Adams uzavírá, že uvažované proměnné vysvětlily pouze 10 % variace doby zaměstnání ve sledovaném území. Proto je zapotřebí pro přesnější úvahy uvažovat a měřit v budoucnu další proměnné jako rodinný stav učitele, roční plat, aprobaci učitele atd.

V poslední době jsme svědky rostoucího zájmu o metody pro analýzu historie událostí. Ukázali jsme pouze ty nejjednodušší postupy, jež se však používají především v medicínských aplikacích nejčastěji. Pracují pouze s jedním typem události. Ve výzkumu se uplatňuje celá škála nových modelů pro analýzu dat s několika různými událostmi. Nalézají uplatnění v epidemiologii, při zpracování dat o klinických pokusech, v demografii a v sociálních vědách. Další třídu modelů používáme pro analýzu procesů s tzv. **neabsorpčními stavy,** které nejsou tradičními modely brány v úvahu. Neabsorpční stav má tu vlastnost, že je možné se z něj přemístit do jiného stavu s určitou pravděpodobností. Jednou z aplikací takového modelu je popis časů mezi jednotlivými infarkty u monitorovaného pacienta. Tyto náročnější postupy popisuje práce Claytona (1988).

13.6 Shluková analýza

Metody shlukové analýzy se vyvinuly z potřeby analyzovat informaci obsaženou v datech, která jsou generována množinou objektů, o jejíž struktuře toho víme jen velmi málo. Strukturou se přitom myslí rozdělení objektů do určitého systému kategorií, jež zachycuje podobnost objektů patřících do téže kategorie na jedné straně a nepodobnost objektů patřících do různých kategorií na straně druhé. Jestliže nám není známa kategorizace objektů, je naším cílem najít takovou kategorizační strukturu, jež je ve shodě s poměry v datech. Hledáme „přirozené skupiny". Metody shlukové analýzy (cluster analysis) nám k tomu poskytují algoritmy, které toto hledání provádějí automaticky za pomoci počítačů. Algoritmy můžeme použít k měření starých klasických kategorizačních struktur nebo k nalézání nových. Přibližme si trochu problém, jímž se shluková analýza zabývá. Ukážeme, že bez počítače by jeho řešení bylo těžko dosažitelné.

Výzkumník má obvykle představu o tom, jaké vlastnosti by mělo mít navržené řešení. Dovede rozlišovat mezi špatnou a dobrou strukturou kategorií. Proto se může pokusit najít dělení příslušného materiálu ručně. Představme si, že by měl k dispozici 25 vektorů měření, které reprezentují 25 objektů. Jestliže by si položil za úkol najít rozdělení těchto objektů reprezentovaných vektory měření do 5 smysluplných kategorií, musel by posoudit 2 436 684 974 110 751 různých možností. Jestliže ani neví, do kolika kategorií by měl materiál rozložit, musí

provést více než 4×10^{18} posouzení – skutečně velký počet, jenž by se těžko zvládl i na počítači. Algoritmy shlukové analýzy jsou navrženy tak, aby se tento počet podstatně zmenšil. Přesto však se i u nich k rozumnému dělení materiálu musí udělat velké množství operací, které zaručují optimálnost nalezené struktury.

Uveďme popis základní situace shlukové analýzy: Je dáno N objektů. Na každém objektu je naměřeno k charakteristik, takže získáváme Nk-rozměrných vektorů měření $x_1, x_2, \ldots, x_N$. Můžeme ztotožnit pozorování a příslušné objekty, takže v dalším nazýváme vektory x_i objekty. Označme X množinu všech těchto objektů. Úkolem shlukové analýzy je seskupit objekty x_i do n shluků $S_1, S_2, \ldots, S_n$ (tuto množinu shluků značíme S) tvořících rozklad množiny X tak, aby si objekty patřící do téhož shluku byly v jistém smyslu podobné či blízké, kdežto od objektů patřících do různých skupin požadujeme, aby byly odlišné či vzdálené. Přitom obvykle chceme, aby počet shluků n byl podstatně menší než počet objektů N. Někdy je úloha analýzy formulována obecněji v tom smyslu, že shluky nemusí být disjunktní (mohou se překrývat). Podle cíle můžeme rozeznávat tři druhy úloh shlukové analýzy:

1. Cílem je nalezení předem definovaného množství shluků.
2. Cílem je nalezení množiny shluků, přičemž jejich počet není specifikován.
3. Cílem je vytvořit **hierarchický strom**.

Formálně můžeme hierarchický strom vymezit jako posloupnost množin shluků $S^t, t = 1, 2, \ldots, r$. Za první množinu shluků S^1 považujeme vlastní objekty ($n = N$). Finální množina S^r je tvořena jedním shlukem zahrnujícím všechny objekty. Každá skupina S^t je zjemněním skupiny S^{t+1}. Zjemněním se rozumí, že shluky v množině S^t vznikly pouze rozdělením některých shluků množiny S^{t+1}. Shlukování podle tohoto předpisu odpovídá Linnéově kategorizaci v biologii.

Shlukování se provádí pro omezený počet objektů. Jestliže se nalezne optimální systém shluků pro výběrovou množinu objektů, vzniká otázka, jak tento výsledek přenést na celou populaci, ze které byly objekty vybrány. Teprve v této fázi analýzy se začnou uplatňovat klasické prostředky matematické statistiky, jako jsou třeba metody testování hypotéz.

Poznamenejme, že kromě objektů, reprezentovaných vektory měření, můžeme podrobit shlukové analýze i vlastní proměnné. Shlukování proměnných má společné znaky s faktorovou analýzou. Obě metodologie se snaží prozkoumat vztahovou strukturu mezi proměnnými, ačkoli využívají úplně rozdílné techniky (viz analýza hlavních komponent).

Shlukovací metody jsou většinou založeny na využití **měr nepodobnosti** (resp. podobnosti) objektů a shluků, jež odrážejí naše intuitivní požadavky. Odpovídající míry pro shluky jsou obvykle odvozeny od měr pro objekty. Jednou

z nejpoužívanější měr nepodobnosti je **euklidovská vzdálenost** v mezi dvěma vektory Y a Z:

$$v_{YZ} = \sqrt{\sum_{i=1}^{k} (y_i - z_i)^2}$$

Shluky se v jednotlivých krocích považují za nové objekty a podrobují se shlukování podle stejných principů jako původní objekty. Primárním podkladem pro shlukovací procedury je matice vzdálenosti (v_{rs}) jednotlivých párů objektů (značených r a s). Uvádíme některé běžně používané míry podobnosti shluků značené v, přičemž S^h a S^k značí h-tý a k-tý shluk v dané fázi shlukování a n_h a n_k je počet objektů v příslušných shlucích.

Vzdálenost nejbližšího souseda:

$$v(S^h, S^k) = \min(v_{ij}) \qquad i \in S^h,\ j \in S^k$$

Vzdálenost nejvzdálenějšího souseda:

$$v(S^h, S^k) = \max_{ij}(v_{ij}) \qquad i \in S^h,\ j \in S^k$$

Průměrná vzdálenost mezi sousedy:

$$v(S^h, S^k) = \frac{1}{n_h n_k} \sum_i \sum_j v_{ij} \qquad i \in S^h,\ j \in S^k$$

PŘÍKLAD 13.7

Postup shlukové analýzy

Na schematickém příkladu si popišme metodu sekvenčního, aglomerativního, hierarchického a disjunktního shlukování. (Metoda je sekvenční proto, že objekty a shluky se shlukují postupně; aglomerativní – začíná se od objektů jako izolovaných shluků, které se shlukují ve stále větší množiny; hierarchická a disjunktní – definováno výše.) Metoda se realizuje užitím euklidovské vzdálenosti mezi shluky. Mějme objekty A, B, C, D a E popsané čtyřmi proměnnými X_1, X_2, X_3, X_4. Tvar matice měření uvádí tabulka 13.13.

Tvar matice vzdálenosti V_0 ukazuje tabulka 13.14 (uvádíme pouze jednu z jejích symetrických částí). V matici vzdáleností hledáme nejmenší v_{rs}. V naší matici přísluší objektům E a D. Objekty s nejmenší vzdáleností spojíme.

Přepočítáme novou matici vzdáleností V^1, která bude o jeden řádek a jeden sloupec menší. Shluk (E a D) považujeme za objekt D', jehož vzdálenosti od ostatních objektů přepočítáme metodou průměrné vzdálenosti od ostatních objektů. Např. vzdálenost vytvořeného shluku od objektu C bude $V_{CD'} = (v_{CE} + v_{CD})/2 = (78{,}1 + 80{,}6)/2 = 79{,}3$. Vzniká matice vzdáleností V^1 (tab. 13.15). Tato matice indikuje spojení objektů B a C. Nová matice V^2 má tvar, který ukazuje tabulka 13.16. Tato matice indikuje spojení A a shluku B'. Dostáváme matici V^3 (tab. 13.17).

Tab. 13.13 Příklad shlukové analýzy – matice měření

	A	*B*	*C*	*D*	*E*
X_1	100	80	80	40	50
X_2	80	60	70	20	10
X_3	70	50	40	20	20
X_4	60	40	50	10	10

Tab. 13.14 Příklad shlukové analýzy – původní matice vzdáleností

	A	*B*	*C*	*D*	*E*
A	0				
B	40,0	0			
C	38,7	17,3	0		
D	110,4	70,7	78,1	0	
E	111,4	72,1	80,6	14,1	0

Tab. 13.15 Příklad shlukové analýzy – upravená matice vzdáleností, v níž byly dva nejbližší objekty (*D* a *E*) nahrazeny jedním objektem *D'*

	A	*B*	*C*	*D'*
A	0			
B	40,0	0		
C	38,7	17,3	0	
D'	110,9	71,4	79,3	0

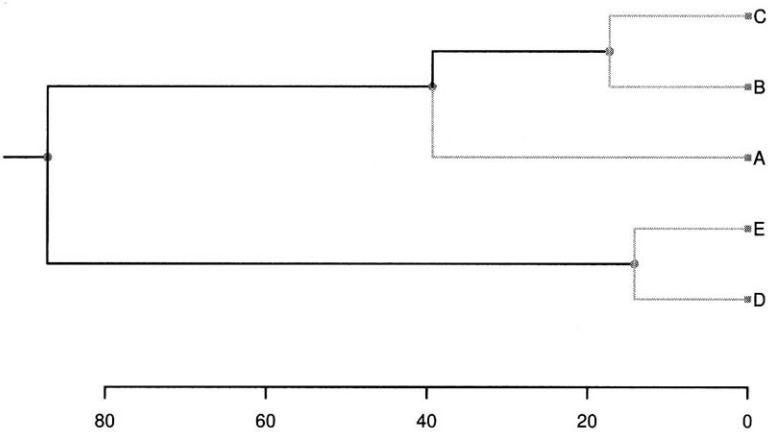
Tab. 13.16 Příklad shlukové analýzy – další krok úpravy matice vzdáleností

	A	*B'*	*C'*
A	0		
B'	39,3	0	
D'	110,9	75,3	0

Tab. 13.17 Příklad shlukové analýzy – poslední krok úpravy matice vzdáleností

	A'	*D'*
A'	0	
D'	43,1	0

Obr. 13.14 Dendrogram pro data modelového příkladu



Poslední krok spočívá ve spojení shluků A' a D'. Celý průběh shlukování lze zobrazit **dendrogramem** (obr. 13.14), který se sestrojuje až po uskutečnění shlukování. V našem případě objekty A, B, C, D, E jsou již uvedeny v pořadí, v němž se objeví na dendrogramu. Toto pořadí je však obecně dáno, až když je znám celý průběh shlukování. Každá úroveň shlukování se hodnotí minimální vzdáleností, při které došlo ke spojení.

Uveďme několik poznámek k metodologii užití výsledků shlukové analýzy.

a) Danou množinu objektů lze často roztřídit mnoha rozdílnými způsoby, přičemž každá získaná struktura má svůj význam. Každá klasifikace může odrážet jiné aspekty dat.
b) Shluková metoda je metodou pro navrhování hypotéz. Klasifikace objektů nebo proměnných získaná shlukovou analýzou nemá žádnou vnitřní validitu. Oprávněnost každé klasifikace a z ní vycházející explanační struktury je zdůvodnitelná shodou se známými fakty bez vztahu ke způsobu jejího získání.

c) Množina získaných shluků není konečným výsledkem, ale pouze možným návrhem struktury.

d) Výsledky shlukovací metody jsou směsí struktury, kterou do dat vkládá sám algoritmus procedury, a té, jež je v datech skutečně přítomná. Používané procedury obsahují operace, které na jedné straně systematicky opomíjejí určité (pro každou proceduru individuální) rysy dat a na druhé straně protežují ostatní.

e) Při analýze často opomeneme dvě možnosti: 1. data neobsahují shluky (absence jakýchkoli diskriminačních proměnných, rovnoměrné rozložení bodů); 2. data tvořící pouze jeden shluk (absence diskriminačních proměnných a vzájemná příbuznost údajů).

f) Základní znalosti o populaci, zvláště nejsou-li známy podmínky sběru dat, mohu způsobit značné divergence výsledků. Například předpokládejme, že v populaci je pět skupin. Jestliže jedna z nich je v datech jen náhodně zastoupena, budou v nich pravděpodobně jen čtyři skupiny. Jestliže pak předepíšeme algoritmu vytvořit v datech pět skupin, získáme shluky, které nebudou odpovídat našim předpokladům. Stejný výsledek dostaneme, když analyzujeme proměnné, jež mají silné diskriminační vlastnosti, avšak vzhledem k jinému problému, než který nás zajímá.

PŘÍKLAD 13.8

Shluková analýza socioekonomických dat

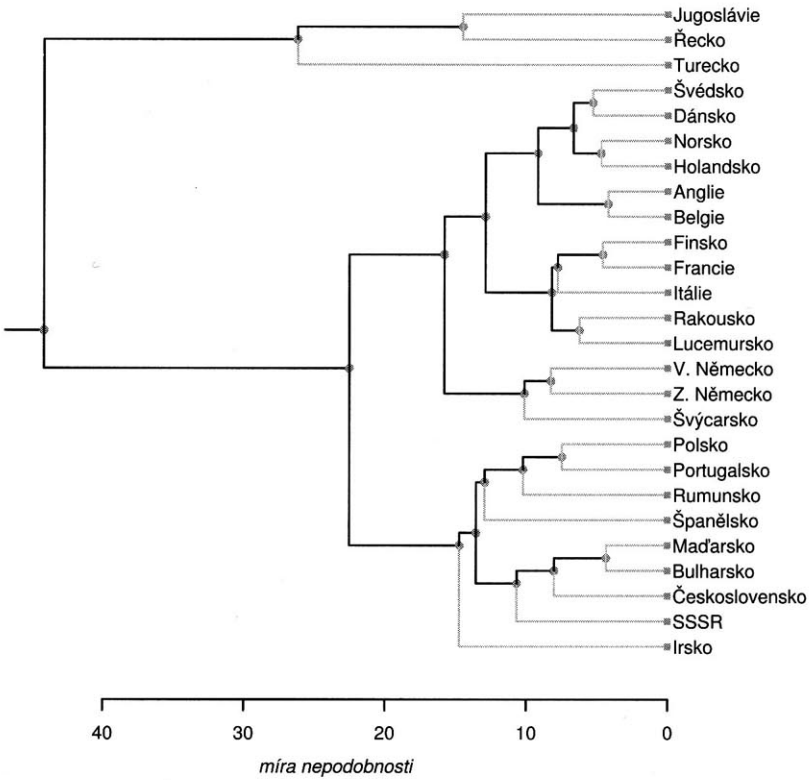
Pomocí shlukové hierarchické aglomerativní analýzy metodou nejbližšího souseda budeme analyzovat socioekonomická data o evropských zemích z roku 1979, tedy z doby, kdy Evropa ještě byla rozdělena na demokratické státy, komunistické státy a kapitalistické státy s diktaturami. Máme k dispozici procenta zastoupení činného obyvatelstva v různých pracovních kategoriích: zemědělství (1), těžba (2), průmyslová výroba (3), energetika (4), stavebnictví (5), místní hospodářství (6), finance (7), služby (8), doprava a komunikace (9) v jednotlivých evropských zemích. Počítáme euklidovskou vzdálenost. Data jsou uvedena v tabulce 13.18 na následující straně.

Analýza pomocí hierarchické shlukové analýzy ukazuje, že země se dělí do tří skupin v závislosti na politickém zaměření systému. První skupinu tvoří země komunistického bloku (22, 21, 20, 25, 24, 23, 19, 14, 12) se Španělskem (26) a Portugalskem (14), druhou skupinu demokratické kapitalistické země a třetí skupinu Jugoslávie (26), Řecko (5) a Turecko (18), které sdílejí charakteristiky obou systémů, podobající se spíše asijským zemím. Irsko (5) je singulární entita.

Tab. 13.18 Příklad shlukové analýzy – socioekonomická data o evropských státech

ID	Stát	(1) Agro.	(2) Doly	(3) Průmysl	(4) Energetika	(5) Stavebn.	(6) Míst.hosp.	(7) Finance	(8) Služby	(9) Doprava
1	Belgie	3,3	0,9	27,6	0,9	8,2	19,1	6,2	26,6	7,2
2	Dánsko	9,2	0,1	21,8	0,6	8,3	14,6	6,5	32,2	7,1
3	Francie	10,8	0,8	27,5	0,9	8,9	16,8	6	22,6	5,7
4	Záp. Německo	6,7	1,3	35,8	0,9	7,3	14,4	5	22,3	6,1
5	Irsko	23,2	1	20,7	1,3	7,5	16,8	2,8	20,8	6,1
6	Itálie	15,9	0,6	27,6	0,5	10	18,1	1,6	20,1	5,7
7	Lucembursko	7,7	3,1	30,8	0,8	9,2	18,5	4,6	19,2	6,2
8	Nizozemsko	6,3	0,1	22,5	1	9,9	18	6,8	28,5	6,8
9	Velká Británie	2,7	1,4	30,2	1,4	6,9	16,9	5,7	28,3	6,4
10	Rakousko	12,7	1,1	30,2	1,4	9	16,8	4,9	16,8	7
11	Finsko	13	0,4	25,9	1,3	7,4	14,7	5,5	24,3	7,6
12	Řecko	41,4	0,6	17,6	0,6	8,1	11,5	2,4	11	6,7
13	Norsko	9	0,5	22,4	0,8	8,6	16,9	4,7	27,6	9,4
14	Portugalsko	27,8	0,3	24,5	0,6	8,4	13,3	2,7	16,7	5,7
15	Španělsko	22,9	0,8	28,5	0,7	11,5	9,7	8,5	11,8	5,5
16	Švédsko	6,1	0,4	25,9	0,8	7,2	14,4	6	32,4	6,8
17	Švýcarsko	7,7	0,2	37,8	0,8	9,5	17,5	5,3	15,4	5,7
18	Turecko	66,8	0,7	7,9	0,1	2,8	5,2	1,1	11,9	3,2
19	Bulharsko	23,6	1,9	32,3	0,6	7,9	8	0,7	18,2	6,7
20	Československo	16,5	2,9	35,5	1,2	8,7	9,2	0,9	17,9	7
21	Vých. Německo	4,2	2,9	41,2	1,3	7,6	11,2	1,2	22,1	8,4
22	Maďarsko	21,7	3,1	29,6,0	1,9	8,2	9,4	0,9	17,2	8
23	Polsko	31,1	2,5	25,7	0,9	8,4	7,5	0,9	16,1	6,9
24	Rumunsko	34,7	2,1	30,1	0,6	8,7	5,9	1,3	11,7	5
25	SSSR	23,7	1,4	25,8	0,6	9,2	6,1	0,5	23,6	9,3
26	Jugoslávie	48,7	1,5	16,8	1,1	4,9	6,4	11,3	5,3	4

Obr. 13.15 Zobrazení výsledků shlukové analýzy ekonomických údajů o evropských zemích



13.7 Analýza hlavních komponent

Analýza hlavních komponent (**PCA**, principal component analysis) se zabývá možností redukce počtu proměnných pomocí tzv. hlavních komponent, kterými popisujeme variabilitu všech proměnných a vztahy mezi nimi. Na rozdíl od regresní analýzy neexistuje při této analýze dělení na závisle a nezávisle proměnné. Všechny proměnné mají stejný status. Tuto techniku poprvé popsal Karl Pearson v roce 1901. Metody analýzy hlavních komponent dále vyvinul Harold Hotelling v třicátých letech minulého století. Jako prostředek analýzy dat se však rozšířila teprve s rozvojem výpočetní techniky.

Hlavní komponenty vznikají jako lineární kombinace původních proměnných. Zkoumání hodnot nových proměnných (hlavních komponent) místo původních hodnot nám mnohdy umožňuje snadněji porozumět posuzovaným datům.

Analýza hlavních komponent patří k nejjednodušším vícerozměrným metodám. Cílem analýzy je z p proměnných X_i vytvořit nové proměnné Z_j, jež jsou nekorelované. Nekorelovanost je užitečná vlastnost, protože znamená, že každá z nových proměnných Z_j měří jinou vlastnost dat (dimenzi dat). Nové proměnné jsou navíc uspořádány podle svého rozptylu a to tak, že $Var(Z_1) > Var(Z_2) > \cdots > Var(Z_p)$. Proměnné Z_j nazýváme **hlavní komponenty**. Když provádíme analýzu hlavních komponent, doufáme, že pouze několik z nich má nezanedbatelný rozptyl. Ostatní pak můžeme při analýze zanedbat. Tak dosáhneme úspornější popis chování původních p proměnných X_i pomocí menšího počtu proměnných Z_j. V datech však musí být pro tuto redukci předpoklady. Především to znamená, že musí být mezi sebou silně korelovaná.

Analýza hlavních komponent začíná tabulkou dat pro p proměnných u n jedinců (viz tab. 13.19). **První hlavní komponenta** je lineární kombinaci proměnných $X_1, X_2, \ldots, X_p$

$$Z_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p,$$

Tab. 13.19 Tabulka pro vícerozměrnou analýzu

	Proměnné			
Jedinci	X_1	X_2	$\cdots$	X_p
1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	x_{21}	x_{22}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	x_{n1}	x_{n2}	$\cdots$	x_{np}

která má co největší variabilitu mezi jedinci za podmínky, že konstanty a_{ij} splňují rovnici

$$a_{11}^2 + a_{12}^2 + \cdots + a_{1p}^2 = 1.$$

Tato podmínka je zavedena proto, aby se rozptyl $Var(Z_1)$ nemohl zvětšit pouhým zvětšováním jednotlivých konstant a_{ij}. Takže rozptyl proměnné Z_1 je pokud možno největší za této omezující podmínky. Druhá hlavní komponenta

$$Z_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$$

je taková, že $Var(Z_2)$ je pokud možno největší opět za podmínky

$$a_{21}^2 + a_{22}^2 + \cdots + a_{2p}^2 = 1.$$

Dále musí být splněno, že proměnné Z_1 a Z_2 jsou nekorelované. Třetí hlavní komponenta

$$Z_3 = a_{31}X_1 + a_{32}X_2 + \cdots + a_{3p}X_p$$

je taková, že $Var(Z_3)$ je pokud možno největší za podmínky

$$a_{31}^2 + a_{32}^2 + \cdots + a_{3p}^2 = 1$$

a omezení, že Z_3 je nekorelovaná s oběma hlavními komponentami Z_1 a Z_2. Další hlavní komponenty vznikají stejným procesem. Jestliže východiskem je p proměnných, může tak vzniknout p hlavních komponent. Obvykle však větší rozptyl má pouze několik prvních hlavních komponent. Ty slouží pro rekonstrukci všech proměnných X_i.

Metodou analýzy hlavních komponent transformujeme p souřadných os pro původní proměnné do nových p os, přičemž velikost rozptylů pro body na nových osách je určena velikostí korelací proměnných mezi sebou. Pokud korelují všechny proměnné s korelací blízkou jedné v absolutní hodnotě, je možné celkový rozptyl zachytit pouze jednou hlavní komponentou (novou osou). V případě dvou proměnných leží v tomto případě všechny body na přímce, která je identická s novou osou – první hlavní komponentou (srov. obr. 7.14, s. 287). Jestliže jsou naopak všechny korelace malé, potřebujeme k vysvětlení celkového rozptylu tolik komponent, kolik je původních proměnných. V tomto případě odpovídají hlavní komponenty původním proměnným – každá komponenta vysvětluje variabilitu jedné proměnné.

Popsané rovnice pro vyjádření hlavních komponent pomocí původních proměnných se dají obrátit s cílem vyjádřit původní proměnné pomocí hlavních komponent. Původní proměnné jsou tak vyjádřeny pomocí p nezávislých hlavních komponent. Z nich pouze několik prvních má významně veliké rozptyly.

Komponenty s malými rozptyly lze zanedbat. Korelace mezi proměnnou a hlavní komponentou se nazývá faktorová zátěž. Součet faktorových zátěží pro všechny proměnné λ_j u zvolené komponenty j odpovídá celkovém rozptylu, který komponenta j vysvětluje. Celkový rozptyl všech proměnných se rovná jejich počtu. Hodnotě λ_j říkáme vlastní hodnota hlavní komponenty j.

Rao (1978, s. 635) ukazuje, že jestliže chceme predikovat původní proměnné pomocí menšího počtu umělých proměnných vytvořených jako lineární kombinace původních proměnných, pak to nejlépe dokážeme právě pomocí hlavních komponent.

PŘÍKLAD 13.9

Analýza hlavních komponent

Vrátíme se k příkladu 10.1 z kapitoly 10.1 o regresní analýze. Z korelační matice je vidět, že přírodovědné či literární vědomosti, koncentrační schopnost a logické myšlení spolu silně korelují. Usuzujeme, že tyto čtyři testy částečně měří totéž. Uvádíme příslušnou korelační matici ještě jednou, a to pouze jednu z jejich symetrických částí (tab. 13.20). Metodou hlavních komponent lze zjistit, že první hlavní komponenta vystihuje 91 % variability v datech. Usuzujeme, že měření se dají popsat prakticky jednou proměnnou, kterou však nemůžeme měřit, ale dopočítáváme ji pomocí hodnot proměnných X_i. Pomocí této nové proměnné dobře odhadneme původní hodnoty testových výsledků. Uvádíme vzorce pro výpočet první hlavní komponenty a pro výpočty odhadů původních proměnných pomocí hlavní komponenty:

$$Z_1 = 0{,}34X_1 + 0{,}30X_2 + 0{,}27X_3 + 0{,}31X_4 - 5{,}6$$

Rekonstrukce proměnných pomocí první hlavní komponenty:

$$X_1 = 1{,}20Z_1 + 4{,}33$$
$$X_2 = 1{,}10Z_1 + 4{,}18$$
$$X_3 = 0{,}99Z_1 + 4{,}00$$
$$X_4 = 1{,}14Z_1 + 5{,}00$$

Kvalitu aproximace X_i (i = 1, 2, 3, 4) lze posoudit korelačním koeficientem se Z_1. Hodnoty těchto korelačních koeficientů jsou postupně; 0,95; 0,964; 0,974; 0,935.

Tab. 13.20 Příklad analýzy hlavních komponent – korelační matice popisující data

	X_1	X_2	X_3	X_4
X_1 (přírodovědné vědomosti)	1,00	0,907	0,87	0,87
X_2 (literární vědomosti)		1,00	0,96	0,82
X_3 (schopnost koncentrace)			1,00	0,89
X_4 (logické myšlení)				1,00

13.7.1 Postup při analýze hlavních komponent

1. **Počáteční analýza.** Jako při každé analýze je důležitý explorační pohled na dat. Počítáme popisné statistiky a sestrojujeme grafy, které mají pomocí získat „cit pro data". Zvláště si všímáme vztahů mezi proměnnými.
2. **Průzkum korelační matice.** Na korelacích mezi proměnnými záleží, zda lze provést redukci jejich počtu pomocí hlavních komponent. Proto před vlastní aplikací procedur pro analýzu hlavních komponent vypočítáme korelační matici a kontrolujeme, zda v ní existují silně korelované proměnné.
3. **Provedení základních procedur analýzy hlavních komponent a rozhodnutí o vhodném počtu hlavních komponent.** Rozhodujeme se, kolika hlavními komponentami je možné popsat data bez ztráty informace. Obvykle stačí, aby vybrané hlavní komponenty vysvětlily 80–90 % variability dat. Sestrojujeme graf, jenž zobrazuje závislost vysvětlené variability dat na počtu vybraných hlavních komponent – **scree graf**.
4. **Interpretace hlavních komponent.** Pro popis dat zachováme q hlavních komponent ($q < p$).

Dále postupujeme takto:

a) Vypočítáme hodnoty hlavních komponent pro každý řádek v matici dat. Následně sestrojujeme bodové grafy pomocí dvou až tří hlavních komponent.
b) Pokusíme se interpretovat hlavní komponenty. To není obvykle jednoduché. Zkoumáme přitom, jak jsou jednotlivé komponenty vytvořené z analyzovaných proměnných.
c) Někdy využíváme hlavní komponenty pro regresní analýzu jako nezávisle proměnné. Je to výhodné proto, že nejsou korelované. Tím se vyhneme problému kolinearity proměnných, která způsobuje nestabilitu řešení při hledání regresní rovnice.
d) Kromě vizuální analýzy bodových grafů sestrojených pomocí hlavních komponent také využíváme procedur shlukové analýzy.
e) Výsledky analýzy hlavních komponent mohou sloužit jako podklad pro faktorovou analýzu.

PŘÍKLAD 13.10

Analýza hlavních komponent

Vraťme se k datům z příkladu shlukové analýzy údajů o evropských zemích (viz příklad 13.8, s. 465). Každý stát byl popsán procentuálními údaji o zastoupení obyvatelstva činného v různých odvětvích. Nejdříve uvádíme základní tabulku analýzy hlavních komponent, která obsahuje vlastní hodnoty, rozsah variability v datech vysvětlené každou hlavní komponentou a kumulativní vysvětlený podíl variability prvními hlavními komponentami. Počet hlavních

Obr. 13.16 Části variabilty vysvětlené jednotlivými hlavními komponentami

Vlastní hodnoty

Poř.	Vlastní h.	Individuální procento	Kumulativní procento
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

Obr. 13.17 Výpis obsahující zátěže proměnných v prvních dvou faktorech

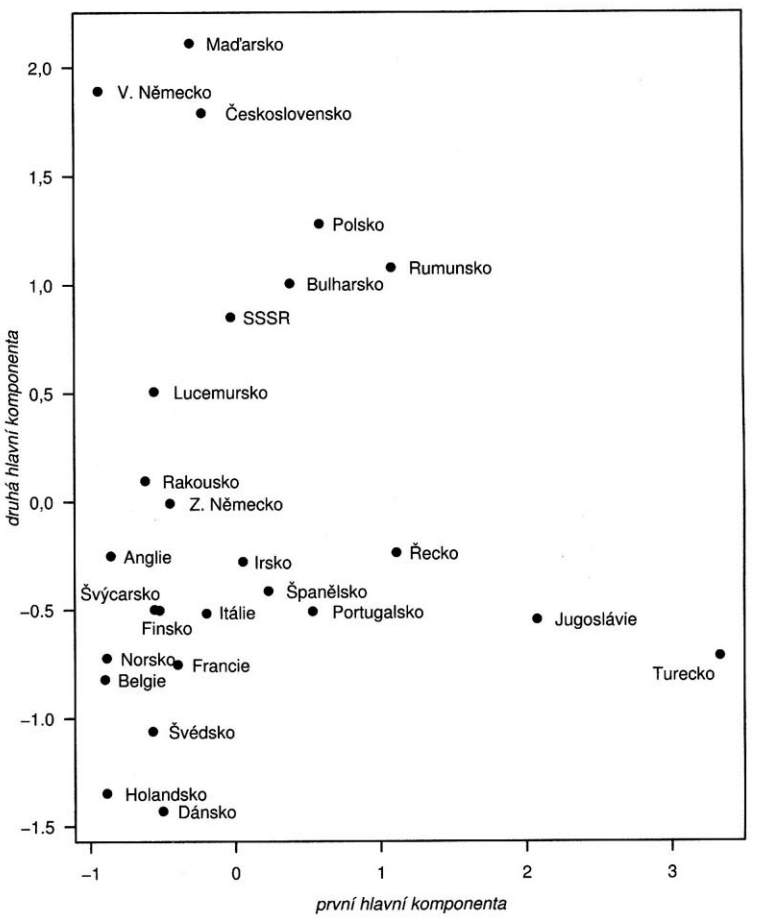
	Hlavní komponenty	
Proměnné	HK1	HK2
Agro	0,523791	0,053594
Doly	0,001323	0,617807
Průmysl	-0,347495	0,355054
Energetika	-0,255716	0,261096
Stavebnictví	-0,325179	0,051288
Místní hospodářství	-0,378920	-0,350172
Finance	-0,074374	-0,453698
Služby	-0,387409	-0,221521
Doprava	-0,366823	0,202592

komponent se rovná počtu proměnných. Velikost vlastních hodnot je úměrná schopnosti hlavní komponenty vysvětlit variabilitu v datech. První dvě hlavní komponenty vysvětlují 38,7 % a 23,7 % variability dat. Třetí hlavní komponenta vysvětluje 12 % (srov. obr. 13.16).

Podíváme se podrobněji na zátěže prvních dvou hlavních komponent, které jsou popsané obrázkem 13.17. První hlavní komponenta HK1 má kladnou zátěž na proměnné Agro (tj. podíl obyvatel zaměstnaných v zemědělství) a zátěž zápornou nebo nulovou na všechny ostatní proměnné. Tato hlavní komponenta diferencuje země podle rozsahu zemědělské výroby a rozlišuje zemědělské a průmyslové země. Je patrné, že Turecko i Jugoslávie mají velké hodnoty této hlavní komponenty.

Druhá hlavní komponenta má zápornou zátěž u místního hospodářství, financí a sociálních služeb. Tato hlavní komponenta diferencuje státy s velkým a malým sektorem služeb.

Obr. 13.18 Zobrazení dat ekonomik zemí pomocí prvních dvou hlavních komponent



Kapitalistické země mají menší hodnoty druhé hlavní komponenty než komunistické státy, což naznačuje, že západní ekonomiky měly oblast služeb rozvinutější.

Hodnoty prvních dvou hlavních komponent znázorníme obrázkem 13.18, který se používá pro odhalení podobností a nepodobností sledovaných objektů (v našem případě států).