Pokročilé metody analýzy dat v neurovědách

RNDr. Eva Janoušová

doc. RNDr. Ladislav Dušek, Dr.

Jaro 2015

Blok 5

Ordinační analýzy I

Osnova

1. Principy redukce dimenzionality dat
2. Selekce a extrakce proměnných
3. Analýza hlavních komponent (PCA)
4. Faktorová analýza (FA)

Principy redukce dimenzionality dat

Schéma analýzy a klasifikace dat

Data

↓

Předzpracování

↓

Redukce

↓

Klasifikace

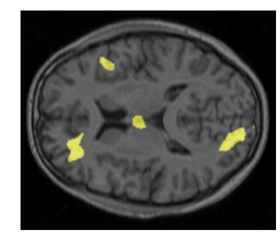
Ukázka - kognitivní data apod.

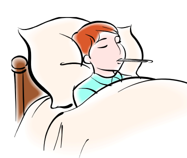
	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M		90
4	3	26	Z	178	70

	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M	167	90
4	3	26	Z	178	70

	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M	167	90
4	3	26	Z	178	70

Ukázka - obrazová data



 ? nebo  ?

Proč používat redukci dat?

Obrazová data

Klasifikace

X voxely

subjekty

	x_1 x_2 …
I_1	
I_2	270 x 1 000 000
…	

subjekty

I_1	pac.
I_2	kon.
…	

Proč používat redukci dat?

Obrazová data

→ Redukce dat → Klasifikace

X voxely

subjekty	x_1 x_2 …
I_1	
I_2	270 x 1 000 000
…	

→

voxely

subjekty	x_1 x_5 …
I_1	
I_2	270 x 1 000
…	

→

subjekty	
I_1	pac.
I_2	kon.
…	

Proč používat redukci dat?

- zjednodušení další práce s daty
- možnost použití metod analýzy dat, které by na původní data nebylo možno použít
- umožnění vizualizace vícerozměrných dat – může být nápomocné k nalezení vztahů v datech či k jejich interpretaci
- redukce dat může být i cílem analýzy (např. identifikace oblastí mozku, kde se nejvíce liší od sebe liší skupiny subjektů)

Volba a výběr proměnných – úvod

- počáteční volba proměnných je z velké části empirická, vychází ze zkušeností získaných při empirické klasifikaci člověkem a závisí kromě rozboru podstaty problému i na technických (ekonomických) možnostech a schopnostech hodnoty proměnných určit

- kolik a jaké proměnné?
  - málo proměnných – možná nízká úspěšnost klasifikace či jiných následných analýz
  - moc proměnných – možná nepřiměřená pracnost, vysoké náklady

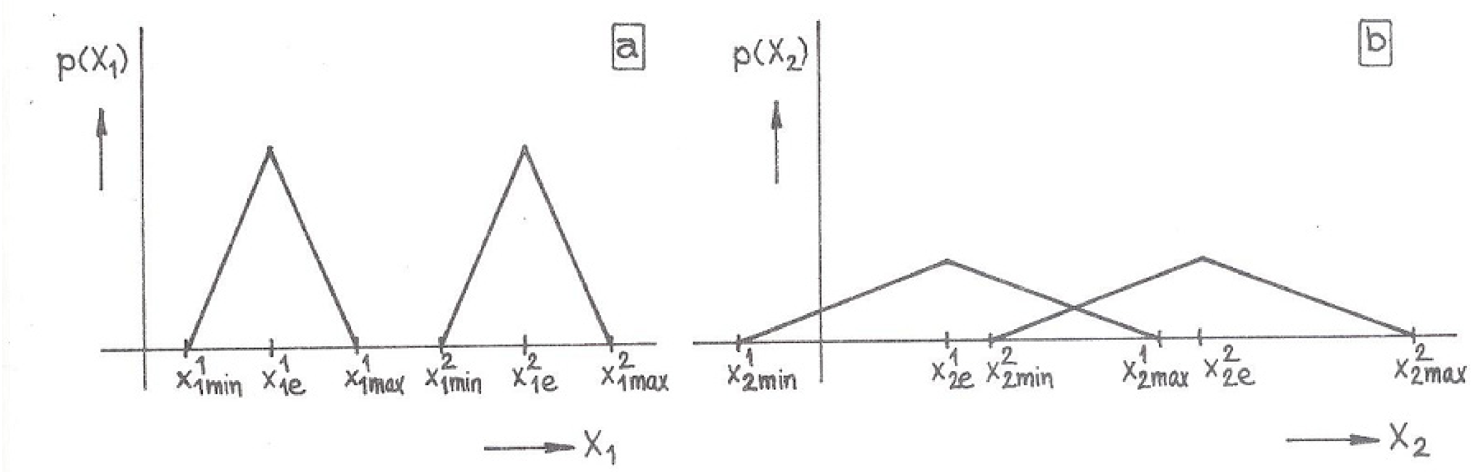
$\Downarrow$

KOMPROMIS

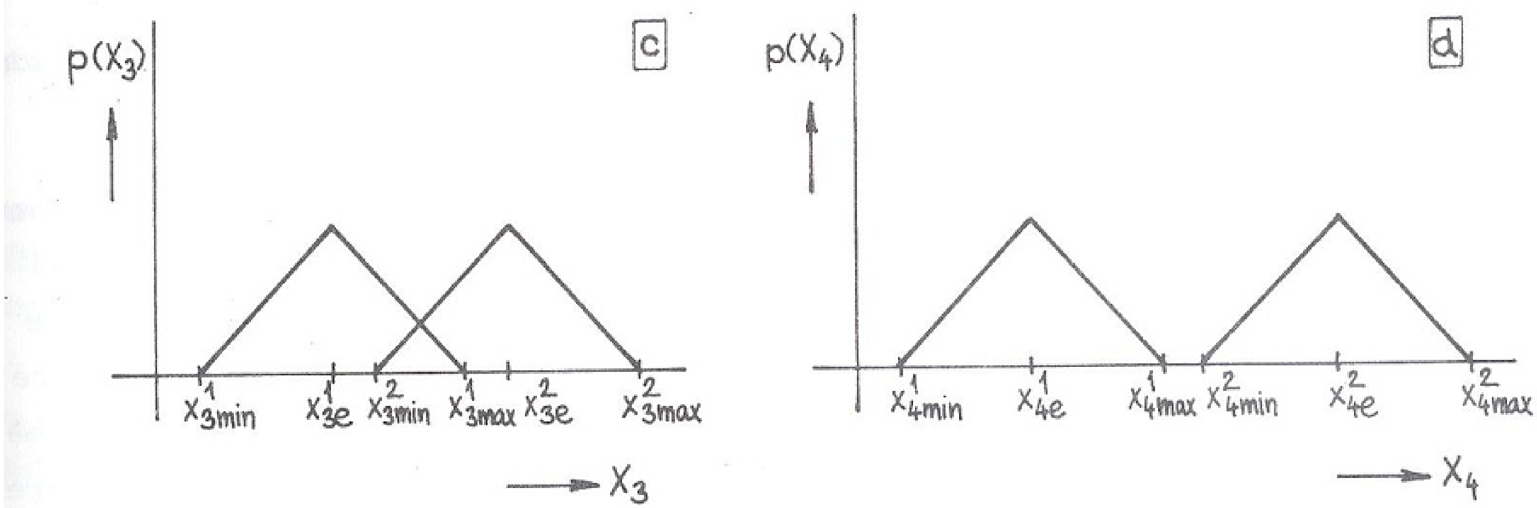
(určit ty proměnné, jejichž hodnoty nesou nejvíce informace z hlediska řešené úlohy, tj. např. ty proměnné, kterou jsou nejefektivnější pro vytvoření co nejoddělenějších klasifikačních tříd)

Zásady pro volbu proměnných I

- výběr proměnných s minimálním rozptylem uvnitř tříd



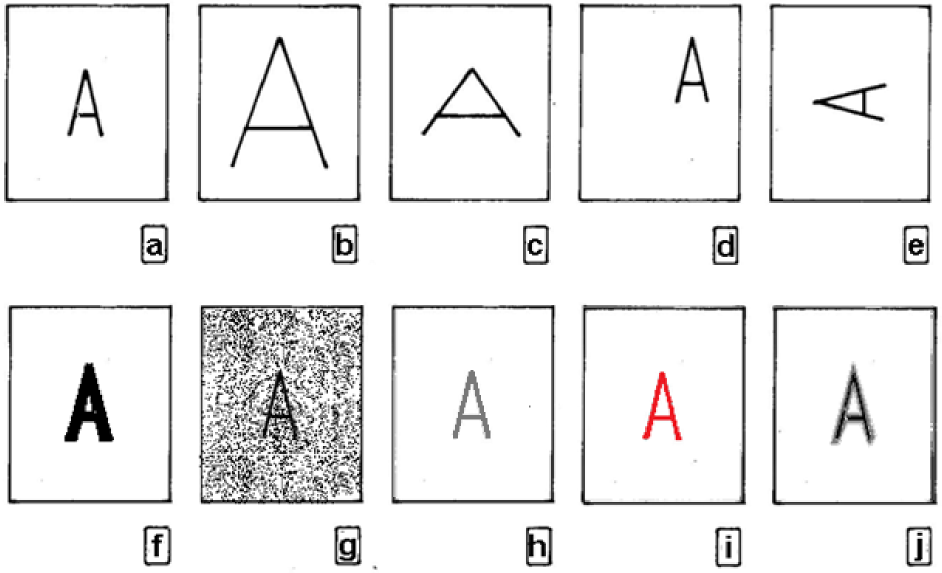
- výběr proměnných s maximální vzdáleností mezi třídami



Zásady pro volbu proměnných II

- výběr vzájemně nekorelovaných proměnných
  - pokud jsou hodnoty jedné proměnné závislé na hodnotách druhé proměnné, pak použití obou těchto proměnných nepřináší žádnou další informaci – stačí jedna z nich, jedno která

- výběr proměnných invariantních vůči deformacím
  - volba elementů formálního popisu závisí na vlastnostech původních i předzpracovaných dat a může ovlivňovat způsob předzpracování

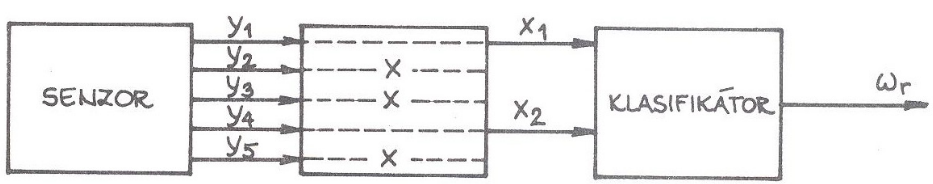


Selekce a extrakce proměnných

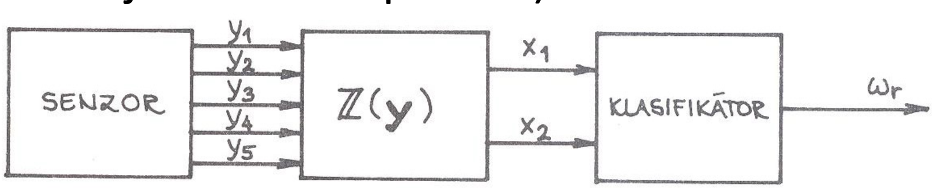
Selekce a extrakce proměnných

- formální popis objektu původně reprezentovaný *m* rozměrným vektorem se snažíme vyjádřit vektorem *n* rozměrným tak, aby množství diskriminační informace obsažené v původním vektoru bylo v co největší míře zachováno
- dva principiálně různé způsoby:

1. **selekce** – nalezení a odstranění těch proměnných, které přispívají k separabilitě klasifikačních tříd nejméně



2. **extrakce** – transformace původních proměnných na menší počet jiných proměnných (které zpravidla nelze přímo měřit a často nemají zcela jasnou interpretaci)



Selekce proměnných

- cílem je výběr proměnných, které jsou nejužitečnější pro další analýzu (např. při klasifikaci výběr takových proměnných, které nejlépe od sebe dokáží oddělit skupiny subjektů/objektů)

- metod selekce je velké množství, nejpoužívanější metody jsou:
  - výběr proměnných na základě statistických testů
  - výběr oblastí mozku (ROI) podle atlasu
  - algoritmy sekvenční selekce (dopředné či zpětné nebo algoritmus plus p mínus q)

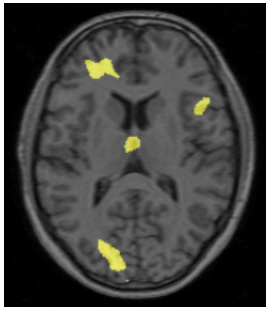
Výběr proměnných na základě statistických testů

Princip: Výběr statisticky významných proměnných pomocí dvouvýběrového t-testu či Mannova-Whitneyova testu.

proměnné

subjekty

		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	…
I_1	pac.									
I_2	pac.									
I_3	kont.									
I_4	pac.									
I_5	kont.									
…										
p-hodnoty:		0,34	0,02	0,09	0,01	0,25	0,63	0,03	0,12	



Výhody:

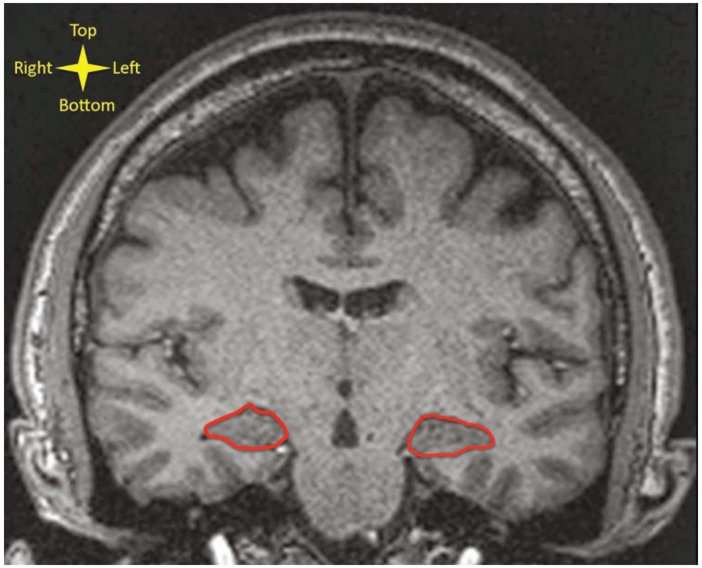
+ rychlé
+ u obrazů mozku výhodou, že je analýza provedena na celém mozku

Nevýhody:

- jednorozměrná metoda (výběr proměnných bez ohledu na ostatní proměnné)
- potřeba použít metody korekce pro mnohonásobné testování (např. FDR)

Výběr oblastí mozku (ROI) podle atlasu

Princip: Výběr oblastí mozku s využitím atlasu mozku podle expertní znalosti daného onemocnění (tzn. výběr oblasti postižené danou nemocí).



Výhody:

+ anatomicky/funkčně relevantní – snadnější interpretace
+ zpravidla rychlé

Nevýhody:

- ne vždy dopředu víme, která z oblastí je vhodná pro odlišení skupin osob
- některá onemocnění postihují celý mozek (např. schizofrenie)

Algoritmy sekvenční selekce

- algoritmus sekvenční dopředné selekce:
  - algoritmus začíná s prázdnou množinou, do které se vloží proměnná s nejlepší hodnotou selekčního kritéria
  - v každém následujícím kroku se přidá ta proměnná, která s dříve vybranými veličinami dosáhla nejlepší hodnoty kritéria
- algoritmus sekvenční zpětné selekce:
  - algoritmus začíná s množinou všech proměnných
  - v každém následujícím kroku se eliminuje ta proměnná, která způsobuje nejmenší pokles kriteriální funkce

Výhody : + dopředný algoritmus je výpočetně jednodušší, protože pracuje maximálně v n-rozměrném prostoru
+ zpětný algoritmus umožňuje průběžně sledovat množství ztracené informace

Nevýhody : - dopředná selekce – nelze vyloučit ty veličiny, které se staly nadbytečné po přiřazení dalších veličin
- zpětná selekce – neexistuje možnost opravy při neoptimálním vyloučení kterékoliv proměnné

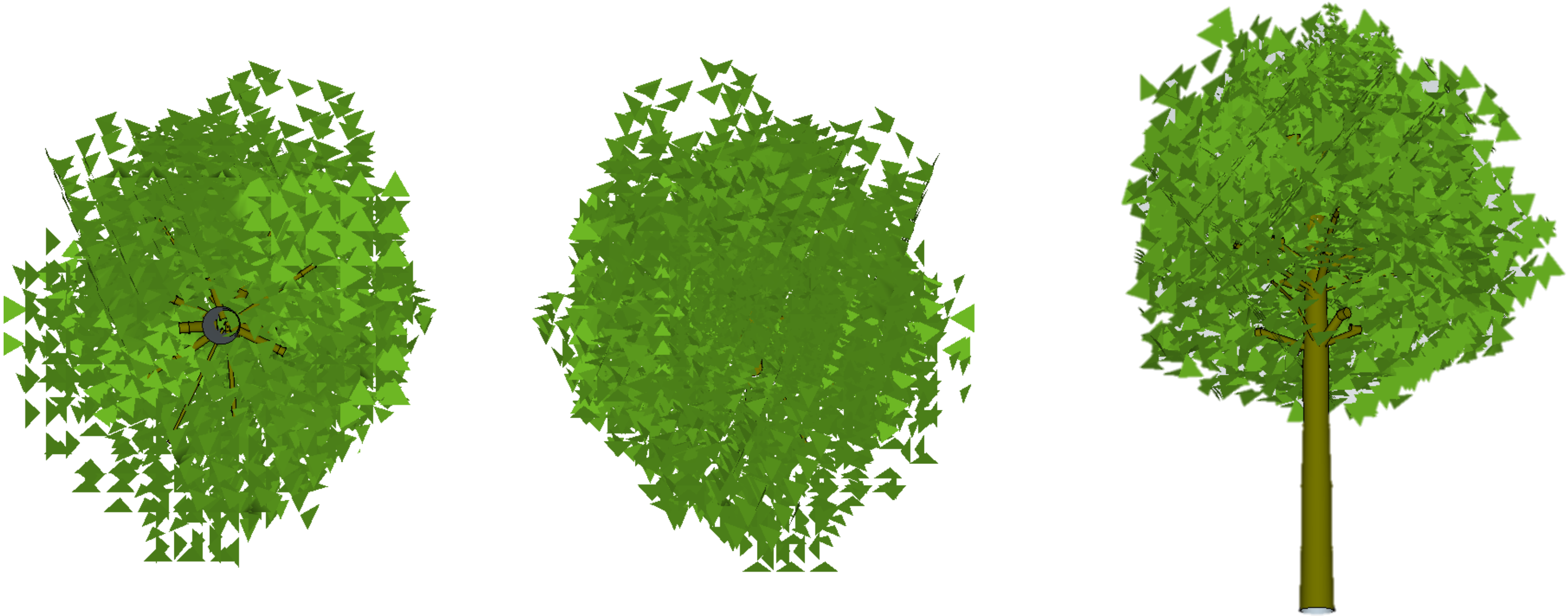
- algoritmus plus p mínus q:
  - po přidání p veličin se q veličin odstraní;
  - proces probíhá, dokud se nedosáhne požadovaného počtu příznaků

Extrakce proměnných

- jednou z možných přístupů redukce dat
- transformace původních proměnných na menší počet jiných proměnných $\Rightarrow$ tzn. hledání (optimálního) zobrazení Z, které transformuje původní *m*-rozměrný prostor (obraz) na prostor (obraz) *n*-rozměrný ($m \geq n$)
- pro snadnější řešitelnost hledáme zobrazení Z v oboru lineárních zobrazení
- metody extrakce proměnných:
  - analýza hlavních komponent (PCA)
  - faktorová analýza (FA)
  - analýza nezávislých komponent (ICA)
  - korespondenční analýza (CA)
  - vícerozměrné škálování (MDS)
  - redundanční analýza (RDA)
  - kanonická korelační analýza (CCorA)
  - manifold learning metody (LLE, Isomap atd.)
  - metoda parciálních nejmenších čtverců (PLS)
- metody extrakce proměnných často nazývány jako metody ordinační analýzy

Ordinační analýza dat = pohled ze správného úhlu

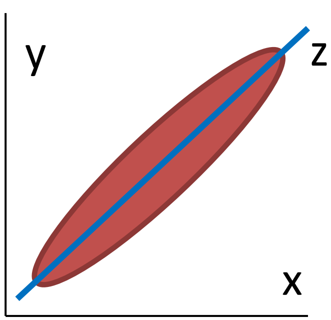
- Vícerozměrná analýza nám pomáhá nalézt v x-dimenzionálním prostoru nejvhodnější pohled na data poskytující maximum informací o analyzovaných objektech



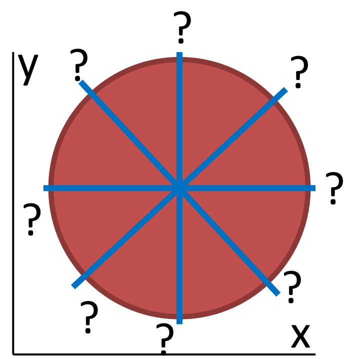
Všechny obrázky ukazují stejný objekt z různých úhlů v 3D prostoru.

Obecný princip redukce dimenzionality dat pomocí extrakce

- V převážné většině případů existují mezi dimenzemi korelační vztahy, tedy dimenze se navzájem vysvětlují a pro popis kompletní informace v datech není třeba všech dimenzí vstupního souboru
- Všechny tzv. ordinační metody využívají principu identifikace korelovaných dimenzí a jejich sloučení do souhrnných nových dimenzí zastupujících několik dimenzí vstupního souboru
- Pokud mezi dimenzemi vstupního souboru neexistují korelace, <u>nemá smysl hledat zjednodušení vícerozměrné struktury takovéhoto souboru !!!</u>



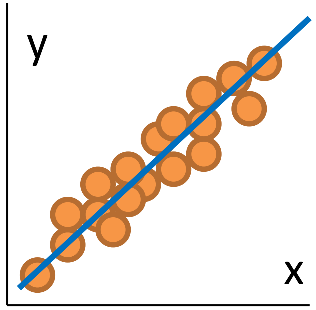
Jednoznačný vztah dimenzí x a y umožňuje jejich nahrazení jedinou novou dimenzí z



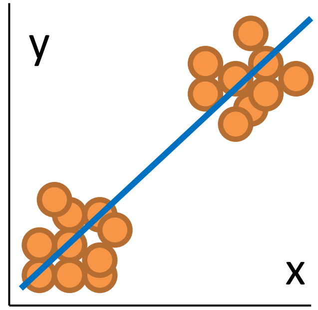
V případě neexistence vztahu mezi x a y nemá smysl definovat nové dimenze – nepřináší žádnou novou informaci oproti x a y

Korelace jako princip výpočtu vícerozměrných analýz

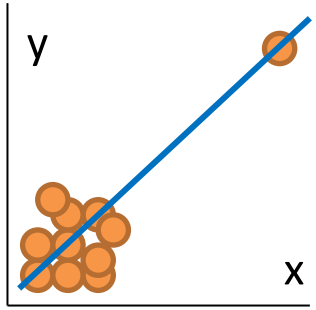
- Kovariance a Pearsonova korelace je základem analýzy hlavních komponent, faktorové analýzy jakož i dalších vícerozměrných analýz pracujících s lineární závislostí proměnných
- Předpokladem výpočtu kovariance a Pearsonovy korelace je:
  – Normalita dat v obou dimenzích
  – Linearita vztahu proměnných
- Pro vícerozměrné analýzy je nejzávažnějším problémem přítomnost odlehlých hodnot



Lineární vztah – bezproblémové použití Pearsonovy korelace



Korelace je dána 2 skupinami hodnot – vede k identifikaci skupin objektů v datech



Korelace je dána odlehlou hodnotou – analýza popisuje pouze vliv odlehlé hodnoty

Typy ordinační analýzy

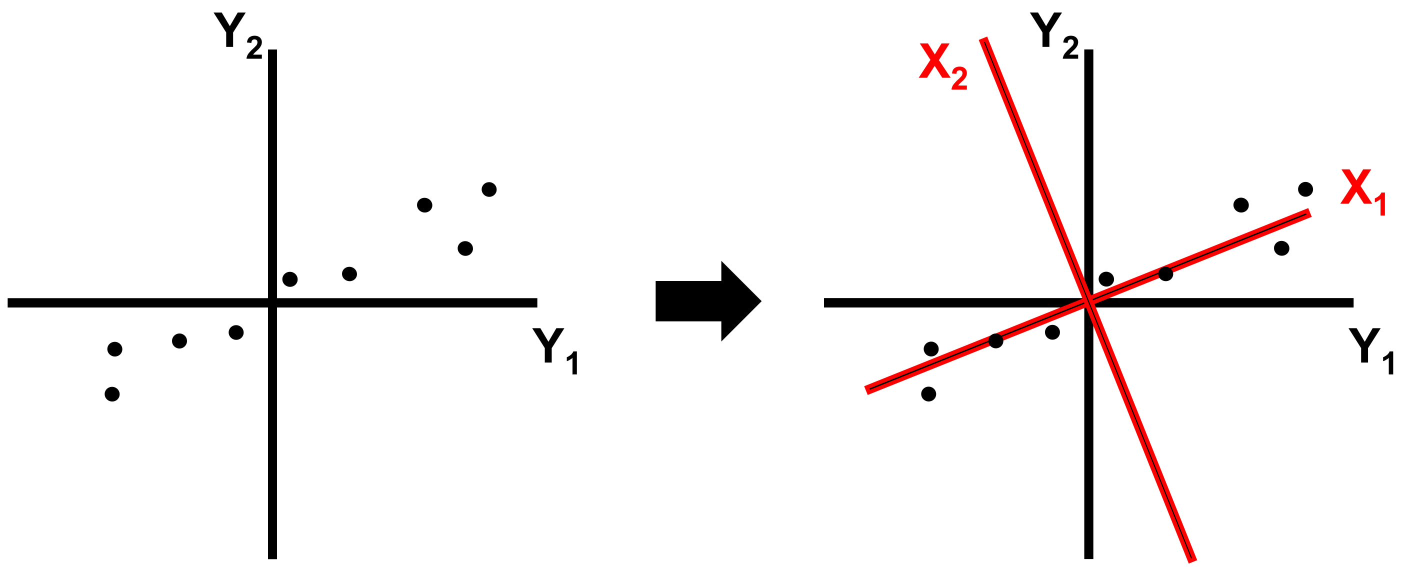
- Ordinačních analýz existuje celá řada, některé jsou spjaty s konkrétními metrikami vzdáleností/podobností
- V přehledu jsou uvedeny pouze základní typy analýz, nikoliv jejich různé kombinace hodnotící vztahy dvou a více sad proměnných (CCA, kanonická korelace, RDA, co-coordinate analysis, co-inertia analysis, diskriminační analýza apod.)

Typ analýzy	Vstupní data	Metrika
Analýza hlavních komponent (PCA)	NxP matice	Korelace, kovariance, Euklidovská
Faktorová analýza (FA)	NxP matice	Korelace, kovariance, Euklidovská
Analýza nezávislých komponent (ICA)	NxP matice	Korelace, kovariance, Euklidovská
Korespondenční analýza (CA)	NxP matice	Chi-square vzdálenost
Analýza hlavních koordinát (PCoA)	Asoc. matice	libovolná
Nemetrické mnohorozměrné škálování (MDS)	Asoc. matice	libovolná

Analýza hlavních komponent (PCA)

Analýza hlavních komponent

- anglicky Principal component analysis (PCA)
- snaha redukovat počet proměnných nalezením nových latentních proměnných (hlavních komponent) vysvětlujících co nejvíce variability původních proměnných
- nové proměnné (X_1, X_2) lineární kombinací původních proměnných (Y_1, Y_2)



Analýza hlavních komponent – cíle

- Popis a vizualizace vztahů mezi proměnnými
- Výběr neredundantních proměnných pro další analýzy
- Vytvoření zástupných faktorových os pro použití v dalších analýzách
- Identifikace shluků v datech spjatých s variabilitou dat
- Identifikace vícerozměrně odlehlých objektů

Analýza hlavních komponent – předpoklady

- vstupem do analýzy datová matice *n* x *p* obsahující kvantitativní proměnné (s normálním rozdělením)
- předpoklady obdobné jako při výpočtu korelací a kovariancí:
  - nepřítomnost odlehlých hodnot (s výjimkou situace, kdy analýzu provádíme za účelem identifikace odlehlých hodnot)
  - nepřítomnost více skupin objektů (s výjimkou situace, kdy analýzu provádíme za účelem detekce přirozeně existujících shluků spjatých s největší variabilitou souboru)
- datový soubor by měl mít více objektů než proměnných, pro získání stabilních výsledků se doporučuje alespoň 10x tolik objektů než proměnných, ideální je 40-60x více objektů než proměnných

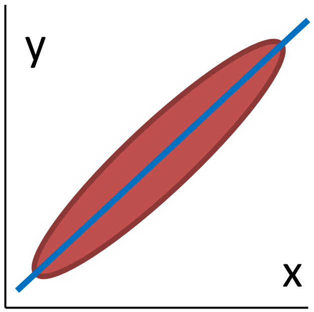
Analýza hlavních komponent – volba asociační matice

- **autokorelační matice** – data nejsou nijak upravena (zohledňována průměrná hodnota i rozptyl původních dat)
- **kovarianční (disperzní) matice** – data centrována (od každé příznakové proměnné odečtena její střední hodnota) – zohledňován rozptyl původních dat
- **matice korelačních koeficientů** – data standardizována (odečtení středních hodnot a podělení směrodatnými odchylkami) – použití pokud mají proměnné různá měřítka
- **každou úpravou původních dat ale přicházíme o určitou informaci !!!**

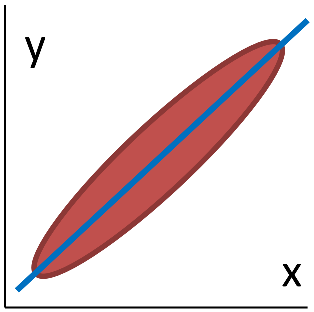
Analýza hlavních komponent – volba asociační matice

- s jakými daty PCA pracuje v případě použití různých asociačních matic:

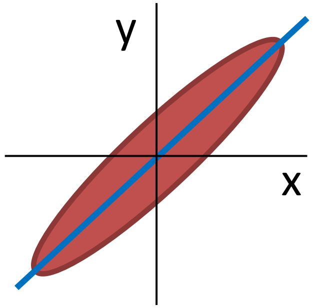
původní data



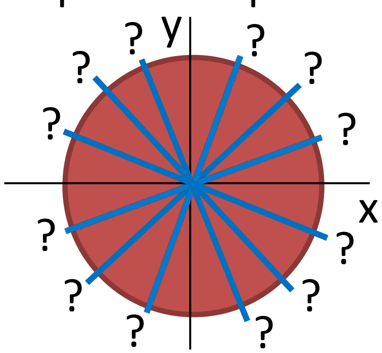
autokorelační matice
(data nijak neupravována)



kovarianční matice
(odečten průměr)



matice korelačních koeficientů
(odečten průměr a podělení SD)



Analýza hlavních komponent – postup

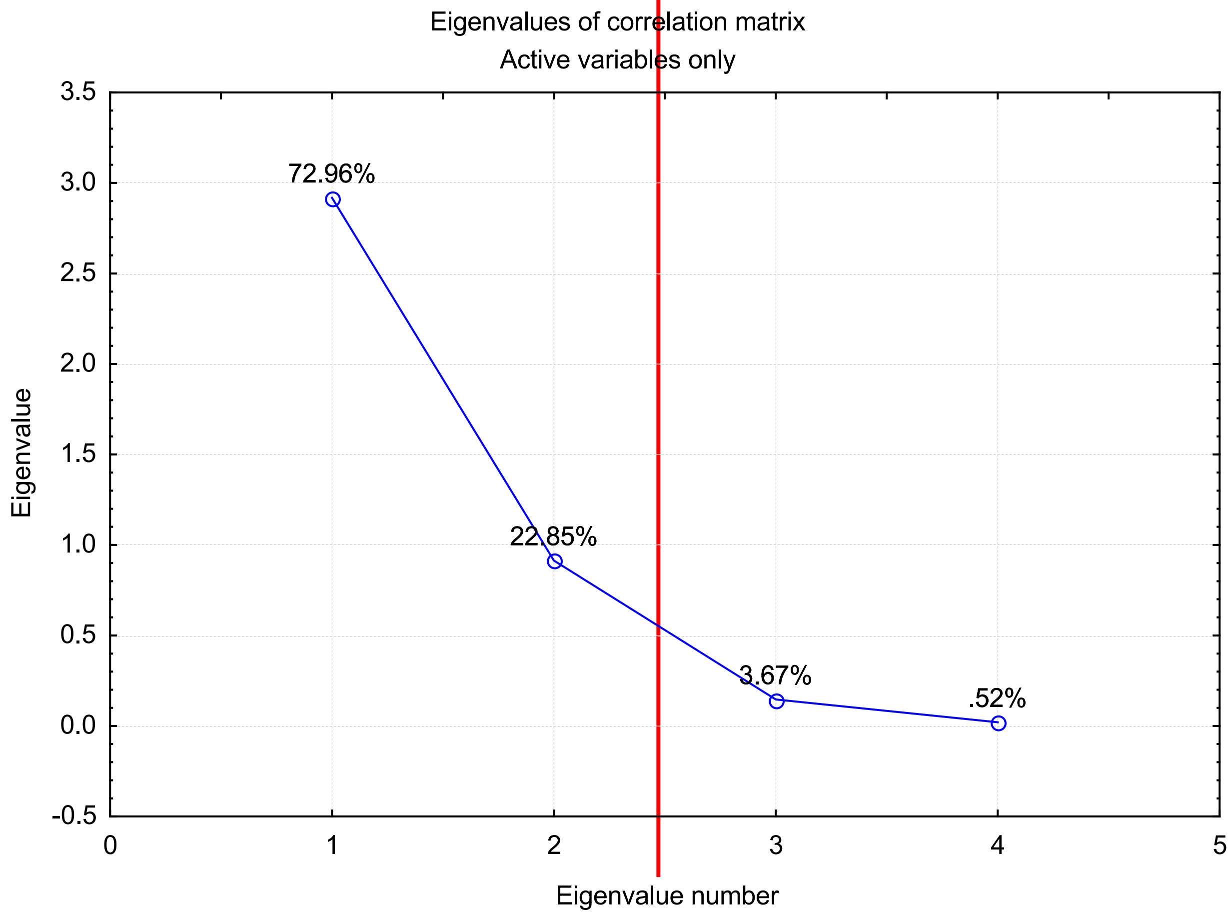
1. Volba asociační matice (autokorelační, kovarianční nebo kor. koeficientů)
2. Výpočet vlastních čísel a vlastních vektorů asociační matice:
   - vlastní vektory definují směr nových faktorových os (hlavních komponent) v prostoru
   - vlastní čísla odrážejí variabilitu vysvětlenou příslušnou komponentou
3. Seřazení vlastních vektorů podle hodnot jim odpovídajících vlastních čísel (sestupně)
4. Výběr prvních *m* komponent vyčerpávajících nejvíce variability původních dat

Identifikace optimálního počtu hlavních komponent pro další analýzu

- pokud je cílem ordinační analýzy vizualizace dat, snažíme se vybrat 2-3 komponenty
- pokud je cílem ordinační analýzy výběr menšího počtu dimenzí pro další analýzu, můžeme ponechat více komponent (např. u analýzy obrazů MRI je úspěchem redukce z milionu voxelů na desítky)
- kritéria pro výběr počtu komponent:
  1. Kaiser Guttmanovo kritérium:
     - pro další analýzu jsou vybrány osy s vlastním číslem >1 (při analýze matice korelačních koeficientů) nebo větším než průměrná hodnota vlastních čísel (při analýze kovarianční matice)
     - logika je vybírat osy, které přispívají k vysvětlení variability dat více, než připadá rovnoměrným rozdělením variability
  2. Sutinový graf (scree plot)
     - grafický nástroj hledající zlom ve vztahu počtu os a vyčerpané variability
  3. Sheppardův diagram
     - grafická analýza vztahu mezi vzdálenostmi objektů v původním prostoru a redukovaném prostoru o daném počtu dimenzí

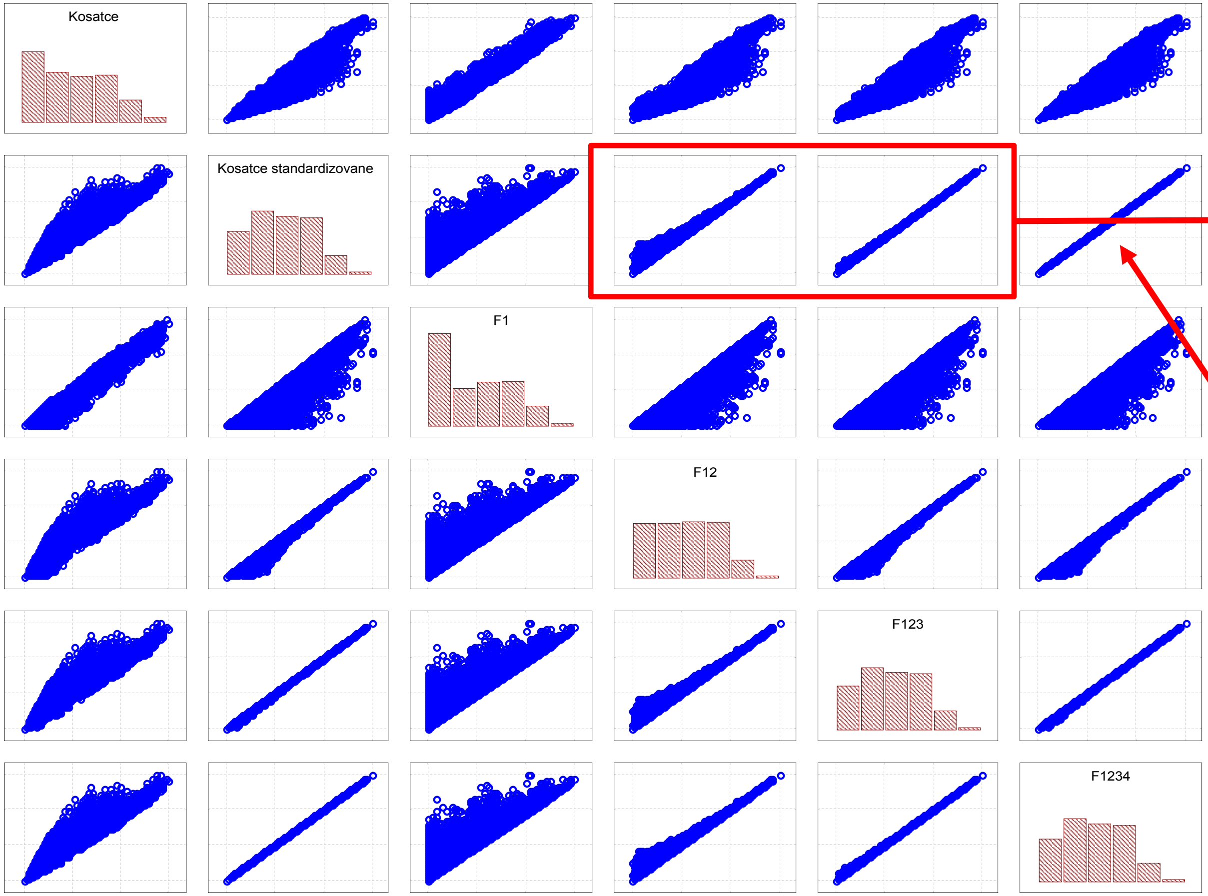
Sutinový graf (scree plot)

Zlom ve vztahu mezi počtem vlastních čísel a jimy vyčepanou variabilitou – pro další analýzu použity první dvě faktorové osy



Sheppardův diagram

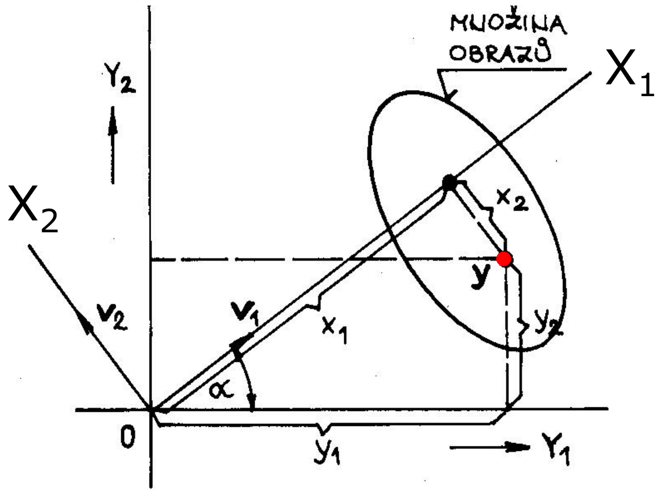
- Vztahuje vzdálenosti v prostoru původních proměnných ke vzdálenostem v prostoru vytvořeném PCA
- Je třeba brát ohled na typ PCA (korelace vs. kovariance)
- Obecná metoda určení optimálního počtu dimenzí v ordinační analýze (třeba respektovat použitou asociační metriku)



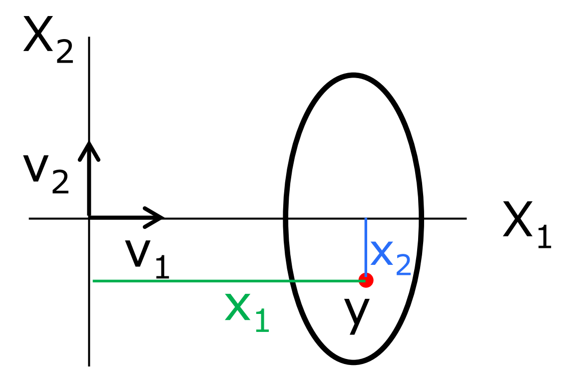
Za optimální z hlediska zachování vzdáleností objektů lze považovat dvě nebo tři dimenze

Při použití všech dimenzí jsou vzdálenosti perfektně zachovány

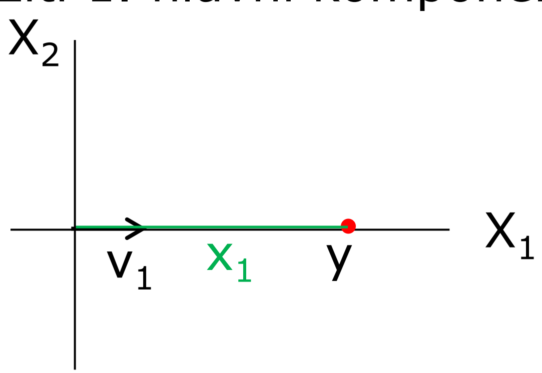
PCA – geometrická interpretace



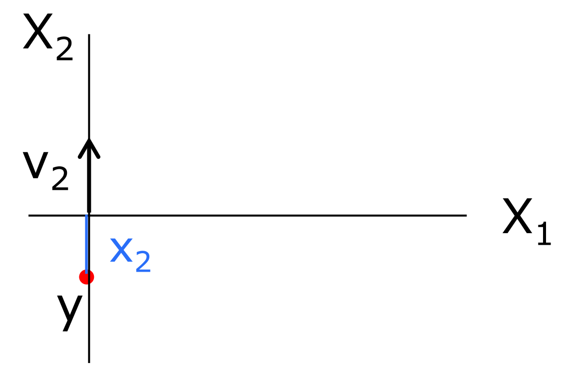
použití obou hlavních komponent



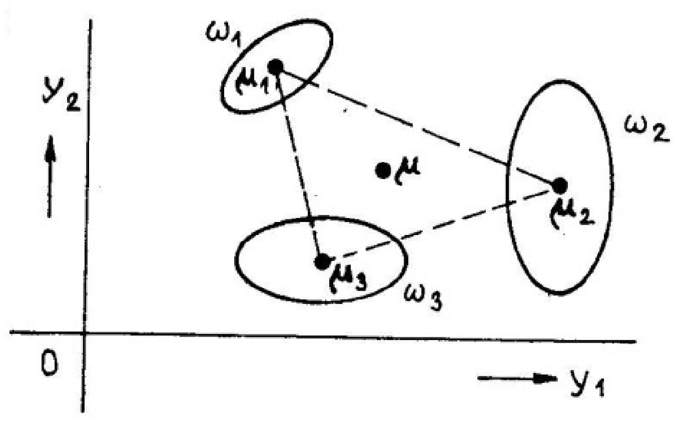
použití 1. hlavní komponenty



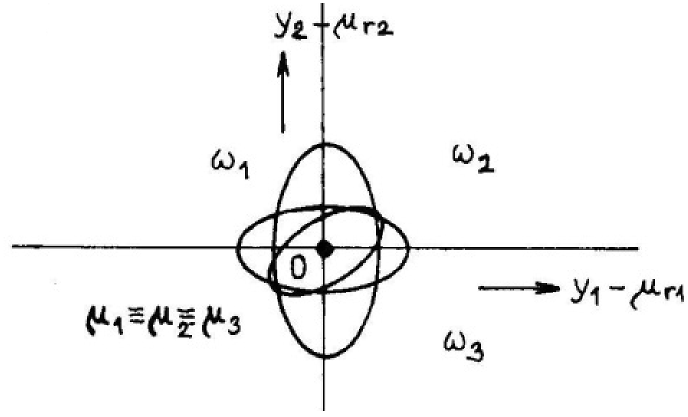
použití 2. hlavní komponenty



PCA – rozdělení do tříd

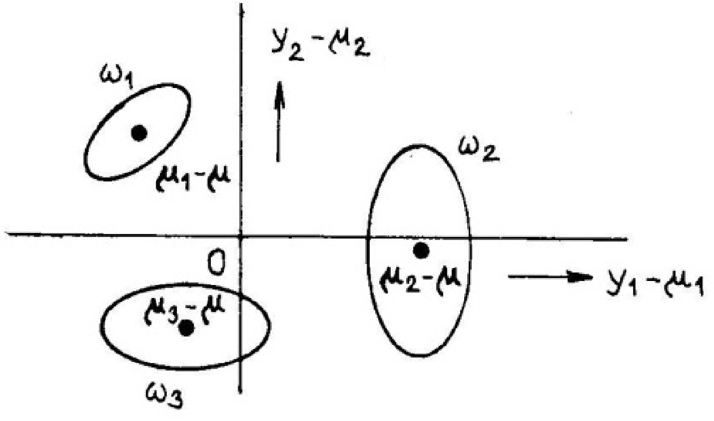


odečtení průměru každé skupiny zvlášť



→ není vhodné

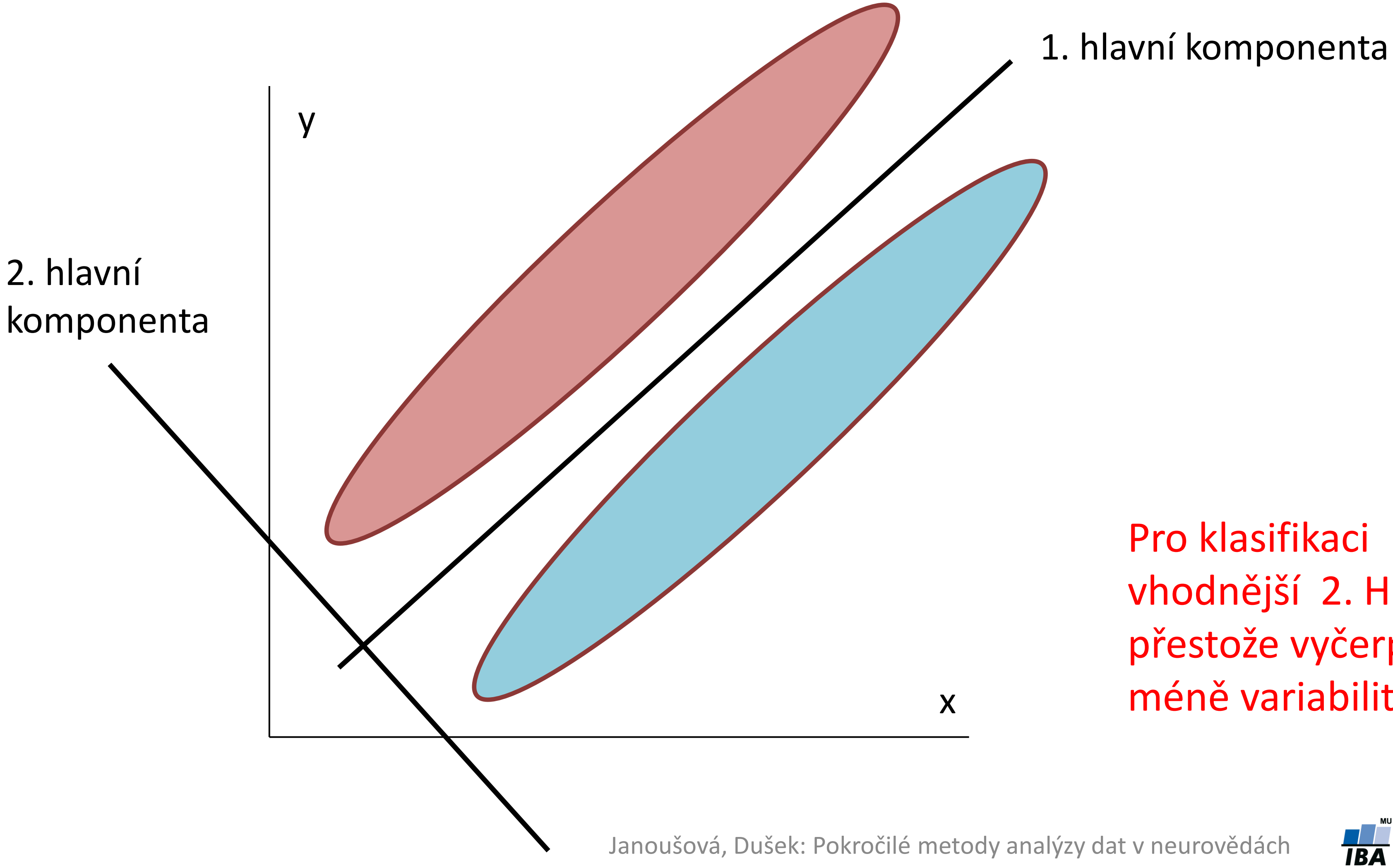
odečtení celkového průměru



→ je vhodné

PCA a klasifikace

- PCA často nebývá vhodnou metodou redukce dat před klasifikací



Pro klasifikaci vhodnější 2. HK, přestože vyčerpává méně variability!

PCA – rozšiřující poznatky

- výpočet PCA, když je m >> K
- souvislost se singulárním rozkladem (SVD – Singular Value Decomposition)

Faktorová analýza (FA)

Faktorová analýza

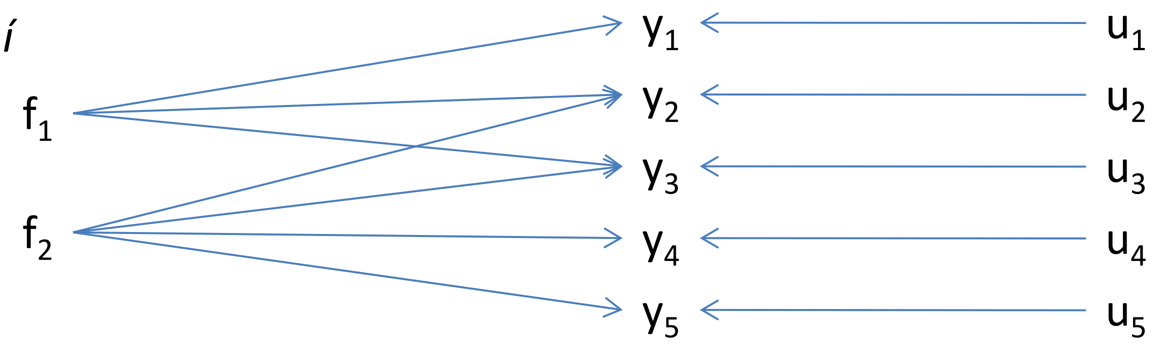
- faktorová analýza se snaží vysvětlit strukturu dat pomocí tzv. společných faktorů vysvětlujících sadu původních proměnných
- cíle, předpoklady, vstupní data a většina výpočtů obdobná jako u analýzy hlavních komponent
- čím se principielně liší od analýzy hlavních komponent?
  - Analýza hlavních komponent – vysvětlení maxima variability v datech
  - Faktorová analýza – vysvětlení maxima kovariance mezi popisnými proměnnými
- čím se prakticky liší od analýzy hlavních komponent?
  - Hlavním praktickým rozdílem je rotace proměnných tak, aby se vytvořené faktorové osy daly dobře interpretovat
  - Výhodou je lepší interpretace vztahu původních proměnných
  - Nevýhodou je prostor pro subjektivní názor analytika daný výběrem rotace
- typy faktorové analýzy
  - Vysvětlující (Explanatory) – snaží se identifikovat minimální počet faktorů pro vysvětlení dat
  - Potvrzující (Confirmatory) – testuje hypotézy ohledně skryté struktury v datech

Společné faktory a základní možné rotace

Společný faktor *Pozorovaná proměnná* *Unikátní faktor*

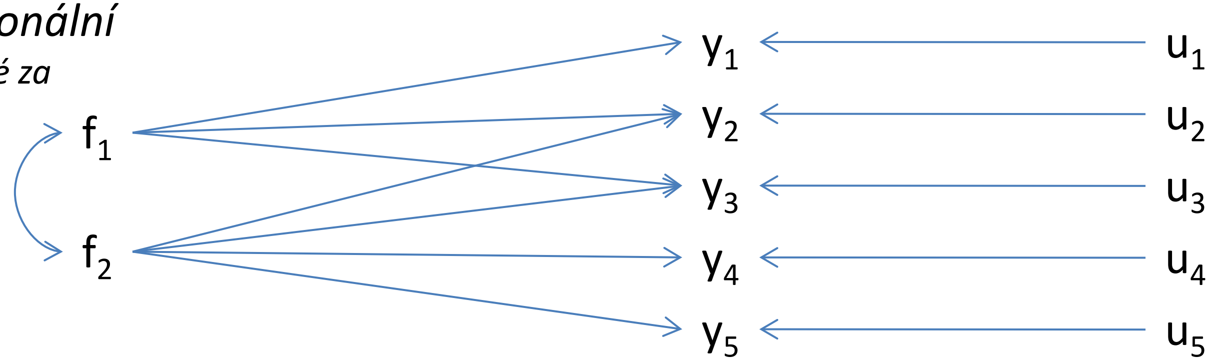
Rotace ortogonální

- Nezávislé faktory



Rotace neortogonální

- Faktory jsou závislé za účelem zvýšení intepretovatelnosti



Faktorová analýza – postup výpočtu

1. extrakce prvotních faktorů z kovarianční matice (analogie vlastních vektorů v PCA)
   - oproti PCA pracuje pouze s částí variability každé proměnné (tzv. communality), která je sdílena společnými faktory
   - několik možných algoritmů – principal factoring, metoda nejmenších čtverců, maximum likelihood apod.
   - výsledkem je komplexní struktura faktorů (obdobná PCA), kde řada faktorů má významné loadings (vztahy) k původním proměnným, počet takových faktorů je tzv. komplexita faktorů

2. v druhém kroku je rotací dosaženo zjednodušení struktury faktorů, tj. vztah mezi společnými faktory a původními proměnnými je zjednodušen (každá původní proměnná má hlavní vztah s jedním faktorem nebo malým počtem faktorů)
   - dva hlavní typy rotace:
     - ortogonální – faktory nemohou být korelovány, jsou tedy zcela nezávislé
     - neortogonální – faktory mohou být korelovány, nejsou tedy zcela nezávislé; vzhledem ke korelacím obtížnější interpretace

Faktorová analýza - rotace

- Ortogonální rotace
  - Quartimax – minimalizuje sumu čtverců loadings původních proměnných na faktorových osách, tedy zjednodušuje řádky matice loadings (=každá původní proměnná má největší loadings na jedné faktorové ose)
  - Varimax – zjednodušuje sloupce matice loadings
  - Equimax – zjednodušuje řádky i sloupce matice loadings
  - Biquartimax – varianta equimax

- Neortogonální rotace
  - Oblimax
  - Quartimin
  - Oblimin
  - Covarimin
  - Biquartimin
  - Atd.

Poděkování

Příprava výukových materiálů předmětu
„DSAN02 Pokročilé metody analýzy dat v neurovědách“
byla finančně podporována prostředky projektu FRMU
č. MUNI/FR/0260/2014 „Pokročilé metody analýzy dat
v neurovědách jako nový předmět na LF MU“