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

1. Основные числовые характеристики главных компонент

Пусть исходные данные Z (размерности $n \times p$) стандартизованы (центрированы и нормированы). Матрица корреляций исходных признаков — R_x.
Главные компоненты определяются как линейные комбинации исходных признаков: $F = ZU$, где U — матрица собственных векторов R_x.
(про собственные векторы можно будет узнать в 30-м билете).

1) Математическое ожидание

Математическое ожидание главных компонент равно нулю.

Доказательство:
Так как исходные данные Z стандартизованы, их математическое ожидание равно нулю ($M(Z) = 0$). Следовательно:

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

2) Ковариационная матрица (Свойство некоррелированности)

Ковариационная матрица векторов главных компонент S_F является диагональной. На главной диагонали стоят собственные числа λ_i матрицы корреляций R_x.
Это означает, что главные компоненты **не коррелированы** и **ортогональны**.

Доказательство:
По определению ковариационная матрица равна:

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

Раскроем скобки (учитывая, что $(AB)^T = B^T A^T$):

$$= M[U^T Z^T Z U] = U^T M[Z^T Z] U$$

Так как данные стандартизованы, $M[Z^T Z]$ представляет собой корреляционную матрицу исходных данных R_x:

$$= U^T R_x U$$

Известно соотношение для собственных векторов и чисел матрицы R_x:

$$R_x U_k = \lambda_k U_k \quad \Rightarrow \quad R_x U = U\Lambda$$

где Λ — диагональная матрица собственных чисел.

Подставим это в выражение для S_F:

$$S_F = U^T (U\Lambda) = (U^T U)\Lambda$$

Так как собственные векторы ортонормированы ($U^T U = I$ — единичная матрица), получаем:

$$S_F = I \cdot \Lambda = \Lambda = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{pmatrix}$$

3) Сумма дисперсий (Сохранение полной информации)

Сумма дисперсий всех главных компонент равна сумме дисперсий исходных признаков.

Доказательство:
Сумма дисперсий равна следу (trace) ковариационной 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 $UU^T = I$:

$$= tr(R_x \cdot I) = tr(R_x)$$

У стандартизованных данных дисперсии всех p признаков равны 1, поэтому сумма диагональных элементов R_x равна p:

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

2. Свойства оптимальности главных компонент

Метод главных компонент решает задачу «сжатия» информации: перехода от большого числа исходных признаков p к меньшему числу новых переменных p', при этом стараясь потерять как можно меньше информации.

В статистике **мерой информации** считается **дисперсия** (вариация, разброс). Если признак не меняется (дисперсия = 0), он не несет информации о различии объектов. Чем больше разброс данных, тем больше информации они содержат.

Функционал информативности (I):
Для оценки качества сжатия используется функционал, показывающий, какую долю суммарного разброса (информации) мы сохранили:

$$I_{p'}[F(Z)] = \frac{\sum_{k=1}^{p'} D(F_k)}{\sum_{j=1}^{p} D(Z_j)} = \frac{D(F_1) + ... + D(F_{p'})}{p} \longrightarrow max$$

Где:

- $D(F_k)$ — дисперсия k-й главной компоненты (мера информации, которую несет эта новая переменная).
- $\sum D(Z_j) = p$ — полная дисперсия (полная информация) исходной системы признаков.
- Смысл критерия: Максимизировать долю сохраненной информации при уменьшении количества переменных.

Исходя из этого, формулируется свойство оптимальности:

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 (то есть берет на себя максимум информации).
Линейная комбинация имеет вид $f_1 = ZU_1$.

Доказательство (Вывод через задачу оптимизации):
Требуется найти вектор коэффициентов U_1, максимизирующий дисперсию новой переменной при условии нормировки вектора (чтобы решение было единственным):

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

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 ее дисперсия равна этому числу ($D(f_1) = \lambda_1$).

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 $D(f_k) = \lambda_k$.