Классификация и факторный анализ

Н.П.Алексеева

СПбГУ, мат-мех ф-т

Многомерный анализ данных

- $X = (X_1, \ldots, X_k)^T$ вектор признаков,
- $\mathbb{E}X$ вектор средних значений,
- $\Sigma = \mathbb{E}(X - \mathbb{E}X)(X - \mathbb{E}X)^T = \{\sigma_{ij}\}_{i,j=1}^{k}$ ковариационная матрица,
- $\lambda_1 \geq \ldots, \geq \lambda_k$ и $\alpha_1, \ldots, \alpha_k$ соответственно собственные числа и собственные вектора матрицы Σ, то есть $\Sigma\alpha_i = \lambda_i\alpha_i$.

Описание признаков

- L продолжительность жизни
- M количество чиновников
- P процент бедных
- A число автомобилей
- V объемы продаваемой водки

Выявление ведущих факторов

Первой главной компонентой Y_1 признаков $\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{k}$ называется линейная комбинация исходных признаков

$$\mathrm{Y}_1 = \alpha_{11}\mathrm{X}_1 + \ldots + \alpha_{\mathrm{k}1}\mathrm{X}_\mathrm{k},$$

где коэффициенты $\alpha_1 = (\alpha_{11}, \ldots, \alpha_{\mathrm{k}1})^\mathrm{T}$ выбраны таким образом, чтобы дисперсия $\mathrm{D}(\mathrm{Y}_1) = \lambda_1$ была максимальной. Вторая главная компонента

$$\mathrm{Y}_2 = \alpha_{12}\mathrm{X}_1 + \ldots + \alpha_{\mathrm{k}2}\mathrm{X}_\mathrm{k},$$

где коэффициенты $\alpha_2 = (\alpha_{12}, \ldots, \alpha_{\mathrm{k}2})^\mathrm{T}$ выбраны таким образом, что компоненты Y_1 и Y_2 некоррелированы, а дисперсия $\mathrm{D}(\mathrm{Y}_2) = \lambda_2$ является максимальной из всех линейных комбинаций, некоррелированных с Y_1. Остальные главные компоненты строятся аналогично.

Факторные нагрузки

Дисперсия главной компоненты $\mathbb{D}Y_j = \lambda_j$.

Корреляция $\beta_{ij} = \mathrm{cor}(X_i, Y_j) = \dfrac{\alpha_{ij}\sqrt{\lambda_j}}{\sigma_{ii}}$ между признаком X_i и главной компонентой Y_j называется факторной нагрузкой.

Суммарная дисперсия $\sum_{i=1}^{k} \sigma_{ii} = \sum_{j=1}^{k} \lambda_i$.

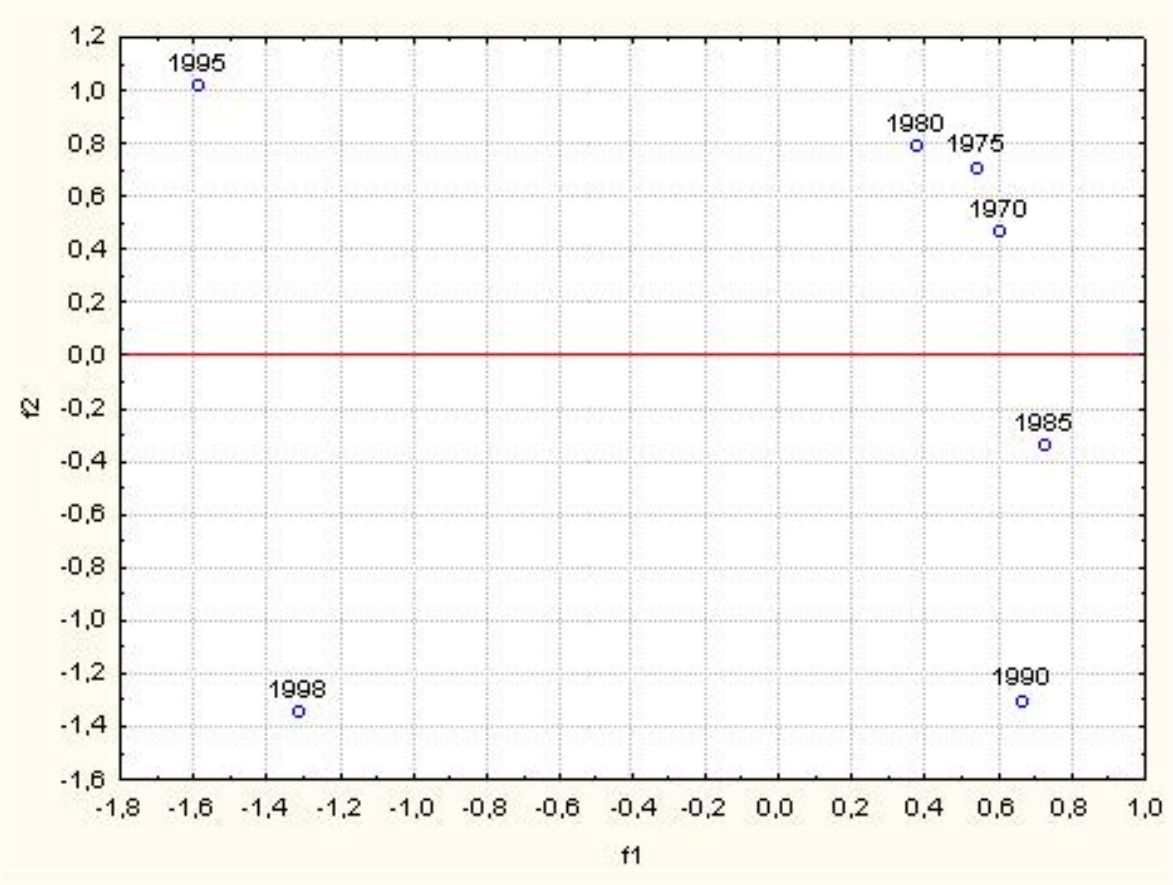
факторные нагрузки

	Factor 1	Factor 2
L	0.896	−0.398
M	−0.815	−0.564
P	0.905	−0.045
A	−0.847	−0.486
V	−0.772	0.613
Дисп.гл.комп. λ	3.60	1.06
вклад в дисп.	72%	21.8%

Значения факторов

годы	f_1	f_2
1970	0.600	0.465
1975	0.540	0.710
1980	0.375	0.790
1985	0.724	−0.339
1990	0.662	−1.301
1995	−1.587	1.019
1998	−1.314	−1.345

Двумерная диаграмма факторов



Формула восстановления

Коэффициенты $A = \{\alpha_{ij}\}$ выражаются через собственные векторы ковариационной матрицы, $Y = A^T X$, $X = AY$.

$$X_i = \sum_{j=1}^{k} \alpha_{ij} Y_j$$

Figure: Восстановление средней продолжительности жизни по первым двум факторам. $\widehat{L} = (0.896 f_1 + (-0.398 f_2)) \cdot 1.64 + 67.8$.

ПРИМЕР. W_1 - больные бронхитом, W_2 - больные пневмонией, x_1 серосодержащие белки в восстановительной форме (SH1), x_2 скорость оседания эритроцитов (SOE1) в начале заболевания.

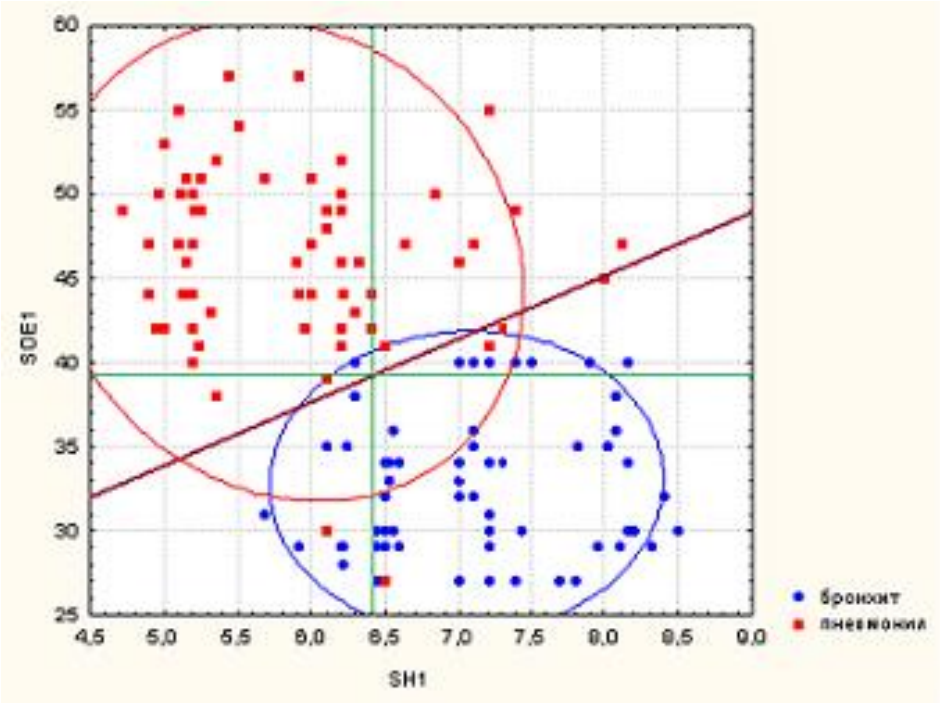


Figure: Классифицирующая прямая $\mathrm{SOE1} = 3.7\mathrm{SH1} + 15.5$.

Классификация в случае более двух популяций

Задача состоит в том, чтобы отнести вектор наблюдений $x = (x_1, \ldots, x_p)^T$ с ковариационной матрицей Σ к одной из популяций $W_1, \ldots, W_k$ с векторами средних $\mu_1, \ldots, \mu_k$.

- $q_i = P(W_i)$ априорная вероятность того, что $x \in W_i$,
- $f_i(x)$ плотность распределения популяции W_i,
- апостериорная вероятность имеет вид

$$P(W_j|x) = \frac{q_j f_j(x)}{\sum\limits_{i=1}^{k} q_i f_i(x)} = \frac{e^{\delta_i}}{\sum\limits_{j=1}^{k} e^{\delta_j}},$$

где $\delta_i = \alpha_i^T x + \gamma_i + \log q_i$, $\alpha_i = \Sigma^{-1}\mu_i$, $\gamma_i = -\frac{1}{2}\mu_i{}^T\Sigma^{-1}\mu_i$.

Наблюдение относится к той популяции, для которой $P(W_j|x)$ наибольшее.

Информационное разнообразие

$n = n_1 + n_2 + \ldots + n_m$ индивидов, отличающихся a_i, $i = 1, \ldots, m$, градациями признака X_1. Информационное разнообразие этой совокупности измеряется формулой:

$$I = I_1 = n \ln n - \sum_{i=1}^{m} n_i \ln n_i.$$

Если признаков несколько, то разнообразия складываются.

Информационный выигрыш от объединения от совокупностей A и B

$$\Delta I_{A+B} = I_{A+B} - I_A - I_B.$$

Кластеризация осуществляется по принципу минимума информационного выигрыша

Пример кластеризации наблюдений.

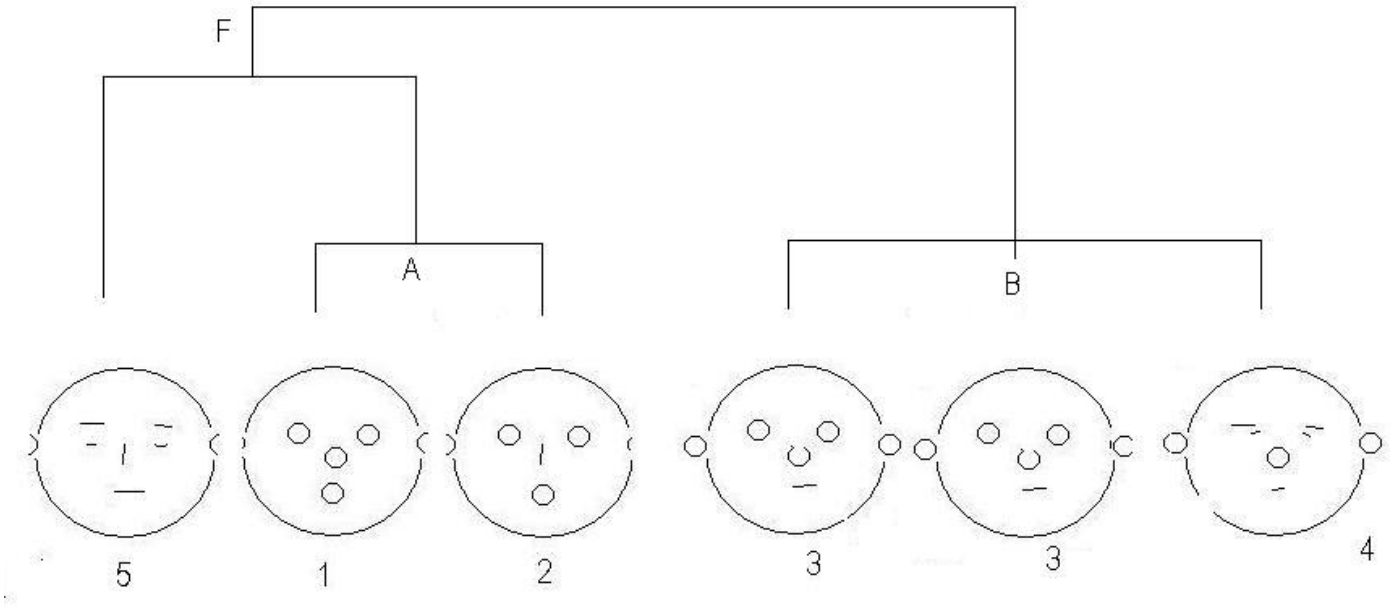


Figure: Дерево кластеризации.