Metoda hlavních komponent a faktorová analýza

David Hampel

Ústav statistiky a operačního výzkumu,
Mendelova univerzita v Brně

Kurz pokročilých statistických metod
Global Change Research Centre AS CR, 5.–7. 8. 2015

Tato akce se koná v rámci projektu: Vybudování vědeckého týmu environmentální metabolomiky a ekofyziologie a jeho zapojení do mezinárodních sítí (ENVIMET; r.č. **CZ.1.07/2.3.00/20.0246**) realizovaného v rámci Operačního programu Vzdělávání pro konkurenceschopnost

Obsah

Motivace

- Vícerozměrná data se oproti jednorozměrným datům podstatně hůře hodnotí i prezentují.
- Při interpretaci zejména regresních charakteristik je vážnou překážkou korektního hodnocení a interpretace silná lineární závislost (multikolinearita) vysvětlujících proměnných.

Motivace

- Metoda hlavních komponent umožní provést potřebná hodnocení pomocích umělých proměnných, které vysvětlují stejnou variabilitu jako původní proměnné.
- Faktorová analýza umožní popsat problém pomocí menšího počtu umělých proměnných, než původních proměnných.
- Mezivýpočty obou uvedených metod mohou posloužit k prezentaci vícerozměrných dat.

Cíle metody hlavních komponent

- U mnoha výzkumných úloh se lze setkat se situací, kdy výchozí počet proměnných, sledovaných u zkoumaných jevů a procesů, je značný a pro interpretaci nepřehledný.
- Pro zjednodušení analýzy a snadnější hodnocení výsledků je často vhodné zkoumat, zda by studované vlastnosti pozorovaných objektů nebylo možné nahradit menším počtem jiných (třeba i umělých) proměnných s co nejmenší ztrátou informace.
- Metoda hlavních komponent (Principal Component Analysis, PCA) definuje nové (umělé, neměřitelné, latentní) proměnné (zde nazývané komponenty), které zcela vysvětlují původní variabilitu.

Cíle metody hlavních komponent

- Tyto nově vytvořené proměnné jsou lineární kombinací původních měřitelných proměnných a jejich hlavní vlastností je, že jsou navzájem nezávislé (nekorelované).
- Komponenty lze použít namísto původních proměnných při budování vícerozměrného regresního modelu při detekci multikolinearity.
- Výstupy PCA slouží ke grafické ilustraci vztahů jednak mezi jednotlivými proměnnými, jednak mezi popisovanými objekty.
- Analýza hlavních komponent je často využívána u vícerozměrných metod jako první krok při velkém počtu proměnných typicky s úkolem provést jejich redukci pomocí faktorové analýzy.

Ilustrativní popis metody hlavních komponent

- Základem je lineární transformace původních znaků na nové, nekorelované proměnné, zvané hlavní komponenty.
- Základní charakteristikou každé hlavní komponenty je její míra variability – rozptyl.
- Hlavní komponenty jsou seřazeny dle důležitosti – od největšího rozptylu k nejmenšímu rozptylu.
- Většina informace o variabilitě dat je soustředěna do první komponenty a nejméně informace je obsaženo v poslední komponentě.
- Má-li nějaký původní znak malý či dokonce žádný rozptyl, není schopen přispívat k rozlišení mezi objekty.

Ilustrativní popis metody hlavních komponent

- Pomocí PCA získáme informaci o skutečné dimenzi úlohy. Tato informace se uplatní při redukci dimenze původní úlohy bez velké ztráty informace v rámci faktorové analýzy.
- Nevyužité faktory (u navazující faktorové analýzy) obsahují malé množství informace, protože jejich rozptyl je příliš malý.
- Hlavní komponenty jsou nekorelované.
- První hlavní komponenta je například vhodným ukazatelem jakosti.
- První dvě resp. tři hlavní komponenty se využívají především pro zobrazení vícerozměrných dat v projekci do roviny resp. prostoru.

Podmínky nasazení metody hlavních komponent

- Mezi analyzovanými původními proměnnými musí být významné korelace.
- Faktorovou analýzu ani PCA nemá smysl použít, pokud jsou původní proměnné nekorelované.
- Faktorová analýza pak nemá co objasnit a PCA povede k hlavním komponentám totožným s původními proměnnými.

- Počet objektů musí být podstatně vyšší než počet proměnných (podobně jako u regrese).

Rozptyl vektoru náhodných veličin

Definujme p-rozměrný náhodný vektor

$$\mathbf{X} = (X_1, X_2, \ldots, X_3).$$

Jeho kovarianční matice je

$$\mathbf{V} = [cov(X_i, X_j)], \quad i, j = 1, 2, \ldots, p.$$

Tato matice je symetrická a tzv. pozitivně semidefinitní. Z toho plyne, že její vlastní čísla jsou reálná a nezáporná. Můžeme je označit a uspořádat následujícím způsobem:

$$\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p \geq 0.$$

Rozptyl vektoru náhodných veličin

Matici $\mathbf{V}$ můžeme vyjádřit jako

$$\mathbf{V} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T,$$

kde

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix}$$

a matice $\mathbf{U}$ je matice vlastních vektorů matice $\mathbf{V}$ (k-tý sloupec matice $\mathbf{U}$ je vlastní vektor příslušný k vlastnímu číslu λ_k a značíme jej $\mathbf{v}_k$).

Rozptyl vektoru náhodných veličin

- Pomocí vlastních vektorů a původního vektoru $\mathbf{X}$ můžeme definovat náhodné veličiny

$$Y_1 = \mathbf{X}\mathbf{v}_1, \quad \ldots, \quad Y_p = \mathbf{X}\mathbf{v}_p,$$

  které budeme nazývat první až p-tá *hlavní komponenta*.

- Tyto hlavní komponenty jsou nezávislé (nekorelované) a jejich rozptyl je postupně

$$\lambda_1, \ldots, \lambda_p.$$

Rozptyl vektoru náhodných veličin

- Celkový rozptyl vektoru $\mathbf{X}$ budeme definovat jako

$$\sigma^2 = var(X_1) + var(X_2) + \cdots + var(X_p),$$

  což je vlastně součet diagonálních prvků kovarianční matice $\mathbf{V}$.

- Platí, že

$$\sigma^2 = var(Y_1) + var(Y_2) + \cdots + var(Y_p) = \sum_{i=1}^{p} \lambda_i.$$

Rozptyl vektoru náhodných veličin

- Dále můžeme definovat relativní podíl celkové variability vysvětlený i-tou hlavní komponentou jako

$$\lambda_i/\sigma^2.$$

- Prvních k hlavních komponent vysvětluje

$$\sum_{i=1}^{k} \lambda_i/\sigma^2$$

celkové variability.

Rozptyl vektoru náhodných veličin

- Pokud chceme potlačit vliv jednotek, ve kterých měříme jednotlivé náhodné veličiny, použijeme namísto kovarianční matice vektoru $\mathbf{X}$ jeho korelační matici.
- Celkovou variabilitu pak dostaneme jako

$$\sigma^2 = \sum_{i=1}^{p} \lambda_i = p.$$

- Tato celková variabilita odpovídá variabilitě *po složkách standardizovaného* vektoru $\mathbf{X}$, nikoliv původního vektoru $\mathbf{X}$.

Praktické nalezení hlavních komponent

- Vyjdeme ze souboru dat, kde jsme u celkově n objektů neboli případů (rostlin, vzorků půdy, států) měřili p charakteristik (u rostliny např. výška, počet listů, hmotnost tisíce semen).

- Formálně můžeme psát

$$\begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}.$$

Praktické nalezení hlavních komponent

- Pokud budeme používat kovarianční matici data po sloupcích centrujeme: od každé hodnoty konkrétního sloupce odečteme průměr sloupce.
- Pokud budeme používat korelační matici data po sloupcích standardizujeme: od každé hodnoty konkrétního sloupce odečteme průměr sloupce a potom podělíme odchylkou hodnot sloupce.
- Dále budeme pracovat s upravenými daty, značení však necháme původní!

Praktické nalezení hlavních komponent

- Spočítáme výběrové rozptyly a kovariance sloupců a zformujeme tak odhad kovarianční matice.
- Pokud budeme požadovat korelační matici, pak spočítáme výběrové korelace sloupců a získáme tak odhad korelační matice.
- Pro odhad kovarianční či korelační matice získáme vlastní čísla a vlastní vektory.

Praktické nalezení hlavních komponent

- První hlavní komponentu získáme pomocí vlastního vektoru

$$\mathbf{v}_1 = (v_{11}, v_{12}, \ldots, v_{1p})^T$$

jako

$$\begin{bmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n1} \end{bmatrix} = v_{11} \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix} + v_{12} \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{bmatrix} + \cdots + v_{1p} \begin{bmatrix} x_{1p} \\ x_{2p} \\ \vdots \\ x_{np} \end{bmatrix}$$

Praktické nalezení hlavních komponent

- Druhou hlavní komponentu získáme pomocí vlastního vektoru

$$\mathbf{v}_2 = (v_{21}, v_{22}, \ldots, v_{2p})^T$$

jako

$$\begin{bmatrix} y_{12} \\ y_{22} \\ \vdots \\ y_{n2} \end{bmatrix} = v_{21} \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix} + v_{22} \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{bmatrix} + \cdots + v_{2p} \begin{bmatrix} x_{1p} \\ x_{2p} \\ \vdots \\ x_{np} \end{bmatrix}$$

Praktické nalezení hlavních komponent

- Poslední, p-tou hlavní komponentu získáme pomocí vlastního vektoru

$$\mathbf{v}_p = (v_{p1}, v_{p2}, \ldots, v_{pp})^T$$

jako

$$\begin{bmatrix} y_{1p} \\ y_{2p} \\ \vdots \\ y_{np} \end{bmatrix} = v_{p1} \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix} + v_{p2} \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{bmatrix} + \cdots + v_{pp} \begin{bmatrix} x_{1p} \\ x_{2p} \\ \vdots \\ x_{np} \end{bmatrix}$$

Faktorová analýza

- Faktorová analýza je metoda zaměřená na vytváření nových proměnných a na snížení dimenze dat s co nejmenší ztrátou informace.
- Nové proměnné jsou latentní, skryté, nepřímo pozorovatelné.
- Jedním ze základních cílů faktorové analýzy je posoudit strukturu vztahů sledovaných proměnných a zjistit tak, zda dovoluje jejich rozdělení do skupin.

Faktorová analýza

- Tyto skupiny dáváme do relace s tzv. faktory, které by měly umožnit lepší pochopení vstupních proměnných.
- Povaha faktorové analýzy je spíše heuristická a průzkumná (explorativní) než ověřovací (konfirmační).
- Faktorová analýza je často kritizovaná. Pochybnosti se týkají nejednoznačnosti řešení v důsledku subjektivity mnoha kroků i cílů, přibližnosti výsledků a mlhavé interpretace.

Faktorová analýza

Princip faktorové analýzy lze přiblížit pomocí vyjádření

$$x_{ij} = a_{j1} f_{i1} + a_{j2} f_{i2} + \cdots + a_{jm} f_{im} + e_{ij}, \quad m < p,$$

kde naměřenou a normalizovanou či standardizovanou hodnotu j-té charakteristiky pro i-tý objekt vysvětlujeme jako vážený součet m faktorů a reziduální složky e_{ij}, která těmito faktory vysvětlit nelze.

Faktorová analýza

V tomto kontextu hovoříme o matici

- $A = [a_{jk}]$, $j = 1, \ldots, p$, $k = 1, \ldots, m$ jako o matici *faktorových zátěží* (též se používají výrazy sycení, saturace, loadings);
- $F = [f_{ik}]$, $i = 1, \ldots, n$, $k = 1, \ldots, m$ jako o matici *faktorů* (někdy též faktorových skórů, označení je nejednotné);
- $E = [e_{ij}]$, $i = 1, \ldots, n$, $j = 1, \ldots, p$ jako o matici *reziduí*.

Faktorová analýza

- Požadujeme, aby faktory (sloupce matice F) byly tzv. ortonormální (nezávislé a jednotkové délky). Pak bude platit $F^T F = I$, kde I je jednotková matice.
- Dále definujeme matici $U = E^T E$. U je diagonální matice a na diagonále má hodnoty $u_1, u_2, \ldots, u_p$.
  - Platí, že
$$u_j = \sum_{i=1}^{n} e_{ij}^2, \quad j = 1, \ldots, p.$$
  - Jedná se o variabilitu, kterou nelze vysvětlit faktory, a označujeme ji jako *specificitu* či *specifický faktor*.

Faktorová analýza

- Pomocí specificit definujeme tzv. *komunality*

$$h_j = 1 - u_j = \sum_{k=1}^{m} a_{jk}^2, \quad j = 1, \ldots, p,$$

které můžeme interpretovat jako variabilitu vysvětlenou faktory.

Faktorová analýza

Hledají se takové faktorové zátěže a faktory, aby

- specificita byla co nejmenší;
- faktorové zátěže byly v absolutní hodnotě blízké buď nule nebo jedničce (pro dobrou identifikovatelnost faktorů s původními charakteristikami);
- počet faktorů byl co nejmenší (prakticky podstatně menší než počet původních charakteristik).

Tyto požadavky jsou ve vzájemném rozporu a algoritmy pro faktorovou analýzu hledají kompromisní řešení.

Faktorová analýza

K dosažení použitelných výsledků se uplatňují následující postupy:

- Určení počtu faktorů (někdy se hovoří o extrakci faktorů) – viz dále.
- Požaduje se takové nastavení, aby komunality splňovaly nerovnost $R_j^2 \leq h_j \leq 1$, $j = 1, \ldots, p$, kde R_j^2 je koeficient determinace lineárního regresního modelu, kde vysvětlujeme j-tou charakteristiku pomocí ostatních charakteristik.
- Používá se tzv. rotace faktorů, která se snaží docílit buď v absolutní hodnotě velkých hodnot nebo téměř nulových hodnot faktorových zátěží.

Faktorová analýza

Počet faktorů lze stanovit dle následujících kritérií:

- Počet vlastních čísel korelační matice větších než 1 či jiná vhodná konstanta.
- Použijeme tolik komponent, které vysvětlují určité procento (např. 90 %) původní variability.
- Použijeme Scree (též sutinový) graf (je vytvořen sestupně z vlastních čísel komponent), kde hledáme bod zlomu od rychlého klesání k pozvolnému.

Rotace faktorů

- Rotace faktorů spočívá v nalezení transformace matice A na matici B s jednodušší, jednoznačnější strukturou.
- Prakticky jde o nalezení tzv. ortogonální matice T, pomocí níž transformaci provedeme: $B = AT$.
- Takových transformací existuje nekonečně mnoho a je otázkou, kterou budeme považovat za optimální.
- Existuje více přístupů, které mohou dávat výrazně jiné výsledky.

Rotace faktorů

- Nejběžnější metodou rotace je tzv. *varimax*, která je založena na maximalizaci výrazu

$$\frac{1}{p}\sum_{j=1}^{m}\sum_{i=1}^{p}(a_{ij}^2 - a_{\cdot j}^2)^2,$$

kde $a_{\cdot j}^2 = \frac{1}{p}\sum_{i=1}^{p} a_{ij}^2$, tedy průměr kvadrátů faktorových zátěží pro j-tý faktor.

Rotace faktorů

- Další používanou rotací je *quartimax*, kde kritériem je funkce, která je součtem čtvrtých mocnin faktorových zátěží.
  - Metoda quartimax produkuje obecný faktor, protože na rozdíl od metody varimax je rozptyl je počítán přes celou matici a nikoli postupně pro všechny sloupce.
  - Zátěže zbývajících faktorů pak bývají nižší než při použití metody varimax.

Porovnání faktorové analýzy a analýzy hlavních komponent

- Obě metody se snaží o vyjádření původních proměnných pomocí latentních proměnných.
- Od latentních proměnných se v obou metodách požaduje, aby maximálně reprezentovaly (vysvětlovaly) původní proměnné.
- Konkretizace tohoto požadavku je v obou metodách odlišná:
  - V metodě PCA latentní proměnné (komponenty) vysvětlují maximum celkového rozptylu původních proměnných.
  - V metodě FA latentní proměnné (faktory) vysvětlují především vzájemné souvislosti mezi pozorovanými proměnnými.

Porovnání faktorové analýzy a analýzy hlavních komponent

- Faktorová analýza pracuje podobně jako PCA s korelační nebo kovarianční maticí a nalézá první hlavní faktor tak, aby vysvětloval největší část rozptylu datové matice.
- Další faktory jsou konstruovány takovým způsobem, aby byly nezávislé, čili nekorelované, a vyčerpávaly sestupně maximum celkového rozptylu.
- Faktorová analýza se pokouší objasnit kovariance a korelace původních proměnných pomocí několika málo společných faktorů, zatímco PCA objasňuje pouze rozptyl původních proměnných.

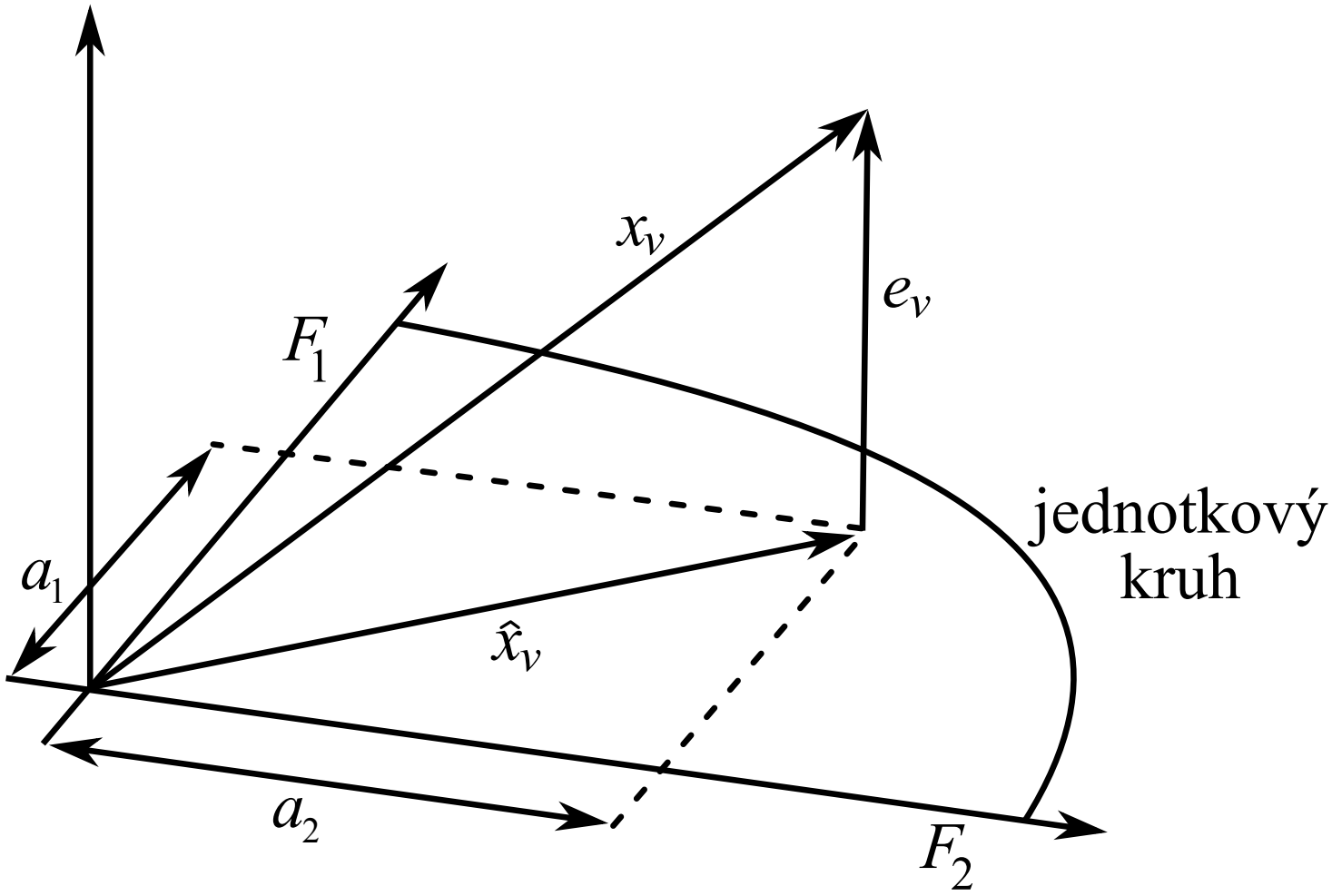
- Výpočet u PCA je přímočarý, jednoduchý.
- U faktorové analýzy je výpočet faktorového skóre daleko komplexnější a byla pro něj navržena řada postupů.

Porovnání faktorové analýzy a analýzy hlavních komponent

- Rozdíl mezi faktorovou analýzou a analýzou hlavních komponent je i v posledním kroku analýzy.
- U faktorové analýzy jsou faktory rotovány tak, aby co nejjednodušeji popisovaly proměnné.
- Ve srovnání s metodou hlavních komponent hledá faktorová analýza vzájemné souvislosti vstupních proměnných.

Faktorová analýza – interpretace

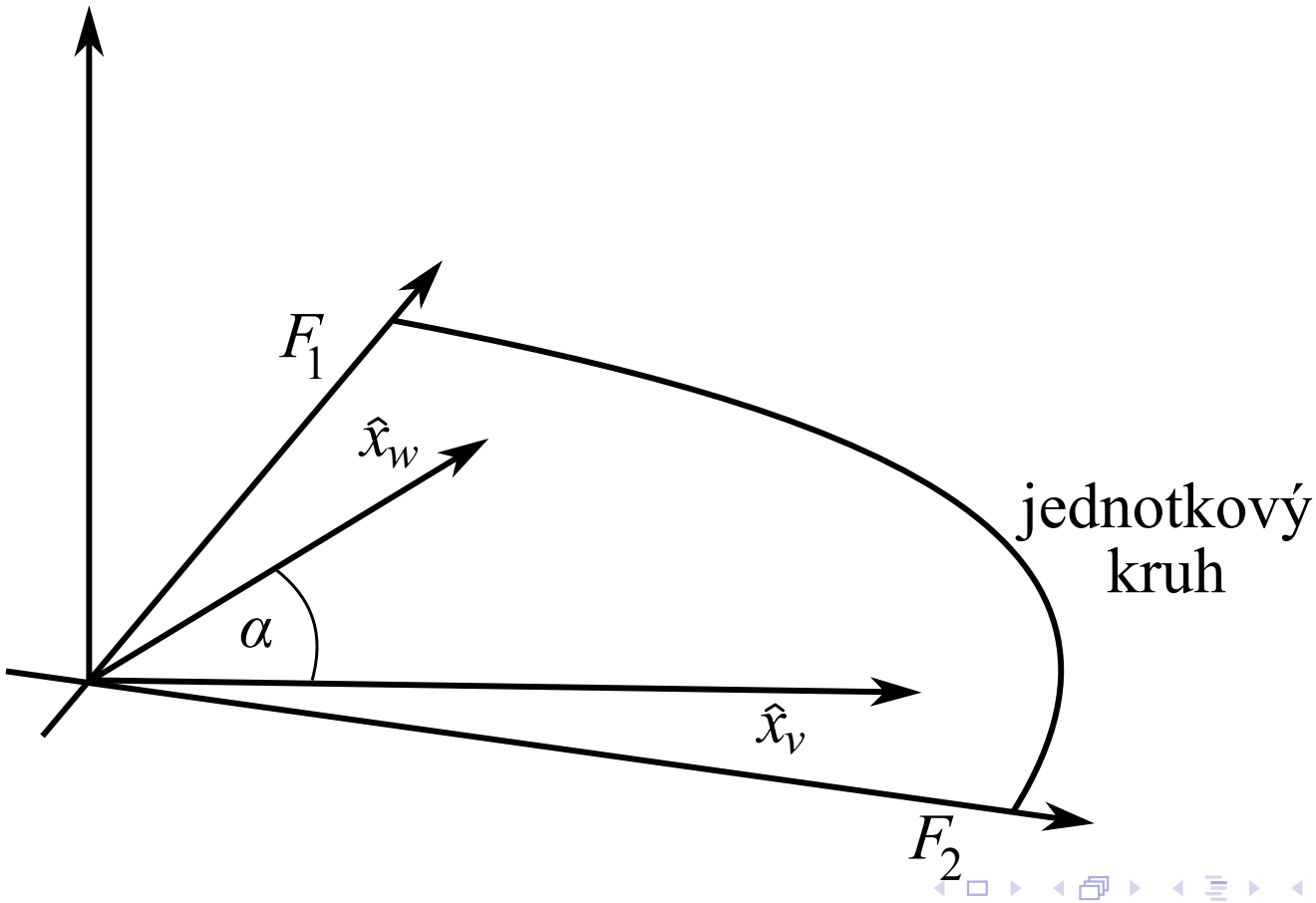
Charakteristika v popsaná pomocí prvních dvou faktorů.



Faktorová analýza – interpretace

Charakteristiky v a w popsané pomocí prvních dvou faktorů.
Platí, že korelace $r_{vw} = \cos(\alpha)$.

$$\cos(0^\circ) = 1, \quad \cos(90^\circ) = \cos(270^\circ) = 0, \quad \cos(180^\circ) = -1$$



Faktorová analýza – graf komponentního skóre

- Souřadnice každého objektu na osách hlavních komponent nazýváme skóre (komponentní skóre).
- Graf komponentního skóre je zobrazení dvou skórových vektorů vynesených v systému kartézských os jeden proti druhému.
- Na x-ové ose obvykle použijeme první komponentu popř. faktor a na ose y další komponenty či faktory.

Faktorová analýza – graf komponentního skóre

- Umístění objektů: daleko od počátku jsou extrémy. Objekty nejblíže počátku jsou nejtypičtější.
- Podobnost objektů: objekty blízko sebe si jsou podobné, objekty daleko od sebe jsou si nepodobné.
- Objekty v shluku: umístěné zřetelně v jednom shluku jsou si podobné a nepodobné objektům v ostatních shlucích. Jsou-li shluky blízko sebe, znamená to značnou podobnost objektů.
- Osamělé objekty: izolované objekty mohou být odlehlé.

Faktorová analýza – graf komponentních zátěží

- Můžeme identifikovat podobné (korelované) charakteristiky či jejich shluky.
- Je vidět relace mezi původními charakteristikami a komponentami (faktory).
- Pokud jsou komponenty či faktory „pojmenovány“ (v biologickém, chemickém, ekonomickém či jiném kontextu), lze usuzovat na příslušnost charakteristik k této věcné kategorii.

Faktorová analýza – biplot

- Jedná se vlastně o současné vykreslení grafu komponentního skóre a komponentních zátěží.
- Pokud je objekt umístěn v blízkosti určité charakteristiky, lze říci, že tento objekt „obsahuje“ velké množství této charakteristiky.
- Můžeme hovořit o interakci ve smyslu vícefaktorové analýzy rozptylu.
- Biplot je třeba interpretovat velmi opatrně.

Faktorová analýza – příklady

- `Desetiboj.sta`
- `data_PCA.sta`