Факторный анализ

Н.П. Алексеева,

СПбГУ, математико-механический факультет
ск Статистический анализ категориальных признаков

2021 г.

Статистический анализ категориальных данных зависит от того, какого рода задачи решаются.

- Изучение зависимости между двумя категориальными признаками
- Изучение влияния совокупности категориальных признаков $X_1, \ldots, X_p$ на явление, которое описывается 1) одной количественной, 2) одной категориальной, 3) совокупностью количественных, 4) совокупностью категориальных, 5) смесью количественных и категориальных
- все наоборот

Вариант многомерного шкалирования через факторный анализ матриц инцидентности

Замена многомерных категориальных данных метрическими меньшей размерности при условии наилучшего сохранения структуры расстояний между объектами.

Основной подход – замена одной категориальной переменной с p градациями на p столбцов инцидентности и дальнейшее снижение размерности с наименьшей потерей информации.

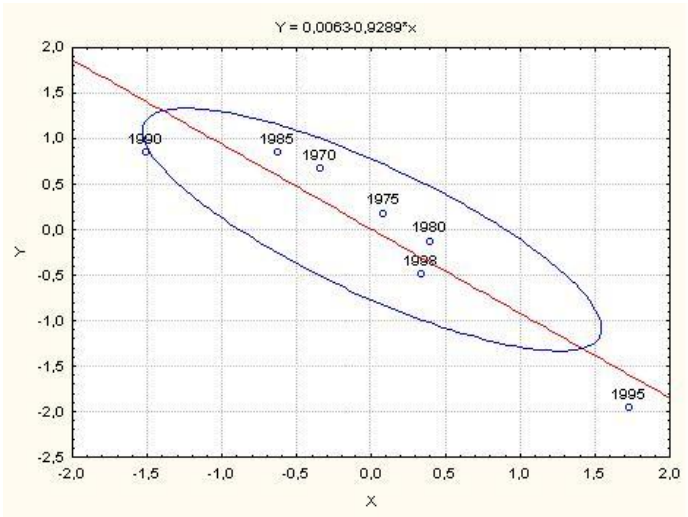
Факторный анализ на основе метода главных компонент — инструмент редукции размерности с наименьшими потерями информации в смысле различимости объектов.

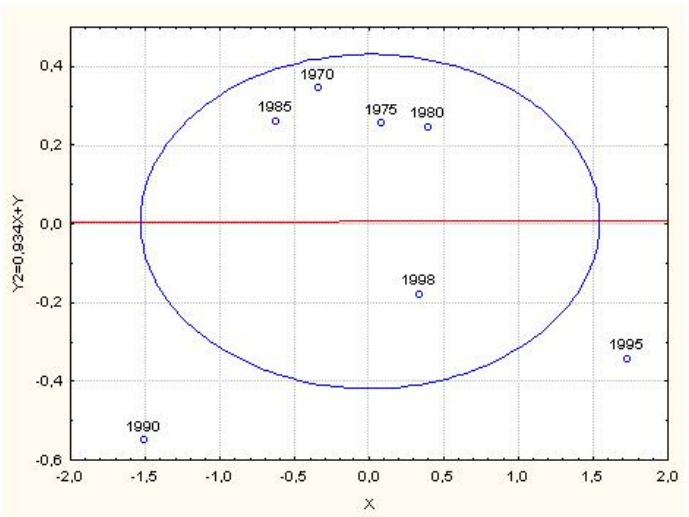
Линейные преобразования, приводящие к некоррелированным компонентам

X, Y центрированные и нормированные величины:
Y – продолжительность жизни X – объемы продаж водки,

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

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





Преобразование, сохраняющее суммарную дисперсию

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

Главные компоненты

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

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

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

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

Свойства преобразования

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

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

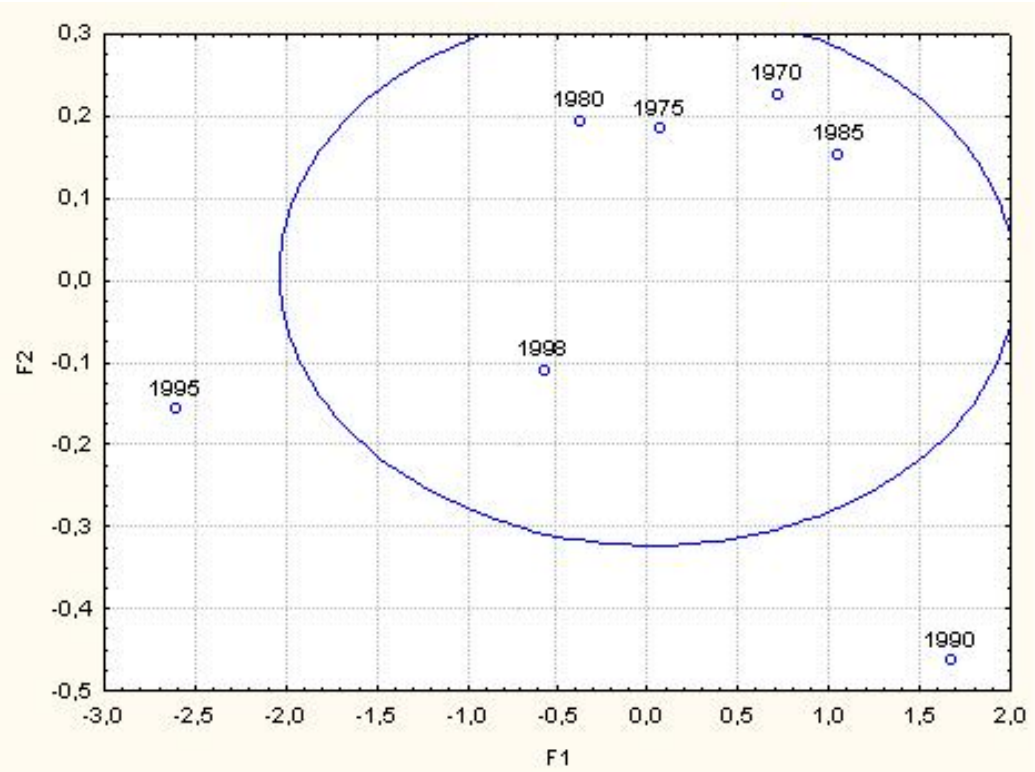
Расстояние между точками $\varrho(a, b) = (a - b)^T(a - b)$

$$\varrho(a', b') = (a' - b')^T(a' - b') = (Ca - Cb)^T(Ca - Cb) =$$
$$= (a - b)^T C^T C(a - b) = \varrho(a, b)\,.$$

Интерпретация главных компонент

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

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



Экстремальные свойства собственных векторов

Задача 1

Показать, что для любого нормированного вектора $[a, b], a^2 + b^2 = 1$,

$$[a, b] \cdot \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \cdot \begin{bmatrix} a \\ b \end{bmatrix} \leq 1 + r$$

Задача 2

Показать, что для для любого нормированного вектора $[a, b], a^2 + b^2 = 1$, дисперсия линейной комбинации $aX_1 + bX_2$ будет не больше максимального собственного числа матрицы $\Sigma = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$, равного $1 + r$.

Идея метода главных компонент

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.

Определение главных компонент

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

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

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

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

Вторая главная компонента также является линейной комбинацией исходных признаков:

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

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

Определение главных компонент

Аналогично строятся остальные главные компоненты

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

Суммарная дисперсия остается неизменной:

$$V = \mathbb{D}(X_1) + \ldots \mathbb{D}(X_k) = \lambda_1 + \ldots + \lambda_k.$$

Значимость главных компонент определяется долей объясняемой ими дисперсии, равной $\frac{\lambda_i}{V} \cdot 100\%$.
Факторами называются нормированные главные компоненты: $Y_j/\sqrt{\lambda_i}$.

Вычисление главных компонент

Пусть $X = (X_1, \ldots, X_k)^T$, $\mathbb{E}X_i = 0$ и $\Sigma = \mathbb{E}XX^T$. Обозначим через A_i собственные векторы матрицы Σ, соответствующие собственным числам λ_i

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

Упорядочим собственные числа: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$.
и определим j-ю главную компоненту как

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

Формула восстановления

Соберем собственные вектора A_i в одну ортогональную матрицу

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

и главные компоненты в один вектор

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

Теорема

.

$$1.\ \mathbb{E}Y_iY_j = \begin{cases} \lambda_j, & \text{при } i = j, \\ 0, & \text{при } i \neq j \end{cases}, \qquad 2.\ \sum_{i=1}^{k} \mathbb{D}X_i = \sum_{j=1}^{k} \mathbb{D}Y_j.$$

$$1.\mathbb{E}Y_iY_j = \mathbb{E}Y_iY_j^T \overset{(4)}{=} \mathbb{E}A_i^TX(A_j^TX)^T = A_i^T\Sigma A_j \overset{(3)}{=}$$

$$= A_i^T\lambda_jA_j = \begin{cases} \lambda_j, & \text{при } i = j, \\ 0, & \text{при } i \neq j. \end{cases} \qquad \mathbb{D}Y_j = \lambda_j\,.$$

$$2.\ \mathbb{D}X_1 + \ldots + \mathbb{D}X_k = \sum_{i=1}^{k} \mathbb{E}X_i^2 = \mathbb{E}X^TX =$$

$$= \mathbb{E}Y^T\mathcal{A}^T\mathcal{A}Y = \mathbb{E}Y^TY = \sum_{j=1}^{k} \mathbb{E}Y_j^2 = \mathbb{D}Y_1 + \ldots + \mathbb{D}Y_k.$$

Факторы и факторные нагрузки

Факторы

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

Факторная нагрузка

Корреляция $\beta_{ij} = \mathrm{cor}(X_i, Y_j)$ 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 β_{ij} имеет вид

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

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

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

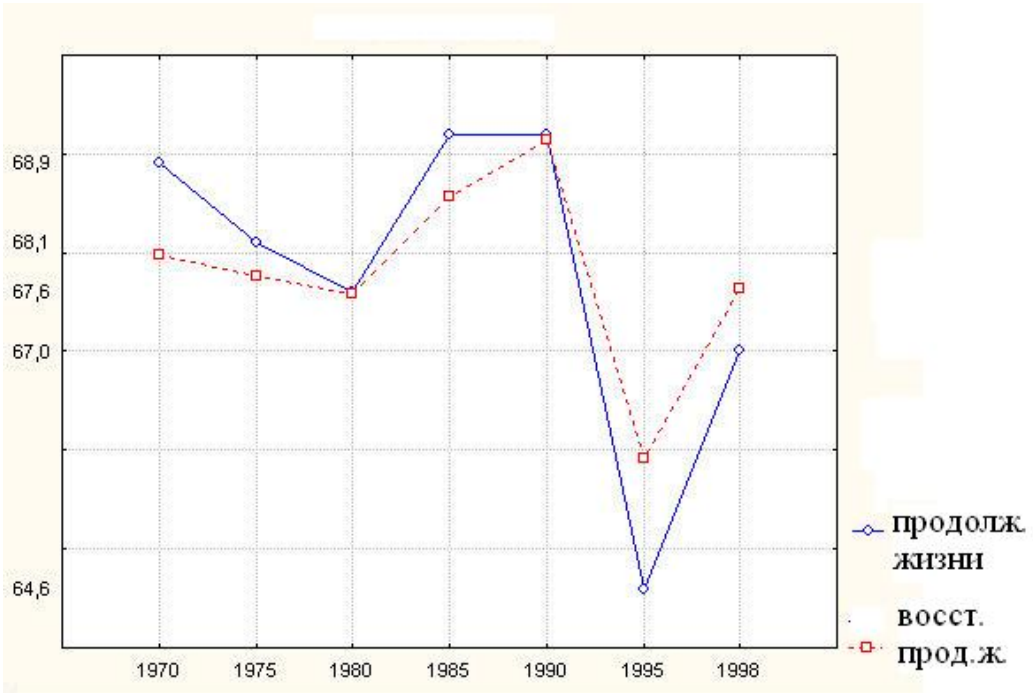


Figure: Восстановление средней продолжительности жизни по первому и второму факторам.

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

Интерпретация результатов

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

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.

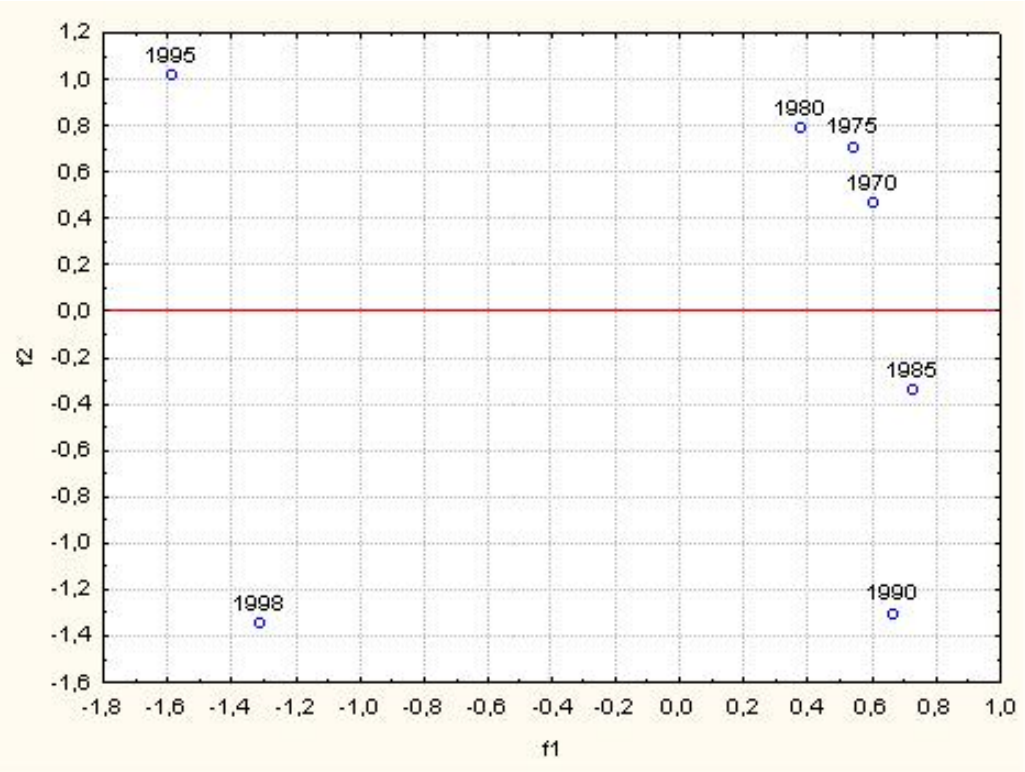


Figure: Двумерная диаграмма фактора 1 (браки) и фактора 2 (разводы).

X матрица инцидентности данных

	ξ_1	ξ_2	ξ_3
1	a	c	u
2	b	c	u
3	a	d	u
4	b	c	v

	a	b	c	d	u	v
1	1	0	1	0	1	0
2	0	1	1	0	1	0
3	1	0	0	1	0	1
4	0	1	1	0	0	1

```
pc<-princomp(X[,c(1,3,5)])
pc$sdev
```

	Comp.1	Comp.2	Comp.3
pc.sdev	0.63	0.50	0.20

pc$loadings[,seq(2)]

	Comp.1	Comp.2
a	-0.54	0.71
c	0.64	-0.00
u	0.54	0.71

pc$scores[,seq(2)]

	Comp.1	Comp.2
1	0.16	0.71
2	0.70	0.00
3	-1.02	0.00
4	0.16	-0.71

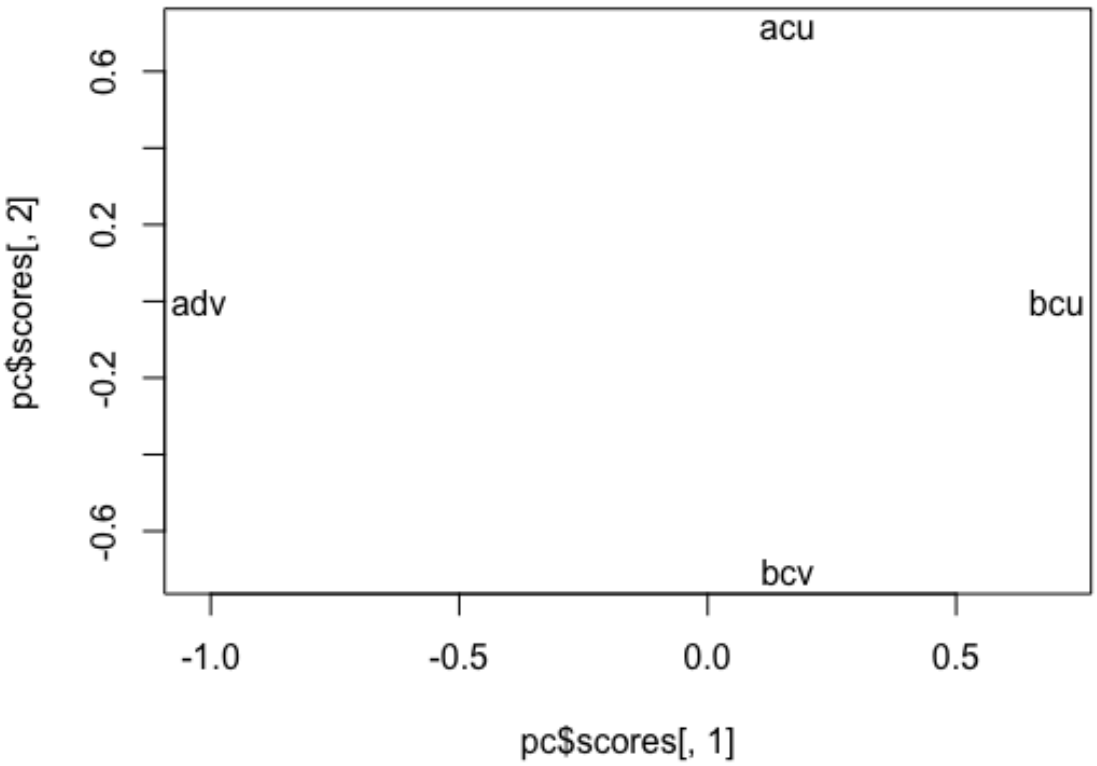


Figure: Двумерная диаграмма факторов