POROVNANIE VIACROZMERNÝCH KALIBRAČNÝCH TECHNÍK PRI STANOVENÍ RNA-NUKLEOTIDOV

MILOSLAV FOLTIN[a], JANA ŠUĽAKOVÁ[a], MARTIN FOLTIN[b] a IVAN SEKAJ[b]

[a]*Katedra analytickej chémie Prírodovedeckej falkulty Univerzity Komenského, Mlynská dolina, 842 15 Bratislava,* [b]*Katedra automatizovaných systémov riadenia, Fakulta elektrotechniky a informatiky, Slovenská technická univerzita, Ilkovičova 3, 812 19 Bratislava*
Miloslav.Foltin@fns.uniba.sk

Došlo 10.1.03, prijaté 18.2.04.

Kľúčové slová: Viacrozmerná kalibrácia, genetický algoritmus, nukleotidy

Úvod

Tvorba kalibračného modelu je kľúčovým krokom v procese analýzy komplexných vzoriek. Súčasná inštrumentálna technika (HPLC-DAD, GC-MS, GC-FTIR a iné) je schopná produkovať viacrozmerné analytické signály, ktoré obsahujú viac informácií, ale ich spracovanie si vyžaduje nové chemometrické prístupy. Pre tento účel boli vyvinuté viaceré viacrozmerné kalibračné metódy (MLR, PCR, PLS a iné), ktoré sú stále predmetom skúmania analytických chemikov.

Teoretická časť

Množinu viacrozmerných kalibračných dát môžeme zapísať vo forme matíc **C**, **R**, kde matica **C** ($n \times p$) obsahuje $p>1$ nezávislých premenných (koncentrácií jednotlivých zložiek) pre každú z n vzoriek. Maticu **R** ($n \times q$) tvorí n analytických signálov (spektier) nameraných v q polohách (vlnových dĺžkach). Kalibračný model v najvšeobecnejšej forme vyjadríme vzťahom:

$$\mathbf{R} = \mathrm{f}(\mathbf{C}; \mathbf{B}) + \mathbf{E_R} \qquad (1)$$

kde **B** ($p \times q$) je matica parametrov modelu a $\mathbf{E_R}$ ($n \times q$) je matica chýb, pričom platí, že funkcia f je lineárnou funkciou parametrov.

Kalibračný model sa vytvorí zmeraním analytického signálu kalibračných vzoriek, ktorých zloženie je známe, a ktoré viac alebo menej reprezentujú skutočné zloženie neznámych vzoriek.

Na predikciu koncentrácie v neznámych vzorkách je potrebné použiť model v inverznom tvare:

$$\hat{\mathbf{C}} = \mathrm{f}^{-1}(\mathbf{R_{new}}, \hat{\mathbf{B}}) \qquad (2)$$

Problémom pri použití kalibračného modelu na predikciu koncentrácie analytov v neznámych vzorkách je nájdenie inverznej funkcie f^{-1}, pretože funkcia f je stochastická a jej inverzia nie je triviálna. Preto niektorí autori doporučujú vytvárať kalibračný model priamo v inverznom tvare:

$$\mathbf{C} = \mathrm{F}(\mathbf{R}; \mathbf{B}) + \mathbf{E_C} \qquad (3)$$

Pri použití tohoto kalibračného modelu na predikciu koncentrácií nie je potrebná inverzia funkcie F:

$$\hat{\mathbf{C}} = \mathrm{F}(\mathbf{R_{new}}, \hat{\mathbf{B}}) \qquad (4)$$

Ak predpokladáme, že aj závislosť medzi **C** a **R** je lineárna, potom klasický resp. inverzný kalibračný model môžeme zapísať jednoduchými rovnicami:

$$\mathbf{R} = \mathbf{C}\,\mathbf{B} + \mathbf{E_R} \qquad (5)$$
$$\mathbf{C} = \mathbf{R}\,\mathbf{B} + \mathbf{E_C} \qquad (6)$$

Príkladom takéhoto lineárneho modelu je napríklad Lambertov-Beerov zákon.

Úlohou kalibrácie je potom určiť (odhadnúť) maticu **B** z kalibračných dát **C, R**. Na tento účel sa najčastejšie používa metóda označená ako MLR (multiple linear regression), ale v poslednom čase sa stále viac presadzujú metódy, ktoré sú založené na rozklade matíc zdrojových dát na maticu hlavných komponentov (matica skóre) a maticu záťaží (loadings). Tieto metódy sú v literatúre známe pod názvami regresia hlavných komponentov (PCR), resp. metóda parciálnych najmenších štvorcov (PLS). Princípy a použitie týchto metód sú opísané vo viacerých monografiách[1–3], preto ich tu neuvádzame.

Napriek tomu, že matematický princíp týchto metód je známy a algoritmy riešenia sú prepracované, mnohé štatistické vlastnosti nie sú dostatočne preskúmané[4].

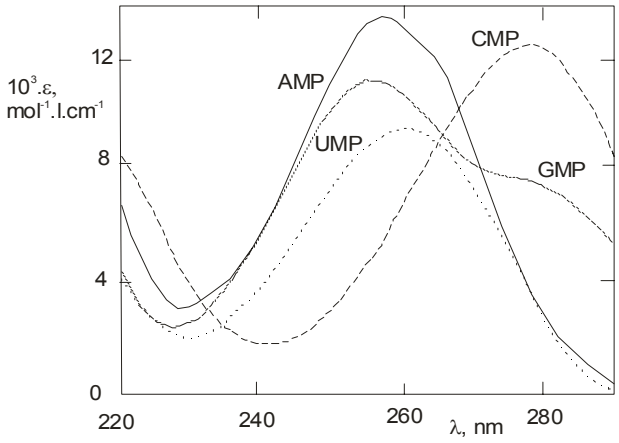
Cieľom práce bolo preveriť možnosti a štatistické vlastnosti viacrozmerných kalibračných postupov (MLR, PCR, PLS) pri kvantifikácii 4 nukleotidov adenínmonofosfátu, (AMP), guanidínmonofosfátu (GMP), uridinmonofosfátu (UMP) a cytidínmonofosfátu (CMP), ktorých prítomnosť a vzájomné pomery v analyzovaných vzorkách môže slúžiť na identifikáciu biologického zdroja alergénov[5,6] v prachových časticiach ovzdušia.

Experimentálna časť

Na testovanie vlastností kalibračných postupov boli použité súbory 15 modelových kalibračných vzoriek a 12 validačných vzoriek. Zloženie kalibračných vzoriek bolo určené metódou štatistického plánovania SIMPLEX tak, aby sa minimalizovala možnosť multikolinearity v kalibračných dátach a aby vzorky pokrývali rovnomerne celý kalibračný priestor. Zloženie kalibračných vzoriek (matica $\mathbf{C}_{15\times4}$) je uvedené v tabuľke I. Spektrá nukleotidov

Tabuľka I
Zloženie kalibračného súboru

$10^5\ c$ [mol.l^{-1}]			
CMP	AMP	GMP	UMP
0,15	0,00	0,00	0,00
0,00	0,15	0,00	0,00
0,00	0,00	0,15	0,00
0,00	0,00	0,00	0,15
7,50	7,50	0,00	0,00
7,50	0,00	7,50	0,00
0,00	7,50	7,50	0,00
7,50	0,00	0,00	7,50
0,00	7,50	0,00	7,50
0,00	7,50	7,50	7,50
4,95	4,95	4,95	0,00
4,95	0,00	4,95	4,95
4,95	4,95	0,00	4,95
0,00	4,95	4,95	4,95
3,75	3,75	3,75	3,75



Obr. 1. **Spektrá nukleotidov**

(matica $\mathbf{S}_{4\times36}$) boli prevzaté z literatúry[7] a sú prezentované na obr. 1.

Analytické signály kalibračných vzoriek boli vypočítané ako súčin $\mathbf{R}_{15\times36}= \mathbf{C}_{15\times4}*\mathbf{S}_{4\times36}$. Aby vypočítané signály simulovali reálne experimentálne dáta, boli zaťažené náhodným šumom na 4 úrovniach so štatistickými parametrami μ, σ : N1(0; 0,001), N2(0; 0,005), N3(0; 0,01), N4(0; 0,02). Matice šumu boli generované funkciou RANDN v programe Matlab.

Súbor neznámych vzoriek, označený ako validačný súbor, bol vytvorený analogickým postupom. Koncentračné zloženie vzoriek validačného súboru je v uvedené v tabuľke II.

Funkčnosť kalibračných modelov získaná metódami MLR, PCR a PLS bola testovaná vyhodnotením správnosti predikcie koncentrácií nukleotidov vo validačnom súbore. Na kvantifikáciu predikčnej chyby boli použité hodnoty

$$\mathrm{MSEP}=\sqrt{\frac{\sum_i (\hat{c}_i - c_i)^2}{n-p}} \qquad (7)$$

resp. jej relatívne hodnoty vztiahnuté na priemerné hodnoty koncentrácií daného nukleotidu vo validačných vzorkách.

Tabuľka II
Zloženie validačného súboru

$10^5\ c$ [mol.l^{-1}]			
CMP	AMP	GMP	UMP
0,150	2,070	2,670	2,540
2,280	2,600	0,700	1,240
2,310	1,890	0,919	2,520
2,480	2,210	1,050	0,808
0,376	2,180	1,540	1,250
0,050	3,000	1,770	1,610
0,375	0,375	0,375	0,375
3,000	0,750	0,750	0,075
0,075	1,800	1,280	3,000
1,050	3,000	1,950	0,075
2,400	0,300	0,135	0,300
0,900	0,450	0,600	0,750

Na spracovanie kalibračných dát metódami MLR, PCR a PLS bol použitý program Matlab ver. 5.6 (The MathWorks, Inc., Natick, Mass., USA) doplnený chemometrickým toolboxom. Na optimalizáciu počtu vlnových dĺžok pre kalibračný model spracovaný metódou MLR bola použitá technika genetických algoritmov[8,9]. Na jej spracovanie bol použitý vlastný toolbox napísaný v jazyku Matlab.

Výsledky a diskusia

Predikčná schopnosť kalibračných modelov, získaných rôznymi viacrozmernými metódami, je prezentovaná v tabuľke IV a V: Správnosť predikcie je vyjadrená hodnotou strednej kvadratickej chyby predikcie MSEP, resp. jej relatívnej hodnoty. Z tabuľky IV je zrejmé, že uspokojivú správnosť predikcie je možné dosiahnuť iba v prípade dát,

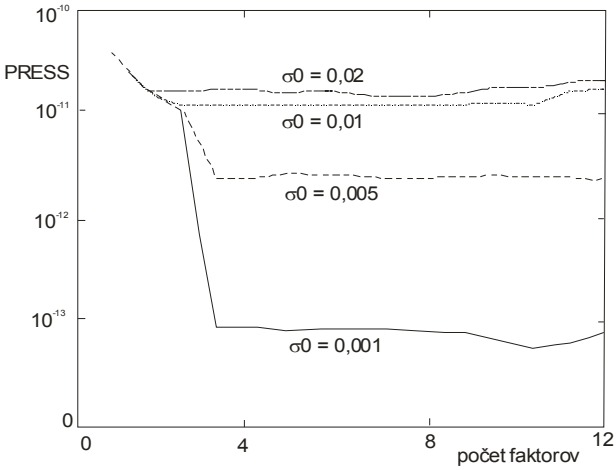
Tabuľka III
Korelačná matica spektier RNA nukleotidov

	CMP	AMP	GMP	UMP
CMP	1,00	– 0,24	0,12	–0,08
AMP	–0,24	1,00	0,79	0,98
GMP	0,12	0,79	1,00	0,82
UMP	0,08	0,98	0,82	1,00

Tabuľka IV
Relatívna stredná chyba predikcie koncentrácií nukleotidov vo validačných vzorkách

Nukleotidy	σ_0	$MSEP_{rel}$ [%]			
		CMP	AMP	GMP	UMP
MLR	0,001	0,80	2,86	1,94	5,73
	0,005	4,78	13,99	8,76	28,69
	0,01	11,76	42,14	22,79	75,45
	0,02	26,52	39,36	60,45	79,22
MLR-GA	0,001	0,78	1,63	2,14	2,42
	0,005	3,16	5,46	8,40	8,82
	0,01	7,79	17,85	16,58	36,44
	0,02	18,14	29,14	27,08	64,50
PCR	0,001	0,81	2,93	1,91	5,88
	0,005	5,11	15,74	8,92	32,61
	0,01	10,39	37,47	22,77	68,04
	0,02	28,16	38,00	54,41	70,77
PLS	0,001	0,81	2,91	1,91	5,86
	0,005	5,10	14,85	9,15	30,18
	0,01	11,84	40,16	24,18	71,13
	0,02	26,88	38,00	52,83	75,34

ktoré sú zaťažené šumom, ktorého smerodajná odchýlka je menšia ako $\sigma_0 < 0,01$. Pri zaťažení analytických signálov väčším šumom žiadny z testovaných kalibračných modelov nie je schopný predikovať koncentrácie validačných vzoriek s akceptovateľnou správnosťou.



Obr. 2. **Závislosť súčtu štvorcov predikovaných rezíduí (PRESS) od počtu faktorov v kalibračnom PCR modeli**

Najväčšie hodnoty $MSEP_{rel}$ boli získané pri predikcii nukleotidu UMP, a to vo všetkých kalibračných modeloch. Nesprávna predikcia tohoto nukleotidu je spôsobená tým, že jeho spektrum v rozsahu 220–290 nm vykazuje veľké hodnoty korelačných koeficientov so spektrami AMP aj GMP (viď. tabuľka III). Spektrá AMP a GMP sú vzájomne menej podobné ($\rho = 0,79$). Významný vplyv podobnosti spektier na správnosť predikcie koncentrácie potvrdzujú najmenšie hodnoty $MSEP_{rel}$ pre nukleotid CMP, ktorého spektrum sa výrazne odlišuje od zvyšných troch nukleotidov.

Korelačné vzťahy spektier analytov môžu teda významne ovplyvňovať správnosť predikcie koncentrácie neznámych vzoriek. Tento záver potvrdzujú aj výsledky získané kombináciou metódy MLR s optimalizáciou počtu vlnových dĺžok pomocou procedúry genetického algoritmu (GA). Významné zlepšenie predikcie (pokles hodnôt $MSEP_{rel}$ asi o 50%) môžeme vysvetliť tým, že vlnové dĺžky, pri ktorých je korelácia spektier najvyššia, sú z riešenia GA vynechané. Zníženie dimenzie matice signálov z pôvodných $\mathbf{R_{12\times36}}$ na $\mathbf{R_{12\times17}}$, resp $\mathbf{R_{12\times22}}$ konečný výsledok významne neovplyvní.

Použitie metód PCR a PLS sa všeobecne doporučuje na spracovanie viacrozmerných analytických signálov, v ktorých buď spektrá analytov, alebo ich elučné profily, sa silne prekrývajú[10], alebo sú silne zašumené[11]. Vhodne voleným počtom relevantných faktorov je možné do určitej miery eliminovať vplyv korelácie v dátach (prekrytie spektier) alebo šumu. Za ideálnych podmienok je počet relevantných faktorov rovný počtu analytov v kalibračnom modeli. Vlastné čísla kalibračnej matice s poradovým číslom väčším ako je počet analytov už reprezentujú iba neužitočnú informáciu t.j. šum.

Pri aplikácii týchto metód na naše modelové dáta sa ukázalo, že pri zaťažení šumom, ktorého $\sigma_0 > 0,02$, kalibračný model modeluje prevážne šumovú informáciu. Počet požadovaných relevantných faktorov nebolo možné jednoznačne určiť. Minimum optimalizačných funkcií

Tabuľka V
Stredná chyba predikcie koncentrácií nukleotidov vo validačných vzorkách

Nukleotidy	σ_0	10^5.MSEP				
		CMP	AMP	GMP	UMP	celková
MLR	0,001	0,0103	0,0491	0,0222	0,0695	0,0377
	0,005	0,0615	0,2404	0,1003	0,3477	0,1871
	0,01	0,1514	0,7243	0,2609	0,9144	0,5137
	0,02	0,3413	0,6766	0,6921	0,9600	0,5992
MLR-GA	0,001	0,0100	0,0280	0,0245	0,0293	0,0206
	0,005	0,0406	0,0938	0,0961	0,1069	0,0752
	0,01	0,1002	0,3067	0,1899	0,4416	0,2469
	0,02	0,2335	0,5008	0,3100	0,7817	0,4291
PCR	0,001	0,0104	0,0504	0,0218	0,0712	0,0386
	0,005	0,0658	0,2705	0,1021	0,3952	0,2107
	0,01	0,1337	0,6440	0,2607	0,8245	0,4633
	0,02	0,3625	0,6532	0,6230	0,8577	0,5530
PLS	0,001	0,0104	0,0501	0,0218	0,0711	0,0385
	0,005	0,0657	0,2553	0,1048	0,3657	0,1973
	0,01	0,1524	0,6902	0,2768	0,8621	0,4898
	0,02	0,3459	0,6530	0,6049	0,9131	0,5634

používaných na výber počtu faktorov v programe Matlab (CROSS, PRESS)[11,13] bolo veľmi ploché v rozsahu hodnôt 3–15. Hodnoty oboch funkcií sa výrazne neznížili ani pri použití všetkých 15 faktorov, čo je celkový počet vlastných čísel kalibračnej matice. Je teda zrejmé, že hodnota šumu $\sigma_0 \approx 0{,}02$ v spektrálnych dátach nukleotidov predstavuje hraničnú hodnotu pre kvantifikáciu RNA-nukleotidov metódami PCR a PLS v intervale vlnových dĺžok 220–290 nm.

Závislosť funkcie PRESS od počtu faktorov, pre dáta zaťažené rôznym šumom, je na obr. 2. Z obrázku je vidieť, že pre hodnoty šumu $\sigma_0 < 0{,}01$ funkcia PRESS vykazuje minimum pri hodnote 4, čo zodpovedá počtu analytov v modeli. Pri vyšších hodnotách šumu minimum je neurčité a s rastúcim počtom použitých faktorov sa hodnota PRESS neznižuje. Značí to, že užitočná informácia je silne prekrytá neužitočnou informáciou t.j. šumom.

Záverom môžeme konštatovať, že viacrozmerné kalibračné techniky MLR, PCR a PLS sú pre spracovanie spektrálnych dát nukleotidov CMP, AMP, GMP a UMP približne rovnocenné ak šum spektrálnych dát neprekročí hodnotu $\sigma_0 \approx 0{,}02$. Pri optimalizácii počtu vlnových dĺžok, napríklad s použitím genetických algoritmov, môžeme chybu predikcie metódy MLR znížiť asi o 50 %.

Ďalšie zníženie predikčnej chyby kalibračných modelov je možné dosiahnuť s použitím vhodných procedúr na filtráciu šumu ešte pred tvorbou kalibračného modelu. Tento prístup bude predmetom ďalšej práce.

Táto práca vznikla za finančnej podpory VEGA MŠ a SAV v rámci projektu č.1/6222/99.

LITERATÚRA

1. Martens H., Naes T.: *Multivariate calibration*. Wiley, Chichester 1989.
2. Vandenginste B. M. G., Massart L. M., Buydens L. M. C., De Jong S., Lewi P. J., Smeyers-Verbeke J.: *Handbook of Chemometrics and Qualimetrics*, díl B. Elsevier, Amsterdam 1998.
3. Meloun M., Militký J.: *Kompendium statistického zpracování dat.* Academia, Praha 2002.
4. Booksh S. K., Kowalski R. B.: Anal. Chem. *66*, 782A (1994).
5. Reijenga J. C., Verheggen Th. P. E. M, Everaerts F. M.: J. Chromatogr. *267*, 75 (1983).
6. De Bruijn C. H. M. M, Reijenga J. C., Aben G. V. A., Verheggen Th. P. E. M., Everaerts F. M.: J. Chromatogr. *320*, 205 (1985).
7. Lorber A.: Anal. Chem. *58*, 1171 (1986).
8. Goldberg D. E.: *Genetic Algorithms in Search, Optimisation, and Machine Learning.* Addison-Wesley

Publishing Company, New York 1989.
9. Michalewicz Z.: *Genetic Algorithms + Data Structures = Evolution Programs*. Springer-Verlag, Berlin 1996.
10. Sentellas S., Saurina J., Hernández-Cassou S., Galceran T. M., Pignou L.:J. Chromatogr. *909*, 259 (2001).
11. Brereton R. G.: Analyst *125*, 2125 (2000).
12. Faber N. M., Bro R.: Chemom. Intell. Lab. Syst. *61*, 133 (2002).
13. Kramer R.: *Chemometrics Toolbox for Use with Matlab*. The MathWorks, Natick 1994.

M. Foltin[a], J. Šuľaková[a], M. Foltin[b], and I. Sekaj[b]
([a]Department of Analytical Chemistry, Faculty of Science, Comenius University, Bratislava, and [b]Department of Automatic Control Systems, Faculty of Electrical Engineering and Information Technology, Slovak University of Technology): **Comparison of Multiway Calibration Techniques in Determination of RNA Nucleotides**

The quality of several multiway calibration models of spectral data of four RNA nucleotides with different noise is compared. An own genetic algorithm for optimization of the number of wavelengths in spectral matrix for multiple linear regression methods (MLR) was used. Useful analytical information was obtained only for data with noise $\sigma_0 < 0.02$. The prediction error of analyte concentration using a combination of MLR and genetic algorithm was decreased by ca. 50 %.