Собственные векторы и собственные значения корреляционной матрицы и их использование для получения матрицы нагрузок в методе главных компонент. Свойства элементов матрицы нагрузок.

Метод главных компонент (Principal Component Analysis, PCA) — это метод снижения размерности данных, который позволяет перейти от большого числа коррелированных признаков к меньшему числу новых некоррелированных переменных — **главных компонент**.
Для этого используются **собственные векторы и собственные значения корреляционной (или ковариационной) матрицы**.

в методе главных компонент **собственные векторы корреляционной матрицы** — это направления, вдоль которых данные имеют наибольший разброс, и они служат базисом для нового, более компактного представления данных.

Собственный вектор — это такой ненулевой вектор U_k, который при умножении на матрицу корреляций R_z изменяется только скалярно (умножается на число), а не меняет направления в пространстве.

Формально это записывается так:

$$R_z U_k = \lambda_k U_k,$$

где

- R_z — корреляционная матрица стандартизованных данных,
- λ_k — **собственное значение** (скаляр), соответствующее собственному вектору U_k.

1. Собственные векторы и собственные значения корреляционной матрицы

В методе главных компонент корреляционная матрица R_z стандартизованных данных Z имеет p собственных значений $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ и соответствующих им собственных векторов $U_1, U_2, \ldots, U_p$, которые нормированы ($U_k^T U_k = 1$) и ортогональны ($U_j^T U_k = 0$ при $j \neq k$).
Каждая главная компонента $f_k = ZU_k$, а её дисперсия $D(f_k) = \lambda_k$.

3. Как получаются главные компоненты?

1. Исходные стандартизованные данные Z (матрица $n \times p$) имеют корреляционную матрицу R.
2. Находятся собственные значения λ_i и собственные векторы u_i матрицы R.
3. Главные компоненты F вычисляются как:

$$F = ZU,$$

где U — матрица, столбцами которой являются собственные векторы $u_1, u_2, \ldots, u_p$.

4. Первая главная компонента $f_1 = Zu_1$ имеет наибольшую дисперсию λ_1, вторая — λ_2, и так далее.

1.4. Матрица «нагрузок»

Матрица «нагрузок» является важной характеристикой главных компонент. Во-первых, с ее помощью можно построить обратную зависимость $Z = F_H \cdot (??)$, а во-вторых, свойства элементов этой матрицы позволяют содержательно интерпретировать главные компоненты.

Из уравнения (13) следует, что ненормированные значения главных компонент могут быть выражены через нормированные значения $F = F_H \Lambda^{1/2}$.

2. Получение матрицы нагрузок

Матрица нагрузок A определяется как

$$A = U\Lambda^{1/2},$$

где $\Lambda^{1/2}$ — диагональная матрица с элементами $\sqrt{\lambda_k}$.

1)***Коэффициенты матрицы нагрузок*** a_{ij} ***определяют удельный вес влияния*** j ***-й нормированной главной компоненты*** f_{Hj} ***на*** i***-й исходный признак***:

$$Z_i = a_{i1}f_{H1} + a_{i2}f_{H2} + \cdots + a_{ip}f_{Hp}. \tag{14}$$

2)***Коэффициенты*** a_{ij} ***матрицы нагрузок определяют величину парного коэффициента корреляции между*** Z_i ***и*** f_j***, т.е.*** $a_{ij} = r(z_i, f_j)$:

$$\begin{aligned} r(Z_i, f_j) &= \frac{M[(Z_i - M(Z_i))(f_j - M(f_j))]}{\sqrt{D(Z_i)G(f_j)}} = \frac{M[Z_i f_j]}{\sqrt{\lambda_j}} = M\left(Z_i \frac{f_j}{\sqrt{\lambda_j}}\right) = \\ &= M(Z_i f_{Hj}) = M\left[(a_{i1}f_{H1} + ai2f_{H2} + \cdots + a_{ipf}Hp)f_{Hj}\right] = \\ &= a_{ij}M\left(f_{Hj}^2\right) = a_{ij}. \end{aligned} \tag{15}$$

При выводе соотношения (15) использовались нормированность и центрированность переменных Z_i и f_{Hj}.

Это свойство коэффициентов матрицы нагрузок дает основание придать главной компоненте Z_j содержательный смысл (присвоить имя), соответствующий исходному признаку Z_i (признакам), для которых коэффициент a_{ij} достигает наибольшего значения (обычно принимают во внимание $|a_{ij}| > 0,6$).

Отметим еще два свойства элементов матрицы нагрузок.

3) Из определения матрицы нагрузок A следует:

$$A^T A = (U\Lambda^{1/2})^T(U\Lambda^{1/2}) = \Lambda^{1/2}U^T U\Lambda^{1/2} = \Lambda^{1/2}\Lambda^{1/2} = \Lambda.$$

Это означает, что ***сумма квадратов элементов*** j***-го столбца матрицы нагрузок равна дисперсии*** j***-й главной компоненты*** — λ_j.

4) ***Сумма квадратов элементов любой*** i***-й строки матрицы нагрузок равна единице***.

Для доказательства этого свойства возведем в квадрат выражение (14), и вычислим математическое ожидание от полученной величины:

$$(Z_i^2) = D(Z_i) = M(a_{i1}f_{H1} + a_{i2}f_{H2} + \cdots + a_{ip}f_{Hp})^2 = \sum_{j=1}^{p} a_{ij}^2 M(f_{Hj}) = \sum_{j=1}^{p} a_{ij}^2.$$

Учитывая, что $D(Z_i) = 1$ (Z_i — нормированная величина), получаем $\sum_{j=1}^{p} a_{ij}^2 = 1$.

5) С помощью матрицы нагрузок можно восстановить матрицу корреляций исходных признаков — $AA^T = R_x$:

$$AA^T = (U\Lambda^{1/2})(U\Lambda^{1/2})^T = U\Lambda^{1/2}\Lambda^{1/2}U^T = U\Lambda U^T = U(U^T R_x U)U^T = R_x.$$