29. Основные числовые характеристики и свойства оптимальности главных компонент в методе главных компонент.

1. Свойства оптимальности главных компонент

Метод главных компонент основан на переходе к новой системе координат с целью «сжатия» информации. Критерием качества (информативности) служит максимизация дисперсии новых переменных.

Функционал информативности:

$$I_{p'}[F(Z)] = \frac{D(F_1) + ... + D(F_{p'})}{D(Z_1) + ... + D(Z_p)} \longrightarrow max$$

Определение и вывод первой главной компоненты

Определение: Первой главной компонентой $f_1(Z)$ называется такая нормированно-центрированная линейная комбинация исходных показателей Z, которая обладает **наибольшей дисперсией**.
Линейная комбинация имеет вид $f_1 = ZU_1$, где U_1 — вектор коэффициентов.

Доказательство (Вывод через задачу оптимизации):
Требуется найти вектор U_1, максимизирующий дисперсию при условии нормировки:

$$\begin{cases} D(ZU_1) \longrightarrow max \\ U_1^T U_1 = 1 \end{cases}$$

1. **Выразим дисперсию:**

$$D(ZU_1) = M[(ZU_1)^2] = M[U_1^T Z^T Z U_1] = U_1^T M(Z^T Z) U_1 = U_1^T R_x U_1$$

(где R_x — корреляционная матрица стандартизованных данных).
Задача принимает вид:

$$\begin{cases} U_1^T R_x U_1 \longrightarrow max \\ U_1^T U_1 = 1 \end{cases}$$

2. **Функция Лагранжа:**

   Составим функцию для поиска условного экстремума:

$$\varphi(U_1, \lambda) = U_1^T R_x U_1 - \lambda(U_1^T U_1 - 1)$$

3. **Необходимое условие экстремума:**

   Найдем производную по вектору U_1 и приравняем её к нулю:

$$\frac{\partial \varphi}{\partial U_1} = 2R_x U_1 - 2\lambda U_1 = 0$$

   Сократив на 2, получаем систему линейных однородных уравнений:

$$R_x U_1 - \lambda U_1 = 0 \quad \text{или} \quad (R_x - \lambda I)U_1 = 0$$

4. **Решение системы:**

   Чтобы система имела ненулевое решение, определитель матрицы должен быть равен нулю:

$$|R_x - \lambda I| = 0$$

   Это характеристическое уравнение. Следовательно, λ — это собственное число матрицы R_x, а U_1 — соответствующий собственный вектор.

5. **Максимизация дисперсии:**

   Докажем, чему равна дисперсия полученной компоненты. Умножим уравнение $R_x U_1 = \lambda U_1$ слева на U_1^T:

$$U_1^T R_x U_1 = U_1^T \lambda U_1 = \lambda(U_1^T U_1)$$

   Так как $U_1^T U_1 = 1$, получаем:

$$D(f_1) = U_1^T R_x U_1 = \lambda$$

   Чтобы дисперсия была **максимальной**, необходимо выбрать **наибольшее** собственное число λ_1.

Вывод: Первая главная компонента соответствует наибольшему собственному числу λ_1, а вектор коэффициентов U_1 — это собственный вектор, соответствующий λ_1.

k-я главная компонента

Аналогично определяется k-я главная компонента ($f_k = ZU_k$), которая должна иметь максимальную дисперсию и **не коррелировать** с предыдущими.
Решение этой задачи приводит к тому, что U_k — это собственный вектор, соответствующий k-му по величине собственному числу λ_k. Дисперсия этой компоненты равна $D(f_k) = \lambda_k$.

2. Основные числовые характеристики главных компонент

Определим характеристики вектора главных компонент $F = ZU$, где U — матрица собственных векторов.

1) Математическое ожидание

$$M(F) = M(ZU) = U \cdot M(Z) = 0$$

(так как исходные данные Z центрированы, $M(Z) = 0$).

2) Ковариационная матрица (Доказательство некоррелированности)

Ковариационная матрица S_F главных компонент является диагональной.

Доказательство:

$$S_F = M(F^T F) = M[(ZU)^T (ZU)] = M[U^T Z^T ZU] = U^T M[Z^T Z]U = U^T R_x U$$

Известно, что собственные векторы и числа связаны соотношением (из уравнения (10) лекции):

$$R_x U_k = \lambda_k U_k$$

Домножим это равенство слева на вектор U_j^T:

$$U_j^T R_x U_k = U_j^T \lambda_k U_k = \lambda_k (U_j^T U_k)$$

Собственные векторы матрицы R_x обладают свойствами ортогональности и нормированности:

- При $j \neq k$ (ортогональность): $U_j^T U_k = 0$.
- При $j = k$ (нормировка): $U_k^T U_k = 1$.

Следовательно:

$$U_j^T R_x U_k = \begin{cases} 0, & j \neq k \\ \lambda_k, & j = k \end{cases}$$

Таким образом, матрица S_F имеет вид:

$$S_F = U^T R_x U = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{pmatrix}$$

Это доказывает, что главные компоненты не коррелированы (внедиагональные элементы равны нулю).

3) Сумма дисперсий

Сумма дисперсий всех главных компонент равна сумме дисперсий исходных признаков (равной p для стандартизованных данных).

Доказательство:
Сумма дисперсий есть след (trace) матрицы S_F:

$$\sum_{k=1}^{p} D(f_k) = tr(S_F) = tr(U^T (R_x U))$$

Используем свойство следа матрицы $tr(AB) = tr(BA)$. Переставим U^T в конец:

$$tr((R_x U)U^T) = tr(R_x (UU^T))$$

Так как матрица собственных векторов U ортогональна, то $UU^T = I$ (единичная матрица):

$$= tr(R_x I) = tr(R_x) = \sum_{k=1}^{p} 1 = p$$

(так как на диагонали корреляционной матрицы R_x стоят единицы).