НИУ ВШЭ НН. Анализ данных. Charge de cours: Калягин В.А.

Рабочий лист по теме "метод главных компонент". **Пример 1:** исследование клиентов банка

Исходные данные (n=15, m=5)

$$X=\begin{bmatrix} 9 & 6 & 9 & 2 & 2 \\ 4 & 6 & 2 & 6 & 7 \\ 0 & 0 & 5 & 0 & 0 \\ 2 & 2 & 0 & 9 & 9 \\ 6 & 9 & 8 & 3 & 3 \\ 3 & 8 & 5 & 4 & 7 \\ 4 & 5 & 6 & 3 & 6 \\ 8 & 6 & 8 & 2 & 2 \\ 4 & 4 & 0 & 8 & 8 \\ 2 & 8 & 4 & 5 & 7 \\ 1 & 2 & 6 & 0 & 0 \\ 6 & 9 & 7 & 3 & 5 \\ 6 & 7 & 1 & 7 & 8 \\ 2 & 1 & 7 & 1 & 1 \\ 9 & 7 & 9 & 2 & 1 \end{bmatrix}$$

Центрированные данные

$$Y=\begin{bmatrix} 4{,}60 & 0{,}67 & 3{,}87 & -1{,}67 & -2{,}40 \\ -0{,}40 & 0{,}67 & -3{,}13 & 2{,}33 & 2{,}60 \\ -4{,}40 & -5{,}33 & -0{,}13 & -3{,}67 & -4{,}40 \\ -2{,}40 & -3{,}33 & -5{,}13 & 5{,}33 & 4{,}60 \\ 1{,}60 & 3{,}67 & 2{,}87 & -0{,}67 & -1{,}40 \\ -1{,}40 & 2{,}67 & -0{,}13 & 0{,}33 & 2{,}60 \\ -0{,}40 & -0{,}33 & 0{,}87 & -0{,}67 & 1{,}60 \\ 3{,}60 & 0{,}67 & 2{,}87 & -1{,}67 & -2{,}40 \\ -0{,}40 & -1{,}33 & -5{,}13 & 4{,}33 & 3{,}60 \\ -2{,}40 & 2{,}67 & -1{,}13 & 1{,}33 & 2{,}60 \\ -3{,}40 & -3{,}33 & 0{,}87 & -3{,}67 & -4{,}40 \\ 1{,}60 & 3{,}67 & 1{,}87 & -0{,}67 & 0{,}60 \\ 1{,}60 & 1{,}67 & -4{,}13 & 3{,}33 & 3{,}60 \\ -2{,}40 & -4{,}33 & 1{,}87 & -2{,}67 & -3{,}40 \\ 4{,}60 & 1{,}67 & 3{,}87 & -1{,}67 & -3{,}40 \end{bmatrix}$$

нормы 10,66 10,92 11,65 10,46 12,07

Нормированные данные (норма стобца равна 1)

$$Z=\begin{bmatrix} 0{,}43 & 0{,}06 & 0{,}33 & -0{,}16 & -0{,}20 \\ -0{,}04 & 0{,}06 & -0{,}27 & 0{,}22 & 0{,}22 \\ -0{,}41 & -0{,}49 & -0{,}01 & -0{,}35 & -0{,}36 \\ -0{,}23 & -0{,}31 & -0{,}44 & 0{,}51 & 0{,}38 \\ 0{,}15 & 0{,}34 & 0{,}25 & -0{,}06 & -0{,}12 \\ -0{,}13 & 0{,}24 & -0{,}01 & 0{,}03 & 0{,}22 \\ -0{,}04 & -0{,}03 & 0{,}07 & -0{,}06 & 0{,}13 \\ 0{,}34 & 0{,}06 & 0{,}25 & -0{,}16 & -0{,}20 \\ -0{,}04 & -0{,}12 & -0{,}44 & 0{,}41 & 0{,}30 \\ -0{,}23 & 0{,}24 & -0{,}10 & 0{,}13 & 0{,}22 \\ -0{,}32 & -0{,}31 & 0{,}07 & -0{,}35 & -0{,}36 \\ 0{,}15 & 0{,}34 & 0{,}16 & -0{,}06 & 0{,}05 \\ 0{,}15 & 0{,}15 & -0{,}35 & 0{,}32 & 0{,}30 \\ -0{,}23 & -0{,}40 & 0{,}16 & -0{,}26 & -0{,}28 \\ 0{,}43 & 0{,}15 & 0{,}33 & -0{,}16 & -0{,}28 \end{bmatrix}$$

SVD разложение, p=min{n,m}=5.

$$U=\begin{bmatrix} -0{,}28 & 0{,}24 & -0{,}33 & -0{,}21 & -0{,}08 \\ 0{,}25 & 0{,}03 & 0{,}00 & 0{,}22 & -0{,}11 \\ -0{,}21 & -0{,}55 & 0{,}03 & 0{,}25 & -0{,}15 \\ 0{,}47 & -0{,}21 & -0{,}28 & -0{,}36 & 0{,}41 \\ -0{,}16 & 0{,}26 & 0{,}22 & 0{,}14 & 0{,}53 \\ 0{,}11 & 0{,}10 & 0{,}44 & -0{,}17 & -0{,}15 \\ 0{,}00 & -0{,}01 & 0{,}10 & -0{,}55 & -0{,}46 \\ -0{,}24 & 0{,}18 & -0{,}25 & 0{,}01 & -0{,}11 \\ 0{,}40 & -0{,}05 & -0{,}29 & 0{,}16 & 0{,}03 \\ 0{,}18 & 0{,}05 & 0{,}48 & 0{,}02 & 0{,}23 \\ -0{,}24 & -0{,}40 & 0{,}12 & 0{,}29 & -0{,}02 \\ -0{,}07 & 0{,}27 & 0{,}25 & -0{,}02 & -0{,}13 \\ 0{,}32 & 0{,}18 & -0{,}16 & 0{,}39 & -0{,}36 \\ -0{,}22 & -0{,}36 & -0{,}07 & -0{,}29 & 0{,}16 \\ -0{,}31 & 0{,}28 & -0{,}27 & 0{,}12 & 0{,}21 \end{bmatrix}$$

$$S=\begin{bmatrix} 1{,}66 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 1{,}33 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}61 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}25 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}17 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 & 0{,}00 \end{bmatrix}$$

$$V=\begin{bmatrix} -0{,}18 & 0{,}64 & -0{,}70 & 0{,}04 & -0{,}27 \\ 0{,}03 & 0{,}69 & 0{,}61 & 0{,}37 & 0{,}09 \\ -0{,}56 & 0{,}21 & 0{,}15 & -0{,}69 & 0{,}36 \\ 0{,}57 & 0{,}17 & -0{,}28 & -0{,}08 & 0{,}75 \\ 0{,}57 & 0{,}20 & 0{,}19 & -0{,}61 & -0{,}48 \end{bmatrix}$$

Сингулярные числа

1,66 1,33 0,61 0,25 0,17

Квадраты сингулярных чисел сумма

2,75 1,77 0,38 0,06 0,03 5,00

Вклад главных факторов в вариацию данных в %

55,09 35,50 7,54 1,30 0,57 100,00

Два главных фактора вносят 90,59% в вариац данных

Ошибка аппроксимации в норме Фробениуса 8,41%

НИУ ВШЭ НН. Анализ данных. Charge de cours: Калягин В.А.

Рабочий лист по теме "метод главных компонент". **Пример 1:** исследование клиентов банка

Главные факторы

	-0,28		0,24
	0,25		0,03
	-0,21		-0,55
	0,47		-0,21
	-0,16		0,26
u_1=	0,11	u_2=	0,10
	0,00		-0,01
	-0,24		0,18
	0,40		-0,05
	0,18		0,05
	-0,24		-0,40
	-0,07		0,27
	0,32		0,18
	-0,22		-0,36
	-0,31		0,28

Аппроксимация столбцов матрицы Z

z^1=	s_1*v_{1,1}*u_1+s_2*v_{1,2}*u_2	=	-0,30	*u_1+	0,85	*u_2					
z^2=	s_1*v_{2,1}*u_1+s_2*v_{2,2}*u_2	=	0,05	*u_1+	0,92	*u_2					
z^3=	s_1*v_{3,1}*u_1+s_2*v_{3,2}*u_2	=	-0,94	*u_1+	0,28	*u_2					
z^4=	s_1*v_{4,1}*u_1+s_2*v_{4,2}*u_2	=	0,95	*u_1+	0,23	*u_2					
z^5=	s_1*v_{5,1}*u_1+s_2*v_{5,2}*u_2	=	0,94	*u_1+	0,27	*u_2					

Квадрат евкл нормы ошибки

	0,18
	0,15
	0,04
	0,05
	0,04
сумма	**0,47**
провер	0,47

Интерпретация главных факторов

u_1 отношение к индивидуальному обслуживанию в банке

u_2 отношение к размеру банка

Вектора ошибок аппроксимации

z^j - s_1*v_{j,1}*u_1-s_2*v_{j,2}*u_2

Z - Z_2 =

0,14	-0,15	0,00	0,05	0,00
0,01	0,02	-0,05	-0,02	-0,03
0,00	0,03	-0,05	-0,03	-0,02
0,10	-0,13	0,06	0,11	-0,01
-0,12	0,10	0,03	0,03	-0,04
-0,18	0,15	0,06	-0,09	0,09
-0,03	-0,02	0,08	-0,06	0,13
0,11	-0,09	-0,03	0,03	-0,02
0,12	-0,09	-0,05	0,05	-0,06
-0,22	0,19	0,06	-0,05	0,03
-0,05	0,07	-0,04	-0,03	-0,03
-0,10	0,09	0,02	-0,06	0,04
0,09	-0,03	-0,11	-0,03	-0,05
0,02	-0,05	0,05	0,04	0,02
0,11	-0,09	-0,03	0,07	-0,07

Перевод глав факторов в шкалу 0-9 (карта клиентов)

инд обсл	размер банка
0	9
6	6
1	0
9	4
2	9
5	7
4	6
1	8
8	5
6	7
1	2
3	9
7	8
1	2
0	9

НИУ ВШЭ НН. Анализ данных. Charge de cours: Калягин В.А.

Рабочий лист по теме "метод главных компонент". **Пример 1:** исследование клиентов банка

Аппроксимация (два главных фактора)

Z ≈ Z_2 = U_2*	S_2*	V_2^T =	-0,28	0,24	*	1,66	0,00	*	-0,18	0,03	-0,56	0,57	0,57				
			0,25	0,03		0,00	1,33		0,64	0,69	0,21	0,17	0,20				
			-0,21	-0,55													
			0,47	-0,21		-0,30	0,05	-0,94	0,95	0,94							
			-0,16	0,26		0,85	0,92	0,28	0,23	0,27							
			0,11	0,10													
			0,00	-0,01													
			-0,24	0,18													
			0,40	-0,05													
			0,18	0,05													
			-0,24	-0,40													
			-0,07	0,27													
			0,32	0,18													
			-0,22	-0,36													
			-0,31	0,28													

новые координаты объектов (карта объектов)

$$V\_2 = \begin{bmatrix} -0,18 & 0,64 \\ 0,03 & 0,69 \\ -0,56 & 0,21 \\ 0,57 & 0,17 \\ 0,57 & 0,20 \end{bmatrix}$$

координаты признаков (карта признаков)

$$z^1 \approx -0,30 * u\_1 + 0,85 * u\_2$$
$$z^2 \approx 0,05 * u\_1 + 0,92 * u\_2$$
$$z^3 \approx -0,94 * u\_1 + 0,28 * u\_2$$
$$z^4 \approx 0,95 * u\_1 + 0,23 * u\_2$$
$$z^5 \approx 0,94 * u\_1 + 0,27 * u\_2$$

$$Z \approx Z\_2 = [u\_1\ u\_2] * \begin{bmatrix} -0,30 & 0,05 & -0,94 & 0,95 & 0,94 \\ 0,85 & 0,92 & 0,28 & 0,23 & 0,27 \end{bmatrix}$$

$$G = \begin{bmatrix} 0,71 & -0,71 \\ 0,71 & 0,71 \end{bmatrix}$$

Вращения в пространстве главных факторов, переход к другому базису $Z\_2 = ([u\_1\ u\_2] * G) * (G^T * A)$

$$Z\_2 = [u'\_1\ u'\_2] * \begin{bmatrix} 0,39 & 0,69 & -0,47 & 0,83 & 0,86 \\ 0,81 & 0,62 & 0,86 & -0,51 & -0,47 \end{bmatrix}$$

Структура матрицы нагрузок ухудшилась

$$G^T = \begin{bmatrix} 0,71 & 0,71 \\ -0,71 & 0,71 \end{bmatrix}$$

НИУ ВШЭ НН. Анализ данных. Charge de cours: Калягин В.А.

Рабочий лист по теме "метод главных компонент". **Пример 1:** исследование клиентов банка

Матричное разложение

$(S\_2)^{(1/2)}=$

1,29	0,00
0,00	1,15

$Z \approx Z\_2 = (U\_2*(S\_2)^{(1/2)}) * ((S\_2)^{(1/2)}*V\_2^T) =$

-0,36	0,28
0,32	0,04
-0,26	-0,64
0,61	-0,24
-0,20	0,30
0,14	0,11
0,00	-0,01
-0,31	0,21
0,51	-0,06
0,23	0,06
-0,31	-0,46
-0,09	0,31
0,41	0,21
-0,29	-0,42
-0,40	0,32

*

-0,23	0,04	-0,72	0,73	0,73
0,74	0,80	0,24	0,20	0,23