POČÍTAČOVĚ INTENZIVNÍ METODY VE ZPRACOVÁNÍ VÝSLEDKŮ ANALYTICKÝCH MĚŘENÍ

JIŘÍ MILITKÝ ,
Katedra textilních materiálů, Technická universita v Liberci,
461 17 Liberec

MILAN MELOUN,
Katedra analytické chemie, Universita Pardubice, Pardubice

Souhrn

Příspěvek je zaměřen na problematiku zpracování výsledků měření z oblasti životního prostředí. Jsou popsány základní postupy počítačové intenzívní analýzy jednorozměrných výběrů vycházející z principu generace simulovaných výběrů (Bootstrap). Je ukázáno použití této techniky pro konstrukci intervalu spolehlivosti střední hodnoty pro případ asymetrických rozdělení, resp. rozdělení výrazně odlišných od normálního.

1. ÚVOD

Zpracování dat v analytické praxi využívá kombinace poznatků klasické *analytické chemie*, *matematické statistiky a informatiky* na jedné straně a speciálních postupů. *chemometrie* na straně druhé. Důležitou součástí analýzy dat jsou metody k získávání relevantních informací z experimentů a pozorování.

Stále větší počet výkonných osobních počítačů třídy PC podporuje na pracovištích trend decentralizace a interaktivnosti při zpracování experimentálních dat a interpretaci výsledků. To klade větší nároky na pracovníky, kteří již těžko obhájí jednoduché postupy vyhodnocování dat, založené mnohdy na zjednodušených nebo i nesprávných předpokladech. Nabídka a možnosti počítačově orientovaného statistického zpracování dat nutí experimentátora k hlubší analýze, což vede většinou i k radikální změně pohledu na rutinně prováděnou výzkumnou práci.

Existuje celé spektrum méně či více dokonalých a komplexních programů a programových systémů pro statistické vyhodnocování dat. Jiné jsou budovány jako univerzálně použitelné, i když zaměřené na specifické oblasti (chemometrie, biometrie, ekonometrie, medicínská statistika, obchodní statistika, statistika pro sociology, psychology, atd.). Přes současnou dostupnost personálních počítačů prakticky ve všech laboratořích (i domácnostech) se jejich využití omezuje na metody, které byly běžně používány v předpočítačové éře. To vede k omezení chyb lidského subjektu, nahrazení rutinních výpočtů strojem a zejména urychlení analýzy dat. Na druhé straně však počítač nepřináší nové informace a v konečném důsledku se stává práce etapa vyhodnocení experimentů nejslabším článkem metrologického řetězce. Klasickým případem, kdy je počítač nenahraditelný v metrologickém řetězci jsou počítačové intenzívní metody. Tyto metody jsou výhodné pro zpracování dat, kde je á priori možno předpokládat, že nebudou splněny podmínky pro klasickou statistickou analýzu. To jsou např. data v oblasti monitorování životního prostředí, kde:

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

(c) rozdělení dat jen zřídka odpovídá normálnímu běžně předpokládanému ve standardní statistické analýze,
(d) v datech se vyskytují podezřelá 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"

S výhodou je možné využívat i alternativních postupů statistické analýzy včetně robustních, počítačově intenzivních a adaptivních metod.

V tomto sdělení jsou na příkladu intervalu spolehlivosti střední hodnoty ukázány základní principy metody Bootstrap využívající simulovaných výběrů. Vychází se z N-tice výsledků experimentů, t.j.dat $\{x_i\}$ i = 1,...N. Je ukázáno jak efektivně realizovat Bootstrap výběry v jazyce MATLAB. Celý postup je demonstrován na jednoduchém příkladu. S ohledem na rozsah příspěvku jsou vynechány detaily a odvození . Jejich přehled je uveden v knize [3]

2. Metoda BOOTSTRAP

Výše uvedené zvláštnosti dat z oblasti monitorování životního prostředí se projevují na asymetrii výběrového rozdělení. Ta pak omezuje použití různých technik založených na průzkumové analýze a identifikaci vybočujících měření. Také robustní techniky obyčejně selhávají, protože eliminují extrémy, které zde nejsou chybami ale důsledkem zešikmení rozdělení dat.

Je známo, že pro konstrukci intervalu spolehlivosti populačního parametru p_s je třeba znát rozdělení $g(p)$ jeho odhadu p. Pro některá rozdělení (např. normální) a parametry (střední hodnota, rozptyl) jsou rozdělení odhadů nebo jejich funkcí známy a intervaly spolehlivosti je možné konstruovat relativně snadno. Pro odhad intervalu střední hodnoty z aritmetického průměru x_A a výběrového rozptylu s^2 není normalita tak striktní požadavek.. Je známo, že pokud zpracovávaný výběr velikosti N prochází z ne - normálního rozdělení se střední hodnotou μ a rozptylem σ^2 má tzv. Studentova náhodná veličina

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

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

$$P(LC \le \mu \le UC) = 1 - \alpha \qquad (2)$$

Symbol $P(.)$ označuje pravděpodobnost a α je tzv. hladina významnosti. Obyčejně se volí α = 0.05 nebo α = 0.01 s tím, že čím je α menší, tím je interval (LC, UC) širší. . Pokud není σ^2 známo lze použít vztah

$$x_A - t_{1-\alpha/2}(N-1) * \frac{s}{\sqrt{N}} \le \mu \le x_A - t_{\alpha/2}(N-1) * \frac{s}{\sqrt{N}} \qquad (3)$$

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 má interval (3) přesně $100(1-\alpha)$ % ní pokrytí střední hodnoty.To znamená, že jen v $100\alpha/2$ % případů je *střední hodnota menší než CI* (nejistota *NP* zprava) a v $100\alpha/2$ % 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

Pro neznámé rozdělení výběru $\boldsymbol{x} = (x_1..x_N)$ a libovolný parametr *ps* lze s výhodou použít techniku Bootstrap, která umožňuje jak nalezení rozdělení výběrové statistiky *p*, tak i konstrukci intervalu spolehlivosti. Základní myšlenka metody Bootstrap je jednoduchá[8,9]. Spočívá v generaci *M*-tice simulovaných výběrů $\boldsymbol{v}_1..\boldsymbol{v}_M$ označovaných jako Bootstrap výběry. Jejich rozdělení odpovídá rozdělení původního výběru $\boldsymbol{x}$, charakterizovaného hustotou pravděpodobnosti *g(x)*. Z těchto výběrů se určí *M*-tice odhadů $p_i = p(\boldsymbol{x})$ hledaného parametru *ps* . Z této *M*-tice hodnot lze počítat intervaly spolehlivosti pomocí celé řady metod.

2.1 Odhad z asymptotické normality

Jde o nejjednodušší postup založený na představě, že *M* je dostatečně veliké a p_i $i = 1..N$ lze zpracovat jako výběr z normálního rozdělení. Pro tzv. Bootstrap odhad střední hodnoty parametru ps platí

$$p_B = \frac{1}{M}\sum_{i=1}^{M} p_i \qquad (4)$$

a odpovídající rozptyl má tvar

$$s_B^2 = \frac{1}{M}\sum_{i=1}^{M} (p_i - p_B)^2 \qquad (5)$$

Pro *100(1-a)* %ní interval spolehlivosti parametru *ps* se pak použije známý vztah

$$p_B - u_{1-\alpha/2} * s_B \le ps \le p_B + u_{1-\alpha/2} * s_B \qquad (6)$$

kde $u_{1-\alpha/2}$ je kvantil normovaného normálního rozdělení.

2.2 Percentilový odhad

Tento postup je založen na neparametrickém odhadu mezí intervalu spolehlivosti vycházejícím z pořádkových statistik $p_{(i)}$,kde $p_{(i)} <= p_{(i+1)}$ jsou pořádkové statistiky, pro které platí, že jsou *d* %ním kvantilem rozdělení odhadu *p* pro

$$d = \frac{i}{M+1}$$

Dolní mez *100(1-a)* %ní intervalu spolehlivosti je pak

$$LC = p_{(k1)} \text{ kde } k1 = int[\alpha * (M+1)/2] \qquad (7)$$

a pro horní mez platí

$$UC = p_{(k2)} \text{ kde } k2 = int[(1-\alpha/2)*(M+1)] \quad (8)$$

Zde *int (x)* je celá část čísla *x*.

2.3 Studentizovaný odhad

Tento odhad vychází z jednoduché transformace vedoucí na Studentizovanou náhodnou veličinu t_i

$$t_i = \frac{p_i - p_B}{s_{Bi}}$$

kde s_{Bi} je výběrová směrodatná odchylka počítaná pro *i* - tý Bootstrap výběr v_i. Pro *100(1-a)* %ní interval spolehlivosti pak platí

$$p_B - t_D * s_B \le ps \le p_B + t_D * s_B \quad (9)$$

kde pořádková statistika $t_D = t_{(int[\alpha*(M+1)/2])}$ a pořádková statistika $t_H = t_{(int[(1-\alpha/2)*(M+1)])}$

2.4 Vyhlazený odhad

Obecně lze na základě hodnot p_i sestavit odhad hustoty pravděpodobnosti jejich rozdělení *fe(p)* např. s využitím histogramu nebo jádrového odhadu. Při znalosti funkce *fe(p)* se snadno konstruuje interval spolehlivosti přímo z definice (2). Pro meze tohoto intervalu pak platí, že

$$\alpha/2 = \int_{-\infty}^{LC} fe(p)dp$$

a

$$\alpha/2 = \int_{UC}^{\infty} fe(p)dp$$

Podle typu odhadu *fe* může jít o úlohu numerické nebo analytické integrace.

2.5 Generace Bootstrap výběrů

Základním předpokladem úspěšnosti celého postupu je generace Bootstrap výběrů,. Pro tento účel je třeba buď znát nebo volit rozdělení *g(x)*. Standardní technika neparametrického Bootstrap vychází z neparametrického odhadu *g(x)* ve tvaru

$$g(x) = \frac{1}{N}\delta(x - x_i) \quad (10)$$

kde Diracova funkce $\delta(x - x_i) = 1$ pro $(x = x_i)$ a všude jinde je. $\delta(x - x_i) = 0$.

Toto rozdělení pokládá pravděpodobnost *1/N* v každém bodě. Simulované výběry se pak realizují jako náhodné výběry složené z prvků původního výběru ***x*** s vracením (tj. jeden prvek původního výběru se může v simulovaném výběru vyskytovat i opakovaně**).**

Další možností je konstruovat vhodný parametrický model *g(x)*, odhadnout jeho parametry a generovat simulované výběry standardními postupy. Tento přístup naráží na celou řadu problémů souvisejících s možnou nehomogenitou, vybočujícími body, heteroskedasticitou a autokorelací.

Bootstrap metody obecně poskytují informace jak o bodových odhadech, tak i intervalech spolehlivosti. Uvažujme standardní neparametrický Bootstrap (v_i jsou výběry s vracením) pro $ps = m$, tj. jde o střední hodnotu a její interval spolehlivosti střední hodnoty. Lze snadno určit, že v tomto případě je Bootstrap průměr totožný s aritmetickým průměrem původních dat a Bootstrap rozptyl je M-krát menší než rozptyl původních dat. Liší se však intervaly spolehlivosti zejména tam, kde se rozdělení dat výrazně odchyluje od normálního rozdělení.

Kromě standardního Bootstrap lze použít také dvojitý Bootstrap (Bootstrap aplikovaný na výběry v_i), blokový Bootstrap (realizace výběru s vracením na bloky homogenních dat a sestavení celkového Bootstrap výběru spojením výsledků). [9]

3. Realizace postupu Bootstrap

Z hlediska realizace metod Bootstrap na počítači je základem generace simulovaných výběrů. Velmi jednoduše se dá tato operace provést v jazyku MATLAB s využitím vektorového triku. Úsek programu má tvar

```
ar=load('conc.txt');[c s]=size(ar); b=800;
if c ==1
   ar=ar';c=s;
end
B=ar(ceil(c*rand(c,b)));
```

Předpokládá se , že data jsou v souboru *conc.txt* a *b* – tice Bootstrap výběrů je v poli *B*.
Pro výpočet odhadu p_i se používá standardních postupů. Výpočet intervalů spolehlivosti je pak závislý na volbě přístupu (viz. 2.1-2.4). Program BOOTM v jazyce MATLAB počítá interval spolehlivosti střední hodnoty z předpokladu normality (2.1), Studentizace (2.3) a percentilové metody (2.2).

4. Příklad . *Určení koncentrace ethyl parathionu v ovzduší*

V rámci monitorování toxických látek byl monitorován toxický ethyl parathionu v ovzduší u Herber Station v Californii (data byla publikována v [8]). Získané koncentrace v mg/m^3 jsou

0.0090 0.0090 0.0090 0.0090 0.0180 0.0320 0.0120 0.0150 0.0090 0.0780 0.0920 0.0230 0.0180 0.0100

Limita detekce přístroje je limd = 0.01 a hodnoty 0.09 jsou tedy pouze dosazeny Místo nich mohou být nuly, či jiná čísla od 0 do 0.01.Účelem je stanovit 95 procentní interval spolehlivosti střední hodnoty .

A. Bootstrap analýza pro původní dat

S využití programu BOOTM bylo určeno:
Průměr = 0.0245 a výběrový rozptyl = 0.000709

Klasická normalita

95 % ní interval spolehlivosti UC = 0.0384 LC = 0.0106

.Bootstrap normalita

95 % ní interval spolehlivosti UC = 0.0360 LC = 0.0069

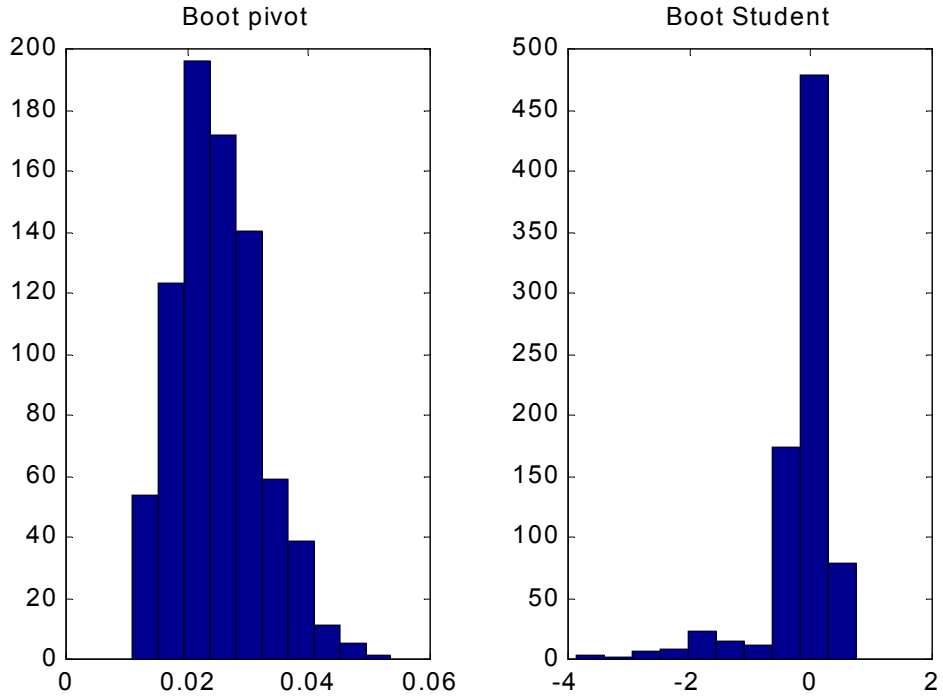
Bootstrap pivot

95 % ní interval spolehlivosti UC = 0.0396 LC = 0.0126

Bootstrap Student

95 % ní interval spolehlivosti UC = 0.0278 LC = 0.0072

Na obr. 1 je uvedeno rozdělení veličin p_i a t_i Jsou patrné odchylky od normálního rozdělení Je patrné, že Studentizovaný Bootsrap poskytuje výrazně nižší horní mez UC.



Obr, 1 Rozdělení veličin p_i a t_i (původní data)

B. Bootstrap analýza pří nahrazení hodnot po limitou detekce nulou

S využití programu BOOTM bylo určeno:

Průměr = 0.02213 a výběrový rozptyl = 0.000837

Klasická normalita

95 % ní interval spolehlivosti UC = 0.0364 LC = 0.00614

.Bootstrap normalita

95 % ní interval spolehlivosti UC = 0.0360 LC = 0.0069

Bootstrap pivot

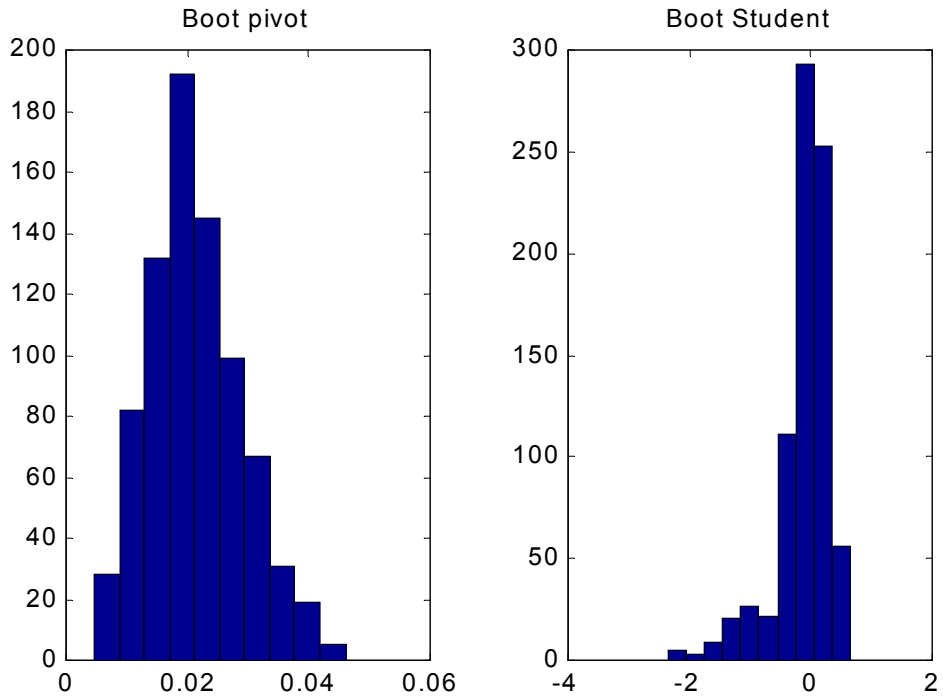
95 % ní interval spolehlivosti UC = 0.0369 LC = 0.0082

Bootstrap Student

95 % ní interval spolehlivosti UC = 0.0246 LC = 0.0103

Na obr. 2 je uvedeno rozdělení veličin p_i a t_i Jsou opět vidět odchylky od normálního rozdělení Je patrné, že Studentizovaný Bootsrap poskytuje výrazně nižší horní mez UC než ostatní metody a nahrazení podlimitních hodnot nulou má za důsledek snížení všech horních mezí. Právě Studentizovaný Bootstrap je často považován za výhodný a doporučován pro komplexnější rozdělení dat.[9].

Je pochopitelně výhodnější zpracovávat tato data modelem, který uvažuje limitu detekce a tento příklad pouze demonstruje rozdíly mezi jednotlivými možnostmi.



Obr, 2 Rozdělení veličin p_i a t_i (nahrazení podlimitních hodnot nulou)

5. Závěr

Je patrné, že pro statistické zpracování dat v analytické chemii a speciálně ve stopové analýze může být využito počítačově intenzivních metod bez větších problémů. Ve shodě s koncepcí „***satistical methods mining***“ je často nezbytné kombinovat různé přístupy.

Poděkování:

Tato práce vznikla s podporou grantu MŠMT č. VS 97084, grantu GAČR . 106/99/1184 a výzkumného záměru MŠMT č.J11/98:244101113

9. Literatura

[1] Meloun M., Militký J.*: Zpracování experimentálních dat*, East Publishing Praha 1998
[2] Shuway, R.M., Atazi, A.S., Johnson, P.: Technometrics 31, 347 (1989)
[3] Boos D.D. , Hughes-Oliver J. M.: Amer. Statist. **54**, 121 (2000)
[4] Hall, P.: J.R. Stat. Sor. 54, 221 (1992)
[5] Chen L. : Environmetrica **6,** 181 (1995)
[6] Chen L.: J. Appl. Statist. **25**, 739 (1998)
[7] Shumway R. H. a kol.: Technometrics, **31**, 347-356 (1989)
[8] Wekrens, R. a kol.: Chem.Int. Lab. Systems **54**, 35-52 (2000)
[9] Davidson, A., Hinkley, D.V.,: Bootstrap Methods and Their Applications, Cambridge Univ. Press, Cambridge, 1997