

Podrobnější studium obou uvedených metod by ukázalo, že faktorová analýza je zobecněním analýzy komponentní. V dalším výkladu se omezíme na analýzu komponentní. Zobecnění je jednou z nabízejících se možností pro vás, ambiciózní studenty kurzu.

12.3 Analýza hlavních komponent – PCA

Metoda analýzy hlavních komponent patří k nejstarším metodám vícerozměrné statistické analýzy. Ve skutečnosti, jak jsme už naznačili výše, je součástí většiny ostatních metod analýzy multidimenzionálních pozorování. Jako metodologie byla zavedena K. Pearsonem (1901), původně jako popisná statistická metoda podporující redukci dimenze multimenzionálních dat. Pearson definoval „rovinu nejtěsnější shody“ jako podprostor minimalizující součet čtverců vzdáleností od všech datových položek. Samotný pojem „hlavní komponenta“ použil poprvé Hotelling (1933).

Nechť \vec{X} je p -rozměrný náhodný vektor s kovarianční maticí $C = (\sigma_{ij})$. Řešíme úlohu nahrazení proměnných x_1, \dots, x_p nějakým menším počtem nových náhodných veličin z_i . Těmto veličinám se říká hlavní komponenty vektoru \vec{X} , přitom z_i je i -tá hlavní komponenta.

Analýza hlavních komponent PCA

- Statistická metoda vyhledávající ve vícerozměrných datech skryté závislosti, za účelem získat kompaktnější popis těchto dat.
- Provádí redukci dat z dimenze m do dimenze n ($m > n$).

Geometrické vlastnosti hlavních komponent

Každá proměnná původního vektoru \vec{X} představuje souřadnicovou osu procházející počátkem. Těchto p os tvoří p -rozměrný prostor, v němž je každý vektor \vec{X} určen p -souřadnicemi. Analýza hlavních komponent hledá takové natočení těchto os, kdy první nová osa (proměnná z_1) prochází směrem maximálního rozptylu shluku původních vektorů. Promítá se tedy na ní maximum informace o struktuře rozložení původních dat. Druhá nová osa, představující proměnnou z_2 , je k první ortogonální a je opět orientována ve směru maximálního rozptylu. Totéž analogicky platí pro další osy.

Analýza hlavních komponent není invariantní vůči změnám měřítek analyzovaných proměnných. Je tedy zapotřebí, aby byly všechny složky vektoru \vec{X} měřeny ve stejných jednotkách.

Aniž bychom zacházeli do přílišných podrobností (zájemce odkazujeme na učebnice Matematické statistiky), uvedeme následující fakta, resp. jejich důsledky:

1. Definice: Charakteristickým polynomem (mnohočlenem) matice C řádu n nazýváme determinant matice $\lambda E - C$. Jeho kořeny jsou právě **vlastní** (charakteristická) čísla matice C .

Z této definice lze vyvodit vztah pro výpočet vlastních čísel v podobě charakteristického polynomu:

$$0 = \det(\lambda E - C) = (-1)^n \lambda^n + b_1 \lambda^{n-1} + \dots + b_{n-1} \lambda + b_n.$$

2. Právě vlastní čísla charakteristického polynomu (viz výše) jsou hledanými hlavními komponentami.
3. Vlastní čísla takto určují tzv. **vlastní** (charakteristické) **vektory**: vlastním vektorem matice C a její vlastní číslo λ je vektor α vyhovující rovnicím

$$(C - \lambda E)\alpha = 0, \quad \alpha^T \alpha = 1.$$

4. Klasické prostředky, resp. postupy pro výpočet těchto vlastních čísel:

„Ručně“ podle definice. Tento postup je snadný, ale jen pro malá n ($n = 2, 3$). Pro velká n je však výpočetní složitost příliš vysoká.

Metodou LU-rozkladu. Nazývá se též *LR-transformací* či *LR-algoritmem*. Metoda je iterační, využívá vlastností podobných matic. Má ale následující nedostatky: obecně pomalu konverguje, pro matice vyšších řádů je počet operací vysoký a pro obecnou matici může být tento algoritmus nestabilní.

Metodou ortogonálních transformací. Opět iterační algoritmus, opět podobné matice. V některých případech vede k *Jacobiho* metodě.