Statistické metody a zpracování dat

IX Úvod do vícerozměrných metod

Petr Dobrovolný

Úvod do vícerozměrných metod

O řadě jevů či procesů máme k dispozici ne jeden statistický znak, ale znaků několik.

Př. Struktura obyvatelstva, vlastnosti povodí, klimatické poměry místa, …

Vstupní data: Statistické jednotky (např. městské obvody) a k nim několik charakteristik (např. demografická data).

	A	B	C	D	E	F
1	obec	Průměrný počet osob na byt	průměrný počet osob na obyt. míst. $8m^2$	Prům.počet m^2 p.ba byt	Prům.počet m^2 obyt.p.ba osobu	Prům.počet obyt.míst. na byt
2	Žďár nad Sázavou	2.78	1.10	43.0	15.4	2.52
3	Balíny	3.35	1.16	60.2	17.5	2.97
4	Blažkov	3.34	1.07	58.7	17.6	3.11
5	Blízkov	3.61	1.20	60.7	16.8	3.00
6	Bobrová	3.06	1.02	60.1	19.3	3.03
7	Bobrůvka	3.13	1.11	55.2	17.6	2.82
8	Bohdalec	3.34	1.00	60.6	18.1	3.33
9	Bohutínu	3.25	1.06	61.2	18.7	3.15

K čemu jsou metody dobré?

- **redukovat počet proměnných**
- **detekovat strukturu vztahů mezi proměnnými (klasifikovat, vytvořit typologii dat)**

Analýza hlavních komponent (Principal Component Analysis – **PCA**)

Shluková analýza

Literatura:

Heřmanová, E. (1991): Vybrané vícerozměrné statistické metody v geografii. SPN, Praha, 133 s.

Hendl, J. (2004): Přehled statistických metod zpracování dat. Portál, Praha, 583 s.

http://www.statsoft.cz/textbook/stathome.html

PCA - Ilustrativní příklad

Vstupní data: Podíl zaměstnaných v devíti odvětvích ve 26 evropských zemích (údaje z konce 70. let 20. století)

1. AGR = agriculture
2. MIN = mining
3. MAN = manufacturing
4. PS = power suplies
5. CON = construction
6. SER = service industries
7. FIN = finance,
8. SPS = social and personal services
9. TC = transport and communications

Vstupní matice: 9 řádků (proměnných – odvětví) a 26 sloupců (případy – státy)

Cíl: Redukce počtu proměnných a odhalení typických znaků v zaměstnanosti jednotlivých států

Příklad – typický výstup PCA I.

No.	Eigenvalue	Individual Percent	Cumulative Percent	Scree Plot
1	3.487151	38.75	38.75	||||||||
2	2.130173	23.67	62.41	|||||
3	1.098958	12.21	74.63	|||
4	0.994483	11.05	85.68	|||
5	0.543218	6.04	91.71	||
6	0.383428	4.26	95.97	|
7	0.225754	2.51	98.48	|
8	0.136790	1.52	100.00	|
9	0.000046	0.00	100.00	|

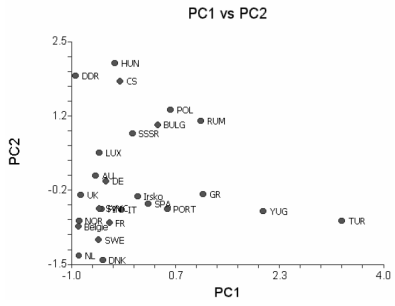
• pořadové číslo nové proměnné (PC - **hlavní komponenty**)
• tzv. **vlastní hodnota** – část z celkového rozptylu původních dat vysvětlená každou z nových komponent
• procentuální vyjádření množství **rozptylu vysvětleného** komponentou
• **kumulativní** hodnota procentuálního podílu vysvětleného příslušnými komponentami (např. první 4 komponenty vysvětlují 85,68 % celkové variability původních dat)
• tzv. **sutinový graf** slouží k určení počtu významných komponent

Příklad – typický výstup PCA II.

Variables	Factor1	Factor2	Factor3	Factor4
AGR	0.523791	0.053594	0.048674	0.028793
MIN	0.001323	0.617807	-0.201100	0.064085
MAN	-0.347495	0.355054	-0.150463	-0.346088
PS	-0.255716	0.261096	-0.561083	0.393309
CON	-0.325179	0.051288	0.153321	-0.668324
SER	-0.378920	-0.350172	-0.115096	-0.050157
FIN	-0.074374	-0.453698	-0.587361	-0.051567
SPS	-0.387409	-0.221521	0.311904	0.412230
TC	-0.366823	0.202592	0.375106	0.314372

Tzv. **zátěže** (loadings) - představují míru korelace mezi původními a novými proměnnými

Příklad – typický výstup PCA



Struktura zaměstnanosti jednotlivých zemí vyjádřená polohou v grafu hodnot prvních dvou (nejvýznamnějších) hlavních komponent.

Princip PCA

Charakteristiky, které na jednotkách měříme, jsou jen určitou formou projevu tzv. **skrytých veličin**, které přímo měřit nemůžeme.

Řada měřených charakteristik spolu do značné míry **souvisí** – vypovídá o stejné vlastnosti, **koreluje** spolu (mezi proměnnými existují „překryvy“).

Cílem metody je **eliminování duplicit**, **zhuštění informace** obsažené v původních proměnných do menšího počtu vzájemně nekorelovaných proměnných.

Tyto nové proměnné (**hlavní komponenty**) popisují soubor jednotek syntetičtěji a úsporněji.

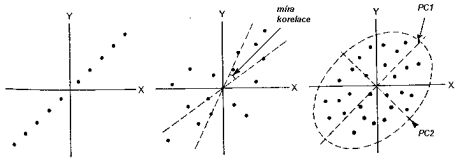
Základní východiska

Princip redukce dat a „skryté“proměnné (interpretace následujícího obrázku)

Máme-li pro soubor znaků dvě proměnné a ty spolu vzájemně korelují – potom vypovídají z velké části o tomtéž – jsou **redundantní**.

Pokud takového dvě (korelované) proměnné vyneseme do grafu a **nějak** proložíme rovnicí přímky – potom tuto přímku můžeme považovat za osu, na niž jsou vyneseny hodnoty **nové proměnné**, která ponese podstatnou informaci z obou proměnných původních.

Základní východiska



Základní východiska

Tedy – dvě původní proměnné redukujeme do jedné nové proměnné – do tzv. hlavní komponenty (PC).

Hlavní komponenta je lineární kombinací původních proměnných.

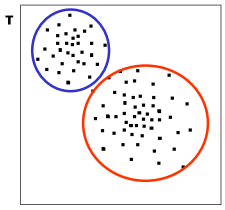
Uvedený princip lze zobecnit na větší počet proměnných

Metody PCA se používají k analýze vztahů závislosti ve vícerozměrném (obecně r-rozměrném) ortogonálním (pravoúhlém) prostoru.

Shluková analýza

Ilustrativní případ pro m=2

Klimatické poměry n stanic jsou charakterizovány dvěma proměnnými (m=2): Průměrnou roční teplotou vzduchu (T) a ročním úhrnem srážek (S)



INTERPRETACE: Stanice s vysokými srážkami a nízkými teplotami tvoří shluk stanic vysokohorských, stanice s nízkými úhrny srážek a vysokými teplotami tvoří shluk stanic níže položených. Ve většině případů není vymezení shluků takto triviální.

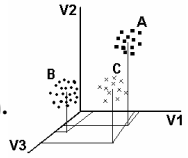
Shluková analýza (Cluster analysis)

Je to skupina metod, jejichž cílem je rozdělení souboru jednotek na několik navzájem vylučujících se relativně stejnorodých podmnožin (**shluků** = clusters).

Rozdělení jednotek je provedeno tak, aby **jednotky patřící do téhož shluku si byly co nejvíce „podobné“,** zatímco jednotky pocházející z různých shluků by měly být co nejvíce odlišné.

Charakteristika metody I.

Shluková analýza je vícerozměrnou metodou. K charakterizování jednotek, kterých je obecně *n* využívá většího počtu znaků (*m*>=*2*).

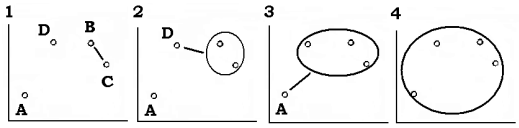


Jednotky představují body v n-rozměrném prostoru, jehož osy tvoří hodnoty jednotlivých znaků (v1, v2, v3).

V takto definovaném prostoru tvoří jednotky s podobnými hodnotami znaků PŘIROZENÉ shluky.

Jednotlivé metody shlukové analýzy řeší problém definice a výpočtu „podobnosti“ či „odlišnosti“ jednotek a jejich PŘÍSLUŠNOST k určitým shlukům.

Princip shlukování a míry vzdálenosti



Kritériem víceznakové podobnosti ve shlukové analýze je VZDÁLENOST.

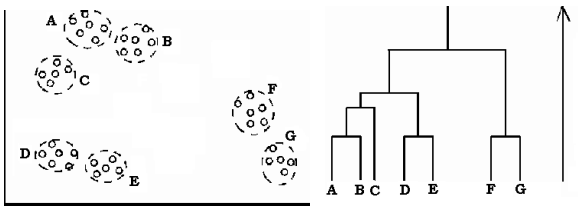
Čím blíže se nacházejí body v m-rozměrném prostoru, tím jsou si podobnější. Nulová vzdálenost znamená identitu – tedy maximální podobnost.

Charakteristiky dendrogramu

Postup shlukování lze prezentovat pomocí tzv. **dendrogramu**.

Rozdělíme-li dendrogram na jakékoliv úrovni pomyslným řezem, vždy dostaneme homogenní shluky.

Čím později ho rozdělíme, tím méně podobné jednotky jsou spojeny v jednom shluku. Šipka značí pokles míry podobnosti jednotek ve shlucích



Analýza vzdáleností

Spočívá ve výpočtu zvoleného typu vzdálenosti mezi všemi jednotkami a v jejich sestavení do symetrické čtvercové matice, která má na diagonále nuly (tj. maximální podobnost).

Matice vzdáleností

Obvod Prahy	1	2	3	4	5	6	7	8	9	10
1	0	4,01	10,58	9,68	13,06	9,05	7,51	11,39	18,85	10,38
2	4,01	0	6,67	6,31	9,68	9,90	3,66	7,50	15,25	6,69
3	10,58	6,67	0	5,34	6,43	13,88	3,44	3,00	10,27	3,58
4	9,68	6,31	5,34	0	7,90	10,25	4,88	3,60	11,99	3,35
5	13,06	9,68	6,43	7,90	0	16,60	6,47	5,33	5,93	4,88
6	9,05	9,90	13,88	10,25	16,60	0	11,96	13,44	21,32	12,46
7	7,51	3,66	3,44	4,88	6,47	11,96	0	4,36	11,76	3,63
8	11,39	7,50	3,00	3,60	5,33	13,44	4,36	0	8,94	1,75
9	18,85	15,25	10,27	11,99	5,93	21,32	11,76	8,94	0	9,30
10	10,38	6,69	3,58	3,35	4,88	12,46	3,63	1,75	9,30	0
Σa_{ij}	94,51	69,68	63,18	63,31	76,29	118,85	57,67	59,32	113,61	56,03

1. krok - nalezení minimálního prvku v původní matici

Obvod Prahy	1	2	3	4	5	6	7	8	9	10
1	0	4,01	10,58	9,68	13,06	9,05	7,51	11,39	18,85	10,38
2	4,01	0	6,67	6,31	9,68	9,90	3,66	7,50	15,25	6,69
3	10,58	6,67	0	5,34	6,43	13,88	3,44	3,00	10,27	3,58
4	9,68	6,31	5,34	0	7,90	10,25	4,88	3,60	11,99	3,35
5	13,06	9,68	6,43	7,90	0	16,60	6,47	5,33	5,93	4,88
6	9,05	9,90	13,88	10,25	16,60	0	11,96	13,44	21,32	12,46
7	7,51	3,66	3,44	4,88	6,47	11,96	0	4,36	11,76	3,63
8	11,39	7,50	3,00	3,60	5,33	13,44	4,36	0	8,94	1,75
9	18,85	15,25	10,27	11,99	5,93	21,32	11,76	8,94	0	9,30
10	10,38	6,69	3,58	3,35	4,88	12,46	3,63	1,75	9,30	0

Ve výchozí matici je minimální vzdálenost mezi prvky 8 a 10:

$d_{8,10}$ = 1,75.

Tyto dvě jednotky se sloučí.

Vypočítáme vzdálenosti tohoto nového shluku k stávajícím jednotkám (příklad pro pro jednotku 1):

$$d_{(8+10,1)} = \frac{d_{(8,1)} + d_{(10,1)}}{2} = \frac{11{,}39 + 10{,}38}{2} = 10{,}88$$

Analogicky se vypočtou nové vzdálenosti mezi novým shlukem a zbylými jednotkami, tedy $d_{(8+10,2)}$, $d_{(8+10,3)}$ ..... $d_{(8+10,9)}$

Výsledkem je nová matice vzdáleností.

2. krok - nová matice vzdáleností

	8+10	1	2	3	4	5	6	7	9
8+10	(1,75)	10,88	7,10	3,29	3,48	5,11	12,95	4,00	9,12
1		0	4,01	10,58	9,68	13,06	9,05	7,51	18,85
2			0	6,67	6,31	9,68	9,90	3,66	15,25
3				0	5,34	6,43	13,88	3,44	10,27
4					0	7,90	10,25	4,88	11,99
5						0	16,60	6,47	5,93
6							0	11,96	21,32
7								0	11,76
9									0

Hodnota v závorce vyjadřuje vzdálenost, při které dochází ke sloučení a využívá se ke konstrukci tzv. dendrogramu (viz. dále).

Opět se najde minimální hodnota a celý výpočet se opakuje tak, jak je naznačeno v dále uvedených maticích vzdáleností ...

3. krok

	8+10+3	1	2	4	5	6	7	9
8+10+3	(2,78)	10,78	6,96	4,10	5,55	13,26	3,81	9,50
1		0	4,01	9,68	13,06	9,05	7,51	18,85
2			0	6,31	9,68	9,90	3,66	15,25
4				0	7,90	10,25	4,88	11,99
5					0	16,60	6,47	5,93
6						0	11,96	21,32
7							0	11,76
9								0

4. krok

	8+10+3	2+7	1	4	5	6	9
8+10+3	(2,78)	5,38	10,78	4,10	5,55	13,26	9,50
2+7		(3,66)	5,76	5,60	8,08	10,93	13,50
1			0	9,68	13,06	9,05	18,85
4				0	7,90	10,25	11,99
5					0	16,60	5,93
6						0	21,32
9							0

5. krok

	8+10+3+4	2+7	1	5	6	9
8+10+3+4	(3,44)	5,44	10,51	6,14	12,51	10,12
2+7		(3,66)	5,76	8,08	10,93	13,50
1			0	13,06	9,05	18,85
5				0	16,60	5,93
6					0	21,32
9						0

6. krok

	8+10+3+4+2+7	1	5	6	9
8+10+3+4+2+7	(4,52)	8,92	6,78	11,98	11,25
1		0	13,06	9,05	18,85
5			0	16,60	5,93
6				0	21,32
9					0

7. krok

	8+10+3+4+2+7	5+9	1	6
8+10+3+4+2+7	(4,52)	9,02	8,92	11,98
5+9		(5,93)	15,96	18,96
1			0	9,05
6				0

8. krok

	8+10+3+4+2+7+1	5+9	6
8+10+3+4+2+7+1	(5,78)	10,01	11,56
5+9		(5,93)	18,96
6			0

9. krok

	8+10+3+4+2+7+1+5+9	6
8+10+3+4+2+7+1+5+9	(7,43)	13,20
6		0

Ukončení shlukování a prezentace jeho průběhu

V posledním kroku dochází ke sloučení všech jednotek do jednoho shluku na vzdálenosti 8,58.

To je průměrná vzdálenost mezi dvěma jednotkami.

Průběh shlukování se obvykle zaznamenává do dendrogramů – hierarchicky uspořádaných „stromů“).

Dendrogram