Grafická analýza vícerozměrných dat[#]

Miroslav Plašil – Petr Vlach[*]

1. Úvod

Prakticky od svých počátků zahrnují výstupy statistické analýzy dat také prostředky vizualizace (různé typy grafů, tabulek, případně jiných schémat) a i dnes je zapojení grafického aparátu běžnou součástí převážné většiny socioekonomických výzkumů. Zřejmé výhody grafické reprezentace dat shrnul již ve 30. letech dvacátého století významný statistik R. Fisher slovy: *„I když grafy nic nedokazují, pomáhají odhalit význačné vzory v datech“* nebo *„Grafy nejsou náhražkou statistických testů, ale jsou cenné v poznání, kdy které testy použít a v pochopení závěrů založených na těchto testech*.“

Nelze než souhlasit, že využití grafických metod při analýze dat je přínosné, a to zejména v případě hledání základních vlastností dat, jakými jsou existence přirozených shluků, rozdílné vztahy mezi veličinami v různých kategoriích, výskyt odlehlých pozorování, ověřování tvaru rozdělení či vztahu mezi proměnnými. Je však zřejmé, že od Fisherova výroku již uplynula řada let a možnosti moderní vizualizace posunuly původní využití grafů na vyšší analyticko-explorační úroveň.

Zásady moderní vizualizace dat se začínají rodit pravděpodobně v 70. letech minulého století. Přestože jsou samy o sobě velmi jednoduché až triviální, je jejich využití mnohdy velmi efektivní a působivé. Rychlejšímu rozšíření metod do reálných úloh kromě nedostatečné výpočetní techniky bránilo také to, že zásady vizualizace měly spíše podobu filozofických principů, než konkrétně uchopitelných matematických postupů.

Důvodem pro použití grafických nástrojů v analýze dat jsou zejména přirozené a velmi silně rozvinuté vizuálně-kognitivní schopnosti člověka – čtenáře informací. Ten dokáže pomocí *řeči grafických symbolů* porozumět i velmi komplexním problémům, jejichž řešení mu pouhý imaginární prostor čísel může jen stěží zprostředkovat. Neopomenutelná je v tomto smyslu i komunikace mezi zadavatelem analýzy a řešitelem, kdy je nutné zadavateli k jeho pozdějšímu rozhodování poskytnout srozumitelné a interpretovatelné výsledky.

Dalším důvodem rozšíření a rostoucí popularity grafických metod v posledních desetiletích je rozvoj výkonné počítačové techniky. Za zmínku stojí zejména tyto

[#] Článek je zpracován jako jeden z výstupů výzkumného projektu *Nové možnosti využití vícerozměrných statistických metod v ekonomických aplikacích* registrovaného u Interní grantové agentury VŠE pod evidenčním číslem IG410055.

[*] Ing. Miroslav Plašil – doktorand; Katedra statistiky a pravděpodobnosti, Fakulta informatiky a statistiky, VŠE v Praze.

Ing. Petr Vlach – doktorand, Katedra statistiky a pravděpodobnosti, Fakulta informatiky a statistiky, VŠE v Praze, vlach@vse.cz .

aspekty: výkonná výpočetní technika podporuje praktickou aplikaci vizualizačních metod v reálném čase včetně efektivních metod, které jsou zpravidla založeny na poměrně náročné kombinatorické[1] optimalizaci. Dalším aspektem je sama výpočetní technika, která umožňuje i tzv. *dynamickou grafickou analýzu* dat, kdy uživatel pomocí speciálního software dynamicky mění výslednou grafickou prezentaci a pomocí vhodných otázek rychle analyzuje strukturu vícerozměrných dat. K popularitě grafických nástrojů přispívá také velké množství volně dostupných programů, jejichž obsluha není o mnoho náročnější než u běžně používaných kancelářských balíků.

Grafická analýza dat však s sebou nese i určitá omezení. Statistické jednotky (objekty) bývají zpravidla popsány velkým počtem různých vlastností (proměnných). Říkáme, že statistický soubor má vícerozměrnou (vícedimenzionální) povahu. Vzhledem k tomu, že lidské vnímání je omezeno třírozměrným prostorem, musí výsledné grafické zobrazení objektů toto omezení respektovat.

Metody vícerozměrné grafické analýzy dat, kterým se budeme v tomto příspěvku věnovat, se zabývají způsoby, jež činí vícerozměrná data lidskému vnímání sémanticky srozumitelnější. Grafická reprezentace obvykle vychází z různě pojaté *projekce* bodů/objektů z n-rozměrného prostoru do prostoru k-rozměrného, kde k je v optimálním případě rovno dvěma, nebo v méně ideálním případě třem. Kritériem kvality bývá geometrická *interpretace*, ale tento způsob není, jak uvidíme později, jediný možný.

Smyslem článku není systematický popis či didaktické rozdělení grafických metod, ale spíše seznámení s některými zajímavými, přínosnými, ale zatím málo známými moderními metodami, se kterými se čtenář pravděpodobně ještě nesetkal. Jejich představení zároveň čtenáře seznámí se základními myšlenkami a postupy, které jsou typické pro veškerou moderní vizualizaci dat.

Článek je rozdělen na dvě základní části: v první části jsou představeny některé z možností vizualizace vícerozměrných dat, druhá část se zabývá implementací algoritmů podporující optimální řešení.

2. Možnosti vizualizace vícerozměrných dat

2.1 Bertinovy (permutační) matice

Základní myšlenkou jednoho z průkopníků moderní vizualizace dat J. Bertina (viz Bertin, 1967) je, že grafický symbol kromě uchování své hodnoty (velikosti, čísla) umožňuje zároveň její okamžitou analýzu. Bertinova matice v zásadě není nic jiného, než tradiční datová matice, ve které jsou hodnoty čísel nahrazeny grafickým symbolem, nejčastěji velikostí sloupku. Pro snadnější orientaci a čtení grafu jsou navíc sloupky, které mají vyšší než zvolenou prahovou hodnotu (nejčastěji průměr), barevně zvýrazněny (viz obrázek 1). Takto zkonstruovaný graf nejenže nese informaci, která je co do uchování hodnot srovnatelná s informací obsaženou v původní datové matici,

[1] Kombinatorickou optimalizací myslíme nalezení maxima (minima) nespojité účelové funkce, jejíž hodnota je jednoznačně přiřazená každé možné reprezentaci (konfiguraci) dat. Možných konfigurací zpravidla může být už i v případě malých datových souborů velmi mnoho, a to v řádu bilionů i více.

navíc ale dovoluje okamžitou *vizuální analýzu*, která může odhalit dosud skryté vlastnosti dat.

I když je výše popsaný graf úplný, nemusí být ještě zcela přehledný a plně využitelný. Po konstrukci původní Bertinovy matice je nutno obsah informace dále zpřehlednit. Hlavní princip Bertinovy strategie spočívá v přeskupení informace pomocí transformace původní matice do homogennější, interpretovatelné, ale zatím neznámé struktury. Transformační úpravy zahrnují zejména permutace řádků a sloupců. Je zřejmé, že čtení grafu se zásadním způsobem zjednoduší, pokud v něm graficky (tedy i hodnotově) podobné řádky a sloupce leží vedle sebe. Toho lze dosáhnout právě záměnou jejich pořadí. Uvedené úpravy obsah informace tedy nemění, činí ji pouze srozumitelnější.

Různé permutace řádků a sloupců mohou vést k různým uspořádáním matice (nazývaným konfigurace); každé uspořádání přitom může odhalovat jiné vlastnosti dat. Zvolená transformace matice dává odpovědi na různé otázky kladené při zpracování dat – nejsme přitom omezeni pouze jedinou strategií (jinak se uspořádají prvky matice při zkoumání závislostí mezi proměnnými, jinak při hledání přirozených shluků). Velkou předností Bertinových matic je možnost *simultánní analýzy vztahů* mezi případy i proměnnými. Při ní je totiž získána cenná informace o vzájemném propojení řádkových a sloupcových objektů, které při použití tradičních přístupů lze získat jen velmi obtížně. V jednom okamžiku je například získána informace o skupinách respondentů s obdobným nákupním chováním včetně odpovídající množiny nakupovaných výrobků.

Manuální zaměňování řádků a sloupců je pochopitelně u větších úloh nemyslitelné a permutace jsou automatizovány podle předem stanovených algoritmů. Pro bližší seznámení s postupem je nutné grafy Bertinových matic poněkud formalizovat.

Uvažujeme výchozí datovou matici $\mathbf{X}_0$ (n případů a p proměnných). Na základě matice $\mathbf{X}_0$ je získána odpovídající grafická reprezentace. Hodnoty překračující průměr

$$\mathbf{X}_0 [i,j] > \textit{průměr} (\mathbf{X}_0 [.,j]) \tag{1}$$

jsou barevně zvýrazněny. Pro jednoduchost matici $\mathbf{X}_0$ ztotožníme s výchozím grafem (výchozí Bertinovou maticí). Pracujeme-li s výchozím grafem, lze k jeho úpravě použít množinu permutací $\prod$. Velikost množiny $\prod$ je zpravidla dána násobkem počtu řádkových a sloupcových permutací[2] $\prod = \prod_{\textit{řádky}} \cdot \prod_{\textit{sloupce}}$. Symbolem $\mathbf{X}$ je označena matice (a odpovídající graf), která vznikne po provedení permutací: $\mathbf{X} = \pi\mathbf{X}_0$ pro určitou permutaci π, tedy $\mathbf{X}[i,j] = \pi\mathbf{X}_0 [i,j]$.

Vzhledem ke kritériu nalezení homogenní struktury je třeba dále definovat tzv. funkci pročištění (*purity function*) $\Phi = \Phi(\mathbf{X})$, která každé konfiguraci přiřadí hodnotu, která odpovídá *jednoduchosti*, resp. homogenitě konfigurace. Jednoduchost konfigurace může být pochopitelně definována různě, formálně však hledáme takové úpravy, které maximalizují $\Phi(\pi\mathbf{X}_0)$. Volba funkce pročištění závisí na konkrétním cíli a strategii zkoumání. V tomto okamžiku je především důležitá představa, že úpravy počátečního

[2] Jen při úloze o rozsahu 15 pozorování a 5 proměnných dostáváme bezmála neuvěřitelných 157 bilionů možných konfigurací.

grafu do homogennější struktury představují jistý optimalizační problém hledání maxima.

Nejčastěji funkce pročištění porovnává hodnotu prvku matice se sousedními hodnotami, problémem proto zůstávají různé škály měření použité u jednotlivých proměnných. Porovnání mezi případy je snadné, avšak při porovnávání hodnot mezi proměnnými je na místě jistá opatrnost. Před samotnými úpravami pro výchozí datovou matici tedy zpravidla použijeme vhodný typ normalizace.

Některé zvolené strategie odpovídají velmi komplexním funkcím pročištění a vedou k výpočetně velmi náročným optimalizačním algoritmům. Proto se někdy pro jednoduchost omezujeme na méně komplikované funkce, kdy lze problém řazení rozložit samostatně na řádkové a sloupcové permutace – je však nutné si uvědomit, že v tomto případě je uvažována pouze množina permutací[3] $\prod_{řádky} + \prod_{sloupce}$, která je podstatně menší. V následujícím oddíle budou představeny způsoby a algoritmy pokrývající i velmi komplexní funkce pročištění.

V tomto příspěvku jsme se doposud zaměřili na problematiku, jak *mechanickými* úpravami výchozího stavu získat smysluplnější grafickou reprezentaci. Pro konečné příjemce informací je tento aspekt patrně nejdůležitější.

V Bertinových maticích je však v zásadě možná i statistická analýza, i když ta je už poněkud složitější. Je založena na teorii permutací. Může nás například zajímat, zda výsledné grafické zobrazení skutečně odráží nějakou (strukturní) informaci, nebo zda je pouze výrazem nahodilosti. Uvedeme jeden z možných testů:[4] nechť $\Phi^* = \Phi(\pi^*\mathbf{X}_0)$ je maximální hodnota funkce pročištění $\Phi(\pi\mathbf{X}_0)$. Pročištění je statisticky významné na hladině významnosti α, pokud

$$\#\{\pi: \Phi(\pi X_0) \geq \Phi^*\} / \#\Pi \leq \alpha, \tag{2}$$

kde # – představuje počet (konfigurací).

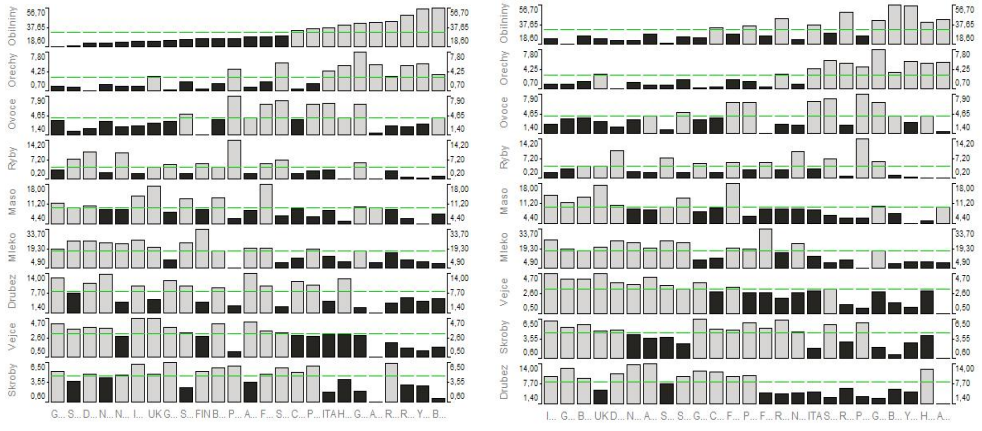
Na závěr dodejme, že možnosti vizualizace pomocí Bertinových matic se v žádném případě neomezují pouze na zobrazování tradiční datové matice nebo určitý druh proměnných. Zobrazitelné a analyzovatelné jsou proměnné prakticky libovolného typu, místo datové matice je možné analyzovat data například ve formě kontingenční tabulky. V tomto ohledu se ukazuje, že Bertinovy matice mohou sloužit jako alternativa, nebo velmi vhodný doplněk korespondenční analýzy. Benzécri například navrhuje uspořádat řádky a sloupce v Bertinově matici podle skóre získaných na základě korespondenční analýzy (viz Benzécri, 1973). V tomto případě jde o alternativní vizualizační nástroj ke klasické korespondenční mapě.

Ilustrativní příklad vychází z dat zachycujících spotřebu devíti potravinových typů ve 25 zemích Evropy (viz Rencher, 2002, s. 483). Smyslem použití Bertinových matic je explorační analýza dat a hledání vztahů v datech.

[3] Ve výše zmíněné úloze o rozsahu 15 pozorování a 5 proměnných se počet permutací, které ve skutečnosti prohledáváme, snížil na méně než jedno procento velikosti původní množiny.

[4] Přes svou teoretickou jednoduchost je jeho praktické využití velmi obtížné.

Obr. č. 1: Bertinova matice



Na obrázku 1 je výsledek dvojího typu (výstup z programu *Visulab*). Schéma v levé části ukazuje, jakým způsobem lze z Bertinovy matice identifikovat korelace – zde mezi proměnnými *ořechy* a *obilniny*, nebo mezi proměnnými *vejce* a *drůbež*, které mají velmi podobný průběh profilů. S růstem hodnot jedné proměnné rostou i hodnoty proměnné druhé. Obdobných vztahů bychom objevili více.

Na schématu vpravo je zobrazena Bertinova matice po permutaci řádků i sloupců, jejíž cílem je dosažení *maximálního stupně homogenity*. V levé dolní části se vytváří skupina zemí, ve kterých převažuje spotřeba *škrobů, vajec, mléka a masa*, v pravé horní části pak nacházíme země, ve kterých převažuje spotřeba *obilnin, ořechů, ovoce a ryb*.

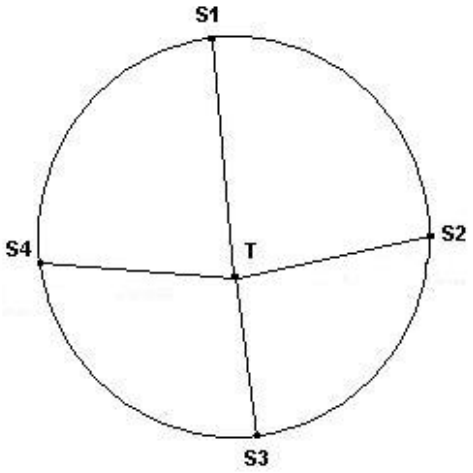
Metoda Bertinových matic je dostupná v několika převážně nekomerčních, volně dostupných grafických nástrojích. Jmenujme například pod Excelem pracující *Visulab* (www.visulab.com) nebo *Voyager* (www.statlab.uni-heidelberg.de/projects/bertin/../voyager/), které lze získat volně na internetu.

2.2 Metoda RADVIZ (Radial Visualization)

K pochopení této metody je vhodné nejdříve použít analogii z fyziky. Představme si p bodů (v případě, že je uvažováno p proměnných), které jsou rovnoměrně rozmístěny po odvodu kružnice. Tyto body označme S_1, S_2, … , S_p. Dále předpokládejme, že je na každý bod zavěšena pružina, jejíž druhý konec je připevněn k tělesu T. Situaci popisuje obrázek 2.

Nakonec předpokládejme, že j-tá pružina přitahuje těleso T k bodu S_j silou x_{ij} ($i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$). Pokud těleso uvolníme a necháme na něj působit pouze síly vyvíjené pružinami, zastaví se nakonec v bodu rovnováhy, ve kterém už nepůsobí žádné dodatečné síly na změnu jeho pozice. Souřadnice pozice rovnováhy $u_i = (u_{i1}, u_{i2})^T$ jsou potom projekcí bodu $(x_{i1}, \ldots, x_{ip})^T$ do dvourozměrného prostoru. Projekci p-rozměrného datového souboru do roviny dosáhneme výpočtem souřadnic u_i pro všechny objekty (p-rozměrná pozorování) $i = 1, 2, \ldots, n$ a následným zakreslením bodů do grafu.

Obr. č. 2: Fyzikální povaha metody RADVIZ



Výpočetní aspekt metody je poměrně snadný: těleso zůstane v rovnovážné pozici, pokud se síly, které na něj působí, vzájemně vykompenzují. Z fyziky je známo, že síla je dána jako násobek směru (vektoru) jejího působení a koeficientu tuhosti. Aby byl jejich součet nulový, musí zřejmě platit:

$$\sum_{j=1,p} \left(S_j - u_i \right) x_{ij} = 0. \tag{3}$$

Proměnné v závorce udávají směr síly, člen x_{ij} pak její velikost. Řešením pro u_i dostáváme:

$$u_i = \sum_{j=1,p} w_{ij} S_j, \tag{4}$$

kde S_j jsou souřadnice jednotlivých bodů S_j a

$$w_{ij} = \left(\sum_{j=1,p} x_{ij} \right)^{-1} x_{ij}. \tag{5}$$

Na výslednou projekci bodu do dvourozměrného prostoru můžeme tedy pohlížet jako na vážený průměr S, kde váhy představuje p normalizovaných proměnných. Tato normalizace přitom činí projekci nelineární.

Z principu tohoto typu nelineární projekce lze vyvodit některé zajímavé vlastnosti. Z hlediska běžného uživatele jsou zřejmě nejdůležitější tyto:

- Body, které mají relativně stejné hodnoty ve všech dimenzích, budou ležet blízko středu kruhu. Tato vlastnost je poněkud zvláštní, protože blízko středu budou ležet jak body, které mají ve všech dimenzích nízké hodnoty, tak body, které mají ve všech dimenzích hodnoty relativně stejně vysoké. To může být v určitých případech velmi nevýhodné – tuto vlastnost lze do jisté míry kompenzovat použitím barevného rozlišení.

- Body mající obdobné hodnoty u proměnných ležících na kruhu proti sobě, budou zobrazeny v blízkosti středu kruhu. Z této vlastnosti vyplývá, že rozmístění (pořadí) proměnných podél kruhu má na výslednou konfiguraci bodů v rovině zásadní vliv. Ne všechny projekce jsou proto stejně zajímavé a vhodnou strategií je nutné vybrat optimální.
- Body, které dosahují vyšších hodnot než ostatní v jedné nebo dvou dimenzích, budou ležet blíže těmto dimenzím umístěným po obvodu kruhu.

2.3 Metoda Projection pursuit

Metody explorační analýzy prováděné nástroji vícerozměrné statistiky s sebou přináší omezení vyplývající z metody použitých datových vstupů. Mnoho pokročilejších vícerozměrných metod například vychází z korelační struktury – ty jsou pak méně vhodné např. při posuzování existence shluků. Obecnější metody, jakou je *projection pursuit*, dovolují individuální nastavení požadované struktury u každé úlohy.

Techniku metody jako první navrhl Kruskal (1969, 1972), později se metoda objevuje i ve studiích dalších autorů jako např. Switzer (1970), nebo Friedman a Tukey (1974). Posledně jmenovaní se zasloužili o její první implementaci a dali metodě i její název.

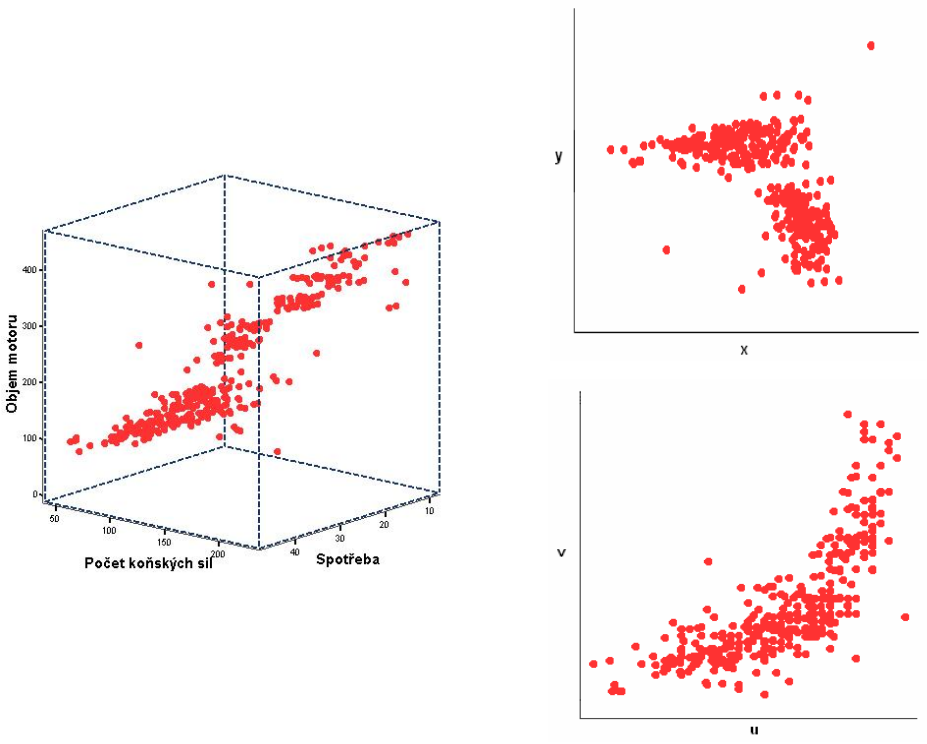
Projection pursuit je původně explorační metoda, ve které je hledána optimální projekce bodů do prostoru nižší dimenze tak, aby bylo možné zobrazit určitou strukturu v datech. Uveďme příklad: Na obrázku 3 jsou data (výukový soubor *Cars* z programu SPSS) zobrazena v prostoru dimenze R_3. Projekcí do roviny získáme lepší pohled na jejich strukturu, např. ve smyslu lepší separace objektů do shluků. Na obrázku jsou zobrazeny projekce stejných objektů do roviny definované nejprve osami *x*-*y* a potom osami *u*-*v*. V prvním případě je zřejmá separace objektů do dvou skupin, ve druhém případě vyniká silná pozitivní závislost. Různé projekce tak odhalí různé aspekty datové struktury. Nutno upozornit, že v mnoha případech nebude zřejmá struktura žádná. Teoreticky však existuje nekonečně mnoho projekcí a smyslem metody projection pursuit je nalézt projekci optimální (nejlepší).

Každá projekce je podřízena určitému kritériu, které určuje její smysl. Kritérium je definováno speciální funkcí, kterou nazýváme *index projekce*. Index projekce může například nabývat vysokých hodnot v případě objevení zajímavých struktur a nízkých hodnot v ostatních případech. Projekce optimální pak bude mít hodnotu indexu projekce nejvyšší (maximální).

Hledání maxima funkce vzhledem k omezujícím podmínkám je zpravidla řešeno iteračními metodami. Z výpočetního hlediska uživatel zadává vstupní parametry, případně podmínky pro konvergenci. Při postupném hledání globálních extrémů závisí kvalita řešení na vhodném nastavení výchozích podmínek konvergence tak, aby nedošlo k dosažení lokálních extrémů. V metodě projection pursuit však mohou i lokální extrémy reprezentovat zajímavá řešení.

Aplikační rozmach a další vývoj nastal v souvislosti s rozvojem výpočetní techniky. Exploračí použití metody bylo rozšířeno o nástroje *Projection Pursuit Regression* (PPR), *Projection Pursuit Classification* (PPC) a *Projection Pursuit Density Estimation* (PPDE).

Obr. č. 3: Zobrazení objektů v prostoru R_3 a v rovině



Pro nalezení optimální projekce je třeba vyjádřit optimálnost numericky. Z koncepčního hlediska se odlišují dva pohledy: (1) abstraktní pohled pracující s *p*-rozměrným náhodným vektorem daným sdruženou hustotou pravděpodobnosti a (2) aplikační pohled vycházející z výběrových dat. Obě verze mohou být do značné míry identické, abstraktní pohled však pracuje s hladkými funkcemi.

Výpočetní hledisko metody a použitá symbolika vychází z původních prací (viz Huber, 1985, s. 439). Projekce z *p*-rozměrného prostoru R_p do *k*-rozměrného prostoru R_k je vždy lineární, nelineární struktury jsou proto metodou jen těžko zachytitelné. Označíme-li $\mathbf{X}=\{x_{ij}\}$ *p*-rozměrný náhodný vektor, resp. datovou matici typu *n* x *p* reprezentující náhodný výběr, zapíšeme lineární projekci z R_p do R_k jako

$$\mathbf{Z} = \mathbf{A}\mathbf{X}^T, \ \mathbf{X} \in \mathbf{R}_p, \mathbf{Z} \in \mathbf{R}_k, \tag{6}$$

kde **A** – matice typu *k* x *p* definující lineární transformaci,
Z – matice typu *k* x *n* reprezentující zobrazení projektovaných bodů.

Hledání projekce spočívá v nalezení prvků matice **A**. Pokud jsou prvky matice **A** ortogonální, potom hovoříme o ortogonální projekci. U jednorozměrných projekcí pracujeme nikoli s maticí **A**, ale s vektorem **a** typu 1 x *n*.

Dále je třeba definovat hledanou vlastnost dat – závislost, podobnost, atd. Index projekce měří sílu hledané vlastnosti v aktuálně projektovaných datech, označujeme jej symbolem *I* a vzhledem k (6) píšeme

$$I(\mathbf{Z}) = I(\mathbf{A}\mathbf{X}^T) = I(\mathbf{A}). \tag{7}$$

Vycházíme-li z předpokladu známé sdružené hustoty pravděpodobnosti f vektoru **Z**, potom tato hustota závisí na **A**. Při řešení úloh je třeba stanovit index projekce, který je funkcí hustoty **Z** a proto může být projekční index zapsán také výrazem $I(f)$. Pracujeme-li s výběrem, nahradíme hustotou f jejím výběrovým odhadem.

Ze známých indexů projekce podrobněji uvádíme jen původní indexy *Tukey-Friedmanův* a *Jones-Sibsonův*. S ostatními se čtenář blíže seznámí v referenční literatuře.

1. Tukey-Friedmanův index projekce

První dva koncepty indexu projekce pro případy jednorozměrných a dvourozměrných projekcí navrhli ve své práci Tukey s Friedmanem v roce 1974. Oba indexy měří míru shody dosažené lineární projekce s požadovanou strukturou. K optimalizaci používají iterační algoritmus *hill-climbing*. Pro jednorozměrnou projekci navrhli index

$$I(\mathbf{a}) = s(a)d(a), \tag{8}$$

kde $s(a)$ – měří obecné rozptýlení hodnot,
$d(a)$ – měří lokální hustotu dat po projekci na vektor **a**.

2. Jones-Sibsonův index projekce

Na rozdíl od Tukeyho a Friedmana spočívá index Jonese a Sibsona v maximalizaci divergence cílové projekce od *nezajímavých* projekcí. Provedli podrobnou analýzu indexu (8) a zjistili, že *d(a)* je odhadem

$$\int f(x)^2 dx, \tag{9}$$

kde f – hustota pravděpodobnosti projektovaných dat.

Již dříve Hodges a Lehman zjistili (viz Hodges – Lehman, 1956, s. 324–335), že funkcionál (9) lze minimalizovat pouze použitím parabolických hustot (mezi všemi hustotami s nulovou střední hustotou a jednotkovým rozptylem). Dále ukázali, že Tukey-Friedmanův index (8) není invariantní vůči rotaci souřadného systému. To znamená, že k jeho změnám nedochází jen jinak provedenou projekcí, ale zároveň i změnou v nastavení souřadného systému. To s sebou přináší problémy hlavně při snaze o vzájemné porovnání více řešení.

Jones se Sibsonem uvádí (viz Jones – Sibson, 1987, s. 1–36), že (9) je monotónní funkcí míry entropie druhého řádu. Míry entropie řádu α zavedl A. Rényi (1964), Jones se Sibsonem navrhují míru entropie prvního řádu jako základ pro index projekce ve tvaru

$$-\int f(x) \log f(x) dx, \tag{10}$$

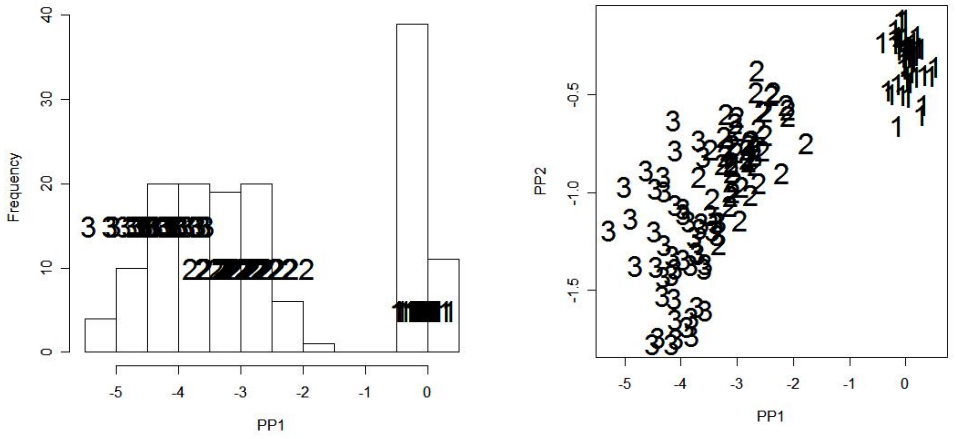
která je známá také jako negativní Shannonova entropie. Její užitečnost spočívá v tom, že svého maxima dosahuje při použití hustoty normálního rozdělení, jehož použití je mnohem příjemnější než použití hustot parabolického typu v indexu (9). Při výpočtu je třeba odhadnout empirickou hustou pravděpodobnosti *est*(*f*) a numericky integrovat výraz (10).

Z dalších indexů uveďme heslovitě: *Hallův index*, *Mortonův index*, *Posseho index* nebo *index Yeniukovův*. Podrobný výklad uvedených indexů v literatuře (viz např. Nason, 1992).

Projection pursuit dnes nachází uplatnění v celé řadě zajímavých úloh: od klasifikačních a regresních úloh po hledání odlehlých pozorování či objevování skrytých závislostí. Metoda se zatím objevuje v softwarových produktech spíše nekomerčního charakteru, jakými jsou například originální zpracování ve FORTRAN autorů Jonese a Sibsona, nebo jako součást grafického balíku *XGobi*. Dostupná je rovněž ve volně dostupném programu *R*.

Příklad: použitím metody projection pursuit při klasifikaci objektů bylo cílem nalezení projekce, která maximálně separuje skupiny do tříd. Data pochází ze známého Fisherova souboru *Irisdata*, často používaném při výkladu diskriminační analýzy. K výpočtu byl použit systém R a knihovna *classPP* (*classification Projection Pursuit*). Projekce do jednorozměrného a dvourozměrného grafu je pak na obrázku 4.

Obr. č. 4: Projection pursuit



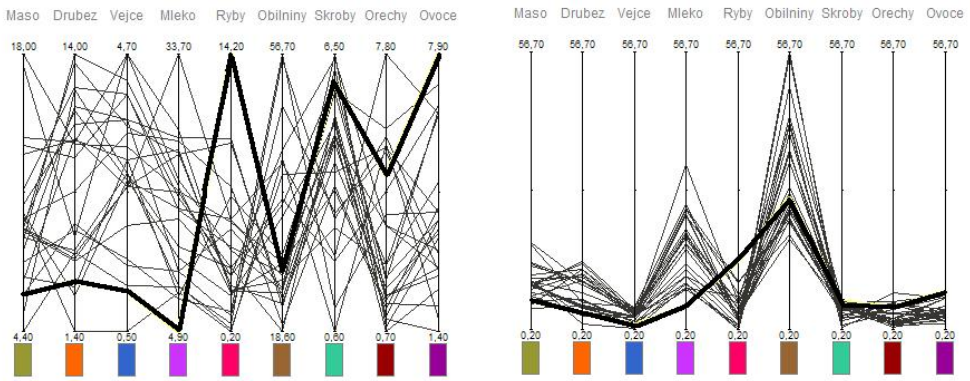
2.4 Paralelní osy

Paralelní osy (*Parallel coordinates*) jsou poměrně mladou grafickou metodou, kterou zavedl v 80. letech minulého století A. Inselberg. Svou jednoduchostí a minimálními nároky na zpracování se řadí mezi nejintuitivnější nástroje pro zpracování dat.

Hlavní myšlenou metody je projekce *n*-rozměrného vektoru pozorování do rovinného diagramu. Na rozdíl od tradičního souřadného systému kartézského jsou jednotlivé osy zobrazeny paralelně vedle sebe. Jsou-li hodnoty proměnných zapsány formou *n*-složkového vektoru $\mathbf{x} = (x_1, x_2, ..., x_n)$, potom je odpovídající grafická reprezentace formou paralelních os tvořena spojnicí *n* paralelních os, které jsou spojeny v bodech $\{1, x_1\}$, $\{2, x_2\}$, ..., $\{n, x_n\}$. Každá linka představuje jeden bod ve vícerozměrném prostoru a může být proto chápána jako *profil* odpovídajícího případu. Konkrétní podoba profilu s sebou obvykle nese dostatek informace k učinění závěru o obecné struktuře dat.

Na obrázku 5 je ukázka grafu paralelních os, vytvořená ve statistickém systému *Visulab*. Data vychází opět z příkladu použitého výše. Jednotlivé osy tak charakterizující míru spotřeby devíti druhů potravin v 25 zemích Evropy. V obrázku je zvýrazněný průběh jedné z linek (*Portugalsko*) pro dosažení lepší čitelnosti odpovídajícího případu. Na první pohled je vidět, ve kterých složkách potravin je spotřeba v porovnání s ostatními zeměmi nízká (*mléko*, *vejce*) a kde naopak vysoká (*ryby*). Obrázek v levé části používá u všech os původní jednotky měření, osy v grafu vpravo jsou normalizovány na délku *min – max*.

Obr. č. 5: Paralelní osy



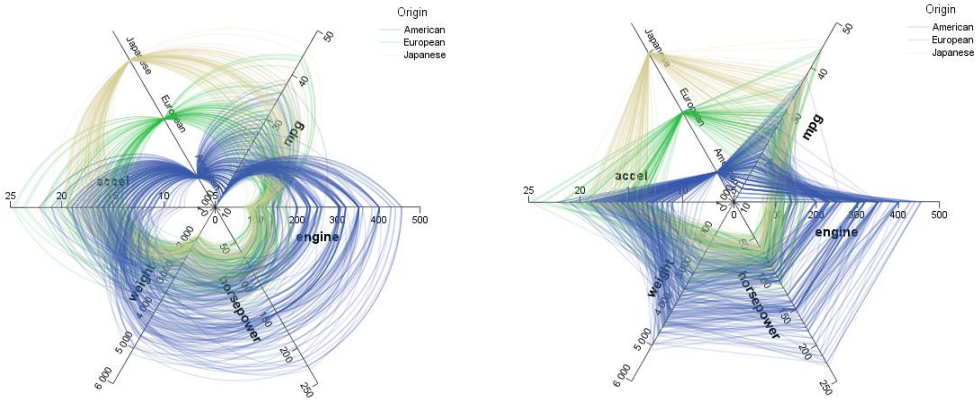
Paralelní osy slouží jako užitečný nástroj pro detekci odlehlých pozorování. Ty identifikujeme podle podezřelého nebo unikátního průběhu jednoho nebo více profilů. Dále metodu oceníme při zjišťování vztahů mezi proměnnými nebo při hledání interpretovatelných shluků, které poznáme podle většího počtu linek s podobným průběhem. Počet úloh je jistě větší, jmenujme například problém diskriminace mezi dvěma či více skupinami.

Metoda má stejně jako ostatní i své omezující faktory. Jedním je počet případů. S růstem velikosti datové matice se stává struktura stále více nepřehledná a při vizuální analýze může být pokus o objevení zajímavých struktur obtížný. Je proto vhodné používat interaktivní prostředí, které dovoluje zvýrazňovat skupinky linek se stejným chováním, měnit počet zobrazených dimenzí či jinou úpravu vizuálního prostředí.

Softwarová podpora je v případě paralelních os poměrně silná. K dispozici je celá řada nekomerčních produktů, které metodu nabízí zdarma. Jmenujme například známý systém *R* (*www.r-project.org*), již zmiňovaný *Visulab* a další. Obliba metody je ale zřejmá i u profesionálů. Např. nám dostupný systém SPSS (*www.spss.cz*) obsahuje paralelní osy včetně polárního zobrazení. Na rozdíl od paralelního uspořádání mají osy počátek ve stejném bodě a nabízí tím jiný náhled na data.

Na obrázku 6 je znázornění ilustračních dat (soubor *Cars*) v systému SPSS s využitím polárních os. V obou obrázcích vynikají tři kategorie objektů, které lze identifikovat podobným průběhem profilů. Zároveň jsou zřejmé nejvíce a nejméně diskriminující proměnné.

Obr. č. 6: Polární osy



3. Algoritmy pro vizualizaci dat

Ve druhém oddíle jsme se zaměřili na principy různých grafických metod a jejich praktickou stránkou jsme se často podrobněji nezabývali. Nutno bohužel poznamenat, že ani dostupný software se touto oblastí příliš nezabývá, nebo jen velmi okrajově a nedostatečně. Přesto je tento aspekt velmi důležitý, protože určuje kvalitu zobrazené informace ve výsledné konfiguraci. Je tomu tak proto, že smyslem většiny výše uvedených grafických metod je data pouze reprezentovat, nikoli však vzhledem k nějakému konkrétnímu (maximalizačnímu) kritériu. To sice dává metodám velkou obecnost, protože lze od počátku sledovat více strategií (více aspektů dat), zároveň však tato volnost přestavuje značnou výpočetní zátěž, protože optimalizace v rámci grafických metod je mnohem náročnější než u tradičních metod vícerozměrné statistiky. U grafických metod nejsou účelové funkce zpravidla spojité, a tak musíme každou konfiguraci ohodnotit samostatně a potom vybrat tu nejlepší vzhledem ke sledovanému cíli. Možných konfigurací přitom nezřídka bývá řádově několik set milionů, někdy až set bilionů či více.

Připomeňme, že v Bertinových maticích je nutné řádky a sloupce nejdříve uspořádat, abychom získali pro zvolenou strategii optimální konfiguraci. Stejně tak v metodách RADVIZ a paralelní osy hraje uspořádání proměnných klíčovou roli. V prvně jmenované metodě má uspořádání proměnných na konfiguraci vliv vzhledem k povaze projekce, u druhé metody optimální pořadí proměnných snižuje nepřehlednost grafu v případě velkých datových souborů a zároveň umožňuje lépe zkoumat vztahy mezi nimi – lépe se zkoumají proměnné, které leží v paralelních osách vedle sebe.

Manuální uspořádání je při úlohách většího rozsahu nepraktické a vzhledem k potřebnému času také zcela neefektivní. Základním požadavkem na zpracování dat je proto automatizovaný průběh podle dostatečně výkonných algoritmů, uživatel pouze nastavuje svoje požadavky a klade datům správné otázky. V tomto článku je blíže popsáno použití *evolučních algoritmů* (EA). Pro jednoduché účelové funkce existují i jiné a snadnější možnosti, ale evoluční algoritmy mají výhodu velké univerzálnosti (lze je v modifikacích použít na všechny výše zmíněné metody) a zároveň relativní efektivnosti. V dalším textu se proto zaměříme výhradně na ně.

Evoluční algoritmy (viz např. Holland, 1992 nebo Kvasnička – Pospíchal – Tiňo, 2000) představují silný optimalizační nástroj v úlohách, kdy neexistuje specifická metoda nalezení optima a kdy ani případné heuristické postupy nenabízí uspokojivá řešení. Jejich koncepce je velmi flexibilní a obecná, proto nachází uplatnění při řešení různých typů úloh. Z důvodu místa představíme pouze základy teorie EA, které jsou nutné k pochopení podstaty algoritmů a jejich aplikace na nalezení zajímavé konfigurace. Obecně není nalezení globálního maxima účelové funkce pomocí EA zaručeno, ale v případě grafických metod jsou pro další analýzu často zajímavá i suboptimální řešení (konfigurace).

Hlavní myšlenkou EA je napodobení evolučního procesu a uplatnění přírodních zákonů křížení, mutace a selekce jedinců, kdy jedinci zastupují jednotlivá přípustná řešení úlohy. Provádění operací křížení a mutací vede k vytvoření nových jedinců, kteří dále soutěží s původními o přežití a místo v další generaci. Slabší jedinci jsou v procesu postupně nahrazováni silnějšími, perspektivní jedinci jsou dále šlechtěni, aby vznikly jedinci (přípustná řešení) s ještě lepšími vlastnostmi (vyšší hodnotou účelové funkce). Vyspívání populace jedinců je možné ztotožnit s postupným přibližováním se optimu. Mutace a křížení je podřízeno náhodným mechanizmům, tzn. že evoluční algoritmus stochasticky prohledává množinu přípustných řešení a při tom uchovává ta slibná. Obecné fungování EA můžeme popsat pomocí schématu:

```
začátek
      VYTVOŘ počáteční populaci z náhodně vybraných přípustných
      řešení (kandidátů)
      OHODNOŤ každého člena populace
      OPAKUJ, DOKUD není splněna podmínka pro ukončení algoritmu
      {
      VYTVOŘ nového potomka (potomky) z párů vybraných rodičů
      ZMUTUJ vybrané členy populace
      OHODNOŤ nově vzniklé členy populace
      VYBER jedince pro další generaci
      }
konec
```

Aplikace EA v praktických úlohách vyžaduje konkretizaci jednotlivých kroků pomocí vhodné matematické formulace. Ze všeho nejdříve je nutné nalézt vhodnou reprezentaci původních přípustných řešení, na kterou by bylo možné v EA aplikovat. Tento krok bývá v rámci použití EA tím nejobtížnějším. Přípustná řešení původního problému jsou nazývána *fenotypy*, jejich zakódované protějšky v EA *genotypy*. Genotypy jsou kódované pomocí vektorů.

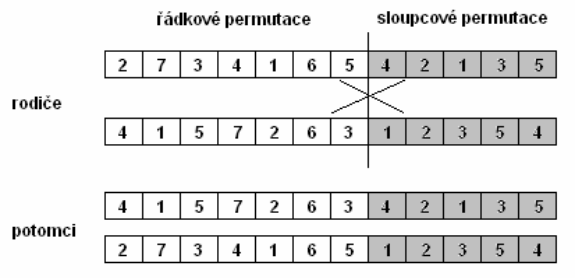
Ukažme si nejdříve použití EA pro uspořádání Bertinovy matice do homogennější struktury. Náš přístup vychází z práce S. Niermanna (viz Niermann, 2005, s. 41–46)[5]. V úloze je genotyp definován jako vektor délky $n + p$, kde prvních n prvků představuje

[5] Zde navržený algoritmus není sice přímo spojován s Bertinovými maticemi, ale jeho rozšíření na ně je logické a možnost aplikace evidentní.

informaci o uspořádání (pořadí) řádků a posledních p prvků kóduje pozici sloupců. Množina permutací řádků a sloupců $n!p!$ zahrnuje všechna možná uspořádání matice a libovolnou konfiguraci tak lze popsat pomocí reprezentace $((\pi(n), \pi(p))$, kde symbol $\pi(.)$ označuje permutace řádků, resp. sloupců.

Praktický postup může vypadat například následovně (lze si přestavit různé další varianty postupu). Počáteční populace je nastavena na 20 náhodně vygenerovaných jedinců – přípustná řešení (jedinci) se skládají z dvou náhodných permutací ve struktuře $((\pi(n), \pi(p))$. Alternativně můžeme do populace zařadit také nenáhodná *dobrá* řešení, které jsme získali pomocí jednodušších metod. Operace křížení je provedena pomocí náhodného vybrání páru přípustných řešení a záměny prvních n prvků prvního z nich s n prvky druhého a naopak. Tím vzniknou dvě nová řešení (obr. 7).

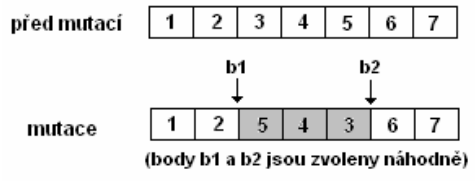
Obr. č. 7: Princip křížení



Operace křížení je opakována a je možné, aby někteří jedinci byli pro křížení vybráni více než jednou. Počet opakování je určen náhodnou veličinou s binomickým rozdělením $Bi(20; 1/2)$.

Mutace přípustného řešení znamená, že uvnitř vektoru dojde s určitou pravděpodobností ke změně hodnoty na jinou. V naší úloze však musíme zabezpečit, aby mutace nenarušila správnou reprezentaci genotypu a řešení po mutaci tak opět přestavovalo permutaci. Nelze mít například ve vektoru $(\pi(n), .)$ dvakrát stejné číslo, protože to by znamenalo, že některý ze řádků zahrneme do uspořádání vícekrát a některý ani jednou. Taková konfigurace neodpovídá původní matici a z pochopitelných důvodů ztrácí smysl. Řešení spočívá v použití tzv. *2-opt-operátoru*, který vzniku podobné situace zabraňuje. Operátor funguje tak, že mezi dvěma náhodně vybranými body se pořadí prvků vektoru obrátí (obr. 8).

Obr. č. 8: Princip mutace



Mutaci tedy definujeme tak, že pro každé přípustné řešení uplatníme s určitou pravděpodobností *2-opt-operátor*, a to jak na vektor řádků, tak sloupců. (Niermann tuto pravděpodobnost nastavuje na 1/2).

Pro další kolo algoritmu musíme nakonec rozhodnout, které jedince necháme vstoupit do nové generace. Všechna řešení v daném kole algoritmu jsou ohodnocena podle své *vhodnosti přežít* a jedinci s vyšší vhodností se stávají členy další generace (vylučuje se buď na základě stanoveného počtu přežívajících, nebo na základě porovnávání k sobě náhodně přidružených dvojic). Vhodnost přežít bezprostředně souvisí s hodnotou použité účelové funkce a tedy se zajímavostí konfigurace. Protože vizuální analýza dat je jednodušší, pokud hodnoty matice leží vedle hodnot vzájemně si podobných, budeme se snažit minimalizovat míru nepodobnosti, resp. vzdálenosti prvků v rámci matice. Tato účelová funkce je nazývána *STRESS* funkcí a měří agregovanou vzdálenost prvků matice od svého okolí[6]. Definic vzdálenosti i okolí se nabízí více, zde budeme vycházet z euklidovských vzdáleností pro *Moorovo* okolí (to je tvořeno osmi okolními prvky matice). Lokální *STRESS* pro prvek matice $\mathbf{X}(i, j)$ je definován vztahem

$$s(i,j) = \sum_{l=max(i,i-1)}^{min(r,i+1)} \sum_{m=max(1,j-1)}^{min(c,j+1)} \left(x_{ij} - x_{lm}\right)^2. \tag{11}$$

Celková STRESS funkce, kterou se snažíme algoritmem minimalizovat (odpovídá záporné hodnotě funkce pročištění) je pak určena součtem lokálních STRESSů pro jednotlivé prvky

$$STRESS = \sum_{i=1}^{r} \sum_{j=1}^{c} s(i,j). \tag{12}$$

Celý proces evolučního algoritmu opakujeme tak dlouho, dokud není rozdíl mezi hodnotami účelové funkce ve dvou po sobě jdoucích kolech menší než předem stanovená hodnota, nebo dokud neproběhne předem zvolený počet kroků algoritmu.

Optimalizovat lze samozřejmě i jiné účelové funkce a čtenář by jistě sám dokázal pomocí účelových funkcí formulovat další strategie, které chce uplatnit. Po úpravách algoritmu můžeme EA volně aplikovat i na další metody. U metody RADVIZ i u paralelních os může být genotyp kódován velmi jednoduše jednorozměrným *p*-prvkovým vektorem udávajícím pořadí proměnných. Na ten potom použijeme např. pouze operaci mutace pomocí *2-opt-operátoru*.

U metody RADVIZ můžeme podle zvoleného cíle maximalizován libovolný index přiřazený ke konfiguraci, který měří jistou strukturu v datech – v úvahu připadají obdobné indexy jako u metody projection pursuit, nebo maximalizace Wardova kritéria známého ze shlukové analýzy, nebo v případě rozpadu souboru podle kategorií maximalizace diskriminace mezi skupinami.

U paralelních os lze uspořádat osy tak, aby byla minimalizována suma vzájemných vzdáleností mezi proměnnými, které jsou měřeny korelačním koeficientem. Opět lze použít i míry maximalizující jiná kriteria, která vedou k odhalování shluků, či diskriminaci mezi skupinami.

[6] Nezapomeňme přitom na nutnost normalizace proměnných.

Závěr

V příspěvku byla provedena studie moderních nástrojů grafické analýzy dat. Jejich rozmach podpořený výkonnou výpočetní technikou a dostupností softwarových nástrojů dává tušit jejich další rozšíření. Předností všech popsaných metod je relativně snadná interpretace získaných výsledků pro statistické odborníky ale i pro laického čtenáře. Univerzálnost spočívá v minimálních předpokladech korektního použití. Jejich komplementární zapojení nejen do přípravné (explorační), ale také do analytické a prezentační fáze při zpracování větších objemů dat lze jen doporučit.

Literatura

[1] BENZÉCRI, J.-P., 1973: *L´analyse des données*. 1973, Vol. 2, Paris, Dunod.

[2] BERTIN, J., 1967: *Sémiologie Graphique. Les diagrammes, les réseaux, les partes*. Paris, La Haye, Mouton, Gauthier-Villars, 1967.

[3] BRUNSDON, C. – FOTHERINGAM, A. S. – CHARLTON, M. E., 1999: *An Investigation of Methods for Visualizing Highly Multivariate Datasets*. Advisory Group of Computer Graphics.

[4] De FALGUEROLLES, A. – FRIEDRICH, F. – SAWITZKI, G., 1973: *A Tribute to J. Bertin´s Graphical Data Analysis*. *Advances in Statistical Software*, Stuttgart, Lucius and Lucius, 1973, s.11–20.

[5] FRIEDMAN, J. H. – TUKEY, J. W., 1974: *A projection pursuit algorithm for exploratory data analysis*. *IEEE Trans. Comput.*, C23 (9), 1974, s. 881–890.

[6] HOLLAND, J. H., 1992: *Adaption in Natural and Artificial Systems 5th edition.* MIT Press, 1992.

[7] HODGES, J. L. – LEHMANN, E. L., 1956: The efficiency of some non-parametric competitors of the t-test. *Ann. Math. Statist.*, 1956, roč. 27, s. 324–335.

[8] HUBER, P. J., 1985: Projection Pursuit. *The Annals of Statistics*, 1985, roč. 13, č. 2.

[9] JONES, M. C. – SIBSON, R., 1987: What is projection pursuit? (with discussion). *J. R. Statistical Society*, 1987, roč. 150, s. 1–36.

[10] KRUSKAL, J. B., 1969: *Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new 'index of condensation'*. New York, R. C. Milton and J. A. Nelder Statistical Computation Academic Press, 1969.

[11] KRUSKAL, J. B., 1972: Linear transformations of multivariate data to reveal clustering. In Seminar Press: Multidimensional Scaling: Theory and Application in the Behavioural Sciences I, 1972.

[12] KVASNIČKA, V. – POSPÍCHAL, J. – TIŇO, P., 2000: *Evolúčne algoritmy*, Bratislava, STU, 2000.

[13] NASON, G. P., 1992: *Design and choice of projection indices*. Thesis, Univ. of Bath, 2000.

[14] NIERMANN, S., 2005: Optimizing the Ordering of Tables With Evolutionary Computation. *The American Statistician*, 2005, roč. 5, s. 41–46.

[15] PLAŠIL, M. – VLACH, P., 2005: *Visualization of Multivariate Data*. AMSE 2006 – 8th International Scientific Conference "Applications of Mathematics and Statistics in Economy", Wroclaw, Wroclaw University of Economics, 2005.

[16] PLAŠIL, M. – VLACH, P., 2006: *Možnosti grafického zpracování dat*. Sborník prací účastníků vědeckého semináře doktorského studia, Praha, Fakulta informatiky a statistiky Vysoké školy ekonomické, 2006, s. 176–186.

[17] RENCHER, A. 2002: *Methods of Multivariate Analysis 2nd edition*, New York, Wiley-Interscience, 2002.

[18] RÉNYI, A., 1964: *Probability Theory*. Amsterdam, North-Holland, 1964 .

[19] SWITZER, P., 1970: *Numerical Classification*. Geostatistics, 1970.

Grafická analýza vícerozměrných dat

Miroslav Plašil – Petr Vlach

Abstrakt

Příspěvek se zabývá novými nástroji pro grafické zpracování a analýzu dat. Demonstrovány jsou základní použité algoritmy a grafické výstupy jednotlivých metod. Podrobnějšímu zkoumání jsou podrobeny metody Bertinových permutačních matic, RADVIZ, projection pursuit a metoda paralelních os. Ilustrativní příklady ukazují jejich praktické využití včetně komentáře výstupů a vysvětlení univerzálnosti jejich použití při zpracování vícerozměrných dat.

Klíčová slova: vizualizace dat; Bertinovy permutační matice; RADVIZ; Projection Pursuit; paralelní osy.

Visual Analysis of multivatiate data

Abstract

The article presents and investigates new possibilities of multivariate data visualization and their analytical convenience. We demonstrate elementary principles, algorithms and graphical outputs of modern visualization methods with particular focus on Bertin matrices, RADVIZ, Projection pursuit and parallel coordinates. Illustrative examples show their practical implementation into the process of multivariate data analysis, hence providing the reader with an idea of the wide range of their application.

Key words: multivariate data visualization; Bertin matrix; RADVIZ; Projection pursuit; parallel coordinates.

JEL classification: C61, C81, C88