VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

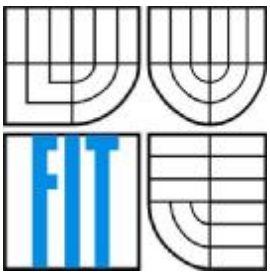
VIZUALIZACE VÍCEROZMĚRNÝCH DAT

BAKALÁŘSKÁ PRÁCE
BACHELOR´S THESIS

AUTOR PRÁCE MAREK JAMBOR
AUTHOR

BRNO 2008

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

VIZUALIZACE VÍCEROZMĚRNÝCH DAT

VISUALIZATION OF MULTIDIMENSIONAL DATA

BAKALÁŘSKÁ PRÁCE
BACHELOR´S THESIS

AUTOR PRÁCE — MAREK JAMBOR
AUTHOR

VEDOUCÍ PRÁCE — ING. MICHAL VYSKOČIL
SUPERVISOR

BRNO 2008

Vysoké učení technické v Brně - Fakulta informačních technologií

Ústav počítačové grafiky a multimédií Akademický rok 2007/2008

Zadání bakalářské práce

Řešitel: **Jambor Marek**
Obor: Informační technologie
Téma: **Vizualizace vícerozměrných dat**
Kategorie: Počítačová grafika

Pokyny:

1. Seznamte se s metodami pro vizualizacemi vícerozměrných dat včetně oblastí a možností jejich využití.
2. Seznamte se s programovacími technikami pro tvorbu těchto vizualizačnic aplikací.
3. Na základě zkušeností vytvořte vizualizační aplikaci
4. Vytvořte testy na různých datových sadách a na jejich základě navrhněte pokračování projektu a vyhodnoťte získané zkušenosti

Literatura:

- Po domluvě s vedoucím

Při obhajobě semestrální části projektu je požadováno:

- 1, 2

Podrobné závazné pokyny pro vypracování bakalářské práce naleznete na adrese http://www.fit.vutbr.cz/info/szz/

Technická zpráva bakalářské práce musí obsahovat formulaci cíle, charakteristiku současného stavu, teoretická a odborná východiska řešených problémů a specifikaci etap (20 až 30% celkového rozsahu technické zprávy).

Student odevzdá v jednom výtisku technickou zprávu a v elektronické podobě zdrojový text technické zprávy, úplnou programovou dokumentaci a zdrojové texty programů. Informace v elektronické podobě budou uloženy na standardním nepřepisovatelném paměťovém médiu (CD-R, DVD-R, apod.), které bude vloženo do písemné zprávy tak, aby nemohlo dojít k jeho ztrátě při běžné manipulaci.

Vedoucí: **Vyskočil Michal, Ing.**, UPGM FIT VUT
Datum zadání: 1. listopadu 2007
Datum odevzdání: 23. ledna 2008

L.S.

doc. Dr. Ing. Pavel Zemčík
vedoucí ústavu

LICENČNÍ SMLOUVA
POSKYTOVANÁ K VÝKONU PRÁVA UŽÍT ŠKOLNÍ DÍLO

uzavřená mezi smluvními stranami

1. Pan

Jméno a příjmení: **Marek Jambor**
Id studenta: 80409
Bytem: Strnadova 2376/11, 628 00 Brno
Narozen: 02. 12. 1984, Brno
(dále jen "autor")

a

2. Vysoké učení technické v Brně

Fakulta informačních technologií
se sídlem Božetěchova 2/1, 612 66 Brno, IČO 00216305
jejímž jménem jedná na základě písemného pověření děkanem fakulty:

.....
(dále jen "nabyvatel")

Článek 1
Specifikace školního díla

1. Předmětem této smlouvy je vysokoškolská kvalifikační práce (VŠKP):
bakalářská práce

Název VŠKP: Vizualizace vícerozměrných dat
Vedoucí/školitel VŠKP: Vyskočil Michal, Ing.
Ústav: Ústav počítačové grafiky a multimédií
Datum obhajoby VŠKP: .....

VŠKP odevzdal autor nabyvateli v:
tištěné formě počet exemplářů: 1
elektronické formě počet exemplářů: 2 (1 ve skladu dokumentů, 1 na CD)

2. Autor prohlašuje, že vytvořil samostatnou vlastní tvůrčí činností dílo shora popsané a specifikované. Autor dále prohlašuje, že při zpracovávání díla se sám nedostal do rozporu s autorským zákonem a předpisy souvisejícími a že je dílo dílem původním.
3. Dílo je chráněno jako dílo dle autorského zákona v platném znění.
4. Autor potvrzuje, že listinná a elektronická verze díla je identická.

Článek 2
Udělení licenčního oprávnění

1. Autor touto smlouvou poskytuje nabyvateli oprávnění (licenci) k výkonu práva uvedené dílo nevýdělečně užít, archivovat a zpřístupnit ke studijním, výukovým a výzkumným účelům včetně pořizování výpisů, opisů a rozmnoženin.
2. Licence je poskytována celosvětově, pro celou dobu trvání autorských a majetkových práv k dílu.
3. Autor souhlasí se zveřejněním díla v databázi přístupné v mezinárodní síti:
   - ☐ ihned po uzavření této smlouvy
   - ☐ 1 rok po uzavření této smlouvy
   - ☐ 3 roky po uzavření této smlouvy
   - ☐ 5 let po uzavření této smlouvy
   - ☐ 10 let po uzavření této smlouvy

   (z důvodu utajení v něm obsažených informací)
4. Nevýdělečné zveřejňování díla nabyvatelem v souladu s ustanovením § 47b zákona č. 111/1998 Sb., v platném znění, nevyžaduje licenci a nabyvatel je k němu povinen a oprávněn ze zákona.

Článek 3
Závěrečná ustanovení

1. Smlouva je sepsána ve třech vyhotoveních s platností originálu, přičemž po jednom vyhotovení obdrží autor a nabyvatel, další vyhotovení je vloženo do VŠKP.
2. Vztahy mezi smluvními stranami vzniklé a neupravené touto smlouvou se řídí autorským zákonem, občanským zákoníkem, vysokoškolským zákonem, zákonem o archivnictví, v platném znění a popř. dalšími právními předpisy.
3. Licenční smlouva byla uzavřena na základě svobodné a pravé vůle smluvních stran, s plným porozuměním jejímu textu i důsledkům, nikoliv v tísni a za nápadně nevýhodných podmínek.
4. Licenční smlouva nabývá platnosti a účinnosti dnem jejího podpisu oběma smluvními stranami.

V Brně dne: .....

.....	.....
Nabyvatel	Autor

Abstrakt

Práce se zabývá problematikou zobrazení n-dimenzionálních dat, pro n>3, která zahrnuje rešerši základních používaných metod, vč.možností a oblastí jejich využití. Dále práce zahrnuje problematiku spojenou s návrhem a implementací modulu pro zobrazení n-dimenzionálních dat. Implementace byla realizována v prostředí programovacího jazyku Java s nadstavbou Java 3D.

Klíčová slova

Vícerozměrná data, nD zobrazení, 3D zobrazení, Java, Java 3D.

Abstract

This work deals with the problems connected to the visualization n-Dimensional data, where n>3, includes the retrieval of elementary methods, incl.field possibilities of an application. Further this work deals with problems connected to the projection and implementation of the module of visualization of n-dimensional data. Implementation was applied in programming environment Java with upgrade Java 3D.

Keywords

Multidimensional data, nD visualization, 3D visualization, Java, Java 3D.

Citace

Jambor Marek: Vizualizace vícerozměrných dat. Brno, 2008, bakalářská práce, FIT VUT v Brně.

Vizualizace vícerozměrných dat

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením Ing.Michala Vyskočila.
Další informace mi poskytl RNDr. Ladislav Mareček, CSc.
Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

……
Marek Jambor
21.1.2008

Poděkování

Rád bych poděkoval svému vedoucímu bakalářské práce Ing.Michalu Vyskočilovi, který ochotně přijal zadání mé práce. Dále bych rád poděkoval RNDr.Ladislavu Marečkovi, který zadání vymyslel, umožnil mi jeho využití jako bakalářské práce a po celou dobu mi poskytoval odbornou pomoc.

Obsah

Úvod

V běžné řeči se slovem statistika často míní znázorňování číselných údajů přehlednou formou. V této podobě se s ní setkáváme např. v médiích v souvislosti s volbami, průzkumy veřejného mínění nebo při zprávách o vývoji ekonomiky. Statistická analýza dat nabývá stále na větším významu a stává se často jedním ze základních přístupů nejen v řadě sociálních, lékařských, technických a přírodovědných věd. Interpretace těchto dat patří mezi neustále se rozvíjející směry zkoumání, ležící na pomezí matematiky a informatiky.

Abychom mohli získaná data nějakým způsobem interpretovat, je nejdříve potřeba tato data analyzovat, odhalit tak jejich zvláštnosti a ověřit předpoklady pro statistické zpracování. Vedle jednorozměrných analytických informací se vyskytují i analytické informace vícerozměrné.

Pro grafickou interpretaci vícerozměrných dat se používá různých technik, umožňující jejich zobrazení ve dvourozměrném a třírozměrném souřadnicovém systému. Takto interpretovaná data nám mnohdy umožňují lépe identifikovat složky, které se jeví jako vybočující a indikovat různé struktury, které ukazují na různorodost výběru nebo přítomnost různých dílčích výběrů s odlišným chováním.

1 Analýza vícerozměrných dat

V praxi se vedle jednorozměrných analytických informací vyskytují i vícerozměrné analytické informace. Příklady vícerozměrných informací jsou:

- vyjádření vlastností produktů potravin, olejů, slitin atd. pomocí rozličných analytických metod
- hodnocení spekter pomocí poloh a velikostí plochy absorpčních pásů pro charakterizaci a identifikaci chemických sloučenin
- sledování složení surovin, produktů, popř. odpadů v závislosti na čase nebo místě výskytu

Na základě provedených analýz je pak k dispozici výběr velikosti n. Tento výběr je tvořen n-ticí vektorů $x_j^T=(x_{j,1},...,x_{j,m})$, které lze chápat jako souřadnice n bodů v m-rozměrném prostoru. Tento výběr lze vyjádřit maticí rozměru $(n\times m)$

$$X=\begin{bmatrix} x_1^T \\ M \\ x_j^T \\ M \\ x_n^T \end{bmatrix}=\begin{bmatrix} x_{1,1} & \cdots & x_{1,i} & \cdots & x_{1,m} \\ M & & M & & M \\ x_{j,1} & \cdots & x_{j,i} & \cdots & x_{j,m} \\ M & & M & & M \\ x_{n,1} & \cdots & x_{n,i} & \cdots & x_{n,m} \end{bmatrix},$$

pro kterou platí, že počet bodů (velikosti výběru) n je větší než počet složek m.

Řádky matice X často představují jisté objekty (vzorky, produkty, odpady, jedince), na kterých se výzkum provádí. Sloupce matice X představují zkoumané znaky, resp. vlastnosti (charakteristiky objektů), které se na objektech zkoumají. Pokud se znaky rozdělují na skupinu vysvětlovaných proměnných (závisle proměnných) a proměnných vysvětlujících (nezávisle proměnných), označuje se submatice vysvětlovaných proměnných jako $Y(n\times p)$ a matici Z rozměru $n\times(m-p)$ pak tvoří skupinu vysvětlujících proměnných. Při dalším strukturování se matice Y dělí do několika skupin vysvětlovaných proměnných $Y_1(n\times p_1),Y_2(n\times p_2),...,Y_o(n\times p_o)$, kde $p=\sum_{i=1}^{o} p_i$. Pokud se pro statistickou analýzu použije celá matice X, jde o tzv. nestrukturovaná dat. Strukturování vícerozměrných dat úzce souvisí s problémy, které jsou pomocí vícerozměrných statistických metod řešeny. Vlastní metody vícerozměrné statistické analýzy také závisí na škále, ve

které jsou data měřena. Podle množství informací obsažených v jednotlivých škálách jsou na nich definovány různé typy operátorů:

Normální škála má zaveden pouze operátor rovnosti $(=)$ nebo nerovnosti $(\neq)$. Rozlišeny jsou tedy pouze různé stavy. Ze statistického hlediska jde o kvalitativní proměnné, které mohou být buď binární (definují přítomnost 1 nebo nepřítomnost 0 nějakého znaku), nebo vícestavové (kódované obyčejně čísly 0, 1, 2, …). Příkladem vícestavové náhodné veličiny je typ katalyzátoru, druh přístroje, barva objektu (pokud se vyjadřuje subjektivně) atd. Tyto vícestavové kvalitativní proměnné se obyčejně převádějí na umělé binární proměnné.

Ordinální škála je škála, kde k operátoru rovnosti a nerovnosti přistupují ještě operátory typu menší $(<)$ nebo větší $(>)$. Tento typ škály se často vyskytuje, když jsou znaky hodnoceny subjektivně a lze provést logické uspořádání do stupnice od nejhoršího k nejlepšímu. Příkladem jsou stupnice stálostí na světě a senzorická hodnocení vzhledu, omaku, vůně atd. Ze statistického hlediska jde o semikvantitativní znaky, kde jsou sice kategorie uspořádány, ale nemají mezi sebou konstantní, resp. měřitelné rozdíly.

Kardinální škála je škála, v níž je zavedena metrika (vzdálenost), takže lze provádět matematické operace jako je sčítání, odečítání, násobení, dělení atd. Kardinální škála se ještě člení na intervalovou škálu a poměrovou škálu. V intervalové škále lze provádět také sčítání a odečítání. Není zde však zaveden přirozený nulový bod. V poměrové škále je možné vyjádřit i poměr mezi objekty (dělení), tj. je zaveden přirozený počátek. Ze statistického hlediska jde buď o diskrétní kvantitativní znaky (kategorizované do disjunktních kategorií), nebo kvalitativní znaky spojité.

Platí, že vyšší typ škály v sobě zahrnuje vlastnosti všech nižších typů a dá se převést (při ztrátě informací) na nižší typ škály. V některých případech je však tento převod z různých důvodů výhodný. Např. pořadová transformace dat před vícerozměrnou statistickou analýzou „přirozeně“ odstraní problém s vybočujícími hodnotami. Na druhé straně mohou vznikat potíže v případech, kdy jsou použité metody založeny na jiných předpokladech (normalitě atd.). Z metrologického hlediska je pochopitelně žádoucí pracovat s daty z poměrové škály obsahující maximum informací.

V současné době existuje celá řada metod vícerozměrné statistické analýzy, které jsou často modifikovány pro speciální účely (taxonometrie, ekologie, časové řady, chemometrie) a v nichž se využívá speciálních zvláštností dat (např. jde o signály prostorově závislá data, respektive časové řady, kde je další proměnnou čas prostorové souřadnice atd.). S ohledem na orientaci v tom, které metody se pro dané účely hodí, je vhodné využít dělení na strukturovaná a nestrukturovaná data.

1.1 Nestrukturovaná data

Výchozí pro vícerozměrnou analýzu jde zde matice $X(n \times m)$, v níž se nepředpokládá žádná speciální struktura sloupci matice X.

1.1.1 Kvantitativní a semikvantitativní data

Podobně jako u jednorozměrných výběrů se zde provádí standardní statistická analýza založená na parametrech plochy (vektoru průměrů) a rozptýlení (kovarianční respektive korelační matici). Zkoumá se přítomnost vybočujících bodů, předpoklady normality a provádějí se standardní statistické testy. Problém, že znaků je více a lze obecně jen těžko použít standardní zobrazovací postupy, se řeší využitím tzv. ordinačních metod. Pro tento typ dat je standardní metodou analýza hlavních komponent. Její základní myšlenka je prostá, spočívá v lineární transformaci původního souřadnicového systému do souřadnicového systému tzv. hlavních komponent, které jsou vzájemně ortogonální (nekorelované) a vybrané tak, aby postihovaly maximální množství informací vyjádřené variabilitou mezi objekty. Relativní pozice objektů zůstává zachována. Nový ortogonální souřadnicový systém je natočen do směrů, které postihují maximální variabilitu minimální vzdálenosti objektů od hlavních komponent. Někdy se hlavní komponenty označují jako faktory, singulární vektory nebo zátěže. Každý objekt má přiřazeny nové souřadnice (projekce do hlavních komponent), které se běžně označují jako skóre. Výsledky analýzy hlavních komponent se často prezentují v grafické formě a slouží buď ke snížení rozměrnosti problému (náhrada původních m znaků menším počtem hlavních komponent, které jsou tvořeny lineární kombinací původních znaků), nebo k zobrazení vícerozměrných dat (projekce do prvních dvou, posledních dvou, resp. jiných kombinací hlavních komponent).

1.1.2 Kvalitativní a semikvalitativní data

Kvalitativní data jsou standardně ve tvaru kontingenčních tabulek (lineární proměnné kódované 0 a 1). Zobecněním analýzy hlavních komponent pro kontingenční tabulky je korespondenční analýza, která využívá ortogonálního rozkladu χ^2-statistiky, vyjadřující míru asociace. Sloupce a řádky u korespondenční analýzy jsou co do informací symetrické a lze je vyjádřit jedním grafem. Korespondenční analýza se také označuje jako duální, popř. optimální, škálování nebo jako reciproké průměrování. Pokud se analyzuje několik binárních proměnných, volí se vícenásobná korespondenční analýza. Pro vyjádření podobnosti, resp. vzdálenosti mezi objekty se používá celá

řada různých koeficientů. Pokud řádky a sloupce matice dat reprezentují stejný objekt, je možné k vyjádření vzdáleností resp. podobností mezi objekty použít vícerozměrné škálování nebo shlukovou analýzu. Metoda vícerozměrného škálování se používá k znázornění objektů na mapě tak, že euklidovská vzdálenost zde odpovídá přibližně původním koeficientům podobnosti resp. vzdálenosti. Klasické vícerozměrné škálování je použito pro vzdálenosti a nemetrické vícerozměrné škálování pro podobnosti. Shluková analýza využívá znázornění ve stromové struktuře (dendogramy).

1.2 Strukturovaná data

1.2.1 Jedna skupina závisle proměnných

Pro tento případ máme výchozí matici závisle proměnných Y rozměru $n \times p$ a nezávisle proměnných Z rozměru $n \times (m-p)$. V případě, že jsou všechny znaky kvantitativní, Z obsahuje nastavované hodnoty a $p=1$, jde o klasickou vícenásobnou regresi. Je-li $p=1$ a Y je binární proměnná, jedná se o logistickou regresi. Je-li $p>1$, jde o vícerozměrovou regresi, která se obyčejně zužuje na vícerozměrnou lineární regresi. Pokud jsou sloupce matice Y ortogonální, čili znaky jsou nekorelované, je možné použít standardní vícenásobné regrese pro každý faktor zvlášť. Pro případ multikolinearity, kdy se vyskytují vysoké korelace mezi faktory v matici Z (např. matice Z obsahuje nadbytečné množství faktorů), používá se celá řada speciálních regresních metod. Metoda parciálních nejmenších čtverců kombinuje analýzu hlavních komponent a vícerozměrnou lineární regresi. Využívá tzv. latentních vektorů (analogie hlavních komponent) k vyjádření jak závisle, tak i nezávisle proměnných. Regrese na hlavních komponentách jako nezávislé proměnné jednotlivé hlavní komponenty. V jistém smyslu inverzní k regresi na hlavních komponentách je tzv. redundantní analýza, v níž se určí hlavní komponenty pro matici Y a příslušné skóry se pak užijí pro sérii vícenásobných regresí. Analogií analýzy rozptylu pro vícerozměrná data je vícenásobná analýza rozptylu. Pro predikci toho,do které ze skupin daný objekt, na základě znaků v matici Z, patří, se volí diskriminační analýza. Skupiny definuje nominální závisle proměnná.

1.2.2 Více skupin závisle proměnných

Pro tento případ je ještě matice Y rozměru $n \times p$ dělena na dílčí matice Y_1 rozměru $n \times p_1$, Y_2 rozměru $n \times p_2$ atd. Kanonická korelační analýza využívá kombinace vektorů $Y_1, Y_2, \ldots Y_o$ k hledání nových proměnných (kanonických proměnných), které mají vyšší korelace. Analogií faktorové analýzy je zde vícerozměrná faktorová analýza. Do této skupiny ještě patří celá řada metod se speciální strukturou dat, jako je PARAFAC, TUCKER3, STATIS. Zajímavá je tzv. Prokrustova analýza. Jejím principem je srovnání tabulek vzdáleností pro stejné objekty. V první fázi se vytvoří MDS a pak se hledají transformace, které přiblíží body na obou mapách co nejblíže k sobě ve smyslu nejmenších čtverců.

Toto dělení metod vícerozměrné statistické analýzy může být sice určitým vodítkem, ale v praxi se často volí různé kombinace podle toho, co se od analýzy očekává.

1.3 Pojem vícerozměrné náhodné veličiny

Vícerozměrná náhodná veličina ξ je jednoznačně určena svou sdruženou distribuční funkcí F(x), která je definována jako pravděpodobnost, že všechny složky ξ_i vektoru ξ budou menší než složky x_i zadaného (náhodného) vektoru x

$$F(x) = P(x_1 \le x_1 \cap x_2 \le x_2 \cap \ldots x_m \le x_m). \qquad (1.1)$$

Symbol $\cap$ označuje logický součin a vyjadřuje současnou platnost uvedených podmínek. Sdružená distribuční funkce F(x) má stejné vlastnosti jako distribuční funkce jedné náhodné veličiny. Je neklesající funkcí svých argumentů, ne nezáporná a maximálně rovna jedné.

Marginální (okrajová) distribuční funkce $F(x_i)$, složky ξ_i, je pak zvláštním případem sdružené distribuční funkce F(x), u které jsou všechny ostatní složky náhodného vektoru na horní mezi svého definičního intervalu, obyčejně $\xi_j = \infty$ pro $j \neq i$.

Speciálním typem rozdělení jsou jednoduché podmíněné distribuční funkce $F(x \mid x_i)$ vyjadřující pravděpodobnost, že všechny složky vektoru ξ kromě i-té budou menší než odpovídající složka vektoru x. Pro složku vektoru ξ_i platí, že je přibližně konstantní, tj. leží v nekonečně malém intervalu $x_i \le \xi_i \le dx + x_i$. Lze tedy psát

$$F(x \mid x_i) = P(x_1 \le x_1 \cap \ldots x_i \le x_i \le (x_i + dx_i) \ldots \cap x_m \le x_m). \qquad (1.2)$$

V případě, že jsou složky vektoru ξ nezávislé, nezávisí podmíněné distribuční funkce na podmínce. Sdružená distribuční funkce se dá pak vyjádřit v jednoduchém součinovém tvaru

$$F(x) = \prod_{i=1}^{m} F(x_i). \tag{1.3}$$

Hlavní roli mezi vícerozměrnými náhodnými rozděleními má vícerozměrné normální rozdělení, jehož sdružená hodnota pravděpodobnosti má tvar

$$f(x) = (2p)^{-m/2} (\det C)^{-1/2} \exp\left(-\frac{1}{2}(x-m)^{\mathrm{T}} C^{-1}(x-m)\right). \tag{1.4}$$

Symbol detC označuje determinant matice Ca symbol x^T označuje transponovaný vektor x. Parametry tohoto rozdělení jsou vektor středních hodnot μ a kovarianční matice C s prvky $C_{ij} = \mathrm{cov}(x_i, x_j)$. Kovarianční matice je obyčejně pozitivně definitivní, takže existuje matice k ní inverzní. K označení inverzního normálního rozdělení se používá symbol N(μ,C).

Pokud vektor x pochází z rozdělení N(μ, C), platí, že veličina (kvadratická forma)

$$Q(x) = (x-m)^{\mathrm{T}} C^{-1}(x-m) \tag{1.5}$$

má rozdělení χ^2-rozdělení s m stupni volnosti. Pro případ, dvou náhodných veličin ξ_1, ξ_2 lze určit, že

$$\det C = s_1^{\,2} s_2^{\,2} (1 - r_{12}^2), \tag{1.6}$$

kde $s_1^{\,2}, s_2^{\,2}$ jsou rozptyly veličin ξ_1 a ξ_2 a ρ_{12} je párový korelační koeficient. Dále platí, že

$$C^{-1} = \begin{bmatrix} \dfrac{1}{s_1^{\,2}(1-r_{12}^2)}, & \dfrac{-r_{12}}{s_1 s_2 (1-r_{12}^2)} \\ \dfrac{-r_{12}}{s_1 s_2 (1-r_{12}^2)}, & \dfrac{1}{s_2^{\,2}(1-r_{12}^2)} \end{bmatrix}. \tag{1.7}$$

Po dosazení do rovnice (1.4) dostáváme sdruženou hustotu pravděpodobnosti *f*(x) ve tvaru

$$f(x_1, x_2) = A \exp\left\{-B\left[\frac{(x_1-m_1)^2}{s_1^{\,2}} - \frac{2r_{12}(x_1-m_1)(x_2-m_2)}{s_1 s_2} + \frac{(x_2-m_2)^2}{s_2^{\,2}}\right]\right\}, \tag{1.8}$$

$$A = \frac{1}{2p s_1 s_2 \sqrt{1-r_{12}^2}}, \qquad B = \frac{1}{2\left(1-r_{12}^2\right)}, \tag{1.9}$$

kde symboly μ_1 a μ_2 označují střední hodnoty náhodné veličiny ξ_1 a ξ_2.

Mezi důležité vlastnosti vícerozměrného normálního rozdělení patří:

a) odpovídající marginální i podmíněná rozdělení jsou také normální,
b) jsou-li všechny složky vektoru ξ vzájemně nekorelované (tj. všechny párové korelační koeficienty jsou nulové), znamená to, že složky ξ_j, j = 1, … , m, jsou nezávislé,

c) pokud má vektor ξ vícerozměrné normální rozdělení, mají libovolné lineární kombinace jeho složek ξ_j také normální rozdělení.

Z uvedeného textu plyne, že předpoklad normality zde usnadňuje analýzu a umožňuje poměrně jednoduché zpracování úloh souvisejících s náhodným vektorem ξ.

1.4 Předúprava vícerozměrných dat

Předúprava dat je důležitý krok v řadě technik vícerozměrné analýzy dat. V některých případech musí být provedena před vlastním použitím metody. V jiných případech tvoří autoškálovací procedura součást metody, takže na vstupu mohou být i neupravená data.

1.4.1 Různé formy standardizace dat

Standardizace dat znamená přiřazení vhodné apriorní důležitosti všem znakům zdrojové matice. Po provedené standardizaci můžeme pomocí vah přiřadit znakům potřebnou důležitost. Standardizace tvoří často první krok v předúpravě vícerozměrných dat. Obecný termín škálování vystihuje, že operace se týká jak měřících jednotek veličin, tak i počátku stupnice. Škálování může být použito na znaky, na objekty nebo na obojí. Škálování by mělo zahrnout:

a) posun počátku souřadného systému,
b) protažení nebo zkrácení měřítka na osách.

Po posunu počátku a centrování sloupců se vzdálenost mezi 2 objekty nezmění. To však neplatí při změně měřítka. Oba znaky před škálováním v prostoru objektů, dobře oddělené, budou po škálování totožné. Z tohoto hlediska mohou být některé škálovací techniky při transformaci dat daleko od reality.

Kromě škálování se také často používá buď logaritmická transformace (eliminace pozitivního zešikmení dat) nebo transformace pořadová, kdy se data nahradí svým pořadím při vzestupném uspořádání. Pořadová transformace je přirozeně robustní, ale za cenu ztráty informace. Výsledky vícerozměrných statistických metod pak mohou být značně odlišné.

Uvedeme nejběžnější škálovací techniky, ve kterých bude y_{ij} představovat transformací upravený, čili škálovaný znak, v j-tém sloupci původního prvku x_{ij}.

1.4.1.1 Sloupcové centrování

Novým počátkem stupnice znaku v j-tém sloupci je průměr prvků znaku $\bar{x}_j$ před jejím centrováním. Sloupcově centrovaná data y_{ij} vzniknou dle vztahu $y_{ij} = x_{ij} - \bar{x}_j$, kde $\bar{x}_j$ je průměr prvků j-tého sloupce vyčíslený dle vztahu

$$\bar{x}_j = \sum_{i=1}^{n} \frac{x_{ij}}{n}. \tag{1.10}$$

1.4.1.2 Sloupcová standardizace

Prvky znaku původních dat v j-tém sloupci x_{ij} jsou děleny svou směrodatnou odchylkou dle vzorce $y_{ij} = x_{ij} / s_j$, kde s_j je směrodatná odchylka střední hodnoty prvků j-tého sloupce vyčíslená dle vztahu

$$s_j = \sqrt{\frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}{n-1}}. \tag{1.11}$$

1.4.1.3 Autoškálování

Je názvem běžně užívaným k označování kombinace sloupcového centrování a sloupcové standardizace. Jde vlastně o tzv. studentizaci

$$y_{ij} = (x_{ij} - \bar{x}_j) / s_j, \tag{1.12}$$

která je analogická Z-transformaci pro velké výběry, kdy předpokládáme, že známe μ_j a σ_j

$$y_{ij} = (x_{ij} - m_j) / s_j. \tag{1.13}$$

Autoškálování užívá odhadů jak střední hodnoty, tak i směrodatné odchylky.

1.4.1.4 Škálování sloupcovým rozsahem

Znaky jsou škálovány, aby bylo získáno minimum každého znaku rovné 0 a maximum rovné 1 dle vztahu

$$y_{ij} = \frac{x_{ij} - \min_j x_{ij}}{\max_j x_{ij} - \min_j x_{ij}}. \tag{1.14}$$

1.4.1.5 Řádkové centrování

Znaky jsou škálovány dle vzorce $y_{ij} = x_{ij} - \bar{x}_i$.

1.4.1.6 Řádková standardizace

Znaky jsou škálovány dle vzorce $y_{ij} = x_{ij} / s_i$.

1.4.1.7 Celkové centrování

Znaky jsou škálovány dle vzorce $y_{ij} = x_{ij} - \bar{x},$ kde $\bar{x}$ je celkový průměr vyčíslený pro celou zdrojovou matici dat o rozměru $n \times m.$

1.4.1.8 Celková standardizace

Znaky jsou škálovány dle vzorce $y_{ij} = x_{ij} / s,$ kde s je směrodatná odchylka od průměru $\bar{x}$ pro všechny prvky zdrojové matice rozměru $n \times m.$

1.4.1.9 Dvojité centrování

Znaky jsou škálovány nejdříve sloupcovým centrováním a následně řádkovým centrováním.

1.4.1.10 Řádkové profily

Znaky jsou škálovány dle vzorce $y_{ij} = x_{ij} / (\bar{x}_i m).$ Tento poněkud zvláštní případ škálování se užívá hodně v chemii, kdy je znak relativní a je vyjádřen v procentech. Součet řádku je pak 1.

1.4.1.11 Sloupcové profily

Znaky jsou škálovány dle vzorce $y_{ij} = x_{ij} / (\bar{x}_j n).$

2 Zobrazení vícerozměrných dat

Pro účely průzkumové analýzy vícerozměrných dat se používá různých technik, umožňujících jejich grafické zobrazení ve dvourozměrném souřadnicovém systému. Toto zobrazení umožňuje:

a) identifikovat vektory x_i nebo jejich složky, které se jeví jako vybočující,
b) identifikovat různé struktury v datech, jako jsou shluky, které ukazují na heterogenitu použitého výběru nebo přítomnost různých dílčích výběrů s odlišným chováním.

Na základě těchto informací a výsledků testů normality (popř. 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.

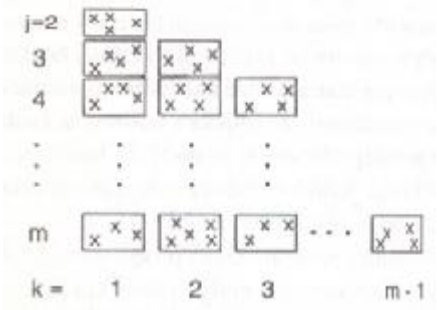
Většina používaných technik pro zobrazení vícerozměrných dat se dá zařadit do jedné ze dvou základních skupin, kterými jsou zobecněné rozptylové diagramy a symbolové grafy.

Pro základní případ dvojice náhodných znaků (m = 2) lze konstruovat rozptylové grafy, které umožňují sledovat statistické zvláštnosti dat. Ke konstrukci výběrového rozdělení nebo jeho porovnání s rozděleními teoretickými je možné použít i histogramy, neparametrické odhady hustoty a jiné grafy. Problémy však nastávají u vícerozměrných dat pro m > 2, kdy je třeba buď volit několik různých grafů, či vhodným způsobem provést transformace na dvoudimenzionální data.

Nelze obecně říci, která metoda zobrazení vícerozměrných dat je nejlepší. Závisí to na počtu znaků m vektorů x_i, i = 1,…n, počtu měření a specifických zvláštnostech dat.

2.1 Zobecněné rozptylové grafy

Pro případ dvou složek x_{i1} a x_{i2} vektorů $\overline{x}_i$ představuje rozptylový diagram (graf) závislost mezi znakem x_{i1} zakresleným na osu x a znakem x_{i2} umisťovaným na osu y. Z takového rozptylového grafu lze snadno identifikovat vybočující (body), struktury v datech (shluky bodů) a míru párové závislosti mezi těmito složkami.



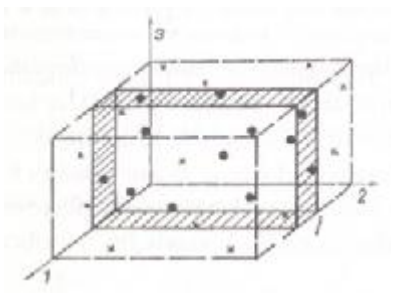
Obr. 2.1 Schematické znázornění párových rozptylových diagramů

Pro případ m-rozměrných dat je nejjednodušší konstruovat rozptylové diagramy pro všechny dvojice složek x_{ij}, x_{ik} vektorů x_i, i = 1…, n. Nejvhodnější je uspořádání těchto diagramů do pole velikosti $(m-1)(m-1)$. V tomto poli je (j, k)-tý rozptylový diagram závislosti složky $x_{i,j+1}$ na x_{ik}.

Vzhledem k symetrii postačuje znázornění pouze $(m-1)\,m/2$ grafů. S růstem m roste počet grafů, a to úměrně s m^2.

Pro větší m (obvykle větším než 10) je již použití této techniky problematické. V takových případech se vyberou pouze některé složky vektorů x_i, pro které se znázorňují rozptylové diagramy.

Pro případ tří složek (m = 3) vektorů x_i je možné rozdělit celou n-tici bodů na několik skupin s ohledem na hodnoty jednoho znaku, a pak pro každou skupinu konstruovat rozptylový diagram. Rozdělení je u diskrétních znaků jednoduché. Pro spojité znaky se postupuje stejně jako při konstrukci histogramu, tj. do j-té skupiny se zařazují všechny body ležící ve zvoleném intervalu. Schematicky je postup znázorněn na obrázku Obr. 2.2. Označuje se jako okénkový graf.



Obr. 2.2 Princip konstrukce okénkového grafu. Kolečka označují body, které leží ve vyšrafovaném j-tém intervalu proměnné

Ke zjednodušení interpretace se často používá rozptylových grafů v modifikovaných souřadnicích, které souvisí s 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 X silně korelovány.

Metoda hlavních komponent patří mezi lineární projekční metody. Důležité je to, že projekce odpovídá minimu součtu čtverců odchylek od hlavních komponent. Obyčejně se pro 2D projekci volí první dvě hlavní komponenty, i když lze často získat zajímavé informace i z posledních hlavních komponent. Nevýhodou hlavních komponent jako prostředku 2D projekce je fakt, že není nikterak vzato v úvahu to, že potřebujeme optimální projekci s ohledem na odkrytí struktur v datech. Tuto nevýhodu odstraňují techniky lineární projekce vícerozměrných dat,

optimalizující zvolený index projekce. Formálně se tedy hlídají vektory projekce C_i, i = 1, 2, maximalizující projekci IP(C_i) při omezení $C_i^T C_i = 1$.

Projekcí na tyto vektory je pak $C_i^T X$. Lze ukázat, že index IP, odpovídající metodě hlavních komponent, má tvar

$$IP(C) = \max(C_i^T S C_i) \text{ při } C_i^T C_i = 1, \qquad (2.1)$$

kde S je výběrová kovarianční matice. C_i (splňující podmínku (2.1)) je vlastním vektorem matice S, kterému odpovídá i-té největší vlastní číslo λ_i, i = 1, 2. Navíc jsou C_1 a C_2 ortogonální. Index IP(C) odpovídá minimum ze všech projekcí C maxima logaritmu věrohodnostní funkce pro normálně rozdělená data $N(c^T m, c^T C c)$. 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 například 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 . Odchylky od normality, charakterizované hustotou pravděpodobnosti Φ(x), vyjadřuje index

$$IP(C) = \int \Phi(x) \left[f_p(x) - \Phi(x) \right]^2 dx. \qquad (2.2)$$

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

$$E = \frac{1}{\sum_{i<j}^{n} d_{ij}^*} \sum_{i<j}^{n} \frac{(d_{ij}^* - d_{ij})^2}{d_{ij}^*}. \qquad (2.3)$$

Pro normalizaci této funkce se používá iterativní Newtonova metoda. Jak je patrné, souvisí volba IP s tím, co se od projekce očekává. Pro optimální projekci se využívá různých typů heuristických optimalizačních algoritmů (generické algoritmy atd.). Jedna technika spočívá ve zobecnění analýzy

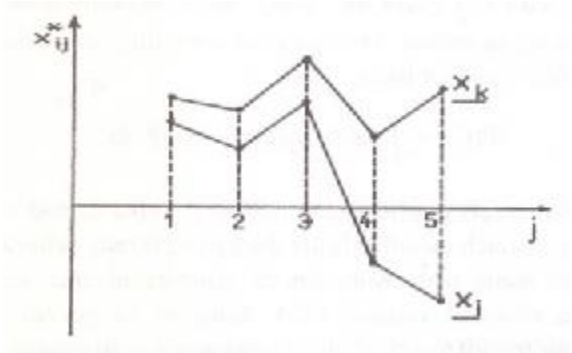
hlavních komponent, kdy je matice S nahrazena vhodnou robustní verzí S_R. To vede k projekci, která je schopna indikovat buď vybočující body nebo shluky v datech.

2.2 Symbolové grafy

Myšlenka použití symbolových grafů je následující. Jednotlivé znaky jsou „kódovány“ s ohledem na jejich konkrétní hodnoty do určitých geometrických tvarů či symbolů. Každému objektu x_i pak odpovídá jistý obrazec složený z těchto symbolů . Vlastnosti dat se posuzují s ohledem na vizuální rozdíly mezi obrazci nebo symboly. Kódování do symbolů je možné užít ke konstrukci rozptylových grafů. Tím lze v jednom grafu rozlišit více znaků x_j, j = 1, …, m. Poměrně jednoduše lze provádět kódování do symbolů pro případ, že složky vektorů x_i nabývají diskrétních hodnot. Jde-li o spojité náhodné veličiny, provede se nejdříve lineární transformace, například do intervalu [0, 1], a ten se pak rozdělí na požadovaný počet úseků

2.2.1 Profily

Profily představují jednoduchou možnost dvourozměrného zobrazení m-rozměrných dat. Každý bod x_i je charakterizován m vertikálními úsečkami nebo sloupci. Jejich velikost je úměrná hodnotě odpovídající složky x_{ij}, j = 1, …, m. Na osu x se vynášejí indexy dané složky j.



Obr. 2.3 Schematické znázornění profilů pro dva objekty x_i, x_k, kde m = 5.

Profil pak vzniká spojením koncových bodů těchto úseček či sloupců. Je vhodné použít škálované znaky

$$x_{ij}^{*} = \frac{x_{ij}}{\max\limits_{i} \left| x_{ij} \right|}, \qquad (2.4)$$

kde $\max\left|x_{ij}\right|$ je maximální hodnota absolutní velikosti složky x_j vektoru x přes všechny body, i = 1, …, n. Schematické znázornění profilu je na obrázku Obr. 2.1.

Profily jsou jednoduché a umožňují určení rozdílů mezi jednotlivými body x_i, x_k i v dílčích složkách. Snadno lze tedy identifikovat vybočující složku objektu, popř.skupiny objektů s téměř shodným chováním.

2.2.2 Polygony

Polygony jsou vlastně profily v polárních souřadnicích. Zde každá složka x_{ij} vektoru x_i odpovídá délce paprsku vycházejícího z jednoho středu. Paprsky jsou rozmístěny ekvidistantně (ve stejných vzdálenostech) na kružnici. Délka j-tého paprsku x_{ij} musí být kladná. Proto se provádí lineární transformace do intervalu [a, 1], kde a je zvolená spodní mez, většinou a = 0. Pro tuto transformaci platí, že

$$x_{ij}^{*} = \frac{(1-a)(x_{ij} - \min\limits_{i} x_{ij})}{\max\limits_{i} x_{ij} - \min\limits_{i} x_{ij}} + a, \qquad (2.5)$$

kde min x_{ij} je minimální a max x_{ij} maximální hodnota j-té složky vektoru x přes všechny objekty x_{ij}, i = 1, …, n.

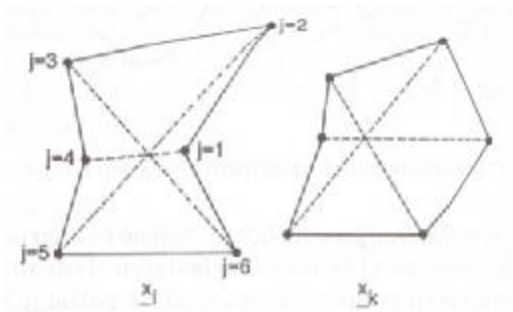
K určení směrů jednotlivých paprsků lze definovat jejich úhel α_j, pro který platí

$$a = \frac{2p\,(j-1)}{m}, \qquad j = 1, \ldots, m. \qquad (2.6)$$

Jako společný střed paprsků se obyčejně volí počátek souřadnic. Pokud má být maximální délka paprsků rovna R (obyčejně R = 1), je polygon pro bod x_i spojnicí m bodů p_{ij} o souřadnicích

$$p_{ij} = \left(x_{ij} R \cos a_j\,, x_{ij} R \sin a_j\right). \qquad (2.7)$$

Aby vznikl uzavřený obrazec, spojuje se ještě bod p_{i1} a p_{im}.

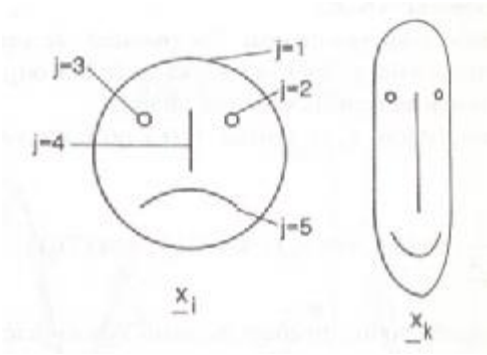


Obr. 2.4 Znázornění polygonu pro dva body x_i, x_k, kdy m = 6.

Schematické znázornění polygonu je uvedeno na obrázku Obr. 2.4. Při interpretaci polygonů se hodnotí jejich podobnost či lokální tvarové změny způsobené hodnotami x_{ij} pro konkrétní j (složka vektoru x).

2.2.3 Tváře

Tváře charakterizují každou složku x_{ij} vektoru x_i nějakým znakem, který je součástí schematizované tváře. Mezi znaky patří tvar tváře, délka nosu, velikost očí, tvar úst apod. Tento typ znaků je znázorněna na obrázku Obr. 2.5. Tvar tváře závisí na použitém pořadí znaků, které ovlivňuje snadnost interpretace dat.

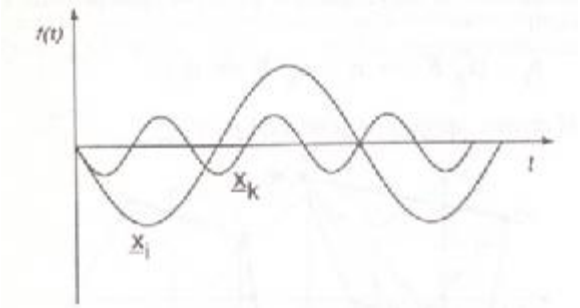


Obr. 2.5 Znázornění tváří pro dva body x_i, x_k, kdy m = 5.

2.2.4 Křivky

Křivky využívají transformace každého objektu x_i na spojitou křivku, která je lineární kombinací všech jeho složek. Pro vyjádření křivky f_i odpovídající objektu x_i volíme konečnou Fourierovu řadu

$$f_{x_i}(t) = f_i = \frac{x_{i1}}{\sqrt{2}} + x_{i2}\sin(t) + x_{i3}\cos(t) + x_{i4}\sin(2t) + x_{i5}\cos(2t) + \dots \ . \qquad (2.8)$$



Obr. 2.6 Schematické znázornění křivek pro body x_i, x_k.

Křivky f_i, i = 1, …, n, se vynášejí jako funkce proměnné t v intervalu $-p \le t \le p$. Použití funkcí f_i, definovaných rovnicí (2.8), má řadu vhodných vlastností:

1 Funkce f_i zachovávají průměr. To znamená, že pokud je průměr $\bar{x}$ z celkem n vícerozměrných dat x_i, je funkce

$$f_{\bar{x}}(t) = \frac{1}{n}\sum_{i=1}^{n} f_{x_i}(t). \qquad (2.9)$$

Funkce $f_{\bar{x}}(t)$ je pak „průměrná“ křivka;

2 Funkce f_i zachovávají vzdálenosti. To znamená, že celková vzdálenost mezi křivkami f_i, f_j, definovaná jako integrální kvadratická odchylka, odpovídá vzdálenosti mezi x_i a x_j. Blízké křivky ukazují na nepříliš vzdálené objekty;

3 Pro zvolenou hodnotu t_0 je funkce $f_{x_i}(t_0)$ projekcí vektoru x_i na p_0 o složkách

$$p_0 = \left(\frac{1}{\sqrt{2}}, \sin(t_0), \cos(t_0), \sin(2t_0), \cos(2t_0), \ldots\right). \qquad (2.10)$$

Tato projekce do jednoho bodu umožňuje odhalení vybočujících bodů či skupin bodů, které mohou být ve více dimenzích špatně identifikovatelné. Křivka $f_{x_i}(t)$ je složena ze všech projekcí na daném intervalu hodnot t;

4 Funkce f_i zachovávají rozptyl. To znamená, že pokud jsou složky x_j vektoru x nekorelované náhodné veličiny s rozptylem σ^2, je

$$D(f_i) = s^2(0.5 + \sin^2(t) + \cos^2(t) + \sin^2(2t) + \cos^2(2t) + \ldots). \qquad (2.11)$$

Pro liché m obdržíme z rovnice $l = \det V_s / \det(V_s + V_c)$ $\mathrm{D}(f_i) = 0.5 s^2 m$ a pro sudé m získáme $0.5 s^2 (m-1) < D(f_i) < 0.5 s^2 (m+1)$. Rozptyl funkce f_i je přibližně konstantní v celém rozmezí veličiny t.

V praktických úlohách je však běžné, že složky vektoru x jsou silně korelované a mají nestejné rozptyly. Pak je výhodné převést vektory původních dat x_i na vektory y_i, kde y_{ij} odpovídá transformaci do j-té hlavní komponenty. Veličiny y_{ij} jsou již nekorelované. Snadno lze provést i jejich standardizaci tak, aby měly konstantní rozptyly.

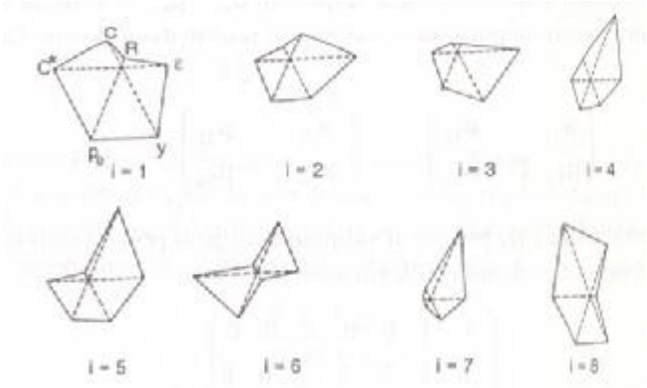
Nevýhodou křivek je to, že jejich tvar závisí na pořadí složek. Na druhou stranu lze snadno indikovat vybočující body nebo skupiny bodů a konstruovat i konfidenční křivky. Schematicky jsou křivky pro dva body znázorněny na obrázku Obr. 2.6.

Pro větší počty bodů (n > 10) dochází ke splývání křivek, což zatěžuje jejich interpretaci. Pak je možné vynášet pouze zvolené podskupiny bodů nebo volit i jiné úpravy.

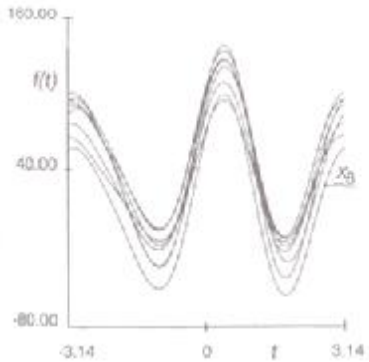
2.2.5 Stromy

Stromy jsou vhodné pro případy, kdy je počet složek m vektoru x veliký. Jednotlivé složky x_j představují délku větví schematického stromu. Jeho struktura, čili rozmístění, větví se volí na základě předběžného hierarchického shlukování znaků (shlukové analýzy).

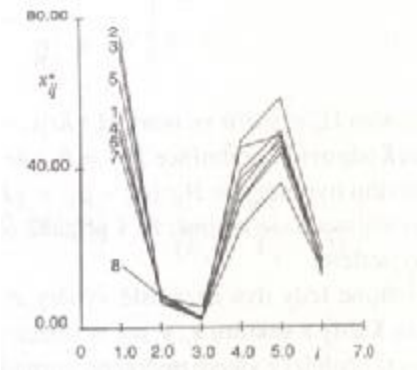
Předběžná shluková analýza se dá použít při výběru pořadí složek vektoru x při konstrukci ostatních symbolových grafů.



Obr. 2.7 Polygony



Obr. 2.8 Křivky



Obr. 2.9 Profily

3 Implementace

Při vývoji modulu bylo potřeba zvážit využití vhodných prostředků pro implementaci. Velký důraz hrála efektivnost, rozšiřitelnost a možnosti programovacího jazyka. V celkové koncepci jsem se zaměřil i na jednoduché a intuitivní uživatelské rozhraní.

3.1 Vývojové prostředky

K implementaci modulu pro vizualizaci vícerozměrných dat byly použity programové prostředky jazyka Java s nadstavbou Java 3D.

3.1.1 Programovací jazyk Java

Java je univerzální (tzn. není určen výhradně pro specifickou aplikační oblast) objektově-orientovaný jazyk se statickou typovou kontrolou, založený na principech C a C++, je však jednodušší než C++. Má méně syntaktických konstrukcí a méně nejednoznačností v návrhu. Program v Javě je meziplatformě přenositelný na úrovni zdrojového i přeloženého kódu.

K Javě je zdarma velké množství knihoven pro různé aplikační oblasti. Kromě knihoven existuje i množství kvalitních vývojových prostředí, komerčních i nekomerčních (např. NetBeans nebo JBuilder).

Využití Javy není pouze pro vícevláknové aplikace nebo přenositelné aplikace s GUI (Graphical User Interface – Grafické uživatelské prostředí), ale také pro výkonné aplikace běžící na serverech (Java Enterprise Edition) nebo přenosných a vestavěných systémech (Java Micro Edition). Java také umožňuje zpracování semistrukturovaných dat (XML), vývoj webových aplikací (servlety, JSP) nebo aplikací distribuovaných po síti (applety, Java Web Start).

3.1.1.1 Java Platforma

Jednou z nejvíce oceňovanou vlastností Javy je plná přenositelnost programů na libovolnou platformu (resp. počítač s operačním systémem) bez nutnosti překladu na této platformě. Této přenositelnosti je dosaženo pomocí bajtkódu (byte-code), jehož interpretace je pak úkolem speciálních programů, nazývaných souhrnně Java platforma, které jsou pro tuto platformu předpřipraveny.

Java se skládá ze dvou hlavních částí, Java Virtual Machine (JVM) a Java Core API. Java Virtual Machine zajišťuje vazbu na hardware a interpretuje bajtkód.

Java Core API obsahuje značné množství knihovních tříd, které jsou považovány za standardní, tzn. musí se vyskytovat v každém prostředí, kde se Java používá.

3.1.2 Nadstavba Java3D

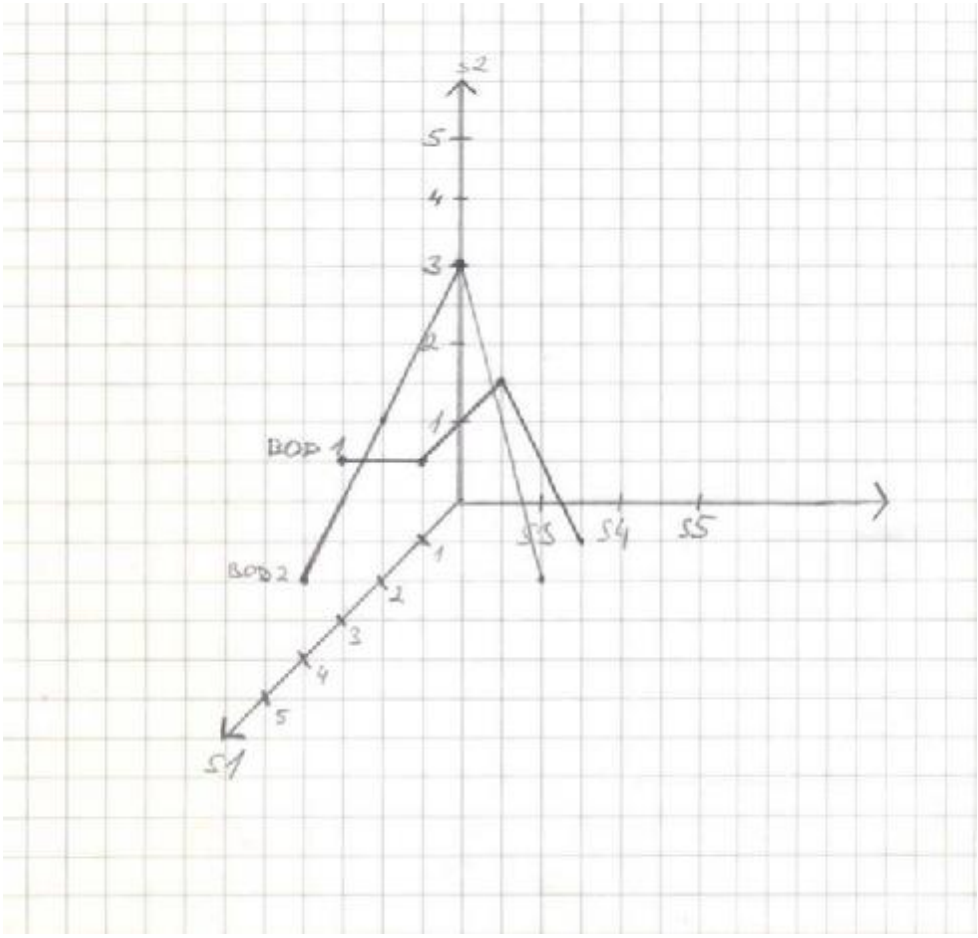
Java3D API je volitelný balíček přímo od společnosti Sun, pomocí kterého Java umí pracovat s 3D grafikou. Pomocí vysokoúrovňových konstrukcí můžeme vytvářet scény s texturami, světly apod. Poskytuje mechanismy pro animování scén a pro definování „chování“ objektů.

Java 3D staví na existujících technologiích jako je DirectX a OpenGL, a také umožňuje začlenit objekty vytvořené 3D modelovacími nástroji jako je TrueSpace nebo VRML modely.

3.2 Metoda zobrazování dat

Data jsou zobrazována do 3D systému souřadnic. Osa 1 (směrem, kterým se díváme – „do obrazovky“) představuje osu X a zobrazuje první souřadnici bodu (x_{i1}). Osa 2 (vertikální) je osou hodnot a zobrazuje souřadnici x_{ij}, j = 2 … n. Osa 3 (horizontální) je osa rovin zobrazení.

Máme-li dva body o n rozměrech, kde n = 5, bod1 má souřadnice {3; 2; 2; 3; 1} a bod2 má souřadnice {4; 1; 3; 5; 2}, můžeme hodnoty převést do n-1 trojic, které zobrazíme do 3D systému souřadnic (Bod1 = [0; 2; 3] [1; 2; 3] [2; 3; 3] [3; 1; 3] a Bod2 = [0; 1; 4] [1; 3; 4] [2; 5; 4] [3; 2; 4]). Tyto trojice zobrazíme na osy v pořadí [Osa 3; Osa 2; Osa 1] a spojíme.



Obr. 3.1 Schematické znázornění dvou objektů Bod1 a Bod2, kde n = 5.

4 Závěr

Zadáním této práce bylo seznámit se se základními používanými metodami pro zobrazení vícerozměrných dat a na základě této rešerše navrhnout a implementovat metodu vizualizace dosud nepopsanou žádnou dostupnou literaturou.

Počínaje definováním specifikace modulu, jeho rozborem, implementací a analýzou jsme se seznámili se skutečnými požadavky spjatými s programovacími technikami pro tvorbu vizualizační aplikace. Jelikož vznikala nová metoda, řešili jsme nejen otázky použitelnosti a jednoduchosti pro koncového uživatele, ale také požadavky na přesnou, odpovídající a výstižnou terminologii. Prošli jsme řadou etap vývoje nejen aplikace, ale i názvosloví, následovanou testováním a laděním. Nakonec se nám požadavky podařilo bez výjimky splnit.

Přínos celé práce pro mě, jako autora, je značný. Bylo mi umožněno seznámení se s problematikou tvorby vizualizční aplikace a jejího možného praktického nasazení.

Jelikož se nejedná o žádný projekt, který bude po dokončení nevyužitý někde ležet, zohlednili jsme i jeho možná rozšíření, mezi která patří především propojení s modulem Regrese, analyzujícím zobrazená data. Především z pohledu komunikace z dalšími moduly je potřeba naši aplikaci dále rozvíjet a zdokonalovat, aby byla schopna plnit požadavky na ni kladené i do budoucna.

Literatura

[1] Herout, P. *Učebnice jazyka Java.* České Budějovice, Kopp 2003.

[2] Hendl, J. *Přehled statistických metod zpracování dat – Analýza a metaanalýza dat.* Praha, Portál 2004.

[3] Meloun M., Militký J. *Statistická analýza experimentálních dat.* 2.vydání, Praha, Academica 2004

[4] *Interval.cz* [online]. Dostupný z WWW: http://interval.cz/

Seznam příloh

Příloha 1. CD

Příloha 2. Uživatelská dokumentace (na přiloženém CD)