этот принцип можно распространить на выявление зависимости между двумя наборами 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

$$U_1 = \beta_{10} + \beta_{11} X_1 + \ldots + \beta_{1r} X_r \quad \text{и} \quad V_1 = \alpha_{10} + \alpha_{11} X_{r+1} + \ldots + \alpha_{1s} X_{r+s},$$

чтобы корреляция $R_1 = cor(U_1, V_1)$ между U_1 и V_1 была максимальной. Случайные величины U_1 и V_1, а также их коэффициенты называются *каноническими.*

1.2. Метод нахождения канонических коэффициентов

Не умаляя общности, будем считать центрированными компоненты случайного вектора

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

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

$EU_2U_1^T$, $\Sigma_{12} = \Sigma_{21}$ ковариационные матрицы подвекторов, образующие общую ковариационную матрицу

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

Канонические случайные величины обозначим через $V_1 = L^T U_1$, $V_2 = M^T U_2$, где $L = (l_1, \ldots, l_r)^T$, $M = (m_1, \ldots, m_s)^T$ векторы канонических коэффициентов. Поскольку

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

то экстремальная задача нахождения L и M таких, что

$$cor(V_1, V_2) = cov(V_1, V_2) \ \rightarrow \ max \quad \text{при} \quad DV_1 = DV_2 = 1 \tag{1}$$

сводится к задаче:

$$L^T \Sigma_{12} M \ \rightarrow \ max \quad \text{при} \quad L^T \Sigma_{11} L = 1 \text{ и } M^T \Sigma_{22} M = 1. \tag{2}$$

Для решения этой задачи используем метод множителей Лагранжа. Введем функцию Лагранжа

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

и построим систему нормальных уравнений:

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

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

$$\begin{cases} L^T\Sigma_{12}M - \lambda_1 L^T\Sigma_{11}L = 0 \\ M^T\Sigma_{21}L - \lambda_2 M^T\Sigma_{22}M = 0 \\ L^T\Sigma_{11}L = 1 \\ M^T\Sigma_{22}M = 1 \end{cases} \iff \quad \lambda_1 = \lambda_2 = cor(V_1, V_2) = \varrho\,.$$

Итак, имеем систему уравнений:

$$\begin{cases} \Sigma_{12}M - \varrho\Sigma_{11}L = 0 \\ \Sigma_{21}L - \varrho\Sigma_{22}M = 0 \end{cases}. \tag{3}$$

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

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

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

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

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

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

Количество ненулевых корней равно рангу матрицы Σ_{12} В первом случае количество корней равно рангу матрицы Σ_{22}, то есть s, во втором рангу матрицы Σ_{11}, то есть r. Количество ненулевых корней равно m рангу

матрцы Σ_{12}, а кратности нулевых корней различны. Таким образом, имеем $m \leq min(r, s)$ канонических случайных величин $V_{1i} = L_i^T U_2$ и $V_{2i} = M_i^T U_2$, упорядоченных в соответствии с корнями $\varrho_1^2 \geq \varrho_2^2 \geq \ldots \geq \varrho_m^2$. Далее для удобства будем считать канонические корреляции ϱ_i положительными, иначе можно поменять знаки канонических коэффициентов M или L.

1.3. Свойства канонических случайных величин

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

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

Обозначим через A симметричную матрицу $\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, M_i и M_j собственные векторы, соответствующие разным собственным числам ϱ_i^2 и ϱ_j^2. В системе

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

первое уравнение умножим на M_j^T, второе на M_i^T, и уравнения сложим:

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

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

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

$$cov(V_{1i}, V_{2j}) = EL_1^T U_1 U_2^T M_j = L_i^T \Sigma_{12} M_j$$

Рассмотрим первое уравнение из системы (3) в случае $\varrho_i \neq 0$ и умножим его обе части на L_i^T:

$$\Sigma_{12}M_i = \varrho_i\Sigma_{11}L_i \iff L_i^T\Sigma_{12}M_i = \varrho_i L_i^T\Sigma_{11}L_i = \varrho_i,$$

при умножении на L_j^T получим $L_j^T\Sigma_{12}M_i = \varrho_i L_j^T\Sigma_{11}L_i = 0$. Если $\varrho_i = 0$, тогда $\Sigma_{12}M_i = 0 \iff L_i^T\Sigma_{12}M_i = 0$.

Для проверки значимости равенства нулю канонических корреляций используются статистики хи-квадрат.

1.4. Канонический анализ в R

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

Индекс тяжести	психологические характеристики
*asi*1_*med* – медицинский статус	*rabdru* – рискованность
*asi*4_*dr* – наркотический статус	*bdi* – уровень депрессии
*asi*5_*leg* – проблемы с законом	*sstati* – уровень тревоги
*asi*7_*psy* – психиатрический статус	

Используем данные

addicts <- read_ excel("Documents/share/minimumA/addicts.xls")

Выбираем переменные и задаем левое и правое множества.

$XY < -na.omit(addicts[, c(8, 11, 12, 14, 22, 25, 26)]);$

$X < -scale(XY[, seq(4)]); Y < -scale(XY[, c(5 : 7)]);$

Применяем канонический корреляцияонный анализ.

$library(CCA)$

$cc. < -cc(X, Y)$

Канонические коэффициенты корреляции

$cc.\$cor;$

0.3545431 0.1896925 0.1438964

Веса канонических величин

cc.$xcoef

	X1	X2	X3
asi1_med	0.49	-0.48	-0.78
asi4_dr	0.38	-0.15	0.61
asi5_leg	-0.24	0.45	-0.67
asi7_psy	0.61	0.73	0.09

cc.$*ycoef*

	X1	X2	X3
rabdru	0.11	0.98	0.21
bdi	0.37	0.08	-1.11
sstati	0.75	-0.19	0.88

Первая каноническая корреляция указывает на взаимосвязь между повышением уровня тревоги при более тяжелом медицинском статусе, при более высокой степени наркотической зависимости и более высоком психиатрическом статусе.

Вторая каноническая корреляция между каноническими величинами вида указывает на взаимосвязь между рискованностью и тяжестью наркотической зависимости в форме проявления психиатрического статуса и проблем с законом, что свидетельствует о том, что психически неустойчивые наркоманы, которые имеют проблемы с законом, более рискованы и позволяют себе использовать чужие шприцы.

Значения канонических величин, например, первой можно получить непосредственно

x.<-X %%as.matrix(cc.$xcoef[,1],ncol=1);*

y.<-Y %%as.matrix(cc.$ycoef[,1],ncol=1)*

или

x.<-cc.$scores$xscores[,1]

x.<-cc.$scores$yscores[,1]

Для проверки можно использовать функцию

cor.test(x.,y.)

cor.test(cc.$scores$xscores[,1],cc.$scores$yscores[,1])

чтобы убедиться, чтобы убедиться, что коэффициент корреляции между этими каноническими величинами совпадает с первой канонической корреляцией.