4.9 Úlohy

Při vyšetřování jednotlivých úloh je třeba postupovat dle následujícího postupu:

Postup analýzy vícerozměrných dat

1. *Standardizace:* vícerozměrné analýze obvykle předchází standardizace čili škálování proměnných.

2. *Odhady parametrů polohy, rozptýlení, tvaru a intenzita vztahu mezi proměnnými:*
vyčíslení výběrové střední hodnoty každé proměnné, odhad kovarianční matice $\boldsymbol{S}$ a její normované podoby - korelační matice $\boldsymbol{R}$, odhadu vícerozměrné šikmosti $g_{1,m}$ a vícerozměrné špičatosti $g_{2,m}$. Matice $\boldsymbol{R}$ obsahuje Pearsonovy párové korelační koeficienty ρ_{ij}, které se diskutují. Užitečný je především diagram korelační matice.

3. *Exploratorní analýza dat EDA:* (a) posoudí podobnost objektů pomocí vizuálních rozptylových diagramů typu casement plot, draftsman plot, dále symbolových a profilových grafů (hvězdičky, sluníčka, obličeje, křivky, stromy), (b) nalezne vybočující objekty nebo vybočující proměnné, mnohdy k nevhodné analýze, (c) stanoví, zda platí předpoklad lineárních vazeb, (d) testuje všechny předpoklady o datech (normalitu, nekorelovanost, homogenitu). Ověřování normality je založeno na vícerozměrné šikmosti $g_{1,m}$ a vícerozměrné špičatosti $g_{2,m}$, kdy se testuje simultánní platnost nulových hypotéz H_{01}: $g_{1,m} = 0$ a H_{02}: $g_{2,m} = m(m + 2)$.

4. *Určení vhodného počtu latentních proměnných:* matice $\boldsymbol{S}$ nebo $\boldsymbol{R}$ se rozloží na vlastní čísla λ_i a vlastní vektory $\boldsymbol{v}_i$. Z indexového grafu úpatí vlastních čísel (Scree plot) se určí vhodný počet latentních proměnných (pro zobrazení v rovině se obvykle dává přednost prvním dvěma latentním proměnným), které ještě dostatečně popisují proměnlivost v datech. Když se latentní proměnné podaří pojmenovat a dát jim i fyzikální, biologický či jiný věcný význam, jedná se o faktory. V opačném případě jde o hlavní komponenty.

5. *Určení struktury v proměnných (PCA a FA):* hledání struktury a vzájemných vazeb (korelace) proměnných se provede v grafu komponentních vah (Plot of components weights, loadings). Hledání struktury v objektech a třídění objektů do shluků se provede v rozptylovém diagramu komponentního skóre (Plot of principal components). Dvojný graf (Biplot) je přehledným spojením obou předešlých grafů a ukáže interakci objektů a proměnných.

6. *Určení struktury a vzájemných vazeb v objektech:* klasifikační postupy zařadí v diskriminační analýze analyzovaný objekt do jednoho již existujícího a předem zadaného shluku. Neutříděnou skupinu objektů lze uspořádat do shluků a výsledek třídění zobrazit dendrogramem v analýze shluků. V hierarchickém postupu je třeba k vytvoření shluků vybrat vzdálenost mezi objekty (Eukleidovskou, Manhattanovskou, Mahalanobisovu) a jednu z nabídnutých metod: průměrovou, centroidní, nejbližšího souseda, nejvzdálenějšího souseda, mediánovou, Wardovu. Nehierarchické postupy rozdělí objekty do shluků, v nichž jsou předem umístěni typičtí reprezentanti.

7. *Soulad nalezené struktury objektů a vzájemných vazeb v dendrogramu a PCA (či FA) grafech:* je třeba vyšetřit a komentovat nalezenou strukturu a vazby *jednotlivých proměnných,* nalezenou jednak v PCA (či FA) a jednak v dendrogramu podobnosti proměnných analýzou vzniklých shluků. Dále je třeba komentovat také strukturu a vazby *klasifikovaných objektů,* nalezenou v PCA a v dendrogramu podobnosti objektů.

Využitím modulu Vícerozměrná data programového systému ADSTAT, resp. programu STATGRAPHICS, SCAN, MINITAB, STATISTICA, S-Plus atd. je třeba analyzovat dále uvedené úlohy. Úlohy jsou rozděleny do pěti kapitol: B4 (farmakologická a biochemická data), C4 (chemická a fyzikální data), E4 (environ-mentální, potravinářská a zemědělská data), H4 (hutní a mineralogická data) a S4 (ekonomická a sociologická data).