Autor: RNDr. Eva Janoušová
Institut biostatistiky a analýz, Masarykova univerzita, Brno

Analýza hlavních komponent – příklad

Bylo provedeno měření objemu šedé hmoty (v cm^3) a objemu likvoru (v cm^3) u pěti dětí. Naměřené hodnoty byly zaznamenány do matice :

U tohoto datového souboru proveďte analýzu hlavních komponent.

Řešení:

U analýzy hlavních komponent potřebujeme nejprve spočítat kovarianční matici . Pro výpočet kovarianční matice potřebujeme znát průměrný objem šedé hmoty a likvoru u dětí:

Jednotlivé prvky kovarianční matice poté spočítáme následujícím způsobem:

Rozptyl objemu šedé hmoty:

Rozptyl objemu likvoru:

Kovariance objemu šedé hmoty a objemu likvoru:

Kovarianční matice je tedy: .

Nyní spočítáme vlastní čísla a vlastní vektory kovarianční matice – tzn., spočítáme následující determinant:

Vypočteme charakteristický polynom:

A jeho kořeny, které odpovídají vlastním číslům:

Následně spočítáme vlastní vektor odpovídající prvnímu vlastnímu číslu :

; → ; např. pro pak dostáváme: , který je po normalizaci roven . Kontrola, že vektor má jednotkovou délku: .

Spočítáme vlastní vektor odpovídající druhému vlastnímu číslu :

; → ; např. pro pak dostáváme: , který je po normalizaci roven . Kontrola, že vektor má jednotkovou délku: .

Vlastní vektory můžeme uspořádat do matice , přičemž pořadí vlastních vektorů odpovídá pořadí vlastních čísel seřazených od největšího k nejmenšímu.

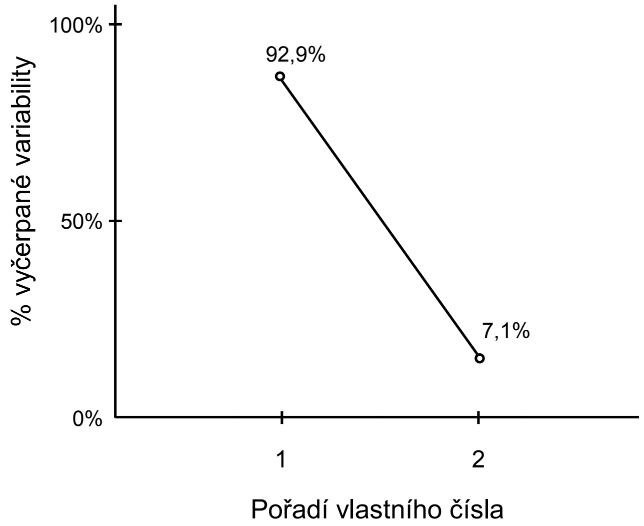
Nyní vyjádříme hlavní komponenty odpovídající vlastním číslům seřazeným od největšího k nejmenšímu – hlavní komponenty jsou lineární kombinace původních proměnných, přičemž koeficienty jsou souřadnice příslušného vlastního vektoru:

1. hlavní komponenta: (pro)
2. hlavní komponenta: (pro)

Výpočet procent vyčerpané variability:

1. hlavní komponenta vyčerpává: (tzn., 92,93% variability v datech)
2. hlavní komponenta vyčerpává: (tzn., 7,07% variability v datech)

Vyčerpanou variabilitu můžeme znázornit i pomocí sutinového grafu:



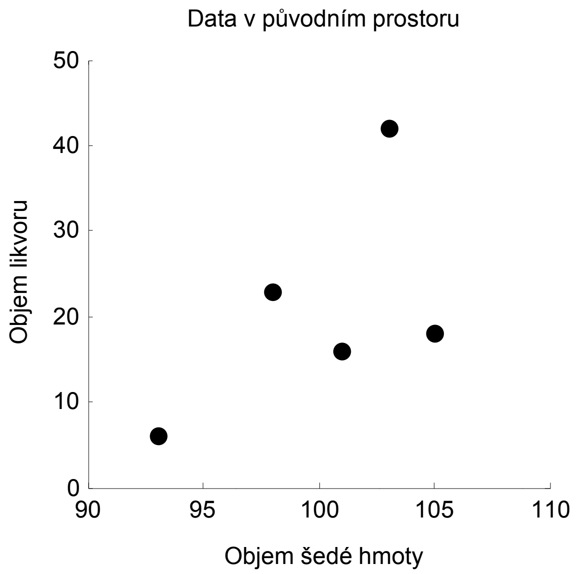
Dále spočítáme korelace hlavních komponent s původními proměnnými:

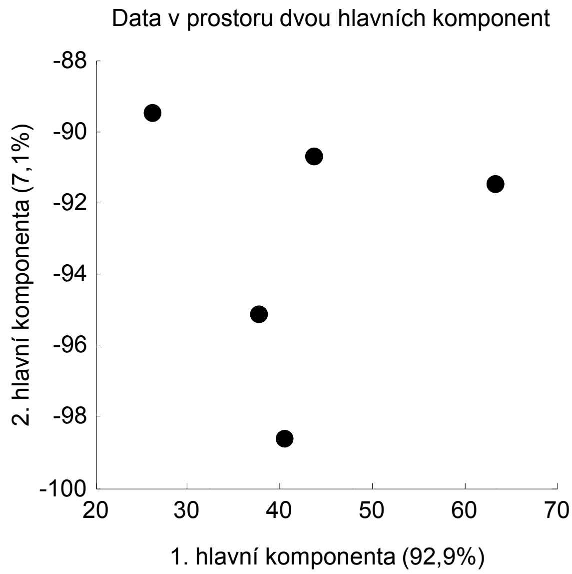
První hlavní komponenta je vysoce korelována s objemem likvoru a středně korelována s objemem šedé hmoty. Druhá hlavní komponenta je středně záporně korelována s objemem šedé hmoty.

Na závěr vypočítáme nové souřadnice původních bodů po transformaci pomocí obou hlavních komponent spočítaných pomocí PCA:

Souřadnice subjektů můžeme přímo získat i z hlavních komponent – např. pro první subjekt:

Původní data i data po transformaci pomocí PCA si znázorníme:





Pokud bychom k transformaci použili pouze první vlastní vektor, získáváme data v prostoru první hlavní komponenty: