МЕТОД КАНОНИЧЕСКИХ КОРРЕЛЯЦИЙ

Метод канонических корреляций позволяет находить максимальные корреляционные связи между двумя группами случайных величин, являясь как бы обобщением корреляционного анализа на случай двух множеств случайных величин. Эти связи определяются при помощи новых аргументов - **канонических переменных**, определяемых как независимые линейные комбинации исходных признаков по каждой из групп. Канонические величины должны максимально коррелировать между собой, а их число определяется по числу переменных в меньшем множестве (если число переменных в них не одинаково). При этом можно ограничиться рассмотрением небольшого числа наиболее коррелированных линейных комбинаций, тем самым сократить объем исходных данных.

Например, эффективность предприятия оценивается такими показателями как: прибыль, рентабельность, фондоотдача и др., которые зависят от таких исходных факторов как: материальные затраты, трудоемкость, стоимость основных фондов, оборачиваемость и др. Задача состоит в выявлении максимальных связей между двумя этими группами показателей.

Пусть имеются две группы исходных переменных: $\vec{\eta} = (\eta_1, \eta_2, \ldots, \eta_m)$ - результирующие показатели и $\vec{\xi} = (\xi_1, \xi_2, \ldots, \xi_k)$ - определяющие факторы (возьмем для определенности $m \le k$). Будем считать, что $\vec{\xi}$ и $\vec{\eta}$ центрированные величины. Составим блочную матрицу ковариаций: $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$, $A_{11} = E(\vec{\eta}\vec{\eta}^T)$ - ковариационная матрица, характеризующая взаимосвязь результирующих показателей, $A_{22} = E(\vec{\xi}\vec{\xi}^T)$ - ковариационная матрица, характеризующая взаимосвязь исходных факторов, $A_{12} = A_{21}^T = E(\vec{\eta}\vec{\xi}^T)$ - ковариационная матрица, характеризующая взаимосвязь показателей первой и второй групп.

Положим:

$$u_i = \vec{\theta}^{(i)T}\vec{\eta} = \theta_1\eta_1 + \theta_2\eta_2 + \ldots + \theta_m\eta_m,$$

$$v_i = \vec{\beta}^T\vec{\xi} = \beta_1\xi_1 + \beta_2\xi_2 + \ldots + \beta_k\xi_k, \quad i = \overline{1,m}.$$

Будем называть u_i и v_i каноническими переменными. Задача метода заключается в нахождении пар $\{u_i, v_i\}$, с наибольшей возможной корреляцией между элементами пары. Соответственно, значения u_i будет наилучшим образом предсказываться значениями v_i.

Так как $\vec{\xi}$ и $\vec{\eta}$ центрированные величины, то $Eu_i = 0$, $Ev_i = 0$. Выберем $\vec{\theta}^{(i)}$ и $\vec{\beta}^{(i)}$ таким образом, чтобы $Du_i = 1$, $Dv_i = 1$:

$$D(u_i) = E(u_i^2) = \vec{\theta}^{(i)T} A_{11} \vec{\theta}^{(i)} = 1, \qquad D(v_i) = E(v_i^2) = \vec{\beta}^{(i)T} A_{22} \vec{\beta}^{(i)} = 1.$$

Коэффициент корреляции между u_i и v_i (учитывая, что u_i и v_i центрированные и нормированные величины): $\rho_{u_i, v_i} = E(u_i v_i) = \vec{\theta}^{(i)T} A_{12} \vec{\beta}^{(i)}$.

Задачу отыскания канонических величин можно свести к задаче на собственные значения и собственные векторы матрицы $A_{11}^{-1} A_{12} A_{22}^{-1} A_{12}^T$. Векторы коэффициентов преобразования $\vec{\theta}^{(i)}, i = \overline{1, m}$, удовлетворяющие условиям задачи поиска канонических переменных, являются собственными векторами матрицы $A_{11}^{-1} A_{12} A_{22}^{-1} A_{12}^T$, соответствующими собственным значениям $\lambda_2^2, \lambda_2^2, ..., \lambda_m^2$, упорядоченным по убыванию значений, и удовлетворяющими условию: $\vec{\theta}^{(i)T} A_{11} \vec{\theta}^{(i)} = 1$. Векторы $\vec{\beta}^{(i)}, i = \overline{1, m}$ связаны с векторами $\vec{\theta}^{(i)}$ соотношениями $\vec{\beta}^{(i)} = \frac{1}{\lambda_i} A_{22}^{-1} A_{21} \vec{\theta}^{(i)}$, $i = \overline{1, m}$. Коэффициент корреляции для каждой пары канонических переменных $\rho_i = E(u_i v_i) = \lambda_i$.

Полученные канонические переменные обладают следующими свойствами:

они являются независимыми линейными комбинациями исходных показателей соответствующих групп;

канонические переменные выбраны таким образом, чтобы соответствующие канонические корреляции были максимальными;

канонические переменные упорядочены по мере убывания соответствующих канонических корреляций;

канонические переменные из разных пар не коррелированы.

Оценка канонических корреляций на основе выборочных данных строится на основе выборочных матриц ковариаций (или матрицы корреляций, если используются нормированные исходные данные).

Для статистической проверки значимости найденных канонических переменных можно использовать критерий отношения правдоподобия. При этом, если определены m оценок канонических корреляций $\hat{\rho}_i = \hat{\lambda}_i$, $i = \overline{1, m}$, то для каждого $p = \overline{1, m}$ следует проверить гипотезу:

$$H_0 : \rho_p = 0, \rho_{p+1} = 0, \ldots, \rho_m = 0,$$

т.е., что все корреляции, начиная с ρ_p, равны нулю, при альтернативе $H_1 : \rho_p \neq 0$.

Отношение правдоподобия в данном случае можно выразить через отношение определителей:

$$\lambda = \left(\frac{\Delta_1}{\Delta_0} \right)^{n/2},$$

где Δ_0, Δ_1 - оценки определителей матрицы ковариаций канонических переменных при истинности H_0 и H_1 соответственно.

Определитель матрицы ковариаций канонических переменных: $\Delta = \prod_{i=1}^{m} (1 - \rho_i^2)$, оценка определителя матрицы ковариаций канонических переменных при истинности H_0: $\Delta_0 = \prod_{i=1}^{p-1} (1 - \hat{\lambda}_i^2)$, оценка определителя матрицы ковариаций канонических переменных при истинности H_1: $\Delta_1 = \prod_{i=1}^{m} (1 - \hat{\lambda}_i^2)$. Соответственно, отношение правдоподобия:

$$\lambda = \left(\frac{\Delta_1}{\Delta_0} \right)^{n/2} = \left(\prod_{i=p}^{m} (1 - \hat{\lambda}_i^2) \right)^{n/2}.$$

При истинности H_0 статистика

$$\eta = -\left\{ n - 1 - \frac{1}{2}(m + k + 1) \right\} \cdot \ln \left(\prod_{i=p}^{m} \left(1 - \lambda_i^2\right) \right)$$

асимптотически стремится к распределению χ^2 с числом степеней свободы $\nu = (k - p + 1)(m - p + 1)$.

Доля дисперсии каждого исходного признака, объясняемая канонической переменной, равна квадрату ковариации между канонической переменной и исходным фактором (учитывая, что канонические переменные нормированы). Вектор ковариаций между канонической переменной и исходными факторами через параметры модели можно найти как: $\mathrm{cov}(\vec{\eta}, u_j) = A_{11}\vec{\theta}_i$ (для первой группы), $\mathrm{cov}(\vec{\xi}, u_j) = A_{22}\vec{\beta}_i$ (для второй группы), либо вычислить выборочные ковариации непосредственно по значениям переменных.

Векторы $\vec{\Theta}^{(i)} = A_{11}\vec{\theta}^{(i)}$, $\vec{\mathrm{B}}^{(i)} = A_{22}\beta^{(i)}$, $i = \overline{1,m}$, будем называть векторами канонических нагрузок соответствующих канонических переменных. Квадрат длины вектора канонических нагрузок равен величине объясняемой (извлеченной) канонической переменной дисперсии для соответствующего множества:

$$V_1^{(i)} = \left|\vec{\Theta}^{(i)}\right|^2, \quad V_2^{(i)} = \left|\vec{\mathrm{B}}^{(i)}\right|^2, \quad i = \overline{1,m}.$$

Доли суммарных дисперсий исходных признаков, объясняемые p каноническими переменными:

$\sum_{j=1}^{p} V_1^{(j)} \Big/ Sp(A_{11})$ - для 1-го множества, $\sum_{j=1}^{p} V_2^{(j)} \Big/ Sp(A_{22})$ - для 2-го множества.

Коэффициент канонической корреляции $\rho_i = \lambda_i$ при возведении в квадрат дает долю дисперсии, общей для каждой из канонических переменных в паре. Если умножить эту долю на соответствующую долю извлеченной канонической переменной дисперсии и просуммировать по всем каноническим переменным, то можно получить меру избыточности множества переменных, т.е., величину, показывающую, насколько избыточно одно множество переменных, если задано другое множество.

Избыточность первого множества переменных при заданном втором множестве:

$$\frac{1}{Sp(A_{11})} \sum_{j=1}^{p} \rho_j^2 V_1^{(j)},$$

избыточность второго множества переменных при заданном первом множестве:

$$\frac{1}{Sp(A_{22})} \sum_{j=1}^{p} \rho_j^2 V_2^{(j)}.$$