Национальный исследовательский ядерный университет «МИФИ»

Факультет бизнес-информатики и управления комплексными системами

Кафедра экономики и менеджмента в промышленности (№ 71)

Математические и инструментальные методы обработки статистической информации

ЛЕКЦИЯ 2

ПОДГОТОВКА ДАННЫХ. ФАКТОРНЫЙ АНАЛИЗ

Киреев В.С.,

к.т.н., доцент

v.kireev@inbox.ru

Москва,
2017

Нормализация

- Десятичное масштабирование
- Минимаксная нормализация
- Нормализация с помощью стандартного преобразования
- Нормализация с помощью поэлементных преобразований

Десятичное масштабирование

$$V_i' = \frac{V_i}{10^k},\ \max(V_i') < 1$$

Минимаксная нормализация

$$V_i' = \frac{V_i - \min\limits_i (V_i)}{\max\limits_i (V_i) - \min\limits_i (V_i)}$$

Нормализация с помощью стандартного отклонения

$$V_i' = \frac{V_i - \overline{V}}{\sigma_V}$$

$\overline{V}$ - **выборочное среднее**

σ_V - **выборочное среднее квадратическое отклонение**

Нормализация с помощью поэлементных преобразований

$$V_i' = f(V_i)$$

$$V_i' = 1/\log(V_i), \quad V_i' = \log(V_i)$$

$$V_i' = \exp(V_i)$$

$$V_i' = V_i^y, \quad V_i' = 1/V_i^y$$

Факторный анализ

Факторный анализ (ФА) представляет собой совокупность методов, которые на основе реально существующих связей анализируемых признаков, связей самих наблюдаемых объектов, позволяют выявлять скрытые (неявные, латентные) обобщающие характеристики организационной структуры и механизма развития изучаемых явлений, процессов.

Методы факторного анализа в исследовательской практике применяются главным образом с целью сжатия информации, получения небольшого числа обобщающих признаков, объясняющих вариативность (дисперсию) элементарных признаков (**R-техника** факторного анализа) или вариативность наблюдаемых объектов (**Q-техника** факторного анализа).

Алгоритмы факторного анализа основываются на использовании редуцированной матрицы парных корреляций (ковариаций). Редуцированная матрица – это матрица, на главной диагонали которой расположены не единицы (оценки) полной корреляции или оценки полной дисперсии, а их редуцированные, несколько уменьшенные величины. При этом постулируется, что в результате анализа будет объяснена не вся дисперсия изучаемых признаков (объектов), а ее некоторая часть, обычно большая. Оставшаяся необъясненная часть дисперсии — это характерность, возникающая из-за специфичности наблюдаемых объектов, или ошибок, допускаемых при регистрации явлений, процессов, т.е. ненадежности вводных данных.

Классификация методов ФА

Метод главных компонент

Метод главных компонент (МГК) применяется для снижения размерности пространства наблюдаемых векторов, не приводя к существенной потере информативности. Предпосылкой МГК является нормальный закон распределения многомерных векторов. В МГК линейные комбинации случайных величин определяются характеристическими векторами ковариационной матрицы. Главные компоненты представляют собой ортогональную систему координат, в которой дисперсии компонент характеризуют их статистические свойства. МГК не относят к ФА, хотя он имеет схожий алгоритм и решает схожие аналитические задачи. Его главное отличие заключается в том, что обработке подлежит не редуцированная, а обычная матрица парных корреляций, ковариаций, на главной диагонали которой расположены единицы.

Пусть дан исходный набор векторов X линейного пространства L^k. Применение метода главных компонент позволяет перейти к базису пространства L^m ($m \leq k$), такому что: первая компонента (первый вектор базиса) соответствует направлению, вдоль которого дисперсия векторов исходного набора максимальна. Направление второй компоненты (второго вектора базиса) выбрано таким образом, чтобы дисперсия исходных векторов вдоль него была максимальной при условии ортогональности первому вектору базиса. Аналогично определяются остальные векторы базиса. В результате, направления векторов базиса выбраны так, чтобы максимизировать дисперсию исходного набора вдоль первых компонент, называемых главными компонентами (или главными осями). Получается, что основная изменчивость векторов исходного набора векторов представлена несколькими первыми компонентами, и появляется возможность, отбросив менее существенные компоненты, перейти к пространству меньшей размерности.

Метод главных компонент. Схема

Пусть имеется матрица переменных **X** размерностью ($I \times J$), где I – число образцов (строк), а J – это число независимых переменных (столбцов), которых, как правило, много ($J \gg 1$). В методе главных компонент используются новые, формальные переменные $\mathbf{t}_a$ ($a=1,\ldots A$), являющиеся линейной комбинацией исходных переменных $\mathbf{x}_j$ ($j=1,\ldots J$)

$$\mathbf{t}_a = \mathbf{p}_{a1}\mathbf{x}_1 + \ldots + \mathbf{p}_{aJ}\mathbf{x}_J \qquad (1)$$

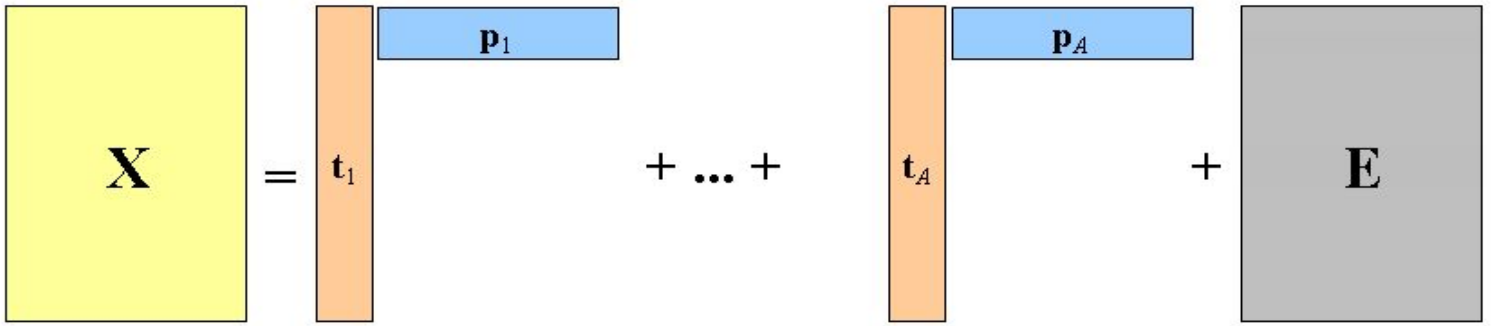
С помощью этих новых переменных матрица **X** разлагается в произведение двух матриц **T** и **P** –

$$\mathbf{X} = \mathbf{T}\mathbf{P}^t + \mathbf{E} = \sum_{a=1}^{A} \mathbf{t}_a \mathbf{p}_a^t + \mathbf{E} \qquad (2)$$

Матрица **T** называется матрицей *счетов* (scores). Ее размерность ($I \times A$).

Матрица **P** называется матрицей *нагрузок* (loadings). Ее размерность ($J \times A$).

E – это матрица *остатков*, размерностью ($I \times J$).

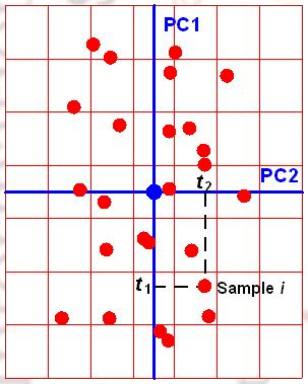


Новые переменные $\mathbf{t}_a$ называются *главными компонентами* (Principal Components), поэтому и сам метод называется методом главных компонент (PCA). Число столбцов – $\mathbf{t}_a$ в матрице **T**, и $\mathbf{p}_a$ в матрице **P**, равно A, которое называется *числом главных компонент* (PC). Эта величина заведомо меньше числа переменных J и числа образцов I.

Метод главных компонент. Матрица счетов

Матрица счетов **T** дает нам проекции исходных образцов (*J* –мерных векторов $\mathbf{x}_1,\ldots,\mathbf{x}_I$) на подпространство главных компонент (*A*-мерное). Строки $\mathbf{t}_1,\ldots,\mathbf{t}_I$ матрицы **T** – это координаты образцов в новой системе координат. Столбцы $\mathbf{t}_1,\ldots,\mathbf{t}_A$ матрицы **T** – ортогональны и представляют проекции всех образцов на одну новую координатную ось.

При исследовании данных методом PCA, особое внимание уделяется графикам счетов. Они несут в себе информацию, полезную для понимания того, как устроены данные. На графике счетов каждый образец изображается в координатах ($\mathbf{t}_i$, $\mathbf{t}_j$), чаще всего – ($\mathbf{t}_1$, $\mathbf{t}_2$), обозначаемых PC1 и PC2. Близость двух точек означает их схожесть, т.е. положительную корреляцию. Точки, расположенные под прямым углом, являются некоррелированными, а расположенные диаметрально противоположно – имеют отрицательную корреляцию.

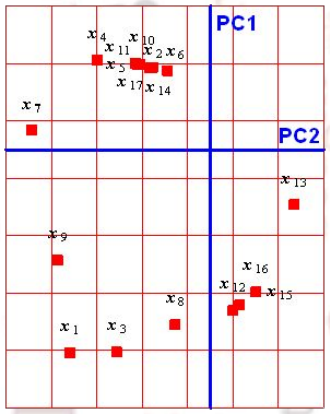


$$\mathbf{T}=\begin{pmatrix} t_{11} & t_{12} & \ldots & t_{1a} & \ldots & t_{1A} \\ t_{21} & t_{22} & \ldots & t_{2a} & \ldots & t_{2A} \\ t_{i1} & t_{i2} & \ldots & t_{ia} & \ldots & t_{iA} \\ t_{I1} & t_{I2} & \ldots & t_{Ia} & \ldots & t_{IA} \end{pmatrix}$$

Метод главных компонент. Матрица нагрузок

Матрица нагрузок **P** – это матрица перехода из исходного пространства переменных $\mathbf{x}_1, \ldots \mathbf{x}_J$ (*J*-мерного) в пространство главных компонент (*A*-мерное). Каждая строка матрицы **P** состоит из коэффициентов, связывающих переменные **t** и **x**. Например, *a*-я строка – это проекция всех переменных $\mathbf{x}_1, \ldots \mathbf{x}_J$ на *a*-ю ось главных компонент. Каждый столбец **P** – это проекция соответствующей переменной $\mathbf{x}_j$ на новую систему координат.

График нагрузок применяется для исследования роли переменных. На этом графике каждая переменная $\mathbf{x}_j$ отображается точкой в координатах ($\mathbf{p}_i$, $\mathbf{p}_j$), например ($\mathbf{p}_1$, $\mathbf{p}_2$). Анализируя его аналогично графику счетов, можно понять, какие переменные связаны, а какие независимы. Совместное исследование парных графиков счетов и нагрузок, также может дать много полезной информации о данных.



$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1j} & \cdots & p_{1J} \\ p_{21} & p_{22} & \cdots & p_{2j} & \cdots & p_{2J} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{a1} & p_{a2} & \cdots & p_{aj} & \cdots & p_{aJ} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ p_{A1} & p_{A2} & \cdots & p_{Aj} & \cdots & p_{AJ} \end{pmatrix}$$

Особенности метода главных компонент

В основе метода главных компонент лежат **следующие допущения**:

- допущение о том, что размерность данных может быть эффективно понижена путем линейного преобразования;
- допущение о том, что больше всего информации несут те направления, в которых дисперсия входных данных максимальна.

Можно легко видеть, что эти условия далеко не всегда выполняются. Например, если точки входного множества располагаются на поверхности гиперсферы, то никакое линейное преобразование не сможет понизить размерность (но с этим легко справится нелинейное преобразование, опирающееся на расстояние от точки до центра сферы). Это недостаток в равной мере свойственен всем линейным алгоритмам и может быть преодолен за счет использования дополнительных фиктивных переменных, являющихся нелинейными функциями от элементов набора входных данных (т.н. kernel trick).

Второй недостаток метода главных компонент состоит в том, что направления, максимизирующие дисперсию, далеко не всегда максимизируют информативность. Например, переменная с максимальной дисперсией может не нести почти никакой информации, в то время как переменная с минимальной дисперсией позволяет полностью разделить классы. Метод главных компонент в данном случае отдаст предпочтение первой (менее информативной) переменной. Вся дополнительная информация, связанная с вектором (например, принадлежность образа к одному из классов), игнорируется.

Пример данных для МГК

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
39			**Autoscaled Data**											
40			**Height**	**Weight**	**Hair**	**Shoes**	**Age**	**Income**	**Beer**	**Wine**	**Sex**	**Strength**	**Region**	**IQ**
41	**1**	**MN**	2.47	1.81	-0.98	2.08	1.43	1.97	1.88	-0.34	-0.98	2.25	-0.98	-1.24
42	**2**	**MN**	1.08	1.29	-0.98	1.05	-0.15	0.62	1.11	-0.60	-0.98	1.43	-0.98	1.22
43	**3**	**MN**	0.98	1.22	-0.98	1.05	0.27	0.73	0.78	-0.68	-0.98	1.30	-0.98	0.98
44	**4**	**MN**	0.88	1.02	-0.98	0.54	0.06	0.29	1.64	-1.35	-0.98	0.48	-0.98	2.04
45	**5**	**MN**	0.68	1.02	-0.98	0.79	0.16	0.29	1.53	-1.39	-0.98	0.34	-0.98	1.14
46	**6**	**MN**	0.98	1.09	-0.98	0.54	0.27	0.85	1.05	-1.75	-0.98	1.16	-0.98	-0.83
47	**7**	**MN**	0.68	1.15	-0.98	1.05	0.90	1.07	1.16	-1.00	-0.98	0.89	-0.98	-0.50
48	**8**	**MN**	0.68	1.09	-0.98	1.05	1.21	1.63	1.24	-0.84	-0.98	0.61	-0.98	-0.17
49	**9**	**MS**	1.18	1.15	-0.98	1.31	-0.89	-1.28	0.50	0.98	-0.98	1.43	0.98	-0.50
50	**10**	**MS**	1.38	1.29	-0.98	1.56	-0.78	-1.22	0.55	0.94	-0.98	1.84	0.98	0.32
51	**11**	**MS**	0.39	0.03	-0.98	0.28	-0.89	-1.06	-0.45	0.57	-0.98	0.61	0.98	0.40
52	**12**	**MS**	0.68	0.50	-0.98	0.79	-0.15	-0.94	-0.15	0.88	-0.98	0.48	0.98	-0.01
53	**13**	**MS**	0.78	0.69	-0.98	0.79	0.79	0.40	-0.57	0.59	-0.98	0.20	0.98	-0.83
54	**14**	**MS**	0.29	0.23	-0.98	0.54	1.64	0.96	-0.60	0.92	-0.98	0.07	0.98	-1.57
55	**15**	**MS**	0.19	0.17	0.98	0.54	2.16	1.18	-0.71	1.12	-0.98	-0.20	0.98	-0.83
56	**16**	**MS**	0.48	0.69	-0.98	0.54	-0.47	-0.38	-0.51	1.54	-0.98	-0.07	0.98	0.24
57	**17**	**FN**	-0.71	-1.15	-0.98	-1.00	-0.26	0.06	0.23	-1.08	0.98	-0.89	-0.98	-0.26
58	**18**	**FN**	-0.31	-0.29	0.98	-0.49	-1.20	-0.83	0.69	-0.66	0.98	-0.07	-0.98	-0.42
59	**19**	**FN**	-0.11	-0.03	0.98	-0.23	-1.10	-0.61	0.65	-0.82	0.98	0.07	-0.98	-1.08
60	**20**	**FN**	-0.41	-0.89	0.98	-1.00	-1.10	-0.50	0.01	-0.86	0.98	-0.48	-0.98	-1.41
61	**21**	**FN**	-0.51	-0.82	0.98	-0.75	-0.78	-0.44	0.12	-0.92	0.98	-0.48	-0.98	-1.24
62	**22**	**FN**	-1.60	-1.15	0.98	-1.00	-0.26	0.51	-0.16	-0.80	0.98	-1.57	-0.98	0.98
63	**23**	**FN**	-0.91	-0.95	0.98	-0.49	0.69	0.73	0.06	0.05	0.98	-0.75	-0.98	-1.16
64	**24**	**FN**	-1.11	-1.02	0.98	-0.75	0.58	0.73	0.17	-0.15	0.98	-0.89	-0.98	-0.59
65	**25**	**FS**	-0.51	-0.95	0.98	-0.75	1.53	0.73	-0.88	0.61	0.98	-0.75	0.98	1.63
66	**26**	**FS**	-0.71	-1.02	0.98	-1.00	-1.41	-1.50	-1.10	2.29	0.98	-0.89	0.98	0.65
67	**27**	**FS**	-1.50	-1.22	0.98	-1.52	-0.47	-1.06	-1.43	-0.23	0.98	-1.57	0.98	0.32
68	**28**	**FS**	-1.01	-0.95	0.98	-1.00	-1.73	-1.84	-1.18	0.09	0.98	-0.89	0.98	-1.08
69	**29**	**FS**	-1.11	-0.95	0.98	-1.00	-1.52	-1.78	-1.29	0.29	0.98	-1.02	0.98	1.39
70	**30**	**FS**	-0.81	-0.89	0.98	-1.00	0.16	-0.16	-1.42	-0.05	0.98	-0.75	0.98	0.89
71	**31**	**FS**	-1.21	-1.08	0.98	-1.26	0.69	0.45	-1.47	1.30	0.98	-0.89	0.98	0.40
72	**32**	**FS**	-1.31	-1.08	0.98	-1.26	0.58	0.40	-1.45	1.34	0.98	-1.02	0.98	1.14
73	**mean**		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
74	**STD**		1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
75														

К. Эсбенсен. Анализ многомерных данных, сокр. пер. с англ. под ред. О. Родионовой, Из-во ИПХФ РАН, 2005

Пример данных для МГК. Обозначения

Height	Рост: в сантиметрах
Weight	Вес: в килограммах
Hair	Волосы: короткие: –1, или длинные: +1
Shoes	Обувь: размер по европейскому стандарту
Age	Возраст: в годах
Income	Доход: в тысячах евро в год
Beer	Пиво: потребление в литрах в год
Wine	Вино: потребление в литрах в год
Sex	Пол: мужской: –1, или женский: +1
Strength	Сила: индекс, основанный на проверке физических способностей
Region	Регион: север : –1, или юг: +1
IQ	Коэффициент интеллекта, измеряемый по стандартному тесту

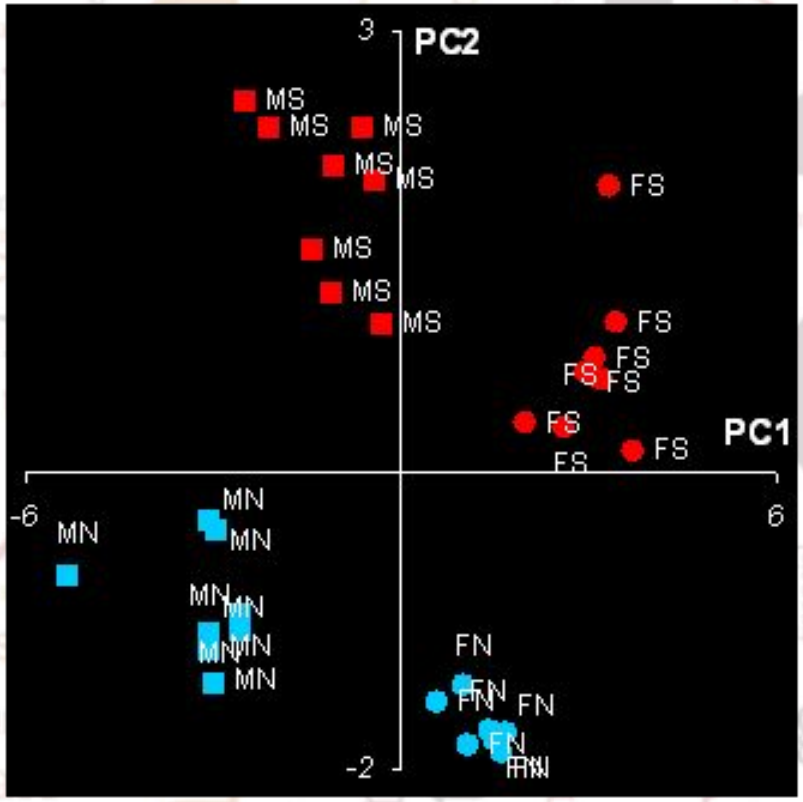
Матрица счетов

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		**PCA**												
2		T Scores												
3		PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	
4	1	=ScoresPCA(Xraw,12,3)				1.060	-0.017	0.563	0.085	0.280	0.078	0.109	0.133	
5	2	-3.114	-0.293	0.671	1.310	0.435	0.119	0.109	-0.456	-0.088	-0.034	0.051	-0.009	
6	3	-2.997	-0.360	0.212	1.117	0.204	-0.017	0.120	-0.469	-0.149	-0.018	0.184	-0.208	
7	4	-2.591	-0.928	0.863	2.321	-0.095	-0.076	-0.270	0.306	0.240	-0.192	-0.168	0.005	
8	5	-2.588	-1.037	0.686	1.461	-0.347	-0.098	-0.447	0.345	0.083	0.021	-0.043	-0.005	
9	6	-3.027	-1.400	0.284	-0.440	-0.633	-0.538	0.252	-0.343	0.042	-0.132	-0.247	0.020	
10	7	-3.095	-1.069	-0.472	-0.144	-0.304	-0.059	-0.147	-0.025	-0.217	0.150	-0.077	-0.088	
11	8	-3.113	-1.181	-1.068	0.242	-0.232	0.141	-0.211	-0.008	-0.038	0.203	-0.060	0.087	
12	9	-2.130	2.356	1.503	-0.858	0.284	0.070	0.020	0.246	-0.171	0.095	-0.123	0.060	
13	10	-2.510	2.525	1.542	-0.105	0.653	-0.087	0.222	0.228	-0.305	0.078	0.024	0.035	
14	11	-0.463	1.992	1.090	0.206	-0.617	0.084	0.138	-0.186	-0.202	-0.421	0.069	0.174	
15	12	-1.098	2.094	0.496	-0.198	-0.376	0.127	-0.126	0.299	-0.071	-0.122	0.112	-0.115	
16	13	-1.438	1.527	-1.059	-0.788	-0.734	-0.124	0.015	-0.060	0.264	0.185	0.156	-0.057	
17	14	-1.137	1.241	-2.200	-1.404	-0.953	0.200	0.033	0.063	-0.075	-0.002	-0.058	-0.009	
18	15	-0.334	1.034	-2.931	-0.866	0.420	-0.654	-0.781	0.037	0.016	-0.283	0.060	-0.004	
19	16	-0.652	2.360	0.125	0.055	-0.334	0.674	-0.398	-0.283	0.359	0.192	-0.007	-0.007	
20	17	1.084	-1.845	0.409	0.123	-1.323	0.872	0.704	0.328	0.032	-0.029	0.078	-0.026	
21	18	0.981	-1.434	1.645	-0.526	0.714	-0.035	-0.096	0.127	-0.005	0.003	-0.124	-0.053	
22	19	0.567	-1.551	1.474	-1.154	0.677	-0.234	-0.047	0.031	0.068	0.213	-0.058	-0.052	
23	20	1.663	-1.762	1.122	-1.394	0.018	-0.081	0.085	-0.218	0.224	-0.284	0.028	0.004	
24	21	1.486	-1.813	0.904	-1.208	0.070	-0.147	-0.003	-0.023	0.043	-0.160	0.109	-0.034	
25	22	2.464	-2.040	-0.222	1.202	-0.140	0.268	-0.491	-0.185	-0.091	0.166	0.191	0.233	
26	23	1.396	-1.736	-1.130	-1.041	0.367	0.487	-0.170	0.133	-0.220	-0.007	0.123	-0.074	
27	24	1.622	-1.894	-0.992	-0.419	0.287	0.463	-0.204	0.141	-0.225	-0.048	-0.033	0.015	
28	25	2.005	0.344	-2.085	1.549	0.603	-0.377	0.509	0.471	0.105	-0.061	0.070	0.020	
29	26	3.335	1.956	0.851	0.063	0.884	0.897	-0.141	-0.048	0.252	-0.128	-0.057	-0.046	
30	27	3.711	0.147	0.195	0.279	-0.774	-0.624	-0.091	0.120	0.002	0.081	-0.114	-0.074	
31	28	3.207	0.630	1.602	-1.358	-0.420	-0.480	-0.057	-0.047	-0.001	0.168	-0.026	0.112	
32	29	3.423	1.021	1.482	1.036	0.016	-0.395	-0.084	-0.024	-0.035	0.124	0.177	-0.027	
33	30	2.615	0.320	-0.600	0.784	-0.038	-0.789	0.455	-0.109	0.045	0.093	0.108	-0.007	
34	31	2.958	0.688	-1.728	0.267	0.268	0.181	0.320	-0.246	-0.082	0.016	-0.249	-0.012	
35	32	3.102	0.788	-1.603	0.988	0.359	0.249	0.220	-0.232	-0.078	0.055	-0.203	0.009	
36														

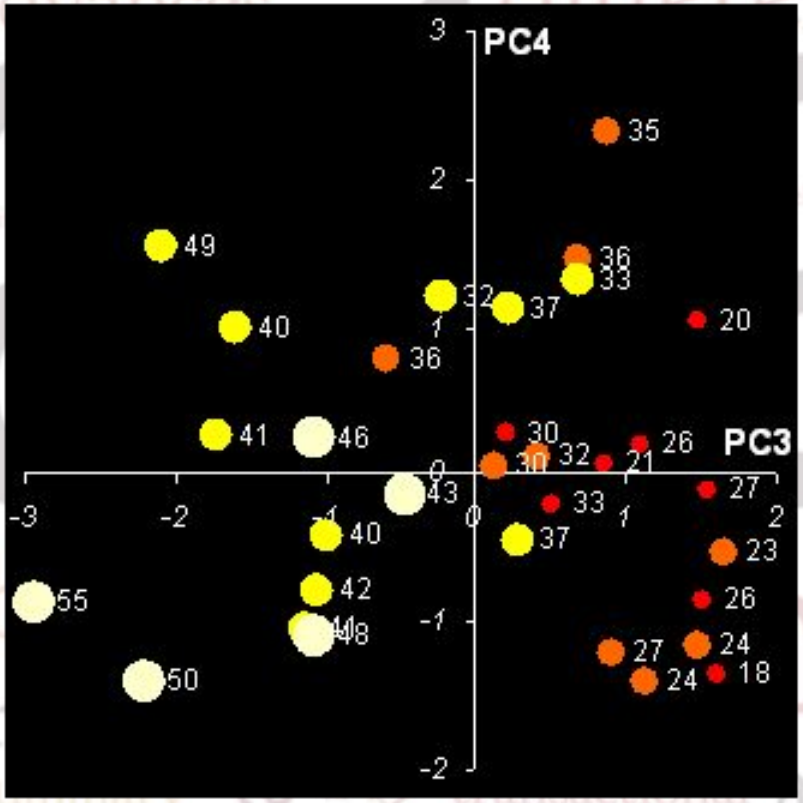
Матрица нагрузок

	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB
2			P Loadings												
3			PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	
4		Height	=LoadingsPCA(Xraw,12,3)				0.186	-0.124	0.268	0.118	0.729	-0.307	0.255	-0.065	
5		Weight	-0.381	0.111	0.068	0.033	0.100	-0.192	-0.224	-0.219	0.190	0.572	-0.415	-0.395	
6		Hair	0.338	-0.150	-0.079	-0.114	0.660	-0.489	-0.368	-0.081	0.041	-0.140	0.007	0.078	
7		Shoes	-0.378	0.151	0.001	-0.066	0.152	-0.031	-0.234	0.171	-0.280	0.376	0.685	0.183	
8		Age	-0.143	-0.061	-0.720	0.055	-0.029	-0.165	0.043	0.435	-0.174	-0.134	-0.036	-0.430	
9		Income	-0.190	-0.287	-0.586	0.085	0.063	0.137	0.129	-0.434	0.180	0.167	-0.038	0.492	
10		Beer	-0.325	-0.308	0.188	0.040	0.231	0.239	-0.170	0.567	-0.015	-0.049	-0.420	0.350	
11		Wine	0.124	0.554	-0.212	-0.125	0.415	0.638	-0.120	-0.040	0.024	-0.054	-0.095	-0.093	
12		Sex	0.352	-0.232	0.052	-0.051	0.313	0.098	0.580	0.254	0.078	0.529	0.084	-0.124	
13		Strength	-0.365	0.112	0.135	-0.081	0.336	-0.160	0.512	-0.258	-0.530	-0.232	-0.165	-0.001	
14		Region	0.144	0.595	-0.130	-0.022	-0.151	-0.402	0.161	0.265	0.050	0.180	-0.255	0.476	
15		IQ	0.044	0.123	0.062	0.969	0.180	-0.010	0.024	0.001	-0.006	-0.033	0.076	-0.010	
16															

Объекты выборки в пространстве новых компонент

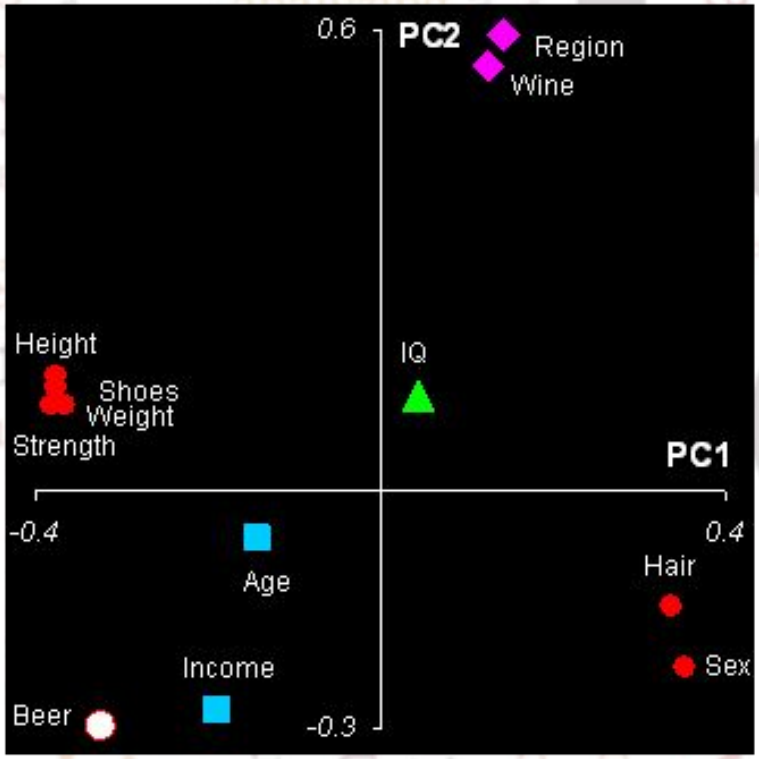


Женщины (F) обозначены кружками ● и ●, а мужчины (M) – квадратами ■ и ■. Север (N) представлен голубым ■, а юг (S) – красным цветом ●.



Размер и цвет символов отражает доход – чем больше и светлее, тем он больше. Числа представляют возраст

Исходные переменные в пространстве новых компонент



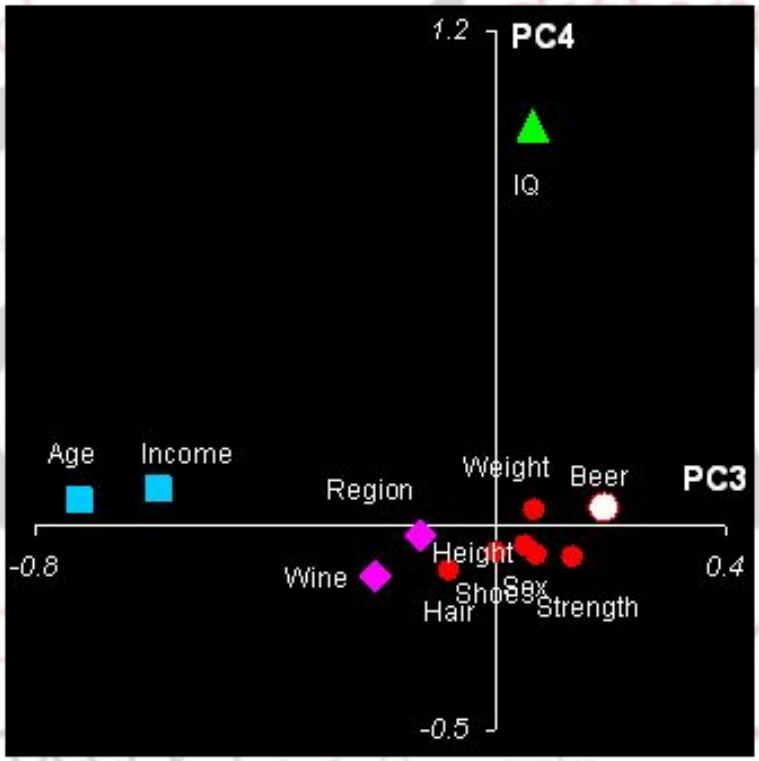
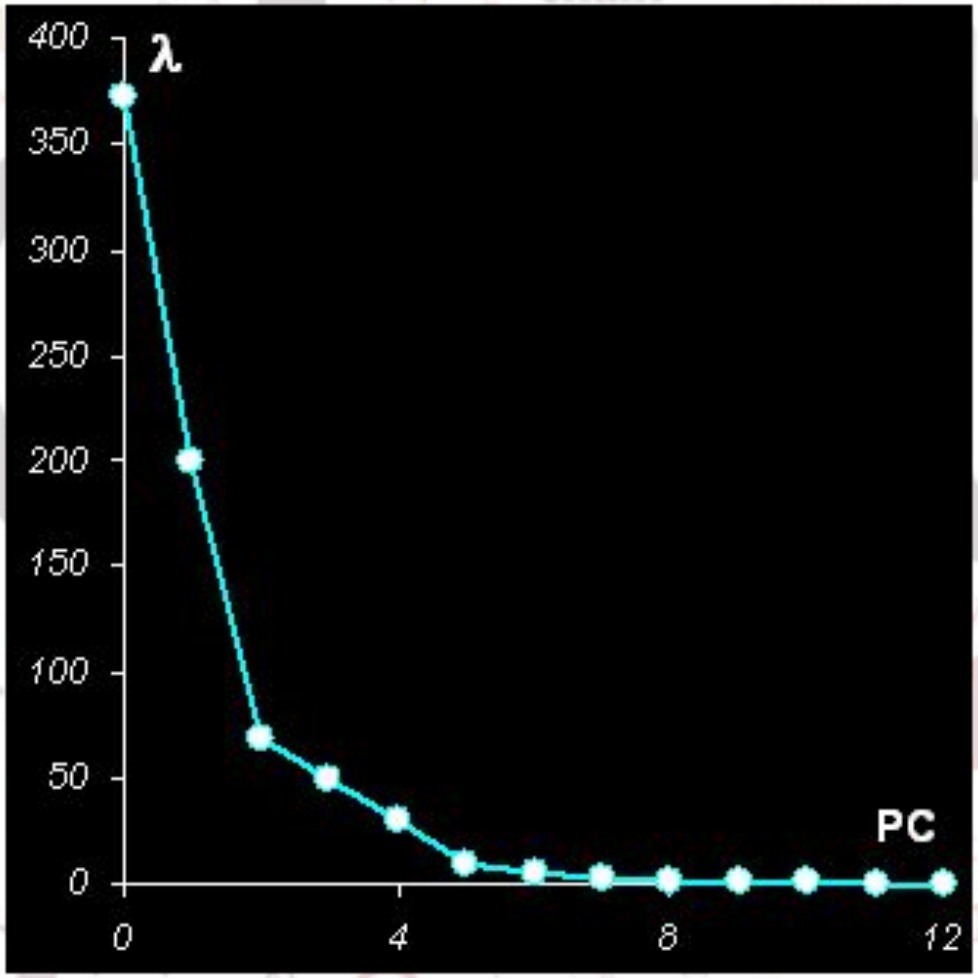


График «каменистой осыпи» (scree plot)



Метод главных факторов

В парадигме метода главных факторов задача снижения размерности признакового пространства выглядит так, что n признаков можно объяснить с помощью меньшего количества m-латентных признаков - общих факторов, где m<<n. Различия между исходными признаками и введёнными общими факторами (линейными комбинациями) учитывают с помощью так называемых характерных факторов.

Конечная цель статистического исследования, проводимого с привлечением аппарата факторного анализа, как правило, состоит в выявлении и интерпретации латентных общих факторов с одновременным стремлением минимизировать как их число, так и степень зависимости от своих специфических остаточных случайных компонент .

Каждый признак является результатом воздействия m гипотетических общих и одного характерного факторов:

$$\begin{cases} X_1 = a_{11} f_1 + a_{12} f_2 + \ldots + a_{1m} f_m + d_1 V_1 \\ X_2 = a_{21} f_1 + a_{22} f_2 + \ldots + a_{2m} f_m + d_2 V \\ \ldots \\ X_n = a_{n1} f_1 + a_{n2} f_2 + \ldots + a_{nm} f_m + d_n V_n \end{cases}$$

a_{ij} - весовые коэффициенты;

f_j - общие факторы, которые подлежат определению;

V_i - характерный фактор для i-ого исходного признака;

d_i - весовой коэффициент при i-ом характерном факторе.

Вращение факторов

Вращение - это способ превращения факторов, полученных на предыдущем этапе, в более осмысленные. Вращение делится на:

- графическое (проведение осей, не применяется при более чем двухмерном анализе),
- аналитическое (выбирается некий критерий вращения, различают ортогональное и косоугольное) и
- матрично-приближенное (вращение состоит в приближении к некой заданной целевой матрице).

Результатом вращения является вторичная структура факторов. Первичная факторная структура (состоящая из первичных нагрузок (полученных на предыдущем этапе) - это, фактически, проекции точек на ортогональные оси координат. Очевидно, что если проекции будут нулевыми, то структура будет проще. А проекции будут нулевыми, если точка лежит на какой-то оси. Таким образом, можно считать вращение переходом от одной системы координат к другой при известных координатах в одной системе(первичные факторы) и итеративно подбираемых координатах в другой системе (вторичные факторы). При получении вторичной структуры стремятся перейти к такой системе координат, чтобы провести через точки (объекты) как можно больше осей, чтобы как можно больше проекции (и соответственно нагрузок) были нулевыми. При этом могут сниматься ограничения ортогональности и убывания значимости от первого к последнему факторам, характерные для первичной структуры.

Ортогональное вращение

Ортогональное вращение подразумевает, что мы будем вращать факторы, но не будем нарушать их ортогональности друг другу. Ортогональное вращение подразумевает умножение исходной матрицы первичных нагрузок на ортогональную матрицу R(такую матрицу, что $|R| = 1, R * R^{-1} = E, R = r \times r$)

Алгоритм ортогонального вращения в общем случае таков:

0. B - матрица первичных факторов.

1. Ищем ортогональную матрицу R^T размера 2*2 для двух столбцов (факторов) b_i и b_j матрицы B такую, что критерий для матрицы $[b_i b_j]$ R максимален.
2. Заменяем столбцы b_i и b_j на столбцы $[b_i b_j] \times R$
3. Проверяем, все ли столбцы перебрали. Если нет, то переход на 1.
4. Проверяем, что критерий для всей матрицы вырос. Если да, то переход на 1. Если нет, то конец алгоритма.

Варимаксное вращение

Этот критерий использует формализацию сложности фактора через дисперсию квадратов нагрузок переменной:

$$v_j = \frac{n\sum_{i=1}^{n}(b_{ij}^4 - (\sum_{i=1}^{n} b_i^2)^2}{n^2}$$

Тогда критерий в общем виде можно записать как:

$$V = \frac{\sum_{j=1}^{r}(n\sum_{i=1}^{n}(b_{ij}^4) - \sum_{j=1}^{r}(\sum_{i=1}^{n} b_i^2)^2}{n^2}$$

При этом, факторные нагрузки могут нормироваться для избавления от влияния отдельных переменных.

Квартимаксное вращение

Формализуем понятие факторной сложности q i-ой переменной через дисперсию квадратов факторных нагрузок факторов:

$$q_i = \frac{1}{r}\sum_{j=1}^{r}(b_{ij}^2 - \overline{b_i^2})^2$$

где r - число столбцов факторной матрицы, b_{ij} - факторная нагрузка j-го фактора на i-ю переменную, - среднее значение. Критерий квартимакс старается максимизировать сложность всей совокупности переменных, чтобы достичь легкости интерпретации факторов (стремится облегчить описание столбцов):

$$Q = \sum_{i=1}^{N} q_i$$

Учитывая, что $\sum_{j=1}^{r}\overline{b_i^2}$ константа (сумма собственных чисел матрицы ковариации) и раскрыв среднее значение (а также учтя, что степенная функция растет пропорционально аргументу), получим окончательный вид критерия для максимизации:

$$Q = \sum_{i=1}^{N}\sum_{j=1}^{r} b_{ij}$$

Критерии определения числа факторов

Главной проблемой факторного анализа является выделение и интерпретация главных факторов. При отборе компонент исследователь обычно сталкивается с существенными трудностями, так как не существует однозначного критерия выделения факторов, и потому здесь неизбежен субъективизм интерпретаций результатов. Существует несколько часто употребляемых критериев определения числа факторов. Некоторые из них являются альтернативными по отношению к другим, а часть этих критериев можно использовать вместе, чтобы один дополнял другой:

- **Критерий Кайзера** или критерий собственных чисел. Этот критерий предложен Кайзером, и является, вероятно, наиболее широко используемым. Отбираются только факторы с собственными значениями равными или большими 1. Это означает, что если фактор не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается.

- **Критерий каменистой осыпи (англ. scree)** или критерий отсеивания. Он является графическим методом, впервые предложенным психологом Кэттелом. Собственные значения возможно изобразить в виде простого графика. Кэттел предложил найти такое место на графике, где убывание собственных значений слева направо максимально замедляется. Предполагается, что справа от этой точки находится только «факториальная осыпь» — «осыпь» является геологическим термином, обозначающим обломки горных пород, скапливающиеся в нижней части скалистого склона.

Критерии определения числа факторов. Продолжение

- **Критерий значимости**. Он особенно эффективен, когда модель генеральной совокупности известна и отсутствуют второстепенные факторы. Но критерий непригоден для поиска изменений в модели и реализуем только в факторном анализе по методу наименьших квадратов или максимального правдоподобия.
- **Критерий доли воспроизводимой дисперсии**. Факторы ранжируются по доле детерминируемой дисперсии, когда процент дисперсии оказывается несущественным, выделение следует остановить. Желательно, чтобы выделенные факторы объясняли более 80 % разброса. Недостатки критерия: во-первых, субъективность выделения, во-вторых, специфика данных может быть такова, что все главные факторы не смогут совокупно объяснить желательного процента разброса. Поэтому главные факторы должны вместе объяснять не меньше 50,1 % дисперсии.
- **Критерий интерпретируемости и инвариантности**. Данный критерий сочетает статистическую точность с субъективными интересами. Согласно ему, главные факторы можно выделять до тех пор, пока будет возможна их ясная интерпретация. Она, в свою очередь, зависит от величины факторных нагрузок, то есть если в факторе есть хотя бы одна сильная нагрузка, он может быть интерпретирован. Возможен и обратный вариант — если сильные нагрузки имеются, однако интерпретация затруднительна, от этой компоненты предпочтительно отказаться.

Пример использования МГК

Пусть имеются следующие показатели экономической деятельности предприятия: трудоемкость (x_1), удельный вес покупных изделий в продукции (x_2), коэффициент сменности оборудования (x_3), удельный вес рабочих в составе предприятия (x_4), премии и вознаграждения на одного работника (x_5), рентабельность (y). Линейная регрессионная модель имеет вид:

$$y = b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 + b_4*x_4 + b_5*x_5$$

x_1	x_2	x_3	x_4	x_5	y
0,51	0,2	1,47	0,72	0,67	9,8
0,36	0,64	1,27	0,7	0,98	13,2
0,23	0,42	1,51	0,66	1,16	17,3
0,26	0,27	1,46	0,69	0,54	7,1
0,27	0,37	1,27	0,71	1,23	11,5
0,29	0,38	1,43	0,73	0,78	12,1
0,01	0,35	1,5	0,65	1,16	15,2
0,02	0,42	1,35	0,82	2,44	31,3
0,18	0,32	1,41	0,8	1,06	11,6
0,25	0,33	1,47	0,83	2,13	30,1

Пример использования МГК

Построение регрессионной модели в статистическом пакете показывает, что коэффициент X4 не значим (p-Value > α = 5%) , и его можно исключить из модели.

Multiple Regression Analysis

Dependent variable: Y

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-53,3379	13,2303	-4,03148	0,0157
X1	7,74608	3,43543	2,25476	0,0872
X2	14,351	4,73443	3,0312	0,0387
X3	29,0017	5,90555	4,91092	0,0080
X4	5,94883	9,81096	0,606345	0,5770
X5	13,784	1,17534	11,7277	0,0003

После исключения X4 снова запускается процесс построения модели.

Multiple Regression Analysis

Dependent variable: Y

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	-47,7167	8,82211	-5,40877	0,0029
X1	8,42709	3,03442	2,77716	0,0390
X2	13,0458	3,94113	3,31017	0,0212
X3	27,8688	5,23601	5,32254	0,0031
X5	14,319	0,725652	19,7326	0,0000

Пример использования МГК

Критерий Кайзера для МГК показывает, что можно оставить 2 компоненты, объясняющие около 80% исходной дисперсии.

```
Data input: observations
Number of complete cases: 10
Missing value treatment: listwise
Standardized: yes

Number of components extracted: 2
```

Principal Components Analysis

Component Number	Eigenvalue	Percent of Variance	Cumulative Percentage
1	1,77677	44,419	44,419
2	1,40431	35,108	79,527
3	0,493571	12,339	91,866
4	0,325347	8,134	100,000

Для выделенных компонент можно построить уравнения в исходной системе координат:

U1 = 0,41*x1 - 0,57*x2 + 0,49*x3 - 0,52*x5 U2 = 0,61*x1 + 0,38*x2 - 0,53*x3 - 0,44*x5

Table of Component Weights

	Component 1	Component 2
X1	0,406712	0,606432
X2	-0,567732	0,385238
X3	0,487689	-0,534899
X5	-0,523856	-0,444652

Пример использования МГК

Теперь можно построить в новых компонентах новую регрессионную модель:

$$y = 15{,}92 - 3{,}74*U1 - 3{,}87*U2$$

Multiple Regression Analysis

Dependent variable: Y

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	15,92	1,70013	9,364	0,0000
U1	-3,73522	1,34445	-2,77825	0,0274
U2	-3,87768	1,51227	-2,56415	0,0373

Метод сингулярного разложения (SVD)

Beltrami и Jordan считаются основателями теории сингулярного разложения. Beltrami – за то, что он первым опубликовал работу о сингулярном разложении, а Jordan – за элегантность и полноту своей работы. Работа Beltrami появилась в журнале “Journal of Mathematics for the Use of the Students of the Italian Universities” в 1873 году, основная цель которой заключалась в том, чтобы ознакомить студентов с билинейными формами.Суть метода в разложении матрицы **A** размера n x m с рангом **d = rank (M) <= min(n,m)** в произведение матриц меньшего ранга:

$$A = UDV^T,$$

где матрицы **U** размера n x d и **V** размера m x d состоят из ортонормальных столбцов, являющихся собственными векторами при ненулевых собственных значениях матриц AA^T и A^TA соответственно и $U^TU = V^TV = I$, а **D** размера d x d - диагональная матрица с положительными диагональными элементами, отсортированными в порядке убывания. Столбцы матрицы **U** представляют собой, ортонормальный базис пространства столбцов матрицы **A**, а столбцы матрицы **V** – ортонормальный базис пространства строк матрицы **A**.

Метод сингулярного разложения (SVD)

Важным свойством SVD-разложения является тот факт, что если для $k<d$ преобразовать матрицу D в матрицу D_k, состоящую только из k наибольших диагональных элементов, а также оставить в матрицах U и V только k первых столбцов, то матрица

$$A_k=U_kD_kV_k^T$$

будет являться лучшей аппроксимацией матрицы A относительно нормы Фробениуса среди всех матриц с рангом k.

Это усечение во-первых уменьшает размерность векторного пространства, снижает требования хранения и вычислительные требования к модели.

Во-вторых, отбрасывая малые сингулярные числа, малые искажения в результате шума в данных удаляются, оставляя только самые сильные эффекты и тенденции в этой модели.