

0.1 Канонические корреляции

Пусть у нас имеется два набора случайных величин:

$$\boldsymbol{\eta} = (\eta_1, \dots, \eta_q)^T \quad (1)$$

$$\boldsymbol{\xi} = (\xi_1, \dots, \xi_p)^T \quad (2)$$

Тогда

Определение. *Первой канонической корреляцией называется*

$$r_1^2 = \max_{A, B} \rho^2(A^T \boldsymbol{\xi}, B^T \boldsymbol{\eta}), \quad (3)$$

где $A \in \mathbb{R}^p, B \in \mathbb{R}^q$. i -ой каноническая корреляция определяется аналогично с условием, что максимум берется по некоррелированным с предыдущими случайным величинам:

$$r_i^2 = \max_{\substack{A, B \\ \rho(A\boldsymbol{\xi}, A_j\boldsymbol{\xi})=0 \\ \rho(B\boldsymbol{\eta}, B_j\boldsymbol{\eta})=0 \\ 1 \leq j < i}} \rho^2(A^T \boldsymbol{\xi}, B^T \boldsymbol{\eta}) \quad (4)$$

Канонических корреляций будет, очевидно, $s = \min(p, q)$.

Так как считаются корреляции, это не важно, случайные величины центрированы или нет, стандартизованы или нет.

На выборочном языке, первая каноническая корреляция — это квадрат косинуса угла между подпространствами, натянутыми на первый и второй наборы признаков (центрированные). Собственно, угол между подпространствами так и определяется как минимальный угол между (ненулевыми) векторами из подпространств.

Множественная корреляция — частный случай, когда один из наборов переменных состоит только из одного признака. Это следует из того, что множественный коэффициент корреляции в квадрате — это тоже квадрат косинуса угла между вектором и подпространством.

0.2 Канонические переменные

Смотрим на (4). У нас каждая каноническая корреляция определяется с помощью двух новых признаков, $\tilde{\eta}_i = B_i^T \boldsymbol{\eta}$ и $\tilde{\xi}_i = A_i^T \boldsymbol{\xi}$. Они определяют новые ортогональные системы признаков в пространстве, натянутом на признаки из одного набора, и в пространстве, натянутом на признаки из другого набора. Новые признаки $\tilde{\eta}_i = B_i^T \boldsymbol{\eta}$ и $\tilde{\xi}_i = A_i^T \boldsymbol{\xi}$ называются правыми и левыми каноническими переменными (или X- и Y- каноническими переменными; на выборочном языке, если есть два набора признаков в матрицах $\mathbf{X}$ и $\mathbf{Y}$, то новые признаки имеют вид их линейных комбинаций $\tilde{X}_i = \mathbf{X}A_i$ и $\tilde{Y}_i = \mathbf{Y}B_i$). Признаки $\tilde{\eta}_i$ между собой ортогональны и $\tilde{\xi}_i$ между собой также ортогональны. Они составляют ортогональные системы в соответствующих пространствах (вообще говоря, не обязательно базисы, если $p \neq q$).

Задача нахождения канонических переменных сводится к обобщенным задачам на собственные собственные числа. А именно, r_i^2 , A_i и B_i находятся из

$$\boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_{yy}^{-1} \boldsymbol{\Sigma}_{yx} A = r^2 \boldsymbol{\Sigma}_{xx} A, \quad \boldsymbol{\Sigma}_{yx} \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xy} B = r^2 \boldsymbol{\Sigma}_{yy} B, \quad (5)$$

причем собственные числа у обеих задач совпадают. Здесь $\boldsymbol{\Sigma}_{xx} = \mathbb{E} \boldsymbol{\xi} \boldsymbol{\xi}^T$, $\boldsymbol{\Sigma}_{xy} = \mathbb{E} \boldsymbol{\xi} \boldsymbol{\eta}^T$, $\boldsymbol{\Sigma}_{yy} = \mathbb{E} \boldsymbol{\eta} \boldsymbol{\eta}^T$ (в предположении центрированности исходных признаков).

Скаттерплот, на котором лучше всего видна линейная зависимость между двумя наборами признаков — это скаттеплот, построенный по первым левой и правой канонической переменной. Корреляция между ними равна первой канонической корреляции.

Интерпретация канонических переменных через канонические функции такая же, как интерпретация канонических переменных через дискриминантные функции, так как это просто коэффициенты линейной комбинации, показывающей, как новые признаки выражаются через старые.

Как обычно, чтобы избавиться от масштаба, нужно стандартизовать признаки, разделив их на выборочное стандартное отклонение, тем самым, сами коэффициенты умножатся на то же число. Получатся стандартизованные канонические функции.