METODA HLAVNÍCH KOMPONENT A EXPLORATORNÍ ANALÝZA VÍCEROZMĚRNÝCH DAT

JIŘÍ MILITKÝ ,
Katedra textilních materiálů, Technická universita v Liberci, Hálkova 6
461 17 Liberec, e- mail: jiri.miliky@vslib.cz

MILAN MELOUN,
Katedra analytické chemie, Universita Pardubice, Pardubice

Motto: *Nic není universální*

Abstrakt:
Jsou popsány základní logické postupy realizace metody hlavních komponent (dále PCA) vycházející z různých hledisek. Je pojednáno o způsobech předběžné analýzy dat, vlastní realizaci PCA, možnostech vizualizace výstupů a různých způsobech omezení vlivu vybočujících bodů. Je diskutováno o vhodnosti využití PCA v průzkumové analýze a možných úskalích. Některé typy transformací a škálování dat jsou demonstrovány graficky.

1.Úvod

Jednou ze základních úloh analytické chemie je simultánní monitorování úrovně různých **látek** (proměnných) v materiálech, ovzduší, vodě a půdě. Cílem je často zjištění, zda dané látky (celkem *p*) nepřekračují zadané úrovně. Problémem je, že se jednotlivé látky navzájem ovlivňují a v řadě případů spolu silně souvisí, takže se často špatně samostatně interpretují. Navíc se informace o koncentracích těchto látek získávají z různých **zdrojů** (míst), které nejsou nezávislé. To vše vede k požadavku zkoumání struktur v datech a hledání vazeb mezi látkami resp. zdroji. Obyčejně je požadováno :

1. Nalézt kombinace původních proměnných, které lépe vystihují data než původní proměnné a objasnit jejich význam
2. Nalézt struktury a souvislosti v datech, které charakterizují jednotlivém zdroje a jejich možné vazby
3. Identifikovat nevýznamné kombinace složek (snížení dimense problému a eliminace šumů) a vybočující zdroje (indikace resp., eliminace atypických zdrojů)

Také celá řada dalších úloh z oblasti analytické chemie vede na zpracování vícerozměrných výběrů. Podobné problémy se vyskytují také v jiných oborech, kde se zkoumá chování systémů ovlivněných simultánně řadou souvisejících faktorů resp. při konstrukci modelů predikujících **vlastnosti výrobků** z **vlastností surovin** atd. Vše je komplikováno tím, že se vychází z experimentálních dat , která mají v těchto případech standardně některé specifické zvláštnosti:
(a) rozsahy zpracovávaných dat nejsou obyčejně velké,
(b) v datech se vyskytují výrazné nelinearity, neaditivity a struktury, které je třeba identifikovat a popsat,
(c) rozdělení dat jen zřídka odpovídá normálnímu běžně předpokládanému ve standardní statistické analýze,
(d) v datech se vyskytují vybočující měření a různé heterogenity,

(e) statistické modely se často tvoří na základě předběžných informací z dat (datově orientované přístupy),
(f) parametry statistických modelů mají mnohdy definovaný fyzikální význam, a musí proto vyhovovat velikostí, znaménkem nebo vzájemným poměrem,
(g) existuje jistá neurčitost při výběru modelu, popisujícího chování dat.

Z hlediska použití statistických metod je proto žádoucí mít možnost zkoumat statistické zvláštnosti dat (průzkumová analýza), ověřovat základní předpoklady o datech a hodnotit kvalitu výsledků s ohledem na základní schéma [1]

"data - model - statistická metoda"

Toto schéma se považuje za základ interaktivní tvorby statistických modelů všeho druhu. Při jeho praktickém použití však nastávají problémy zejména v případech, kdy se jedná o vícerozměrné úlohy. Již samotné znázornění dat vyžaduje použití různých projekcí, které však vzhledem k multikolinearitě , nelinearitám a dimensi problému nemusí dobře indikovat např. tzv. **vybočující hodnoty** (body), jejichž přítomnost může mít katastrofické důsledky s ohledem na interpretaci výsledků a praktické závěry. Standardně se pro průzkumovou analýzu vícerozměrných dat používá metoda hlavních komponent (PCA), která je dnes běžnou součástí prakticky všech programových systémů pro vícerozměrná data. To vede ke stavu, že je rutinně využívána tak, jak je naprogramována, což může často způsobit potíže tam, kde je vhodné volit alternativní cesty.

V této práci je pojednáno o způsobech předběžné analýzy dat, vlastní realizace PCA, možnostech vizualizace výstupů a různých způsobech interpretace výsledků. Je diskutováno o vhodnosti využití PCA v regresní analýze a možných úskalích. Jednotlivé postupy a metody jsou demonstrovány na datech z textilního oboru.

2. Metoda PCA

Většina metod vícerozměrné analýzy dat vychází z náhrady **původních proměnných** (látek, faktorů), které jsou korelované tzv. **hlavními komponentami**, které jsou párově nekorelované (ortogonální). Hlavní komponenty jsou většinou tvořeny lineární kombinací původních proměnných a při jejich konstrukci se obyčejně definují další omezení určující jednoznačně jejich polohy. Jedním ze základních požadavků bývá výběr takových směrů, které vždy vedou k maximálnímu snižování celkové variability dat [1]. U metody PCA je vstupem matice dat ***X*** *(Nxp)* obsahující hodnoty *N* měření (vzorků) pro *p* původních proměnných. Výstupem je matice ***Z*** *(Nxp)* , obsahující hodnoty *N* měření (vzorků) pro *p* hlavních komponent. Předpokládejme nejdříve, že matice X je sloupcově centrovaná, tj. sloupcové průměry jsou rovny nule (důvod tohoto centrování je uveden v kap.3). Matice ***Z*** je tvořena sloupci hlavních komponent, které jsou lineární kombinací sloupců matice ***X,*** což znamená, že platí

$$\mathbf{Z} = \mathbf{X} * \mathbf{A} \qquad (1)$$

kde ***A*** musí být ortogonální matice. Je zřejmé, že i matice ***Z*** je sloupcově centrovaná. Z geometrického hlediska tvoří řádky matic ***Z*** a ***X*** body v p rozměrném prostoru (souřadnicovém systému) proměnných resp.hlavních komponent. Existuje také inverzní transformace, která je vzhledem k ortogonalitě matice A dána vztahem

$$\mathbf{X} = \mathbf{Z} * \mathbf{A}^T \qquad (2)$$

Na základě vzájemných lineárních transformací lze určit, že $\mathbf{X} * \mathbf{X}^T = \mathbf{Z} * \mathbf{Z}^T$. Z této rovnosti, tj. invariance matic skalárních součinů, plyne, že obou souřadnicových systémech jsou zachovány Eukleidovské vzdálenosti mezi body a velikosti úhlů, které svírají vektory

spojující tyto body s počátkem souřadnic. Vzdálenosti a úhly definované maticemi skalárních součinů se často souhrnně označují jako **konfigurace**. Je tedy patrné, že matice ***A*** způsobuje pouze rotaci kolem počátku souřadnic. Nechť je symbolem ***G*** označená taková matice ***A*** způsobující rotaci kolem počátku souřadnic, pro kterou jsou hlavní komponenty vzájemně nekorelované. S použitím matice ***G*** vede transformace rov. (1) ke tvaru

$$\mathbf{Z} = \mathbf{X} * \mathbf{G} \qquad (3)$$

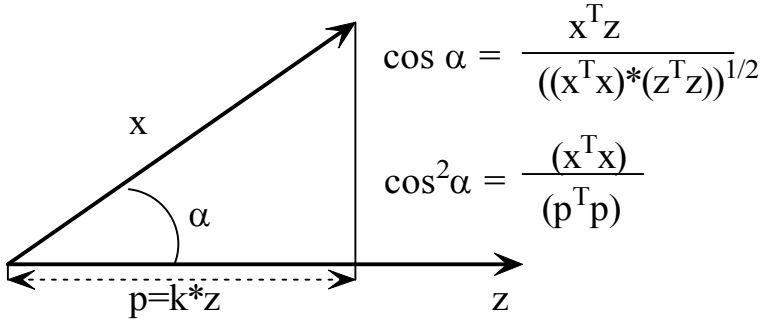
Sloupce matice ***Z*** se pak označují jako **skóry hlavních komponent** (dále skóry) a řádky definují souřadnice bodů vzhledem k tomuto souřadnému systému **hlavních os.** Protože je matice $\mathbf{X}^T\mathbf{X}$ až na násobivou konstantu rovna matici kovarianční matici výběru musí platit, že

$$\mathbf{Z}^T * \mathbf{Z} = \mathbf{G}^T * \mathbf{X}^T\mathbf{X} * \mathbf{G} = \mathbf{L}^2 \qquad (4)$$

Jak bude ukázáno dále, obsahuje matice ***G*** jako sloupce vlastní vektory a diagonální matice $\boldsymbol{L^2}$ obsahuje vlastní čísla matice $\mathbf{X}^T\mathbf{X}$. Standardně jsou vlastní čísla setříděná sestupně tj. $L^2_{j+1} \le L^2_j$. Předpokládejme, že matice ***X*** je tvořena p sloupci původních proměnných $\mathbf{X} = (\mathbf{x}_1, .. \mathbf{x}_p)$. Pak lze pro jednotlivé hlavní komponenty tj. sloupce matice $\mathbf{Z} = (\mathbf{z}_1, .. \mathbf{z}_p)$ a původní proměnné psát, že

$$\mathbf{z}_j = \sum_{i=1}^{p} G_{ij} * \mathbf{x}_i \quad \text{resp.} \quad \mathbf{x}_j = \sum_{j=1}^{p} G_{ij} * \mathbf{z}_j \qquad (5)$$

Tyto relace ukazují na vztah mezi původními a novými proměnnými. V prostoru proměnných je vektor $\boldsymbol{z}_j$ součet složkových vektorů $G_{ij} * \mathbf{x}_i$. Délka tohoto vektoru je součet projekcí vektorů $\boldsymbol{x}_i$ do směru dané hlavní osy. Schematicky je kolmá projekce vektoru ***x*** na vektor ***z*** znázorněna na obr. 1.



Obr. 1 Projekce vektoru ***x*** na vektor ***z***

Z obr 1. je zřejmé, že konstanta úměrnosti $k = (\mathbf{x}^T\mathbf{z})/(\mathbf{z}^T\mathbf{z})$. Vektor ***p*** je projekcí vektoru ***x*** na vektor ***z*** . Existuje tedy projekční matice ***Q,*** pro kterou je

$$\mathbf{p} = \mathbf{Q} * \mathbf{x} = k * \mathbf{z} = \mathbf{z}^T * \mathbf{x} * \mathbf{z} / (\mathbf{z}^T\mathbf{z}) = [\mathbf{z} * \mathbf{z}^T (\mathbf{z}^T\mathbf{z})] * \mathbf{x}.$$

Matice $\mathbf{Q} = \mathbf{z} * \mathbf{z}^T / (\mathbf{z}^T\mathbf{z})$ je ortogonální projekční matice (tj, symetrická a idempotentní). Délka vektoru ***p*** je dána vztahem

$$\|\mathbf{p}\| = \sqrt{\mathbf{p}^T\mathbf{p}} = \cos\alpha * \|\mathbf{x}\| = \mathbf{x}^T\mathbf{z} / \|\mathbf{z}\| = \mathbf{x}^T * \mathbf{Q} * \mathbf{x} = k * \|\mathbf{z}\| \qquad (6a)$$

V kontextu PCA je vektor ***z*** vyjádřen jako součet definovaný rov (5) a příspěvek jednoho vektoru ***x*** k tomuto součtu je úměrný konstantě ***k*** . Při zkoumání vazeb mezi vektory $\boldsymbol{z}_j$ a $\boldsymbol{x}_i$ umožňuje tato analýza lépe porozumět geometrickým souvislostem.

Je zřejmé, že délka vektoru $\boldsymbol{z}_j$ je dána výrazem $\|\mathbf{z}_j\| = \sqrt{\mathbf{z}_j^T * \mathbf{z}_j} = L_j$, kde L_j je odmocnina z j-tého vlastního čísla. To plyne přímo z rov. (3). Délka projekce vektoru $\boldsymbol{x}_i$ na vektor $\boldsymbol{z}_j$ je dán vztahem (viz. rov (6a))

$$p_{ij} = \frac{\mathbf{x}_i^T * \mathbf{z}_j}{\|\mathbf{z}_j\|} = G_{ij} * L_j \qquad (6)$$

V této rovnici bylo využito ortogonality sloupců matice $\boldsymbol{Z}$ tj. $\mathbf{z}_j * \mathbf{z}_k = 0 \quad pro \quad j \neq k$. Při konstrukci vektoru z se sčítají složkové vektory vektorů $G_{ij} * \mathbf{x}_i$, takže je celková délka vektoru $\boldsymbol{z}_j$ vyjádřitelná vztahem [2]

$$\|\mathbf{z}_j\| = L_j = \sum_{i=1}^{p} G_{ij} * p_{ij} = \sum_{i=1}^{p} G_{ij}{}^2 * L_j \qquad (7)$$

Rov. (7) ukazuje, že příspěvek každé původní proměnné (v přítomnosti ostatních) k délce vektoru $\boldsymbol{z}_j$ je úměrný čtverci G_{ij}. Veličina L_j je úměrná směrodatné odchylce nové proměnné (hlavní komponentě). Zajímavé je také určení vazby mezi vektory $\boldsymbol{z}_j$ a $\boldsymbol{x}_i$, kdy pro odpovídající korelační koeficient r_{ij} platí

$$r_{ij} = \frac{\mathbf{x}_i^T * \mathbf{z}_j}{\|\mathbf{x}_i\| * \|\mathbf{z}_j\|} = \frac{G_{ij} * L_j}{\|\mathbf{x}_i\|} \qquad (8)$$

Ve statistické terminologii je délka centrovaného vektoru úměrná směrodatné odchylce, protože $\mathbf{x}_i^T * \mathbf{x}_i = \sum_{j=1}^{N} x_{ji}^2$ zde odpovídá součtu čtverců odchylek hodnot odpovídajících i-té proměnné. Z rov. (8) a (6) plyne ,že $p_{ij} = G_{ij} * L_j = \|\mathbf{x}_i\| * r_{ij}$. Pokud je $\|\mathbf{x}_i\| = 1$, platí že $p_{ij} = r_{ij}$. K této situaci dojde v případě, že data jsou standardizovaná, tj. $\mathbf{X}^T\mathbf{X} = \mathbf{R}$ je rovna korelační matici. V maticovém vyjádření je projekční matice projekcí **x**-vektorů do **z**-vektorů ve tvaru $\boldsymbol{P} = \boldsymbol{G}*\boldsymbol{L}$ a korelační matice ve tvaru $\boldsymbol{R} = \boldsymbol{S}^{-1}*\boldsymbol{G}*\boldsymbol{L}$.

V některých případech se při rekonstrukci matice $\boldsymbol{X}$ využívá pouze omezeného počtu $\boldsymbol{k} < \boldsymbol{p}$ těch hlavních komponent, které nejvíce přispívají k snížení celkové variability dat. Lze ukázat, že platí

$$\sum_{j=1}^{k} G_{ij}^2 \leq 1 \quad tj. \quad \sum_{j=1}^{k} p_{ij}^2 / L_j^2 \leq 1 \qquad (9)$$

Projekce vektoru $\boldsymbol{x}_i$ do *k*- rozměrného pod-prostoru pouze prvních *k* hlavních komponent leží uvnitř hyper-elipsoidu jehož poloosy jsou L_j.

Při konstrukci matice $\boldsymbol{Z}$ se běžně používá pouze omezený počet hlavních komponent, takže platí model

$$\mathbf{x}_i = \sum_{j=1}^{k} G_{ij} * \mathbf{z}_j + \mathbf{e}_i \qquad (10)$$

kde chybový člen $\boldsymbol{e}$ souvisí s hlavními komponentami, které nebyly použity při rekonstrukci vektoru $\boldsymbol{x}_i$, tedy

$$\mathbf{e}_i = \sum_{j=k+1}^{p} G_{ij} * \mathbf{z}_j \qquad (11)$$

Délka vektoru $\boldsymbol{e}_i$ je maximálně L_{k+1}, protože platí, že $\sum_{j=k+1}^{p} G_{ij}^2 \leq L^2{}_{k+1}$.

V *N* rozměrném prostoru měření (vzorků) je rov. (5) interpretovatelná jako lineární regresní model , kde vysvětlující proměnné jsou původní proměnné a vysvětlovaná proměnná je hlavní komponenta $\boldsymbol{z}_j$. To je alternativní motivace pro PCA.

Je také možné uvažovat každou původní proměnnou jako lineární kombinaci všech ortogonálních hlavních komponent $\boldsymbol{z}_j$ *j = 1..p*. Délky těchto složek jsou projekce p_{ij}, které poskytují souřadnice pro $\boldsymbol{x}_i$ s ohledem na směry hlavních komponent.

Ze statistického hlediska se PCA uvažuje jako popisná vícerozměrná metoda založená na spektrálním rozkladu kovarianční matice $\mathbf{\Sigma}$ definovaném vztahem (viz. rov. (3))

$$\mathbf{\Sigma} = \mathbf{G} * \mathbf{L}^2 * \mathbf{G}^T \qquad (12)$$

Předpokládá se uspořádání vlastních čísel podle velikosti, takže j – tému vlastní číslo $L_j^2 = \lambda_j$ je co do velikosti na j-tém místě a odpovídá mu j-tý vlastní vektor $\boldsymbol{G}_j$, tj. j-tý sloupec matice $\boldsymbol{G}$.

Častým důvodem použití PCA je snížení rozměrnosti problému, kdy se místo původních *p* proměnných vybere jenom *k* hlavních komponent odpovídajících největším vlastním číslům, které objasňují největší podíl variability v datech. Pro účely průzkumové analýzy se vybírají dvě nebo tři hlavní komponenty a data se znázorňují v prostoru těchto hlavních komponent graficky. To umožňuje relativně snadno odhalit struktury v datech jako jsou skupiny bodů, isolované body atd. Pro posouzení struktur v datech je možné použít i jiné dvojice resp. trojice hlavních komponent a chápat PCA jako jeden ze způsobů 2D resp. 3D projekce dat. Standardně se tvoří graf skórů tj. sloupců matice $\boldsymbol{Z}$.Tento graf je pochopitelně silně ovlivněn transformací dat. Základní omezení naznačeného postupu spočívají v tom, že:

1. komponenty které objasňují malou část variability dat mohou být z hlediska analýzy vícerozměrných dat významné
2. Nelze a priori odhadnout, jaká část variability dat je již nevýznamná
3. Při použití ve spojení s regresními modely nesouvisí často vůbec variabilita vysvětlujících proměnných s variabilitou objasňovanou regresním modelem.

Standardní postup PCA pro průzkumové účely se dá rozdělit do těchto kroků [4]:

1. Transformace dat
2. Rozklad kovarianční resp. korelační matice
3. Určení počtu významných hlavních komponent
4. Vizuální zobrazení vícerozměrných dat

Standardně se vychází z vícerozměrných výběrů obsahujících N měření $(\boldsymbol{x}_1 ..... \boldsymbol{x}_N)$. Vektor $\boldsymbol{x}_j$ pro *j* té měření obsahuje složky $(x_{j1}, x_{j2} ... x_{jp})$. Výsledkem měření je tedy matice dat $\boldsymbol{X}$ řádu $N \times p$ obsahující *N* řádků (měření) a *p* sloupců (látek).

Určení počtu významných hlavních komponent je velmi kontroverzní úloha, protože významnost charakterizovaná velikostí vlastních čísel nijak nemusí souviset s významností pro popis datových struktur. To je dobře patrné např. v oblasti použití PCA v regresi. Přehled vybraných metod pro určování významných hlavních komponent podává práce [4]. V případě, kdy se PCA používá pro průzkumovou analýzu se provádí projekce do dvou resp. tří hlavních komponent a není obtížné vyzkoušet různé kombinace.

Standardním výstupem PCA je graf skórů (sloupců matice $\boldsymbol{Z}$) pro vybrané dvojice hlavních komponent. Někdy se tento graf doplňuje o vektory projekcí jako řádků matice $\mathbf{P} = \mathbf{G} * \mathbf{L}$ a vzniká **kombinovaný graf**.

3. Transformace dat

Transformace dat může mít řadu příčin a důsledků. Obyčejně souvisí se specifikou jednotlivých proměnných a jejich rozdělením. Speciálním případem transformace je lineární transformace nazývaná **standardizace**.

Jak již bylo ukázáno v kap. 2, vychází standardní PCA z sloupcově centrovaných dat (kovarianční matice $\mathbf{C} = \mathbf{X}^T\mathbf{X}$). Je však možné použít také normovaná data vedoucí ke korelační matici $\boldsymbol{R}$. Rozdíly v těchto dvou standardizacích jsou způsobeny různými vahami

jednotlivých původních proměnných při tvorbě matic skalárních součinů. Při použití kovarianční matice jsou sloupce matice $\boldsymbol{X}$ tj. původní proměnné "váženy" s ohledem na jejich délku $\|\mathbf{x}_i\|$, tj. úměrně směrodatné odchylce v původních jednotkách. Při použití korelační matice jsou sloupce matice $\boldsymbol{X}$ normovány tak, aby měly jednotkovou délku (nulový průměr a jednotkový rozptyl). Váhy všech proměnných jsou tedy stejné , protože délka všech proměnných je jednotková. Běžně se uvádí, že pro případ proměnných v různých jednotkách je vhodnější použití korelační matice. Bro a Smilde [3] rozebírají podrobně různé varianty centrování a normování. Obecně platí, že **centrování** odstraní absolutní člen v modelech a tím sníží počet odhadovaných parametrů a vede k omezení numerických potíží. Přitom nedochází ke změně struktury konfigurace (jen se posune se do počátku souřadnic). Normování se používá k odstranění závislosti na jednotkách a heteroskedasticitě u původních proměnných. Normování ovlivní kritérium odhadu parametrů (vážené nejmenší čtverce). Na druhou stranu je normování zcela nevhodné pro proměnné, které jsou na úrovni šumu (podíl signál/šum je velmi nízký). Zde dochází k nevítanému zvýraznění významnosti. V práci [6] se doporučuje použití vah 1/s (s je směrodatná odchylka dané proměnné) pro proměnné s výraznou převahou signálu. Pokud je signál a šum na stejné úrovni jsou doporučeny váhy 1/(4s) a tam, kde je šumová složka převládající se doporučuje vypuštění proměnné resp. váha 1/(20s). U proměnných, kde některé hodnoty leží pod mezí detekce *d* se určuje podíl signál/šum (S/N) ze vztahu

$$S/N = \frac{\sum I(x_i \geq d) * x_i}{d * N_d}$$

kde *I(.)* je indikátorová funkce a N_d je počet hodnot pod limitou detekce *d*.Pokud je S/N<2 je proměnná prakticky šum. Pro 0,2 <S/N<2 je proměnná málo odlišná od šumu. Prakticky toznamená, že přibližné konstantní hodnoty proměnné ve všech vzorcích indikují její nevhodnost.

V řadě případů jsou výchozí data vyjádřená jako **podíly z celku** (např. relativní zastoupení různých sloučenin a prvků). V celé řadě oblastí (např. stopové analýze) je běžné používat logaritmickou transformaci dat. Tato transformace má obecně některé výhody:

1. Omezuje působení extrémních hodnot
2. Snižuje pozitivní zešikmení dat běžné u řady výsledků měření
3. Stabilizuje nestejný rozptyl proměnných (heteroskedasticitu)

To znamená , že logaritmicky transformovaná data již není třeba dále normovat (postačuje sloupcové centrování).

Pro případ, že rozdělení dat je velmi vzdálené od normality, nebo jsou v datech skupiny vybočujících bodů doporučuje se použít pořadové transformace (hodnoty se nahradí jejich pořadími). Pak lze místo korelačních koeficientů na bázi momentů použít Spearmanovy pořadové korelační koeficienty. Na základě porovnání těchto transformací se standardizací resp. kombinace transformace a standardizace došel Baxter [5] k závěru, že logaritmická transformace a pořadová transformace jsou výhodné zejména tam, kde se vyskytují vybočující hodnoty. Žádná transformace nevyšla jako optimální pro všechny případy. V chemometrické literatuře se vyskytují ještě další speciální transformace vhodné pro speciální účely [4] .

4. Konstrukce hlavních komponent

Jak je patrné z rov. (3) je základem konstrukce hlavních komponent spektrální rozklad

kovarianční resp. korelační matice na vlastní čísla a vlastní vektory. Jde o jednu ze základních úloh lineární algebry. S ohledem na přesnost a spolehlivost se používá přímo rozkladu matice ***X*** pomocí metody SVD (singular value decomposition). Metoda SVD, rozkládá libovolnou obdélníkovou matici ***X*** (Nxp) na tři matice tj.

$$\mathbf{X} = \mathbf{U} * \mathbf{S} * \mathbf{V}^T \qquad (13).$$

Obyčejně se provádí tzv. zkrácená SVD kterou uvažujeme v dalším (pro zkrácenou SVD se mění rozměry matic **U** a **S**) Pro zkrácenou SVD je matice **S** (pxp) diagonální a obsahuje na diagonále tzv. singulární čísla matice **X.** Pokud má matice **X** hodnost r (tj. obsahuje pouze r lineárně nezávislých sloupců) je právě r kladných nenulových singulárních čísel seřazených dle velikosti, tj. $S_{11} \geq S_{22} \geq S_{33} \geq ... \geq S_{rr}$. Matice **U** (Nxp) a **V** (pxp) jsou ortogonální a normované, takže platí $\mathbf{U}^T\mathbf{U} = \mathbf{E}$ a $\mathbf{V}^T\mathbf{V} = \mathbf{E}$, kde **E** je jednotková matice.

Pro zkrácenou SVD platí, že kladná singulární čísla jsou odmocniny z vlastních čísel matice $\mathbf{X}^T\mathbf{X}$ (ale také matice $\mathbf{X}\mathbf{X}^T$), sloupce $\boldsymbol{u}_i$ matice **U** jsou vlastní vektory matice $\mathbf{X}\mathbf{X}^T$ a řádky $\boldsymbol{v}_i$ matice $\mathbf{V}^T$ jsou vlastní vektory matice $\mathbf{X}^T\mathbf{X}$. Platí, že singulární čísla jsou odmocniny z vlastních čísel, tedy $S_{ii} = L_i$ a matice $\boldsymbol{V}^T$ je rovna matici vlastních vektorů ***G***. S využitím SVD lze rov. (3) vyjádřit ve tvaru

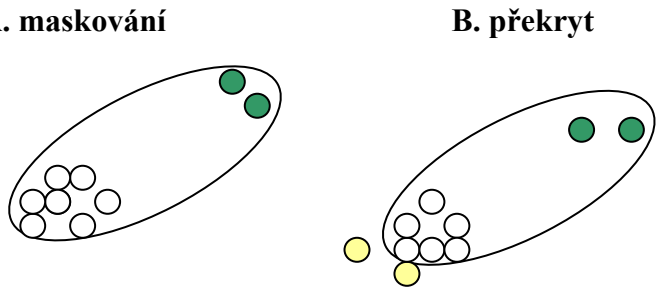
$$\mathbf{Z} = \mathbf{U} * \mathbf{S}$$

Důležitou vlastností SVD je že matice

$$\mathbf{X}_{(k)} = \sum_{i-11}^{k} \mathbf{u}_i * S_{ii} * \mathbf{v}_i$$

je nejbližší matice řádu *k* k matici ***X*** ve smyslu nejmenších čtverců odchylek. Je tedy minimalizováno kritérium $\sum_{ij}(X_{ij} - X_{(k)ij})^2$. Je tedy patrná úzká souvislost s metodou nejmenších čtverců. Samostatným problémem souvisejícím s rozkladem na vlastní čísla vlastní vektory je citlivost na vybočující body. Existují v zásadě dvě možnosti jak realizovat PCA v přítomnosti silně vybočujících bodů. První spočívá v jejich identifikaci a odstranění a druhý v použití robustních metod. Ukažme si základní problémy s identifikací vybočujících bodů.

Techniky indikace vybočujících bodů jsou citlivé na tzv. „**maskování**“, kdy vybočující se jeví jako korektní (díky zvětšení kovarianční matice) nebo „**překryt**“, kdy přítomnost vybočujících měření způsobí, že některá správná měření leží mimo akceptovatelnou oblast.(díky zkreslení kovarianční matice). Schematicky jsou tyto situace znázorněny na obr. 2 (vybočující body jsou tmavé).



Obr. 2 Příklad maskování (A) a překrytu (B)

Znázornění na obr 2. vychází z faktu, že čtverce zobecněných vzdáleností mají χ_p^2 rozdělení (elipsa je tedy hraniční oblast oddělující dobrá (D) a vybočující (V) data.

Řada metod pro identifikaci vybočujících bodů funguje jen pro některé situace nebo modely datových struktur. Příkladem jsou techniky uvažující pouze jedno vybočující měření (testy založené na odchylkách od průměru atd.) nebo speciální metody pro regresní modely.

Samostatným problémem je interpretace vybočujících hodnot. Existují dvě mezní situace:

A. Vybočující měření je chybné. To je třeba. případ, kdy vznikne chyba při měření, resp. zpracování dat (např. místo 0.74 je použita hodnota 74).

B. Vybočující měření je správné. To je případ, kdy byl použit nesprávný předpoklad o rozdělení dat (např. normalita pro případ, že reálné rozdělení je silně zešikmené) nebo jde o tzv. řídké jevy (které se u malých výběrů mohou jevit jako vybočující).

V realitě nelze často rozhodnout, o který případ se vlastně jedná. Problém je také v tom, co s vybočujícími hodnotami dělat. Přímá možnost, tj. jejich odstranění je nebezpečná ze dvou důvodů:

a) data se upravují tak, aby vyhovovala předpokládanému modelu a nelze tedy dobře posoudit jeho vhodnost,

b) variabilita dat vyjde extrémně nízká, což se může negativně projevit při porovnání s novými daty, resp. informacemi

Jednotný postup zde neexistuje a záleží na experimentátorovi, resp. zpracovateli jakou variantu zvolí.Vzhledem k tomu, že vybočující body jsou většinou extrémně vlivné vede zde nevhodná manipulace ke ztrátě informací a nesprávným závěrům.

Předpokládejme pro jednoduchost, že data mají p rozměrné normální rozdělení $N(\mu, \Sigma)$, kde μ je vektor středních hodnot a Σ je kovarianční matice. Vybočující měření leží v **oblasti**

$$out(\alpha_{1-\alpha}\mu, \Sigma) = \left(x \in R^p : (x-\mu)^T \Sigma^{-1}(x-\mu) > \chi^2_{1-\alpha}\right)$$

Tato oblast pokrývá celý prostor E^p s vyloučením vícerozměrného elipsoidu kolem vektoru středních hodnot. Vybočující body jsou tedy příliš vzdáleně od střední hodnoty.

Oblast vybočujících bodů OR pro výběr velikosti N je určena výrazem

$$OR(\alpha_{N,1-\alpha_N}, x) = \left(x \in R^p : (x-x_A)^T \mathbf{C}^{-1}(x-x_A) > c(p, N, \alpha_N)\right)$$

kde $\alpha_N = (1-\alpha)^N$ pro $\alpha = 0.05,\ 0.1$. Vše co leží v OR je vybočující. Oblast vybočujících bodů úzce souvisí se zobecněnou (Mahalanobisovou) vzdáleností resp. jejich čtvercem

$$d^2{}_i = (x_i - x_A)^T \mathbf{C}^{-1}(x_i - x_A)$$

Jako vybočující se pak identifikují ty body, pro které je $d_i > c(p, N, \alpha_N)$

Pro případ vícerozměrného normálního rozdělení a velké výběry je $c(p, N, \alpha_N)$ dáno kvantilem chí kvadrát rozdělení

$$c(p, N, \alpha_N) = \chi^2_p(1-\alpha/N)$$

Pro malé výběry je lépe použít modifikovaný koeficient

$$c(p, N, \alpha_N) = \frac{p*(N-1)^2 * F_{p,N-p-1}(1-\alpha/N)}{N*(n-p-1+p*F_{p,N-p-1}(1-\alpha/N))}$$

Aby bylo možno použít zobecněné vzdálenosti pro identifikaci vlivných bodů, je třeba určit „čisté odhady „ $\boldsymbol{x_A}$ a $\boldsymbol{C}$. Pro robustní odhad kovarianční matice se často volí [1]:

- M odhady
- S odhady minimalizující *det C* s omezením
- Odhady minimalizující objem konfidenčního elipsoidu

Při průzkumové analýze se vlastně očekává, že vybočující body budou výrazné na grafech , ale zkreslení hlavních komponent jako souřadnicového systému je nežádané. Pokud získáme „čisté odhady „ zejména kovarianční matice lze přímo sestavit nezkreslené hlavní komponenty a pak jsou vybočující body lépe identifikovatelné na grafech. Je tedy patrné, že robustní metody úzce souvisí s identifikací vlivných bodů. Z celé řady robustních metod navržených pro PCA jsou často používané techniky, kdy se hledají hlavní komponenty maximalizující robustní odhad rozptýlení dat. Příkladem je postup robPCA [7] resp. RAPCA. Jednoduché jsou metody stanovení „čisté podmnožiny dat“ složený z těchto kroků:

1. Výběr základní podmnožiny bud na základě
   - Mahalanobisovy vzdálenosti a uřezání podezřelých dat
   - Vzdálenosti od mediánu

Výsledkem je podmnožina „čistých dat“ s parametry x_{Ac} C_c

2. Výpočet reziduí

$$d_i = (x_i - x_{AC})^T \mathbf{C}_C{}^{-1} (x_i - x_{AC})$$

3. doplnění „čisté podmnožiny“ o body s reziduem menším než $c * \chi^2_\alpha$, kde

$c_1 = \max(0, (h-r)/(h+r))$; $h = (n+p+1)/2$

$c_2 = 1 + (p+1)/(n-p) + 2/(n-1-3p)$

$c = c_1 + c_2$

4. Skončení procesu v okamžiku, kdy se již nic nepřidává ani neubírá

Poměrně jednoduchá je metoda využívající kombinace identifikace potenciálně vybočujících bodů a uřezaných odhadů. V i té iteraci se určí uřezané odhady x_{RC} a $\boldsymbol{C}_C$, kde se uřezává definované procento (obyčejně 30%) bodů s nejvyššími zobecněnými vzdálenostmi z vektoru $\boldsymbol{d}^2_{i-1}$ vypočítaného v i-1 té iteraci. Z takto získaných odhadů se vypočte vektor opravených zobecněných vzdáleností $\boldsymbol{d}^2_i$ a přechází se na i+1 ní iteraci.

Proces je ukončen, když se ve dvou následujících iteracích nemění odhady parametrů x_{RC} a $\boldsymbol{C}_C$. Po získání finálních odhadů již postačuje použít klasickou PCA na matici $\boldsymbol{C}_C$.

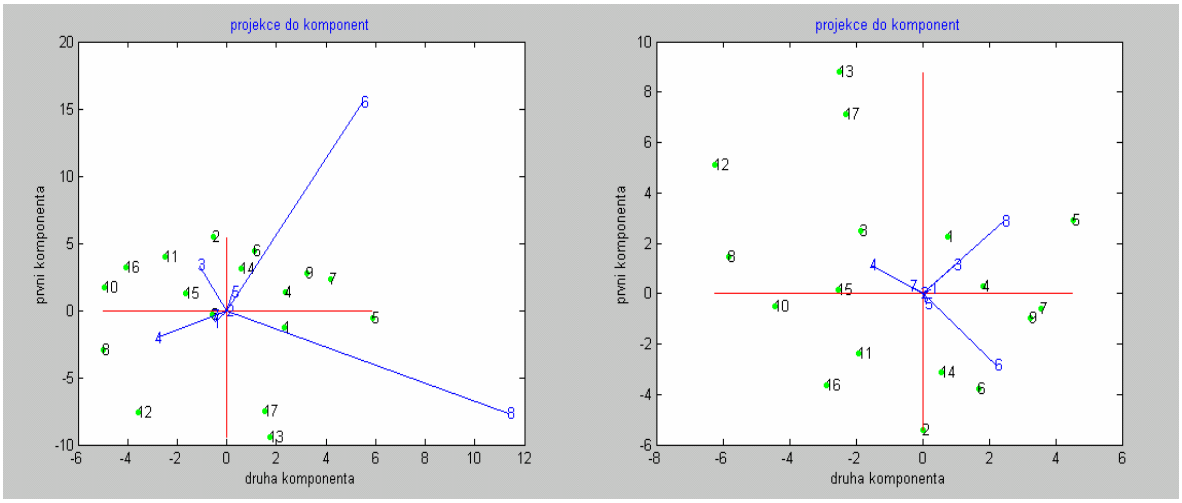
5. Program PCA1

Program PCA1 v jazyce MATLAB byl sestaven tak aby umožnil různé typy transformace dat, standardizace a případně robustní odhad hlavních komponent. Vychází se z SVD metody. Program obsahuje tyto základní volby:

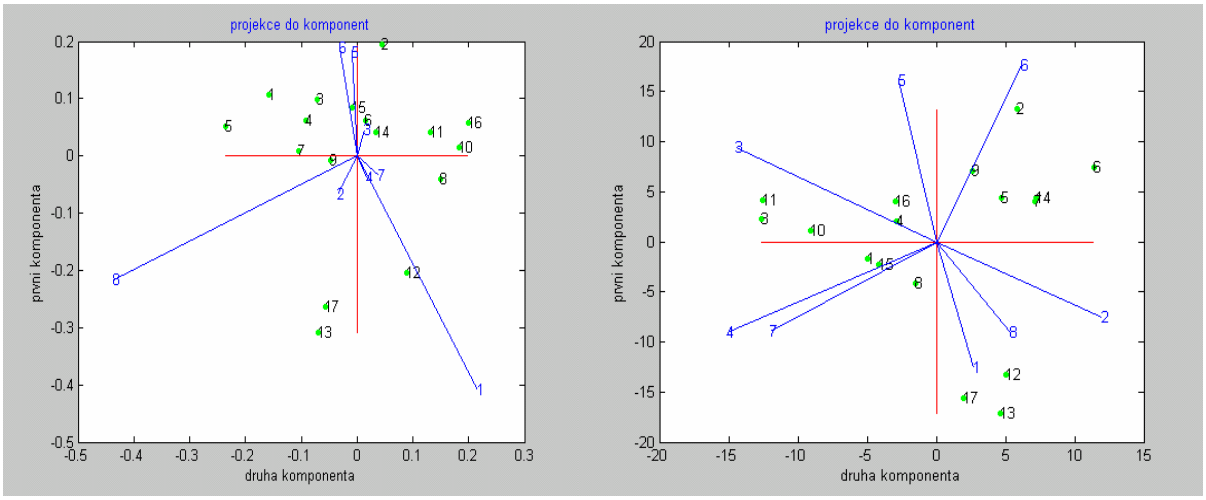
1. **Typ transformace** (bez transformace, logaritmická transformace a pořadová transformace)
2. **Typ škálování** (sloupcové centrování, vážení pomocí směrodatných odchylek, a normování)
3. **Druh odhadu hlavních komponent** (standardní metoda a robustní RAPCA)

Grafickým výstupem je především kombinovaný graf skórů a projekcí.

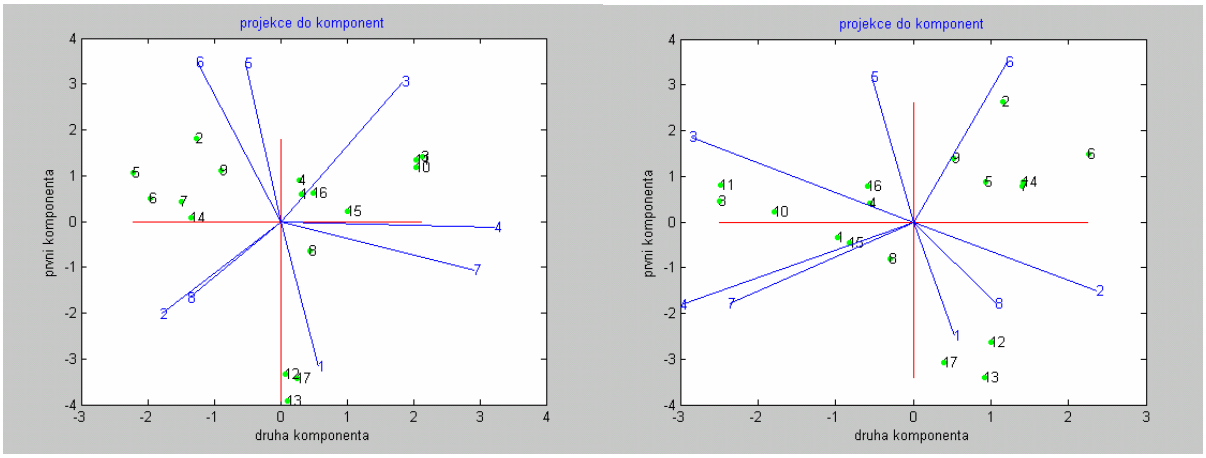
Pro ilustraci programu jsou na obr 1-4 ukázány kombinované grafy pro různé typy voleb. Byl zvolen příklad z oblasti porovnání 8 vlastností bavlněných vláken. Data jsou popsána v práci [8].



A **B**



C **D**



E F

Obr. 1 Vliv volby transformace a metody odhadu korelační matice na kombinované grafy A (centrovaná data, klasická metoda), B(centrovaná data robustní metoda), C(centrovaná data , logaritmy, klasická metoda), D (centrovaná data , pořadová transformace, klasická metoda), E (jako A ale normovaná data), F (jako D ale normovaná data)

8. Závěr

Je patrné, že metoda PCA má celou řadu specifických zvláštností. V řadě případů je třeba i ve zdánlivě jednoduchých situacích používat poměrně speciální postupy. Formální aparát PCA resp. transformace dat bez hlubšího rozboru zde může vést ke zkresleným informacím.

Poděkování:

Tato práce vznikla s podporou výzkumného centra Textil LN00B090 a grantu NB/7391-3.

9. Literatura

[1] Meloun M., Militký J.*: Zpracování experimentálních dat*, East Publishing Praha 1998

[2] Arnold A., Collins A., J.: Appl. Statist. **42**,381, (1993)

[3] Bro R., Smilde A, K.: J. Chemometrics **17**,16 (2003)

[4] Johnson G.W., Ehlich R.: Environmental Forensic **3**,59 (2002)

[5] Baxter M.,J.: Appl. Statist. . **44,** 513 (1995)

[6] Paatero P., Hopke P. K.: Analytica Chimica Acta 1-13 (2003) v tisku

[7] Smolinski A., Walczak B., Einax J., V.: Chemosphere **49**, 233, (2002)

[8] El Mogahzy E., Broughton R.M.: *Text.Res.J.***59**, 440 (1989)