# 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:** *Všechno není jinak*

**Abstrakt:**
*Jsou popsány základní problémy vznikající v důsledku použití vícerozměrných dat. Je pojednáno o vybraných projekcích pro obecně korelovaná data a jsou uvedeny některé možnosti snižování dimense ve statistické analýze vícerozměrných dat.*

## 1.Úvod

Jedním ze základních rysů vícerozměrných dat je jejich rozměr (dimense), která komplikuje jak jejich statistickou tak i průzkumovou analýzu. Pro znázornění vícerozměrných dat je pak třeba hledat vhodné projekce, velikosti výběrů musí být značně větší než u jednorozměrných dat a řada statistických technik naráží na obtíže. Je tedy vždy účelné uvažovat o redukci dat, která může být výhodná zejména pokud platí, že:

a) řada proměnných má variabilitu na úrovní šumu takže jsou s ohledem na statistickou analýzu nadbytečná.
b) V datech existují výrazné lineární vazby (korelace), které jsou způsobeny buď použitím redundantních proměnných nebo vazbami vyplývajícím i z jejich podstaty. Tyto proměnné lze často bez újmy na přesnosti nahradit podstatně menším počtem proměnných, které jsou již nekorelované.

Základním důvodem pro redukci počtu proměnných však může být tzv. *„prokletí vícerozměrnosti“* (multivariate curse), které se projevuji již v tom , že počet dat potřebných k zachování stejné přesnosti při odhadu funkce několika proměnných je exponenciálně rostoucí funkcí jejich počtu. V případě vícerozměrných dat se také vyskytuje problém tzv. *„prázdného prostoru“*, který klade další nároky na zvětšování velikosti výběru.
Samostatným problémem je vyjádření vzdálenosti, které se používá v řadě metod jak průzkumové tak i konfirmativní statistické analýzy dat. Místo klasické Eukleidovské vzdálenosti je výhodné použít vzdálenost Mahalanobisovu, která bere v úvahu i korelační vazby.

Tento příspěvek je zaměřen na diskusi ke třem výše zmíněným problémům vícerozměrné statistické analýzy dat, tedy:

1. Důsledky růstu dimense vícerozměrných dat
2. Snižování rozměrnosti dat
3. Vybrané projekce vícerozměrných dat

Protože je snahou demonstrovat základní a přitom opomíjené skutečnosti je omezen matematický aparát resp. citování původních prací. Čtenář nalezne oboje v knihách [1-5].

## 2. Základní postupy

Vyjděme z jednoduchého zadání, kdy je k dispozici pro $n$ objektů (bodů) celkem m znaků (proměnných) vyjádřených v kardinální škále . Zdrojová matice $\boldsymbol{X}$ má rozměr $n \times m$, kde standardně platí, že $n$ je podstatně vyšší než $m$. Předpokládejme, že veličina $m$ definuje dimensi problému (určující počet znaků).

Před vlastní aplikací vhodné metody vícerozměrné statistické analýzy je třeba vždy provést *exploratorní (průzkumovou) analýzu dat*, která umožňuje:

(a) posoudit *podobnost objektů* resp. jejich tendenci ke shlukování,
(b) nalézt *vybočující objekty*, resp. jejich znaky,
(c) stanovit, zda lze použít předpoklad *lineárních vazeb,*
(d) ověřit *předpoklady o datech* (normalitu, nekorelovanost, homogenitu).

Pro účely průzkumové analýzy vícerozměrných dat se používá různých technik, umožňujících jejich grafické zobrazení ve dvourozměrném souřadnicovém systému a u novějších výpočetních prostředků i v třírozměrném souřadnicovém systému. Toto zobrazení umožňuje zejména:

a) identifikaci objektů nebo jejich složek, které se jeví jako vybočující,
b) indikaci různých struktur v datech, které ukazují na heterogenitu použitého výběru nebo přítomnost různých dílčích výběrů s odlišným chováním.

Na základě těchto informací a výsledků testů normality (příp. grafických ekvivalentů těchto testů) pak může být před vlastní statistickou analýzou provedena řada různých korekcí, vedoucích k odstranění nehomogenity výběru a přiblížení se k vícerozměrné normalitě.
Většina používaných technik pro zobrazení vícerozměrných dat se dá zařadit do jedné ze dvou základních skupin, a to *zobecněné rozptylové diagramy* nebo *symbolové grafy*.

Pro základní případ dvojice náhodných znaků ($m = 2$) lze snadno a bez zkreslení konstruovat rozptylové grafy, které umožňují sledovat statistické zvláštnosti dat. Je možné použít i histogramy, neparametrické odhady hustoty a jiné grafy pro konstrukci výběrového rozdělení nebo jeho porovnání s rozděleními teoretickými.

Problémy však nastávají u vícerozměrných dat pro $m > 2$, kdy je možno buď volit několik různých grafů či vhodným způsobem provést redukci na dvou dimenzionální data. V obou případech nelze vyloučit zkreslení způsobené vícerozměrným charakterem dat, které může vést k mylným závěrům.

Jednoduchou možnost dvourozměrného zobrazení m-rozměrných dat představují *profily*. Každý bod $\boldsymbol{x}_i$ je zde charakterizován $m$ vertikálními úsečkami nebo sloupci. Jejich velikost je úměrná hodnotě odpovídající složky $x_{ij}$ $j = 1, ..., m$. Na $x$-ovou osu se vynáší index dané složky $j$. Profil pak vzniká spojením koncových bodů těchto úseček či sloupců. Je vhodné použít škálované znaky kde max max abs($x_{ij}$) je maximální hodnota absolutní velikosti složky $x_j$ vektoru $\boldsymbol{x}$ přes všechny body, $i = 1, ..., n$. Profily jsou jednoduché a umožňují určení rozdílů mezi jednotlivými body $\boldsymbol{x}_i$, $\boldsymbol{x}_k$ i v dílčích složkách. Snadno lze tedy identifikovat vybočující složku objektu, respektive skupiny objektů, s prakticky shodným chováním.

Ke zjednodušení interpretace a omezení artefaktů se často používá rozptylových grafů v *modifikovaných souřadnicích*, které souvisí se zavedením tzv. *latentních* proměnných resp. vhodnou projekcí vícerozměrných dat do dvou dimenzí. Z řady různých technik, jsou velmi často využívány techniky založené na metodě *hlavních komponent*, která je vhodná pro případy, kdy jsou sloupce matice $\boldsymbol{X}$ silně korelovány.

## 3. Vícerozměrná geometrie

Výše uvedené a řada dalších postupů grafického zobrazení dat ¨mohou být pro případ většího počtu znaků (větší rozměrnosti úlohy) omezeně použitelné. To lze snadno demonstrovat na případu dat která se budou koncentrovat v $m$ dimensionálním prostoru mimo jeho souřadnice. Vyjděme z představy, že data leží v hyper-krychli $[-1,1]^m$ na diagonálním vektoru $\boldsymbol{v}$ vycházejícím ze středu a končícím v jejím rohu. Úhel mezi tímto vektorem a souřadnicovou osou $\boldsymbol{e}_i$ je roven

$$\cos\theta_i = \frac{\mathbf{v} * \mathbf{e}_i}{\|\mathbf{v}\| * \|\mathbf{e}_i\|} = \frac{\pm 1}{\sqrt{m}} \qquad (1)$$

Je zřejmé, že pro dostatečně vysoké $m$ vyjde tento kosínus přibližně roven nule a tedy , diagonální vektory budou přibližně kolmé na všechny souřadné osy. Při použití rozptylových diagramů se pak shluky bodů ležících ve směru těchto diagonálních vektorů promítnou do počátku a nebude je možné je

identifikovat. Naopak shluky bodů ležící přibližně ve směrech jednotlivých os se v rozptylových diagramech promítnou prakticky ne zkresleně.
Dalším problémem souvisejícím s vícerozměrnými daty je fenomén koncentrace objemu, který není zřejmý z klasického 2 3 dimensionální geometrie.
Lze dokázat, že objem hyper-koule poloměru *r* v *m* rozměrném prostoru má tvar

$$Vk = \frac{\pi^{m/2}}{\Gamma(m/2+1)} r^m \qquad (2)$$

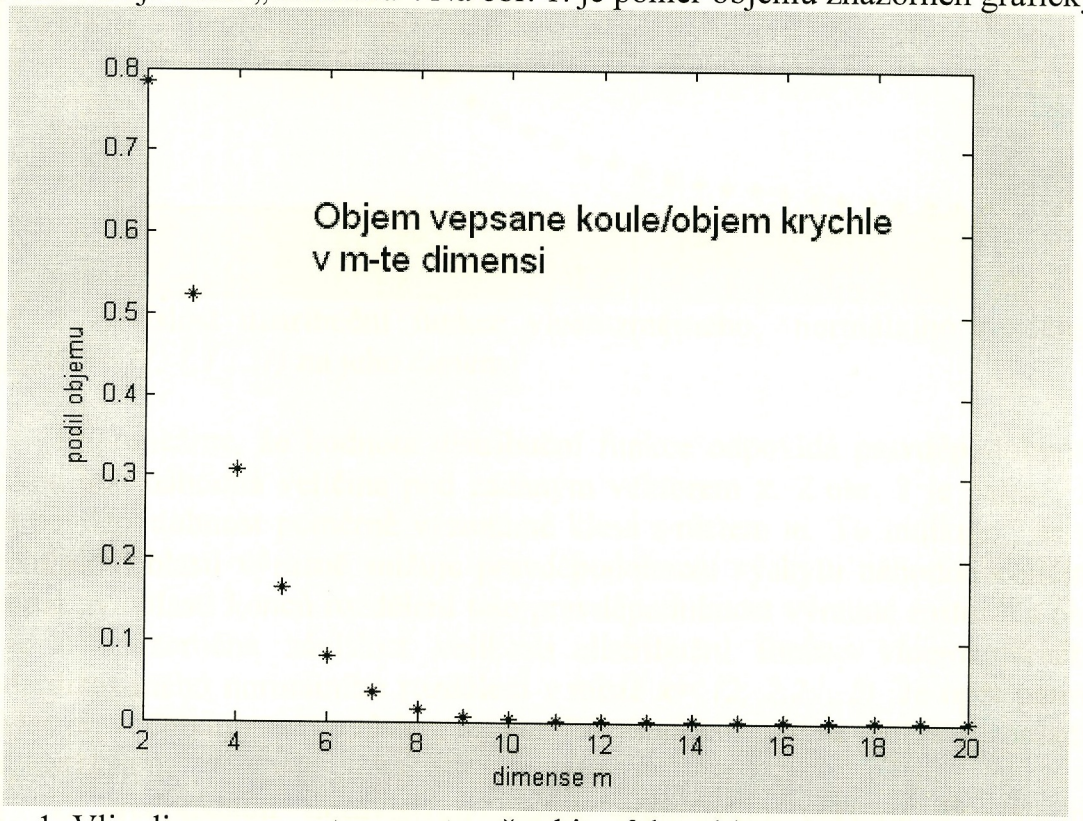
Pro objem hyper kryche o velikosti strany *2r* platí, že

$$Vh = 2^m * r^m \qquad (3)$$

Předpokládejme hyper- kouli vepsanou do hyper-krychle. Poměr objemů pak bude roven

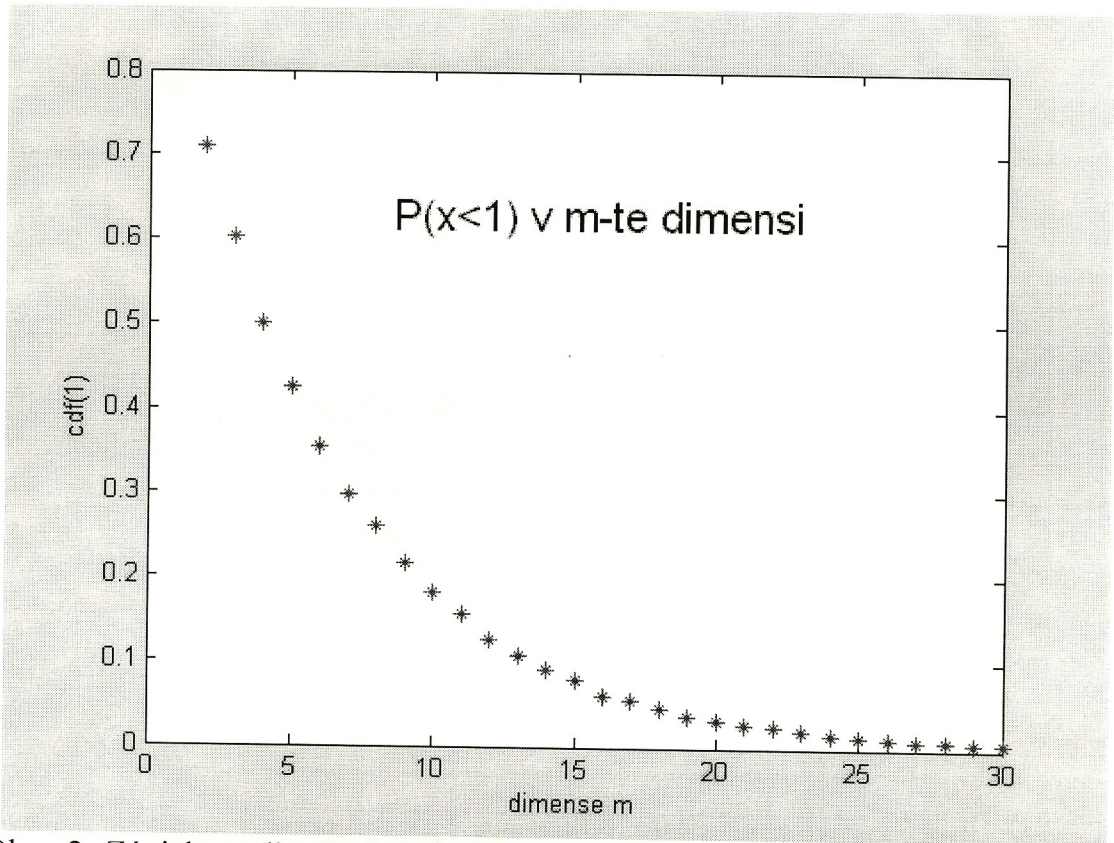
$$\frac{Vk}{Vh} = \frac{\pi^{m/2}}{2^m * \Gamma(m/2+1)} \rightarrow 0 \text{ pro } m \rightarrow \infty \qquad (4)$$

S rostoucí rozměrností se tedy objem hyper-krychle koncentruje v rozích a střední část je málo „obsazená“. Na obr. 1. je poměr objemů znázorněn graficky.




Obr. 1. Vliv dimense prostoru na poměr objemů krychle a vepsané koule.

Je patrné, že už pro $m = 8$ je objem koule zanedbatelně malý vůči objemu krychle. To znamená, že se i data ve vícerozměrném prostoru budou koncentrovat spíše v oblasti konců než středu a bude potřeba mít dostatečně veliké výběry aby se data významnou měrou vyskytovala i o oblastech blízko módu. Dobře e to patrné při použití vícerozměrného normálního rozdělení s nulovým vektorem středních hodnot a jednotkovou kovarianční maticí. Na obr. 2 je znázorněna závislost velikosti distribuční funkce vícerozměrného stadardizovaného normálního rozdělení v místě $\boldsymbol{x}= (1, 1,1....1)$

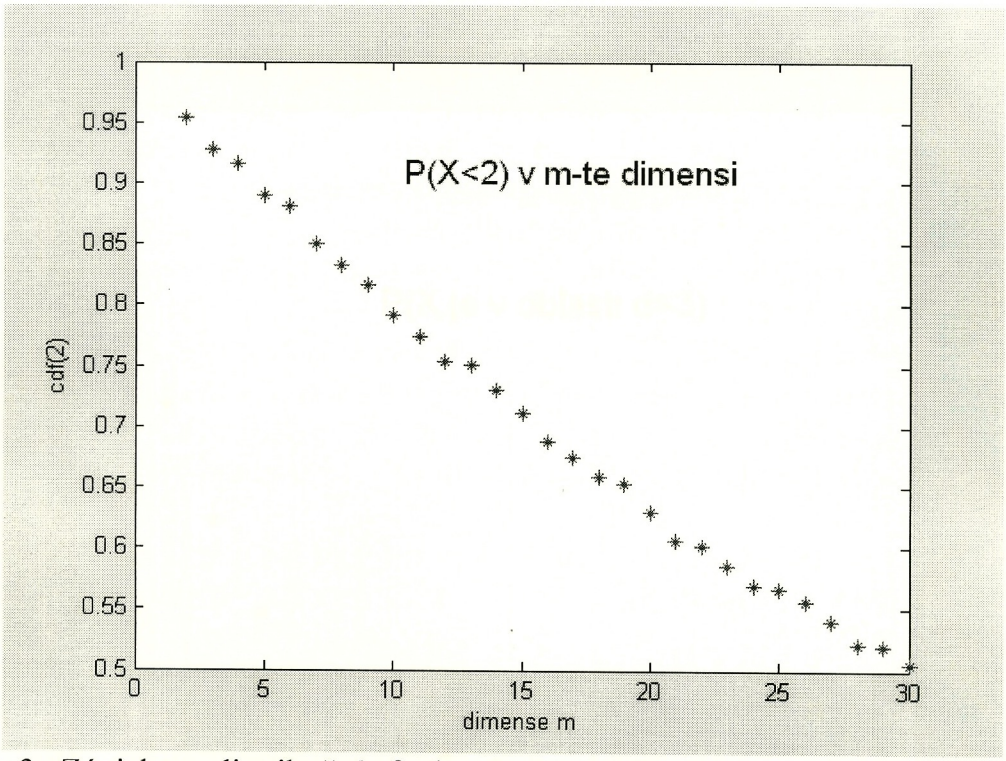



Obr. 2 Závislost distribuční funkce vícerozměrného normálního rozdělení v místě $\boldsymbol{x}= (1, 1,1....1)$ na jeho dimensi

Připomeňme, že hodnota distribuční funkce odpovídá pravděpodobnosti s jakou leží náhodná veličina pod zadaným vektorem $\boldsymbol{x}$. Z obr. 2 je patrné, že tato pavděpodobnost poměrně významně klesá s růstem $m$. To indikuje , že se v centrální oblasti výrazně snižuje pravděpodobnost výskytu náhodné veličiny Naopak, v oblasti konců rozdělení tato pravděpodobnost výrazně roste. Na obr. 3 je e znázorněna závislost velikosti distribuční funkce vícerozměrného stadardizovaného normálního rozdělení v místě $\boldsymbol{x}= (2, 2,2....2)$. Je opět patrný pokles pravděpodobnosti s růstem $m$, který demonstruje, že se objem „hromadí" v oblasti krajů.

Z předpokladů nezávislých složek náhodného vektoru **x** majícího standardizované vícerozměrné normální rozdělení resultuje známý výsledek, že

součet čtverců složek má chí kvadrát rozdělení.




Obr. 3 Závislost distribuční funkce vícerozměrného normálního rozdělení v místě $\boldsymbol{x}$= *(2, 2,2....2)* na jeho dimensi

Tedy

$$\|x\|^2 = \sum_{i=1}^{m} x_i^2 \approx \chi_m^2 \qquad (5)$$

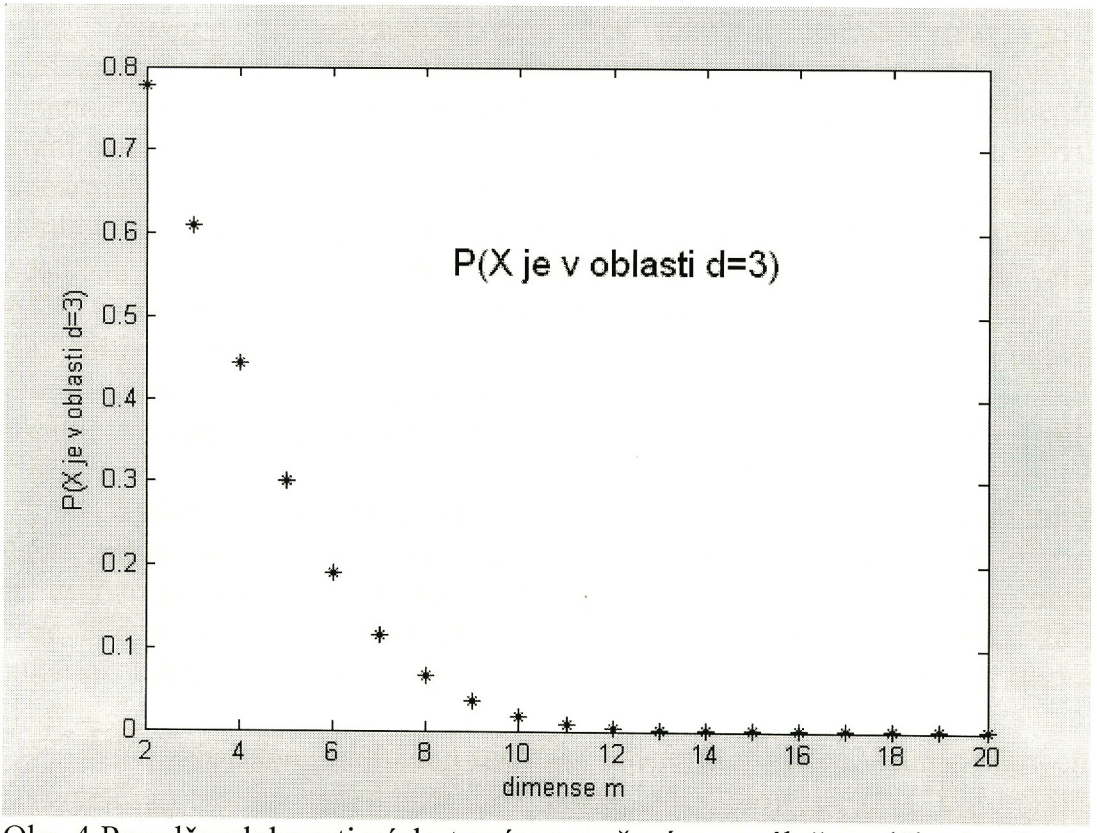
Protože je střední hodnota rovna nulovému vektoru definuje tato norma čtverec vícerozměrné vzdálenosti. Pro výpočet pravděpodobnosti výskytu vícerozměrné náhodné veličiny v hyper kouli s počátkem v nule a poloměrem $r$ pak platí, že.

$$P(\|\mathbf{x}\|^2 \le r) = P(\chi_m^2 \le r) \qquad (6)$$

Z tohoto vztahu lze tedy snadno určit pravděpodobnost s jakou se data budou vyskytovat v centrální oblasti.

Na obr. 4 je znázorněna závislost výskytu vícerozměrné normálně rozdělené náhodné veličiny v oblasti o poloměru 3. na dimensi náhodné veličiny. Je patrný rychlý pokles směrem k nule. To ukazuje, že pro případy od $m = 8$ bude třeba počítat s tím, že výskyt hodnot jednotlivých objektů v oblastí středu je málo pravděpodobný.Může se tedy snadno stát, zejména pro menší $m$,

že celý výběr bude ležet v oblasti konců. Proti běžné intuici zde tedy dochází k výraznému posílení vlivu konců rozdělení na úkor centra.




Obr. 4 Pravděpodobnosti výskytu vícerozměrné, normálně rozdělené náhodné veličiny na hyper kouli o poloměru 3.

Z uvedeného je patrné, že s růstem dimense dochází k paradoxním jevům, které pochopitelně výrazně ovlivní jak průzkumovou analýzu dat, tak i funkci řady statistických metod.

## 4. Redukce počtu proměnných

V řadě případů lze zjistit, že rozměrnost vektoru objektu je silně ovlivněna šumy, nepřesnostmi a nevýznamnými znaky. Pak je logické hledat redukovanou velikost m, tedy omezený počet znaků.Pro případ zobrazení dat se spíše hledá vhodná projekce resp. se původní znaky nahrazují latentními proměnnými, které jsou jejich lineární kombinací ale jsou vzájemně nezávislé.

Na bázi latentních proměnných je založena *metoda hlavních komponent,* která je podrobně popsána v knize [1]. Tato metoda patří mezi lineární projekční metody. Důležité je to, že projekce odpovídá minimu součtu čtverců odchylek od hlavních komponent (latentních proměnných).

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 a jsou vzájemně nekorelované. Latentní proměnné jsou nazvány *hlavními komponentami* a jsou to lineární kombinace původních proměnných: *první hlavní komponenta* $y_1$ popisuje největší část 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. Druhá hlavní komponenta má největší rozptyl mezi zbytkem lineárních kombinací a je kolmá na komponentu první. Analogicky jsou definovány ostatní hlavní komponenty. Pro řešení úlohy hlavních komponent postačuje rozklad kovarianční matice dat na vlastní čísla vlastní vektory. Analýza hlavních komponent je standardní 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.

Obyčejně se pro 2D projekci volí první dvě hlavní komponenty, i když lze často získat zajímavé informace i z posledních hlavních komponent. Výhodou je, že tyto projekce zachovávají vzdálenosti a úhly mezi jednotlivými objekty. Nevýhodou hlavních komponent jako prostředku 2D projekce je fakt, že není nikterak zohledněna potřeba optimální projekce s ohledem na odkrytí struktur v datech.

Tuto nevýhodu odstraňují techniky lineární projekce vícerozměrných dat *(projection pursuit)*, optimalizující zvolený index projekce $IP(\boldsymbol{C}_i)$. Formálně se tedy hledají vektory projekce $\boldsymbol{C}_i$, maximalizující funkci $IP(\boldsymbol{C}_i)$ při omezení $\boldsymbol{C}_i^T\boldsymbol{C}_i = 1$. Projekce na tyto vektory je pak $\boldsymbol{C}_i^T\boldsymbol{X}$. Lze snadno ukázat, že index $IP$, odpovídající metodě hlavních komponent má tvar

$$IP(\mathbf{C}) = \max(\mathbf{C}_i^T\mathbf{S}\mathbf{C}_i) \quad \text{pro } \mathbf{C}_i^T\mathbf{C}_i = 1 \tag{7}$$

kde je $\boldsymbol{S}$ výběrová kovarianční matice. Jak je ukázáno v knize [1]je $\boldsymbol{C}_i$ splňující podmínku maxima rov. (7) vlastním vektorem matice $\boldsymbol{S}$ kterému odpovídá $i$-té největší vlastní číslo $\lambda_i$, $i = 1, 2$. Navíc jsou $\boldsymbol{C}_1$ a $\boldsymbol{C}_2$ortogonální. Lze ukázat [1], že index $IP(\boldsymbol{C})$ odpovídá minimu ze všech projekcí $\boldsymbol{C}$ maxima logaritmu věrohodnostní funkce pro normálně rozdělená data. $N(\boldsymbol{c}^T\mu, \boldsymbol{c}^T\ \boldsymbol{C}\ \boldsymbol{c})$. Tedy za předpokladu normality dat je statisticky odvoditelná jako *optimální* projekce do prvních dvou hlavních komponent.

Častým požadavkem bývá vyhledávání shluků v projekci. Pro tento účel se používá celá řada indexů. Jednoduchý je např. poměr mezi průměrnou meziobjektovou vzdáleností $D$ a průměrnou vzdáleností nejbližších sousedů $d$. Řada indexů využívá odhadu hustoty rozdělení dat v projekci $f_P(x)$. Jako odhad $f_P(x)$ ) se obyčejně volí jádrový odhad hustoty [1]. Často se doporučuje volba

$$IP(\mathbf{C}) = \int f_p^2(x)dx \qquad (8)$$

Tento index však ukazuje spíše na odchylky od parabolického tvaru než na shluky. Odchylky od normality, vyjádřené hustotou pravděpodobnosti $\varphi(x)$ vyjadřuje index

$$IP(\mathbf{C}) = \int \phi(x)[f_p(x) - \phi(x)]^2 dx \qquad (9)$$

Základním problémem při použití těchto indexů je velká časová náročnost výpočtů. Kromě lineárních projekčních metod existuje dnes již celá řada nelineárních projekčních metod [5] Mezi známé patří Kohonenova samoorganizující mapa a generativní topografická mapa, nelineární varianty PCA, založené na použití neuronových sítí. Sammonův algoritmus provádí projekci z původního prostoru do prostoru menšího rozměru tak, aby byly pokud možno zachovány vzdálenosti mezi objekty. Pokud jsou $d_{ij}^*$ vzdálenosti mezi dvěma objekty v původním prostoru a $d_{ij}$ vzdálenosti v redukovaném prostoru je cílová funkce $E$ (která má být minimální) ve tvaru

$$E = \frac{1}{\sum_{i<j} d_{ij}^*} \sum_{i<j} \frac{(d_{ij}^* - d_{ij})^2}{d_{ij}^*} \qquad (10)$$

Pro minimalizaci této funkce se používá iterativní Newtonova metoda nebo řada heuristických optimalizačních postupů. Jak je patrné, souvisí volba *IP* s tím, co se od projekce očekává.

Jedna poměrně jednoduchá projekční technika spočívá v zobecnění PCA, kdy je matice $\boldsymbol{S}$ nahrazena vhodnou robustní verzí $\boldsymbol{S_R}$. To vede k projekci, která je schopna indikovat buď vybočující body nebo shluky v datech. Pro určení $\boldsymbol{S_R}$ je možné použít řady technik, popsaných v knize [1]

## 5. Závěr

Je patrné, že vícerozměrná data mají celou řadu specifických zvláštností. V řadě případů je třeba i ve zdánlivě jednoduchých situacích používat poměrně speciální postupy.

**Poděkování:**
Tato práce vznikla s podporou výzkumného centra Textil LN00B090

## 6. Literatura

[1] Meloun M., Militký J.: *Zpracování experimentálních dat*, East Publishing Praha 1998 resp nové přepracované vydání Academia Praha 2004

[2] Bellman R.: *Adaptive control processes*, Princeton 1961

[3] Silverman B.W.: *Density estimation for statistics and data analysis*, Champman and Hall, London 1986

[4] Spivak M.: *Calculus on Mainfolds*, Addison Wesley 1965

[5] Perpinan M.,A.C.: *A review of dimension reduction techniques,* Tec.h Rept. CS-96-09, Sheffield University, 1996