VYBOČUJÍCÍ HODNOTY VE VÍCEROZMĚRNÝCH DATECH

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 je jinak*

Abstrakt:
Jsou popsány základní postupy pro určování odlehlých měření ve vícerozměrných datech, které jsou modifikací postupů pro jednorozměrná data. Jsou vybrány především metody, které při své jednoduchosti umožňují zpracovat výběry obsahující skupiny vlivných bodů, kde se může projevit jak maskování tak i nesprávné zařazení korektních bodů. Jsou uvedeny jednoduché grafy pro identifikaci skupin vlivných bodů.

1.Úvod

Jednou ze základních úloh analytické chemie je simultánní monitorování úrovně různých látek v materiálech, ovzduší, vodě a půdě. Cílem je často zjištění, zda dané látky (celkem p) nepřekračují zadané úrovně. Standardně se vychází z vícerozměrných výběrů obsahujících N měření $(\boldsymbol{x}_1 ..... \boldsymbol{x}_N)$. Vektor $\boldsymbol{x}_j$ pro j té měření obsahuje složky $(x_{j1}, x_{j2} ... x_{jp})$. Výsledkem měření je tedy matice dat $\boldsymbol{X}$ řádu $N \times p$ obsahující N řádků (měření) a p sloupců (látek). Účelem je v nejjednodušším případě stanovení vhodného odhadu vektoru středních hodnot μ a porovnání se zadanými úrovněmi μ_o. S ohledem na variabilitu měření je vhodné ověřit, zda vektor μ_o padne do oblasti spolehlivosti CI vektoru parametrů μ či nikoliv. Tato úloha může být komplikována jak nestejnými rozptyly tak i vzájemnou korelací mezi látkami (sloupci matice $\boldsymbol{X}$). Také celá řada dalších úloh z oblasti analytické chemie vede na zpracování vícerozměrných výběrů.

Problém komplikuje to, že data z oblasti analytické chemie mají standardně některé specifické zvláštnosti:

(a) rozsahy zpracovávaných dat nejsou obyčejně velké,
(b) v datech se vyskytují výrazné nelinearity, neaditivity a struktury, které je třeba identifikovat a popsat,

(c) rozdělení dat jen zřídka odpovídá normálnímu běžně předpokládanému ve standardní statistické analýze,
(d) v datech se vyskytují vybočující měření a různé heterogenity,
(e) statistické modely se často tvoří na základě předběžných informací z dat (datově orientované přístupy),
(f) parametry statistických modelů mají mnohdy definovaný fyzikální význam, a musí proto vyhovovat velikostí, znaménkem nebo vzájemným poměrem,
(g) existuje jistá neurčitost při výběru modelu, popisujícího chování dat.

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

"data - model - statistická metoda"

Toto schéma se považuje za základ interaktivní tvorby statistických modelů všeho druhu. Při jeho praktickém použití však nastávají problémy s tzv. **vybočujícími hodnotami** (body) a to nejen s jejich indikací, ale zejména interpretací a omezením jejich vlivu. Všeobecně se soudí, že většina statistických metod je **výrazně negativně** ovlivněna přítomností vybočujících bodů**. Ve skutečnosti** záleží na tom, o jaké vybočující body jde, a kde se vyskytují. (např. v regresní analýze mohou vybočující body jak zvětšovat tak i zmenšovat korelační koeficient). V některých případech je třeba rozhodnout, zda podezřelé hodnoty považovat za správné (a pracovat např. se zešikmenými rozděleními) nebo za chybné a jejich vliv eliminovat. Toto rozhodnutí nelze učinit bez znalosti způsobu získávání dat a realizace experimentů. Nesprávné rozhodnutí může mít katastrofické důsledky s ohledem na interpretaci výsledků a praktické závěry.

V této práci popsány postupy pro určování odlehlých měření ve vícerozměrných datech, které jsou modifikací postupů pro jednorozměrná data. Jsou uvedeny jednoduché grafy pro identifikaci skupin vlivných bodů.

2. Vybočující body

Pojem vybočující body evokuje představu, že jde o body, které lze vizuálně určit na základě vhodného zobrazení. To platí pro jednorozměrné výběry, kdy vybočující znamená také odlehlé. Ve vícerozměrných případech jsou vybočující hodnoty buď odlehlé co do hodnot od ostatních nebo neodpovídající strukturám v ostatních datech. Pro vybočující body obecně platí, že:

- zkreslují výsledky
- „nelíbí se“ vypadají nepatřičně
- zhoršují přesnost
- neumožňují selekci modelu

Pro identifikaci odlehlých měření je obecně třeba:

- definovat „čistá data“
- určit pravděpodobnostní model dat (a často i vybočujících bodů)
- odhadnout parametry tohoto modelu

Při analýze vybočujících bodů se množina indexů *I = (1,2,3,...N)* rozkládá na podmnožinu potenciálně dobrých dat *D* a potenciálně vybočujících bodů *V*. Tedy *I = (D,V)*. Počet potenciálně dobrých dat je N_D a počet potenciálně vybočujících bodů je N_V. Podíl vybočujících bodů je pak $e = N_V/N$. Nechť je rozdělení podílu *1 - e* dobrých bodů charakterizováno distribuční funkcí $G(\mu_0, \Sigma_0)$ s vektorem středních hodnot μ_0 a kovarianční maticí Σ_0 a rozdělení podílu *e* potenciálních vybočujících bodů je $H(\mu + \mu_0, \Omega)$ s vektorem středních hodnot $\mu + \mu_0$ a kovarianční maticí Ω. Očekávaná hodnota výběrového průměru $\boldsymbol{x_p}$ ze všech dat je pak

$$E(x_p) = \mu_0 + e\,\mu$$

a očekávaná hodnota výběrové kovarianční matice **S** je

$$E(S) = (1-e)\,\Sigma_0 + e\,\Omega + e\,(1-e)\,\mu^T \mu$$

Je tedy patrné, že výběrové průměry a kovarianční matice ze všech dat jsou závislé jak na podílu vybočujících bodů tak i na jejich parametrech. To může vésti k situaci, kdy se z odhadů získaných ze všech dat nedají určit vybočující body. Nejhorší situace z hlediska indikace vybočujících bodů je případ, kdy obě kovarianční matice mají stejný tvar. Tento typ vybočujících bodů se označuje jako „posunuté vybočující body“.
Pro indikaci vybočujících měření se často s výhodou používá definice zobecněné vzdálenosti

$$d_i = \sqrt{(x_i - x_{AD})^T * [w(D,p) * S_D]^{-1} * (x_i - x_{AD})}$$

kde x_{AD} a S_D jsou vektor aritmetických průměrů a kovarianční matice pro potenciálně dobrá data. Korekční faktor *w(D,p)* byl zaveden ve tvaru [4]

$$w(D,p) = \left[1 + \frac{2}{N_D - 1 - 3p} + \frac{p+1}{N_D - p}\right]^2$$

Většina metod pro indikaci vybočujících bodů vychází z představy vícerozměrné normality, kdy

$$G(\mu_0, \Sigma_0) = N(\mu_0, \Sigma_0)$$

a

$$H(\mu + \mu_0, \Omega) = N(\mu + e\,\mu_0, k\,\Omega).$$

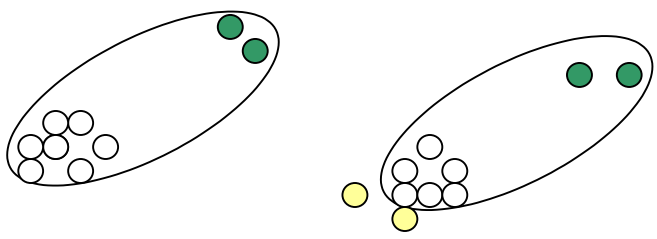
Tato představa je potřebná pro určení kritických mezí oddělujících dobrá a špatná data. Podíl e ovlivňuje také špičatost rozdělení měření $\boldsymbol{x}$. Pro jednorozměrné výběry asymptoticky (pro $s_D^2 / s_V^2 \to \infty$) platí, že špičatost g_2 je dána vztahem

$$g_2 = \frac{e^3 + (1-e)^3}{e\,(1-e)}$$

Pro $e = 0{,}5$ je $g_2 = 1$ (minimum) a pro e blízké nule je g_2 rostoucí nade všechny meze. Nejmenší počet e_N dobrých bodů je

$$e_N = \mathrm{int}[(N + p + 1)/2]$$

Techniky indikace vybočujících bodů jsou citlivé na tzv. „**maskování**“, kdy vybočující jeví jako korektní (díky zvětšení kovarianční matice). Dalším problémem je „**překryt**“, kdy přítomnost vybočujících měření způsobí, že některá správná měření leží mimo akceptovatelnou oblast.(díky zkreslení kovarianční matice) [1]. Schematicky jsou tyto situace znázorněny na obr. 1 (vybočující body jsou tmavé).



A. maskování **B. překryt**

Obr. 1 Příklad maskování (A) a překrytu (B)

Znázornění na obr 1. vychází z faktu, že čtverce zobecněných vzdáleností mají χ^2_p rozdělení (elipsa je tedy hraniční oblast oddělující dobrá (D) a vybočující (V) data.

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

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

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

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

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

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

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

Jednotný postup zde neexistuje a záleží na experimentátorovi, resp. zpracovateli jakou variantu zvolí.Vzhledem k tomu, že vybočující body jsou většinou extrémně vlivné vede zde nevhodná manipulace ke ztrátě informací a nesprávným závěrům. V souvislosti s vybočujícími body je možné definovat tyto základní paradoxy:

1. známe - li rozdělení dat a model můžeme určit vybočující hodnoty. Model a rozdělení dat se však hledá.

2. pro posouzení vybočujících měření potřebujeme znát „čistá data“. Robustními metodami však dostaneme data „čistá“ s ohledem na základní model

3. Ne vše co vypadá jako vybočující skutečně vybočuje a naopak (maskování, překryt)

4. Platí, že co je vybočující pro jeden model může být akceptovatelné pro jiný model

3. Jednorozměrné výběry

Standardní způsob zpracování jednorozměrných výběrů spočívá ve výpočtu aritmetického průměru x_A a výběrového rozptylu s^2. Je známo, že pokud zpracovávaný výběr velikosti N prochází z ne - normálního rozdělení se střední hodnotou μ a rozptylem σ^2 $(<\infty)$ má náhodná veličina

$$Z = \sqrt{N} * (x_A - \mu) / \sigma \qquad (1)$$

asymptoticky normální rozdělení. Pokud není σ^2 známo, nahrazuje se výběrovou směrodatnou odchylkou s. Pak má tzv. Studentova náhodná veličina

$$t = \sqrt{N} * (x_A - \mu) / s \qquad (2)$$

Studentovo rozdělení s $(N - 1)$ stupni volnosti. Asymptotická normalita veličiny Z resp. Studentovo rozdělení veličiny t umožňuje konstrukci intervalu spolehlivosti střední hodnoty μ. Při tzv. frekventistickém přístupu je 100 $(1 - \alpha)$ % na interval spolehlivosti CI definován vztahem

$$P(CID \le \mu \le CIH) = 1 - \alpha \qquad (3)$$

Symbol P (.) označuje pravděpodobnost a α je tzv. hladina významnosti. Obyčejně se volí $\alpha = 0.05$ nebo $\alpha = 0.01$ s tím, že čím je α menší, tím je interval (CID, CIH) širší. Při znalosti rozptylu σ^2 je možno interval spolehlivosti CI vyjádřit ve tvaru

$$x_A - z_{1-\alpha/2} * \frac{s}{\sqrt{N}} \le \mu \le x_A - z_{\alpha/2} * \frac{s}{\sqrt{N}} \qquad (4)$$

kde $z_{1-\alpha/2} = -z_{\alpha/2}$ jsou kvantily normovaného normálního rozdělení. Pokud není σ^2 známo lze použít vztah

$$x_A - t_{1-\alpha/2}(N-1) * \frac{s}{\sqrt{N}} \le \mu \le x_A - t_{\alpha/2}(N-1) * \frac{s}{\sqrt{N}} \qquad (5)$$

kde $t_{1-\alpha/2}(N-1) = -t_{\alpha/2}(N-1)$ jsou kvantily Studentova rozdělení s *N-1* stupni volnosti. Pro případ normálního rozdělení mají intervaly (4) resp. (5) přesně 100*(1-α)* % *ní pokrytí střední hodnoty*.To znamená, že jen v 100*α/2* % případů je *střední hodnota menší než CI* (nejistota *NP* zprava) a v 100*α/2* % případů je větší než *CI* (nejistota *NL* zleva).Pro případ ne-normálního rozdělení platí tyto intervaly pouze asymptoticky tedy pro dostatečně vysoká *N*. Dostatečná velikost *N* závisí silně na šikmosti $g_1(x)$ rozdělení z kterého data pocházejí [3].

Pro kvantifikaci vlivu šikmosti na rozdělení náhodné veličiny *Z* definované rov. (1) je možno použít prvního členu Edgeworthova rozvoje pro, který platí

$$P(Z \le x) = F_n(x) - \frac{g_1(x) * (x^2 - 1)}{6\sqrt{N}} f_n(x) \qquad (6)$$

Zde $F_n(x)$ je distribuční funkce normovaného normálního rozdělení a $f_n(x)$ je odpovídající hustota pravděpodobnosti. Šikmost náhodné veličiny *Z* je dána vztahem

$$g_1(Z) = g_1(x) / \sqrt{N} \qquad (7)$$

Čím je $g_1(Z)$ blíže k nule, tím je rozdělení veličiny *Z* bližší normálnímu. Z rov. (6) je patrné, že pro rozdělení dat zešikmené k vyšším hodnotám (tj. $g_1(x)$ kladné), je také rozdělení náhodné veličiny *Z* zešikmené k vyšším hodnotám (tj. $g_1(Z)$ kladné). Interval spolehlivosti (4) pak má vyšší horní mez *CIH* a vyšší dolní mez *CID* než odpovídá reálnému rozdělení statistiky *Z*. Např. pro výběr rozsahu *N=10* ze standardizovaného exponenciálního rozdělení, kdy je $g_1(x)=2$, je 97.5 % ní kvantil rozdělení veličiny *Z* určený z rov. (6) roven 2.24 a odpovídající kvantil normovaného normálního rozdělení je pouze 1.96. Podobně lze určit, že 2.5 % ní kvantil *Z* je pouze –1.65 oproti odpovídajícímu kvantilu normovaného normálního rozdělení -1.96. Interval spolehlivosti definovaný rov. (4) je tedy celý <u>posunut doprava</u> oproti skutečnému [3].

Také pro kvantifikaci vlivu šikmosti na rozdělení náhodné veličiny *t* definované rov. (2) je možno použít prvního členu Edgeworthova rozvoje

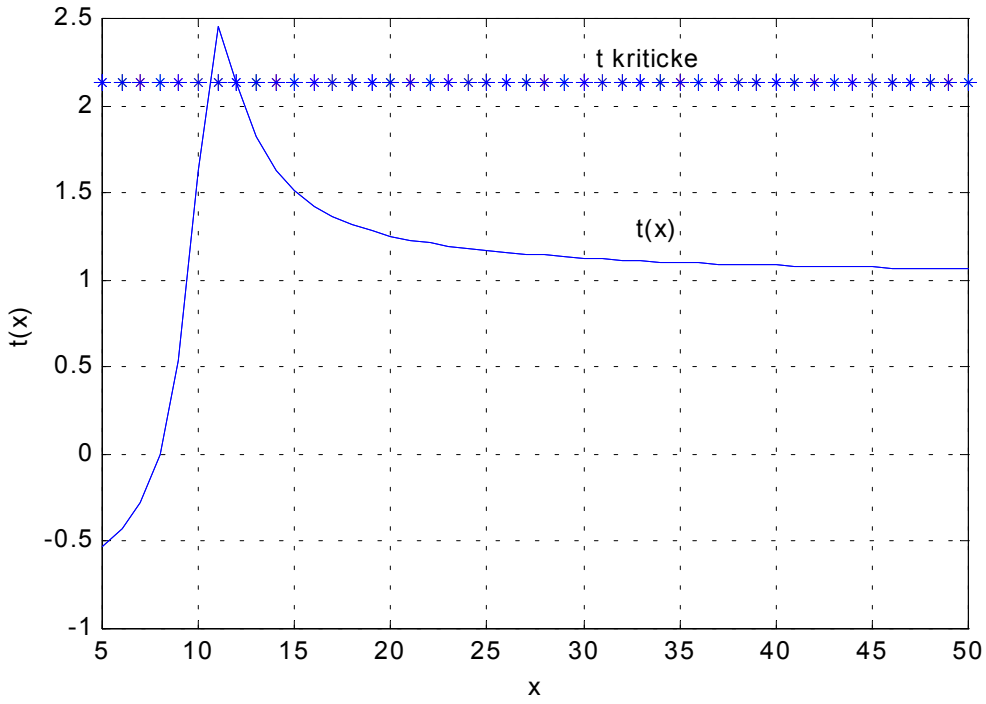
$$P(t \le x) = F_n(x) + \frac{g_1(x) * (2x^2 + 1)}{6\sqrt{N}} f_n(x) \qquad (8)$$

Zde je opět $F_n(x)$ distribuční funkce normovaného normálního rozdělení a $f_n(x)$ je odpovídající hustota pravděpodobnosti. Při porovnání s rov. (6), je patrné

opačné znaménko korekčního členu, což znamená, že pro rozdělení dat zešikmené k vyšším hodnotám (tj. $g_1(x)$ kladné), je rozdělení náhodné veličiny t zešikmené k nižším hodnotám (tj. $g_1(t)$ záporné). Interval spolehlivosti (5) pak má nižší horní mez *CIH* a nižší dolní mez *CID* než odpovídá reálnému rozdělení statistiky t. Interval spolehlivosti definovaný rov. (5) je tedy celý *posunut doleva* oproti skutečnému [3]. To je zvláště nepříjemné u dat silně zešikmených vpravo a vede to k přehnaně optimistickým závěrům.
Při testování hypotéz o střední hodnotě se používá statistika

$$t(x) = \frac{(x_A - \mu_0)\sqrt{N}}{s} \qquad (9)$$

Zde x je uvažováno jako poslední hodnota ve výběru tj. $x = x_N$. Obecně přijímaný předpoklad je, že stačí přítomnost jednoho vybočujícího bodu x aby došlo ke zvýšení $t(x)$ a tedy zamítnutí nulové hypotézy $H_o : \mu = \mu_0$ je nesprávný. Na obr. 2 je zakreslen průběh funkce t(x) pro výběr V = (10, 10 , 11,11, x) pro x = 5,6,…50. Hvězdičkami je znázorněna kritická hodnota Studentova rozdělení $t_{1-\alpha}(N-1) = 2.1318$ pro $\alpha = 0.05$ a N = 5, která platí pro alternativní hypotézu $H_A : \mu > \mu_0$.



Obr. 2 Vliv velikosti posledního bodu x výběru *(10,10,11,11,x)* na velikost t(x)

Je patrné, že pouze malá oblast nad hvězdičkami vede k přijetí alternativní hypotézy a silně vybočující bod naopak podporuje hypotézu nulovou o shodě střední hodnoty s $\mu_0 = 10$. Dá se také ukázat, že [5]

$$\lim_{x \to \pm\infty} t(x) = \pm 1$$

Snadno lze také odvodit, jakému *x* odpovídá maximum *t(x)*. Zaveďme označení $a = \sum_{i=1}^{N-1} x_i$ a $b = \sum_{i=1}^{N-1} x_i^2$. Pak v případě, že $a \neq (N-1)\mu_0$ nabývá *t(x)* nejvyšší hodnoty pro x_c rovno [5]

$$x_c = \frac{b - a\mu_0}{a - (N-1)\mu_0}$$

Je tedy patrné, že silně vybočující měření vede k poklesu *t(x)*. Pokud je $a = (N-1)\mu_0$ je *t(x)*.monotónně rostoucí od –1 do 1.

Příklad

Mějme „čistá“ data D1= (10, 10, 11, 11, 12) a $\mu_0 = 10$. *Pak t(x) = 2,138 . Pro data s vybočujícím měřením D2 = (10,10,11,11,18) je však t(x) = 1,318. Tedy hypotéza Ho:* $\mu_0 = 10$ *je pravděpodobnější. Maximálnímu t(x) odpovídá hodnota* $x_c = 11$.

Pro identifikaci vybočujících hodnot se používá řady testů založených na definici standardizované maximální odchylky od aritmetického průměru (počítaného bez extrémní hodnoty). Jednoduchá momentová metoda předpokládá, že vybočující je takové x_j ,pro které je

$$\left| \bar{x}_A^* - x_j \right| \geq K_c . s^*$$

kde

$$K_c = 1.55 + \left[0.8 . \sqrt{g_2^* - 1} . \log(N/10) \right]$$

$\bar{x}_A^*$, s^* ..... „čisté“odhadnuty střední hodnoty a směrodatné odchylky (bez podezřelého bodu x_j).

g_2^*..... odhad špičatosti bez vybočujícího bodu

$$g_2^* = \frac{N.\sum (x_i - \bar{x})^4}{\left[\sum (x_i - \bar{x})^2\right]^2}$$

Postup je vhodný pro N ≥ 20. Pro N = 20 vyjde pro různá rozdělení:

- Normální rozdělení $g_2 = 3$ $K_c = 1.89$
- Rovnoměrné rozdělení $g_2 = 1.8$ $K_c = 1.77$
- Laplaceovo rozdělení $g_2 = 6$ $K_c = 2.09$

V případě více vybočujících měření se postupně vylučují podezřelé body na základě výše uvedeného kritéria nebo se simultánně určují všechny vybočující hodnoty pro různé kombinace podezřelých bodů.
Obecně je tedy třeba řešit tyto úlohy:

A. Výběr vhodného rozdělení dat
B. určení „čistých odhadů“ ze všech bodů, nebo podmnožiny „čistých bodů“
C. nalezení kritické hodnoty pro selekci vybočujících bodů

Jako vhodné rozdělení dat (A) se většinou uvažuje rozdělení normální, protože umožňuje jednoduché nalezení kritických hodnot (C). Pro určení „čistých odhadů“ se používají především různé robustní metody. Pro nalezení „čistých bodů“ se používá dvou přístupů:

„Brute force“ – kdy se zkouší všechny možné kombinace podvýběrů. Tento postup vede k cíli ale je časově náročný.

„Clean subset“ – kdy se hledají data, která jsou určitě „čistá“ a nezkreslí odhady střední hodnoty a rozptylu.

Je snahou nalézt takové metody, které nevyžadují příliš komplikované výpočty a přitom jsou dostatečně spolehlivé.
Obecně lze tento přístup použít pro libovolně rozměrná data.

4. Vícerozměrná data

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

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

Tato oblast pokrývá celý prostor E^p s vyloučením vícerozměrného elipsoidu kolem vektoru středních hodnot. Vybočující body jsou tedy příliš vzdáleně od střední hodnoty.
Oblast vybočujících bodů OR pro výběr velikosti N je určena výrazem

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

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

$$d^2{}_i = (x_i - x_A)^T S^{-1} (x_i - x_A)$$

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

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

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

Pro malé výběry je lépe použít modifikovaný koeficient

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

Je zajímavé, že pro případ jednoho vybočujícího měření neroste zobecněná vzdálenost nade všechny meze, ale je ohraničená hodnotou

$$d^2{}_{\max} \approx \frac{(N-1)^2}{N}$$

Wilks použil pro určení jednoho vybočujícího bodu ve vícerozměrných datech statistiku

$$R_i = \frac{\det(S_i)}{\det(S)}$$

kde S_i je odhad kovarianční matice s vynecháním *i-tého* bodu a S je odhad kovarianční matice ze všech bodů. Minimální R_i indikuje potenciální vybočující bod. Dá se ukázat, že R_i souvisí se čtvercem zobecněné vzdálenosti vztahem

$$R_i = 1 - \frac{N}{(N-1)^2} d^2{}_i$$

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

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

Pro stanovení **„čisté podmnožiny dat“** se doporučuje BACON algoritmus, složený z těchto kroků:

1. určení základní podmnožiny m > p „čistých dat“
2. odhady parametrů předpokládaného modelu M a nalezení reziduí pro všechny body
3. doplnění základní podmnožiny o data s malými rezidui a vyloučení dat s velkými rezidui
4. iterace kroků 2 a 3 až se překročí pravidlo ukončení
5. body, které nejsou v základní podmnožině jsou vybočující

BACON pro vícerozměrná data (bez struktury) se skládá z těchto kroků:

1.Výběr základní podmnožiny bud na základě

- Mahalanobisovy vzdálenosti a uřezání podezřelých dat
- Vzdálenosti od mediánu

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

2. Výpočet reziduí

$$d_i = (x_i - x_{AC})^T S_C{}^{-1} (x_i - x_{AC})$$

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

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

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

$c = c_1 + c_2$

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

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

Proces je ukončen, když se ve dvou následujících iteracích nemění odhady parametrů x_{RC} a S_C (maximální rozdíl je menší než 10^{-6})

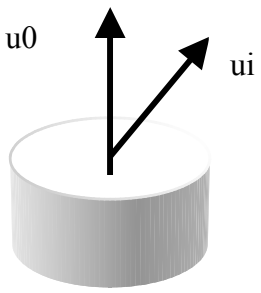
4.Grafická interpretace výsledků

Při identifikaci skupin vybočujících bodů se s výhodou volí různé typy grafů. Mezi základní patří:

A. Indexový graf pro Mahalanobisovy vzdálenosti. Vynáší se $d^2_{(i)}$ proti i a kritická úroveň. Vybočující body leží nad touto úrovní

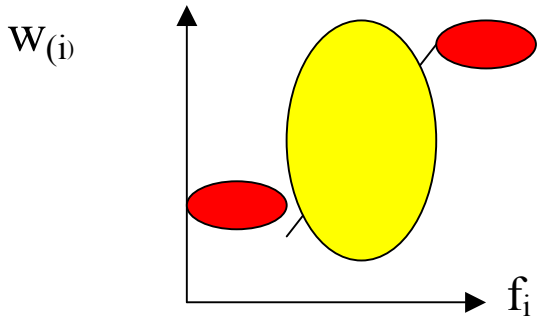
B. Q-Q graf funkcí d^2. Vynáší se $y_i = \frac{F_{p,N-p}(0.5)}{median(d)} d^2{}_{(i)}$ proti $x_i = F_{p,N-p}(\frac{i}{N+1})$. Poměr $y_i/x_i>2$ indikuje vybočující body. Je možné také volit klasický Q-Q graf, kdy se vynáší pořádkové statistiky $d^2{}_{(i)}$ (tj. vzestupně uspořádané čtverce vzdáleností) proti kvantilům χ^2_p rozdělení.

C. Úhlová metoda založená na jednorozměrné projekci dat. Je známo, že pro eliptické rozdělení $X = \Sigma * y + \mu$ má $y_i = S^{-1}(x_i - m)$ cirkulární rozdělení a $u_i = y_i / \|y_i\|$ má vícerozměrné rovnoměrné rozdělení. Pro zvolené u_o má pak veličina $w_i = \cos^{-1}(u_0^T u_i)$ rovnoměrné rozdělení.



Zde u_o je blízké směrům, kde nabývá špičatost maxima resp. minima.

Konstruuje se tedy Q-Q graf pro rovnoměrné rozdělení, kdy se vynáší $w_{(i)}$ vs. $P_i = i/(N+1)$



5. Program EXMUL

Pro identifikaci vybočujících bodů ve vícerozměrných datech byl sestaven program EXMUL v jazyce MATLAB. Program obsahuje tyto části:

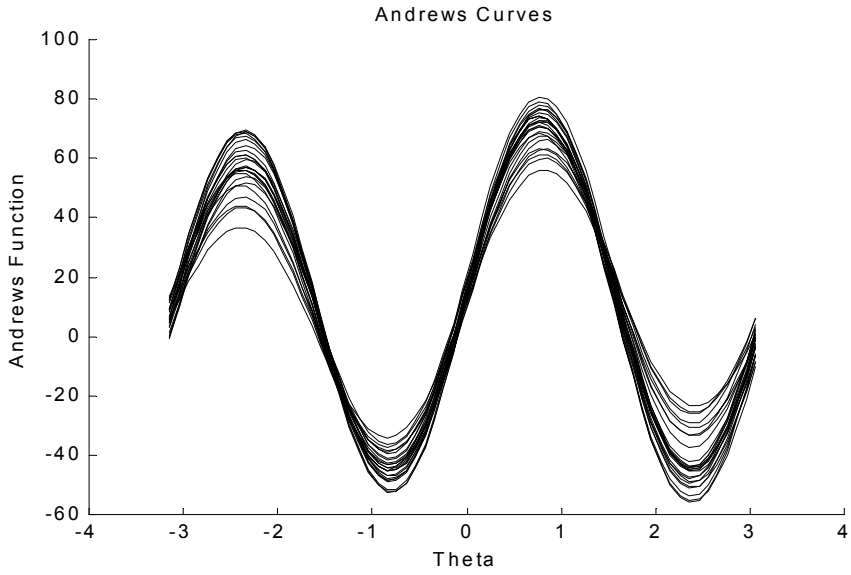
- Exploratorní grafy pro vícerozměrná data [1]
- Robustní odhad $\boldsymbol{x}_A$ a $\boldsymbol{S}$ využívající eliptické vícerozměrné uřezání
  Indexový graf pro $\boldsymbol{d}^2$
- Q-Q graf pro funkci $\boldsymbol{d}^2$
- Úhlový graf

Pro ilustraci činnosti tohoto programu byla použita Hockingova syntetická data určená pro regresní diagnostiku . N=26 , p = 4

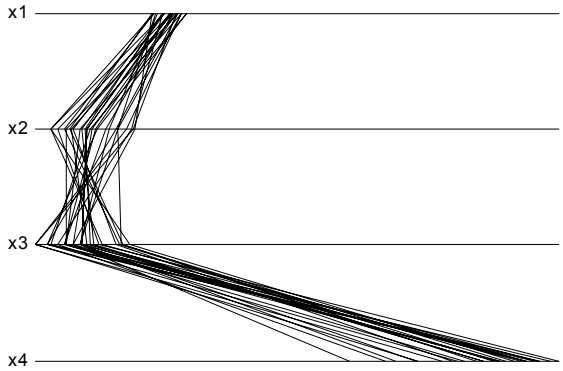
Model: Y = a0 + a1*x1 + a2*x2 + a3*x3

Generace dat : y= 20 + 3*x1 - 2*x2 + eps1 eps1...náhodná čísla z N(0,.25)

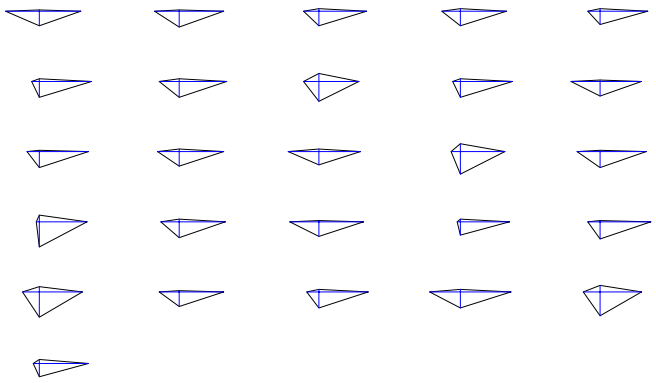
multikolinearita: 2*x3 = 60-3*x1 -1.5*x2 + eps2,
eps2... náhodná čísla z N(0,.16)
Vybočující body : č.11,17,18. Extrém : č.24 (leží mimo rovinu multikolinearity)
Data byla zpracována jako nestrukturalizovaná . Výstupy jsou na obr. 3-9.



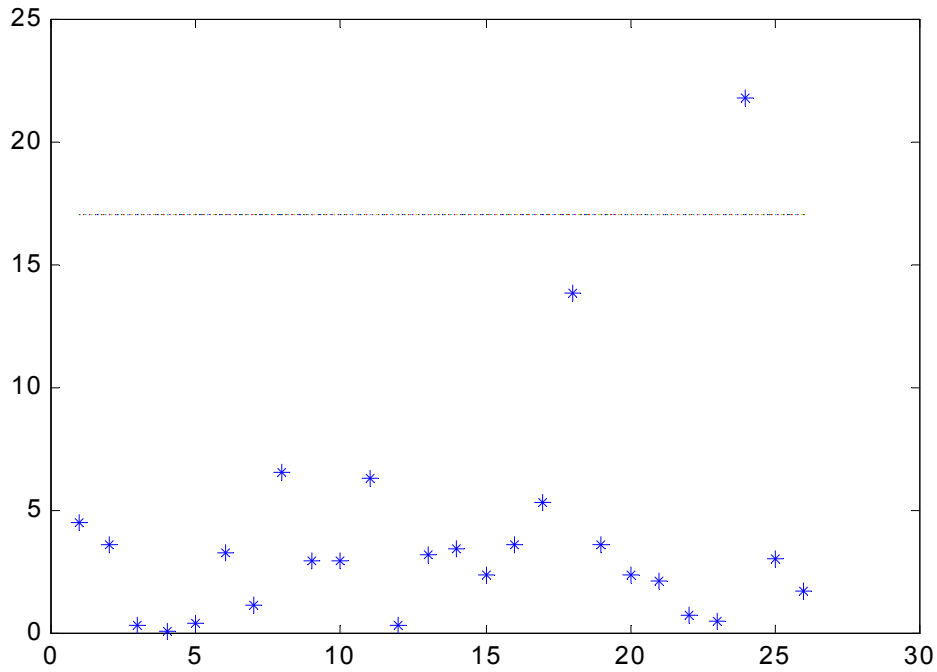
Obr. 3 Andrewsův graf



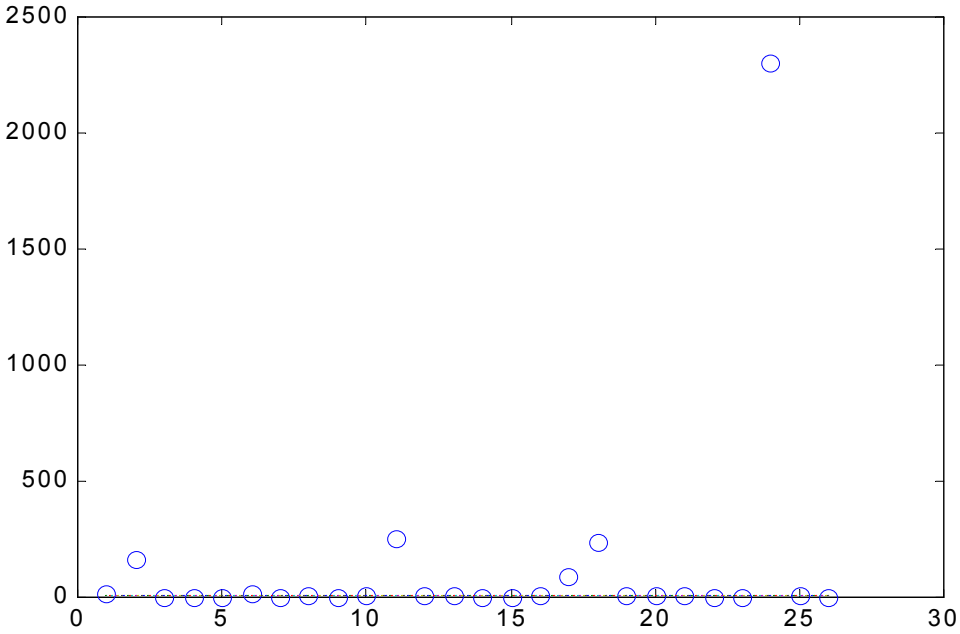
Obr. 4 Paralelní souřadnice



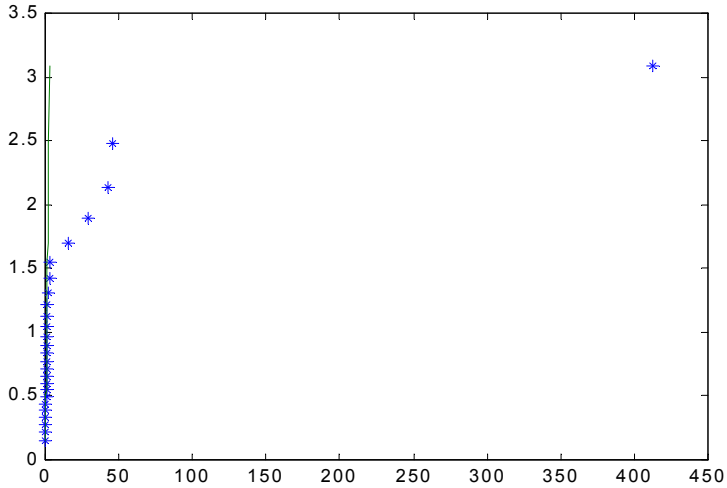
Obr. 5 Hvězdy



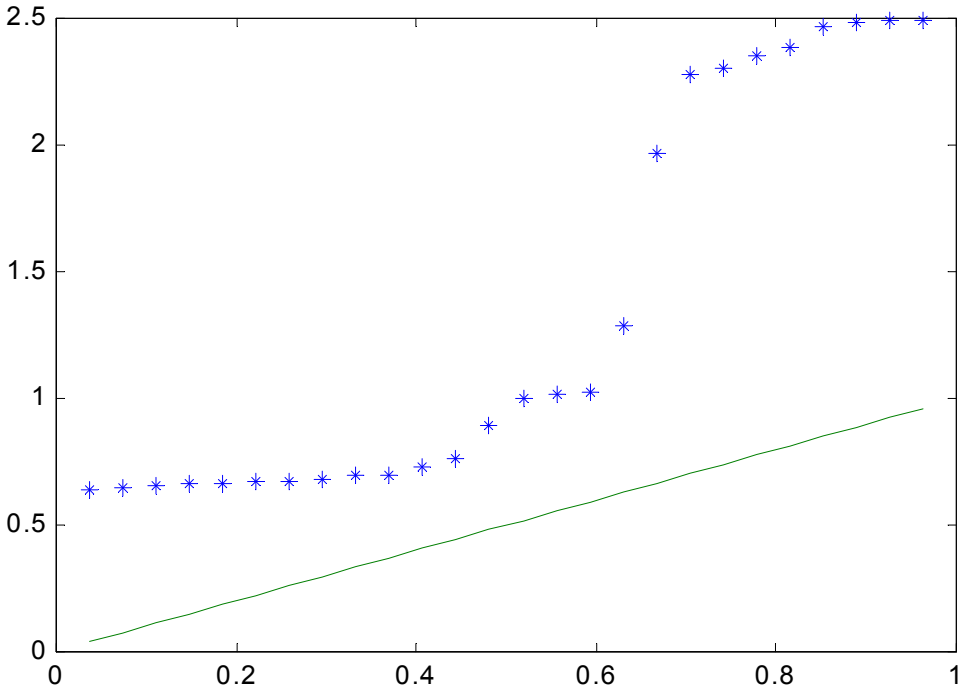
Obr. 6 Zobecněné vzdálenosti (Mahalanobisovy) pro odhady středních hodnot a kovarianční matice ze všech dat (momenty).



Obr. 7 Zobecněné vzdálenosti (Mahalanobisovy) pro robustní odhady středních hodnot a kovarianční matice (30% uřezání).



Obr. 8 Q-Q graf pro funkci d^2



Obr. 9 Úhlová metoda

Je patrné, že:

1. Nejvíce vlivný je bod č. 22
2. Vybočující jsou body 11, 17, 18 ale také 2.
3. Robustní **Mahalanobisova** vzdálenost je vhodná pro vícerozměrná data
4. Indexový graf je jednoduchý a přehledný při postačující schopnosti identifikovat skupiny vybočujících bodů

Stále platí, že je třeba nejen identifikovat potenciální vybočující hodnoty ale také rozhodnout co s nimi dále

8. Závěr

Je patrné, že statistické zpracování dat v analytické chemii má celou řadu specifických zvláštností. V řadě případů je třeba i ve zdánlivě jednoduchých situacích používat poměrně komplikované metody. Formální aparát statistiky resp. přizpůsobení dat potřebám statistické analýzy bez hlubšího rozboru zde může vést ke katastrofickým závěrům.

Poděkování:
Tato práce vznikla s podporou výzkumného centra Textil LN00B090

9. Literatura

[1] Meloun M., Militký J.*: Zpracování experimentálních dat*, East Publishing Praha 1998
[2] Barnett V., Lewis T.: Outliers in statistical data , 3rd. Ed., Wiley, Chicheter 1994
[3] Campbell N.A.: Appl. Statist. **29**, 231 (1980)
[4] Hadi A. S.: J. R. Stat. Soc. **B56**, 393 (1994)
[5] Grimmet D.R., Ridenhorn J.R.: Amer. Statist. **50,** 145 (1996)
[6] Hocking R.R.,Pendleton O.J.: Commun.Statist. A12,497 (1983)