

# La informació física no específica: el best-seller de la quimiometria

## Non-specific physical information: A Chemometric best-seller

Michele Forina,\* Monica Casale i Paolo Oliveri

Dipartimento di Chimica e Tecnologie Farmaceutiche ed Alimentari, Università di Genova, Italia

Conferència inaugural del IX Memorial Enric Casassas (IQS, Barcelona, 2/12/2009, text traduït pel Dr. Xavier Tomàs)

**Resum.** En aquesta conferència, homenatge al Prof. Enric Casassas, el Prof. Forina exposa de manera molt didàctica, tres idees importants a tenir presents per tots els que treballen en Quimiometria o utilitzen tècniques quimiomètriques. L'objectiu sempre ha de ser resoldre una situació química real, procurant emprar eines de qualitat coneguda i no cal tenir recança en aplicar-les a situacions com ara la informació física no específica.

**Paraules clau:** Quimiometria, Informació física no específica, Validació, Procés analític, Valors anòmals, Selecció de variables.

**Abstract.** This lecture, a tribute to Prof. Enric Casassas, Prof. Forina presents in a very didactic way, three important ideas to take into account for all those who are working in chemometrics or use chemometric techniques. The main goal should always be to resolve a real chemistry situation, trying to use tools of known quality and we should not be afraid to apply Chemometrics to situations such as non-specific physical information.

**Keywords:** Chemometrics, Non-specific physical Information, Validation, Analytical Process, Outliers, Selection of variables.

## Introducció

Fa aproximadament uns cinquanta anys es va iniciar la Quimiometria. La Figura 1 mostra un fragment de la carta que Bruce Kowalski i Svante Wold, els fundadors de la Chemometrics Society, van enviar als químics interessats en l'aplicació de mètodes matemàtics i estadístics a la Química, convidant-los a formar part de la Chemometrics Society.

En un principi, els quimiòmetres van emprar informació química o informació física reconeguda abastament com a informació útil. El conjunt de dades ARCH,<sup>1,2</sup> que formava part del primer programari quimiomètric, ARTHUR, és un conjunt de dades de 75 mostres d'obsidiana provinents de quatre localitzacions properes a San Francisco caracteritzades pel contingut en 10 metalls. En el conjunt de dades KETONES, que formava part de la primera versió del programari SIMCA, es descriuen vint cis i trans cetones mitjançant set variables obtingudes de llurs espectres IV i UV, i que descriuen l'efecte del grup carbonil i del doble enllaç sobre l'absorbància.

Molt aviat els quimiòmetres van començar a utilitzar informació física no específica.

Un cromatograma és exactament el registre d'un senyal físic no específic que esdevé informació química quan hom procedeix a assignar els pics cromatogràfics a compostos químics. Al seu torn, la informació química

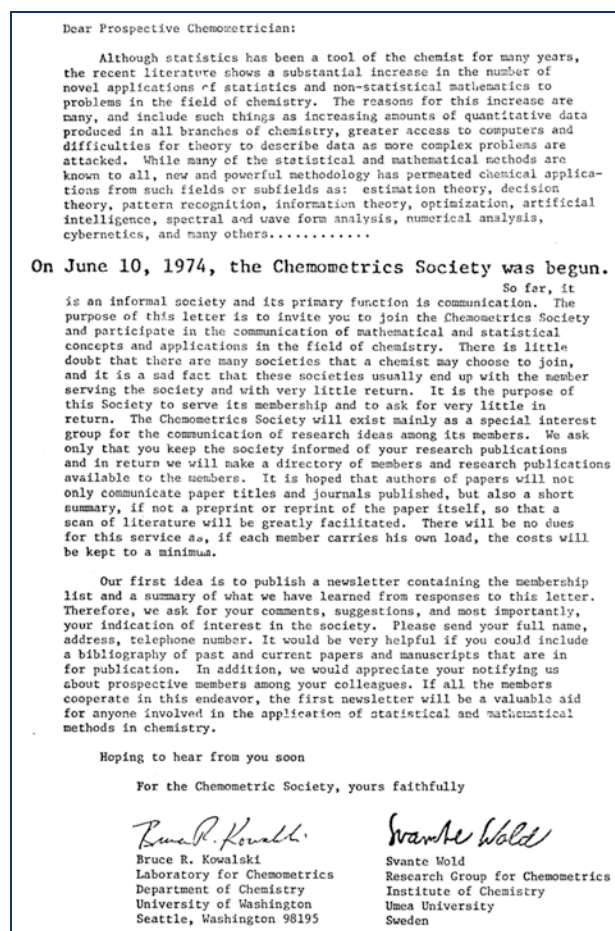


Figura 1. El "certificat de naixement" de la Chemometrics Society.

pot ser o no informació específica. La interpretació química d'un cromatograma pot esdevenir molt difícil, és especialment en el cas dels cromatogrames complexos.

L'any 1978, Bruce Kowalski<sup>3</sup> va utilitzar cromatogrames d'espai de cap «cecs» en un estudi que tenia com a objectiu la caracterització de marques de whisky.

**Correspondència:** Michele Forina. Dipartimento di Chimica e Tecnologie Farmaceutiche ed Alimentari, Università di Genova, Via Brigata Salerno 13, I-16147 Genova, Italia.

Tel.: +39 0103532630. Fax: +39 0103532684.

A/e: forina@dictfa.unige.it

Com que l'objectiu de l'estudi no era augmentar els coneixements químics, la interpretació química dels diferents pics cromatogràfics no era, en conseqüència, necessària.

## Fonts d'informació física no específica

Avui dia, la quimiometria s'aplica amb molta freqüència a l'estudi d'informació física no específica provinent de diferents àmbits. Per a cada mostra, els instruments proporcionen milers (en ocasions, milions) de variables, no totes necessàriament útils. L'objectiu és, doncs, obtenir la informació útil per resoldre tant problemes de classificació com problemes de regressió. Els sectors d'aplicació més importants són el control de qualitat, el control de processos, la traçabilitat dels aliments i la metabolòmica.

Les fonts d'informació física més emprades són l'absorció o la reflectància dels espectres electromagnètics (com el visible, l'ultraviolat, l'infraroig, els raigs X, la fluorescència i el Raman), els espectres de masses (com ara els obtinguts per un «nas artificial»), la potenciometria (dades obtingudes per un conjunt d'elèctrodes), la voltametria (com és el cas d'algunes «llengües artificials»), els espectres de ressonància magnètica nuclear, els diversos tipus de cromatografia, etc.

La informació física que hom pot obtenir d'una mostra pot ser un vector (com ara l'espectre obtingut d'una mostra de composició uniforme), una matriu bidimensional (com és el cas d'un espectre Raman) o també una matriu tridimensional (com és el cas d'una imatge).

Al seu torn, cada tipus de senyal pot correspondre a una informació molt diferent segons les característiques de l'instrument de mesura o els tractaments especials de les dades originals. A tall d'exemple, l'espectre obtingut per espectroscòpia en l'infraroig proper (NIRS) depèn molt de la sonda emprada, dels tipus de senyal mesurats (absorció, reflectància, transreflectància), de l'entorn físic de la mostra (com és el cas del NIR en dues dimensions, en el qual s'aplica una tensió de petita amplitud a la mostra, la qual cosa provoca una fluctuació dinàmica dels senyals IR) i de la mateixa font (com ara en TOF-NIRS, on la font és un làser de polsos).

Segons el tipus de tractament (derivades, eliminació de tendències, transformada de Fourier, etc.), la informació física pot donar resultats molt diferents. És més: a vegades és possible combinar la informació obtinguda amb diferents instruments amb uns resultats excel·lents.<sup>4</sup>

Entre les diferents fonts d'informació física no específica, l'espectroscòpia d'infraroig proper (NIR) és avui la tècnica líder en control de qualitat tant del producte final com dels productes intermedis i de les matèries primeres. Això vol dir que milers de químics i tècnics utilitzen la tècnica NIR cada dia fent milions de determinacions. Per aquesta raó, l'espectroscòpia NIR és un *best-seller* de la quimiometria.

La imatge de la quimiometria no és la pròpia dels quimiòmetres, la de les poques persones que fan servir mètodes de quimiometria avançada i publiquen articles d'alta qualitat, tot i que molt difícils de comprendre per a la majoria dels químics. La imatge pública de la quimiometria és la que presenten els *best-sellers* a la gran majoria dels químics. En dir *best-seller*, ens referim simplement a un producte amb una circulació molt gran, malgrat que no que tingui necessàriament una gran qualitat. La qualitat final d'un *best-seller* pot ser pobra o, si més no, limitada.

Normalment, els químics que fan servir l'espectroscòpia NIR en control de qualitat utilitzen el programari subministrat pels fabricants dels instruments com si fos una «caixa negra». En general, aquest programari té una qualitat bona o almenys acceptable, però sempre ha estat dissenyat com una eina senzilla per a ser emprada per persones sense coneixements d'estadística elemental ni de quimiometria. Per aquesta raó, juntament amb els instruments i el programari, s'acostumen a proporcionar unes «regles d'or» que els usuaris segueixen després fil per randa.

## La qualitat dels "best sellers" quimiomètrics

A continuació presentarem uns pocs exemples de la qualitat dels nostres *best-sellers*. Tots els exemples fan referència a l'ús de l'espectroscòpia NIR. Seguint l'esperit de la dita popular, farem esment del pecat, però no del pecador.

### Exemple 1.

En una conferència, un dels ponents va dir: «En el cas del potassi, el de variància en validació creuada explicada va ser del 7 %. No és gaire, però sí que és *prometedor*».

$Q^2$ , és a dir, el de variància explicada en validació creuada, s'utilitza amb molta freqüència en la presentació de resultats NIR quan hom aplica un calibratge multivariat.  $Q^2$  és un paràmetre enganyós respecte al paràmetre d'interès, la desviació estàndard de l'error en predicció (SDEP) amb la qual es relaciona mitjançant la funció quadràtica inversa:

$$\frac{\text{SDEP}}{s_b} = \sqrt{1 - Q^2/100}$$

Aquí,  $s_b$  és la desviació estàndard abans de la regressió o desviació estàndard sense model, calculada a partir de les diferències entre els valors de la variable resposta i la mitjana de les respostes.

La figura 2 mostra com un valor igual a 7 de  $Q^2$  correspon a un valor de SDEP pràcticament igual a la desviació estàndard abans de la regressió. Aquest valor es deu a una petita correlació (no significativa i fortuïta) entre el potassi i l'espècie química, que té un efecte sobre l'espectre NIR. No hi ha res de prometedor en aquest valor tan petit.

Però fins i tot valors grans de  $Q^2$  no indiquen necessà-

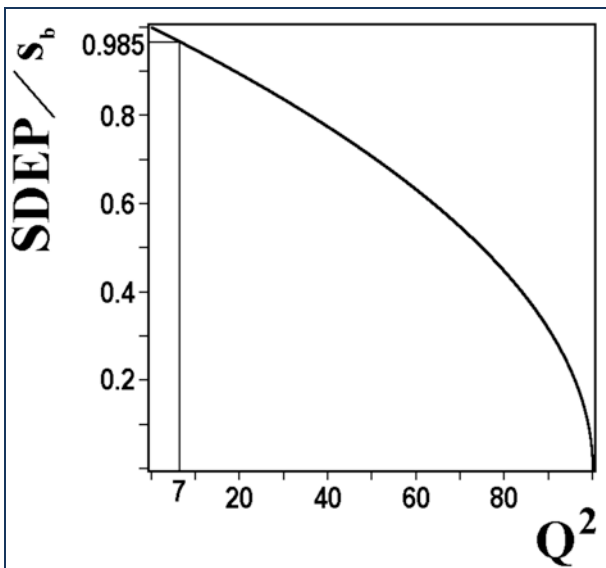


Figura 2. Relació de  $Q^2$  i el quocient entre la desviació estàndard de l'error en predicció (SDEP o SEP) i la desviació estàndard abans de la regressió.

riament que s'hagin obtingut bons resultats. Un valor de  $Q^2$  igual a 50 vol dir que SDEP és aproximadament igual a 0,7 vegades la desviació estàndard abans de la regressió. Aquesta és una quantitat experimental, una estimació del valor real de  $\sigma$  que ens és desconegut.

Ara bé:

$$\frac{\text{SDEP}^2}{\sigma^2} \nu$$

Això es distribueix com una variable  $\chi^2$  amb  $\nu$  graus de llibertat, sent  $\nu$  el nombre de mostres emprades per a la validació. L'interval de confiança de  $\sigma^2$  s'obté a partir dels valors crítics de la distribució de  $\chi^2$ . Així, per  $\nu = 20$ , l'interval de confiança al 95 % és aproximadament  $\pm 35$  %, la qual cosa vol dir que el model de calibratge es pot comportar pitjor (o també millor) del que s'esperaria del valor de la SDEP obtinguda.

Ho sento pels estadistes, però Svante Wold, un dels fundadors de la quimiometria, va escriure el següent:<sup>5</sup>

*Si nosaltres, erròniament, considerem la filosofia de l'estadística més sòlida que la de la química, prendrem el trist i desafortunat camí de la biometria, la psicometria... que avui són de poc o cap interès per als biòlegs, els psicòlegs... Aquesta desgràcia és la conseqüència de considerar més important el «rigor» matemàtic i estadístic que no pas el fet de resoldre els problemes científics. Naturalment, sempre s'ha de ser tan rigorós com sigui possible, però aquest rigor no ha de ser un rigor mortis; el primer de tot sempre és el sentit químic.*

Ara bé, això també vol dir que quasi sempre és necessari un mínim coneixement i ús de l'estadística bàsica.

### Exemple 2.

Generalment, l'índex de maduració de les pomes, paràmetre molt important per a un bon producte després del seu emmagatzematge, es determina mitjançant la prova de iode i midó. Seguidament, un grup d'experts avalua la coloració obtinguda amb el reactiu en una zona de la poma com una puntuació en una escala sensorial.

Els resultats no acostumen a ésser satisfactoris degut al caràcter subjectiu de l'avaluació i a la seva gran dispersió.

Per aquestes raons, un equip d'investigadors van desenvolupar un model objectiu seguint l'esquema que es mostra a la Figura 3 i van fer la prova sobre unes dues mil pomes.

Van construir el model PLS amb prop de cent mil variables predictores (les reflectàncies dels píxels de les imatges NIR obtingudes per cada mostra) i com a resposta les corresponents puntuacions atorgades a les mostres per l'equip d'experts. Els resultats no van ser satisfactoris, més o menys equivalents als obtinguts per l'equip de tastadors.

Malgrat tot, es va continuar l'estudi, el que significa més pomes, més experts, instruments amb un interval de longituds d'ona més ampli, imatges més grans i en conseqüència més diners.

Sembla com si les persones que treballen amb la Quimiometria en la pràctica ignoressin que un model de calibratge mai pot tenir una qualitat superior a la que tenen les respostes que s'han mesurat (en aquest cas les puntuacions sensorials) i que s'han emprat per construir el model.

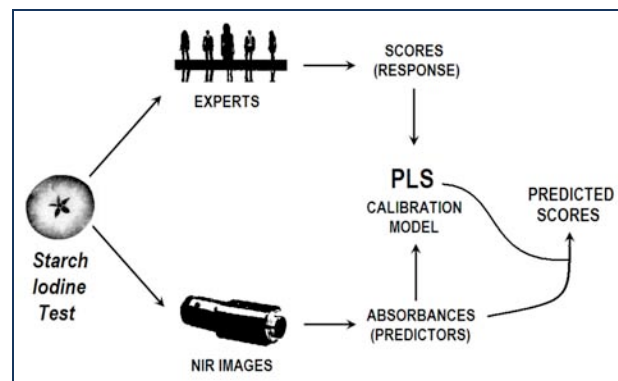


Figura 3. Esquema de calibratge emprat per l'índex de maduració de les pomes.

### Exemple 3.

Una nova tècnica instrumental, TOF-NIR (temps de vol-NIR), es va aplicar a la determinació de sucres en fruites de pell gruixuda. El model es va construir sobre unes cent mostres i la SDEP resultant va ser del 0,22 %. Aquest va ser un resultat esplèndid, gairebé increïble, ja que aquesta precisió és molt poc freqüent en les anàlisis.

Si analitzem els resultats de la predicció que es mostren a la figura 4, observem que:

a) La resposta està expressada en percentatge de sucre. Això implica que el valor indicat de la SDEP és una desviació absoluta, no relativa.

b) L'interval d'error correspon aproximadament al 70 % de l'error de les variables resposta, aproximadament 1,4, és a dir, del 9 % al 10,4 %. La precisió de la tècnica no és pas per entusiasmar-se.

Ara bé, també podríem emprar un model alternatiu. Aquest model es desenvoluparia de la manera següent:

$$y_i^{\text{PREDITA}} = \frac{\sum_{j \neq i} y_j^{\text{MESURADA}}}{N-1}$$

Aquí,  $y$  és la variable resposta mesurada o predita. El valor predit de la mostra  $i$  s'obté a partir de la resposta mesurada en les altres  $N - 1$  mostres, és a dir, el model està construït en validació «total». Aquest model té una SDEP igual a 0,28 %, que no difereix significativament del valor 0,22 % obtingut per TOF-NIR. Aquest model, però, és molt més econòmic, ja que no necessita instruments.

Fem servir aquesta ironia per mostrar que amb massa freqüència s'utilitza un munt de treball i de diners per desenvolupar mètodes inútils i llurs resultats es presenten de manera enganyosa.

Validació «total» és allò que els quimiòmetres denominen *validació un a un* (LOOV). En aquest cas, el model s'avalua tantes vegades com mostres es deixen fora, cada vegada una d'elles. S'utilitza, doncs, el model per predir el valor de la resposta (o de la categoria) de la mostra que ha estat exclosa. Els resultats obtinguts amb la validació LOOV es consideren generalment massa optimistes i molts consideren que és preferible subdividir el conjunt en grups (entre tres i set) i calcular el model tantes vegades com grups s'han definit.

La capacitat de predicció s'obté cada vegada amb el model aplicat a la mostra del grup que s'ha exclòs. Aquest procediment es coneix amb el nom de *validació creuada* (CV).

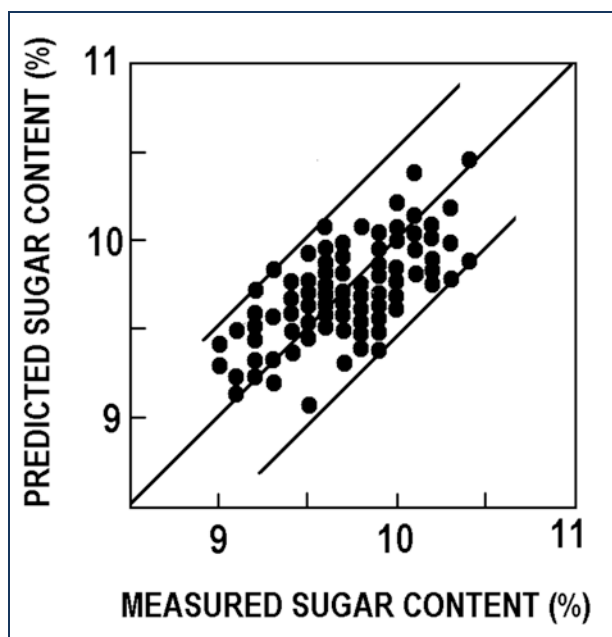


Figura 4. Predicció de sucre mitjançant TOF-NIRS.

## La Quimiometria i el procés químic analític

La majoria dels químics considera la quimiometria simplement com un conjunt d'eines per a l'anàlisi de les dades, just al final del procés de l'anàlisi química. Al contrari, la quimiometria és present en totes les etapes del procés analític (figura 5).

Aquest procés es duu a terme per tal d'obtenir la informació química o física (però relacionada amb la composició química) necessària d'un problema (si és possible, d'un problema real).

La quimiometria és molt més que tot això i ha de ser emprada en cadascuna de les etapes del procés.

La quimiometria ajuda a clarificar els objectius del procés, a definir el problema. La quimiometria és necessària a l'hora de preparar el disseny del pla de mostreig, és a dir, de recollir mostres representatives de totes les fonts de variació del sistema objecte del procés, possiblement, de la manera més econòmica que es pugui. Molt sovint, a la pràctica, les mostres no són gaire representatives, però les seves característiques poques vegades es discuteixen. Les tècniques quimiomètriques són útils per optimitzar les tècniques de mesura, i a la fi del procés són eines per a l'anàlisi de les dades, per extreure de les dades originals la informació útil per als objectius del problema. Cada un d'aquests aspectes constitueix un ampli capítol de la quimiometria.

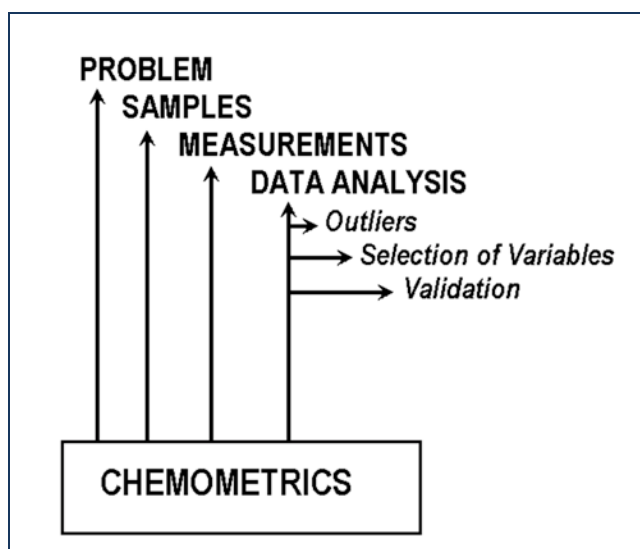


Figura 5. La Quimiometria i el procés d'anàlisi química.

En referència a l'anàlisi de les dades hi ha punts concrets on no és recomanable aplicar les "regles d'or" d'una manera automàtica.

### Valors anòmals, "Outliers"

Els valors anòmals es defineixen com a objectes (mostres) que tenen unes característiques molt diferents a la resta dels objectes del problema. Una bona pràctica, generalment admesa, consisteix a eliminar aquests valors anòmals. El guany segur és que, tant en problemes de classificació com de calibratge, després de l'eliminació, el model proposat funcionarà millor.

En el cas de l'anàlisi exploratòria de dades o en problemes de classificació, l'eliminació dels valors anòmals es realitza mitjançant la prova de Hotelling (prova multivariada anàloga a la prova de Student), que implica el compliment de la hipòtesi que totes les variables es distribueixen segons una llei normal, una circumstància que molt poques vegades es compleix quan hom treballa amb dades reals. En el cas dels problemes de regressió, els valors anòmals són també la causa per la qual l'error en predicció és molt gran.

Provaré d'explicar el que pot passar amb l'eliminació a cegues dels valors anòmals amb un exemple que no és químic. La figura 6 mostra el resultat després d'analitzar la



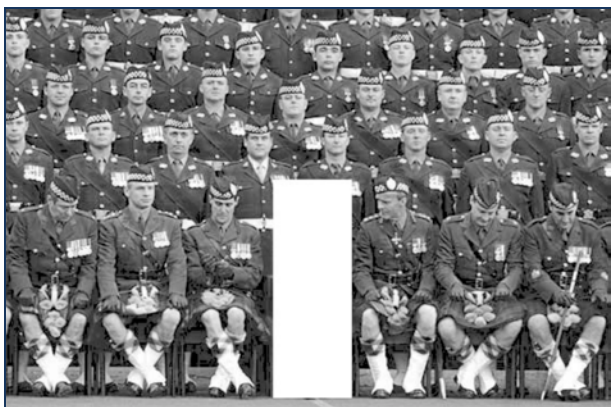


Figura 6. La informació després d'haver eliminat el valor anòmal.

informació, detectar el valor anòmal i eliminar-lo.

Pel que fa a la identificació del valor anòmal és evident que una part de la informació és redundant i per tant es pot eliminar. La figura 7 mostra la informació un cop eliminada la informació redundant.



Figura 7. La informació un cop s'ha eliminat la part redundant a l'hora de detectar el valor anòmal.

La figura 8 mostra el valor anòmal en l'espai de la informació reduïda i la figura 9 mostra el valor anòmal en l'espai d'informació original.



Figura 8. El valor anòmal en l'espai d'informació reduïda.



Figura 9. El valor anòmal en l'espai d'informació original.

Aquest exemple posa de manifest que el valor anòmal pot ser molt important a l'hora de comprendre les dades i que l'eliminació d'informació pot ser útil només en aparença.

### Selecció de variables

Quan hom treballa amb informació no específica, gairebé sempre una part de la informació no és útil, la qual cosa anomenem *soroll*. S'ha demostrat que l'eliminació d'aquesta informació soroll millora la qualitat dels models de classificació o de calibratge.

Hi ha una gran quantitat de tècniques per a l'eliminació de variables predictores inútils. En el cas dels problemes

de classificació, la tècnica més important és l'anàlisi discriminant lineal per etapes (SLDA).

Pel que fa als problemes de calibratge, s'ha desenvolupat una gran quantitat de tècniques que van des de la coneguda regressió per mínims quadrats pas a pas fins a l'aplicació d'algorismes genètics.

Els resultats següents fan referència a un problema de classificació amb tres classes i seixanta objectes disponibles per tal de construir un model de classificació (conjunt de dades FAN). El nombre de variables va ser d'un miler, massa per aplicar directament una tècnica de classificació probabilística com ara l'anàlisi discriminant lineal (LDA). Cal, doncs, procedir a realitzar necessàriament una selecció de variables.

La selecció es va realitzar per SLDA, amb les restriccions habituals (valor de tall de l'estadístic de Wilks i un nombre màxim de variables predictores seleccionades igual a vint, per tal de disposar d'una relació com a mínim igual a tres entre el nombre d'objectes i el nombre de variables).

La figura 10 mostra el resultat representat a l'espai de les dues variables canòniques de l'anàlisi discriminant lineal. La separació de les tres classes és perfecta. D'altra banda, la validació realitzada amb l'estratègia de validació «total» o, fins i tot, amb altres procediments de validació més elaborats indica una capacitat de predicció del 100 %.

Simplement un detall: el nom del conjunt de dades FAN és una abreviatura de FANTASY. Les dades són dades artificials obtingudes d'una distribució normal  $N(0; 1)$ , la mateixa per a les tres classes.

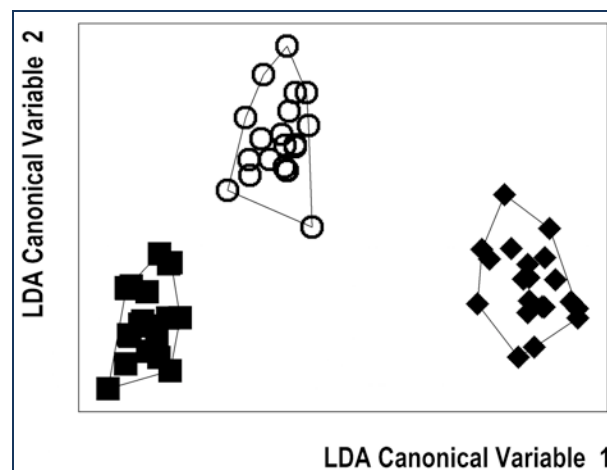


Figura 10. Conjunt de dades FAN: Representació en l'espai de les variables canòniques LDA.

### Validació

Com és possible que s'hagi obtingut aquest resultat?

En l'estratègia habitual de selecció de variables mitjançant SLDA s'utilitzen tots els objectes a la fase de selecció. La validació es realitza en una etapa posterior per validar el model amb les variables seleccionades. Aquest esquema es representa a la figura 11a.

El fet de realitzar una validació completa significa que la selecció de variables s'aplica també als grups definits en validació creuada, tal com mostra l'esquema representat a la figura 11b.

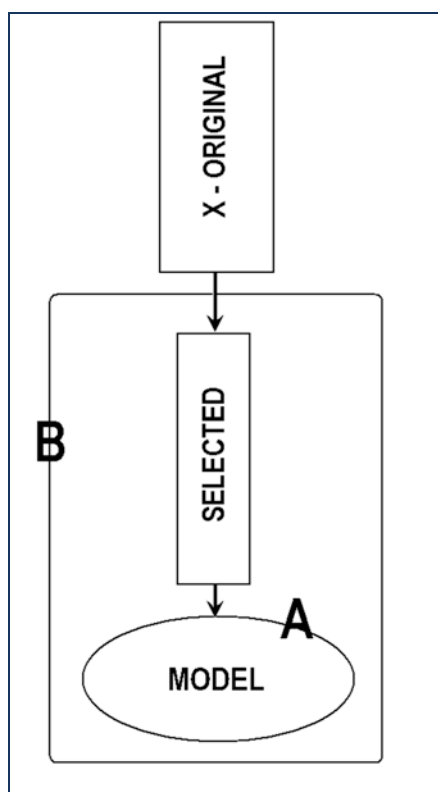


Figura 11. a) Validació habitual; b) Validació completa.

Si fem servir la validació completa sobre el conjunt de dades FAN, s'obtenen els valors de la capacitat de predicció que mostra la figura 12. S'evidencia que la capacitat de predicció és lluny del 100 % i que, quan augmenta el nombre de variables seleccionades, baixa exactament fins a un 33 % el que es pot esperar de la classificació aleatòria en el cas de tres categories.

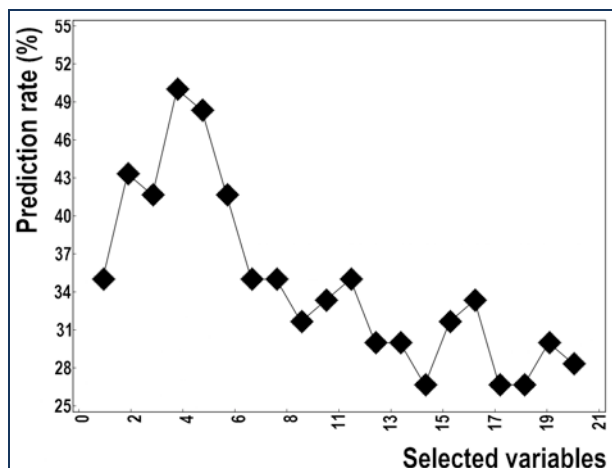


Figura 12. A: Conjunt de dades FAN: Capacitat de predicció avaluada mitjançant validació completa en funció del nombre de variables seleccionades per SLDA.

## Conclusions

La imatge generalitzada de la quimiometria no és la que ofereixen les poques persones que fan servir acuradament les eines d'aquesta disciplina amb una plena comprensió de la teoria en la qual es fonamenten, les seves limitacions i el compliment o no de les hipòtesis subjacents.

La imatge real és la que resulta de la utilització de la quimiometria en milions de determinacions rutinàries, els *best-sellers* de la quimiometria.

En aquest cas, les eines quimiomètriques poden ser aplicades sense un coneixement suficient i, molt sovint, amb una sobrevaloració de les seves possibilitats o altres circumstàncies que poden ser molt crítiques.

Creiem que els veritables quimiòmetres han de fer un esforç per apropiarse als problemes reals, potser de baix nivell, per mostrar la manera correcta de treballar, tant amb el seu exemple com amb la seva pedagogia.

## Agraïments

Volem agrair sincerament als organitzadors d'aquest IX Memorial Enric Casassas la invitació a presentar aquesta conferència.

Hom recorda una persona per la seva qualitat humana, per la seva simpatia o pel temps passat junts.

Per tal de recordar un científic, necessàriament cal alguna cosa més.

Un científic eminent, un excel·lent professor que mereix ser recordat pels seus col·legues i estudiants, ha ser capaç de mirar al futur per adonar-se de quina serà l'evolució de la seva ciència. No és gens senzill, però sí molt poc freqüent.

L'Enric Casassas tenia aquest do.

## Referencias

1. Stevenson, D. F.; Stross, F. H.; Heizer, R. F. *Archaeometry*, 1971, 13, 17.
2. Harper, A. M.; Duewer, D. L.; Kowalski, B. R.; Fashing, J. L. A. Kowalski, B. R. *Chemometrics: Theory and applications*. Washington: American Chemical Society, 1977, p. 14-52. (ACS Symposium Series; 52).
3. Saxberg, B. E. H.; Duewer, D. L.; Booker, J. L.; Kowalski, B. R. *Anal. Chim. Acta* 1978, 103, 201.
4. Casale, M.; Casolino, C.; Oliveri, P.; Forina, M. *Food Chemistry* 2010, 118, 163.
5. Wold, S. *Chemometrics and Intelligent Laboratory Systems* 1995, 30, 115.



M. Forina

M. Forina. Nascut a Torino (Itàlia), es va llicenciar en Química a la Universitat "La Sapienza" de Roma. Catedràtic de Química Analítica a l'Istituto di Analisi e Tecnologia Farmaceutiche ed Alimentari a la Università di Genova. Un dels impulsors de la Quimiometria a nivell internacional, ha tingut especial influència en el naixement i posterior desenvolupament d'aquesta ciència a Catalunya, i també a la resta de l'estat espanyol, mantenint contactes constants amb tots els grups de treball. Juntament amb el Prof. Enric Casassas va crear (1987) el Colloquium Chimometricum Mediterraneum, punt de trobada bianual dels investigadors en Quimiometria de parla llatina. És autor de més d'un centenar de treballs, de diversos llibres i del conegut programari "Parvus", tot un clàssic dins l'àmbit quimiomètric. Ha sigut President de la International Chemometrics Society i membre de l'equip editorial de "Chemometrics and Intelligent Laboratory Systems" i del "Journal of Food Technology".