# METODA HLAVNÍCH KOMPONENT V LABORATORNÍ PRAXI

**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:** *V jednoduchosti je síla*

## Abstrakt:

*Jsou popsány základy realizace metody hlavních komponent (dále PCA) vycházející z různých hledisek. Je pojednáno o možnostech interpretace transformovaných os (hlavních komponent). Na příkladech simulovaných dat s různou korelační strukturou jsou demonstrovány vlivy korelační struktury dat na výsledky PCA.*

## 1.Úvod

Jednou ze základních úloh laboratorní praxe je měření vybraných parametrů (znaků) a interpretace výsledků. Jen zřídka dojde k situaci, kdy se měří pouze jeden parametr resp. vlastnost. Obyčejně jsou měřeny také související parametry resp. jsou k dispozici další informace (o technologii, struktuře, složení, vzorkování, podmínkách měření atd.), které způsobují, že výchozí data jsou vícerozměrná. Požadavkem je pak zkoumání struktur v datech, hledání vazeb a zjednodušení (komprese dat). Obyčejně je třeba :

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 jednotlivé znaky a jejich možné vazby
3. Identifikovat nevýznamné kombinace složek (snížení dimense problému a eliminace šumů) a vybočující data (indikace resp., eliminace atypických výsledků)

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 předpovídajících **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 (jako statisticky postačující se obyčejně uvažuje 100 dat na každý znak),
(b) v datech se vyskytují výrazné statistické vazby 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 graficky statistické zvláštnosti dat, zjednodušovat datové struktury s ohledem na minimalizaci ztráty informace a interpretovat vhodně získané výsledky. Již samotné znázornění dat vyžaduje použití různých projekcí, které však vzhledem k multikolinearitě 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 základních myšlenkách PCA a možnostech interpretace transformovaných os (hlavních komponent). Na příkladech simulovaných dat s různou korelační strukturou jsou demonstrovány vlivy korelační struktury dat na výsledky PCA.

## 2. Metoda PCA

Metoda hlavních komponent (PCA) je jedna z nejstarších a nejvíce používaných metod vícerozměrné analýzy. Poprvé byla zavedena Pearsonem již v roce 1901 a nezávisle Hotellingem v roce 1933. Cílem analýzy hlavních komponent je především zjednodušení popisu skupiny vzájemně lineárně závislých, tedy korelovaných znaků. V analýze hlavních komponent nejsou znaky děleny na závisle a nezávisle proměnné jako v regresi. Techniku lze popsat jako metodu lineární transformace původních znaků na nové, nekorelované proměnné, nazvané ***hlavní komponenty***. Každá hlavní komponenta představuje lineární kombinací původních znaků. Základní charakteristikou každé hlavní komponenty je její míra variability tj. ***rozptyl***. Hlavní komponenty jsou seřazeny dle důležitosti tj. dle klesajícího rozptylu, od největšího k nejmenšímu. Většina informace o variabilitě původních dat je přitom soustředěna do první komponenty a nejméně informace je obsaženo v poslední komponentě. Platí pravidlo, že má-li nějaký původní znak malý či dokonce nulový rozptyl, není schopen přispívat k rozlišení mezi objekty.

Standardním využitím PCA je **redukce počtu znaků** bez velké ztráty informace, a to užitím pouze prvních několika hlavních komponent. Toto snížení dimenze úlohy se netýká počtu původních znaků. Je tedy výhodné především pro možnost zobrazení vícerozměrných dat. Předpokládá se, že nevyužité hlavní komponenty obsahují malé množství informace, protože jejich rozptyl je příliš malý. Tato metoda je atraktivní především z důvodu, že hlavní komponenty jsou nekorelované. Namísto vyšetřování velkého počtu původních znaků s komplexními vnitřními vazbami analyzuje uživatel pouze malý počet nekorelovaných hlavních komponent.

Dále lze vybrané hlavní komponenty využít také k **testu vícerozměrné normality**. Analýza hlavních komponent je rovněž součástí průzkumové analýzy dat.

**Snížení rozměrnosti** je často využíváno při konstrukci komplexních ukazatelů jako lineárních kombinací původních znaků. Např. první hlavní komponenta je vlastně vhodným ukazatelem jakosti pokud původní znaky charakterizují její složky. Využití první hlavní komponenty jako komplexního ukazatele je běžné v oblasti ekonomie, sociologie a mediciny.

První dvě respektive první tři hlavní komponenty se využívají především jako techniky **zobrazení vícerozměrných dat** v projekci do roviny nebo do prostoru. Výhodou je, že tato projekce zachovává vzdálenosti a úhly mezi jednotlivými objekty.

V řadě případů jsou hlavní komponenty pouze jednou z fází komplexnější analýzy. Např. **regrese s využitím hlavních komponent** umožňuje odstranění problémů s multikolinearitou a přebytečným počtem vysvětlujících proměnných. (Pozor, také hlavní komponenty, kterým odpovídá malý rozptyl mohou být v kontextu regrese důležité). Oblíbené je také použití hlavních komponent v oblasti řízení jakosti.

## 3. Podstata analýzy hlavních komponent

Základním cílem PCA je transformace původních znaků $x_i$, $j$=1, ..., $m$, do menšího počtu latentních proměnných $y_j$. Tyto latentní proměnné mají vhodnější vlastnosti:

- je jich výrazně méně,
- vystihují téměř celou ***proměnlivost*** původních znaků
- jsou vzájemně nekorelované.

Latentní proměnné jsou nazvány ***hlavními komponentami***. Jde o 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 znaků, která má největší rozptyl mezi všemi ostatními lineárními kombinacemi. Má tvar

$$y_1 = \sum_{j=1}^{m} V_{1j}\, x_{Cj} = V_1^{\mathrm{T}}\, x_C$$

kde sloupcový vektor původních znaků $\boldsymbol{x_C}$ obsahuje původní znaky v odchylkách od středních hodnot čili centrované hodnoty $x_C = (x_1 - \mu_1, x_2 - \mu_2, \ldots, x_m - \mu_m,)^{\mathrm{T}}$. Je zřejmé, že rozptyl

$$D(y_1) = D(V_1^{\mathrm{T}} x_C) = \mathrm{E}[(V_1^{\mathrm{T}} x_C)(V_1^{\mathrm{T}} x_C)^{\mathrm{T}}] =$$

$$V_1^{\mathrm{T}}\, \mathrm{E}(x_C x_C^{\mathrm{T}})\, V_1 = V_1^{\mathrm{T}}\, C\, V_1$$

je závislý na velikosti vektoru koeficientů $\boldsymbol{V_1}$. Symbol $\boldsymbol{C}$ označuje kovarianční matici. Je tedy třeba zavést vhodné omezení na velikosti $\boldsymbol{V_1}$. Standardním je použití normalizace $\boldsymbol{V_1}^{\mathrm{T}}\, \boldsymbol{V_1} = 1$. Pro vektor koeficientů $\boldsymbol{V_1}^{\mathbf{T}} = (V_{11}, \ldots, V_{1m})^{\mathrm{T}}$ pak platí, že proměnlivost vyjádřená rozptylem $D(y_1)$ je maximální. Druhá hlavní komponenta

$$y_2 = \sum_{j=1}^{m} V_{2j}\, x_{Cj} = V_2^{\mathrm{T}}\, x_C$$

maximalizuje rozptyl $D(y_2) = V_2^{\mathrm{T}}\, C\, V_2$ za těchto omezujících podmínek

$$V_2^{\mathrm{T}} V_2 = 1 \text{ a } V_1^{\mathrm{T}} V_2 = 0$$

Druhá omezující podmínka zajišťuje kolmost obou hlavních komponent. Pro obecně $j$-tou hlavní komponentu $y_i$ platí, že minimalizuje rozptyl $D(y_j) = V_j^{\mathrm{T}}\, C\, V_j$ za celkem $j$-tice omezujících podmínek $V_j^{\mathrm{T}} V_j = 1$ a $V_i^{\mathrm{T}} V_j = 0$ pro všechna $i < j$. Lze snadno zjistit, že podmínky $V_i^{\mathrm{T}} V_j = 0$ zajišťují kolmost hlavních komponent. Pro nalezení vhodných vektorů $\boldsymbol{V_1}$, $\boldsymbol{V_2}$, ..., $\boldsymbol{V_m}$, je třeba řešit sérii maximalizačních úloh s omezeními na parametry ve tvaru rovnosti.

Řešení s využitím metody Lagrangeových multiplikátorů vede ke zjištění, že vektor $\boldsymbol{V_j}$ je vlastní vektor kovarianční matice $\boldsymbol{C}$, kterému odpovídá $j$-té největší vlastní číslo $\lambda_j$. Využívá se tedy známého rozkladu

$$\boldsymbol{C} = \boldsymbol{V}\,\boldsymbol{\Lambda}\,\boldsymbol{V}^{\mathbf{T}} \tag{1}$$

kde $\boldsymbol{V}$ je ($m$ x $m$) matice, obsahující jako sloupce vektory $\boldsymbol{V_j}$ a $\boldsymbol{\Lambda}$ je ($m$ x $m$) diagonální matice, obsahující na diagonále vlastní čísla $\lambda_1 <= \lambda_2 <= ... \lambda_m$ kovarianční matice. Matice $\boldsymbol{V}$ je ortogonální, tj. $\boldsymbol{V^T V} = \boldsymbol{E}$, kde $\boldsymbol{E}$ je jednotková matice. Z rovnice (1) je zřejmé, že rozptyl $D(y_j) = \lambda_j$ je roven $j$-tému vlastnímu číslu. Celkový rozptyl všech hlavních komponent je pak roven

$$\mathrm{tr\,C} = \sum_{\mathrm{i=1}}^{\mathrm{m}} \lambda_{\mathrm{j}}$$

kde tr (.) označuje stopu matice. Podíl variability, objasněný $j$-tou hlavní komponentou $y_j$ je pak

$$P_j = \frac{\lambda_j}{\sum_{i=1}^{m} \lambda_j}$$

Kovariance mezi $j$-tou hlavní komponentou a vektorem znaků $\boldsymbol{x_C}$ jsou rovny

$$\mathrm{cov}\,(x_C\,,\,y_j\,) = \mathrm{cov}\,(x_C\,,V_j^T\,x_C\,) = E\,(\,x_C\;\;x_C^T\,)\,V_j = C\,V_j = \lambda_j\,V_j$$

Platí tedy, že kovariance mezi $i$-tým znakem $x_i$ a $j$-tou hlavní komponentou $y_j$ je cov ($x_{Ci}$, $y_j$) = $\lambda_{\mathrm{j}}\,V_{ji}$, kde $V_{ji}$ je $i$-tý prvek vektoru $\boldsymbol{V_j}$. Pro odpovídající korelační koeficient $r(x_{Ci}, y_j)$ platí, že

$$r(\,x_C\,,\,y_j\,) = \frac{\lambda_j\,V_{ji}}{\sigma_{x_i}\sqrt{\lambda_j}} = \frac{\sqrt{\lambda_j}\,V_{ji}}{\sigma_{x_i}}$$

Je zřejmé, že pokud se místo centrovaných znaků $\boldsymbol{x_C}$ použijí ***standardizované znaky*** nazývané také ***normované*** či ***normalizované znaky*** a označované

$$x_N = \left( \frac{x_1 - \mu_1}{\sigma_{x_1}}, \frac{x_2 - \mu_2}{\sigma_{x_2}}, \ldots, \frac{x_m - \mu_m}{\sigma_{x_m}} \right),$$

vyjde korelační koeficient roven

$$\mathrm{cov}\,(\mathrm{x_N}\,,\,\mathrm{y_j}\,) = \mathrm{r}(x_N\,,\,y_j\;\,) = v_{ji}^{*}\sqrt{\lambda_j^{*}}$$

kde $V_{ji}^{*}$ a $\lambda_{\mathrm{j}}^{*}$ odpovídají rozkladu korelační matice $\boldsymbol{R}$. Použití standardizovaných znaků (tj. náhrada kovarianční matice $\boldsymbol{C}$ maticí korelační $\boldsymbol{R}$) zjednodušuje interpretaci a odstraňuje závislost na jednotkách měření.

## 4. Hlavní komponenty pro dvojrozměrná data

Uvažujme dvojici znaků $x_1$ a $x_2$, kterým odpovídají kovarianční matice $\boldsymbol{C}$ a korelační matice $\boldsymbol{R}$, definované vztahy

$$\mathrm{C} = \begin{pmatrix} \sigma_1^2 & \mathrm{C_{12}} \\ \mathrm{C_{12}} & \sigma_2^2 \end{pmatrix} \qquad \mathrm{R} = \begin{pmatrix} 1 & \mathrm{r} \\ \mathrm{r} & 1 \end{pmatrix}$$

Stanovme PCA pro případ korelační matice. Podmínka k určení vlastních čísel je

$$\det\,(\mathrm{R} - l\,\mathrm{E}\,) = \det \begin{pmatrix} 1-l & r \\ r & 1-l \end{pmatrix} = 0$$

Platí tedy $(1 - l)^2 - r^2 = 0$. Po roznásobení rezultuje kvadratická rovnice

$$l^2 - 2*l\; + 1 - r^2 = 0$$

která má řešení ve tvaru

$$l_1 = \lambda_1 = 0.5\left[2 + \sqrt{4 - 4(1 - r^2)}\right] = 1 + r$$
$$l_2 = \lambda_2 = 0.5\left[2 - \sqrt{4 - 4(1 - r^2)}\right] = 1 - r$$

Pro jednotlivé vlastní vektory je pak třeba řešit rovnice

$$(\mathrm{R} - \lambda_i \mathrm{E})\,\mathrm{V}_i = 0, \quad i = 1, 2$$

Tak pro $i = 1$ rezultuje soustava dvou homogenních rovnic

$$V_{11}(1 - \lambda_1) + V_{12} r = 0$$
$$V_{11} r + V_{12}(1 - \lambda_1) = 0$$

Normalizační podmínka $\boldsymbol{V}_1^{\mathrm{T}}\boldsymbol{V} = 1$ znamená dělení vektoru $\boldsymbol{V}_1$ jeho délkou

$$d = \sqrt{V_1^T V_1} = \sqrt{V_{11}^2 + V_{12}^2}\,.$$

Pro řešení výše uvedené soustavy rovnic je možné zvolit $V_{11} = 1$ a z první rovnice určit

$$V_{12} = (\lambda_1 - 1) / r = 1$$

Délka tohoto vektoru je

$$d_1 = \sqrt{1 + (\lambda_1 - 1)^2 / r^2} = \sqrt{\frac{r^2 + (\lambda_1 - 1)^2}{r^2}} = \sqrt{2}\,,$$

takže normalizovaný vlastní vektor $\boldsymbol{V}_1^*$ má tvar

$$\mathrm{V}_1^* = \begin{Bmatrix} \left[\dfrac{\mathrm{r}^2 + (\lambda_1 - 1)^2}{\mathrm{r}^2}\right]^{-1/2} \\ \\ \dfrac{\lambda_1 - 1}{\sqrt{\mathrm{r}^2 + (\lambda_1 - 1)^2}} \end{Bmatrix} = \begin{pmatrix} \dfrac{1}{\sqrt{2}} \\ \\ \dfrac{1}{\sqrt{2}} \end{pmatrix}$$

Podobně při řešení soustavy homogenních rovnic pro druhou hlavní komponentu vyjdou složky nenormalizovaného vektoru $\boldsymbol{V}_2$ jsou $V_{21} = r/(1 - r - 1) = -1$ a $V_{22} = 1$. Délka tohoto vektoru je rovna

$$d_2 = \sqrt{\frac{r^2 + (1 - r - 1)^2}{(1 - r - 1)^2}} = \sqrt{2}$$

Normalizovaný vektor $V_2^*$ má pak jednoduchý tvar

$$V_2^* = \begin{Bmatrix} \dfrac{-r}{\sqrt{r^2 + (1 - r - 1)^2}} \\ \\ \left[\dfrac{r^2 + (1 - r - 1)^2}{(1 - r - 1)^2}\right]^{-1/2} \end{Bmatrix} = \begin{bmatrix} -\dfrac{1}{\sqrt{2}} \\ \\ \dfrac{1}{\sqrt{2}} \end{bmatrix}$$

První hlavní komponenta je rovna $y_1 = \dfrac{1}{\sqrt{2}}(z_1 + z_2)$ a druhá hlavní komponenta je rovna $y_1 = \dfrac{1}{\sqrt{2}}(z_2 - z_1)$, kde

$$z_1 = (x_1 - E(x_1)) / \sqrt{D(x_1)}$$

a

$$z_2 = (x_2 - E(x_2)) / \sqrt{D(x_2)}$$

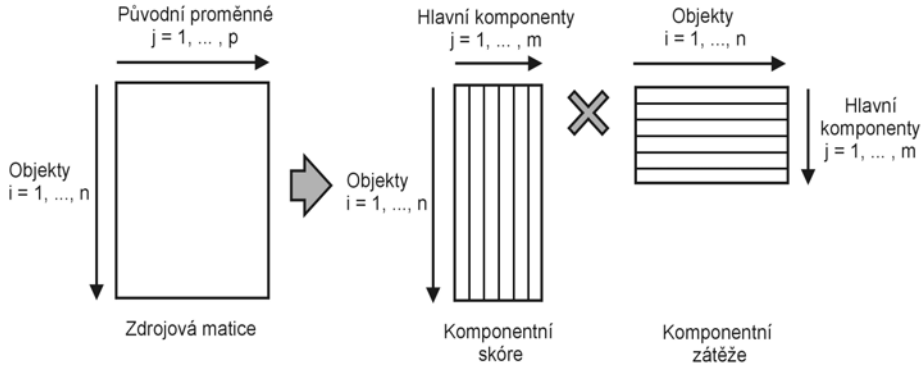
Je zřejmé, že při použití normalizovaných proměnných, znaků ve dvourozměrném případě nezávisí hlavní komponenty na korelaci v původních datech. Také je zřejmé, že dochází k pootočení souřadného systému o úhel $\cos\alpha = 1/\sqrt{2}$ tj. o $45^o$.

Je patrné, že prvky vlastních vektorů představují směrové kosíny nového souřadnicového systému hlavních komponent vzhledem k souřadnicovému systému původních znaků. Pro rozlišení mezi transformovanými znaky a transformovanými objekty se používá označení ***hlavní komponenty*** pro transformované znaky a ***skóry hlavních komponent*** (komponentní skóry) pro transformovaná data (objekty). Komponentními skóry objektů se také často označují ***hlavní*** *osy*.

## 5. Redukce počtu hlavních komponent

Protože platí, že součet rozptylů všech hlavních komponent je roven součtu rozptylů vstupních původních znaků, můžeme z podílu rozptylů jednotlivých hlavních komponent vůči celkovému rozptylu původních znaků, proměnných usuzovat na část proměnlivosti, vysvětlenou dotyčnou hlavní komponentou. Jestliže součet prvních (nejvyšších) $P_j$, $j$= 1, ..., $k$, podílů proměnlivosti (vyjádřených vlastními čísly) je dostatečně blízký jedné, respektive vyjádřeno v procentech 100 % (obvykle však stačí 80 % - 90 %), postačí brát v úvahu právě těchto prvních $k$ hlavních komponent pro dostatečné vysvětlení variability původních znaků. **Indexový graf úpatí vlastních čísel** je vlastně sloupcový diagram vlastních čísel $\lambda_1 <= \lambda_2 <= ... \lambda_m$ v závislosti na indexu $i$. Zobrazuje relativní velikost jednotlivých vlastních čísel. Významné komponenty jsou odděleny zřetelným zlomovým místem a hodnota indexu $i$ tohoto zlomu udává počet významných komponent.

Rozdíl mezi souřadnicemi objektů v původních znacích a v hlavních komponentách čili ztráta informace projekcí do menšího počtu rozměrů se nazývá *mírou těsností proložení* modelu PCA nebo také *chybou modelu PCA*. Na obr.1 je tato situace schematicky znázorněna spolu s použitým označením




Obr 1. Princip metody hlavních komponent

I při velkém počtu původních znaků $m$ může být $k$ velmi malé, běžně 2 až 5. Volba počtu užitých komponent $k$ vede k *modelu hlavních komponent PCA*. Vysvětlení užitých hlavních komponent, jejich pojmenování a vysvětlení vztahu původních znaků $x_j$ , $j$ = 1, ..., $m$, k hlavním komponentám $y_j$ , $j$ = 1, ..., $k$, tvoří dominantní součásti analýzy modelu hlavních komponent PCA.

Z obr.1 je zřejmé, že zdrojová centrovaná matice $\boldsymbol{X_C}$ se rozkládá na matici komponentních skórů $\boldsymbol{T}$ rozměru ($n$ x $k$) a matici komponentních zátěží $\boldsymbol{V_k^T}$ rozměru ($k$ x $m$). Vzhledem k tomu, že k rekonstrukci se obecně používá pouze $k$ z $m$ hlavních komponent, projeví se ztráta informace vznikem chybové matice $\boldsymbol{O}$ rozměru ($n$ x $m$). Platí tedy vztah

$$X_C = T\,V_k + O = \hat{X}_C + O$$

což je vlastně zápis bilineárního regresního modelu, kde se odhadují jak skóry $\boldsymbol{T}$, tak i vlastní vektory $\boldsymbol{V_k}$. Protože platí, že

$$\hat{X}_C = T\,V_k^T = X_C\,V_k\,V_k^T$$

stačí odhadnout jen matici hlavních komponent. Predikce $\hat{X}_C$ matice $\boldsymbol{X_C}$ se dá také vyjádřit jako lineární kombinace sloupců $\boldsymbol{t_i}$ matice komponentních skórů $\boldsymbol{T} = [\boldsymbol{t}_1, ..., \boldsymbol{t_k}]$. Vektor $\boldsymbol{t_i}$ rozměru ($n$ H 1) má tvar

$$t_i = X_C\,V_i$$

Složky vektoru $\boldsymbol{t_i}$ jsou komponentní skóry, odpovídající $i$-tému znaku. Proto $t_{ij} = \boldsymbol{V_i}^T\,\boldsymbol{X_{Cj}}$, kde $\boldsymbol{X_{Cj}} = (x_{C1j}, ..., x_{Cmj})^T$ má složky odpovídající $j$-tému řádku matice $\boldsymbol{X_C}$. Lze ukázat (viz vlastnosti SVD), že matice $\boldsymbol{t_i}\,\boldsymbol{V_i^T}$ rozměru ($n$ x $m$) mají hodnost 1. Matice $\hat{X}_C$ je tedy součet $k$ matic

$$\hat{X}_C = \sum_{i=1}^{k} t_i\,V_i^T$$

a matice reziduí

$$\hat{O} = X_C - \hat{X}_C = X_C - \sum_{i=1}^{k} t_i\,V_i^T = X_C\,(E - V_k\,V_k^T)$$

Pro určení matice $\boldsymbol{V_k}$ lze formálně použít přístup numerické aproximace a minimalizovat vzdálenost dist($\boldsymbol{X_C}$ - $\boldsymbol{T}\,\boldsymbol{V_k}^T$ ) (ve zvoleném smyslu) mezi oběma maticemi. Jednodušší je využití vztahů odvozených výše pro lineární model $f(\boldsymbol{V})$ a vektor $\boldsymbol{x}$ znaků, kdy se minimalizuje kritérium nejmenších čtverců odchylek

$$S(\mu, y_I, V_k) = \sum_{i=1}^{n} (x_i - \mu - V_k\,y_{Ii})^T\,(x_i - \mu - V_k\,y_{Ii})$$

Je tedy zřejmé, že metoda hlavních komponent s redukovaným počtem komponent je také případ speciálního regresního modelu, nebo aproximace kovarianční matice váženým součtem matic hodnosti 1.

## 6. Interpretace transformovaných os

Pro hlubší pochopení souvislostí mezi hlavními komponentami a původními znaky vyjádříme matici $\boldsymbol{X_c} = (\boldsymbol{x_1}, ..., \boldsymbol{x_m})$ jako $m$-tici sloupcových vektorů, které tvoří body v $m$ rozměrném prostoru znaků. Podobně matice *skórů hlavních komponent* $\boldsymbol{T} = (\boldsymbol{t_1}, ..., \boldsymbol{t_m})$ tvoří body v $m$ resp. $k$ rozměrném prostoru hlavních komponent. Je zřejmé, že skóry jsou vzájemně *ortogonální*, tj. $\boldsymbol{t_j^T}\,\boldsymbol{t_i} = 0$ pro $i \# j$ a vlastní vektory (zátěže) jsou *ortonormální* tj. $\boldsymbol{V_j^T}\,\boldsymbol{V_i} = 0$ pro $i \# j$ a $\boldsymbol{V_i^T}\,\boldsymbol{V_i} = 1$.

Pro $j$-tý vektor $\boldsymbol{t_j}$ nového znaku tj. hlavní komponenty platí, že

$$t_j = \sum_{i=1}^{m} V_{ij}\,x_{Ci}$$

kde $\boldsymbol{x_{Ci}}$ je vektor hodnot původního znaku (sloupec matice $\boldsymbol{X_C}$). Podobně lze provést inverzní lineární transformaci a vyjádřit $\boldsymbol{x_{Ci}}$ jako lineární kombinaci

$$x_{Ci} = \sum_{j=1}^{m} V_{ij}\,t_j$$

kde $V_{ij}$ jsou prvky matice $\boldsymbol{V_m}$, resp. $\boldsymbol{V_k}$ pokud se uvažuje jen $k$ komponent. V prostoru znaků je $\boldsymbol{t_j}$ vektorem získaným jako vážený součet vektorů $\boldsymbol{x_{Ci}}$ s vahami $V_{ij}$. Délka tohoto vektoru je součtem projekcí vektorů $V_{ij}\,\boldsymbol{x_i}$ do směru, odpovídající hlavní komponenty $y_j$.

Je zřejmé, že délka vektoru $\boldsymbol{t_j}$ je rovna

$$d(t_j) = \sqrt{t_j^T\,t_j} = \sqrt{v_j^T\,X_C^T\,X_C\,v_j} = \sqrt{\lambda_j}$$

Projekce $\boldsymbol{t_{Pji}}$ vektoru $\boldsymbol{x_{Ci}}$ na $\boldsymbol{t_j}$ je vyjádřena jako $\boldsymbol{t_{Pji}} = \boldsymbol{t_j}\, b$, kde $b$ je faktor úměrnosti. Platí tedy, že

$$t_j^T(x_{Ci} - b\, t_j) = 0 \quad \text{resp.} \quad b = \frac{t_j^T x_{Ci}}{t_j^T t_j} = \frac{t_j^T x_{Ci}}{\lambda_j}$$

Snadno lze určit, že platí rovnost

$$t_j^T x_{Ci} = t_j^T \left[ \sum_{k=1}^{m} V_{ik}\ t_k \right] = V_{ij}\ \lambda_j$$

protože $\boldsymbol{t_j^T}\ \boldsymbol{t_k} = 0$ pro $j$ â $k$ (vektory $\boldsymbol{t_j}$ jsou ortogonální). Pak je zřejmé, že $b = V_{ij}$. Vektor projekce $\boldsymbol{t_{Pji}} = V_{ij}\ \boldsymbol{t_j}$ má délku

$$p_{ij} = \sqrt{t_{Pji}^T\, t_{Pji}} = \sqrt{V_{ij}^2\, \lambda_j} = V_{ij} \sqrt{\lambda_j}$$

Délka vektoru $d(\boldsymbol{t_j})$ je pak součtem projekcí $p_{ij}$ vážených vahami $V_{ij}$ dle

$$d(t_j) = \sum_{i=1}^{m} V_{ij}\ p_{ij} = \sum_{i=1}^{m} V_{ij}^2 \sqrt{\lambda_j}$$

Tato rovnice ukazuje, že příspěvek každého původního znaku k délce vektoru $\boldsymbol{t_j}$ je úměrný čtverci $V_{ij}$. Protože délka tohoto vektoru $\sqrt{\lambda_i}$ je úměrná směrodatné odchylce příslušné hlavní komponenty, je jasné, že variabilita, objasněná $j$-tou hlavní komponentou je složena z příspěvků ***původních znaků*** a významnost těchto příspěvků je dána hodnotami $V_{ij}^2$. Malá hodnota $V_{ij}^2$ znamená, že $i$-tý původní znak přispívá málo k variabilitě $j$-té hlavní komponenty a je v tomto kontextu nevýznamný. Pokud je celý řádek matice $\boldsymbol{V}$ složen z malých hodnot, ukazuje to na nevýznamnost $i$-tého znaku po konstrukci hlavních komponent. Prvky $V_{ij}$ lze interpretovat poměrně zajímavě.

Výhodné je konstruovat ***příspěvkový graf***, jako $m$ skupin $m$ sloupců. Každá skupina odpovídá jedné komponentě a každý sloupec jednomu znaku. Sloupcové diagramy mají výšky odpovídající $V_{ij}^2 * \sqrt{\lambda}$. Výšky $m$ sloupců první skupiny (pro první hlavní komponentu) je normována tak, aby jejich součet byl roven 100 %. (podělení součtem jejich velikostí S1). Také pro další hlavní komponenty se využívá dělení S1 takže z výšky sloupců vychází jejich relativní význam. Příspěvkový graf umožňuje posouzení vlivu původních znaků na variabilitu jednotlivých hlavních komponent.

Minimalizace kolmých vzdáleností mezi $\boldsymbol{x_{Ci}}$ a $j$-tou hlavní komponentou zajišťuje maximalizaci rozptylu této hlavní komponenty. To umožňuje interpretovat PCA jako metodu hledání směrových kosínů vzájemně ortogonálních přímek tak, aby byl součet délek projekcí na tyto přímky *maximální*. Pro posouzení vztahů mezi původními znaky a hlavními komponentami se také využívá korelačních koeficientů mezi $\boldsymbol{x_{Ci}}$ a $\boldsymbol{t_j}$, což odpovídá kosínu jejich vzájemného úhlu $\alpha_{ij}$. Platí, že

$$r_{ij} = \cos \alpha_{ij} = \frac{x_{Ci}^T t_j}{\sqrt{(x_{Ci}^T x_{Ci})\, t_j^T t_j}} \approx \frac{V_{ij} \sqrt{\lambda_j}}{\sigma_i} = \frac{p_{ij}}{\sigma_i}$$

kde $\sigma_i$ je směrodatná odchylka příslušející $i$-tému znaku. Je patrné, že při použití normovaných proměnných (což je náhrada matice $\boldsymbol{S}$ maticí korelační $\boldsymbol{R}$) jsou korelační koeficienty $r_{ij} = p_{ij}$ rovny přímo dílčím projekcím. čím jsou $r_{ij}$ větší, tím jsou větší i projekce. To znamená, že $\boldsymbol{x_i}$ je blíže $\boldsymbol{t_j}$ a přispívá výrazněji k rozptylu $j$-té hlavní komponenty. Malé $r_{ij}$ naopak indikují malou významnost s ohledem na variabilitu hlavních komponent.

## 7. 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 , 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 ***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 ***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 ***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).

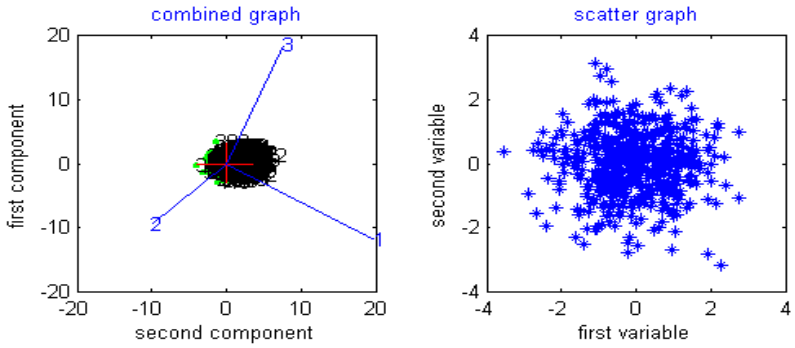
Pro případ, že rozdělení dat je velmi vzdálené od normality, nebo jsou v datech skupiny vybočujících bodů doporučuje se použít pořadové transformace (hodnoty se nahradí jejich pořadími). Pak lze místo korelačních koeficientů na bázi momentů použít Spearmanovy pořadové korelační koeficienty. Na základě 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] .

## 8. PCA pro simulovaná data

Pro ilustraci vlivu korelace v původních proměnných na výsledky PCA byla použita simulovaná data pocházející z tří rozměrného normálního rozdělení se speciálně definovanými korelačními strukturami. Bez újmy na obecnosti se předpokládal nulový vektor středních hodnot a korelační matice odpovídající kovarianční matici s prvky

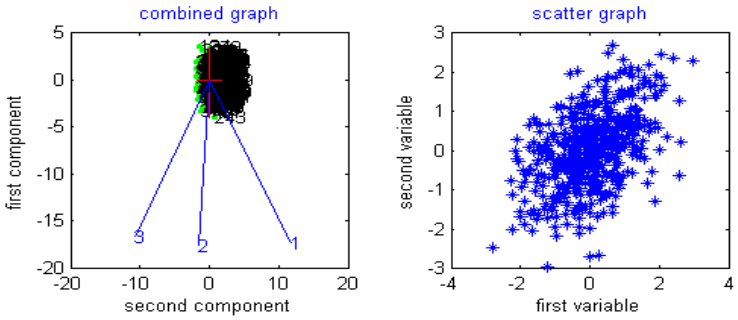
$$\begin{matrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{matrix}$$

Bylo generováno n = 500 dat. Na obr 2. je kombinovaný graf pro první dvě komponenty a rozptylový graf pro nekorelovaná data.



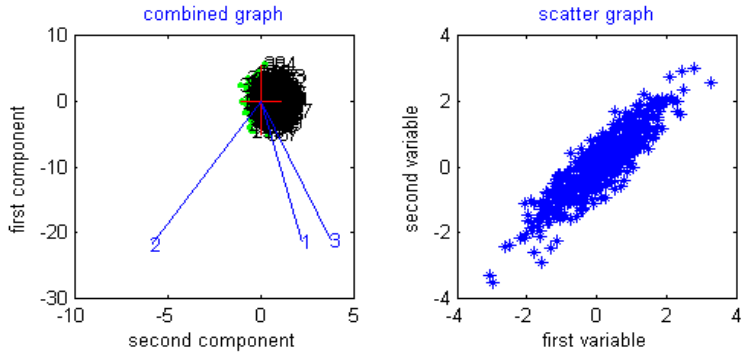

Obr. 2. Kombinovaný graf (všechny korelace nulové)

Na obr 3. je kombinovaný graf pro první dvě komponenty a rozptylový graf pro všechny párové korelace rovné 0,5




Obr. 3. Kombinovaný graf (všechny korelace 0,5)

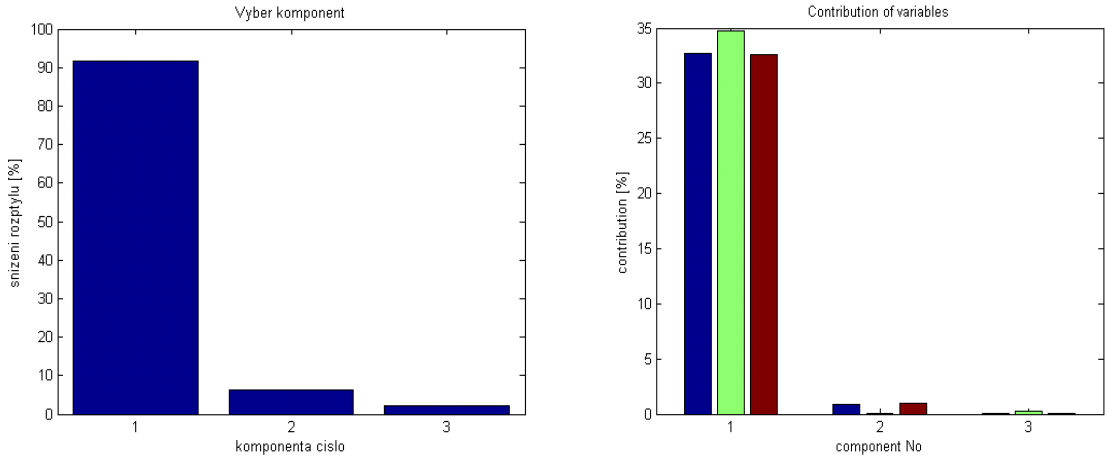
Na obr 4. je kombinovaný graf pro první dvě komponenty a rozptylový graf pro všechny párové korelace rovné 0,9




Obr. 4. Kombinovaný graf (všechny korelace 0,9)

Tyto grafy ukazují jak se mění polohy původních souřadnic a hlavních komponent.

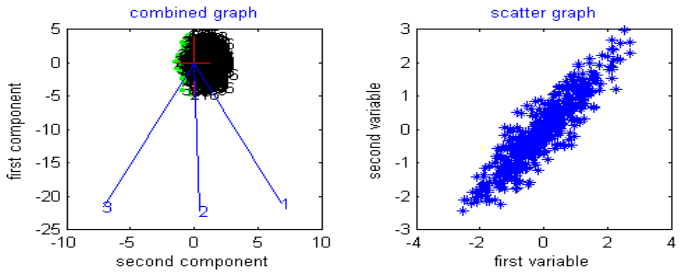
V řadě případů se stává, že struktura párových korelací nesouvisí s e vztahy mezi proměnnými. (falešné korelace). Uvažujme situaci, kdy je mezi x1 x2 vysoká korelace $r_{12} = H$ H $\to 1$ a existuje x3 pro kterou vyjde $r_{13} = H^2$ a $r_{23} = H$. Vícenásobný korelační koeficient je $R_{1(2,3)} = H$ a pro parciální korelační koeficienty platí $R_{1,3(2)} = 0$ a $R_{1,2(3)} = \frac{H}{\sqrt{1+H^2}}$. Je tedy patrné, že proměnná x3 nepřispívá k objasnění variability x1 a je z tohoto pohledu parazitní. Při simulaci bylo zvoleno H = 0,9. Na obr 5a je graf úpatí, a na obr 5b graf příspěvků.




**a)** **b)**

Obr 5. a) graf úpatí , b) příspěvkový graf

Na obr. 6 je znázorněn kombinovaný graf




Obr. 6. Kombinovaný graf

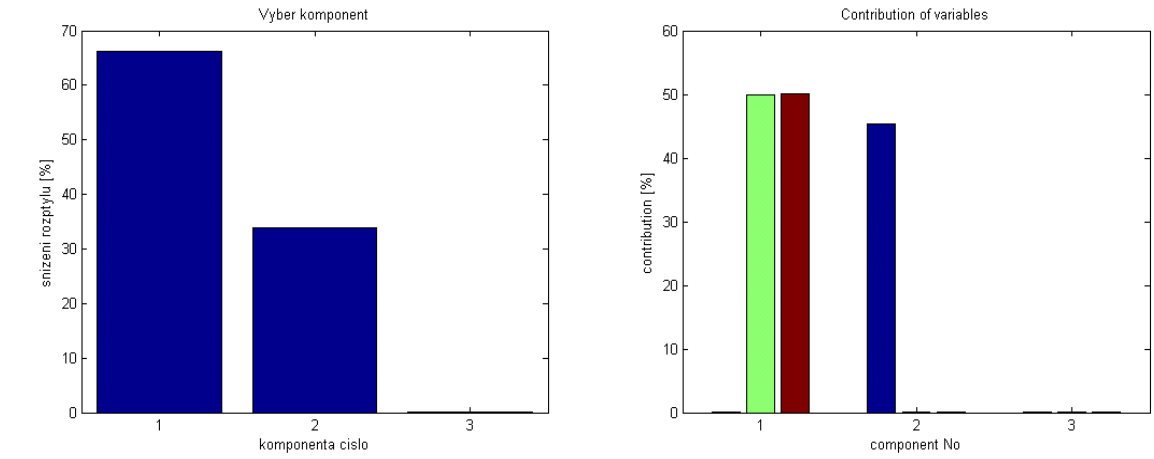
V tabulce 1 jsou uvedeny korelační koeficienty pro korelace mezi hlavními komponentami a původními proměnnými.

**Tabulka 1 Korelace mezi souřadnicovými systémy**

| | **1 komponenta** | **2 komponenta** | **3 komponenta** |
|---|---|---|---|
| **X1** | -0.9450 | 0.3057 | 0.1159 |
| **X2** | -0.9790 | 0.0182 | -0.2029 |
| **X3** | -0.9455 | -0.3107 | 0.0973 |

Je patrné, že příspěvkový graf ukazuje, že pouze jedna hlavní komponenta postačuje pro vyjádření těchto dat. Ve druhé komponentě se promítá nejvíce x1. Proměnná x3 parazitní vzhledem k x1 nebyla objevena protože není parazitní vzhledem k x2.

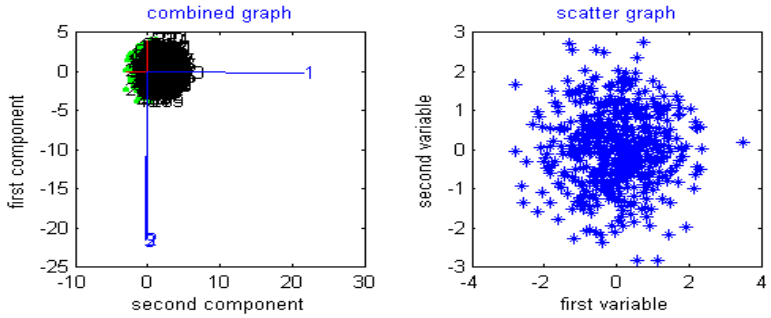
Jako příklad toho, že nízké párové korelační koeficienty mohou vést k vysokým parciálním korelačním koeficientům uvažujme situaci, kdy je mezi x1 x2 nízká korelace $r_{12} = H$ H $\to 0.01$ a existuje x3 pro kterou vyjde $r_{13} = 0$ a $r_{23} = \sqrt{1 - H^2}$. Vícenásobný korelační koeficient je $R_{1(2,3)} = 0.707$ a pro parciální korelační koeficienty platí $R_{1,3(2)} = -0.707$ a $R_{1,2(3)} = 0.707$. Je tedy patrné, že všechny proměnné jsou významné. Při simulaci bylo zvoleno H = 0,01. Na obr 7a je graf úpatí a na obr 7b graf příspěvků.




**a)** **b)**

Obr 7. a) graf úpatí , b) příspěvkový graf

Na obr. 8 je znázorněn kombinovaný graf




Obr. 8. Kombinovaný graf

V tabulce 2 jsou uvedeny korelační koeficienty pro korelace mezi hlavními komponentami a původními proměnnými.

**Tabulka 2 Korelace mezi souřadnicovými systémy**

| | **1 komponenta** | **2 komponenta** | **3 komponenta** |
|---|---|---|---|
| **X1** | -0.0108 | 0.9999 | 0.0001 |
| **X2** | -1.0000 | -0.0003 | -0.0052 |
| **X3** | -0.9999 | -0.0107 | 0.0052 |

Je patrné, že příspěvkový graf ukazuje, že první dvě hlavní komponenty postačují pro vyjádření těchto dat. Ve druhé komponentě se promítá pouze x1. Proměnné x2 a x3 se projevují pouze v první hlavní komponentě.

Z těchto výsledků je patrné, že PCA není schopna nahradit analýzu korelačních struktur. Na druhou stranu je možné provádět kompresi dat a vytvářet nové nekorelované proměnné.

## 9. 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 bez hlubšího rozboru zde může vést ke zkresleným informacím.

## Poděkování:

Tato práce vznikla s podporou výzkumného centra Textil LN00B090

## 10. 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)