

Семинар 10.

Мультиколлинеарность. Метод главных компонент (МГК, PCA).

1. Метод главных компонент (Principal Component Analysis, PCA).

(a) Метод главных компонент. Теория.

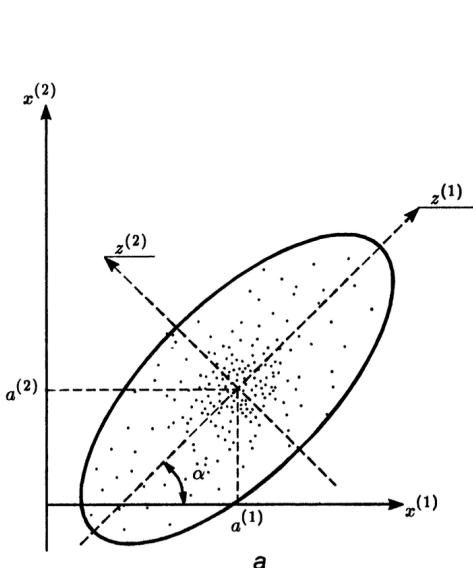
(b) Разберите геометрическую интерпретацию на примере двух нормальных признаков $x_1 \sim N(a_1, \sigma_1^2)$ и $x_2 \sim N(a_2, \sigma_2^2)$.

Рис. 1. Умеренный разброс точек.

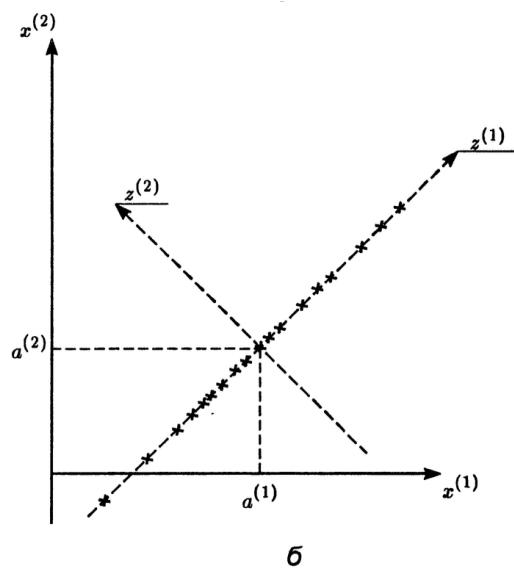


Рис. 2. Отсутствие разброса точек в направлении второй главной компоненты (вырожденный случай).

(c) Определим матрицу факторных нагрузок как $A = C\Sigma_Z^{1/2}$, где Σ_Z — ковариационная матрица некоррелированных главных компонент. Покажите, что элемент (a_{ij}) матрицы A является коэффициентом корреляции между стандартизированной главной компонентой j на стандартизированном признаком x_i

2. Теоретическая регрессионная зависимость и выборочная корреляционная матрица стандартизованных регрессоров X имеют вид:

$$y_i = \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i,$$

$$\widehat{R}(X) = \begin{pmatrix} 1 & 0.95 & 0 \\ 0.95 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

(a) Вычислите все главные компоненты. Сколько главных компонент надо выбрать, чтобы они объясняли не менее 70% общей дисперсии?

- (b) Вычислите матрицу факторной нагрузки. Проинтерпретируйте полученные результаты.

Решение:

- (a) Выборочная корреляционная матрица:

$$\hat{R} = \begin{pmatrix} 1 & 0.95 & 0 \\ 0.95 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Собственные значения:

$$\lambda_1 = 1.95, \quad \lambda_2 = 1, \quad \lambda_3 = 0.05$$

Собственные векторы:

$$v_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{pmatrix}, \quad v_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad v_3 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \\ 0 \end{pmatrix}$$

Главные компоненты:

$$Z_1 = \frac{1}{\sqrt{2}}X_1 + \frac{1}{\sqrt{2}}X_2, \quad Z_2 = X_3, \quad Z_3 = \frac{1}{\sqrt{2}}X_1 - \frac{1}{\sqrt{2}}X_2$$

Доля объяснённой дисперсии первой главной компонентой Z_1 :

$$I(1) = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{1.95}{3} = 0.65 \quad (65\%).$$

Доля объяснённой дисперсии двумя первыми главными компонентами Z_1 и Z_2 :

$$I(2) = \frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3} \frac{1.95 + 1}{3} = 0.9833 \quad (98.33\%)$$

- (b) Матрица факторных нагрузок:

$$A = C \cdot \Sigma_Z^{1/2}$$

$$C = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{1}{\sqrt{2}} \\ 0 & 1 & 0 \end{pmatrix}, \quad \Sigma_Z^{1/2} = \begin{pmatrix} \sqrt{1.95} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \sqrt{0.05} \end{pmatrix}$$

$$A = \begin{pmatrix} \frac{\sqrt{1.95}}{\sqrt{2}} & 0 & \frac{\sqrt{0.05}}{\sqrt{2}} \\ \frac{\sqrt{1.95}}{\sqrt{2}} & 0 & -\frac{\sqrt{0.05}}{\sqrt{2}} \\ 0 & 1 & 0 \end{pmatrix} \approx \begin{pmatrix} 0.987 & 0 & 0.158 \\ 0.987 & 0 & -0.158 \\ 0 & 1 & 0 \end{pmatrix}$$

Проинтерпретируем полученные результаты:

- Z_1 : сильно коррелирована с X_1 и X_2 (0.987)
- Z_2 : тождественна X_3 (нагрузка 1)
- Z_3 : слабо связана с разностью X_1 и X_2