7. LA INSTRUMENTACIÓ MODERNA I ELS MÈTODES DE CALIBRATGE MULTIVARIANT

Romà Tauler i Anna Izquierdo-Ridorsa*

7.1. INTRODUCCIÓ

La utilització dels instruments analítics per a extraure informació quantitativa s'assembla en certs aspectes a la dels instruments musicals. En ambdós casos, per exemple, els instruments s'han d'afinar de manera adequada, i també en ambdós casos s'ha d'aprendre a interpretar allò que els instruments ens proporcionen; en un cas, s'ha d'aprendre a convertir les dades experimentals en informació quantitativa i en l'altre s'ha d'aprendre a apreciar la música. Els instruments més senzills de mesura són els que només tenen un canal de mesura (per exemple, un mesurador de pH), comparable a l'instrument musical d'una sola corda. Però també tenim instruments de cordes múltiples (per exemple, els espectrofotòmetres) i àdhuc disposem d'una orquestra completa d'instruments de mesura per a resoldre problemes analítics més difícils. La seqüència de mostres a estudiar pot comparar-se amb la melodia —un seguit de sons en un camp determinat i amb un tema determinat. La velocitat i la seqüència amb les quals les mostres s'analitzen constitueixen el seu ritme. Així, els resultats obtinguts en la mesura de cada mostra són vistos en el context dels resultats anteriors. A partir de canvis bruscs en la melodia o en el ritme, poden detectar-se les anormalitats en les mostres.

Considerem ara com funcionen els nostres ulls, i comparem-los amb un instrument de mesura de la radiació electromagnètica,

* Departament de Química Analítica, Universitat de Barcelona, Diagonal, 647, 08028 Barcelona.

com, per exemple, un espectrofotòmetre. A la figura 1 es pot veure el model de funcionament acceptat actualment [1, 2]. Quan mirem un vestit de color vermell fem un procés molt similar al que fa un instrument modern multicanal (un espectrofotòmetre). Quan la llum del sol incideix damunt de la roba, penetra una mica dins del material abans d'ésser dispersada enrere de la superfície, a causa de la manca d'homogeneïtat en les propietats refractives del material. Però mentre es troba a l'interior del material, una part de la llum solar és absorbida pels pigments del material, fonamentalment pels pigments no vermells, ja que aquest és de color vermell. La llum dispersada serà radiada de manera difusa des de la superfície del teixit cap a fora; una part d'ella cap als nostres ulls i sobre la retina on es produeix la fotorecepció. La gent que té una visió normal dels colors té a la retina cons de tres tipus diferents, pigmentats i sensibles a la llum. Aquests tres detectors no lineals tenen sensibilitats diferents en el camp del visible de 380-760 nm. Així, a l'ull, l'espectre d'intensitats de llum que prové del tros de roba es tradueix en intensitats de senyal sobre tres components nerviosos. Al cervell, l'espectre d'intensitats de llum «comprimit» en aquestes tres variables latents es combinen de nou i ens donen la nostra percepció del color. A partir d'aquesta percepció podem mesurar les concentracions, per exemple, de colorant vermell sobre el teixit, així com altres característiques associades al color. Com que la distribució espectral de la font de llum afecta de manera molt important l'espectre de la mostra, el cervell també tindrà en compte el color de la llum de la font emissora (el sol), de manera similar al que fa l'espectrofotòmetre, on la intensitat de la font es mesura simultàniament, I_0, i es fa la transformació a transmisió $T = I_t/I_0$ o reflectància $R = I_r/I_0$. Aquestes tres dimensions associades a la percepció del color poden classificar-se en: eix clar/fosc, eix vermell/verd, eix groc/blau. De fet, aquesta descomposició i recombinació que fa la nostra percepció del color és molt similar a la dels procediments de calibratge multivariant emprats en els instruments analítics moderns. La diferència principal està en com s'estableix el procediment de calibratge. Mentre que la visió del color està determinada en gran part per la genètica, i en una altra petita part per l'aprenentatge durant la infància, el calibratge dels instruments analítics necessita un model matemàtic basat en dades empíriques i assumpcions teòriques. L'ull pot percebre més coses que els colors; per exemple, pot percebre les formes i els moviments.

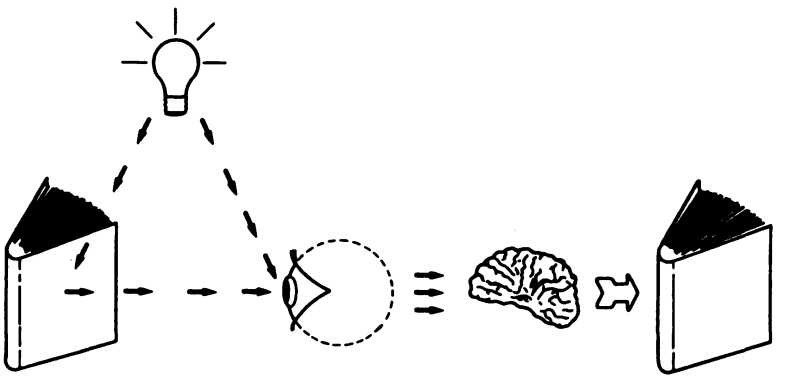

FIGURA 1. Model de funcionament de la visió.

De la mateixa manera, si se substitueix un espectrofotòmetre per una càmera de VDU, el calibratge dels instruments ens pot permetre determinar estructures en l'espai i canvis dinàmics amb el temps. Així mateix, el nas i la llengua poden comparar-se amb un cromatògraf de gasos i un elèctrode selectiu d'ions. Igualment com utilitzem simultàniament les dues orelles, els dos ulls, el nas i la llengua, es poden combinar matemàticament dos o més instruments multicanal en un sistema analític molt poderós que permetrà l'estudi de les mostres reals del nostre món complex.

7.1.1. *Dades multivariants*

El tipus de transformació o procés que s'ha de realitzar amb les mesures experimentals per a extraure'n informació analítica depèn de la manera com s'han adquirit aquestes dades, de la informació analítica que se'n vol extraure i del coneixement previ que es té del material que s'està analitzant. Un senyal, mesurat com a funció d'una sola variable sota control, per exemple, la longitud d'ona o el temps, s'anomena *senyal univariant* (un espectre o un cromatograma). Òbviament, la quantitat d'informació que pot extraure's a partir d'aquest senyal univariant és limitada. Per exemple, en cromatografia HPLC, quan la detecció es fa a una sola longitud d'ona, no es pot deduir el nombre de components coeluïts a partir del perfil d'un

pic cromatogràfic complex. Aquestes dificultats han estat superades amb la introducció dels mètodes de detecció multivariant, tals com la mateixa cromatografia HPLC, però ara amb detecció en diverses longituds d'ona (detector de díode *array*, DAD), o amb l'acoblament de tècniques com la cromatografia de gasos (GC) amb l'espectrometria de masses (MS). Aquests sistemes permeten mesurar les dades en funció de dues o més variables de control, per exemple, temps i longitud d'ona. En altres paraules, la taula o matriu de dades obtingudes d'aquesta manera s'anomena *senyal* o *resposta multivariant.*

A

$$\begin{array}{c} \\ \lambda_1 \\ \lambda_1 \\ \cdot \\ \cdot \\ \lambda_1 \end{array} \begin{array}{c} \begin{array}{cccccc} t_1 & t_2 & . & . & . & t_K \end{array} \\ \begin{pmatrix} a_{11} & a_{12} & . & . & . & a_{1K} \\ a_{21} & a_{22} & . & . & . & a_{2K} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ a_{I1} & a_{I2} & . & . & . & a_{IK} \end{pmatrix} \end{array}$$

B

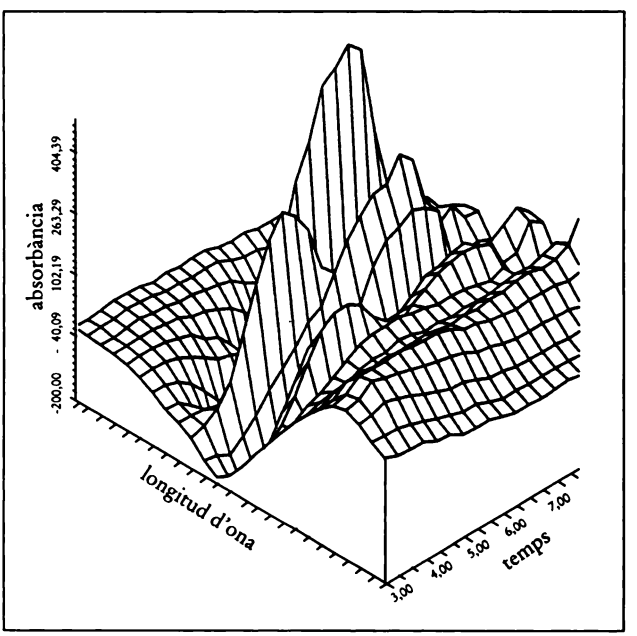

FIGURA 2. A) Exemple d'una matriu de dades obtingudes per aplicació de la tècnica HPLC-DAD. B) Representació gràfica tridimensional de la matriu de dades.

A la figura 2 es mostra un exemple de representació gràfica tridimensional de la matriu de dades obtingudes per aplicació de la tècnica HPLC-DAD; les columnes de la matriu seran els espectres mesurats a intervals de temps espaiats regularment, i les files de la matriu representaran els cromatogrames mesurats en cada longitud d'ona. Tant les files com les columnes de la taula estan relacionades entre si; és a dir, durant l'elució, per exemple, de dos components parcialment separats, la seva concentracio variarà d'acord amb el progrés de l'elució. Si considerem totes les columnes simultàniament, obtindrem més informació que no pas si considerem cada espectre separadament, independentment dels altres. Aquesta és la tasca de les tècniques multivariants: la manipulació simultània de tota la taula de dades. Com que, des d'un punt de vista matemàtic, la taula de dades no és més que una matriu de nombres, les tècniques estadístiques de procés de dades multivariants requeriran la utilització de l'àlgebra de matrius o àlgebra lineal.

Exemples de dades analítiques en dos, tres i quatre dimensions

pacients	tests clínics		
vins	orígens		
mostres	espectres		
long. d'ona d'excitació	long. d'ona d'emissió		
mostres	temps	long. d'ona (HPLC-DAD)	
mostres	temps	m/e (GC-MS)	
long. d'ona d'excitació	long. d'ona d'emissió	temps (HPLC-EEMS)	
mostres	temps	long. d'ona d'emissió	long. d'ona d'excitació

7.1.2. *Problemes derivats de la selectivitat de les dades analítiques*

En l'anàlisi química moltes vegades és difícil obtenir unes mesures que siguin només selectives per a aquells components que es volen determinar. A més del soroll de fons usual, més gran o més

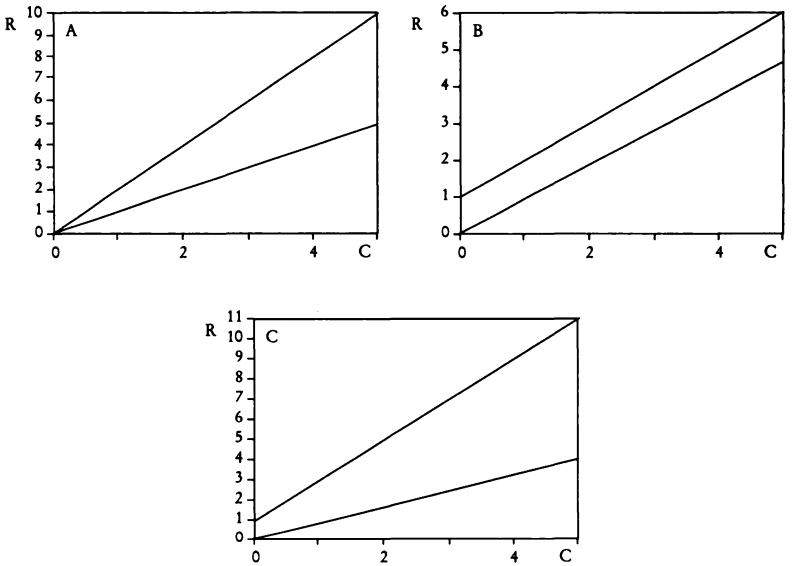

FIGURA 3. A) Efecte de matriu. B) Presència d'interferències. C) Efecte de matriu i presència d'interferències.

petit, les dades es troben afectades per interferències d'origen químic o físic causades per les mostres mateixes, o bé per interferències experimentals causades pel mateix procés de mesura. A més, moltes vegades en el procés de mesura sorgeixen fenòmens no lineals que creen problemes addicionals. Tradicionalment, aquestes interferències s'han d'excloure físicament per tal d'assegurar la selectivitat del procés, mentre a fi d'aconseguir mantenir la linealitat s'utilitza només un tros restringit de les escales de l'instrument. En alguns casos, fer tal cosa resulta prohibitiu des d'un punt de vista econòmic o, simplement, resulta impossible. Altres vegades, l'interès es troba en la quantificació dels constituents interferents, i, per tant, no interessa eliminar-los prèviament.

Amb el calibratge multivariant, els problemes que ocasionen les interferències i la manca de linealitat poden resoldre's millor que amb el calibratge univariant. A més, amb el calibratge multivariant l'avaluació explícita dels fenòmens no lineals no és necessària. Algunes vegades es distingueix entre interferències i efectes de matriu (figura 3). Els efectes de matriu es refereixen als canvis de sensibili-

tat dels sensors, mentre que les interferències afecten la resposta analítica però no la sensibilitat. De totes maneres, en la discussió que segueix ens referirem a les interferències com a qualsevol efecte sistemàtic sobre la resposta analítica causat per qualsevol fenomen físic o químic, que no sigui l'analit.

7.1.3. *El calibratge multivariant*

El calibratge multivariant és una eina general de millora de la selectivitat i la fiabilitat de les mesures analítiques. S'aplica tant per a la determinació dels constituents majoritaris com per a la determinació dels constituents minoritaris; es pot utilitzar en l'anàlisi de dades obtingudes amb un gran nombre d'instruments analítics. Amb el calibratge multivariant, la necessitat de la preparació i del pretractament de la mostra es redueix substancialment. La raó d'això rau en què no es necessiten mesures selectives per a obtenir resultats que siguin selectius. El calibratge multivariant estimula el desenvolupament d'instrumentació nova i n'augmenta les capacitats i el grau de fiabilitat; estén les possibilitats de les anàlisis químiques als processos de control industrial *on-line* (*process analytical chemistry*), a les anàlisis de mostres mèdiques o biològiques sense destruir-les i al seguiment de la pol·lució amb mètodes de cost més baix.

7.2. MODEL LINEAL (ANÀLISI MULTICOMPONENT)

El model de mescla additiva (llei de Beer) pot generalitzar-se així:

$$x_{ik} = \Sigma\,(y_{ij}k_{kj}) + e_{ik},$$

o en forma matricial:

$$X = Y\,K' + E$$

$$\begin{bmatrix} x_{11}\,x_{12} & \cdots & x_{1K} \\ x_{21}\,x_{22} & \cdots & x_{2K} \\ \vdots & \cdots & \vdots \\ x_{I1}\,x_{I2} & \cdots & x_{IK} \end{bmatrix} = \begin{bmatrix} y_{11}xy_{12} & \cdots & y_{1K} \\ y_{21}\,y_{22} & \cdots & y_{2K} \\ \vdots & \cdots & \vdots \\ y_{I1}\,y_{I2} & \cdots & y_{IK} \end{bmatrix} \begin{bmatrix} k_{11}\,k_{12} & \cdots & k_{1K} \\ k_{21}\,k_{22} & \cdots & k_{2K} \\ \vdots & \cdots & \vdots \\ k_{I1}\,k_{I2} & \cdots & k_{IK} \end{bmatrix} + \begin{bmatrix} e_{11}\,e_{12} & \cdots & e_{1K} \\ e_{21}\,e_{22} & \cdots & e_{2K} \\ \vdots & \cdots & \vdots \\ e_{I1}\,e_{I2} & \cdots & e_{IK} \end{bmatrix},$$

on X representa la matriu de dades espectrals per a les mescles $i = 1, 2, ..., I$ en els canals (longituds d'ona) k = 1, 2, ..., K:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdot & \cdot & \cdot & x_{1K} \\ x_{21} & x_{22} & \cdot & \cdot & \cdot & x_{2K} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ x_{I1} & x_{I2} & \cdot & \cdot & \cdot & x_{IK} \end{pmatrix}.$$

K representa els espectres unitaris dels constituents $j = 1, 2, ..., J$:

$$K = \begin{pmatrix} k_{11} & k_{12} & \cdot & \cdot & \cdot & k_{1J} \\ k_{21} & k_{22} & \cdot & \cdot & \cdot & k_{2J} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ k_{K1} & k_{K2} & \cdot & \cdot & \cdot & k_{KJ} \end{pmatrix}.$$

Y representa les concentracions d'aquests J constituents:

$$Y = \begin{pmatrix} y_{11} & y_{12} & \cdot & \cdot & \cdot & y_{1J} \\ y_{21} & y_{22} & \cdot & \cdot & \cdot & y_{2J} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ y_{I1} & y_{I2} & \cdot & \cdot & \cdot & y_{IJ} \end{pmatrix},$$

i la matriu E representa els residuals que apareixen entre els valors calculats amb el model additiu i les dades experimentals, i que seran deguts als errors experimentals o a la manca d'adequació del model (interferències no modelades, manca de linealitat en la resposta...):

$$E = \begin{pmatrix} e_{11} & e_{12} & \cdot & \cdot & \cdot & e_{1K} \\ e_{21} & e_{22} & \cdot & \cdot & \cdot & e_{2K} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ e_{I1} & e_{I2} & \cdot & \cdot & \cdot & e_{IK} \end{pmatrix}.$$

És més realista escriure:

$$X = 1k_0 + Y K' + E,$$

on k_0 representa l'efecte de fons o de línia de base. Per simplicitat, s'utilitza la primera equació, on $Y = (1, Y)$ i $K = (k_0, K)$. Quan els valors espectrals i les concentracions s'expressen com a desviacions respecte al seu valor central (la seva mitjana), l'efecte comú corresponent a la línia base es fa negligible.

7.2.1. *Millora de la selectivitat amb calibratge directe*

Quan es coneixen els espectres unitaris k de tots els constituents de la mostra a partir d'experiències prèvies, i no hi ha cap altra interferència en la matriu de dades X, la solució directa mínimoquadràtica és:

$$Y = X K (K'K)^{-1},$$

que implícitament assumeix la independència lineal de totes les columnes de K o, dit d'altra manera, que la matriu K és de rang complet. Quan la matriu $(K'K)$ no es pot invertir de manera correcta (matriu mal condicionada, amb les columnes de K colineals entre si), significa que un o més dels espectres unitaris dels constituents són similars a alguna de les combinacions lineals possibles dels espectres unitaris dels altres constituents a K. Aquest és un problema de manca de selectivitat analítica.

Tots els problemes de calibratge lineal es poden descriure amb l'equació:

$$Y = X B,$$

on B és la matriu de coeficients estimats al calibratge. Quan es considera la línia de base

$$Y = (X - 1k_0') B$$

$$Y = (X - 1k_0') K (K'K)^{-1},$$

la matriu de coeficients B és, doncs:

$$B = K(K'K)^{-1}.$$

Quan es disposa d'una estimació prèvia de quin és el soroll associat a les mesures o els errors (aleatoris), $V = diag(\sigma_k^2, k = 1, 2, ..., K)$, la solució minimoquadràtica ponderada és:

$$B = V^{-1} K(K'V^{-1}K)^{-1}.$$

La matriu de covariàncies de les concentracions estimades y_i a la mostra i per als $j = 1, 2, ..., J$ analits serà

$$Cov(y_i) = (K'K)^{-1} \sigma_i^2$$

o, en el cas de la solució ponderada (cas de variàncies diferents),

$$Cov(y_i) = (K'V^{-1}K)^{-1}.$$

7.2.2. *Calibratge indirecte: estimació i ús dels espectres obtinguts* in situ

En cas de no conèixer prèviament els espectres dels constituents, també és possible determinar-los en el mateix procés de calibratge. És a dir, s'ha de determinar la matriu K en l'etapa del calibratge. A partir de les concentracions dels analits a les solucions patró, Y, i dels espectres experimentals X, es pot estimar K per regressió de cada columna de X sobre Y en el conjunt de les solucions patró $i = 1, 2, ..., I$. En el cas general en què es vulgui estabilitzar l'estimació en contra del soroll o de l'error associat a les dades de calibratge, es faran més mesures que analits a determinar, i el valor de K serà donat per l'equació:

$$K' = (Y'Y)^{-1} Y'X.$$

Quan es considera l'efecte de la línia base en el model, tenim:

$$(k_0, K)' = ((1, Y)'(1, Y))^{-1}(1, Y)'X.$$

La inversió de la matriu de concentracions Y requereix que les concentracions dels constituents en el conjunt de calibratge variïn

de manera independent, ja que, si no, és impossible de determinar quines variacions de X poden relacionar-se amb els constituents individuals (problema de colinealitat). El model ha d'especificar-se de manera completa; si la matriu X es veu afectada per altres constituents químics o per efectes físics que han estat ignorats, la resolució de la mescla serà incorrecta.

7.2.3. *Calibratge invers*

Tal com ja s'ha explicat per al calibratge univariant en el capítol precedent, en l'anàlisi multicomponent també és possible l'aplicació generalitzada dels mètodes de calibratge invers, i, de fet, són aquests els que s'utilitzen en general actualment. És per aquesta raó que les concentracions s'expressen amb la lletra y i les respostes analítiques amb la lletra x; s'admet, en efecte, que la concentració és funció de la resposta instrumental. Aquest model invers de la llei de Beer es pot escriure:

$$Y = XP + E_Y,$$

on Y representa la matriu de les concentracions dels diversos analits a les solucions problema, X la matriu de respostes, P la matriu dels coeficients de calibratge que relacionen les concentracions amb les respostes instrumentals, i E_Y la matriu dels residuals no explicats pel model proposat. En aquest procediment, hom suposa que els errors en les concentracions dels patrons emprats són més grans que els associats a la mesura instrumental. El procediment de calibratge invers té l'avantatge fonamental que el tractament és invariant respecte al nombre de components que s'inclouen a l'anàlisi. Quan els errors en E_Y són independents, es pot realitzar l'anàlisi per a cada component separadament a partir de l'equació:

$$y = Xp + e_y,$$

on y és el vector de concentracions d'un analit que ens interessa determinar en les solucions patró de calibratge que poden contenir altres constituents i e_y és el vector d'errors residuals en les concentracions no ajustades pel model. En el calibratge, la solució mínimoquadràtica per a p és:

$$p = (X'X)^{-1}X'y$$

i durant la predicció, la concentració d'analit en una solució desconeguda és:

$$y = x'p.$$

Això vol dir que es pot fer l'anàlisi quantitativa d'un dels components independentment de la presència dels altres, àdhuc en el cas que només la concentració d'aquest component sigui coneguda en les solucions dels patrons. Els altres components presents a les solucions problema hauran d'estar també presents en les solucions de calibratge, i es modelaran només de manera implícita (a diferència dels mètodes de calibratge directe i indirecte abans esmentats, on el modelat de l'efecte matriu i de les interferències ha de fer-se de manera explícita). Aquest procediment, anomenat de vegades ILS (*inverse least-squares*) té el desavantatge que l'anàlisi s'haurà de realitzar per a un nombre de variables o canals de mesura (longituds d'ona, freqüències, etc.) igual o inferior al nombre de patrons emprats en l'etapa del calibratge. Això és així perquè el rang de la matriu que s'ha d'invertir és igual al nombre de canals, i aquest nombre no pot ésser superior al nombre de solucions patró analitzades. Treballant amb aquest procediment és fàcil que sorgeixin problemes de colinealitat (relacions quasi lineals entre les respostes analítiques en cada canal) quan el nombre de canals és gran, i de fet es perd precisió en aquests casos en relació amb els mètodes abans esmentats. Generalment, amb aquest mètode s'obtindran resultats menys precisos que amb el mètode de calibratge tradicional (vegeu els apartats anteriors) i els mètodes de compressió de dades que es descriuen a continuació.

7.3. MÈTODES DE COMPRESSIÓ DE DADES

7.3.1. *Compressió de les dades: a partir de moltes variables* X, *es dedueixen uns quants factors* T

Els següents problemes són força comuns quan s'intenta predir Y (matriu de concentracions) a partir de X (matriu de respostes

multivariants o espectres) a partir dels procediments anteriors basats en el model additiu (per exemple, la llei de Beer):

1) *Manca de selectivitat*: no hi ha cap variable a x (longitud d'ona a l'espectre experimental) que, per si sola, ens permeti trobar la concentració d'un component y.

2) *Colinealitat*: redundància i, per tant, intercorrelació entre les variables de la matriu de respostes instrumentals X (espectres).

3) *Manca de coneixement*: el nostre coneixement previ del mecanisme que regula les dades pot ésser incomplert o erroni.

Hi ha, per tant, la necessitat de desenvolupar mètodes de calibratge que estiguin absents dels problemes anteriors, que permetin fer bones prediccions de les concentracions dels analits en mostres reals de manera senzilla i que ajudin a augmentar la comprensió del mecanisme químic que regula les dades experimentals.

Els mètodes més importants proposats fins ara es basen en el que s'anomena *reducció del rang* o *compressió de dades*. L'estructura bàsica d'aquests mètodes es basa en el fet que la informació present en les nombroses variables $x_1, x_2, ..., x_K$, observades es concentra en unes quantes variables latents subjacents, anomenades *components principals*, *factors de regressió* o simplement *factors* $t_1, t_2, ..., t_A$, amb $A << K$, que es troben a partir d'una combinació lineal de les variables originals:

$$(t_1, t_2, ..., t_A)' = h_1[(x_1, x_2, ..., x_K)'].$$

Aquests factors s'utilitzen com a regressors en l'equació per a trobar $y_1, y_2, ..., y_J$, és a dir, per a fer la predicció de les concentracions dels analits a les mostres problema:

$$(y_1, y_2, ..., y_J)' = h_2[(t_1, ..., t_A)'] + f',$$

on f' representa aquelles contribucions a y que no poden explicar-se mitjançant els factors t. Els A factors t representaran la variació sistemàtica observada en els espectres X i seran importants per a fer la predicció de les concentracions y. Les dues funcions h_1 i h_2 formen les equacions de predicció adequades: $y = f(X) = h_2(h_1(x))$.

Generalment s'utilitzen models lineals per a aproximar les relacions existents entre les dades. Aquestes aproximacions lineals es

fan a partir de les dades centrades, ja que així s'evita el problema associat a l'existència de contribucions constants, com poden ésser el soroll de fons o la línia base (*background*):

$$X = X(input) - 1x_m',$$

$$Y = Y(input) - 1y_m',$$

on x_m i y_m són respectivament l'espectre o la resposta multicanal mitjana i la concentració mitjana dels patrons. El models de compressió de les dades es poden escriure segons les dues etapes següents:

a) *Etapa de calibratge*

El model de calibratge emprat es descriu amb els dos conjunts d'equacions lineals següents (model bilineal):

$$X = TP' + E,$$

$$Y = TQ' + F.$$

on:

$$T = XV.$$

La matriu de *loadings P* representa els coeficients de regressió de X a T, i la de *loadings Q*, la dels coeficients de regressió de Y a T. E i F representen la variació no explicada per l'estructura bilineal amb A factors. La matriu V es calcula per a optimitzar un criteri determinat que caracteritza el mètode. Després, o simultàniament, es troben els *scores T*, els *loadings P* i, finalment, els *loadings Q* estimats per regressió multilineal de cada variable x o de cada concentració y sobre els factors T (t_a, $a = 1, ..., A$). En notació matricial:

$$P' = (T'T)^{-1} T'X,$$

$$Q' = (T'T)^{-1} T'Y.$$

i els residuals o variació no modelada:

$$E = X - TP',$$

$$F = Y - TQ'.$$

La determinació del nombre mínim de factors A que descriuen les dades correctament és una etapa de gran importància. El model només ha d'incloure aquells factors que millorin la predicció de Y en mostres d'assaig independents. No ha d'incloure aquells factors que només són conseqüència del soroll de fons en les dades.

Un cop determinades les matrius V i Q per un procediment específic de cada variant del mètode en particular (vegeu més avall), l'estimació del valor de les concentracions desconegudes Y es trobarà a partir de l'equació:

$$Y = XVQ'.$$

b) Etapa de predicció en els mètodes bilineals

En l'etapa de predicció, les concentracions desconegudes de cada analit en cada solució, y_i, es trobaran després de determinar el factor t_i corresponent:

$$t_i' = x_i'V,$$

$$y_i' = t_i'Q',$$

i els residuals:

$$e_i' = x_i' - t_i'P'.$$

La predicció de la concentració desconeguda y_i a partir de les mesures x_i es pot fer de dues maneres diferents.

7.3.2. *Mètodes de calibratge bilineals més corrents*

Les dues aproximacions més importants al calibratge multivariant mitjançant mètodes de compressió de dades són l'anàlisi de

regressió per components principals (*principal component regression*), PCR, i l'anàlisi de regressió per mínims quadrats parcials (*partial least squares regression*), PLSR. El criteri per a trobar la matriu V es basa en les dues idees següents:

PCR: COMPRIMIR LA MATRIU X EN ELS SEUS FACTORS DOMINATS.

PLSR: COMPRIMIR LA MATRIU X EN ELS FACTORS MÉS RELLEVANTS EN LA PREDICCIÓ DE Y.

A la figura 4 es mostra un gràfic comparatiu de l'estructura diferent dels diferents mètodes existents per al calibratge multivariant [3].

7.4. MÈTODE DE REGRESSIÓ PER COMPONENTS PRINCIPALS (PCR)

7.4.1. Anàlisi de components principals PCA: compressió de X *en els factors més significatius*

La importància de l'aplicació del mètode de calibratge per components principals (*principal component analysis*), PCA, es troba sobretot en aquells casos en què les variables X són colineals. Hi ha moltes raons per a trobar-nos amb problemes de colinealitat, per exemple, quan el nombre d'analits i d'interferències és menor que el nombre de respostes mesurades X (menor que el nombre de longituds d'ona mesurades), respostes que, a més, es poden assemblar entre elles. Aquesta colinealitat significa que a la matriu X hi ha certes variables que porten la major part de la informació. La redundància i les altres petites contribucions o soroll de fons que acompanyen les dades experimentals poden eliminar-se a partir de l'anàlisi de components principals, PCA.

La finalitat de l'anàlisi mitjançant components principals, PCA, és expressar la informació principal de les variables a X (x_k, $k = 1, 2, ..., K$) amb un nombre més petit de variables $T = (t_1, t_2, ..., t_A)$, amb $A << K$, que són els anomenats *components principals* de X.

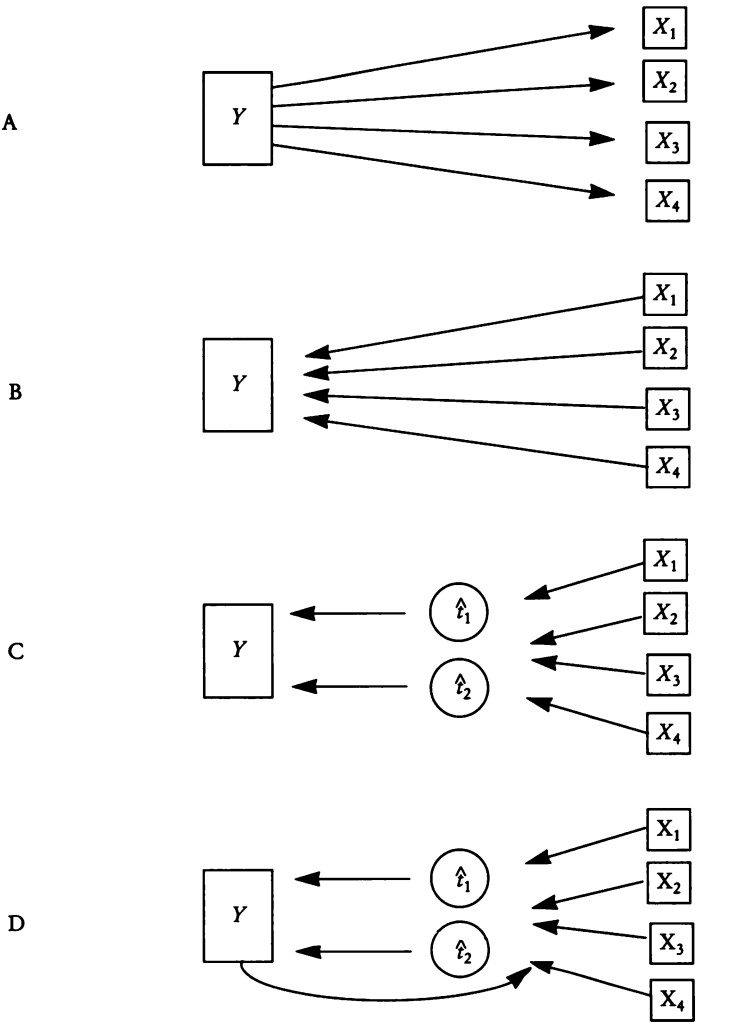

FIGURA 4. Gràfic comparatiu dels diferents procediments de calibratge multivariant: A) anàlisi multicomponent clàssica; B) calibratge invers per regressió múltiple lineal; C) regressió per components principals; D) regressió per mínims quadrats parcials.

Es calcula el primer component principal p_1 com el vector normalitzat a longitud unitària que explica el màxim de la variància de X, o, el que és el mateix, que maximitza $p_1'X'Xp_1 = t_1't_1$. El segon

factor, p_2, es calcula com el vector que maximitza la mateixa quantitat $p_2'X'Xp_2 = t_2't_2$, ara sota la restricció que t_1 i t_2 siguin ortogonals entre si, és a dir, que $t_1't_2 = 0$. El procediment continua d'aquesta manera mantenint sempre el criteri d'ortogonalitat respecte als factors anteriors. Es pot demostrar que:

$$P'P = I \text{ i que } T'T = diag(\lambda_a),$$

on I és la matriu identitat; els elements p_a de la matriu P són els vectors propis de $X'X$ i *diag* és la matriu diagonal que conté els seus valors propis λ_a; això vol dir que els vectors p compleixen l'equació:

$$X'Xp_a = p_a \lambda_a.$$

De la mateixa manera, els *scores* t_a representen els vectors propis de XX', escalats a la longitud $\sqrt{\lambda_a}$.

La matriu de respostes o espectres, aleshores es pot escriure com:

$$X = T P'.$$

Aquesta igualtat és estrictament certa quan es troben tots els vectors propis $A = K$ (amb alguns d'aquests vectors propis que tindran valors propis probablement gairebé nuls). Escollint només els primers components principals, podem aproximar la matriu de dades amb l'equació:

$$X = T P' + E,$$

on ara E representa la matriu dels residuals a X per al model amb el nombre de components seleccionat. El nombre de components s'escollirà amb el criteri que E sigui de l'ordre de l'error experimental. La matriu T està associada als designadors de les files de la matriu X, i la matriu P' està associada als designadors de les columnes de la matriu X (producte de les files de T per les columnes de P'). Una propietat interessant del mètode PCA és que, un cop coneguda la matriu T, es pot calcular fàcilment la matriu P, i viceversa, un cop coneguda P, es pot trobar T amb equacions de càlcul que

queden notablement simplificades a causa de les propietats d'ortogonalitat dels factors entre ells:

$$P = (T'T)^{-1}T'X = diag(\lambda_a)T'X,$$

$$T = X\,\mathrm{P}\,(P'P)^{-1} = X\,P.$$

L'aspecte principal del mètode PCA és que s'ha d'escollir el nombre correcte de factors o components que s'han de considerar. Hi ha diferents alternatives, les més importants són les que es deriven del mètode de validació proposat per Wold [4], i la proposada per Malinowski [5] basada en tests de significació estadística. A partir de les representacions gràfiques dels factors que tenen valors propis més grans, o components principals, es poden conèixer les propietats del sistema.

A continuació es descriu l'algorisme NIPALS, molt emprat segons la bibliografia del calibratge multivariant per a trobar els components principals.

7.4.2. *Algorisme NIPALS per a trobar els components principals*

Aquest algorisme extrau un factor cada vegada. Cada factor s'obté de manera iterativa mitjançant regressions repetides de la matriu X_{a-1} (matriu de dades residual després de la subtracció de la contribució dels primers $a-1$ components principals) sobre els factors (*scores*) t per a obtenir un millor p, i, viceversa, de la matriu X_{a-1} sobre els factors (*loadings*) p per a obtenir un millor t. L'algorisme funciona de la manera següent: les variables de la matriu de dades original, X, s'escalen per tal d'assegurar un nivell de soroll similar en totes elles. Després se centren totes les variables —per exemple, per subtracció de les seves mitjanes— per a donar X_0. Per a un nombre de factors $a = 1, 2, ..., A$ es calculen t_a i p_a a partir de X_{a-1}.

7.4.3. *Inici*

Se seleccionen com a valors inicials de t_a les columnes en X_{a-1} que tenen una suma de quadrats més gran. Es repeteixen els passos següents fins a la convergència:

i) Es millora l'estimació del vector *loading* p_a per projecció de la matriu X_{a-1} sobre t_a:

$$p_a' = (t_a' t_a)^{-1} t_a' X_{a-1}.$$

ii) Es normalitza o es trasllada a escala la longitud de p_a a 1,0 per tal d'evitar l'ambigüitat en la seva longitud:

$$p_a = p_a (p_a' p_a)^{-0,5}.$$

iii) Es millora l'estimació del factor *score* t_a per projecció de la matriu X_{a-1} sobre p_a:

$$t_a = X_{a-1} p_a (p_a' p_a)^{-1}.$$

iv) Es millora l'estimació del valor propi τ_a:

$$\tau_a = t_a t_a'.$$

v) Es comprova la convergència: si la diferència entre el darrer valor de τ_a estimat i l'obtingut en la iteració anterior és més petit que una certa constant petita definida prèviament, per exemple, 0,0001 vegades τ_a, el mètode ha convergit per aquest factor. Si no és així, es torna al pas *i*. Se subtrau l'efecte del factor (en cas d'haver assolit convergència):

$$X_a = X_{a-1} - t_a p_a'$$

i es torna a l'inici del procés per al factor següent.

7.5. MÈTODE DE REGRESSIÓ PER MÍNIMS QUADRATS PARCIALS (PLSR)

Partial least-squares, PLS, és un terme poc concret que s'utilitza per a una família de mètodes de modelat multivariant relacionats tècnicament i que es deriven dels conceptes bàsics desenvolupats per H. Wold a partir de 1975 [6], amb la finalitat pràctica de solucionar problemes concrets de l'econometria, de les ciències socials i, actualment, d'altres àrees, com ara la química. Quan les tèc-

niques de modelat multivariant tradicionals són aplicades a dades reals amb moltes variables colineals i amb un nombre relativament petit d'observacions, sorgeixen greus problemes d'identificació i de convergència. Molts d'aquests problemes es poden solucionar aplicant l'aproximació molt més empírica dels mètodes PLS, basats en una sèrie d'ajusts locals per mínims quadrats, a diferència dels mètodes tradicionals de regressió basats en el principi de màxima probabilitat.

PLSR intenta produir models bilineals de calibratge amb el nombre mínim de dimensions possibles, i procura que aquestes dimensions siguin les més rellevants possibles en la predicció de les concentracions d'analit. L'objetiu del mètode PLSR és comprimir la matriu de respostes X en els factors més «rellevants» en la predicció de la matriu de concentracions dels analits en les solucions de calibratge Y. Aquests factors no coincidiran, en general, amb els factors trobats pel mètode PCR, ja que aquests últims es trobaven tenint en compte només la matriu de respostes X. Mentre que en el mètode PCA els factors s'extrauen amb l'objetiu d'explicar la màxima variància de la matriu de respostes X, en el mètode PLS els factors s'extrauen amb l'objetiu d'explicar la màxima covariància entre la matriu de respostes X i la matriu de concentracions Y.

Com ja s'ha dit, els mètodes de calibratge bilineal permeten l'obtenció de prediccions fiables de les concentracions $Y = f(X)$ mitjançant la projecció de les variables $X = (x_1, x_2, x_3, ..., x_K)$ en un nombre menor de variables $T = (t_1, t_2, ..., t_A)$. És a dir:

$$T = X\,V$$

i la posterior utilització d'aquestes variables T com a regressors de les concentracions y. Com a conseqüència de l'aplicació del mètode PLSR, les causes de variació a les variables X es comprimeixen en un conjunt de variables estabilitzades i més fàcilment interpretables, i es deixa de banda una gran part del soroll. La qüestió més important és, no obstant això, com definir aquests factors més rellevants T que permeten, alhora, una millor interpretació i una millor predicció. El mètode PCR descrit defineix els factors T com les projeccions sobre els fenòmens dominants a X. Per a dades colineals, el mètode PCR ja proporciona una millora substancial respecte al mètode d'anàlisi multicomponent

clàssic mitjançant MLR o CLS. Però també és cert que de vegades es pot millorar el mètode PCR eliminant alguns dels factors més importants de X, si és que no tenen importància en la predicció de les concentracions dels analits Y. Tenint en compte aquesta idea, s'ha desenvolupat el mètode de regressió PLS o PLSR (regressió per mínims quadrats parcials). Com es pot veure a la figura 4, el mètode PLSR difereix del mètode PCR en el fet que s'utilitzen les variables de Y per a fer la descomposició bilineal de X. En fer el balanç de la informació proporcionada per X i per Y, el mètode redueix l'impacte que pot proporcionar la informació irrellevant de X de cara al modelat de Y i a la seva predicció posterior.

El mètode PLSR pot proporcionar models de calibratge una mica més simples que el mètode PCR, encara que normalment el nombre de factors òptims sigui el mateix. Per a dades poc precises, és millor utilitzar el mètode PLSR que el mètode PCR. No obstant això, el preu que s'ha de pagar amb el PLSR és la major complexitat de l'algorisme de càlcul, i la pèrdua d'ortogonalitat dels factors T en comparació amb el mètode PCR. A continuació es descriu el mètode PLSR en la seva versió més senzilla en forma d'algorisme.

7.5.1. *PLSR per a una variable (PLS1) cada vegada*

7.5.1.1. Calibratge

S'aplica successivament a cada una de les variables y (concentració de cada analit que es vol determinar). Es realitza un pretractament de les variables d'entrada X, que s'escalen a nivells de soroll similars. Se centren tant la matriu de respostes X com el vector que conté les concentracions de l'analit d'interès en les solucions patró y, per tal d'obtenir X_0 i y_0 (per subtracció dels valors de les mitjanes). Els passos següents de l'1 al 6 es realitzen per a cada factor $a = 1, 2, \ldots A_{max}$, on A_{max} és el nombre màxim de factors PLSR que s'han de calcular. Aquest nombre, A_{max}, pot ésser més gran que el nombre de fenòmens que s'espera que siguin presents a X, per tal d'incloure també en el model els fenòmens inesperats.

$$X_0 = X - 1x_m' \qquad y_0 = y - 1y_m$$

Pas 1

Es troba el vector *loading weight* (w_a) unitari que maximitza $w_a' X_{a-1}' y_{a-1}$, és a dir, que fa màxima la covariància escalada entre X_{a-1} i y_{a-1} (la covariància entre les respostes instrumentals i les concentracions de l'analit en els patrons):

$$X_{a-1} = y_{a-1}\, w_a' + E,$$

$$w_a = c\, X'_{a-1}\, y_{a-1},$$

on

$$c = (y'_{a-1}\, X_{a-1}\, X'_{a-1}\, y_{a-1})^{-0,5}.$$

Pas 2

Es troben els factors t_a com a projeccions de X_{a-1} sobre w_a (model local):

$$X_{a-1} = t_a\, w_a' + E,$$

$$t_a = X_{a-1}\, w_a.$$

Pas 3

Es fa la regressió de X_{a-1} sobre els factors t_a per a trobar els *loadings* p_a' (model local):

$$X_{a-1} = t_a\, p_a' + E,$$

$$p_a = X_{a-1}'\, t_a / (t_a' t_a).$$

Pas 4

Es produeix a l'estimació del factor (*loading*) «químic» q_a utilitzant el model local:

$$y_{a-1} = t_a\, q_a + f,$$

que té per solució:

$$q_a = y_{a-1}{}'\, t_a \,/\, (t_a{}'t_a).$$

Pas 5

Es troben les noves matrius de residuals a X i y per subtracció de l'efecte ja explicat pel factor considerat.

$$E = X_{a-1} - t_a\, p_a{}' \qquad f = y_{a-1} - t_a q_a$$

Es calculen tests estadístics i de detecció d'aberrants sobre aquests residuals després de considerar a factors.

Se substitueix:

$$X_a = E,$$

$$y_a = y,$$

$$a = a + 1.$$

Pas 6

Es comprova si s'ha assolit el nombre òptim de factors PLS, A, que s'han de retenir en el model de calibratge. S'utilitzaran tests de validació per a veure que aquest és el nombre correcte.

7.5.1.2. Predicció

Un cop establert el model de calibratge, es pot realitzar la predicció i l'anàlisi de mostres desconegudes. Per exemple, es pot utilitzar l'equació:

$$\mathrm{y} = 1b_0 + Xb,$$

on:

$$b = W\,(P'W)^{-1}\,q,$$

$$b_0 = y_m - x_m'b.$$

W, P i q són les matrius i els vectors dels factors PLS obtinguts en el calibratge; y_m és la mitjana de concentracions emprada en el model de calibratge i x_m és la resposta centrada de la mostra problema o incògnita.

Dins de la quimiometria, el mètode PLS constitueix actualment una de les eines més emprades per a la modelització de les relacions lineals entre mesures multivariants. El mètode PLS ha estat introduït en la literatura química en forma d'algorisme, i només recentment han estat plenament enteses les seves propietats estadístiques i numèriques [7-10]. A més de l'algorisme NIPALS descrit en el present treball, han estat presentats altres algorismes que, essencialment, donen resultats molt semblats, però que, de vegades, en permeten una millor interpretació.

Quan les relacions entre les mesures multivariants no són lineals, és a dir, quan les relacions entre les mesures instrumentals (matriu X) i les concentracions dels analits a determinar (matriu Y) no són lineals, s'han desenvolupat tota una altra sèrie de mètodes de calibratge multivariant no lineal que no es troben descrits en el present treball [11].

El tractament de mesures multivariants que tenen una estructura més complexa, i que es troben agrupades en tensors d'ordre superior, com ara matrius (tensors d'ordre 2), cubs (tensors d'ordre 3) o hipercubs (tensors d'ordre superiors a 3), requereix la utilització de mètodes de calibratge d'ordre superior o mètodes de calibratge tensorial [12]. Exemples de mesures multivariants d'ordre superior es troben freqüentment en química analítica quan s'utilitzen les anomenades tècniques «hifenades» d'anàlisi, és a dir, quan la mostra s'analitza mitjançant l'aplicació simultània de dues o més tècniques instrumentals d'anàlisi, com ara la cromatografia i l'espectrometria, o quan s'utilitzen tècniques que ja proporcionen per elles mateixes una estructura superior de dades, com és el cas de la fluorescència, on excitació i emissió forneixen els dos ordres de mesura d'una matriu de dades en l'anàlisi de cada mostra. Aquest és un camp d'aplicacions analítiques actualment molt ric, i encara ho serà

més en un futur proper, ja que l'evolució dels mètodes instrumentals d'anàlisi es produeix simultàniament a l'evolució i el desenvolupament dels nous procediments quimiomètrics.

Tant els mètodes de calibratge no lineal com els mètodes de calibratge tensorial requereixen un tractament detallat i aprofundit, i, per tant, no són descrits en aquest treball, dedicat especialment als mètodes de calibratge lineal.

REFERÈNCIES

1. R. S. HUNTER (1975). *The measurement of the appearance*. Nova York: John Wiley and Sons.
2. R. S. HUNTER (1983). «Quantification of sensory colour differences from physical measurement: Implications for food appearance». A: VALBERG, SEIM, MARTENS [*et al.*] [ed.], *Food Research and Data Analysis*. Londres: Applied Science Publishers.
3. H. MARTENS i T. NAES (1989). *Multivariate Calibration*. Londres: John Wiley and Sons.
4. H. WOLD (1978). *Technometrics*, **20**, 397-406.
5. E. R. MALINOWSKI (1991). *Factor Analysis in Chemistry*. 2a. ed. Nova York: John Wiley and Sons.
6. H. WOLD (1981). «Soft modeling: the basic design and some extensions.» A: K. G. JÖRESKOG i H. WOLD [ed.]. *Systems under indirect observation, casuality-structure prediction*. Amsterdam: North Holland.
7. S. DE JONG (1993). *Chemom. and Intelli. Lab. Systems*, **18**, 251-263.
8. A. LORBER, L. E.WANGEN i B. R. KOWALSKI (1987). *Chemometrics*, **1**, 19-31.
9. R. MANNE (1987). *Chemom. and Intelli. Lab. Systems*, **2**, 283-290.
10. A. HÖSKULDSSON (1987). *J. of Chemometrics*, **2**, 211-228.
11. S. SEKULIK, M. B. SEASHOLTZ, Z. WANG i B. R. KOWALSKI (1993). *Anal. Chem.*, **65**, 835A-845A.
12. E. SÁNCHEZ i B. R. KOWALSKI (1988, 1990). *J. of Chemometrics*, **2**; 247-263, 265-280; **4**, 29-45.