

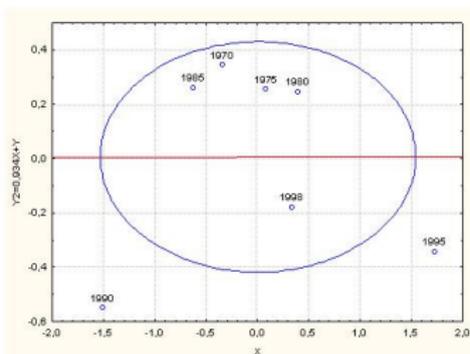
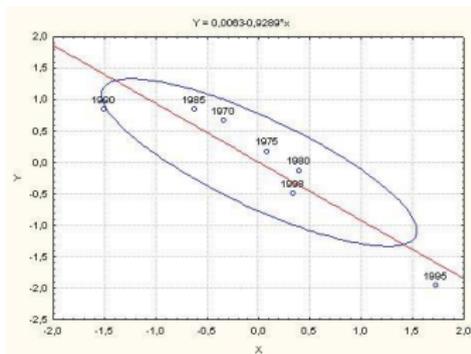
# Линейные преобразования, приводящие к некоррелированным компонентам

$X$ ,  $Y$  центрированные и нормированные величины:

$Y$  - продолжительность жизни  $X$  - объемы продаж водки,

$$\begin{cases} X_0 = 1 \cdot X + 0 \cdot Y \\ Y_0 = -r \cdot X + 1 \cdot Y \end{cases}, \quad (1)$$

$\text{cor}(X, Y) = r = -0.934$ ,  $\mathbb{D}X_0 = 1$ ,  $\mathbb{D}Y_0 = r^2 + 1 - 2r^2 = 1 - r^2$ ,  
 $\text{cov}(X_0, Y_0) = \mathbb{E}X_0Y_0 = \mathbb{E}X(Y - rX) = r - r = 0$ .



$$\mathbb{E}X = \mathbb{E}y = 0, \quad \mathbb{D}X = \mathbb{D}y = 1, \quad \text{cor}(X, Y) = r.$$

Главные компоненты

$$\begin{cases} X_1 = \frac{1}{\sqrt{2}}X - \frac{1}{\sqrt{2}}Y \\ Y_1 = \frac{1}{\sqrt{2}}X + \frac{1}{\sqrt{2}}Y \end{cases} . \quad (2)$$

$$\mathbb{E}X_1Y_1 = \mathbb{E} \left( \frac{1}{\sqrt{2}}X - \frac{1}{\sqrt{2}}Y \right) \left( \frac{1}{\sqrt{2}}X + \frac{1}{\sqrt{2}}Y \right) = 0,$$

$$\mathbb{D}X_1 = \mathbb{D} \left( \frac{1}{\sqrt{2}}X - \frac{1}{\sqrt{2}}Y \right) = \frac{\mathbb{D}X}{2} + \frac{\mathbb{D}Y}{2} - \mathbb{E}XY = 1 - \rho,$$

$$\mathbb{D}Y_1 = \mathbb{D} \left( \frac{1}{\sqrt{2}}X + \frac{1}{\sqrt{2}}Y \right) = \frac{\mathbb{D}X}{2} + \frac{\mathbb{D}Y}{2} + \mathbb{E}XY = 1 + \rho.$$

- 1 главные компоненты  $X_1$  и  $Y_1$  некоррелированы;
- 2 преобразование (2) сохраняет суммарную дисперсию, то есть  $DX_1 + DY_1 = D(X) + D(Y)$ .
- 3 сохраняется расстояния между точками

ортогональная матрица  $C = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$

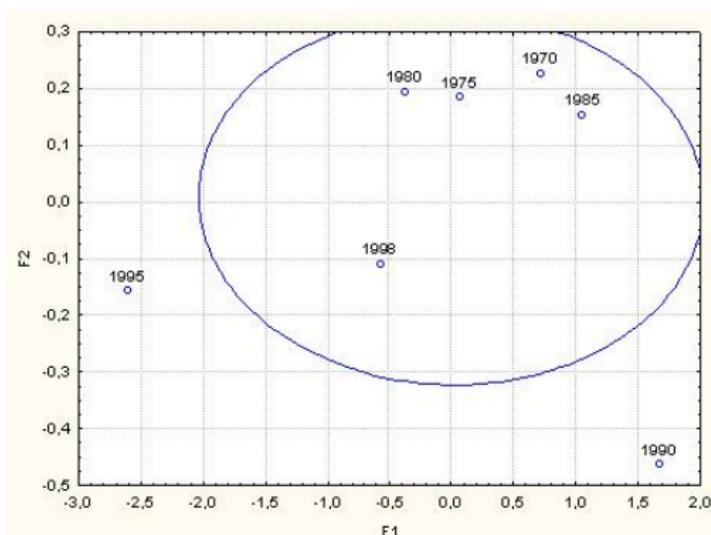
Расстояние между точками  $\varrho(a, b) = (a - b)^T(a - b)$

$$\begin{aligned}\varrho(a', b') &= (a' - b')^T(a' - b') = (Ca - Cb)^T(Ca - Cb) = \\ &= (a - b)^T C^T C (a - b) = \varrho(a, b).\end{aligned}$$

# Интерпретация главных компонент

Благополучные годы 1990 и 1985 (водки меньше, продолжительность жизни больше), неблагоприятный 1995. Вклад первой главной компоненты в общую дисперсию равен

$$\frac{D(X_1)}{D(X_1) + D(Y_1)} = \frac{1 - \rho}{2} = \frac{1 - (-0.934)}{2} = 0.967.$$



Как заменить  $k$ -мерную случайную величину при наименьшей потере информативности на  $m$ - мерную ( $m < k$ ).

Эффективность любого выбора зависит от того, в какой степени эти  $m$  линейных функций дают возможность реконструировать или восстановить  $k$  первоначальных величин.

Оказывается, что наилучшее восстановление по  $m$  линейным функциям осуществляется при помощи первых  $m$  главных компонент.

Пусть имеется  $k$  признаков  $X_1, \dots, X_k$ . Первой главной компонентой  $Y_1$  называется сохраняющая расстояние между точками линейная комбинация исходных признаков

$$Y_1 = \alpha_{11}X_1 + \dots + \alpha_{k1}X_k,$$

где коэффициенты  $\alpha_{11}, \dots, \alpha_{k1}$  выбираются таким образом, чтобы дисперсия  $D(Y_1) = \lambda_1$  была максимальной.

Это соответствует тому, что по первой главной компоненте индивиды должны отличаться наибольшим образом.

Вторая главная компонента также является линейной комбинацией исходных признаков:

$$Y_2 = \alpha_{12}X_1 + \dots + \alpha_{k2}X_k,$$

где коэффициенты  $\alpha_{12}, \dots, \alpha_{k2}$  выбираются таким образом, что компоненты  $Y_1$  и  $Y_2$  некоррелированы, а дисперсия  $D(Y_2) = \lambda_2$  является максимальной из всех линейных комбинаций, некоррелированных с  $Y_1$ .

Аналогично строятся остальные главные компоненты

$$Y_j = \sum_{i=1}^k \alpha_{ij} X_i, \quad j = 1, \dots, k.$$

Суммарная дисперсия остается неизменной:

$$V = D(X_1) + \dots + D(X_k) = \lambda_1 + \dots + \lambda_k.$$

Значимость главных компонент определяется долей объясняемой ими дисперсии, равной  $\frac{\lambda_j}{V} \cdot 100\%$ .

Факторами называются нормированные главные компоненты:  $Y_j / \sqrt{\lambda_j}$ .

Пусть  $X = (X_1, \dots, X_k)^T$ ,  $EX_i = 0$  и  $\Sigma = EXX^T$ . Обозначим через  $A_i$  собственные векторы матрицы  $\Sigma$ , соответствующие собственным числам  $\lambda_i$

$$\Sigma A_i = \lambda_i A_i, \quad A_i^T A_j = \begin{cases} 1, & \text{при } i = j, \\ 0, & \text{при } i \neq j \end{cases}, \quad i = 1, \dots, k. \quad (3)$$

Упорядочим собственные числа:  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k$ .  
и определим главную компоненту как

$$Y_j = A_j^T X = \sum_{i=1}^k a_{ij} X_i. \quad (4)$$

Соберем собственные вектора  $A_i$  в одну ортогональную матрицу

$$\mathcal{A} = [A_1, \dots, A_k] = \begin{bmatrix} a_{11} & \dots & a_{1k} \\ \cdot & \cdot & \cdot \\ a_{k1} & \dots & a_{kk} \end{bmatrix}, \quad \mathcal{A}^T \mathcal{A} = I$$

и главные компоненты в один вектор

$$Y = \begin{bmatrix} Y_1 \\ \cdot \\ Y_k \end{bmatrix} = \begin{bmatrix} A_1^T X \\ \cdot \\ A_k^T X \end{bmatrix} = \mathcal{A}^T X, \quad \text{откуда } X = \mathcal{A} Y. \quad (5)$$

## Теорема

$$1. \ EY_i Y_j = \begin{cases} \lambda_j, & \text{при } i = j, \\ 0, & \text{при } i \neq j \end{cases}, \quad 2. \ \sum_{i=1}^k DX_i = \sum_{j=1}^k DY_j.$$

$$\begin{aligned} 1. \ EY_i Y_j &= EY_i Y_j^T \stackrel{(4)}{=} EA_i^T X (A_j^T X)^T = A_i^T \Sigma A_j \stackrel{(3)}{=} \\ &= A_i^T \lambda_j A_j = \begin{cases} \lambda_j, & \text{при } i = j, \\ 0, & \text{при } i \neq j. \end{cases} \quad DY_j = \lambda_j. \end{aligned}$$

$$\begin{aligned} 2. \ DX_1 + \dots + DX_k &= \sum_{i=1}^k EX_i^2 = EX^T X = \\ &= EY^T \mathcal{A}^T \mathcal{A} Y = EY^T Y = \sum_{j=1}^k EY_j^2 = DY_1 + \dots + DY_k. \end{aligned}$$

## Факторы

Нормированные главные компоненты  $G_i = \frac{Z_i}{\sqrt{\lambda_i}}$ ,  $i = 1, \dots, k$ ,

## Факторная нагрузка

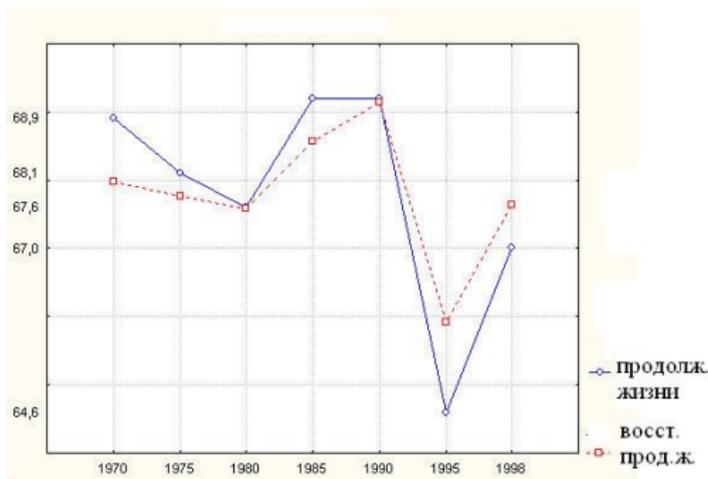
Корреляция  $\beta_{ij} = \text{cor}(X_i, Y_j)$  между признаком  $X_i$  и главной компонентой  $Y_j$

Пусть  $\mathbb{D}X_i = \sigma_i^2$ ,  $\mathbb{D}Y_j = \lambda_j$ . Тогда факторная нагрузка  $\beta_{ij}$  имеет вид

$$\beta_{ij} = \text{cor}(X_i, Y_j) = \frac{EX_i Y_j}{\sigma_i \sqrt{\lambda_j}} = \frac{E \sum_{t=1}^k \alpha_{it} Y_t Y_j}{\sigma_i \sqrt{\lambda_j}} = \frac{\alpha_{ij} \sqrt{\lambda_j}}{\sigma_i} \quad (6)$$

Формулу восстановления  $X_i$  по главным компонентам можно выразить через факторы  $G_j = Y_j / \sqrt{\lambda_j}$ :

$$X_i = \sum_{j=1}^k \alpha_{ij} Y_j = \sigma_i \sum_{j=1}^k \frac{\beta_{ij}}{\sqrt{\lambda_j}} G_j \sqrt{\lambda_j} = \sigma_i \sum_{j=1}^k \beta_{ij} G_j .$$



**Figure:** Восстановление средней продолжительности жизни по первому и второму факторам.

$$\hat{L} = (0.896f_1 + (-0.398f_2)) \cdot 1.64 + 67.8,$$

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. $\lambda$	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

- 1) много бедных, мало чиновников и автомобилей, водки продается мало, продолжительность жизни выше (коррелирует с числом браков) - перестройка.
- 2) много водки и мало автомобилей -1998 - корреляция с числом разводов.

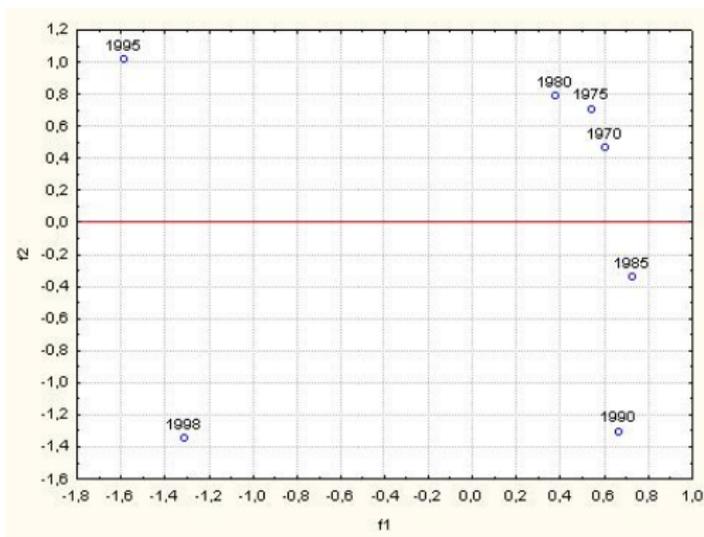


Figure: Двумерная диаграмма фактора 1 (браки) и фактора 2 (разводы).