4.5 Určení struktury a vazeb v proměnných a objektech

Zdrojová matice má rozměr $n \times m$. Před vlastní aplikací vhodné metody vícerozměrné statistické analýzy je třeba vždy provést *exploratorní (průzkumovou) analýzu dat*, která umožňuje

(a) posoudit *podobnost objektů* pomocí rozptylových a symbolových grafů,
(b) nalézt *vybočující objekty*, resp. jejich proměnné,
(c) stanovit, zda lze použít předpoklad *lineárních vazeb,*
(d) ověřit *předpoklady o datech* (normalitu, nekorelovanost, homogenitu).

Jednotlivé techniky k určení vzájemných vazeb se dále dělí podle toho, zda se hledají
(a) *struktura a vazby v proměnných* nebo
(b) *struktura a vazby v objektech*:

(1) Hledání struktury *v proměnných* v metrické škále: *faktorová analýza FA*, *analýza hlavních komponent PCA* a *shluková analýza*.
(2) Hledání struktury *v objektech* v metrické škále: *shluková analýza*.
(3) Hledání struktury *v objektech* v metrické i v nemetrické škále: *vícerozměrné škálování.*
(4) Hledání struktury *v objektech* v nemetrické škále: *korespondenční analýza.*
(5) Většina metod vícerozměrné statistické analýzy umožňuje *zpracování lineárních vícerozměrných modelů*, kde závisle proměnné se uvažují jako lineární kombinace nezávisle proměnných, resp. vazby mezi proměnnými jsou lineární. V řadě případů se také uvažuje normalita metrických proměnných.

Určením struktury a vzájemných vazeb mezi proměnnými ale i mezi objekty se zabývají techniky redukce proměnných na latentní proměnné, metoda *analýzy hlavních komponent (PCA)* a *metoda faktorové analýzy (FA).* Důležitou metodou určení vzájemných vazeb mezi proměnnými je i *kanonická korelační analýza CA*, která se používá ke zkoumání závislosti mezi dvěma skupinami proměnných, přičemž jedna ze skupin se považuje za proměnné nezávislé a druhá za skupinu proměnných závislých.

4.5.1 Analýza hlavních komponent (PCA)

Cílem metody je transformace dat z původních proměnných $x_j, j=1, ..., m$, do menšího počtu latentních proměnných y_j. Tyto proměnné mají vhodnější vlastnosti, je jich výrazně méně, vystihují téměř celou *proměnlivost* původních proměnných a jsou vzájemně nekorelované (korelační koeficient mezi latentními proměnnými $y_1 , ..., y_m$ je 0). Latentní proměnné jsou u této metody nazvány *hlavními komponentami* a jsou to lineární kombinace původních proměnných: *první hlavní komponenta* y_1 popisuje největší část proměnlivosti čili rozptylu původních dat, *druhá hlavní komponenta* y_2 zase největší část rozptylu neobsaženého v y_1 atd. Matematicky řečeno, *první hlavní komponenta* je takovou lineární kombinací vstupních proměnných, která zahrnuje největší proměnlivost mezi všemi lineárními kombinacemi. Má tvar

$$y_1 = \sum_{j=1}^{m} v_{1j} x_j = \boldsymbol{v_1}^{\mathbf{T}} \boldsymbol{x},$$

kde objekt $\boldsymbol{x}$ obsahuje proměnné $x_1, ..., x_m$. Pro vektor koeficientů $\boldsymbol{v}_1^{\mathbf{T}} = (v_{11} , ..., v_{1m})^{\mathrm{T}}$ platí, že proměnlivost vyjádřená rozptylem $D(y_1) = \boldsymbol{v_1}^{\mathbf{T}} \boldsymbol{S} \boldsymbol{v_1}$ je maximální, přičemž $\boldsymbol{S}$ značí kovarianční matici původních dat. Zcela analogicky jsou konstruovány další hlavní komponenty, jejichž celkový počet je roven menšímu ze dvou čísel, a to n (počet objektů) nebo m (počet proměnných). Protože platí, že součet rozptylů všech hlavních komponent je roven součtu rozptylů vstupujících původních proměnných, můžeme z podílu rozptylů jednotlivých hlavních komponent usuzovat na část proměnlivosti vysvětlenou dotyčnou hlavní komponentou. Jestliže součet prvních (nejvyšších) A podílů proměnlivosti je dostatečně blízký jedné, resp. 100 % (obvykle však stačí 80 % - 90 %), postačí brát v úvahu právě těchto prvních A hlavních komponent pro "dostatečné" vysvětlení variability původních proměnných. Rozdíl mezi souřadnicemi objektů v původních proměnných a v hlavních komponentách čili ztráta informace projekcí do menšího počtu rozměrů se nazývá *špatnou mírou těsností proložení* modelu PCA nebo také *chybou modelu PCA*. I při velkém počtu původních proměnných (m) může být A velmi malé, často 2 až 5. Volba počtu užitých komponent A představuje vlastní *model hlavních komponent PCA.* Vysvětlení užitých hlavních komponent, jejich pojmenování a vysvětlení vztahu původních proměnných $x_j , j = 1, ..., m,$ k hlavním komponentám $y_k , k = 1, ..., A,$ tvoří dominantní součásti zvoleného modelu hlavních komponent PCA.

Vlastní *matematický postup PCA* je následující: maximalizací při zavedení normalizační podmínky $\boldsymbol{v}_1^{\mathbf{T}} \boldsymbol{v} = 1$ vyjde, že

$$(\boldsymbol{S} - \lambda_1 \boldsymbol{I}) \boldsymbol{v_1} = \mathbf{0} ,$$

kde **0** označuje nulový vektor, λ_1 je největší *vlastní číslo* a $\boldsymbol{v}_1$ je odpovídající *vlastní vektor* kovarianční matice ***S*** ***a I*** je jednotková matice. Po dosazení vyjde

$$D(y_1) = \boldsymbol{v}_1^{\mathrm{T}} \boldsymbol{S} \boldsymbol{v}_1 = \lambda_1 .$$

Analogicky lze odvodit, že vektor koeficientů $\boldsymbol{v}_2$ ve vztahu $y_2 = \sum_{j=1}^{m} v_{2j} x_j$, maximalizující $D(y_2)$ za podmínky, že $\mathrm{cov}(y_1, y_2) = 0$, odpovídá vlastnímu vektoru, příslušejícímu druhému největšímu vlastnímu číslu λ_2.

Provedeme-li rozklad kovarianční matice ***S*** na vlastní čísla $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m$, jsou odpovídající vlastní vektory $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m$ přímo koeficienty hlavních komponent $y_1, \ldots, y_m$. Hlavní komponenty mají řadu zajímavých vlastností. Lze je interpretovat jako hlavní osy *m*-rozměrného elipsoidu $\boldsymbol{x}^{\mathrm{T}} \boldsymbol{S}^{-1} \boldsymbol{x} = \mathrm{konst}$.

K odstranění závislosti na jednotkách původních proměnných se lépe užívá standardizovaných proměnných $\boldsymbol{x}^*$ s prvky $x_j^* = (x_j - \bar{x}_j)/\sigma_j$. Pro *j*-tou hlavní komponentu pak platí

$$y_j^* = \sum_{k=1}^{m} v_{jk}^* x_k^* ,$$

kde $\boldsymbol{v}_j^*$ je vlastní vektor *korelační* matice ***R*** odpovídající *j*-tému největšímu vlastnímu číslu λ_j^*. Hlavní komponenty y_j^*, určené z korelační matice ***R***, jsou však hůře interpretovatelné. Platí, že $\boldsymbol{v}_j^{*\mathrm{T}} \boldsymbol{v}_j^* = \lambda_j$, nikoliv rovno jedné. Pro účely zobrazení vícerozměrných dat různého měřítka jsou však vhodnější standardizované y_j^* než původní y_j.

PCA umožňuje rozklad matice dat ***X*** na *strukturní část* $\boldsymbol{TP}^{\mathrm{T}}$ a *šumovou část* ***E*** dle vztahu $\boldsymbol{X} = \boldsymbol{TP}^{\mathrm{T}} + \boldsymbol{E}$,

kde ***T*** je *matice komponentního skóre* a ***P*** je *matice komponentních vah,* ***E*** je *matice reziduí*. Modelem hlavních komponent PCA nazýváme součin $\boldsymbol{T}\,\boldsymbol{P}^{\mathrm{T}}$. Matice reziduí ***E*** není součástí modelu, týká se modelem PCA nevysvětlované části dat ***X***. Je to část dat, která není zahrnuta v modelu $\boldsymbol{T}\,\boldsymbol{P}^{\mathrm{T}}$ a představuje *míru netěsnosti proložení* původních reálných dat modelem PCA. Model hlavních komponent PCA

$$\boldsymbol{X} = t_1 p_1^{\mathrm{T}} + t_2 p_2^{\mathrm{T}} + \ldots + t_A p_A^{\mathrm{T}} + \boldsymbol{E}$$

se vypočte postupem:

1. Vypočte se t_1 a p_1 z ***X*** a vyčíslí se $\boldsymbol{E_1} = \boldsymbol{X} - t_1 p_1^{\mathrm{T}}$.
2. Vypočte se t_2 a p_2 z $\boldsymbol{E_1}$ a vyčíslí se $\boldsymbol{E_2} = \boldsymbol{E_1} - t_2 p_2^{\mathrm{T}}$.
3. Vypočte se t_3 a p_3 z $\boldsymbol{E_2}$ a vyčíslí se $\boldsymbol{E_3} = \boldsymbol{E_2} - t_3 p_3^{\mathrm{T}}$ a pokračuje se tak dlouho, až se vyčíslí všech *A* komponent, $A = \min(n, m)$. Výhodou postupu je snížení počtu proměnných, snížení dimenzionality a dále rozdělení původní matice dat ***X*** na část strukturní $\boldsymbol{TP}^{\mathrm{T}}$ a část šumovou ***E***. Když bychom použili celý PCA model, pak je ***E*** rovno nule. Musíme ale hledat optimální počet využitelných hlavních komponent *A* tak, abychom dosáhli nejlepšího proložení a aby matice ***E*** byla téměř nulová a její absolutní velikost byla srovnatelná s experimentální chybou dat. To je ústřední myšlenka vícerozměrné analýzy dat: uživatel musí navrhnout počet hlavních komponent *A* tak, aby byla matice reziduí co nejmenší. Velká hodnota ***E*** znamená špatný model, malá hodnota ***E*** dobrý model. Termíny malý, velký, dobrý, špatný jsou však pouze kvalitativní. Vyhodnocení ***E*** je relativní k centrovanému počátku, tj. střednímu objektu, který má souřadnice v aritmetickém průměru každé proměnné a představuje počátek (0, 0, ..., 0). Je vhodné říkat, že jde o nulu hlavních komponent. První operace je totožná s centrováním proměnných původní matice dat ***X***. To znamená, že pro $A = 0$ bude reziduálová matice $\boldsymbol{E_0}$ totožná s centrovanou maticí ***X***. Zde $\boldsymbol{E_0}$ hraje důležitou roli jako referenční stav, ke kterému budeme přirovnávat velikost klesající hodnoty ***E***, takže pro $A = 0$ bude $\boldsymbol{E_0} = 100\ \%$ a člen $\boldsymbol{TP}^{\mathrm{T}} = 0\ \%$. Velikost ***E*** představuje chybový výraz, který vyčíslíme běžným statistickým způsobem. Buď vyjádříme rezidua jako rezidua objektů (v řádcích) nebo rezidua proměnných (ve sloupcích).

Rezidua objektů (v řádcích): rozptyl reziduí *i*-tého objektu se týká průměru středově centrovaných dat a je dán vztahem

$$e_i = \sqrt{\sum_{k=1}^{m} e_{ik}^2}$$

a odpovídá *vzdálenosti* mezi *i*-tým objektem a hyperplochou *A* hlavních komponent. Je to vzdálenost, která byla minimalizována nejmenšími čtverci, když se určovaly hlavní komponenty. Proto je rozptyl reziduí objektu mírou vzdálenosti mezi prostorem proměnných objektu a reprezentací objektu v prostoru hlavních kompo-nent. Platí přitom pravidlo: čím menší bude tato vzdálenost, tím těsnější bude reprezentace objektu v prostoru hlavních komponent (čili PCA model) vůči původnímu objektu. Řádky v matici ***E*** jsou nepřímo úměrné těsnosti proložení původních dat PCA modelem.

Celkový rozptyl reziduí objektu je suma rozptylu reziduí všech objektů e_{tot} dle vztahu

$$e_{tot} = \sqrt{\sum_{i=1}^{n} e_i^2} .$$

Graf reziduí jednotlivých objektů. Rozptyl reziduí jednoho *i*-tého objektů představu-je vzdálenost mezi objektem a modelem. Je vhodné zobrazovat tuto veličinu pro všechny objekty a odhalit tak odlehlé objekty či jiné anomálie.

Graf celkového rozptylu reziduí. Metodou hlavních komponent PCA se vyčíslí $\boldsymbol{E}_0$, $\boldsymbol{E}_1$, $\boldsymbol{E}_2$, ... a z nich $e_{tot,1}$, $e_{tot,2}$, ... a ty se vynesou do indexového grafu úpatí proti indexu *A*. Protože je snaha o pojmenování a vysvětlení hlavních komponent, je správný počet hlavních komponent *A* velmi důležitý: je-li počet *A* příliš malý, "podceněný model" způsobí povrchní popis datové struktury. Je-li počet *A* příliš velký, "přeceněný model" je ještě horší, protože zahrnuje do své struktury také část šumu. Existuje pravidlo, že velká proměnlivost v datech odpovídá hlavnímu jevu, například proměňované koncentraci, zatímco malá proměnlivost odpovídá spíše šumu. V analýze dat se snažíme odhalit stěžejní jev, o kterém předpokládáme, že dominuje v datech. V PCA platí pravidlo, že "velké" hlavní komponenty odpovídají nejdůležitější informaci, kterou vlastně hledáme, zatímco "malé" hlavní komponenty odpovídají spíše šumu a jsou pro strukturu dat obvykle nepodstatné. Malé hlavní komponenty mohou být proto zahrnuty do matice $\boldsymbol{E}$. Cílem PCA je odfiltrování šumu z dat a soustředění se na strukturní, bezšumovou část dat.

Graficky lze výsledek analýzy hlavních komponent zobrazit v několika grafech hlavních komponent následujícím způsobem:

(a) **Indexový graf úpatí vlastních čísel** (Scree Plot) je vlastně sloupcový diagram vlastních čísel nebo reziduálního rozptylu proti stoupající hodnotě indexu, pořadového čísla *A* (obr. 4.7). Zobrazuje relativní velikost jednotlivých vlastních čísel. Řada autorů ho s oblibou využívá k určení počtu *A* "užitečných" hlavních komponent. Cattel vysvětluje scree jako zlomové místo mezi kolmou stěnou a vodorovným dnem. Vybrané "užitečné" hlavní komponenty (nebo také faktory) pak tvoří kolmou stěnu a "neužitečné" hlavní komponenty (nebo faktory) představují vodorovné dno. Užitečné komponenty jsou tak odděleny zřetelným zlomovým místem a souřadnice *x* tohoto zlomu je hledaná hodnota indexu. Jiným, hrubším kritériem je pravidlo, podle kterého využíváme ty hlavní komponenty, jejichž vlastní číslo je větší než jedna. Graf úpatí se však jeví objektivnějším.

(b) **Graf komponentních vah, zátěží** (Plot Components Weights) zobrazí komponentní váhy pro první dvě hlavní komponenty (obr. 4.8). V tomto grafu se porovnávají vzdálenosti mezi proměnnými. Krátká vzdálenost mezi dvěma proměnnými znamená silnou korelaci. Lze nalézt i shluk podobných proměnných, jež spolu korelují. Tento graf můžeme považovat za most mezi původními proměnnými a hlavními komponentami, protože ukazuje, jakou měrou přispívají jednotlivé původní proměnné do hlavních komponent. Někdy se podaří hlavní komponenty y_1, y_2, ... pojmenovat, vysvětlit a přidělit jim fyzikální, chemický nebo biologický význam. Pak lze názorně vysvětlit, jak jednotlivé původní proměnné x_j, $j = 1,..., m$, přispívají do první hlavní komponenty y_1 nebo do druhé hlavní komponenty y_2. Některé původní proměnné x_j přispívají kladnou vahou, některé zápornou. Bývá zajímavé sledovat kovarianci původních proměnných x_j v prostorovém 3D grafu komponentních vah y_1, y_2 a y_3. Jsou-li proměnné x_j, $j = 1,..., m$, blízko sebe v prostorovém shluku, jde o silnou pozitivní kovarianci. Kovariance však nemusí ještě nutně znamenat korelaci. Výklad grafu komponentních vah lze obecně shrnout do následujících bodů:

1. *Důležitost původních proměnných* x_j, $j = 1,..., m$: proměnné x_j s vysokou mírou proměnlivosti v datech objektů mají vysoké hodnoty komponentní váhy. Ve 2Ddiagramu prvních dvou hlavních komponent pak leží hodně daleko od počátku. Proměnné s malou důležitostí leží blízko počátku. Když určíme *důležitost proměnných*, určíme tím také proměnlivost proměnných: jestliže například y_1 objasňuje 70 % proměnlivosti a y_2 jenom 5 % (přečteno z indexového grafu úpatí vlastních čísel), jsou původní proměnné x_j, j= 1,..., m, s vysokou vahou v y_1 tím pádem mnohem důležitější než proměnné x_j s vysokou vahou v y_2. Proměnné s úhlem 0E mezi průvodiči jsou zcela pozitivně korelované, proměnné s úhlem 90E jsou zcela nekorelované zatímco proměnné s úhlem 180E jsou sice nekorelované, říkáme však lépe že jsou negativně korelované.
2. *Korelace a kovariance:* původní proměnné x_j, j= 1,..., m, jsou blízko sebe, anebo proměnné x_j s malým úhlem mezi svými průvodiči proměnných a na stejné straně vůči počátku mají vysokou kladnou kovarianci a vysokou kladnou korelaci. Naopak, původní proměnné x_j daleko od sebe, anebo s velikým úhlem mezi průvodiči proměnných, jsou negativně korelovány.
3. *Spektroskopická data:* ve spektroskopických datech je 1-rozměrný graf komponentních vah často nejvhodnější. I zde platí pravidlo, že vysoké komponentní váhy představují vysokou důležitost proměnných x_j (vlnových délek).

(c) **Rozptylový diagram komponentního skóre** (Scatterplot) zobrazuje *komponent-ní skóre* čili hodnoty obyčejně prvních dvou hlavních komponent u všech objektů (obr. 4.9). Dokonalé rozptýlení objektů v rovině obou hlavních komponent vede k rozlišení objektů při jejich popisu pomocí y_1 a y_2. V rovině lze snadno nalézt shluk vzájemně podobných objektů a dále objekty odlehlé a silně odlišné od ostatních. Diagram komponentního skóre však může být i ve 3 či více hlavních komponentách a v rovin-ném grafu se pak sleduje pouze jeho průmět do roviny. Tento diagram se užívá k identifikaci odlehlých objektů, identifikaci trendů, tříd, shluků objektů, k objasnění podobnosti objektů atd. Je nemožné analyzovat všechny diagramy, protože jich je velmi mnoho: uvažujme například $m < n$ a pro $m = 10$ proměnných existuje
$m(m\text{-}1)/2 = 45$ diagramů, pro $m = 11$ pak 55 diagramů, pro $m = 12$ pak 66 diagramů, atd. Obvykle vybíráme diagramy y_1 vs. y_2, y_1 vs. y_3, y_1 vs. y_4 atd. Držíme se první hlavní komponenty y_1, protože v ní bývá nějvětší míra proměnlivosti v datech. Interpretace rozptylového diagramu komponentního skore lze shrnout do těchto bodů:

1. *Umístění objektů*. Objekty daleko od počátku jsou extrémy. Objekty nejblíže počátku jsou nejtypičtější.
2. *Podobnost objektů*. Objekty blízko sebe si jsou podobné, objekty daleko od sebe jsou si nepodobné.
3. *Objekty v shluku*. Objekty umístěné zřetelně v jednom shluku jsou si podobné a přitom nepodobné objektům v ostatních shlucích. Dobře oddělené shluky prozrazují, že lze nalézt vlastní model pro samotný shluk. Jsou-li shluky blízko sebe, znamená to značnou podobnost objektů.
4. *Osamělé objekty*. Izolované objekty mohou být odlehlé objekty, které jsou silně nepodobné ostatním objektům. Pravidlo platí, pokud se nejedná o zdánlivou nehomogenitu danou sešikmením dat a odstranitelnou transformací proměnných.
5. *Odlehlé objekty*. V ideálním případě bývají objekty rozptýlené po celé ploše diagramu. V opačném případě je něco špatného v modelu, obyčejně je přítomen silně odlehlý objekt. Odlehlé objekty jsou totiž schopny zbortit celý diagram, ve srovnání se silně vybočujícím objektem jsou ostatní objekty nakumulovány do jediného úzkého shluku. Po odstranění vybočujícího objektu se ostatní objekty roztřídí po celé ploše diagramu a teprve vypovídají o existujících shlucích.
6. *Pojmenování objektů.* Výstižná jména objektů slouží k hledání hlubších souvislostí mezi objekty a mezi pojmenovanými hlavními komponentami. Snadno obkroužíme shluky podobných objektů nebo nakreslením spojky mezi objekty vystihneme tak jejich fyzikální či biologický vztah.
7. *Vysvětlení místa objektu.* Umístění objektu na ploše v diagramu může být porovnáváno s komponentními vahami původních proměnných ve dvojném grafu a pomocí původních proměnných pak i vysvětleno.

(d) **Dvojný graf** (Biplot) kombinuje předchozí dva grafy (obr. 4.10). Úhel mezi průvodiči dvou proměnných x_j a x_k je nepřímo úměrný velikosti korelace mezi těmito dvěma proměnnými. Čím je menší úhel, tím je větší korelace. Každý průvodič má své souřadnice na první a na druhé hlavní komponentě. Délka této souřadnice je úměrná příspěvku původní proměnné x_j do hlavní komponety, čili je úměrná komponentní váze. Kombinace obou grafů v jediném přináší cenné srovnání, jeden graf působí zde doplňkově vůči druhému. Když se ve dvojném grafu nachází objekt v blízkosti určité proměnné x_j, znamená to, že tento objekt "obsahuje" hodně právě této proměnné a je s ní v interakci. Interakce proměnných a objektů umožňuje také vysvětlit umístění objektů vpravo od nuly na ose y_1 (či vlevo od nuly) pomocí pozice proměnných v tomto grafu, resp. umístění nahoře od nuly (či dole od nuly) na ose y_2.

(e) **Indexový graf úpatí rozptylu reziduí** (Residual Scree Plot). V PCA se vyčíslí $\boldsymbol{E_0}$, $\boldsymbol{E_1}$, $\boldsymbol{E_2}$, ..., z nich $e_{\text{tot},1}$, $e_{\text{tot},2}$, ... a ty se vynesou do indexového grafu úpatí proti indexu A. Poslední bod bude pro menší ze dvou čísel n nebo m, $A = \min(n, m)$. Zlom na křivce úpatí $e_{\text{tot},i} = f(i)$ ukazuje na optimální počet hlavních komponent A, na nejlepší dimenzionalitu. V tomto bodě končí struktura a začíná šum. Graf je co do použití naprostou obdobou indexového grafu úpatí vlastních čísel.

(f) **Graf reziduí jednotlivých objektů.** Rozptyl reziduí jednoho i-tého objektu představuje vzdálenost mezi objektem a modelem. Je vhodné zobrazovat tuto veličinu pro všechny objekty a odhalit odlehlé objekty či jiné anomálie.

Diagnostikování častých problémů v PCA: v analýze metodou hlavních komponent se často setkáváme s následujícími problémy:

1. *Data neobsahují předpokládanou informaci.* Vysvětlení grafů a diagramů metody PCA nemá smysl, protože data neobsahují informaci, popisující studovaný problém.

2. *Užito příliš málo hlavních komponent.* V modelu PCA bylo použito příliš málo latentních proměnných. Nedostatečné vysvětlení dat vede ke ztrátě informace. Problém se může vyřešit opětovným rozborem grafu úpatí vlastních čísel.

3. *Užito příliš mnoho hlavních komponent* V modelu PCA bylo zahrnuto příliš mnoho latentních proměnných, což může vyvolat vážnou chybu, protože šum je zahrnut do modelu.
4. *Neodstranění odlehlých objektů.* Odlehlé objekty mohou být důvodem hrubých chyb v datech. Do modelu jsou vtahovány spíše hrubé chyby než zajímavé proměnlivosti v datech objektů.
5. *Odstraněné odlehlé objekty obsahovaly důležitou informaci.* Ztrátou určitých objektů se vytratila důležitá informace z dat a nalezený model je proto zkreslený.
6. *Komponentní skóre je nedostatečně analyzováno.* Nedostatečným rozborem důležitého rozptylového diagramu byly zanedbány důležité rysy v datech.
7. *Vysvětlení komponentních vah se špatným počtem hlavních komponent.* Může vést k vážnému zkreslení výkladu. Může totiž dojít k vyjmutí důležitých proměnných, protože se zdají býti odlehlými. Tento graf je mostem mezi prostorem původních proměnných a prostorem hlavních komponent PC. Když zvolíme špatný prostor PC, tento most už nám mnoho nepomůže.
8. *Přecenění standardních diagnostik v software.* Je třeba hodně rozvažovat a přemýšlet o úloze samé a specifickém problému řešeném před pohodlným přebíráním počítačových výsledků.
9. *Užití špatného předzpracování dat.* Chybná předúprava dat (ve škálování užitého centrování nebo standardizace, transformace logaritmická, mocninná, Boxova-Coxova atd.) může vést ke zkresleným závěrům a neporozumění úloze. Způsob předúpravy dat je obecně dán typem úlohy a druhem instrumentálních dat a může vést ke ztrátě informace.

Postup metody hlavních komponent (PCA)

Problém musí být správně a přesně definován. Je odpovědností řešitele, aby data obsahovala dostatek relevantní informace k řešení problému. Ani nejlepší počítačová metoda nemůže kompenzovat nedostatek informace v datech. Maticový graf korelace proměnných slouží k získání počáteční informace o celém datovém souboru. Odhalí, zda data potřebují škálování. Při prvním seznámení s daty se v rámci exploratorní analýzy, kam také patří metoda hlavních komponent PCA, aplikuje první výpočet touto metodou. Data je obvykle potřeba škálovat nebo alespoň centrovat. Lze vyzkoušet i ostatní formy předúpravy dat. V tomto stadiu se vždy vyčíslují všechny hlavní komponenty. První diagramy komponentního skóre slouží k odhalení odlehlých hodnot, tříd, shluků a trendů. Jsou-li objekty roztříděny do dobře oddělených shluků, je třeba určit způsob, jak je z dat oddělit a shluky pak analyzovat odděleně. Nikdy se nepokoušíme odhalovat a odstraňovat odlehlé proměnné, mohlo by pak dojít k odstranění cenné informace. Po redukci datového souboru na několik podsouborů, kdy shluky jsou modelovány odděleně, se znovu aplikuje metoda hlavních komponent PCA na jednotlivé podsoubory.

1. *Vyšetření indexového grafu úpatí vlastních čísel:* z hrany v tomto diagramu se určí nejlepší počet hlavních komponent.

2. *Výpočet vlastních vektorů pro hlavní komponenty:* vedle číselných hodnot se užívá i názorný čárový diagram hodnot vlastních čísel vektorů, který přehledně informuje o zastoupení původních proměnných x_j , $j = 1, ..., m$, v hlavních kompo-nentách.

3. *Výpočet komponentních vah:* matice párových korelačních koeficientů ukazuje na korelaci původních proměnných s hlavními komponentami. Čárový diagram názorně vysvětluje korelační strukturu mezi oběma druhy proměnných. Uživatel nyní vybere pouze prvních A hlavních komponent a vytvoří tak *PCA model.*

4. *Vyšetření grafu komponentních vah.*

5. *Vyšetření rozptylového diagramu komponentního skóre.*

6. *Vyšetření dvojného grafu.*

7. *Vyšetření grafu úpatí reziduí objektů:* rezidua objektů a rezidua proměnných by měla prokazovat dostatečnou těsnost proložení. Není-li tomu tak, je třeba se navrátit k předúpravě dat a celý výpočet PCA opakovat.