MOCNINNÁ TRANSFORMACE VÍCEROZMĚRNÝCH DAT

JIŘÍ MILITKÝ ,
Katedra textilních materiálů, Technická universita v Liberci, Hálkova 6
461 17 Liberec, e- mail: jiri.miliky@vslib.cz

MILAN MELOUN,
Katedra analytické chemie, Universita Pardubice, Pardubice

Motto: *Transformuj ale ověřuj*

Abstrakt:
Jsou uvedeny postupy analýzy vícerozměrných dat se zvláštním zřetelem na zlepšení jejich rozdělení pomocí mocninné transformace. Kromě série jednorozměrných mocninných transformací je použita také vícerozměrná mocninná transformace. Jsou ukázány rozdíly mezi oběma typy transformací a původními daty pro identifikaci vybočujících bodů, analýzu hlavních komponent a shlukovou analýzu. Většina výsledků je demonstrována graficky.

1.Úvod
Celá řada praktických úloh z oblasti chemie životního prostředí vede na zpracování vícerozměrných výběrů. Podobné problémy se vyskytují také v jiných oborech, kde se zkoumá chování systémů ovlivněných simultánně řadou souvisejících faktorů resp. při konstrukci modelů predikujících úroveň znečištění atd. Vše je komplikováno tím, že se vychází z experimentálních dat , která mají v těchto případech standardně některé specifické zvláštnosti:
(a) rozsahy zpracovávaných dat jsou buď malé nebo extrémně veliké,
(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, upravovat data a hodnotit kvalitu výsledků [1]. Při praktické realizaci však nastávají problémy zejména v případech, kdy se jedná o vícerozměrné úlohy. Již samotné znázornění dat vyžaduje použití různých projekcí, které však vzhledem k multikolinearitě , nelinearitám a dimensi problému nemusí dobře indikovat např. tzv. **vybočující hodnoty** (body), jejichž přítomnost může mít katastrofické důsledky s ohledem na interpretaci výsledků a praktické závěry. Analýza vybočujících dat souvisí úzce s problémem ne normality. Data která se jeví jako vybočující pro normální rozdělení nohou být pro jiná rozdělení přijatelná. To poněkud komplikuje problém transformace dat, kdy může dojít k maskování vybočujících bodů.
Standardně se pro průzkumovou analýzu vícerozměrných dat používá metoda hlavních komponent (PCA), která je dnes běžnou součástí prakticky všech programových systémů pro vícerozměrná data. problémem je, že základním předpokladem PCA je vícerozměrná

normality. Ta je také častým předpokladem dalších vícerozměrných metod. Transformace vedoucí k zlepšení rozdělení dat (ve smyslu přiblížení k normalitě) je tedy na jednu stranu žádaná ale na druhou stranu (přítomnost vybočujících bodů) může vést ke zkreslení dat.
To vede ke stavu, že je mocninná transformace rutinně využívána bez hlubšího rozboru, což může často způsobit potíže tam, kde je vhodné volit alternativní cesty.
V této práci je pojednáno o mocninné transformaci vícerozměrných dat a jejich vlivu na některé metody vícerozměrné statistické analýzy. Kromě série jednorozměrných mocninných transformací je použita také vícerozměrná mocninná transformace. Jsou ukázány rozdíly mezi oběma typy transformací a původními daty pro identifikaci vybočujících bodů, analýzu hlavních komponent a shlukovou analýzu. Většina výsledků je demonstrována graficky.

2. Analýza vícerozměrných dat

Standardním výsledkem měření resp. sběru dat pro případ nestrukturovaných informací je tabulka , která se dá vyjádřit maticí $\boldsymbol{X}$ rozměru ($N \times m$). Řádky matice $\boldsymbol{X}$ často představují jisté *objekty* (vzorky, produkty, odpady, jedince), na kterých se provádí zkoumání. Sloupce matice $\boldsymbol{X}$ představují zkoumané *znaky*, respektive *vlastnosti* (charakteristiky objektů), které se na objektech zkoumají. Vlastní metody vícerozměrné statistické analýzy také závisí na škále, ve které jsou data měřena. Podle množství informace obsažených v jednotlivých škálách jsou na nich definovány různé typy operátorů.
Standardní analýza vícerozměrných dat [1] je založena na analýze matice dat $\boldsymbol{X}$. Podobně jako u jednorozměrných výběrů se zde provádí standardní statistická analýza založená na parametrech polohy (vektoru průměrů) a rozptýlení (kovarianční respektive korelační matici). Zkoumá se přítomnost vybočujících bodů, předpoklady normality a provádějí se standardní statistické testy. Před vlastní aplikací vhodné metody vícerozměrné statistické analýzy je třeba vždy provést *exploratorní (průzkumovou) analýzu dat*, která umožňuje:

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

Pro účely průzkumové analýzy vícerozměrných dat se používá různých technik, umožňujících jejich grafické zobrazení ve dvourozměrném souřadnicovém systému a u novějších výpočetních prostředků i v třírozměrném souřadnicovém systému. Toto zobrazení umožňuje zejména identifikaci objektů nebo jejich složek, které se jeví jako vybočující nebo indikaci různých struktur v datech, které ukazují na heterogenitu použitého výběru nebo přítomnost různých dílčích výběrů s odlišným chováním.
Na základě těchto informací a výsledků testů normality (příp. grafických ekvivalentů těchto testů) pak může být před vlastní statistickou analýzou provedena řada různých korekcí, vedoucích k odstranění nehomogenity výběru a přiblížení se k vícerozměrné normalitě.
Většina používaných technik pro zobrazení vícerozměrných dat se dá zařadit do jedné ze dvou základních skupin, a to *zobecněné rozptylové diagramy* nebo *symbolové grafy*.
Pro základní případ dvojice náhodných znaků ($m = 2$) lze snadno a bez zkreslení konstruovat rozptylové grafy, které umožňují sledovat statistické zvláštnosti dat. Je možné použít i histogramy, neparametrické odhady hustoty a jiné grafy pro konstrukci výběrového rozdělení nebo jeho porovnání s rozděleními teoretickými.
Problémy však nastávají u vícerozměrných dat pro $m > 2$, kdy je možno buď volit několik různých grafů či vhodným způsobem provést redukci na dvou dimenzionální data. V obou případech nelze vyloučit zkreslení způsobené vícerozměrným charakterem dat, které může vést k mylným závěrům.
Jednoduchou možnost dvourozměrného zobrazení m-rozměrných dat představují *profily*. Každý bod $\boldsymbol{x}_i$ je zde charakterizován m vertikálními úsečkami nebo sloupci. Jejich velikost je

úměrná hodnotě odpovídající složky x_{ij} j = 1, ..., m. Na x-ovou osu se vynáší index dané složky j. Profil pak vzniká spojením koncových bodů těchto úseček či sloupců. Je vhodné použít škálované znaky kde max max abs(x_{ij}) je maximální hodnota absolutní velikosti složky x_j vektoru $\boldsymbol{x}$ přes všechny body, i = 1, ..., n. Profily jsou jednoduché a umožňují určení rozdílů mezi jednotlivými body $\boldsymbol{x}_i$, $\boldsymbol{x}_k$ i v dílčích složkách. Snadno lze tedy identifikovat vybočující složku objektu, respektive skupiny objektů, s prakticky shodným chováním.
Ke zjednodušení interpretace a omezení artefaktů se často používá rozptylových grafů v *modifikovaných souřadnicích*, které souvisí se zavedením tzv. *latentních* proměnných resp. vhodnou projekcí vícerozměrných dat do dvou dimenzí. Z řady různých technik, jsou velmi často využívány techniky založené na metodě *hlavních komponent* (principal component analysis, PCA), která je vhodná pro případy, kdy jsou sloupce matice $\boldsymbol{X}$ silně korelovány.
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. Výsledky PCA 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, respektive jiných kombinací hlavních komponent). V případě , kdy vybočující měření tvoří shluky je možné použít také vybrané metody *shlukové analýzy*.

3. Vybočující body

Pojem vybočující body evokuje představu, že jde o body, které lze vizuálně identifikovat 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 odhadů
- 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 vést 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 [4]

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

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 ***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 $e_N = int[(N + p + 1)/2]$

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].

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 pro jiný model akceptovatelné.

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.

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

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

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

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

kde $\alpha_{\mathrm{N}} = (1-\alpha)^{\mathrm{N}}$ 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

$$\mathrm{d}^2_{\mathrm{i}} = (\mathrm{x}_{\mathrm{i}}-\mathrm{x}_{\mathrm{A}})^{\mathrm{T}}\mathrm{S}^{-1}(\mathrm{x}_{\mathrm{i}}-\mathrm{x}_{\mathrm{A}})$$

Jako vybočující se pak identifikují ty body, pro které je $\mathrm{d}_{\mathrm{i}} > \mathrm{c}(\mathrm{p},\mathrm{N},\alpha_{\mathrm{N}})$

Pro případ vícerozměrného normálního rozdělení a velké výběry je $\mathrm{c}(\mathrm{p},\mathrm{N},\alpha_{\mathrm{N}})$ dáno kvantilem chí kvadrát rozdělení $\mathrm{c}(\mathrm{p},\mathrm{N},\alpha_{\mathrm{N}}) = \chi^2_{\mathrm{p}}(1-\alpha/\mathrm{N})$

Pro malé výběry je lépe použít modifikovaný koeficient

$$\mathrm{c}(\mathrm{p},\mathrm{N},\alpha_{\mathrm{N}}) = \frac{\mathrm{p}*(\mathrm{N}-1)^2 * \mathrm{F}_{\mathrm{p},\mathrm{N}-\mathrm{p}-1}(1-\alpha/\mathrm{N})}{\mathrm{N}*(\mathrm{n}-\mathrm{p}-1+\mathrm{p}*\mathrm{F}_{\mathrm{p},\mathrm{N}-\mathrm{p}-1}(1-\alpha/\mathrm{N}))}$$

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

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

Wilks použil pro určení jednoho vybočujícího bodu ve vícerozměrných datech statistiku

$$\mathrm{R}_{\mathrm{i}} = \frac{\det(\mathrm{S}_{\mathrm{i}})}{\det(\mathrm{S})}$$

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 „ x_A a S. Pro robustní odhady se často volí [1,5]:

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

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 d^2_{i-1} vypočítaného v i-1 té iteraci. Z takto získaných odhadů se vypočte vektor opravených zobecněných vzdáleností 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})

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 $y_i = \frac{F_{p,N-p}(0.5)}{\text{median}(d)} d^2_{(i)}$ 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.

4. Mocninná transformace dat

S transformací dat se při zpracování experimentů setkáváme velmi často. Podle příčin můžeme transformaci dělit do dvou základních skupin:

A. Transformace zlepšující rozdělení dat. Zde je transformace žádána a přispívá ke zlepšení rozdělení dat (zjednodušuje jejich zpracování)

B. Transformace jako důsledek matematických operací (obyčejně realizace funkcí) s měřenými veličinami. To je případ, kdy známe u komplikovaných systémů vstupní náhodné veličiny a zajímá nás výstupní náhodná veličina. Patří sem tedy všechny transformace, kdy na základě experimentálních výsledků počítáme jiné veličiny (např. z hodnot poloměru plochu kruhových elementů). Zde je vlastně transformace nežádaná, protože deformuje původní rozdělení dat.

V případě ad A) se hledá vhodná transformace. V případě ad B) se hledají vhodné postupy zpracování dat, které omezují vliv transformace. Tato dualita vede ke stavu, kdy formálně shodné (matematicky správné) metody poskytují značně odlišné výsledky.

Z uvedeného je zřejmé, že transformace může být buď "užitečným nástrojem", nebo "základní překážkou" při statistické analýze dat.

Mocninná transformace je poměrně široce využitelná pro řešení celé řady problémů. Platí, že aditivní i multiplikativní model lze vyjádřit jako speciální případy mocninné třídy modelů která je charakterizována tím, že transformací měřené veličiny x pomocí mocninné funkce $h(.)$ vyjde aditivní model

$$h(x) = h(\mu) + \varepsilon \qquad (1)$$

kde μ je skutečná hodnota měřené veličiny a ε je náhodná chyba měření.

Vhodnou transformací dat lze **stabilizovat rozptyl**, **přiblížit šikmost rozdělení k nule** a tvar rozdělení k **normálnímu rozdělení**. Cílem je na základě znalostí o výběru x_i , $i = 1, ...N$ nalézt vhodnou mocninu, resp. vhodný člen (pokud se použije celá rodina transformací).

Nejjednodušší je prostá mocninná transformace

$hp(x) = sign(x)*abs(x)^P$ pro $P \# 0$

$hp(x) = ln(x)$ pro $P = 0$

kde *abs(x)* je absolutní hodnota a *sign(x)* je znaménková funkce

$sign(x) = 1$ pro $x>0$, $sign(x) = -1$ pro $x<0$, $sign(x) = 0$ pro $x=0$

Tato transformace nezachovává měřítko a ani není vzhledem k všude spojitá. Zachovává však pořadí dat ve výběru (jako všechny mocninné transformace).

Používá se jako jednoduchá symetrizující transformace a proto se hledá optimální mocnina tak, aby byly minimalizovány vhodné míry symetrie výběru. Je možno použít přímo výběrovou šikmost $g_1(y)$, nebo její robustní verzi $g_{R1}(y)$ [1]. Stejně jednoduché je sledovat rozdíl mezi průměrem a mediánem v transformaci.

Pro posouzení kvality transformace, resp. nalezení optimálního je také možno použít grafu rozptýlení s kvantily (GRK), resp. kvantilových grafů (Q -Q grafů), jejichž konstrukce je popsána v [1].

Nevýhody prosté mocninné transformace (zejména nespojitost v okolí nuly a nesrovnatelnost měřítek v transformaci) odstraňuje rodina Box-Coxových transformací h(x), která je lineární transformací prosté mocninné transformace *hp(x)*. Box Coxova třída polynomických transformací má tvar

$$h(x) = \frac{x^\lambda - 1}{\lambda} \quad \lambda \neq 0 \qquad (2)$$

$$h(x) = ln(x) \qquad \lambda = 0$$

kde λ je parametr transformace. Pro $\lambda = 1$ resultuje aditivní model měření a pro $\lambda = 0$ model multiplikativní. S využitím Taylorova rozvoje lze odvodit, že v tomto případě je

$$x \approx \mu + \varepsilon / \mu^{1-\lambda}$$

Pro případ, že rozptyl $D(\varepsilon) = \sigma^2$ je malý jde o aditivní model s nekonstantními chybami, pro který lze použít jako odhad μ vážený aritmetický průměr s vahami úměrnými $\mu^{-(1-\lambda)/2}$.

Lze ukázat, že vhodným odhadem parametru μ (neznámá koncentrace) je výběrový medián, který je invariantní vůči monotónní transformaci dat.

Prostá mocninná transformace je invariantní vůči změně měřítka a Box Coxova transformace není invariantní vůči změně měřítka. Detaily lze nalézt v práci [6]. Pro eliminaci této nevýhody lze použít modifikované transformace

$$h(x,p) = \frac{x^\lambda - p^\lambda}{\lambda} \quad \lambda \neq 0$$

$$h(x,p) = \ln(x/p) \qquad \lambda = 0$$

kde parametr p se volí jako aritmetický průměr, geometrický průměr resp. medián původních dat. Z uvedeného také přímo plyne, že obě transformace jsou závislé na posunu. Tedy mocninná transformace (x+a) poskytne jiné výsledky než mocninná transformace *x*.

Lze se snadno přesvědčit, že:

- rodina transformací definovaných rovnicí (2) je vzhledem k mocnině λ spojitá. V okolí nuly platí $lim\ (x - 1)^\lambda / \lambda = lim\ x\ .\ ln(x) = ln(x)$

všechny transformační závislosti *h(x)* procházejí jedním bodem o souřadnicích y = 0, x = 1 a mají v tomto bodě společnou směrnici (jsou zde, co do průběhu, shodné)

- Mocninné transformace s exponenty -2, -3/2, -1, -0,5, 0, 0,5, 1, 3/2, 2 jsou co do křivosti rovnoměrně rozmístěné.

- Vlivem transformace (2) se však obecně mění charakteristiky polohy a rozptýlení, což komplikuje porovnání různě transformovaných výběrů (nevadí pochopitelně pro přiblížení k normalitě, resp., zesymetričtění výběru).

Pro zajištění toho, aby měla transformovaná data přibližně stejnou polohu a rozptýlení jako data netransformovaná, je možné použít dostatečné lineární transformace (viz [2]).

Z hlediska analýzy dat je transformace vždy žádoucí, pokud $x_{(N)} / x_{(1)} > 20$ (kde se předpokládají kladná data). Rovnice (2) je použitelná pouze pro kladná data. Pokud je znám jiný počátek x_0, pod kterým se data nemohou vyskytovat, volí se zobecněná mocninná transformace

$$h(x) = \frac{(x+c)^{\lambda} - 1}{\lambda} \qquad \lambda \neq 0$$

$$h(x) = \ln(x+c) \qquad \lambda = 0$$

Zde $c > x_0$. Obecně se hledají u této transformace dva parametry. S ohledem na to, že dosavadní transformace platí pro zdola omezené rozdělení dat, není zřejmě možné, aby jejich rozdělení bylo striktně normální. Pro odstranění této (prakticky nepříliš důležité) nevýhody doporučují Bickel a Doksum rozšířenou Box-Coxovu transformaci (pro parametr $\lambda > 0$), která pokrývá celou reálnou osu

$$h(x) = \frac{sign(x) * abs(x)^{\lambda} - 1}{\lambda} \qquad \lambda \neq 0$$

Nevýhodou je, že tato transformace neobsahuje logaritmickou transformaci. Tato transformace je již nezávislá na měřítku. Yeo a Johnson [3] navrhli mocninnou transformaci platnou pro libovolné hodnoty *x*. Tato transformace má tvar

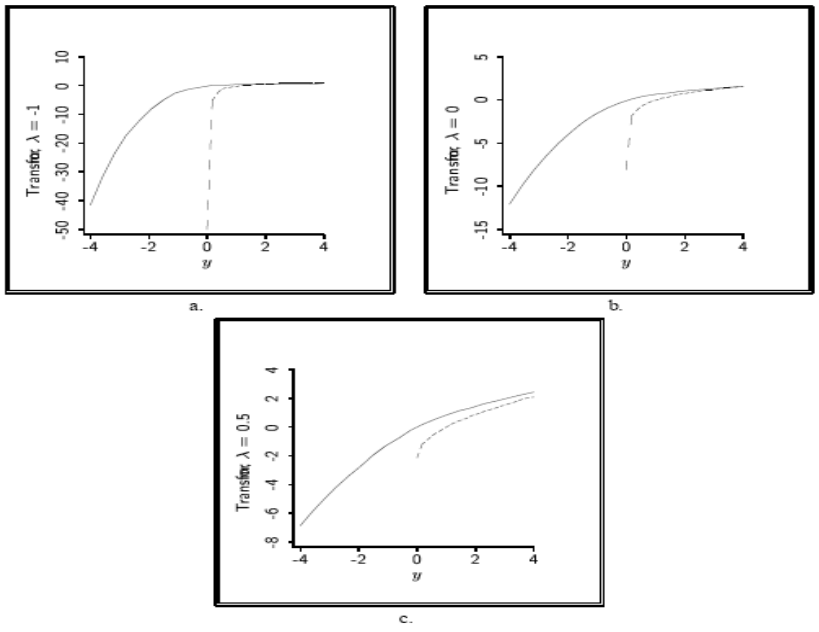
$$h(x) = \left((x+1)^{\lambda} - 1\right)/\lambda \quad \text{pro } \lambda \neq 0 \; x \geq 0$$

$$h(x) = \ln\left(x+1\right) \quad \text{pro } \lambda = 0 \; x \geq 0$$

$$h(x) = -\left((-x+1)^{2-\lambda} - 1\right)/(2-\lambda) \quad \text{pro } \lambda \neq 2 \; x \leq 0$$

$$h(x) = -\ln\left(-x+1\right) \quad \text{pro } \lambda = 2 \; x \leq 0$$

Na obr 1 je porovnán Box Coxova a Yeo Johnsonova transformace pro tři typické parametry transformace.



Obr. 1 Box Coxova (čárkovaně) a Yeo Johnsonova transformace pro různá λ (-1, 0, 0.5)

Pro odhady parametrů v těchto rodinách transformací lze opět použít různých charakteristik šikmosti a špičatosti.

V případě **jednoparametrických rodin transformací** se lze zaměřit pouze na jednu charakteristiku tvaru (obyčejně šikmost). Výhodnější je použití testů normality dat po mocninné transformaci. Známý Shapiro-Wilkův test je úměrný testu významnosti směrnice v Q-Q grafu, takže lze také posuzovat linearitu v Q-Q grafech.

S ohledem na požadavek, aby se rozdělení výběru v transformaci co nejvíce blížilo normálnímu rozdělení, lze pro odhad optimálního použít metodu maximální věrohodnosti.

Pokud platí předpoklady aditivního modelu měření (normalita a nezávislost) má logaritmus věrohodnostní funkce tvar

$$ln\,L(\lambda) = \sum(\lambda - 1) * ln(x_i) - \frac{1}{2\sigma^2}\sum\left[h(x_i) - h(\mu)\right]^2 \qquad (3)$$

Pro pevné λ lze určit maximálně věrohodný odhad rozptylu ve tvaru

$$\sigma_c^2 = \frac{1}{N}\sum\left[h(x_i) - h(\mu)\right]^2 \qquad (4)$$

kde se za $h(\mu)$ dosazuje aritmetický průměr transformovaných dat

$$h(\mu) \approx \frac{1}{N}\sum h(x_i) \qquad (5)$$

Po dosazení do věrohodnostní funkce resultuje vztah

$$ln\,L^*(\lambda) = \sum(\lambda - 1) * ln(x_i) - \frac{N * ln\,\sigma_c^2}{2} \qquad (6)$$

Maximalizací $ln\,L^*(\lambda)$ podle λ (viz.[1]) lze pak snadno určit maximálně věrohodný odhad $\hat{\lambda}$ parametru transformace λ. Je patrné, že je tato úloha ekvivalentní minimalizaci rozptylu v transformovaných proměnných σ_c^2. Na základě Taylorova rozvoje funkce *h(x)* pro pevné vyjde přibližný výraz

$$D\left(\frac{x^\lambda - 1}{\lambda}\right) = \frac{1}{\lambda^2} D(x^\lambda) \approx E(x)^{2\lambda-2} D(x) = E(x)^{2\lambda}\,\delta^2$$

kde δ je variační koeficient. Je zřejmé, že pro pevné λ bude rozptyl v transformaci tím vyšší, čím bude větší rozptýlení dat. To umožní identifikaci extrému (minima). Pro málo rozptýlená data bude rozptyl v transformaci malý a identifikace extrému bude obtížnější. V práci [7] bylo ukázáno, že pro D(x) ~ 0 roste rozptyl zobecněného průměru nade všechny meze. Pro snadnou identifikovatelnost transformace je tedy výhodné mít větší rozptýlení dat jak je např. běžné u výběrů z asymetrických rozdělení.

Formálně lze úlohu maximalizace rov (6) vyjádřit ve tvaru

$$\frac{d\,ln(L)}{d\lambda} = \sum_i ln(x_i) - \frac{1}{\sigma^2}\sum_i\left(h(x_i) - \frac{1}{N}\sum_i h(x_i)\right) * \frac{dh(x_i)}{d\lambda} = 0 \qquad (7)$$

kde $\frac{dh(x_i)}{d\lambda} = \frac{(1 + \lambda * x_i)\,ln(1 + \lambda * x_i) - \lambda * x_i}{\lambda^2}$

Z druhé derivace věrohodnostní funkce lze určit rozptyl maximálně věrohodného odhadu mocninné transformace [8] .

Na základě asymptotického $(1-\alpha)$ % ního intervalu spolehlivosti parametru mocninné transformace lze sestavit nerovnost

$$ln\,L(\lambda) \geq ln\,L(\hat{\lambda}) - 0.5 * \chi^2_{1-\alpha}(1) \qquad (8)$$

Všechna λ splňující tuto nerovnost leží v intervalu spolehlivosti a jsou tedy přijatelná. Toho lze snadno využít pro rozlišení mezi aditivním a multiplikativním modelem měření. V rovnici (8) označuje $\chi^2_{1-\alpha}(1)$ kvantil chí kvadrát rozdělení s 1 stupněm volnosti.

Platí, že:

- pokud obsahuje 95% ní interval spolehlivosti také jedničku, volí se aditivní model.
- pokud obsahuje 95% ní interval spolehlivosti nulu a nikoliv jedničku, volí se multiplikativní model.
- v ostatních případech je možné zvolit pravděpodobnostní model (1) a použít pro další analýzu postup navržený v [1].

S výhodou lze využít grafického záznamu ln L(λ) na se zakresleným (obyčejně 95 %ním) konfidenčním intervalem. Z takového grafu lze již snadno odhadnout jak kvalitu transformace, tak i posoudit, v jakých mezích se může hodnota λ pohybovat. (Platí, že čím jsou tyto meze užší, je kvalita transformace vyšší, pokud v nich neleží $\lambda = 1$).

Parametr mocninné transformace zřejmě souvisí s šikmostí rozdělení dat. V práci [9] je toto odvození provedeno. Výsledek lze zapsat ve tvaru

$$\lambda \approx 1 - \frac{E(x) * \sigma * \beta_1}{6} \qquad (9)$$

kde β_1 je šikmost původních dat. Je patrné, že pro data zešikmená k vyšším hodnotám vyjde parametr transformace podstatně menší než jedna.

Pomocí vztahu (30) můžeme např. snadno posoudit vliv posunu dat na parametr mocninné transformace. Např. pro případ, že data posuneme o konstantu a tj. $y = a*x$ vyjde, že

$$\lambda_y = \lambda_x - (a * \sigma * \beta_1) / 6$$

Jak je patrné, je třeba při použití postupu mocninné transformace brát v úvahu také případné lineární transformace dat a jejich rozmezí.

Speciálně pro účely průzkumové analýzy dat (viz. [1]) byl navržen postup, který umožňuje grafické posouzení vhodnosti mocninné transformace. Je použito jednoduché třídy transformací typu

$$h(x) = a * x^{\lambda} + b \quad \lambda \neq 0 \qquad (10)$$
$$h(x) = c * \ln(x) + d \qquad \lambda = 0$$

Parametry a, b, c, d volí Emerson a Stotto [10] tak, aby byla zachována přibližná linearita transformace v okolí mediánu, tj.

$$med(x^{\lambda}) \approx med(x) \qquad \frac{d}{dx}(med(x^{\lambda}) \approx 1$$

Pro určení vhodné transformace se vychází z výběrových kvantilů x_p a mediánu $x_{0,5}$.

Vynesením $y^* = (x_p + x_{1-p}) / 2$ na $x^* = [(x_{1-p} - x_{0,5})^2 + (x_{0,5} - x_p)^2] / (4\, x_{0,5})$ rezultuje v případě možnosti symetrizační transformace lineární závislost, procházející počátkem typu $y^* = (1 - \lambda)\, x^*$. Ze směrnice této závislosti tedy můžeme přímo nalézt odhad parametru transformace λ. Při praktické aplikaci tohoto postupu se volí jednotlivé písmenové hodnoty (viz [1]), pro které je $P_i = 2^{-(i+1)}$, i = 1, ... Pro robustní odhad směrnice $(1 - \lambda)$ se doporučuje počítat pro všechny body směrnice $k_i = y_i^*/x_i^*$ a jako optimální vzít pak medián ze všech k_i.

Uvedený postup je vhodný pro málo a středně sešikmená rozdělení. Cameron [11] ukázal, že pro silně sešikmená rozdělení a kvantily x_p vzdálené od mediánu vzniká na grafu y^* vs , x^* systematická křivost. Pak je vhodné provádět iterativní hledání optimálního , kdy se výsledek z prvého určení směrnice (y^* vs . x^*) dosadí do transformace (10) a v dalších vyneseních se místo kvantilů proměnné *x* používají transformované kvantily *h(x)* (určené z předchozího grafu). Také je výhodné v prvních fázích brát spíše směrnice určené z kvantilů ($P_i = 0{,}25$). Z toho plyne, že Emerson-Stottův postup není zcela automatický a vyžaduje často iterativní hledání vhodného λ, kde v každé iteraci se konstruuje graf typu y^* na x^*. Na druhé straně je

tento postup velmi jednoduchý a umožňuje posouzení vlivu případných vlivných bodů na výsledek transformace.
Pro případ vícerozměrných dat existují dvě možnosti:

- marginální mocninná transformace (MBC), kdy se použije vhodná mocninná transformace pro jednotlivé vektory dat ($\mathbf{x}_1$, … $\mathbf{x}_m$), kde. $\mathbf{x}_i$ je i- tý vektor s *n* složkami.
- sdružená mocninná transformace (JBC), kdy se hledá vektor mocninné transformace $\boldsymbol{\lambda} = (\lambda_1 \ldots . \lambda_m)$ maximalizací kondenzované věrohodnostní funkce PL

$$PL = -\frac{\ln(\det(\mathbf{S}(\boldsymbol{\lambda})))}{n} + \sum_{j=1}^{m}\left((\lambda_j - 1)\sum_{i=1}^{n}\ln(x_{ij})\right)$$

Kovarianční matice $\mathbf{S}(\boldsymbol{\lambda})$ se počítá pro konkrétní $\boldsymbol{\lambda}$ standardním způsobem.
Sdružená mocninná Box Coxova transformace vede obecně k efektivnějším odhadům vektoru transformace $\boldsymbol{\lambda}$. Na druhé straně jsou numerické hodnoty složek vektoru transformace blízké [12].

5. Experimentální část

K analýze dat byl použit výběr obsahující koncentrace 11 prvků (Ni, Cu, Cr, Zn, Cd, Pb, Be, Co, Mo, V, As) v 80-ti vzorcích půdy z průmyslových oblastí Moravy. Pro analýzu dat byl použit software R, který je volně dostupný na síti . Protože je účelem ukázat na rozdíly mezi původními daty a MBC resp. SBC jsou diskutovány pouze statistické souvislosti. Pro výpočet parametrů mocninné transformace byla použita nelineární optimalizační metoda (derivační).
V tab.1 jsou uvedeny parametry mocninné transformace vypočítané různými postupy. Ve druhém řádku jsou odhady MBC, ve třetím řádku SBC a ve čtvrtém řádku jsou odhady pro simultánní Yeo Johnsonovu transformaci.

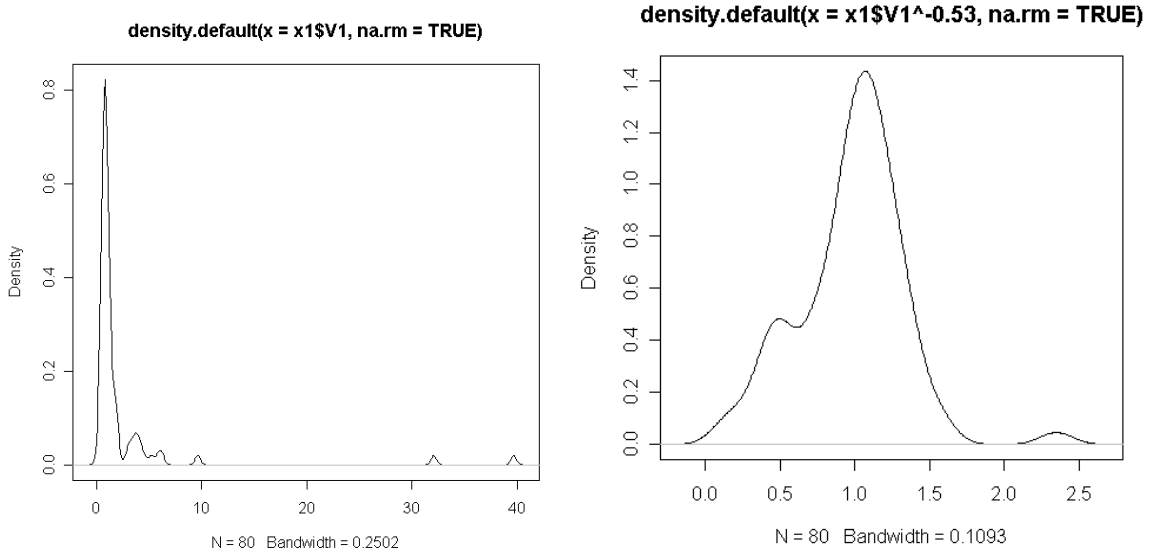
Tabulka 1 Odhady optimální mocninné transformace

Cd	**Pb**	**Ni**	**Cu**	**Co**	**Cr**	**Zn**	**As**	**Be**	**V**	**Mo**
-0,51	-0,70	-1,22	-0,96	-0,46	-1,80	-0,72	-1,02	0.051	0.93	-0.08
-0,53	-0.73	-1.35	-0,88	-0,39	-1,54	-0,69	-0,78	0,33	1.15	-0.097
-1,62	-0,72	-1,42	-0,91	-0,56	-1,59	-0,67	-0,89	-0.,34	1,19	-1,27

Pomocí těchto mocninných (kromě Yeo Johnsonovy) byl původní výběr re transformován na výběr MBC resp. SBC. V dalším jsou analyzovány původní výběr (**PV**), výběr MBC (**MV**) a výběr SBC (**SV**). Pro jednoduchost nebyla použita celá Box Coxova transformace ale pouze prostá mocnina (která je její lineární transformací).

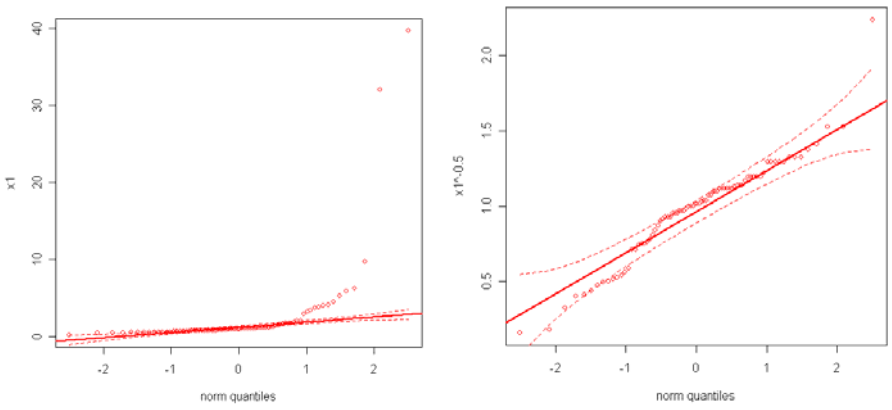
6. Výsledky a diskuse

Použití mocninné transformace vedlo v naprosté většině případů k výraznému zlepšení symetrie rozdělení dat. Na obr. 1 je ukázán neparametrický odhad hustoty pravděpodobnosti pro prvek Cd před (PV) a po simultánní (SV) mocninné transformaci Na obr. 2 jsou odpovídající rankitové grafy (qq graf pro normální rozdělení). Je patrné výrazné zlepšení, i když jsou některá měření stále indikována jako vybočující.
Jako příklad situace, kdy mocninná transformace příliš rozdělení dat neovlivnila je na obr. 3 ukázán neparametrický odhad hustoty pravděpodobnosti pro prvek Cd před (PV) a po simultánní (SV) mocninné transformaci. Na obr. 4 jsou odpovídající rankitové grafy (qq graf pro normální rozdělení). Je patrné, že i po transformaci mají data rozdělení systematicky odchýlené od normálního rozdělení.



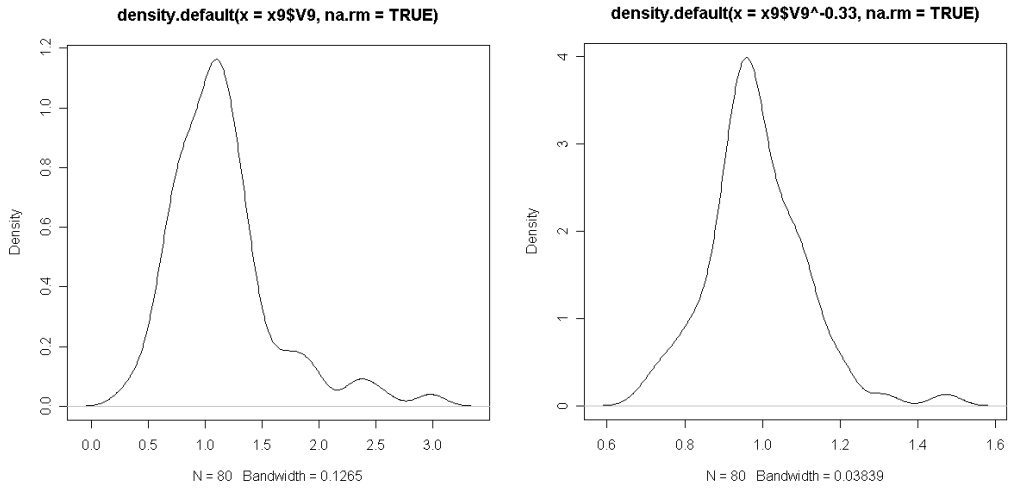
a b

Obr. 1 Hustota pravděpodobnosti prvku Cd pro a) PV a b) SV



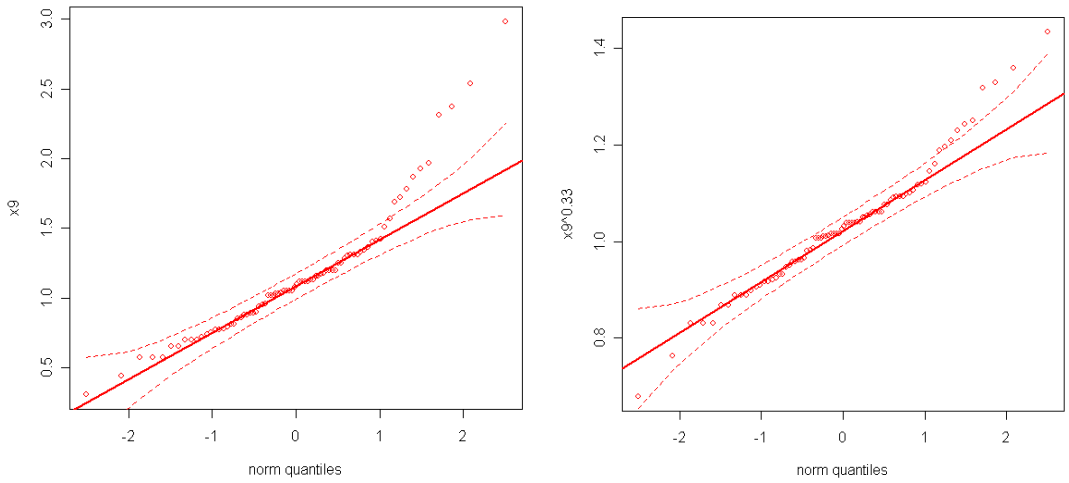
a b

Obr. 2 Rankitový graf prvku Cd pro a) PV a b) SV



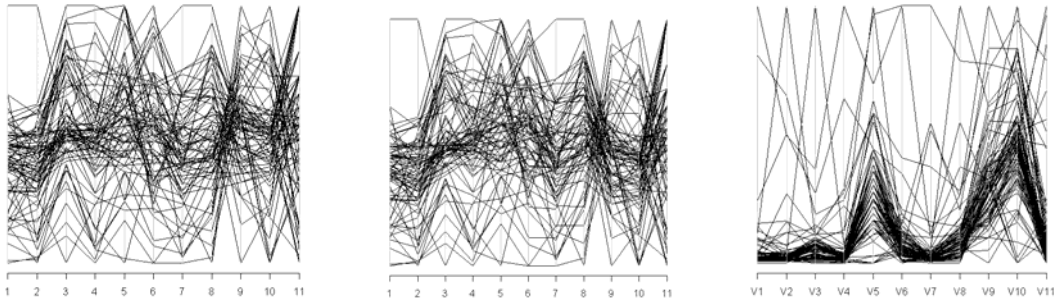
a b

Obr. 3 Hustota pravděpodobnosti prvku Be pro a) PV a b) SV



Obr. 4 Rankitový graf prvku Be pro a) PV a b) SV

Na obr. 5 jsou ukázány profilové grafy pro všechny tři výběry



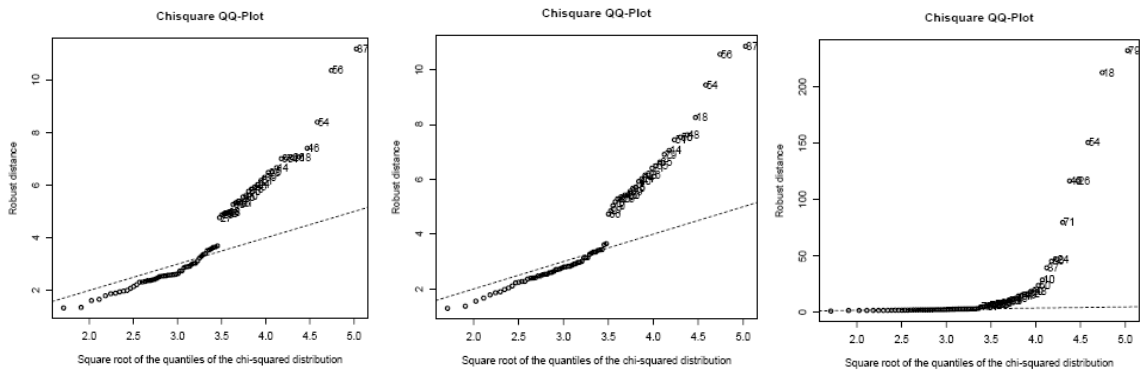
a b c

Obr. 5 profilové grafy pro a) SV, b) MV, c) PV

Je patrné (obr, 5c), že původní výběr se skládá z části podobných dat a části anomálních dat. Po transformaci se tato tendence částečně překrývá.

Identifikace vícerozměrných vybočujících měření byla provedena pomocí robustních Mahalanobisových vzdáleností.

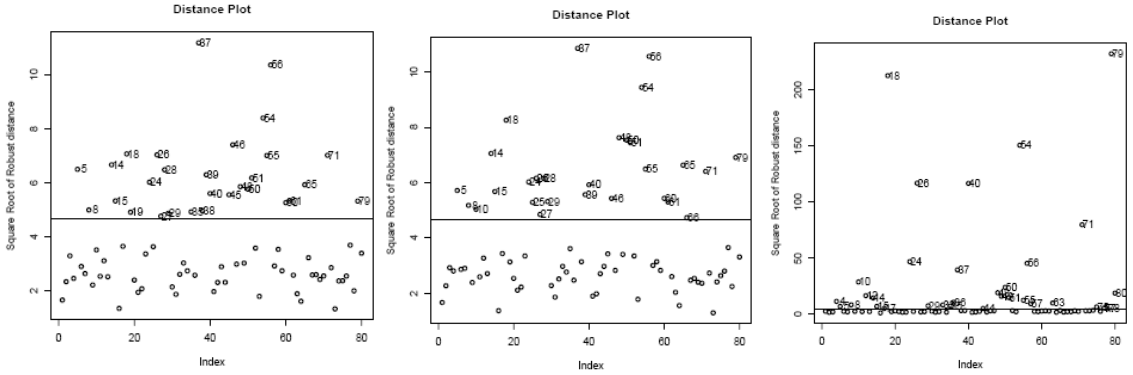
Na obr. 6 jsou znázorněny qq grafy pro robustní Mahalanobisovu vzdálenost.



a b c

Obr. 6 QQ grafy robustních Mahalanobisových vzdáleností pro a) SV, b) MV, c) PV

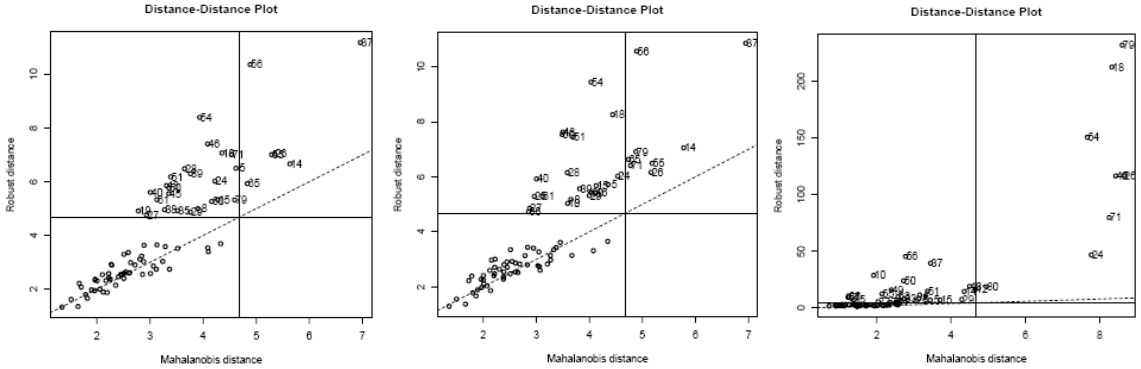
Na obr. 7 jsou znázorněny indexové grafy pro robustní Mahalanobisovu vzdálenost.



a b c

Obr. 7 Indexové grafy robustních Mahalanobisových vzdáleností pro a) SV, b) MV, c) PV

Na obr. 8 jsou znázorněny grafy robustní Mahalanobisovu vzdálenost vs. robustní Euklidova vzdálenost.



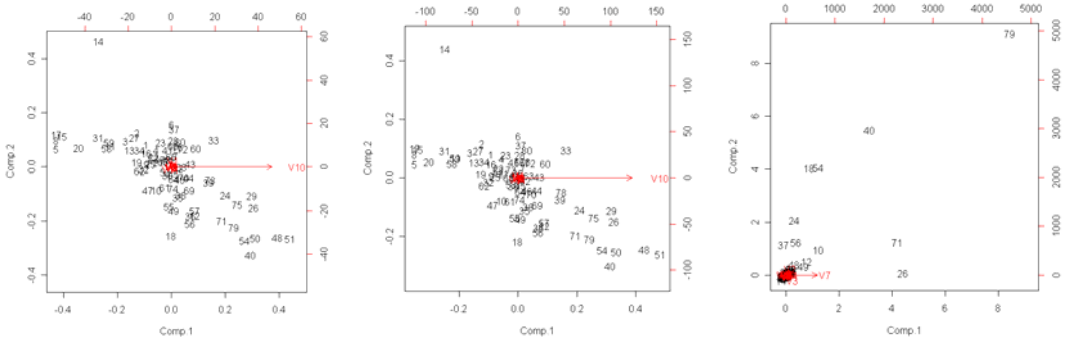
a b c

Obr. 8 Grafy robustní Mahalanobisova vzdálenost vs. robustní Euklidova vzdálenost pro a) SV, b) MV, c) PV

Je opět patrné, že transformace výrazněji odděluje dvě skupiny a částečně maskuje vybočující měření. Je zajímavé sledovat jak, se mění pořadí významnosti vybočujících bodů podle velikosti zobecněné vzdálenosti. Indexy nejvíce vybočujících bodů pro data PV jsou:
79 18 54 26 40 71 24 56 37 10 50 48 80 12 49 51 14 55 4 36 63 57 8 33 29 35 15 5 75 77 44
pro data MV jsou:
37 56 54 18 48 50 51 14 79 65 55 71 26 28 24 40 5 15 39 60 46 29 61 25 8 10 27 66 77 35 17
a pro data SV jsou:
37 56 54 46 18 26 71 55 14 5 28 39 51 24 65 48 50 40 45 61 15 79 60 8 38 35 19 29 27 77 17
Je patrné, že pořadí vybočujících bodů se liší i pro obě varianty Box Coxovy transformace.
Na obr. 9 je ukázán dvojný graf pro klasickou analýzu hlavních komponent (PCA) a standardizovaná data. Je vidět, že standardizace zvětšuje rozptýlení dat a mocninná transformace opět částečně maskuje vybočující hodnoty.
Na obr. 10 je ukázán dvojný graf pro robustní analýzu hlavních komponent (PCA) a nestandardizovaná data. Je vidět, že bez standardizace se zmenšuje rozptýlení dat a mocninná transformace opět částečně maskuje vybočující hodnoty.

a b c

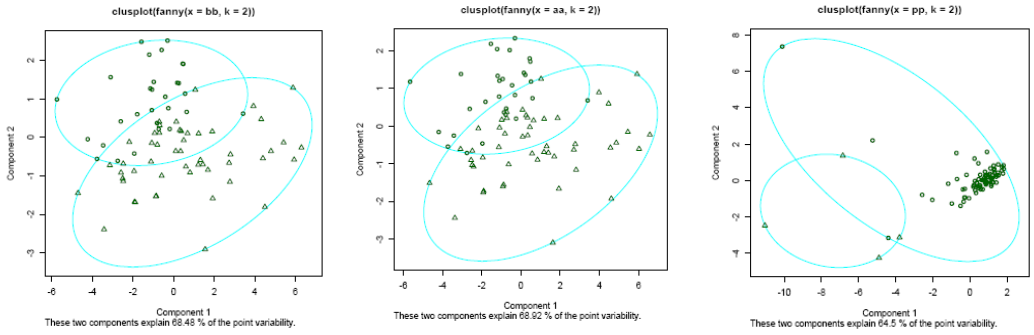
Obr. 9 Dvojné grafy standardizované PCA pro a) SV, b) MV, c) PV



a b c

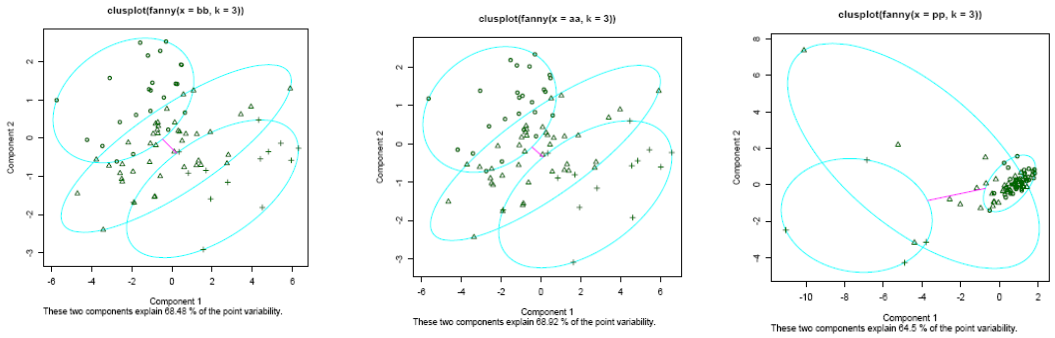
Obr. 10 Dvojné grafy robustní nestandardizované PCA pro a) SV, b) MV, c) PV

Na obr. 11 je znázorněn výsledek „fuzzy“ shlukování pro případ dvou shluků a na obr. 11 výsledek pro tři shluky.



a b c

Obr. 10 Dva „fuzzy“ shluky pro a) SV, b) MV, c) PV



a b c

Obr. 10 Tři „fuzzy“ shluky pro a) SV, b) MV, c) PV

Opět je patrné, že pro netransformovaná data jsou lépe odděleny vybočující body a pro mocninné transformace se projevuje výraznější rozptýlení dat (dvě až tři nepříliš jasně oddělené skupiny). Jsou patrné rozdíly mezi marginální a simultánní Box Coxovou transformací. Např. pro PV jsou dva shluky velikosti 76, 4, pro MV jsou dva shluky velikosti 58, 22 a pro MV jsou dva shluky velikosti 42, 38.

7. Závěr

Je patrné, že oba postupy výpočtu parametru mocninné transformace vedou k poněkud odlišným výsledkům. Podle očekávání maskuje mocninná transformace body, které se jeví jako vybočující s ohledem na normalitu. Bohužel však tato jedno parametrová mocninná transformace neumožňuje akceptovat všechna data a má celou řadu specifických zvláštností.
Vždy je lépe použít simultánní mocninná transformace, která je efektivnější a v některých případech statisticky odlišná od marginálních transformací.
V řadě případů je tedy třeba iterativně řešit problém anomálií a vybočujících hodnot s ohledem na cíl zpracování. Pokud jde o indikaci extrémů jde o zcela jinou úlohu než pokud se hledá „centrální tendence“.

Poděkování:

Tato práce vznikla s podporou výzkumného centra Textil, projekt č. 1M4674788501

9. Literatura

[1] Meloun M., Militký J.*: Zpracování experimentálních dat*, Academia Praha 2004
[2] Militký J., Meloun M.: Konference *Mikroelementy "99* ,Řež u Prahy , listopad 1999
[3] Yeo K., Johnson R.: Biometrika, **87,** 954 (2000)
[4] Barnett V., Lewis T.: Outliers in statistical data , 3rd. Ed., Wiley, Chicheter 1994
[5] Campbell N.A.: Appl. Statist. 29, 231 (1980)
[6] Schlesselman J.: *J. Roy Stat. Soc*. **B33**, 307 (1971)
[7] Bickel P.J., Doksum K.A.: *J. Amer. Stat Assoc.* **76**, 296 (1981)
[8] Draper N.R., Cox D. R.: *J. Roy Stat. Soc*. **B31**, 472 (1969)
[9] Box G. E. P., Cox D. R.: *J. Roy Stat. Soc*. **B26**, 211 (1964)
[10] Emerson J.D., Stotto M.A: *J.Amer.Stat.Assoc*. **77**, 103 (1982)
[11] Cameron M.: *J.Amer. Statist. Assoc*. **79**, 107 (1984)
[12] Andrews D. F. a kol. : Biometrics **27**, 825 (1971)