

UNIVERSITAT AUTÓNOMA DE BARCELONA

**APLICACIÓ I DESENVOLUPAMENT
DE TÈCNIQUES QUIMIOMÈTRIQUES
A L'ANÀLISI AGROALIMENTÀRIA**

Ivan Montoliu i Roura

Grup de Quimiometria Aplicada
Unitat de Química Analítica
Departament de Química
Edifici C
08193 Bellaterra (Barcelona) Spain
Tel: (3) 581 17 12
Fax: (3) 581 23 79
E-mail: gr.quimiometria@uab.es



La **Dra. Hortensia Iturriaga i Martínez**, catedràtica de Química Analítica de la Facultat de Ciències de la Universitat Autònoma de Barcelona, i

La **Dra. Enriqueta Bertran i Gimferrer**, doctora en Química Analítica per la Universitat Autònoma de Barcelona,

CERTIFIQUEN: Que el present treball d'investigació titulat '**Aplicació i Desenvolupament de Tècniques Quimiomètriques a l'Anàlisi Agroalimentària**', que constitueix la memòria presentada per l'**Ivan Montoliu i Roura** per a aspirar al grau de doctor en ciències químiques, ha estat realitzat sota la nostra direcció i recull, des del nostre punt de vista, les condicions que s'exigeixen a aquest tipus de treballs.

I perquè en quedi constància, signen el present certificat a 17 de Setembre del 2001

Dra. H. Iturriaga

Dra. E. Bertran

És tradicional que en treballs d'aquesta mena es faci un seguit d'agraïments més o menys extens que abarca grups de persones, animals i fins i tot objectes que han estat propers a qui ha redactat el treball. Suposo que jo no puc escapar a aquest tòpic. Però a la meua manera, vull intentar-ho.

El gran avantatge d'escriure una pàgina d'agraïments és que s'acostuma a fer al final de la redacció del treball. La satisfacció pel treball gairebé enllestit fa que tendim a simplificar-ho tot una mica, bàsicament per raons de temps. El principal inconvenient és que això ens pot fer ser una mica injustos, tant pel fet de deixar d'esmentar algú com pel fet de fer-ho. No pretenc fer un llistat exhaustiu dels uns i dels altres. Crec que tots plegats saben on encabir s'hi i jo no soc pas ningú per a fer distincions. Es fan soles. De tota manera estic profundament agraït als uns i als altres. Als uns perquè han fet que l'experiència personal de la tesi hagi estat engrescadora i moltes estones divertida. Als altres perquè han fet que hagi hagut d'aprendre a fer les coses al meu aire, cosa que tot i que és dura, a la llarga resulta molt positiva.

A tots ells, una abraçada ben forta.

Als pares, per haver confiat en mi quan les coses no rutllaven gaire.
I a l'Eva, perquè aquesta també ha estat la seva tesi.

ÍNDEX

Objectius de la tesi	1
Apartat A. Quimiometria	
A.1. Introducció	3
A.2. Correccions del senyal	5
A.2.1. Correccions Individuals	5
A.2.2. Correccions de l'escalat	7
A.2.3. Correccions Globals	8
A.3. Tècniques de Reducció de variables	9
A.4. Recursos Quimiomètrics	12
A.4.1. Tècniques Quantitatives	12
A.4.1.1. Regressió Lineal Múltiple (MLR)	14
A.4.1.2. Regressió Lineal Múltiple Inversa (ILS)	15
A.4.1.3. Regressió en l'espai de Components Principals (PCR)	16
A.4.1.4. Regressió Parcial per Mínims Quadrats (PLSR)	17
A.4.1.5. Avaluació de resultats. Determinació de les condicions finals de l'equació de calibració	20
A.4.2. Tècniques Qualitatives	23
A.4.2.1. Tècniques no supervisades	24
A.4.2.2. Classificadors (ANN)	27
A.4.2.3. Tècniques supervisades	29
A.4.2.3.1. Xarxes Neuronals Artificials (ANN)	32
A.4.2.3.2. Regressió Logística	41

Apartat B. Espectroscopia Infrarroja

B.1.	Espectroscopia de Transformada de Fourier (FT-IR)	45
B.1.1.	Absorció de la radiació IR	45
B.1.2.	Instrumentació	50
B.1.3.	Mesures de reflectància Total Atenuada	54
B.1.4.	Aplicacions a l'anàlisi agroalimentària	56
B.2.	Espectroscopia d'infrarroig Proper (NIR)	
B.2.1.	Introducció	59
B.2.2.	Absorció de la radiació NIR	60
B.2.3.	Instrumentació	61
B.2.3.1.	Fonts emissores de radiació policromàtica	62
B.2.3.2.	Sistemes monocromadors	63
B.2.3.3.	Detectors	64
B.2.3.4.	Presentació de la mostra	65
B.2.3.5.	Aplicacions a l'anàlisi alimentària	66

Apartat C. L'oli d'oliva

C.1.	Introducció	71
C.2.	L'olivera a Catalunya	73
C.2.1	L'oliva	75
C.3.	Elaboració de l'oli d'oliva verge. Processos d'elaboració i tractament	77
C.3.1	Recollida	77
C.3.2.	Emmagatzematge	77
C.3.3.	Processat	78
C.4.	Composició de l'oli d'oliva	82
C.4.1.	Components majoritaris. Àcids Grassos i Triglicèrids	82
C.4.2.	Components minoritaris. Matèria Insaponificable	86
C.5.	Qualitat i Legislació	93
C.5.1	Classificació dels Olis d'oliva	93
C.5.2.	Alteracions de la Qualitat. Adulteracions de l'oli d'oliva	97
C.5.3.	Interpretació dels paràmetres analítics	98
C.6.	Denominacions d'origen	100
C.6.1.	DOP Siurana	101
C.6.2.	DOP Garrigues	102

Apartat I. Anàlisi Quantitativa. Resolució de problemes quantitativs. Efecte de la correcció ortogonal del senyal a la determinació de compostos per espectrofotometria NIR

		105
	Metodologia experimental	107

Presentació i discussió dels resultats	111
Conclusions	117
Apartat II. Pretractaments espectrals. Estudi de la correcció ortogonal del senyal (OSC). Control del sobreajust	119
Metodologia experimental	122
Aplicació	127
Conclusions	134
Apartat III. Anàlisi Quantitativa. Resolució de problemes quantitius. Determinació de l'índex d'acidesa en olis d'oliva per espectrofotometria FTIR-ATR	137
Metodologia experimental	139
Presentació i discussió dels resultats	142
Conclusions	149
Apartat IV. Anàlisi Qualitativa. Resolució de problemes qualitius. Estudi de les denominacions d'origen d'olis d'oliva catalans	151
Metodologia experimental	154
Presentació i discussió dels resultats	157
Conclusions	163
Conclusions generals	165

OBJECTIUS DE LA TESI

Quan es planteja un treball de recerca, que posteriorment pot esdevenir el contingut d'una tesi de caràcter aplicat, apareixen tota una sèrie de condicionants. Tant l'interès general del camp d'aplicació, com les necessitats del moment, com l'entorn en el que es portarà a terme són factors que exerciran una influència molt important al llarg del seu desenvolupament. Aquests aspectes de l'entorn són els que, habitualment, modifiquen les trajectòries establertes inicialment per a un determinat treball i fan que aquest acabi d'una manera relativament diferent a la que es va començar. Així es pot dir que el procés de selecció d'objectius al llarg d'una tesi és un procés dinàmic, que s'autoajusta en funció de les noves perspectives que es presenten a l'anar aprofundint en una determinada direcció.

Aquesta tesi n'és un exemple més. Coneguda la importància del sector de l'oli d'oliva al nostre país, així com la del conjunt de normes i regulacions que afecten la seva comercialització, es va pensar que seria una bona iniciativa trobar mètodes d'anàlisi alternatius. Serien mètodes que, sense cap intenció d'arraconar els mètodes oficials, haurien de permetre fer les determinacions de forma més ràpida i amb un índex de fiabilitat alt. Així, donat el rerafons històric del grup d'investigació, per a desenvolupar els nous mètodes es va recórrer a les tècniques espectrofotomètriques d'infraroig mitjà i proper, utilitzades en conjunció amb la quimiometria, com a tècnica d'anàlisi i tractament de les dades obtingudes.

Com a **primer objectiu**, es va mirar d'abordar la determinació simultània de paràmetres de qualitat en olis d'oliva de diferents categories fent servir espectrofotometria FTIR-ATR. El primer dels escollits va ser l'índex d'acidesa, tant pel farragós que resulta la seva determinació mitjançant el mètode de referència, com perquè és un paràmetre de qualitat que permetia aproximar-se a un camp, el dels olis d'oliva, on l'experiència prèvia era pràcticament inexistent.

Els estudis fets en aquest primer treball, i el millor coneixement de les mostres, van mostrar diferents aspectes en l'estructura de les dades. Aquests aspectes permetien pensar que, mitjançant tècniques quimiomètriques adequades, es podien arribar a fer classificacions dels olis verges tenint en compte el seu origen. El fet de no existir una tècnica analítica que permetés assignar de forma clara una mostra a una certa denominació d'origen, va motivar que s'aprofundís en aquesta idea. D'aquesta manera es va fixar un **segon objectiu**, la classificació d'olis d'oliva verge catalans en funció

de les seves denominacions d'origen i fent servir l'espectrofotometria NIR com a tècnica instrumental.

A conseqüència de l'assoliment d'aquest segon objectiu, com a efecte afegit, es va aprofundir en el coneixement de la varietat i abast de les diferents tècniques quimiomètriques fins a tal punt que aquest fet va esdevenir un punt d'inflexió en el desenvolupament de la tesi.

El **tercer objectiu** de la tesi va ser l'inici de la tendència majoritàriament quimiomètrica del treball. Sense perdre de vista quina havia estat la seva motivació inicial, es va realitzar un estudi sobre un nou mètode de pretractament del senyal espectral, la correcció ortogonal del senyal (OSC) que havia aparegut recentment. Es va aplicar sobre un conjunt de dades espectrals d'olis d'oliva verges per a determinar el percentatge de determinats àcids grassos sobre el total i amb la intenció de comprobar si aquest tractament permetia obtenir millors resultats en la determinació de components minoritaris.

Va ser a partir de l'aprofundiment en el coneixement de la correcció ortogonal que es van detectar una sèrie de factors que condicionaven la seva aplicació. Aquest fet va originar un **quart objectiu**, el qual va ser el de la proposta i comprovació de diferents estratègies de validació per aconseguir una aplicació menys iterativa del pretractament. Arribats a aquest punt, la trajectòria de la tesi havia experimentat el final del seu procés de migració des del camp de les aplicacions quimiomètriques fins el dels desenvolupaments semi-teòrics.

La descripció cronològica d'aquest procés, en el que es veu quina ha estat l'evolució del plantejament del treball no és, potser, la millor manera d'organitzar la memòria. És per això que, tenint en compte el contingut de cadascun dels treballs, se n'ha alterat l'ordre. Crec que amb aquesta alteració es guanya en estructuració del text i s'afavoreix una major claredat en l'exposició.

APARTAT A

QUIMIOMETRIA

A.1 Introducció

Fins a final dels anys 60, les dades analítiques s'obtenien majoritàriament a partir de processos d'anàlisi llargs i costosos. Amb l'aparició dels mètodes computeritzats d'anàlisi, que suposaven la connexió de la neixent microinformàtica amb la instrumentació analítica, l'obtenció de dades analítiques de qualitat es va facilitar en gran manera. Des d'aquest moment, els químics analítics varen disposar d'un gran volum d'informació amb el que abordar els diferents problemes.

De forma gairebé immediata, es va veure que era necessària la recerca de mètodes matemàtics que permetéssin l'extracció de la informació rellevant de les dades disponibles. Així es va recórrer a tècniques matemàtiques i estadístiques aplicades a d'altres camps, especialment a les ciències socials. Aviat es va veure que aquestes tècniques permetien extreure la informació que es buscava i, a més, proporcionaven informació no esperada però força útil, com l'existència d'una certa estructura de les dades.

El concepte de *Quimiometria* va ser introduït el 1972 pel suec Swante Wold i pel nord americà Bruce R. Kowalski. La primera descripció d'aquesta nova disciplina es va generar el 1974, època en la que es va fundar la International Chemical Society (ICS). En els anys següents varen anar apareixent publicacions periòdiques dedicades al tema, com el *Journal of Chemometrics* (Wiley) o el *Chemometrics and Intelligent Laboratory Systems* (Elsevier). Actualment es defineix la quimiometria com *aquella*

disciplina de la química que utilitza mètodes matemàtics i estadístics: per a dissenyar o seleccionar mètodes òptims de mesura i experimentació i per a proporcionar el màxim d'informació química a l'analitzar dades químiques¹.

A part del disseny d'experiències, un bon nombre de les aplicacions típiques de la quimiometria estan orientades al desenvolupament de relacions quantitatives entre estructura i activitat (Quantitative Structure-Activity Relationships, QSAR) dels compostos químics. La resta estan relacionades amb el tractament de la informació generada per la instrumentació moderna per a extreure'n la que és rellevant. Constitueixen l'aplicació de la Quimiometria a la química analítica.

Al començament dels anys 80 es dona la popularització de la microinformàtica i d'aquesta manera es posa a l'abast dels usuaris una capacitat de càlcul elevada a uns costos relativament baixos. Aquest fet permet que els químics analítics comencin a manipular un elevat nombre de dades i que la necessitat de tractar-les i extreure'n la informació rellevant també es generalitzi. Aquests dos factors, combinats, fan que la utilització d'eines quimiomètriques de tractament comenci a ser necessària i que els softwares de control dels instruments les incorporin. Malgrat ser eines amb un rerafons matemàtic complex s'ha simplificat força el seu ús, de manera que poden ser utilitzades per un gran nombre d'analistes.

La simplicitat del seu ús, però, és sovint un dels seus principals inconvenients. El relatiu desconeixement en alguns casos dels principis en els que es basen aquests mètodes fa que moltes vegades s'apliquin de forma massa directa, anant a la recerca directa de resultats, en lloc d'aprofitar la seva capacitat per a destriar la informació rellevant de les dades analítiques i així captar nova informació del sistema.

En aquesta memòria s'ha mirat de fer un recorregut per una bona part de les tècniques quimiomètriques relacionades amb l'extracció i aprofitament d'informació rellevant de dades analítiques instrumentals. Aquestes tècniques s'han aplicat a diferents conjunts de dades espectrofotomètriques corresponents a mostres d'origen natural amb diferents objectius analítics. El gran nombre de dades per una banda, i la complexitat de la mostra per l'altra han fet que el recorregut no hagi estat fàcil. Al marge de l'èxit (relatiu o complet) dels treballs realitzats, el camí ha resultat sempre engrescador i ha permès una profunda presa de contacte amb una bona part de les possibilitats que ofereix aquest camp.

Les diferents tècniques quimiomètriques orientades a l'extracció i aprofitament d'informació rellevant poden classificar-se en dos grans grups: per una banda les *tècniques de pretractament* de les dades i per l'altra les eines d'*anàlisi multivariable*.

A.2 Correccions del senyal

La qualitat de les dades de partida és un factor molt important en l'anàlisi multivariable. Aquesta prové dels diferents orígens que poden tenir. Pot venir condicionada per diferents factors, ja siguin derivats de la pròpia tècnica instrumental o bé del dispositiu de mesura. Aquests condicionants són prou importants com per a que l'anàlisi multivariable del conjunt de dades pugui ser molt més complexa del necessari i, inclús en certs casos, comportar conclusions errònies.

Les dades espectrals en el seu conjunt poden ser considerades com a matrius de i files (on i és el número d'espectres) i j columnes (on j és el número de variables). Aquest tractament matricial de les dades, que és pràcticament una constant al llarg de la majoria de tècniques quimiomètriques d'anàlisi multivariable, facilita la introducció d'una classificació entre els mètodes de correcció del senyal. Així, els diferents tipus de correccions poden dividir-se en tres grans grups: les **correccions individuals**, que tracten el senyal de forma independent; les **correccions de l'escalat**, que tracten els senyals de forma independent tenint en compte la dispersió global del conjunt; i les **correccions globals**, que generen un model matemàtic de correcció a partir d'un conjunt de dades determinat i que ha de ser recalculat cada cop que el conjunt es modifica.

A.2.1 Correccions Individuals

Les correccions individuals s'apliquen a cadascun dels espectres de forma independent. Bàsicament es tracta de procediments que miren d'eliminar efectes aleatoris no desitjats (soroll) o bé de desplaçament de la línia de base (o de pendent) sobre el senyal individual.

Promig de senyals

El promig de senyals permet la minimització d'efectes aleatoris (soroll) superposats sobre el senyal analític. Es fonamenta en promitjar, per a cada variable, el major nombre de registres individuals (replicats) amb l'objectiu de compensar la component aleatòria. Aquest és un mètode efectiu quan el nombre de registres per a una única mostra individual és suficientment elevat. La seva principal limitació prové de la pròpia

instrumentació analítica, de si pot o no generar un nombre alt de replicats en un espai de temps suficientment curt; o de la pròpia estabilitat del sistema que es mesura.

Filtrat digital

Hi ha diferents tipus de filtres, destacant entre ells el Boxcar o de finestra mòbil per la seva simplicitat i el de Savitzky-Golay² per la generalització del seu ús.

El filtre Boxcar defineix una finestra de $n+1$ punts amb dues condicions: el nombre de dades ha de ser senar i el mínim de dades a tractar ha de ser de 3. El filtre promitja tots els punts continguts en la finestra i substitueix el valor central per la mitja obtinguda. Un cop s'ha calculat aquest valor es desplaça un punt per a repetir el procés de filtrat, amb una nova finestra de la mateixa mida i un nou punt central. El principal inconvenient a l'hora d'utilitzar un mètode tan senzill prové de l'alta influència del tamany de la finestra, que fa que només resulti eficaç quan el soroll és d'una freqüència determinada.

Per a calcular el valor de senyal corregit y_k^* a partir de dades originals y_k s'utilitza la següent relació

$$y_k^* = \frac{1}{2m+1} \sum_{j=-m}^{j=m} y_{k+j}$$

On $2m+1$ és el tamany de la finestra i m és el nombre de punts a banda i banda del punt central.

A l'inrevés que en el cas del filtre anterior, el filtre de Savitzky-Golay pondera de forma diferent els diferents punts de la finestra, mitjançant un ajust polinomial dels punts que hi són continguts. Les majors diferències entre ambdós filtres s'obtenen en aquells casos en els que els tamany de finestra són relativament grans. Aquest procés d'ajust és relativament lent, pel que els autors van reduir la tècnica a un filtre Boxcar ponderat que fa servir una sèrie de números sencers que es troben tabulats en diferents articles i monografies.

Aleshores, cada punt del vector filtrat es calcula com

$$y_k^* = \frac{1}{\text{NORM}} \sum_{j=-m}^{j=m} c_j y_{k+j}$$

On NORM és el factor de normalització que es calcula a partir de la suma dels coeficients c_j .

Derivació del Senyal

La derivació permet l'emfatització de les diferències entre senyals, ja sia a través del desolapament de senyals com per l'eliminació d'efectes constants o per l'eliminació simultània d'efectes constants i de pendent. L'eliminació d'efectes constants es dóna a l'aplicar la primera derivada. L'eliminació dels efectes constants i de pendent s'aconsegueix a través de la derivada segona.

La seva utilització, però, no està exempta de certs riscos. Així, el progressiu desolapament del senyal comporta una reducció de la relació senyal-soroll i una possible emfatització d'aquest darrer.

A.2.2 Correccions de l'Escalat

L'objectiu de les correccions de l'escalat és fixar la importància relativa de cadascuna de les variables. D'aquesta manera, pot resultar interessant en alguns casos donar la mateixa importància a totes les variables, referir-les a un mateix origen o canviar les longituds dels eixos de coordenades.

El *Centrat* de les dades canvia l'origen espacial de les dades i passa a referir-les totes a un mateix centre. Quan s'aplica a dades espectrals, el més freqüent és fer-ho per a cada columna (variable), mitjançant l'aplicació de l'expressió per a cadascuna de les j variables dels N objectes

$$x_{ij}^* = x_{ij} - \frac{\sum_{i=1}^N x_{ij}}{N}$$

Quan les variables presenten una diferent dispersió entre elles, pot unificar-se aquesta variància fent servir l'*Estandarització* per columnes, amb el que la desviació estàndar per a totes elles és 1. S'aconsegueix a través de l'expressió

$$x_{ij}^* = \frac{x_{ij}}{\sqrt{S_i}}$$

L'*Autoescalat* combina els efectes del centrat i l'escalat. Per una banda canvia l'origen de coordenades de les dades i per l'altra n'homogenitza la variància. A efectes pràctics

canvia la longitud dels eixos de coordenades i és especialment útil en aquells casos en els que les variables s'expressen en diferents unitats o bé en aquells en els que les variances de les diferents variables són molt diferents.

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_i}{s_i}$$

En el cas particular de l'espectrofotometria, no es pot considerar que sigui el pretractament a utilitzar per defecte, ja que el fet d'homogeneitzar la variància comporta assignar una excessiva rellevància al soroll instrumental.

Un cas particular d'Autoescalat és el que es calcula per files de dades espectrals. Rep el nom de *Standard Normal Variate*³ (SNV) i permet la reducció dels efectes de dispersió de la línia de base originats per diferents factors. La seva aplicació correcta demana prestar especial atenció a l'interval de dades a les que s'aplica, amb la intenció de disposar d'una correcció que no pugui eliminar informació rellevant. Recentment ha aparegut una variant d'aquest pretractament, inspirada en mètodes robustos, anomenada *Robust Normal Variate* (RNV)⁴. La seva descripció, així com alguns exemples del seu funcionament, són presents a la bibliografia.

A.2.3. Correccions Globals

Les correccions globals del senyal s'apliquen de forma conjunta a totes les dades espectrals. Els paràmetres o equacions de correcció es calculen a partir de tots els espectres continguts en el conjunt de dades i d'aquesta manera, la correcció depèn de la variabilitat total. Qualsevol canvi del conjunt fa variar en major o menor mesura els paràmetres de la correcció i d'aquesta manera els seus efectes poden diferir lleugerament, en funció de la composició d'aquest.

La correcció *Multiplicative Scatter Correction* (MSC)⁵ permet corregir els efectes dels diferents tamanys de partícula sobre els espectres de mostres sòlides. Aquests efectes acostumen a ser additius i multiplicatius i es manifesten com a desviacions constants de la línia de base (efecte additiu) i desviacions que varien en funció de les diferents longituds d'ona (efecte multiplicatiu). MSC corregeix aquests efectes a partir del càlcul d'una regressió lineal entre els valors d'absorbància d'un interval de longituds d'ona on no absorbeix l'analit i els valors d'absorbància en la mateixa zona d'un espectre de referència (habitualment l'espectre promig del conjunt). Tot i que va ser un mètode desenvolupat inicialment per a tractar efectes multiplicatius, a la llarga s'ha utilitzat per a corregir altres efectes similars, com els que donen les variacions de camí òptic, etc.

L'ajust DT^3 de la línia de base (*Detrending*) és una altra de les vies per a l'eliminació dels efectes de desviació de la línia de base a longituds d'ona superiors. Per a aquesta correcció es genera una funció lineal o quadràtica que s'aplica posteriorment als espectres originals.

Un dels darrers mètodes de correcció que ha aparegut i que es troba en fase de desenvolupament és l'anomenada Correcció Ortogonal del Senyal (OSC, *Orthogonal Signal Correction*)^{6,7,8}. Aquesta correcció pretén eliminar de la informació espectral la major part de la variabilitat no lligada directament a la concentració, o el que és el mateix, la component de les dades espectrals que n'és ortogonal. Per a assolir aquest objectiu, es modela la informació ortogonal mitjançant una tècnica de calibració multivariable (PLS) i es retira aquesta informació de les dades originals.

En un apartat d'aquesta tesi es fa un estudi més acurat i en profunditat d'aquest nou algoritme de pretractament, de les seves virtuts i inconvenients així com de les possibles estratègies a l'hora de la seva utilització.

A.3. Tècniques de Reducció de Variables

Els senyals espectrals poden ser considerats com a dades puntuals situades en un espai de j dimensions, on j és el nombre de variables (longituds d'ona de l'espectre). La distribució dels diferents espectres, els uns envers els altres, dins d'aquest espai multidimensional ens pot donar una idea de la seva similitud de forma global. Com que no totes les dades col·laboren d'igual manera en l'establiment d'aquestes similituds, és fa necessari disposar de sistemes de simplificació del senyal. Aquests han de permetre la representació de les dades espectrals en sistemes de dimensionalitat més petita, conservant el màxim d'informació rellevant.

Una d'aquestes tècniques de reducció és l'Anàlisi en components principals o PCA^{9,10} (*Principal Component Analysis*). El seu objectiu és trobar les direccions, ortogonals entre sí, on es dona la màxima variabilitat del conjunt d'espectres. D'aquesta manera, el PCA és un canvi de sistema de coordenades, on s'agafen com a nous eixos aquells en els que la variabilitat és màxima. Aquests eixos reben el nom de Components Principals (PC).

Matemàticament, l'anàlisi en components principals es basa en la descomposició en vectors propis (*eigenvectors*) de la matriu de covariància o correlació de la matriu de dades espectrals. Si aquesta matriu \mathbf{X} conté i files i j columnes i considerem els

espectres distribuïts en files, la matriu de covariància per a una matriu \mathbf{X} de dades centrades (per columnes) es defineix com

$$\text{cov}(\mathbf{X}) = \frac{\mathbf{X}^T \mathbf{X}}{l - 1}$$

PCA fa la projecció de la matriu \mathbf{X} en un subespai de d dimensions mitjançant una matriu de projecció \mathbf{P}^T . Les files de la matriu \mathbf{P} (\mathbf{p}_i) són els anomenats *eigenvectors* o vectors de *Loading* i proporcionen els cosinus directors dels nous eixos respecte del sistema d'eixos inicial^A. Les columnes de la matriu \mathbf{T} són els anomenats vectors de *Score* i contenen les coordenades dels diferents espectres en el nou subespai. La diferència entre les dades contingudes a \mathbf{X} i el resultat de reconstruir les dades proporciona una matriu de residuals \mathbf{E} . La dimensionalitat màxima de les columnes de la matriu de *scores* ve donada per la dimensió més petita de la matriu \mathbf{X} .

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$

Els vectors de *Score* formen un conjunt de vectors ortogonals entre sí, complint-se que $\mathbf{t}_i^T \mathbf{t}_j = 0$ per a $i \neq j$. Els vectors de *Loading* són ortonormals entre ells ($\mathbf{p}_i^T \mathbf{p}_j = 0$ per a $i \neq j$; $\mathbf{p}_i^T \mathbf{p}_i = 1$ per a $i=j$). Aquests vectors són vectors propis (*eigenvectors*) de la matriu de covariància de \mathbf{X}

$$\text{cov}(\mathbf{X})\mathbf{p}_i = \lambda_i \mathbf{p}_i$$

Els parells de vectors $\mathbf{t}_i \mathbf{p}_i$ s'ordenen en funció del valor de λ_i . Aquest és el valor propi (*eigenvalue*) associat a cadascun dels vectors propis \mathbf{p}_i i resulta un descriptor de la quantitat d'informació (variància) descrita pel parell de vectors. Així, a PCA, el primer parell de vectors, $\mathbf{t}_1 \mathbf{p}_1$ és el que explica la major quantitat d'informació de tots els que integren la descomposició. El següent parell de vectors $\mathbf{t}_2 \mathbf{p}_2$ explica part de la informació restant, no descrita pel primer parell, i així successivament.

Hi ha diferents mètodes de càlcul que permeten fer la descomposició. El més conegut és l'anomenat NIPALS (*Non-Iterative Partial Least Squares*)¹¹. Fins fa poc ha estat un dels més utilitzats, ja que el càlcul dels components principals no és iteratiu i permet calcular-ne un nombre determinat. El fet de poder calcular el nombre de components principals desitjat, estalviant estona de càlcul, ha estat especialment important per a la seva elecció com a mètode per defecte.

^A La traducció del terme *Score* resulta relativament difícil (*puntuació* no sembla un terme adequat), en canvi el terme *Loading* sembla més fàcilment traduïble per *cosinus directors*. Malgrat els esforços per a dotar aquests dos termes d'una traducció al català, encara no ha estat possible trobar termes adequats. És per això que es manté l'ús de la forma anglesa per a la seva definició

Actualment, amb els avenços en la velocitat dels equips informàtics, el càlcul d'un número limitat de components principals ja no és tan necessari. Així han agafat importància d'altres mètodes, com l'anomenada Descomposició en Valors Singulars (SVD, *Single Value Descomposition*)¹². Aquest mètode proporciona la descomposició de la matriu de dades \mathbf{X} en tres matrius \mathbf{U} , \mathbf{W} i \mathbf{V}

$$\mathbf{X} = \mathbf{U}\mathbf{W}\mathbf{V}^T$$

La matriu \mathbf{U} proporciona els mateixos valors que la matriu de *scores* \mathbf{T} , però normalitzats a longitud 1; \mathbf{W} és la matriu diagonal que conté les arrels quadrades dels valors propis quan la matriu \mathbf{X} és simètrica; i \mathbf{V}^T és idèntica a \mathbf{P}^T , matriu de *loadings*.

Independentment del mètode emprat per a calcular els components principals, un dels aspectes més delicats i fonamentals en PCA és la determinació del número òptim de components que descriuen la matriu \mathbf{X} . Per a decidir-ne el número, hom disposa de diferents mètodes heurístics i estadístics, com el percentatge de variància explicada, el criteri de valor propi 1, l'anomenat *scree test* i la validació creuada o *cross validation*¹³.

- El **percentatge de variància explicada** pot fer-se servir si es disposa de suficient experiència en la manipulació de conjunts de dades simil·lars. La fracció de variància acumulada (s_e^2) es calcula a partir de la relació de la suma de d valors propis respecte la suma de tots els valors propis calculats (p). Es fa servir la següent expressió

$$s_e^2 = \frac{\sum_{i=1}^d}{\sum_{i=1}^p}$$

Quan s'expressa com a percentatge, s_e^2 apareix multiplicada per 100. En la utilització d'aquest mètode s'especifica un percentatge de variància fix, que es coneix a partir de l'experiència prèvia sobre conjunts de dades simil·lars.

- El **criteri de valor propi 1** es basa en el fet que el valor promig dels valors propis de dades autoescalades és 1. Això és cert perquè per a dades estandaritzades, la suma de tots els valors propis de la matriu de correlacions és igual al nombre total de components. En l'aplicació d'aquest criteri només es consideren importants aquells components que presenten valors propis més grans que 1.

- El **scree test** és un dels més àmpliament utilitzats. És la representació dels valors propis (λ_i) envers el número de components principals i la representació de la variància residual (o explicada) en funció del número de components. A partir d'aquestes representacions s'acostuma a triar el nombre de components principals més baix on la variància (o els valors propis) presenten una variació significativa. La selecció d'aquest valor acostuma a fer-se de forma gràfica. En els casos en els que la disminució de la variància explicada és pràcticament asimptòtica pot acabar sent poc objectiva.
- **La validació creuada, o cross validation**¹⁴, extreu un espectre (o espectres) del conjunt de dades i calcula **T** i **P** amb la resta de dades. Un cop calculades **T** i **P**, es prediuen els espectres extrets i se'n calcula l'error residual fent servir cada component principal. El nombre de components principals significatius es selecciona a partir de l'error residual mínim.

A.4 Recursos Quimiomètrics

En línies generals, els problemes analítics poden ser abordats de forma qualitativa o de forma quantitativa. La quimiometria ha avançat en la mateixa direcció, proporcionant recursos tant qualitatiu com quantitatiu pel tractament de les dades. D'aquesta manera s'han desenvolupat tot un seguit de tècniques matemàtiques de tractament que permeten aprofitar el major nombre de dades generades per la instrumentació actual.

Aquestes dades acostumen a manipular-se en forma matricial, amb el que es fa necessari fer servir una certa organització i nomenclatura. Per regla general, les diferents variables predictorres corresponents a una mateixa mostra són contingudes en una fila, rebent el nom d'*objectes*. Cadascuna de les columnes fa referència a una *variable* diferent (ja sia *predictora* o *resposta*).

A.4.1. Tècniques Quantitatives

En el procés de calibració s'estableix una relació matemàtica entre la resposta instrumental i una propietat determinada de la mostra que, per regla general en determinacions quantitatives, acostuma a ser la concentració. Aquesta relació matemàtica entre concentració i resposta rep el nom de **model** o **equació** de calibració. La seva representació gràfica rep el nom de **corba** de calibració.

El procés de calibració presenta dues vessants: l'**ajust** i la **capacitat predictiva**. Ambdues vessants presenten efectes contraposats. Mentre l'ajust reflecteix el modelat de les dades i de la seva estructura, la capacitat predictiva fa èmfasi en l'aspecte pràctic del procés de calibració, que no és un altre que el de donar prediccions fiables de les concentracions d'analit en mostres desconegudes. Tot i que no sempre és així, les dues vessants poden donar efectes contraposats. El buscar la millora en l'ajust de les dades comporta moltes vegades l'augment de la complexitat de les equacions de calibració, i aquest fet acostuma a derivar en capacitats predictives poc estables.

Els mètodes de calibració poden classificar-se de diferents maneres, depenent del criteri que es faci servir¹⁵. Els més habituals es poden resumir a la taula següent

Criteri	Mètode de Calibració
Dependent del nombre de variables	Univariable Multivariable
Dependent del tipus de funció matemàtica	Lineal No Lineal
Dependent de com s'obtenen els paràmetres de calibració	Directa Indirecta
Dependent de quin és la variable independent	Clàssica Inversa

Taula A.1. Criteris de Classificació de les tècniques de calibració

La calibració **Univariable** és la més habitual. Estableix la relació matemàtica entre la concentració i una única variable predictor. Quan les variables predictor són més d'una, s'està treballant amb calibració **Multivariable**. Les calibracions **Lineals** són aquelles que relacionen la concentració amb les variables predictor mitjançant equacions lineals o bé amb funcions polinòmiques lineals en els coeficients. Si les funcions matemàtiques no són d'aquest tipus es tracta de calibracions **No Lineals**. Quan els paràmetres de quantificació es coneixen a partir del senyal de l'analit individual, la calibració és **Directa**. Quan els paràmetres es coneixen a partir d'un senyal analític donat per mesclades de components, la calibració és **Indirecta**. La calibració pot considerar-se com a **Clàssica** quan la variable independent és la concentració de l'analit. En cas contrari, estem parlant de calibració **Inversa**.

Independentment del tipus utilitzat, el procés de calibració pot desglosar-se en tres etapes

- Una etapa de **Calibració** pròpiament dita, on a partir d'un conjunt de dades representatiu de la població de mostres que són esperables, es calcula l'equació que millor relaciona el senyal instrumental amb la magnitud a determinar.
- Una etapa de **Validació**, on s'aplica l'equació de calibració obtinguda a l'apartat anterior a un conjunt de dades representatiu amb l'objectiu de conèixer-ne la capacitat predictiva. Els valors de predicció obtinguts es comparen amb els valors de referència, que són coneguts.
- L'etapa final de **Predicció** és l'aplicació pràctica de l'equació de calibració ja validada a un conjunt de mostres problema, amb valors de concentració desconeguts.

A.4.1.1 Regressió Lineal Múltiple (MLR)

La regressió lineal múltiple¹⁶ (*Multiple Linear Regression*, MLR) permet la construcció de calibracions multivariables tant de forma directa com inversa. En espectroscòpia, aquesta tècnica assumeix que els espectres mesurats són la suma dels espectres dels analits purs ponderats per la seva concentració. Matricialment, aquesta relació pot expressar-se com

$$\mathbf{x} = \mathbf{p}\mathbf{S}$$

On \mathbf{x} és el vector de resposta mesurat, \mathbf{S} és la matriu de resposta dels analits purs i \mathbf{p} és el vector que conté els pesos (concentracions) dels analits. Com és de suposar, \mathbf{p} és el vector que interessa determinar, ja que \mathbf{x} és conegut. Això pot fer-se a través de la pseudoinversa de \mathbf{S}

$$\mathbf{p} = \mathbf{x}\mathbf{S}^T(\mathbf{S}\mathbf{S}^T)^{-1}$$

El principal inconvenient de MLR és que és necessari conèixer de forma acurada la matriu \mathbf{S} . Conegudes les concentracions \mathbf{C} de totes les espècies presents i el vector de resposta, els espectres dels analits purs poden ser estimats a partir de

$$\mathbf{S}_{\text{est}} = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{X}$$

Per a poder calcular \mathbf{S}_{est} s'han de complir dues condicions. Han de ser conegudes les concentracions de totes les espècies que contribueixen al senyal espectral i, a més, han de ser linealment independents. Per altra banda, les respostes de cadascun dels components purs han de ser també linealment independents, ja

que si no, no és possible calcular \mathbf{SS}^T . En els casos en els que la colinealitat és força elevada, el producte de $\mathbf{S}^T(\mathbf{SS}^T)^{-1}$ queda molt influenciat per petites variacions degudes al soroll o d'altres interferències. En aquestes condicions, les estimacions de \mathbf{p} són poc precises.

A.4.1.2 Regressió Lineal Múltiple Inversa (ILS)

En moltes ocasions no és possible conèixer totes les espècies presents en una mescla i que alhora contribueixen al senyal espectral. En aquests casos, l'estimació de la matriu \mathbf{S} no és possible i cal buscar una solució alternativa. Aquesta solució passa per assumir que la concentració és funció de l'absorbància, amb el que es segueix un plantejament invers al de la llei de Lambert-Beer. Matemàticament, la concentració passa a ser la variable dependent, mentre que les variables espectrals passen a ser independents. És el que s'anomena Regressió Lineal Múltiple Inversa, ILS^{17,18}.

Matemàticament estem expressant que \mathbf{y} , el vector que conté els valors de concentració, és funció del vector \mathbf{x} , que conté les dades espectrals.

$$\mathbf{x}\mathbf{b} = \mathbf{y}$$

Per a la determinació de \mathbf{b} es segueix un procediment similar al de MLR, calculant la pseudoinversa de \mathbf{X} . Així

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

La utilització d'ILS presenta l'avantatge de no necessitar conèixer la concentració de totes les espècies presents o bé la seva contribució a l'espectre total de la mescla. Només és necessari conèixer la dels analits d'interès. Cal tenir en compte que els analits en els que no estem interessats han de ser igualment presents a totes les mostres.

El seu principal desavantatge acostuma a provenir de la colinealitat de les dades espectrals. En altres casos perquè la matriu de dades espectrals conté menys files (espectres) que columnes (variables). Aquests dos fets fan que sigui un mètode especialment sensible a selecció de variables i, per tant, que aquesta selecció sigui una etapa especialment important si es volen tenir equacions estables i resultats precisos.

A.4.1.3 Regressió en l'espai de Components Principals (PCR)

Els mètodes de regressió basats en la reducció de variables¹⁹, a l'igual que la ILS, permeten la quantificació dels analítics encara que no es coneixi la totalitat dels components que integren la mescla. A diferència de ILS, on el procés de selecció de variables és essencial, el canvi a un subespai de dimensionalitat més reduïda permet utilitzar la informació continguda en un nombre de variables originals més elevat i, eliminar els efectes de colinealitat entre les variables. La major quantitat d'informació disponible ha fet que aquests mètodes hagin vist incrementat progressivament el seu ús. Són mètodes en els que el procés de regressió s'efectua entre les variables del nou subespai (components principals) i el vector (o matriu) de concentracions.

La regressió en l'espai de components principals, (*Principal Components Regression*, PCR) aprofita la descomposició de la matriu \mathbf{X} en el nou subespai realitzant una variant de ILS de la concentració fent servir com a noves variables la matriu de *scores* \mathbf{T} .

Pel càlcul de PCR es parteix de dues matrius, una de variables resposta \mathbf{Y} , amb una dimensionalitat de m files i n columnes; i d'una matriu \mathbf{X} , amb m files i p columnes. Cadascuna de les m files de la matriu \mathbf{X} conté les p variables predictorres de cadascuna de les mostres. Les m files de la matriu \mathbf{Y} contenen les n variables resposta per a cadascuna de les m files de variables predictorres.

El primer pas cerca obtenir la matriu de *Scores* \mathbf{T} a través d'una descomposició PCA de la matriu \mathbf{X} . D'aquesta forma

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E}$$

Un cop escollit el nombre de components principals òptim (D) per a descriure la matriu \mathbf{X} , es fa servir per a calcular la matriu de *scores*

$$\mathbf{T} = \mathbf{XP}$$

Coneguda la matriu de *scores* (\mathbf{T}), ara es tracta de realitzar una regressió lineal inversa

$$\mathbf{Y} = \mathbf{TB}$$

on \mathbf{B} , la matriu de regressors, que es calcula a través de la pseudoinversa de \mathbf{T}

$$\mathbf{B} = (\mathbf{T}^T \mathbf{T})^{-1} \mathbf{T}^T \mathbf{Y}$$

Un cop realitzada l'etapa de calibració i modelat de la informació, que en aquest cas ve donada per la matriu de regressors, convé descriure com es realitza l'etapa de predicció en noves variables predictores (dades espectrals) \mathbf{X}^* . Primerament convé calcular la seva pròpia matriu de *scores* \mathbf{T}^* . Això és possible gràcies a la matriu de *loadings* \mathbf{P} obtinguda en la descomposició en components principals de l'etapa anterior. Per a la transformació es fan servir els *loadings* corresponents als D components principals seleccionats fent la multiplicació

$$\mathbf{T}^* = \mathbf{X}^* \mathbf{P}$$

Per a la determinació del valor de la variable resposta a partir de les noves variables predictores, es fa servir la matriu de regressors \mathbf{B} fent

$$\mathbf{Y}^* = \mathbf{T}^* \mathbf{B}$$

El principal problema que presenta PCR neix de les pròpies condicions amb les que es calcula PCA. Els components principals calculats mitjançant PCA són aquells que millor descriuen el conjunt de variables predictores (\mathbf{X}), però poden no ser els que millor descriu les variables resposta (\mathbf{Y}).

A.4.1.4 Regressió Parcial per Mínims Quadrats (PLSR)

Per a eliminar la manca d'ajust de les variables resposta deguda a la forma en que es calculen els components principals en PCR, així com per a treballar amb conjunts de dades que presenten una elevada colinealitat, es va desenvolupar un nou mètode de regressió en el que el càlcul dels components es fes tenint en compte les variables predictores. Aquest nou mètode va rebre el nom de *Partial Least Squares Regression*, PLSR²⁰.

Les matrius de variables predictores (\mathbf{X}) i de variables resposta (\mathbf{Y}) es centren (o autoescalen, depenent del cas) i es descomposen de forma simultània en dos conjunts de D components principals, segons

$$\begin{aligned} \mathbf{X} &= \mathbf{T} \mathbf{P}^T + \mathbf{E} \\ \mathbf{Y} &= \mathbf{U} \mathbf{Q}^T + \mathbf{F} \end{aligned}$$

On \mathbf{T} i \mathbf{P} són les matrius de *Scores* i *Loadings* per a les variables predictores i \mathbf{U} i \mathbf{Q} són les seves homòlogues per les variables resposta. \mathbf{E} i \mathbf{F} són les matrius de residuals d'ambdós conjunts de variables. A diferència del cas de PCA, els *Loadings* (o cosinus directors) ja no dirigeixen els nous eixos cap a les direccions

de màxima variabilitat en la matriu \mathbf{X} , sinó que estàn corregits per la variabilitat de la matriu \mathbf{Y} . En el cas de PLSR, aquestes noves direccions reben el nom de variables latents o **components PLS**.

Per a calcular la matriu de regressors \mathbf{B} , que es pot substituir a l'expressió general $\mathbf{Y}=\mathbf{XB}$, calen les matrius \mathbf{P} , \mathbf{Q} i \mathbf{W}

$$\mathbf{B} = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{Q}^T$$

El sentit de la matriu \mathbf{W} (o de pesos) pot entendre's millor a través de la descripció de l'algoritme NIPALS modificat per a PLSR. L'estimació del número de variables latents és iterativa i s'atura en un cert valor frontera que cal especificar. Aquest valor pot triar-se tenint en compte la precisió de la màquina.

Partint de dades centrades (o autoescalades), per a calcular la primera variable latent

1. S'utilitza la primera columna de la matriu de resposta \mathbf{Y} per a calcular-ne el primer vector de score \mathbf{u}_1 .

$$\mathbf{u}_1 = \mathbf{y}_1$$

2. Es calculen els pesos de les \mathbf{X} mitjançant

$$\mathbf{w}_1^T = \frac{\mathbf{u}_1^T \mathbf{X}}{\mathbf{u}_1^T \mathbf{u}_1}$$

3. S'escala el vector dels pesos a longitud 1

$$\mathbf{w}_1^T = \frac{\mathbf{w}_1^T}{(\mathbf{w}_1^T \mathbf{w}_1)^{1/2}}$$

4. A partir d'aquests pesos s'estimen els scores de la matriu \mathbf{X}

$$\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1^T$$

5. Es calculen ara els *loadings* de la matriu \mathbf{Y} .

$$\mathbf{q}_1^T = \frac{\mathbf{t}_1^T \mathbf{Y}}{\mathbf{t}_1^T \mathbf{t}_1}$$

6. I es normalitzen

$$\mathbf{q}_1 = \frac{\mathbf{q}_1}{(\mathbf{q}_1^T \mathbf{q}_1)^{1/2}}$$

7. Coneguts els *loadings* de la matriu de variables resposta, es torna a generar el vector de score \mathbf{u}_1 de la variable resposta \mathbf{y} .

$$\mathbf{u}_1 = \mathbf{y}\mathbf{q}_1$$

En aquest moment es fa la comparació del nou valor calculat pel vector \mathbf{u}_1 amb l'anterior. Si es compleix que $\|\mathbf{u}_{1nou} - \mathbf{u}_{1anterior}\| < \epsilon$ *valor frontera no agafa el valor de convergència especificat, es reinicia el procés des del punt 2. Si el valor de convergència s'assoleix, es passa al punt següent.

8. Es calculen els *loadings* de la matriu \mathbf{X} a partir dels valors de score, \mathbf{t}

$$\mathbf{p}_1^T = \frac{\mathbf{t}_1^T \mathbf{X}}{\mathbf{t}_1^T \mathbf{t}_1}$$

9. Es reescalen els vectors de score i de pesos amb els valors de \mathbf{p}_1^T abans de la seva normalització

$$\mathbf{t}_1 = \mathbf{t}((\mathbf{p}_1^T)^T \mathbf{p}_1^T)^{1/2}; \mathbf{w}_1^T = \mathbf{w}_1^T ((\mathbf{p}_1^T)^T \mathbf{p}_1^T)^{1/2}$$

10. I es normalitza \mathbf{p}_1^T segons

$$\mathbf{p}_1^T = \frac{\mathbf{p}_1^T}{((\mathbf{p}_1^T)^T \mathbf{p}_1^T)^{1/2}}$$

11. Es determina l'anomenada relació interna, un escalar que es calcula mitjançant

$$\mathbf{b}_1 = \frac{\mathbf{u}_1^T \mathbf{t}_1}{\mathbf{t}_1^T \mathbf{t}_1}$$

12. Un cop s'han calculat els scores i els *loadings* per la primera variable latent, es determinen els residuals d'ambdues matrius, \mathbf{X} i \mathbf{Y} .

$$\begin{aligned} \mathbf{E}_1 &= \mathbf{X} - \mathbf{t}_1 \mathbf{p}_1^T \\ \mathbf{F}_1 &= \mathbf{Y} - \mathbf{b}_1 \mathbf{t}_1 \mathbf{q}_1^T \end{aligned}$$

Per a calcular la segona variable latent, es substitueixen les matrius \mathbf{X} i \mathbf{Y} per les matrius de residuals \mathbf{E}_1 i \mathbf{F}_1 i es reinicia el càlcul des del punt 2, incrementant tots els subíndexs en 1. I així successivament per a la resta de variables latents.

Coneguda la matriu de regressors, es poden realitzar prediccions de la propietat calibrada a partir de variables predictores externes. Si suposem aquest conjunt de variables \mathbf{X}^* , convenientment centrat i/o escalat en les mateixes condicions que la matriu de calibració, el primer pas és el càlcul dels *scores* dels diferents objectes de predicció. Si s'han seleccionat D variables latents, des de $h=1$ fins a $h=D$, es calcula

$$\begin{aligned}t_h &= \mathbf{E}_{h-1}^* \mathbf{w}_h \\ \mathbf{E}_h^* &= \mathbf{E}_{h-1}^* - t_h \mathbf{p}_h^T\end{aligned}$$

On \mathbf{w}_h és el vector de pesos de la matriu \mathbf{X} , obtingut a la calibració per a la variable latent h ; \mathbf{E}_0^* equival a la matriu de dades de predicció \mathbf{X}^* i \mathbf{E}_{h-1}^* equival als residuals successius de la matriu \mathbf{X}^* per a cada variable latent.

Per a determinar la variable resposta (propietat analítica) es calcula

$$\mathbf{Y}^* = \sum_{h=1}^D b_h t_h \mathbf{q}_h^T$$

I el resultat obtingut es reescala convenientment.

A.4.1.5. Avaluació de Resultats. Determinació de les condicions finals de l'equació de Calibració.

Tant si es fa servir PCR com PLSR es fa necessari algun sistema que, a part de saber si una calibració proporciona una capacitat predictiva apropiada, permeti avaluar la conveniència de fer servir més o menys components principals (o PLS) en una determinada calibració.

En general, l'eina principal que permet l'avaluació d'uns resultats és l'anàlisi dels residuals. D'aquesta manera, el que es busca en tota calibració és generar una equació predictora que els minimitzi, permetent predir els valors de les variables resposta amb mínimes diferències envers el seu valor real.

Hi ha moltes maneres d'estudiar-los, ja sia de forma directa o bé a partir d'estadístics globals. Aquests acostumen a contenir la informació de tots els residuals posant

l'èmfasi en uns o altres aspectes. Acostumen a ser els més utilitzats perquè, si el conjunt de mostres és prou representatiu de la població, el comportament del seu error promig és de l'ordre del que es pot esperar de la població total.

El més utilitzat és l'anomenat PRESS (*Predicted Residual Error Sum of Squares*). És també el més senzill de calcular. Per a n objectes

$$\text{PRESS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Aquest estadístic d'error és la base per d'altres estadístics, com el MSE (*Mean Squared Error*), en el que es promitja el PRESS pel nombre d'objectes

$$\text{MSE} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}$$

En els processos de calibració basats en reducció de variables, és necessari disposar d'un conjunt de dades de prova (o *test set*) que ens permeten avaluar si l'equació de calibració obtinguda és o no acceptable. Quan el nombre d'objectes de què es disposa és prou elevat, la solució és relativament simple. Es divideix el conjunt d'objectes en dos, ambdós suficientment representatius de la població global, i es calculen les equacions de calibració amb un dels conjunts, restant l'altre per a fer-ne la predicció de la variable resposta. Per comparació amb els valors de referència, que són coneguts, poden calcular-se els residuals i els estadístics d'error.

No obstant, això fa necessari l'ús de moltes dades, i aquesta situació no és l'habitual. Quan es treballa amb dades espectrofotomètriques, on el nombre de variables acostuma a ser elevat, és necessari un gran nombre d'objectes amb la intenció que les equacions de calibració siguin estables. Com que això no sempre és possible, ha estat necessari triar altres estratègies de validació. La més habitual acostuma a ser la validació creuada o *Cross-validation*.

El principi de la validació creuada és senzill. Es tracta de fer servir les pròpies mostres de calibració per a comprovar la capacitat predictiva de les equacions calculades. Per a fer-ho, el conjunt de calibració és dividit en diferents segments que contenen un cert nombre d'objectes. A l'hora de fer els càlculs, es calculen les equacions de calibració fent servir tots els segments menys un, que resta com a segment de validació. Les equacions de calibració són calculades tantes vegades com segments es tenen, deixant-ne un fora alternativament. Aquest procés es repeteix per a cada

component principal (o de PLS) afegit, acumulant els valors dels errors de predicció de tots els segments per a cada component calculat. En el cas particular que cada segment contingui un únic objecte, el mètode de validació creuada rep el nom de *Leave One Out*. Hi ha diferents maneres de calcular els errors de predicció, tot i que acostumen a ser variants del MSE. En una d'aquestes variants, es divideix el valor del PRESS pel nombre de graus de llibertat. En aquest cas es té l'anomenada variància residual, que pot expressar-se en termes absoluts o bé en termes relatius. Una altra forma relativament comú de treballar és amb valors de PRESS acumulat.

La utilitat del càlcul d'aquests estadístics globals de validació radica en el comportament que s'espera d'ells. Quan es calculen els valors de MSE pels objectes de calibració (MSEC), és raonable esperar una disminució continuada del seu valor a mesura que s'incrementa el número de components principals (o de PLS) que incorpora l'equació de calibració. Dit d'una altra manera, creix la capacitat d'ajust de l'equació obtinguda. El comportament de l'error que s'obté en el conjunt de validació (MSEV), ja sia un test set, o bé l'acumulació de l'error dels diferents segments de la validació creuada, és diferent. En aquest cas, a l'incrementar-se el nombre de components principals (o PLS), l'error de predicció disminueix progressivament fins a arribar a un valor mínim. Superat aquest valor, el fet d'afegir més components a l'equació de calibració comporta la incorporació de variabilitat no relacionada amb la variable resposta i un desajust dels valors predictius que es coneix amb el nom de *sobreajust*.

Com es pot veure, la selecció del nombre adequat de components és un pas fonamental en l'obtenció d'una equació de calibració que proporcioni una bona capacitat predictiva. Si el comportament del sistema en estudi es correspon amb el teòric, és útil fer servir el sistema suggerit per Wold l'any 1987¹⁴. Es proposa la representació gràfica dels valors de MSE dels objectes de validació envers el número de components principals (o PLS), amb la intenció de determinar gràficament el mínim d'error per les mostres de validació.

Malauradament, aquest mètode és orientatiu però no proporciona resultats òptims²¹. En molts casos, el comportament asimptòtic de la corba d'error de validació no permet la determinació clara d'aquest mínim i obre la porta a possibles sobreajustos de les dades. Per altra banda, la utilització de mínims locals d'error en la determinació, no està exempta de riscos i pot conduir a equacions subajustades. Amb la intenció d'obtenir un mètode general de selecció, Haaland i Thomas²², varen proposar un criteri basat en consideracions estadístiques aplicant una variant del criteri F de comparació de variàncies.

Aquest mètode busca quin és el número de components principals (o PLS) a partir del qual el fet d'afegir-ne de nous no proporciona una disminució significativa del MSE de les mostres de validació interna (MSEV). El punt de partida és el mínim valor de MSEV calculat (per i components). Aquest valor es compara amb l'obtingut en el component immediatament anterior mitjançant un test de Fisher (Test F). És a dir

$$F(i - 1) = \frac{\text{MSEV}(i - 1)}{\text{MSEV}(i)}$$

Aquest valor F empíric es compara amb el valor F tabulat per a un test F unilateral amb un nivell de significació ($1-\alpha$). Els graus de llibertat pel numerador i el denominador són iguals. Vénen donats pel nombre d'objectes quan existeix una única variable resposta (PLS1) o bé pel producte del número de variables resposta pel número d'objectes quan n'hi ha més d'una. L'aspecte que mou més crítiques en l'aplicació d'aquest mètode prové del nivell de significació que fa servir. Aquest s'ha fixat empíricament en $\alpha=0.25$.

Si el valor de $F(i - 1) > F_{\alpha;g_{\text{dli}};g_{\text{dli}}}$, aleshores es compara $F(i - 2)$ amb F tabulat. Es calculen valors de F successius fins que $F(\text{calculat}) < F_{\alpha;g_{\text{dli}};g_{\text{dli}}}$. Aquest valor indica quin dels components comença a ser significatiu i no pot ser obviat de l'equació final.

A.4.2. Tècniques Qualitatives

La instrumentació analítica actual proporciona un nombre molt elevat de dades, de les quals se'n aprofita usualment una petita part. El seguiment de l'absorbància a una única longitud d'ona en espectrofotometria, o bé el seguiment d'una única unitat de massa a GC-MS, deixa sense ús un nombre elevat de dades analítiques.

Amb la intenció d'aprofitar aquest excés aparent d'informació, l'objectiu principal de l'aplicació de tècniques d'anàlisi multivariable ha estat doble. Per una banda, modelar la relació de les dades obtingudes amb dades analítiques de diferent origen (calibracions multivariables) i per l'altra, l'intent d'agrupar i classificar els diferents objectes en raó de les seves similituds. D'aquesta manera, s'han posat a punt tècniques quimiomètriques no supervisades, d'agrupació, i tècniques supervisades, de classificació.

Dins dels mètodes no supervisats més comuns es troben les tècniques d'agrupació, els arbres de mínima expansió i les xarxes neuronals de Kohonen. Dels mètodes supervisats es poden comentar l'anàlisi discriminant, el mètode dels K veïns més

propers, els mètodes basats en mesures de densitat i els mètodes basats en el modelat de les dades.

A.4.2.1. Tècniques no supervisades.

Les tècniques no supervisades permeten establir l'existència d'agrupacions entre variables que pertanyen a una mateixa matriu de dades, sense tenir-ne un coneixement previ. El principal exponent d'aquestes tècniques és l'anomenada **anàlisi de Clusters**²³. Aquests mètodes ordenen seqüencialment els objectes tenint en compte les similituds entre les seves variables, generant *clusters* o agrupacions. Permeten obtenir clusters ordenats de forma jeràrquica o no .

En aquests mètodes, és important el terme de *similitud*. En tècniques de reconeixement de pautes el concepte de similitud apareix lligat al de distància en l'espai de variables. Així, si dos objectes són similars, la distància entre ambdós serà petita. Si no ho són gaire, la distància interobjectes serà més gran.

Hi ha diferents formes d'expressar el concepte de distància. El més general és el que es coneix com a distància de Minkowski o mètrica L_p . S'expressa com

$$d_{ij} = \left(\sum_{k=1}^K |x_{mk} - x_{nk}|^p \right)^{1/p}$$

On K representa el nombre de variables i m i n els diferents objectes. Quan $p=1$, tenim el que es coneix com a distància Manhattan. Quan $p=2$, la distància calculada és l'Euclidiana.

Quan les variables són correlacionades, la mètrica L_p no és una bona interpretació de la distància i tendeix a errar els resultats. Aquest tipus de mètrica no té en compte l'efecte de la correlació i, a més, no té en compte les diferents escales de les dades. La mesura de la distància que té en compte aquests dos efectes és la distància de Mahalanobis. Aquesta distància es calcula mitjançant

$$D_{mn}^2 = (\mathbf{x}_m - \mathbf{x}_n)^T \mathbf{C}^{-1} (\mathbf{x}_m - \mathbf{x}_n)$$

On \mathbf{x}_m i \mathbf{x}_n són els vectors columna per a cadascun dels dos objectes i \mathbf{C} és la matriu de covariància.

Un terme complementari a les mesures de distància són les mesures de similaritat. Prenent com a base la distància de Minkowski, aquestes vénen donades per

$$S_{mn} = 1 - \frac{d_{mn}}{d_{mn}(\max)}$$

Aquest valor de $d_{mn}(\max)$ dóna una idea de la distància màxima esperable, que ve a ser el rang de la variable k per a aquell conjunt de mostres. Quan la similitud dels objectes és completa, $S_{mn}=1$. Quan són totalment dissimilars, $S_{mn}=0$.

Mètodes Jerarquitats

Una de les formes d'establir agrupacions d'objectes en l'espai variables predictores és mitjançant la seva aglomeració jeràrquica. Els objectes són agrupats en funció de les distàncies entre ells i, depenent de si el punt de partida és un únic objecte o bé és el total d'ells, els procediments poden ser **aglomeratius** o **divisius**. La determinació del número d'agrupacions acostuma a fer-se a partir del coneixement previ que es té del sistema i de les classes esperables. Quan això no és possible, es predetermina una distància.

L'agrupació jeràrquica d'objectes es pot realitzar fent servir diferents criteris

Criteri	Expressió
<i>Weighted Average Linkage</i>	$d_{ki} = \frac{d_{Ai} + d_{Bi}}{2}$ <p>Assumeix que els tamanys de les aglomeracions i els seus pesos són iguals.</p>
<i>Single Linkage</i>	$d_{ki} = \frac{d_{Ai} + d_{Bi}}{2} - \frac{ d_{Ai} - d_{Bi} }{2} = \min(d_{Ai}, d_{Bi})$ <p>També conegut com a 'veí més proper'. La distància entre dos grups ve donada per la distància entre el parell d'objectes de cada grup que es troba més proper. En general forma aglomeracions allargades.</p>
<i>Complete Linkage</i>	$d_{ki} = \frac{d_{Ai} + d_{Bi}}{2} + \frac{ d_{Ai} - d_{Bi} }{2} = \max(d_{Ai}, d_{Bi})$ <p>La distància entre els dos grups es mesura entre els dos objectes que es troben més allunyats (veí més llunyà). En general es formen aglomeracions ben separades, petites i compactes.</p>
<i>Unweighted Average Linkage</i>	$d_{ki} = \frac{n_A}{n} d_{Ai} + \frac{n_B}{n} d_{Bi} \text{ on } n = n_A + n_B$ <p>Fa servir el número d'objectes a cada aglomeració per a ponderar distàncies. No s'observen distorsions de l'aglomeració, però poden aparèixer petites aglomeracions d'outliers.</p>
<i>Centroid Linkage</i>	$d_{ki} = \frac{n_{Ai}}{n} d_{Ai} + \frac{n_B}{n} d_{Bi} - \frac{n_A n_B}{n^2} d_{AB}$ <p>Es fa servir el centroid (calculat com a promig de l'aglomeració) com a base de l'agregació.</p>

Criteri	Expressió
<i>Median Linkage</i>	$d_{ki} = \frac{d_{Ai}}{2} + \frac{d_{Bi}}{2} - \frac{d_{AB}}{4}$ <p>Fa servir la mediana per a calcular el centroid. Conserva la importància de les aglomeracions petites.</p>
<i>Ward's Method</i>	$d_{ki} = \frac{n_A + n_i}{n + n_i} d_{Ai} + \frac{n_B + n_i}{n + n_i} d_{Bi} - \frac{n_i}{n + n_i} d_{AB}$ <p>Les agregacions es construeixen de manera que es doni la mínima variació en la suma d'errors quadràtics de l'aglomeració. Aquesta suma fa referència a les desviacions de cada aglomeració respecte el seu centroid. Dóna dendogrames ben estructurats.</p>

Taula A.2. Criteris d'agrupació jeràrquica.

Els arbres de mínima expansió (*Minimal Spanning Tree*, MST)²⁴ busquen la mínima distància que interconnecta tots els objectes. Basats en l'algoritme de Kruskal, agafen aquesta distància mínima com a punt de partida i van afegint connexions pel costat que té un valor menor, de manera que la figura que es va obtenint no es tanqui. Per a la formació de les aglomeracions es trenquen les cadenes més llargues.

Mètodes no jerarquitats

L'altre gran grup d'aglomeracions són les aglomeracions no jeràrquiques. Parteixen d'una assignació inicial de les aglomeracions que permet calcular un primer centroid. Posteriorment, aquest va sent modificat en funció de les distàncies de cada objecte als centroids inicials. El procediment conegut com a mètode de Forgy acostuma a seguir els següents passos

- Una primera divisió en aglomeracions aproximades.
- Es determinen els centroids de cada aglomeració i la distància de cada mostra a aquests centroids.
- Es busca quins objectes són més propers a cada centroid.
- Es calculen els nous centroids tenint en compte la informació anterior.

Els dos darrers passos es repeteixen fins assolir la convergència, que ve donada quan apareixen dues aglomeracions amb els mateixos objectes. Quan en lloc de centroids es fan servir objectes com a nucli d'aglomeració, a aquests se'ls considera centrotips.

A.4.2.2. Classificadors (ANN)

Les xarxes neuronals artificials (ANN) són presents en gran número d'aplicacions de la química. Tot i que es dóna una explicació més esplaiada en un apartat posterior, és bo saber que permeten fer associacions entre conjunts de dades, classificacions, transformacions (o mapeig) i modelat de la informació. Quan treballen com a eines classificadores el seu aprenentatge pot ser supervisat o no supervisat. Un exemple típic de xarxa no supervisada es la basada en la regla d'aprenentatge de Kohonen²⁵, que s'emmarca en el conjunt de les regles d'aprenentatge associatiu.

Aprenentatge associatiu

Quatre són les principals regles d'aprenentatge associatiu: l'aprenentatge Hebbià, l'aprenentatge 'instar', l'aprenentatge 'outstar' i la regla de Kohonen. L'aprenentatge Hebbià és el més senzill i es basa en el concepte d'aprenentatge per reforç: quan dues neurones connectades entre si donen una resposta alta, la fortalesa de la seva resposta s'accentua. Matemàticament aquesta idea s'expressa com

$$W(i, j) = lrA(i)P(j)$$

En aquesta expressió, els pesos \mathbf{W}_{ij} de l'entrada j a la neurona i experimenten un increment positiu i proporcional al producte de l'entrada j i la sortida de la neurona i . En altres paraules, es dóna un **reforç** de la connexió. L'efecte de la velocitat d'aprenentatge^B (lr), permet alentir el procés o accelerar-lo en funció de si es vol o no conservar de forma estable associacions antigues. Per a evitar un increment arbitrari dels pesos \mathbf{W} s'introdueix un terme exponencial de disminució dels pesos, amb el que l'expressió anterior es reescriu com

$$W(i, j) = lrA(i)P(j) - drW(i, j)$$

Els conceptes d'aprenentatge *instar* i *outstar* són introduïts per Grossberg²⁶ per a explicar la visió i d'altres sistemes d'aprenentatge de l'home i els animals. No seran explicats aquí, però ens podem quedar amb la idea que són mètodes d'aprenentatge per reforç, que permeten reconèixer vectors (en el cas de *instar*) i generar-los a partir d'un valor d'entrada (*outstar*). Ambdós estan molt relacionats amb la regla de Kohonen. (De fet, la regla de Kohonen és una modificació de l'aprenentatge *instar*).

L'expressió del canvi de pesos en una capa de S neurones amb aprenentatge *instar* és

$$W(i, j) = lrA(i)(P(j) - W(i, j))$$

^B En textos en anglès, *learning rate*.

On \mathbf{W} és la matriu de pesos, lr és la velocitat d'aprenentatge, \mathbf{A} és el vector de sortides *instar* per a cada neurona i \mathbf{P} és el vector d'entrada. Quan s'utilitza la regla d'aprenentatge de Kohonen, l'expressió que governa el canvi de pesos ve donada per

$$W(i, j) = lr((P(j) - drW(i, j)) \cdot A(i)) \quad i, A(i) = 1$$

Aprenentatge Competitiu

En l'aprenentatge competitiu, les neurones d'una mateixa capa es distribueixen per a reconèixer vectors que es presenten freqüentment. Els pesos d'una xarxa competitiva s'inicialitzen de forma aleatòria abans del procés d'entrenament. Quan es fa servir la regla d'aprenentatge *instar* per a entrenar una xarxa competitiva, només aquella neurona que presenta un vector de pesos més proper al vector d'entrada actualitza el seu valor. D'aquesta forma, aquesta neurona incrementa la seva capacitat per a reconèixer vectors d'entrada simil·lars al que acaba de reconèixer, i augmenta les seves possibilitats de resultar 'guanyadora' quan es presenti un altre vector d'entrada simil·lar. Aplicant aquesta regla *instar* només una neurona (la guanyadora) modifica els seus pesos en cada iteració. Si s'aplica la regla de Kohonen només a la neurona guanyadora el procés és computacionalment més efectiu. És per això que la majoria de les xarxes competitives es calculen fent servir aquesta regla.

Independentment de la regla d'aprenentatge que es faci servir, el resultat que s'obté és que cada neurona de la capa propera a un conjunt de vectors d'entrada ajusta la seva resposta cap a ells. En els casos en els que hi ha un número suficient de neurones, cada agrupació de vectors d'entrada simil·lars pot provocar respostes simil·lars en més d'una neurona de la capa. D'aquesta manera, la capa de neurones mapeja l'estructura de les dades d'entrenament^C.

Per norma general, les anomenades xarxes neuronals de Kohonen per a anàlisi no supervisada es basen en una única capa bidimensional de neurones. Aquest tipus de xarxa pot ser tractada de forma individual o bé acoblada a d'altres capes de neurones que facin servir sistemes d'aprenentatge diferents. El seu aprenentatge és competitiu i manté la topologia de les dades que es fan servir com a conjunt d'entrenament. Pel manteniment de la topologia fa servir funcions de distància que regulen el comportament de les neurones veïnes a la neurona 'guanyadora'. Bàsicament es fan servir funcions de distància de dos tipus: triangulars i de les de barret mexicà. El resultat de les aglomeracions obtingudes es pot utilitzar per a fer classificacions

^C El conjunt de vectors d'entrada que es fa servir per a ajustar els pesos de la xarxa acostuma a rebre el nom de **conjunt d'entrenament**.

afegint una capa addicional que tracti la informació amb una llei associativa. D'aquesta forma s'obté una xarxa classificadora.

A efectes pràctics, les xarxes de Kohonen funcionen com a sistemes de reducció de variables. Poden ser utilitzades com a sistemes de preprocessat de la informació i permeten fer prediccions sobre la pertinença o no d'un determinat objecte a una aglomeració concreta.

A.4.2.3. Tècniques Supervisades

Les tècniques supervisades de classificació presuposen un coneixement previ del sistema i de les possibles classes existents.

L'Anàlisi Discriminant Lineal (LDA) es tracta d'un mètode paramètric que permet trobar un hiperplà de decissió que fa de frontera entre classes. Requereix que cadascuna de les classes tingui una distribució normal. Per a calcular-lo l'objectiu és maximitzar la variància interclasse i minimitzar la variància intraclasse. Aquest fet permet calcular una funció frontera que separa ambdues classes. El concepte d'anàlisi discriminant lineal s'ha generalitzat per a conjunts de dades multivariades, amb el que han aparegut tècniques com el **PLS-DA** (*Partial Least Squares / Discriminant Analysis*). Es tracta d'una proposta relativament recent, exposada al CAC^D de Tarragona de 1996, que comença a aplicar-se amb certa lentitud^{27,28,29}. El mètode codifica les classes mitjançant la creació d'una matriu **Y** d'uns i zeros. Seguidament es calcula l'equació de predicció d'aquests valors codificats envers les variables predictorres (espectres) permetent fer prediccions d'objectes no classificats en funció dels valors codificats que s'obtenen. El principal problema que planteja és la continuïtat de l'interval de predicció (que no permet fer prediccions dicotòmiques) i que proporciona valors intermedis allunyats de 0 i 1.

El mètode dels **K veïns més propers** fa servir les distàncies euclidianes per a classificar els objectes. En la seva versió més senzilla, es calculen les diferents distàncies entre l'objecte problema i la resta d'objectes presents al conjunt de calibració. L'objecte problema s'associa al grup al que pertanyen la majoria dels *k* objectes que es troben a una distància més propera. El valor de *k* acostuma a trobar-se per optimització del conjunt de dades disponibles, però valors petits, de *k*=3 o *k*=5 són els més utilitzats.

^D Chemometrics in Analytical Chemistry, Tarragona, 25-29 de Juny del 1996.

Aquest mètode presenta l'avantatge de la seva senzillesa, i de ser un mètode que permet l'existència de múltiples categories. El principal inconvenient rau en que és sensible al número de mostres que formen els conjunts de calibració de les diferents classes, sobretot si aquests són molt diferents entre ells. Per a resoldre aquests problemes s'acostuma a utilitzar un criteri alternatiu al de la majoria per a assignar l'objecte al grup més proper. L'optimització d'aquest criteri és un pas més que ha de fer-se per a cada conjunt de dades.

Els mètodes basats en mesures de **densitat** suposen l'existència d'un camp de potencial al voltant dels objectes a agrupar. El perfil que agafen aquests camps pot variar, però essencialment està condicionat pel perfil de la funció de base utilitzada. N'hi ha moltes, però les més utilitzades per la seva simplicitat són les funcions triangulars i les gaussianes. A partir de les funcions de potencial a cadascun dels punts individuals es calculen els valors de potencial acumulatiu. Aquest potencial acumulatiu agafa valors diferents de zero al llarg de la classe.

Quan aquest valor de potencial es promitja agafa un valor probabilístic, amb el que es pot fer classificació basada en probabilitats. Quan es vol avaluar la pertinença o no d'un determinat objecte a una classe es calcula el valor de la funció probabilitat a les coordenades de l'objecte. Aquest és assignat a la classe on dona un valor de potencial més elevat. De la mateixa manera es calcula la frontera entre classes, essent aquesta la zona on els valors de posició donen els mateixos valors de probabilitat.

Els mètodes de **modelat** permeten obtenir equacions matemàtiques que descriuen la zona de l'espai multidimensional ocupada per una classe. Les mostres problema s'assignen a una o altra classe en funció de la distància a cadascuna de les zones descrites pels models matemàtics. Els principals exemples d'aquests tipus de classificació són el **SIMCA** (*Soft Independent Modelling of Class Analogy*)³⁰ o les xarxes neuronals (perceptrons multicapa) adaptats a l'anàlisi discriminant.

Quan es porta a terme una classificació mitjançant SIMCA, es realitza el modelat de cada classe a partir d'una descomposició en components principals (PCA). Cadascuna d'aquestes classes queda definida per un número de components principals significatius que no té perquè ser igual al de la resta de classes modelades. D'aquesta manera, les diferents classes queden determinades per cilindres, plans, octaedres o hiperplans de diferent dimensionalitat en funció del número de components principals utilitzat (1,2,3 ó n). La distància de cada punt a la figura de modelat calculada ve donada pel residual \mathbf{e}_{ij}^A . Aquest representa la distància respecte la figura definida pels components principals de l'objecte i a la variable j . La desviació estàndar global

d'aquests residuals permet calcular el gruix de la classe definida pel model. L'expressió de la desviació estàndar global és

$$s_o^A = \frac{\sum_{j=1}^m \sum_{i=1}^{n_A} \left(\epsilon_{ij}^A \right)^2}{(n_A - p_A - 1)(m - p_A)}^{1/2}$$

Per a n_A objectes de la classe A enregistrats per a m variables i p_A components principals utilitzats en calcular la classe A.

Aquest mètode supervisat permet fer prediccions de objectes externs (classificacions). Per a fer-ho, a partir dels vectors de *loading* es calculen els valors de *score* de l'objecte extern U i es calculen les distàncies e_U als diferents models preparats. Aquestes distàncies permeten calcular una desviació estàndar s_U^A que pot ser comparada amb la del propi model a través d'un criteri F per a veure si és o no significativa.

$$s_U^A = \frac{\sum_{j=1}^m \left(\epsilon_{Uj}^A \right)^2}{(m - p_A)}^{1/2} \quad F = \frac{\left(s_U^A \right)^2}{\left(s_o^A \right)^2}$$

Si la diferència no és significativa, es conclou que l'objecte desconegut pertany a la classe. Habitualment es poden calcular d'altres paràmetres, com la distància entre models, el poder modelant de les variables, el seu poder discriminant, les distàncies de l'objecte al model i l'efecte palanca ('Leverage') de l'objecte en la definició de la classe. Alguns d'aquests paràmetres permeten representacions gràfiques força útils com els diagrames de Cooman, que representen la distància dels objectes al centre d'altres models³¹.

A.4.2.3.1. Xarxes Neuronals Artificials (ANN)

Les ANN simulen de forma senzilla el comportament que es coneix de les neurones biològiques³². Aquestes són cèl·lules altament ramificades que es troben interconnectades les unes amb les altres mitjançant terminacions filamentoses de dos tipus diferents: les dendrites, encarregades de recollir el senyal que prové d'altres cèl·lules i l'axó, encarregat de transmetre el senyal cap a d'altres dendrites. El punt de contacte entre dues cèl·lules diferents rep el nom de connexió sinàptica. La força o

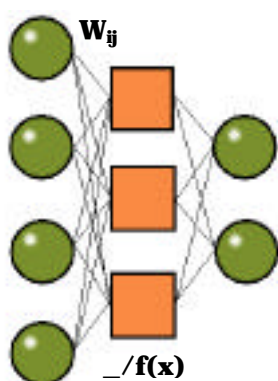


Figura. Simbolització habitual de una ANN. Les esferes corresponen als nodes d'entrada i de sortida. Els requadres corresponen als nodes de la capa oculta

influència del senyal provinent d'una cèl·lula sobre la següent està modulada per l'eficiència de la connexió sinàptica. És el que es coneix com a força sinàptica. Les diferents cèl·lules nervioses es connecten entre si generant teixits nerviosos. Aquests teixits són robustos i tolerants a errors. Alhora són sistemes flexibles, que s'autoajusten a noves condicions. Són sistemes que poden tractar informació difusa, amb pertorbacions o bé probabilística i poden funcionar en paral·lel.

Aquestes característiques de les neurones biològiques han estat simulades matemàticament per tal d'aprofitar-ne les seves propietats. D'aquesta manera, es considera la neurona com un element computacional connectat a d'altres que

accepta uns valors d'entrada i que proporciona uns valors de sortida tenint en compte una ponderació dels valors d'entrada. Es tracta de sistemes tolerants a errors, on l'alteració d'algun element del càlcul (neurona) no fa davallar dràsticament la capacitat predictiva. A l'anar deteriorant les connexions només es dona una disminució progressiva de la capacitat predictiva. Aquest fet és particularment interessant per d'altres aplicacions de les ANN allunyades de la química, com el control de plantes nuclears o el control de vols espaials. L'altra gran propietat que ha afavorit l'interès pel seu ús prové de la seva capacitat per a generalitzar, que li permet avaluar entrades test i proporcionar un valor de resposta apropiat encara que en el seu procés d'entrenament no s'hagi contemplat la variabilitat d'aquest tipus de dades. Tot i així es pot dir que són sistemes que interpolen millor que no pas extrapolen.

De forma anàloga als sistemes biològics, les neurones artificials acostumen a acceptar un nombre i d'entrades (\mathbf{x}_i) i a donar un únic valor de sortida \mathbf{o} . Per a obtenir aquest

valor de sortida, la neurona fa una suma ponderada de tots els valors d'entrada i aplica seguidament una funció de transferència.

$$o = f\left(\sum_i w_i x_i\right)$$

La neurona que opera d'una manera més simple és aquella que funciona com un llindar, transmetent la informació només si es supera un cert valor frontera **u**. És el que es coneix com un **perceptró simple**

$$o = f\left(\sum_i w_i x_i - u\right)$$

On la funció $f(\mathbf{x}) = \{1 \text{ si } \mathbf{x} \geq 0; 0 \text{ si } \mathbf{x} < 0\}$.

Les xarxes neuronals són conjunts d'aquests elements individuals que actuen junts i interconnectats amb la finalitat d'obtenir uns valors de resposta a partir de les variables presentades inicialment. Aleshores, per a una xarxa de j neurones cal modificar la notació, amb el que queda com a expressió de sortida

$$o_j = f_j\left(\sum_i x_i w_{ij} - u_j\right)$$

Les xarxes neuronals s'organitzen en seqüències de nivells (també anomenats capes). La capa d'entrada ve a ser una capa teòrica, perquè no conté cap neurona i només simbolitza la presentació de la informació a la xarxa. La capa de sortida és aquella que dona els valors calculats per la xarxa i també té com a única funció mostrar els resultats. La resta de nivells, compresos entre l'entrada i la sortida, són els que reben el nom de capes ocultes i són les que realment estan composades per grups de neurones individuals. La forma en la que les neurones estan interconnectades rep el nom d'arquitectura o topologia de la xarxa. En quant al seu funcionament, passa per dos processos diferenciats: l'aprenentatge i la resposta.

El procés d'entrenament de la xarxa neuronal comporta la modificació dinàmica dels valors dels pesos, amb l'objectiu de proporcionar uns valors de sortida, en funció d'uns certs valors d'entrada. Per a aconseguir-ho, hi ha tres tipus d'aprenentatge diferenciats: el supervisat, el no supervisat i l'aprenentatge per reforç.

En l'aprenentatge **supervisat**, el que la xarxa intenta és aconseguir la relació entre la entrada i la sortida que s'ha predeterminat. Si la sortida és diferent de l'entrada, la xarxa és heteroassociativa. Si per a tots els vectors d'entrada, la sortida és igual a la d'entrada, la xarxa és autoassociativa. Quan l'aprenentatge és **no supervisat**, no hi

ha resposta coneguda que pugui associar-se als valors d'entrada. Aleshores, la xarxa s'organitza internament de forma automàtica, de manera que respon a grups d'estímuls semblants (aglomeracions)^E. A l'aprenentatge per **reforç**, el valor de referència extern és una indicació de si la resposta és o no correcta. Aquesta indicació reforça les connexions entre neurones o bé les afebleix.

Independentment del tipus d'aprenentatge, l'element essencial de qualsevol xarxa neuronal és la manera en que actualitza els pesos al rebre diferents variables d'entrada. Es el que es coneix com les **regles d'aprenentatge**. En l'apartat anterior se n'han descrit dues: la de l'aprenentatge competitiu (no supervisat) i la de l'aprenentatge associatiu o Hebbià (per reforç). N'hi ha una tercera: la regla delta, corresponent a l'aprenentatge supervisat.

El procés de resposta comporta l'obtenció d'un valor de sortida per a un conjunt de variables fent servir una xarxa concreta. Aquest valor de sortida pot donar els resultats predictius per a un conjunt de mostres incògnita així com formar part del procés d'entrenament (validació de la xarxa amb un conjunt de mostres conegudes), on l'estudi dels errors predictius permet trobar les condicions predictives idònies de la xarxa. Les respostes més simples s'obtenen amb els sistemes de xarxa 'cap endavant' *Feedforward*, on les dades circulen des de la capa d'entrada a la de sortida en un únic pas. Quan les connexions són 'cap endavant' i 'cap endarrera', on la informació es mou al llarg de la xarxa, i només és presentada a l'exterior per la capa de 'output' al superar un cert valor frontera (generalment un error de predicció global), es tracta de xarxes de retropropagació o *Backpropagation*.

Hi ha molts tipus de xarxes neuronals que cobreixen gran nombre d'utilitats en diferents camps^{33,34}, des de la reconstrucció de senyals distorsionats fins el reconeixement de caracters. De totes elles, n'hi ha que serien perfectament vàlides per a realitzar diferents tipus de tasques quimiomètriques. Però la que s'ha consolidat en el camp de la quimiometria ha estat el **perceptró multicapa**. Aquest tipus de xarxa és un perceptró modificat que fa servir l'esquema de retropropagació per a actualitzar els pesos i utilitza la regla delta com a regla d'aprenentatge. Des del punt de vista químic es tracta d'un mètode ideal perquè està plenament adaptat a l'aprenentatge supervisat i permet ajustar tot tipus de dades a respostes predeterminades. Són especialment útils en aquells casos en els que és difícil disposar d'una solució numèrica o teòrica i és necessari disposar de prediccions fiables. Han estat aplicades a problemes quantitius i de classificació³³.

^E Han estat descrites en l'apartat de les tècniques de classificació no supervisades

Perceptrons Multicapa

A l'estructura del perceptró multicapa, la capa d'entrada presenta tantes neurones o nodes com variables té el vector d'entrada. Cada objecte ve descrit per un vector. Si cada objecte que va a passar per la xarxa té k variables, la capa d'entrada té, doncs k nodes. La capa de sortida té tants nodes com dimensions tingui el vector de sortida \mathbf{Y} . Com el seu nom indica, entre els nodes de la capa d'entrada i els de la capa de sortida el perceptró multicapa intercala una o vàries capes ocultes amb diferent número de neurones.

La neurona d'un perceptró pondera els valors d'entrada segons

$$I_j = \sum_i w_{ij} x_i + a_{j1}$$

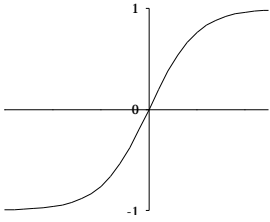
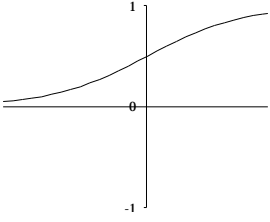
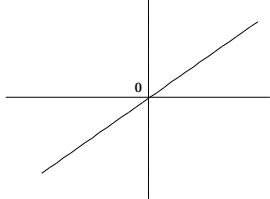
On w_{ij} són els pesos associats dins de la neurona j a la variable i ; x_i és el valor de la variable i a_{j1} és un terme d'ajust que s'afegeix per a cada neurona que rep el nom de biaix. Aquest és un valor constant, no nul, i que inicialment té valor 1. Compleix una funció similar a l'ordenada a l'origen dels sistemes lineals.

Un cop s'ha ponderat el valor de les variables d'entrada, al resultat se li aplica una funció de transferència. Aquesta funció ha de complir dos requisits: a) ser diferenciable en tots els punts del domini i b) ha de créixer (o decreixer) de forma monotònica al llarg de tot el domini de la funció. La capacitat d'ajustar sistemes no lineals prové precisament del tipus de funció seleccionada. Així es poden arribar a fer servir diferents funcions de transferència, tot i que la més utilitzada és la sigmoïdal.

Un cop aplicada la funció de transferència s'obté el valor de sortida. Aquest ve donat per

$$O_j = f(I_j) = f\left(\sum_i x_i w_{ij} + a_{j1}\right)$$

Aquests valors de sortida de les diferents neurones poden formar part de l'entrada de neurones situades a una altra capa, si és que aquesta existeix. I així successivament, el que permet generalitzar l'expressió per a una xarxa neuronal amb una capa oculta.

	Funció	Representació
<p>Tangent Hiperbòlica</p> <p><i>Tanh(x)</i></p>	$f(x) = \frac{1 - e^{-x}}{1 + e^{-x}}$	
<p>Sigmoidal</p> <p><i>Sigm(x)</i></p>	$f(x) = \frac{1}{1 + e^{-x}}$	
<p>Lineal</p>	$f(x) = ax$	

Taula A.3. Exemple de les funcions de transferència més utilitzades en química analítica.

Aquesta generalització s'expressa

$$o = f\left(\sum_{j=1}^J b_j g_j\left(\sum_{i=1}^I x_i w_{ij} + a_{j1}\right) + a_2\right) + e$$

On o és la sortida de la xarxa, x són les i variables d'entrada, w_{ij} són els pesos d'aquestes i variables per a cadascuna de les j neurones, b_j són els pesos de la capa de sortida i g_j és la funció de transferència de la capa de sortida, habitualment una funció lineal o bé la identitat. Els valors de a_{j1} i a_2 corresponen al biaix de cadascuna de les dues capes.

Aprentatge de la xarxa.

L'aprenentatge supervisat estableix una associació entre el vector de variables d'entrada i la variable (o variables) de sortida. És a partir d'aquests valors de sortida i dels valors de referència que és possible calcular un residual que monitoritzi la capacitat predictiva de la xarxa. Aquests residuals, però, presenten una segona propietat, més important que l'anterior: permeten la modificació supervisada del valor del vector de pesos, per a ajustar la resposta de la xarxa. La forma en que ho aconseguen constitueix la base de l'aprenentatge d'aquest tipus de xarxes de retropropagació i és el que es coneix com la *regla delta generalitzada*.

La regla delta

El procés de correcció dels pesos comporta que aquests vagin canviant iterativament a la finalització de cada cicle de càlcul. La regla delta busca que aquest increment de pesos sigui proporcional a cert paràmetre (η) que funciona com una constant de correcció. Alhora, η és proporcional a l'error i a x (entrada que ha donat una resposta errònia) i després de la correcció, el nou vector de pesos ha de donar un error menor. A partir del desenvolupament d'aquestes premisses, s'obté l'expressió estàndar de l'anomenada regla delta

$$w = \eta x$$

On l'increment del vector de pesos ve donat per la diferència d'aquests en dues iteracions consecutives. El valor de η és menor que 1 i rep el nom de velocitat d'aprenentatge.

$$w = w_{(t)} - w_{(t-1)}$$

Generalització.

Quan s'aplica la regla delta a xarxes de més d'una capa cal tenir en compte que a l'entrada de cada neurona no té perquè acudir-hi una variable directa x_i sino que també pot entrar-hi la sortida d'una neurona de la capa precedent. D'aquesta manera, pot expressar-se la regla delta com

$$w_{ij}^n = s_j^{n-1}$$

On s'ha simbolitzat com l'entrada d'una neurona j pertanyent a una capa n rep el resultat de sortida d'una neurona pertanyent a la capa anterior.

El càlcul del paràmetre η per a l'algoritme de retropropagació s'acostuma a calcular mitjançant l'algoritme del gradient descendent. Aquest algoritme es basa en la premisa que l'error envers la variable que l'origina ha de presentar un mínim en algun valor d'aquesta variable. En el cas que ens ocupa, es considera que l'error de predicció està relacionat amb el valor dels pesos. Aleshores, pot escriure's

$$w_{ij}^n = - \frac{E^n}{w_{ij}^n}$$

On η és un número d'escalat de signe positiu. El desenvolupament d'aquesta derivada parcial de l'error es pot trobar àmpliament documentada a les publicacions de Zupan i Gasteiger³³ i proporciona les equacions de correcció dels pesos. Es donen dues situacions diferents

- Per l'última capa de neurones

$$w_{ij}^n = 2 (t_j - S_j^n) f'(Net_j^n) S_i^{n-1}$$

On f' és la derivada de la funció de transferència utilitzada; Net_j^n és el resultat de l'entrada ponderada dels resultats de la penúltima capa ($Net_j^n = \sum_i S_i^{n-1} w_{ij}$). I el paràmetre η es calcula com

$$\delta_j^n = (t_j - S_j^n) f'(Net_j^n)$$

On t_j és el vector de valors de referència, bàsic per a l'aprenentatge supervisat.

- Per a les capes ocultes

$$w_{ij}^n = 2 f'(Net_j^n) \sum_k (\delta_k^{n+1} w_{jk}^{n+1}) S_i^{n-1}$$

On s'està calculant la variació dels pesos entre dues capes n i $n+1$ i k és el nombre d'entrades a la capa $n+1$. Aquesta expressió ens mostra com per a conèixer el canvi de pesos de la capa oculta n , cal informació de tres capes diferents: les entrades ponderades a la pròpia capa n (Net_j^n); la sortida de la capa anterior (S_i^{n-1}) i els valors de δ_k i w_{jk} calculats a la capa següent ($n+1$). El valor de η es calcula a partir de

$$\delta_j^n = f'(Net_j^n) \sum_k (\delta_k^{n+1} w_{jk}^{n+1})$$

Funcionament

En el procés d'entrenament, el vector d'entrada \mathbf{X}_k es propaga a través de la xarxa des de la capa d'entrada a la capa de sortida. En arribar a la sortida, es compara el valor obtingut per les diferents neurones amb el vector de referència \mathbf{T}_k (o *Target*), per tal d'obtenir el vector de pesos \mathbf{w}_{ij} de la darrera capa.

Des de la sortida, i anant de la darrera capa a la primera (retropropagació) es calcula el valor de δ_j^n a partir de δ_j^{n+1} i s'actualitza el vector de pesos \mathbf{w}_{ij}^n a la capa n fent servir l'expressió bàsica de la regla delta generalitzada ($\delta_j^n = \delta_j^{n+1} S_i^{n-1}$). El procés d'aprenentatge segueix pel vector corresponent al següent objecte \mathbf{X}_{k+1} i així successivament fins que s'assoleix un error de predicció dels vectors de *target* (\mathbf{T}_k) acceptable.

Es defineixen les iteracions com el procés d'entrada, correcció dels pesos i avaluació de l'error de predicció per a tots els vectors del conjunt d'entrenament \mathbf{X} . Com es pot suposar, per assolir la convergència d'una xarxa mitjana pot ser necessari un nombre d'iteracions molt elevat, el que incrementa de forma considerable el temps de càlcul. És per això que l'expressió bàsica de la regla delta generalitzada ha estat expandida amb la incorporació de nous termes que miren d'evitar que l'evolució de l'error presenti les mínimes oscil·lacions possibles i es pugui quedar trabada en mínims locals.

$$w_{ij}^n = \delta_j^n S_i^{n-1} + \mu w_{ij}^{t-1}$$

El terme μ rep el nom de momentum i és el responsable d'evitar que la xarxa quedi trabada en mínims locals d'error. És una constant de proporcionalitat que fa participar, en major o menor mesura, la correcció de pesos efectuada en el cicle anterior d'entrenament (δ_j^{t-1}) i permet d'aquesta manera donar una certa inèrcia al canvi de pesos. Habitualment agafa valors entre (0.5 i 0.9). Cal, però, anar amb compte amb el seu ús en xarxes grans, perquè és exigent amb els recursos de computació al necessitar del doble d'espai per a guardar els pesos de dos cicles consecutius.

En general, el procés d'entrenament (o de calibració) de les xarxes segueix les mateixes pautes que en el cas dels sistemes lineals. Per a evitar problemes de sobreajust s'avalua la capacitat predictiva de la xarxa calculada mitjançant un conjunt de variables predictoras externes que no intervé en el procés d'entrenament i de les que es coneix el valor de la resposta (Test Set). És la metodologia de validació de ANN

més utilitzada per la seva simplicitat, tot i que requereix disposar d'un nombre elevat de mostres.

El perceptró multicapa com a eina de classificació

A part de la capacitat dels perceptrons simples com a eina de classificació, les possibilitats d'ajust de vectors de *Target* per part de les xarxes de retropropagació permeten pensar en aquests sistemes per a definir mètodes que, amb el mateix objectiu, involucrin conjunts de dades més complexes³⁵. Amb aquesta finalitat, el que cal és codificar les classes amb vectors de *target* adequats, de forma anàloga a la que s'aplica a PLS-DA.

Per a establir una comparació entre ambdós mètodes, PLS-DA i la xarxa de retropropagació, es pot recórrer a una comparació general d'ambdues metodologies de càlcul.

El vector de sortida d'una xarxa neuronal d'una sola capa oculta es pot representar com

$$S = f\left(\sum_{j=1}^J b_j g_j\left(\sum_{i=1}^I x_i w_{ij} + a_{j1}\right) + a_2\right) + e$$

On el sentit de cadascun dels vectors implicats ja ha estat descrit al parlar dels perceptrons multicapa. Es pot afegir una part lineal a l'equació afegint connexions lineals al model entre la capa d'entrada i la de sortida³⁶

$$S = f\left(\sum_{j=1}^J b_j g_j\left(\sum_{i=1}^I x_i w_{ij} + a_{j1}\right) + a_2\right) + \sum_{j=1}^J d_j x_i + a_3 + e$$

On \mathbf{d} és el vector de pesos que pondera els valors d'entrada. Si es té en compte que en el cas de la regressió lineal múltiple (MLR), es pot relacionar la variable resposta amb les predictores mitjançant

$$y = a + \sum_{i=1}^I b_i x_i + e$$

es pot veure aquesta expressió com un cas particular de l'equació de la xarxa de retropropagació. En aquest cas, s'ha eliminat la capa oculta i s'ha escollit la funció identitat com a funció de transferència. Tot i la similitud entre ambdues expressions, l'ajust dels pesos es fa de forma diferent.

Quan es té en compte altres tipus de regressions com PCR i PLS, partint de dades centrades, l'ajust entre resposta i variables predictores es pot expressar com

$$y = \sum_{j=1}^J b_j \left(\sum_{i=1}^I w_{ij} x_i \right) + e = \sum_{j=1}^J b_j t_j + e$$

On els pesos \mathbf{w} són els vectors de *loading* per als diferents components principals i \mathbf{b}_j són els coeficients de regressió. Aquesta expressió és en certa manera equivalent a agafar l'expressió del perceptró multicapa i considerar \mathbf{f} i \mathbf{g} com a funcions lineals. Aleshores, la principal diferència entre ambdós mètodes d'ajust prové de les restriccions en l'aprenentatge de PCR i PLS. Així com en aquests dos mètodes, els vectors de score \mathbf{t}_j i els vectors de *loading* \mathbf{w}_{ij} han de ser ortogonals per a tots els i components calculats, això no és necessari pel cas de les ANN.

A.4.2.3.2. Regressió Logística

La regressió logística (LR), com qualsevol altre tipus d'anàlisi de regressió mira d'ajustar un conjunt de variables resposta a partir d'una sèrie de variables predictores. Es distingeix de la regressió lineal tant pel tipus de variable resposta que fa servir, ja que mentre la regressió lineal fa servir variables resposta contínues al llarg d'un cert interval, la regressió logística fa servir variables dicotòmiques (binàries); com per les assumpcions que fa.

- En qualsevol regressió lineal, el que es busca en el fons és una bona estimació del valor de la variable resposta en funció de les variables predictores, $E(Y|x)$. Es considera doncs que

$$E(Y|x) = \beta_0 + \beta_1 x_1$$

En LR, al ser dicotòmiques les variables resposta, l'estimació manté el seu valor en l'interval $[0,1]$. En general, al fer la distribució dels valors de resposta al llarg de l'interval comprès per les variables predictores, aquesta agafa un perfil en forma de 'S', similar al que s'obté en una distribució acumulativa d'una variable aleatòria. Hi ha diferents models que descriuen aquest comportament. Un d'ells és la distribució logística. Es selecciona per la seva flexibilitat, la seva facilitat d'ús i la seva fàcil interpretació.

El model d'estimació logístic s'expressa

$$\sigma(\mathbf{x}) = \frac{e^{0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n}}{1 + e^{0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n}}$$

On $\sigma(\mathbf{x})$ equival a $E(Y|\mathbf{x})$.

Per a poder tractar aquest tipus de model de forma anàloga a un model lineal es fa servir la transformació **Logit**. Presenta l'avantatge de donar una funció $g(\mathbf{x})$ que és lineal en els seus paràmetres, contínua i pot agafar qualsevol valor en funció de les variables predictores

$$g(\mathbf{x}) = \ln \frac{\sigma(\mathbf{x})}{1 - \sigma(\mathbf{x})} = 0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

- La segona gran diferència entre la regressió lineal i la regressió logística fa referència a la seva forma d'estimar els errors. El model lineal presuposa que l'error associat a una variable resposta estimada té una distribució normal al voltant del zero i que la variància de les variables predictores és constant. En canvi, quan la variable resposta és dicotòmica, aquesta assumptió és diferent. En aquest cas, l'error associat segueix una distribució binomial.

L'ajust dels paràmetres de la regressió segueix un camí relativament similar al de la regressió lineal, però tenint en compte algunes diferències significatives. L'ajust lineal més utilitzat és el de mínims quadrats, on es calculen els estimadors de la recta que presenta una distància mínima envers els punts experimentals. En les condicions habituals de càlcul, aquests estimadors tenen unes propietats estadístiques que deixen de complir-se quan la variable resposta es dicotòmica. En aquests casos, el mètode que s'utilitza, i que vindria a ser l'equivalent del mètode de mínims quadrats, rep el nom de *Maximum Likelihood*. Aquest mètode busca maximitzar la probabilitat d'obtenir el conjunt de dades de calibració a partir de les variables predictores. Per això és necessari disposar d'una funció (*likelihood function*), que relacioni la probabilitat amb els estimadors i que ha de ser maximitzada. Aquesta funció s'expressa com

$$l(\hat{\mathbf{a}}) = \prod_{i=1}^n (x_i)^{y_i} [1 - (x_i)]^{1-y_i}$$

Matemàticament és més senzill treballar amb la forma logàrítica, pel que habitualment s'utilitza la forma *log likelihood*

$$L(\hat{\alpha}) = \ln[L(\hat{\alpha})] = \sum_{i=1}^n \{y_i \ln[\hat{\pi}(x_i)] + (1 - y_i) \ln[1 - \hat{\pi}(x_i)]\}$$

Per a maximitzar aquesta expressió, es diferencia respecte β_0 i β_1 i s'igualen les expressions a zero. Per a una única variable x resulten

$$\sum_{i=1}^n [y_i - \hat{\pi}(x_i)] = 0$$

$$\sum_{i=1}^n x_i [y_i - \hat{\pi}(x_i)] = 0$$

Ambdues expressions són no lineals pels valors de β_0 i β_1 i per a resoldre-les cal recórrer a mètodes iteratius de càlcul que acaben proporcionant el valor de β estimat ($\hat{\beta}$).

L'ajust d'un model logístic multivariable és una generalització del cas anterior, en el que s'assumeix que es disposa d'un conjunt de n vectors predictors independents amb p variables cadascun. L'ajust del model logístic comporta en aquest cas el càlcul del vector $\beta = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)$, que es realitza de forma similar al cas univariable. Una informació més detallada de la forma de calcular LR amb conjunts de dades multivariables es troba al llibre de D.W. Hosmer i S. Lemeshow³⁷.

-
- ¹ D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michotte i L. Kaufmann, *Chemometrics: A textbook*, Elsevier, Amsterdam, (1988).
- ² A. Savitzky i M.J.E. Golay, *Anal. Chem.*, **36**, 1627, (1964).
- ³ R.J. Barnes, M.S. Dhanoa i S.J. Lister, *Appl. Spectrosc.*, **43**, 772, (1989).
- ⁴ Q. Guo, W. Wu, D.L. Massart, *Anal. Chim. Acta*, **382**, 87, (1999).
- ⁵ P. Geladi, D. McDougall i H. Martens, *Appl. Spectrosc.*, **39**, 491, (1985).
- ⁶ T. Fearn, *Nir News*, **10-2**, 9, (1999).
- ⁷ J. Sjöblom, O. Svensson, M. Josefson, H. Kullberg i S. Wold, *Chemom. Intell. Lab. Syst.*, **44**, 229, (1998).
- ⁸ S. Wold, H. Antti, F. Lindgren, J. Öhman, *Chemom. Intell. Lab. Syst.*, **44**, 175, (1998).
- ⁹ J.E. Jackson, *User's Guide to Principal Components*, Wiley, New York, (1991).
- ¹⁰ S. Wold, K. Esbensen i P. Geladi, *Chemom. Intell. Lab. Syst.*, **2**, 37, (1987).
- ¹¹ H. Wold, *Multivariate Analysis*, Ed. Krishnaiah, P.R. Academic Press, New York, (1989).
- ¹² *Linear Algebra and its applications*, G. Strang, Third Ed. Harcourt Brace Jovanovich Publishers, San Diego, (1988).
- ¹³ M. Otto, *Chemometrics, Statistics and Computer Application in Analytical Chemistry*, Wiley-VCH Verlag, Weinheim, Deutschland (1999).
- ¹⁴ S. Wold, *Technometrics*, **20**, 397, (1987).
- ¹⁵ H. Martens and T. Naes, *Multivariate Calibration*, John Wiley & Sons, Chichester, (1989).
- ¹⁶ R.G. Brereton, *Analyst*, **125**, 2125, (2000).
- ¹⁷ P. Geladi i B.R. Kowalski, *Anal. Chim. Acta*, **185**, 1, (1985).
- ¹⁸ C.W. Brown, P.F. Lynch, R.J. Obremsky i D.S. Lavery, *Anal. Chem.*, **54**, 1472, (1982).
- ¹⁹ K.R. Beebe i B.R. Kowalski, *Anal. Chem.*, **59**, 1007A, (1987).
- ²⁰ H. Wold, 'Soft modelling by Latent Variables; the non-linear Iterative Partial Least Squares Approach' a *Perspectives in Probability and Statistics, Paper in honour of M.S. Bartlett*, Ed. J. Gani, Academic Press, London, (1975).
- ²¹ D.W. Osten, *J. Chemometrics*, **2**, 39, (1988).
- ²² D.M. Haaland i E.V. Thomas, *Anal. Chem.*, **60**, 1193, (1988).
- ²³ N. Bratchell, *Chemom. Intell. Lab. Syst.*, **6**, 105, (1989).
- ²⁴ O. Strouf, *Chemical Pattern Recognition*, Research Studies Press, Ltd., Lechtworth, Hertfordshire, (1986).
- ²⁵ T. Kohonen, *Self-Organization and Associative Memory*, 3 Ed., Springer Verlag, Berlin, (1989).
- ²⁶ S. Grossberg, *Studies of the Mind and Brain*, Dordrecht, Holland: Reidel Press, (1982).
- ²⁷ D.J. Louwse, A.A. Tate, A.K. Smilde, G. L. M. Koot i H. Berndt, *Chemometrics and Intelligent Laboratory Systems*, **46** (2), 197 (1999).
- ²⁸ H. J. M. Verhaar, E.U. Ramos i J. L. M. Hermens, *Journal of Chemometrics* , **10** (2), 149, (1996).
- ²⁹ A.D. Shaw, A. Di Camillo, G. Vlahov, A. Jones, G. Bianchi, J. Rowland, D.B. Kell, *Analytica Chimica Acta*, **348**, 357, (1997).
- ³⁰ S. Wold, *Pattern Recognition*, **8**, 127, (1976).
- ³¹ *The Unscrambler 7.5. Users' Guide*, Oslo, (2000).
- ³² B.J. Wythoff, *Chemom. Intell. Lab. Syst.*, **18**, 115, (1993).
- ³³ J. Zupan and J. Gasteiger, *Neural Networks for Chemists: An Introduction*, VCH, Weinheim, Germany, (1993).
- ³⁴ J. Zupan and J. Gasteiger, *Analytica Chimica Acta*, **248**, 1, (1991).
- ³⁵ C.W. McCarrick, D.T. Ohmer, L.A. Gilliland, P.A. Edwards i H.T. Mayfield, *Anal. Chem.*, **62**, 1791, (1990).
- ³⁶ T. Naes, K. Kvaal, T. Isaksson, i C. Miller, *J. Near Infrared Spectrosc.*, **1**, 1, (1993).
- ³⁷ D.W. Hosmer i S. Lemeshow, *Applied Logistic Regression*, 2^a Ed., Wiley Interscience, New York (2000).

APARTAT B

ESPECTROSCOPIA INFRARROJA

B.1. Espectroscopia de Transformada de Fourier (FT-IR)

B.1.1 Absorció de la radiació IR

La regió infrarroja de l'espectre (IR) es considera que està compresa entre els 780 i els 10^6 nm. És un interval ampli, que ha estat dividit en tres subzones, tant per raons instrumentals com per característiques de la pròpia interacció de la radiació. D'aquesta manera, l'infrarroig proper (Near InfraRed, NIR), comprèn la zona delimitada entre els 700 i els 2500 nm (dels 14300 als 4000 cm^{-1}); l'infrarroig mitjà (Middle InfraRed, MIR) es situa entre els 2500 i els 50000 nm (4000 a 200 cm^{-1}) i l'infrarroig llunyà (Far InfraRed, FIR) es troba entre els 50000 i els 10^6 nm (200 a 10 cm^{-1}).

Tot i que en els darrers anys hi ha hagut un fort creixement en el nombre d'aplicacions analítiques basades en espectrofotometria NIR i FIR, l'espectroscopia en l'infrarroig mitjà ha estat la tècnica més utilitzada en el desenvolupament d'aplicacions analítiques, donada l'elevada selectivitat de l'absorció d'aquest tipus de radiació per part dels enllaços mol·leculars. Aquesta selectivitat ha fet que s'hagi valorat positivament la utilització d'aquest tipus d'espectroscopia a l'hora d'acoblar-la a d'altres tècniques analítiques, com la cromatografia de gasos (GC)^{1,2}.

L'espectroscopia MIR presenta nombroses aplicacions qualitatives i quantitatives. Durant molts anys, ha estat una tècnica bàsica en la identificació de compostos

orgànics donades les bandes característiques que els diferents enllaços presenten en aquesta zona de l'espectre. Tot i la seva àmplia utilització com a eina analítica d'identificació, ha estat darrerament quan les aplicacions quantitatives han experimentat una certa expansió degut a l'elevada selectivitat de les absorcions ja esmentada. Aquesta característica ha fet possible abordar diferents quantificacions en matrius complexes, com els productes naturals, la determinació de compostos contaminants industrials, productes agroalimentaris, etc.

Absorció de la radiació MIR

La radiació infrarroja no és suficientment energètica com per a produir les transicions electròniques que es donen quan la radiació prové d'una font Ultravioleta- visible (UV-vis). D'aquesta manera, només té energia suficient per a provocar transicions entre els nivells rotacionals i vibracionals d'enllaços mol·leculars.

Per a que una mol·lècula presenti absorció de la radiació és necessari que aquesta experimenti un canvi en el seu moment dipolar (μ_D) a conseqüència d'un canvi en l'amplitud dels seus moviments de rotació o de vibració. L'expressió clàssica pel moment dipolar d'una mol·lècula ve donada per l'expressió

$$\mu_D = \sum_{i=1}^n q_i \cdot d$$

en la que es relaciona la magnitud de les càrregues i la distància entre elles. La vibració dels enllaços a temperatura ambient provoca una fluctuació regular del moment dipolar originant un camp elèctric. Aquest pot interactuar amb el camp elèctric de la radiació electromagnètica infrarroja si es compleixen determinades condicions. En el moment en el que la freqüència de la radiació coincideix exactament amb la freqüència de vibració de l'enllaç es dona una transferència d'energia que provoca un canvi en l'amplitud de la vibració mol·lecular. De forma totalment anàloga, els fenòmens de rotació provoquen fluctuacions periòdiques de μ_D .

D'aquesta manera, la radiació IR pot afectar l'enllaç mol·lecular originant dos tipus de transicions: rotacionals i rotacionals-vibracionals. Les primeres vénen donades per radiacions amb un nivell energètic molt baix ($\sim 100 \text{ cm}^{-1}$), que corresponen a les que apareixen a l'IR llunyà. Les transicions vibracionals són fàcilment relacionables amb els moviments vibracionals que experimenten les mol·lècules di- i triatòmiques. A mesura que el nombre d'àtoms creix, el major nombre d'enllaços que poden donar moviments vibratoris, així com l'acoblament entre els diferents moviments, dificulta l'assignació de les diferents transicions a moviments concrets de vibració.

Hi ha dos tipus bàsics de moviments de vibració: tensió i flexió. El primer provoca canvis en la distància interatòmica. El segon provoca una variació de l'angle que presenten dos enllaços. Poden ser de quatre tipus: de tisora, de balanceig, de batre d'ales i de torsió. Aquest tipus de moviments són propis de totes les mol·lècules de més de tres àtoms i poden donar-se fenòmens d'acoblament de les vibracions si dos o més d'aquests moviments impliquen enllaços units a un mateix àtom central.

Aquest procés d'absorció d'energia per part de l'enllaç es sap que està quantitzat. Per això el que seria d'esperar és un espectre de ratlles, si la resolució instrumental fos suficientment elevada, corresponent a les diferents transicions entre nivells. Això és cert, però només en part, ja que aquest tipus de resposta la donen únicament els gasos. La interacció entre mol·lècules en sòlids i líquids és la responsable de l'eixamplament de les línies, generant un espectre continu de bandes.

Com es pot veure, l'explicació del fenomen d'absorció IR és complex. Malgrat tot, inicialment es va intentar fer una modelització física senzilla del fenomen a través de les eines de la mecànica clàssica, originant el que es coneix com a model de l'oscil·lador harmònic.

Oscil·lador Harmònic

Aquest model es basa en els postulats de la mecànica clàssica i proposa assimilar el moviment vibratori de tensió d'una mol·lècula diatòmica al d'un oscil·lador harmònic constituït per dues masses m_1 i m_2 unides per un ressort elàstic. L'expressió que dona la freqüència natural d'oscil·lació per a aquest tipus de sistema és

$$m = \frac{1}{2} \sqrt{\frac{k}{\mu}}$$

On μ és la massa reduïda, que ve donada per $\mu = \frac{m_1 * m_2}{m_1 + m_2}$.

De tota manera, el tractament clàssic del moviment oscil·latori passa per alt algunes característiques de la mecànica de sistemes de partícules amb dimensions atòmiques, com es la quantització de l'energia. Per això, convé tenir en compte les equacions de la mecànica quàntica a l'hora d'explicar alguns fenòmens espectroscòpics observats experimentalment, com l'aparició de sobretons.

El model quàntic parteix dels postulats clàssics de l'oscil·lador harmònic per a descriure'n el comportament. L'expressió de l'energia potencial d'aquest oscil·lador, des del punt de vista quàntic, ve donada per

$$E_p = (v + \frac{1}{2}) * \frac{h}{2} \sqrt{\frac{k}{\mu}}$$

On v és el número quàntic vibracional, que només pot agafar valors sencers positius, incloent el 0, μ la massa reduïda i h la constant de Planck.

El terme entre parèntesi $(v + \frac{1}{2})$ és el que dona la principal diferència respecte la concepció clàssica de l'energia de l'oscil·lador. Introdueix el concepte de la quantització de l'energia potencial de l'enllaç a l'impedir que aquesta agafi qualsevol valor positiu. Com es pot veure, aquesta expressió quàntica conté la definició clàssica de l'energia potencial de l'oscil·lador, amb el que aquesta es pot reordenar com

$$E = (v + \frac{1}{2}) * h * \nu_m$$

On ν_m és la freqüència vibracional obtinguda per la via clàssica.

Si una radiació externa aporta l'energia suficient per a provocar una transició entre nivells vibracionals, al calcular ΔE com la diferència d'energies entre dos nivells consecutius es pot veure com aquest valor ha de ser constant, independentment de si les transicions es donen entre els nivells 0 i 1 ó 1 i 2.

$$\Delta E = h * \nu_m = \frac{h}{2} \sqrt{\frac{k}{\mu}}$$

Tenint en compte aquests aspectes, així com les regles de selecció de la mecànica quàntica, que estableixen un $\Delta v = \pm 1$, pels sistemes basats en un oscil·lador harmònic hauríem d'esperar una única banda d'absorció.

Oscil·lador no harmònic

Un enllaç químic és un sistema real que presenta desviacions respecte el model simple de l'oscil·lador harmònic. L'existència de forces coulòmbiques de repulsió entre àtoms, així com el fet de que la força de recuperació de l'enllaç no sigui totalment constant fan que el comportament observat sigui diferent de l'esperat. Aquests fets introdueixen el que es coneix com **anharmonicitat** del sistema. Aquesta anharmonicitat pot ser descrita per la mecànica quàntica mitjançant equacions força complexes, que depenen de la naturalesa de l'enllaç, així com dels àtoms implicats. Aquesta complexitat fa difícil la seva aplicació quantitativa fora de les mol·lècules més senzilles. Malgrat tot, l'aspecte més interessant està en la similitud que presenten, a nivells d'energia propers al fonamental, les corbes de potencial de sistemes no harmònics i harmònics.

L'efecte de l'anharmonicitat es manifesta amb dos tipus de desviacions:

- El no compliment estricte de les regles de selecció a números quàntics alts, amb el que transicions entre nivells teòricament no permeses ($\Delta v = \pm 2$; $\Delta v = \pm 3$) són possibles. Aquestes transicions són les responsables de l'aparició de les **bandes de sobretò**, que normalment surten a freqüències dues i tres vegades més altes que la fonamental. Origenen bandes d'absorció petites, donada la baixa probabilitat amb la que es donen, amb el que sovint ni es manifesten. Per posar un exemple, el primer sobretò, on $\Delta v = \pm 2$, és de 10 a 100 vegades de menor intensitat que la freqüència fonamental en funció de l'enllaç. A més, els nivells energètics no són espaiats de la mateixa manera. Així, per a nivells d'energia alts, els sobretons apareixen a freqüències lleugerament menors de les que cabria esperar
- La interacció entre moviments vibratoris de mol·lècules veïnes pot portar a l'aparició de bandes d'absorció a freqüències properes a la suma o a la diferència de les freqüències fonamentals dels dos moviments. La intensitat d'aquestes **bandes (combinació/diferència)** acostuma a ser baixa.

$$\nu_{\text{Comb}} = n_1 \nu_1 + n_2 \nu_2 + n_3 \nu_3 + \dots$$

n_i són números sencers i ν_i són les freqüències de les diferents transicions que contribueixen a la banda de combinació. Les combinacions que es presenten més freqüentment són aquelles on $n_1 = n_2 = 1$.

D'aquesta manera, podem veure com l'absorció de la radiació infrarroja està íntimament relacionada amb els moviments de vibració de la mol·lècula. Aquests moviments poden resumir-se bàsicament en dos: Estirament (*Stretching*), moviment oscil·latori on es produeix una elongació de la distància interatòmica al llarg de l'eix de l'enllaç, i Doblegament (*Bending*), on es produeix una variació en l'angle de l'enllaç. Cadascun d'aquests dos modes presenta alhora diferents modes de vibració addicionals. Per a mol·lècules poliatòmiques són inicialment possibles $3N-6$ modes normals de vibració ($3N-5$ en el cas de mol·lècules lineals).

No tots els modes normals de vibració presenten absorció. Alguns d'ells no provoquen canvis en el moment dipolar donada la simetria de la mol·lècula, en altres casos coincideixen en el valor d'energia de diferents modes. Hi ha casos en que l'absorció d'energia és tan petita que no és detectable i, en d'altres, l'energia de vibració necessita una radiació d'una freqüència que és lluny de les prestacions de l'instrument.

En conseqüència, els espectres d'infrarroig d'espècies poliatòmiques són molt difícils d'analitzar. L'acoblament entre moviments vibratoris no permet la determinació amb exactitud de la posició de l'absorció d'un grup funcional, ja que aquesta depèn força de l'entorn molecular. Aquest conjunt d'acoblements personalitzen els espectres dels diferents compostos, el que permet la identificació de productes molt similars.

Perquè es doni acoblament entre moviments vibratoris, s'han de complir una sèrie de condicions

- L'acoblament entre moviments vibratoris d'estirament és fort quan aquests comparteixen algun àtom de l'enllaç.
- Quan s'acoblen dos moviments oscil·latoris de flexió és perquè ambdós moviments comparteixen un enllaç.
- L'acoblament entre moviments de flexió i d'estirament és possible sempre i quan l'enllaç que experimenta els moviments de tensió formi part del sistema que té un moviment de flexió.
- La major interacció es dona quan ambdós moviments de vibració són aproximadament iguals.
- Les vibracions acoblades han de pertànyer al mateix grup puntual de simetria.
- Es dona molt poca interacció entre grups separats per dos o més enllaços.

B.1.2. Instrumentació

Atenent al sistema que utilitzen en la generació de la radiació monocromàtica, els instruments d'infrarroig es poden classificar bàsicament en **dispersius** i **no dispersius**.

En general, els instruments **dispersius** són aquells que utilitzen xarxes de difracció per a generar la radiació monocromàtica. En espectrofotometria IR acostumen a ser instruments de doble feix, que són menys exigents amb les característiques de les fonts i detectors i permeten la compensació de l'absorció de la radiació per part dels gasos de l'atmosfera. L'especial configuració del sistema d'atenuació del doble feix en aquest tipus d'instruments els fa lents de resposta. Això provoca que aquesta s'endarrereixi quan es donen canvis forts de transmitància al voltant de bandes intenses, que puguin donar-se sortides d'escala del sistema enregistrator i que la determinació de la posició nullo sigui difícil a les zones on la transmitància s'apropa a zero. Com a resultat s'obtenen bandes arrodonides i una resposta poc definida per part del detector. Aquestes característiques han fet que, a l'actualitat, els instruments dispersius hagin vist limitades les seves aplicacions a l'anàlisi qualitativa i que els

instruments que s'han anat imposant, per la relació favorable preu/prestacions, hagin estat els no dispersius.

Dins els instruments **no dispersius** es troben equips amb orientacions ben diferents. Per una banda els fotòmetres de filtres, que s'utilitzen en la determinació 'in situ' de paràmetres ambientals i en l'anàlisi de sòlids agraris i industrials. En aquest tipus d'instruments es valora la senzillesa, la robustesa i l'economia de manteniment de l'equip per damunt d'altres factors. Per a determinacions de gasos absorbents en la zona IR de l'espectre existeixen els anomenats fotòmetres sense filtre. Aquests dispositius es basen en l'absorció de la radiació policromàtica per part de dues cel·les en sèrie que contenen el gas de referència i la mescla de gasos a analitzar.

Per l'altra banda, hom disposa d'instruments multicanal (o *multiplex*). En aquests sistemes, el senyal es registra de forma simultània i posteriorment es decodifica amb l'ajut d'un algoritme matemàtic. Bàsicament, s'han aplicat dos procediments de decodificació. Són les anomenades *transformades* de Fourier i de Hadamard. Tot i que ambdues són alternatives vàlides, només la primera s'ha consolidat, aplicant-se a camps diferents de l'espectrofotometria, com la ressonància magnètica nuclear (RMN) i l'espectrometria de masses (MS).

En aquest tipus d'instruments no s'utilitza un sistema monocromador per a dispersar la radiació i mesurar la transmissió a cadascuna de les longituds d'ona individuals, sinó que és un feix policromàtic amb certes característiques el que incideix sobre la mostra. Aquest fet suposa diferents avantatges, que poden resumir-se breument en tres:

- **Avantatge de Jaquinot** (Rendiment). Suposa una elevació considerable de la relació senyal-soroll respecte els instruments dispersius, al ser sistemes que presenten un menor nombre d'elements òptics i no utilitzen esclatxes per a atenuar la radiació.
- **Exactitud en la longitud d'ona**. Els sistemes basats en un làser d'He-Ne presenten un elevada exactitud i precisió en la selecció de la longitud d'ona (inferior a 0.01 cm^{-1} a l'interval entre 4800 i 400 cm^{-1}), el que fa possible el promitjat d'espectres amb el conseqüent increment de la relació senyal-soroll. Els efectes deguts a radiacions paràsites són minimitzats al tractar-se d'un senyal modulad.
- **Avantatge de Fellgett**. S'aconsegueix perquè, a l'arribar totes les components de l'espectre alhora (codificades), el temps entre espectre i espectre pot ser sensiblement petit ($1''$). Aquest aspecte afavoreix el promitjat

d'un elevat nombre d'espectres que també augmenta la relació senyal/soroll de l'espectre.

A diferència de l'espectroscopia convencional, que és de domini de freqüències (s'enregistra la potència radiant rebuda pel detector a cada freqüència), l'espectroscopia de transformada de Fourier és de domini de temps. En aquest tipus d'espectroscopia el senyal obtingut és el resultat de la combinació de les diferents freqüències i rep el nom d'**interferograma**. Aquesta combinació genera un resultat complex que prové de la superposició d'ones que són lleugerament desfassades les unes de les altres. És el que s'anomena espectre de domini de temps.

Per a poder disposar d'un interferograma, cal un sistema que moduli el senyal d'alta freqüència, que conservi les seves característiques en el que respecta al temps (és a dir, que en sigui proporcional) i que permeti el seu enregistrament pels transductors actuals. El sistema rep el nom d'**interferòmetre**. N'hi ha dos que són força populars: l'interferòmetre de Michelson i l'interferòmetre de prismes.

L'interferòmetre de Michelson és un dispositiu òptic que divideix un feix de llum policromàtic en dues meitats aproximadament iguals. Aquests feixos es reflexen en dos miralls interns, un de fix i un de mòbil de velocitat perfectament coneguda. La diferència de distàncies recorregudes pels dos feixos als miralls fix i mòbil rep el nom de retard τ . La recombinació dels dos feixos amb un desfassatge entre ells provoca interferències constructives i destructives que es reflecteixen en la potència radiant que rep el detector. Quan es representa la potència radiant en funció del retard, es disposa de l'**Interferograma**.

Amb tot aquest procés el que s'aconsegueix és modular el senyal. Es disposa d'un senyal amb una freqüència molt inferior a la de la font emissora però que pot ser relacionada amb aquesta de forma senzilla

- Si el mirall mòbil es desplaça amb una velocitat constant v_M i es considera τ com el temps necessari perquè el mirall recorri una distància corresponent a la meitat de la longitud d'ona:

$$v_M = \frac{c}{2\tau}$$

- La freqüència del senyal serà igual a la inversa de τ , és a dir:

$$f = \frac{2v_M}{c}$$

- Per a relacionar la freqüència del senyal amb la freqüència òptica cal substituir adequadament ν , amb el que queda:

On ν és la freqüència òptica i c la velocitat de la llum. Com es desprèn d'aquesta expressió, quan v_M és constant, la freqüència de la potència radiant

$$f = \frac{2v_M}{c}$$

rebuda pel detector és proporcional a la freqüència òptica mitjançant una constant.

Transformada de Fourier

Un senyal enregistrat en el domini de temps pot ser representat com una combinació de funcions periòdiques. Així, l'anàlisi de Fourier és el procediment matemàtic que permet la descomposició d'una corba qualsevol en una suma de termes sinus i cosinus, anomenada sèrie de Fourier.

La potència radiant de la font ($P(\nu)$) i la potència del senyal de l'interferograma ($P(x)$) es poden arribar a descriure segons l'expressió

$$P(x) = B(\nu) \cos(2\pi \nu x)$$

On el terme $B(\nu)$ està relacionat amb la potència de la font de radiació, la divisió del feix per part de l'interferòmetre i la resposta del detector amb la freqüència.

Tenint en compte conceptes com el número d'ona i el retard τ , aquesta expressió pot reescriure's com

$$P(x) = B(\nu) \cos(2\pi \nu \tau)$$

Així, es pot considerar matemàticament l'interferograma complet com

$$P(x) = \int_{-\infty}^{+\infty} B(\nu) \cos(2\pi \nu x) d\nu$$

Quan s'aplica la transformada de Fourier a aquesta expressió es generen dues solucions: una de real i una d'imaginària. En espectroscopia s'utilitza la solució real, que dona el valor de $B(\nu)$ i que representa el valor de la potència radiant en l'espai de freqüències (espectre).

B.1.3. Mesures de Reflectància Total Atenuada

Tradicionalment, l'espectrofotometria IR havia estat bàsicament una tècnica d'anàlisi qualitativa, per la informació que proporciona sobre els enllaços presents als compostos. La tècnica de pastilla de KBr ha estat la tècnica estàndard de suport de mostres sòlides en aquest tipus d'instruments. Conjuntament, han tingut àmplia difusió les suspensions en dissolvents apropiats com el *Nujol* o els *Fluorolubes*. Per a mostres líquides s'han utilitzat cel·les d'acer amb finestres de materials transparents a l'IR, com el NaCl i el BaF₂, amb camins òptics que oscil·len entre els 0.005 i 5mm,

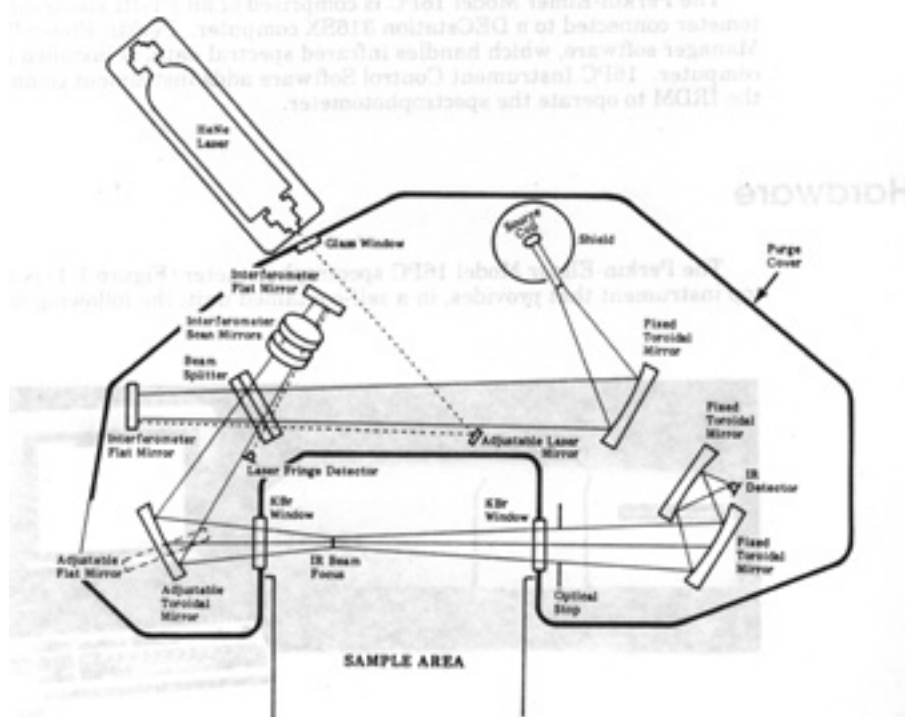


Figura B.1. Esquema Òptic d'un instrument de Transformada de Fourier. (Extret de *Perkin-Elmer 16 PC – User's Guide*).

donada l'elevada absorptivitat molar de la majoria dels compostos en aquesta zona de l'espectre. Per a mantenir aquests camins òptics s'han utilitzat espaiadors de diferents tipus, així com sistemes de camí òptic variable.

Amb l'aparició dels instruments de transformada, les aplicacions quantitatives han experimentat una expansió considerable. L'obtenció d'espectres amb una relació senyal-soroll realment millorada, així com l'elevada precisió en la selecció de la longitud d'ona n'han estat els principals artífexs. Aquest tipus d'aplicacions exigeixen camins òptics altament reproduïbles que han de ser continuament verificats. Malgrat

els avenços en la construcció de cel·les d'IR, amb segellats efectius i camins òptics precisos, consideracions pràctiques com la neteja de la cel·la entre mesures així com la fragilitat del material emprat (higroscòpic en alguns casos, com el NaCl) han originat la recerca de sistemes de mesures diferents, que aprofitin altres tipus d'interaccions radiació/mostra.

En aquest context apareixen les mesures de reflectància interna o reflectància total atenuada (ATR)³. Aquest tipus de mesura aprofita el comportament del feix de radiació IR al passar a través de dos medis d'índex de refracció diferents.

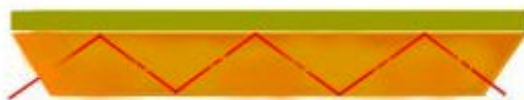


Figura B.2. Esquema d'una mesura per ATR en un sistema de canal. En la interfase superior del cristall es situa la mostra líquida.

En aquests sistemes, el feix IR travessa un cristall tallat d'un material transparent a l'IR (ZnSe) amb un cert angle d'incidència, superior a l'angle crític. A l'arribar a la interfície cristall-mostra el feix es reflexa quasi totalment. Només una

petita part d'aquest feix travessa la interfície i penetra lleugerament a la mostra. Després acaba sent reflectit i retorna havent experimentat el fenomen de l'absorció a les diferents longituds d'ona. La profunditat de penetració del feix IR en el sinus de la mostra és de l'ordre de la pròpia longitud d'ona i és funció de l'índex de refracció d'ambdues bandes de la interfície (mostra (n_s) i cristall (n_p)), de la longitud d'ona del feix incident (λ) i de l'angle d'incidència (θ). L'expressió de la distància de penetració del feix a la mostra és

$$d_p = \frac{\lambda}{2 \left(\sin^2 \theta - \left(\frac{n_s}{n_p} \right)^2 \right)^{1/2}}$$

Els espectres que s'obtenen per aquesta via són similars als obtinguts per transmissió i permeten de forma indistinta el registre de l'espectre de mostres sòlides i líquides amb una elevada simplicitat pràctica. Com que la penetració del feix és molt petita, això obre la porta al registre d'espècies altament absorbents, l'estudi de superfícies i el control de processos.

Els sistemes ATR estan constituïts per un sistema de miralls que dirigeix el feix des de la font emissora fins un extrem del cristall, amb un angle d'incidència que habitualment és de 30, 45 o 60°. A l'altre extrem es situa un altre sistema de miralls que recull la radiació atenuada i la dirigeix cap el detector. Per a conèixer el nombre de reflexions o punts de contacte del feix amb la mostra és necessari conèixer la llargada i el gruix del cristall, així com l'angle d'incidència. Són dispositius òptics on el paral·lelisme i el tallat del cristall són especialment crítics.

Els compartiments de mostra poden ser de molts tipus: de canal, amb sistemes de pressió, termostatitzats, inclús sistemes tipus sonda per a realitzar mesures directes. En aquests sistemes, el feix incident i el feix de retorn circulen pel mateix cristall i incideixen en un extrem, que és la superfície de contacte amb la mostra.

Els cristalls poden ser de diferents materials, tots ells amb índexs de refracció elevats. Al ZnSe (IRTRAN IV), s'hi afegeixen l'AgCl, el Iodur de Tal·li I/ Bromur de Tal·li I (KRS-5), el Germani, el Safir (Al_2O_3), etc. Tots ells són materials amb característiques químiques i físiques que els diferencien i fan que els uns siguin més apropiats que els altres per segons quins tipus de mostra. Així, l'AgCl i el Germani són apropiats per al registre de l'espectre de solucions aquoses i el ZnSe i el KRS-5 són apropiats per a les mesures en compostos orgànics. Aquests compostos han de mantenir una relació apropiada amb l'índex de refracció de la mostra i han de ser utilitzats amb certes precaucions, donades les seves característiques físiques i químiques. Així, el ZnSe és soluble en medi àcid per la formació de H_2Se , el Germani és mecànicament fràgil i el KRS-5 és soluble en solucions bàsiques.

B.1.4. Aplicacions a l'anàlisi agroalimentària

Malgrat l'especificitat de les bandes d'absorció IR, la tècnica ha presentat fins ara una baixa utilització en el camp agroalimentari. Tot i així, les principals limitacions que presentava en l'anàlisi quantitativa (relació senyal-soroll baixa i lentitud en l'anàlisi) van ser superades amb l'aparició dels instruments de transformada de Fourier.

Les millores en la instrumentació han portat al desenvolupament de noves aplicacions dins l'anàlisi agroalimentària, tant quantitatives com qualitatives. D'aquesta manera s'han dut a terme treballs de determinació del contingut de proteïna i greix en carns⁴; la humitat i el greix total en productes d'alt contingut gras, com les salses mahoneses⁵; l'anàlisi del percentatge d'àcids grassos insaturats en conformació cis- o trans- en greixos i olis comestibles⁶; el contingut d'àcids grassos lliures en greixos i olis⁷; l'índex de peròxids en olis vegetals comestibles⁸ i la determinació de la composició en greix, proteïna i lactosa de la llet⁹. Aquestes han estat aplicacions desenvolupades fent servir la tècnica clàssica de mesura de la transmitància, mitjançant cel·les de transmissió amb finestres de CaF_2 , amb camí òptic petit i termostatitzades en alguns casos.

De tota manera, a part de les millores en la instrumentació, un altre dels factors que ha contribuït decisivament al renaixement de l'interès d'aquest camp per la tècnica ha estat l'aparició i perfeccionament dels nous sistemes de mesura basats en fenòmens de reflectància: Reflectància Difusa (DRIFT) i Reflectància total atenuada (ATR). Tot i que la primera ha estat aplicada majoritàriament al control de processos, s'han estat

desenvolupant dispositius que permeten el registre de mostres agroalimentàries sòlides com els cafès solubles i les melmelades. Aquest fet ha obert la porta a estudis comparatius entre sistemes DRIFT i ATR, on s'han abordat problemes d'adulteracions¹⁰ i de discriminació entre varietats¹¹.

L'aplicació dels sistemes ATR a l'anàlisi agroalimentària ha estat diversa. Així, entre d'altres, s'han realitzat estudis quantitius del contingut d'àcids grassos en conformació trans de diversos productes alimentaris¹²; de la composició de llets condensades¹³; la monitorització de la producció de sucres durant la hidròlisi del midó^{14,15}; l'estudi de les concentracions de glucosa i fructosa en mesclades de fermentació en presència de diferents percentatges d'etanol¹⁶; el control de qualitat de xarops de glucosa¹⁷ i el desenvolupament de noves metodiques de dipòsit de la mostra a la superfície del cristall, aplicades al cas de sucres¹⁸. Al marge d'aquestes aplicacions quantitatives, s'han desenvolupat estudis encarats a la detecció d'adulteracions i a l'anàlisi semiquantitativa, particularment en el cas de fruites^{19,20,21,22} i carns²³.

El fet que darrerament s'hagin incrementat el nombre de publicacions basades en aquests sistemes de mesura confirma el renaixement de l'interès per aquesta tècnica i l'expansió de les seves possibilitats.

-
- ¹ M. Praisler, I. Dirinck, J.F. van Bocxlaer, A.P. de Leenheer i D.L. Massart, *Journal of Analytical Toxicology*, **25** (1), 45, (2001).
 - ² N. Ragunathan, K.A. Krock, C. Klawun, T.A. Sasaki i C.L. Wilkins, *Journal of Chromatography, A*, **856** (1-2), 349, (1999).
 - ³ J. Fahrenfort, *Spectrochimica Acta*, **17**, 698, (1961).
 - ⁴ B.Dion, M. Ruzbie, F.R. van der Voort, A.A. Ismail i J.S. Blais, *Analyst*, **119**, 1765, (1994).
 - ⁵ F.R. van de Voort, J. Sedman i A.A. Ismail, *Food Chem.*, **48**, 213, (1993).
 - ⁶ F.R. van de Voort, A.A. Ismail i J. Sedman, *J. Am. Oil Chem. Soc.*, **72**, 873, (1995).
 - ⁷ A.A. Ismail, F.R. van de Voort, G. Emo i J Sedman, *J. Am. Oil Chem. Soc.*, **70**, 335, (1993).
 - ⁸ F.R. van de Voort, A.A. Ismail, J. Sedman, J. Dubois i T. Nicodemo, *J. Am. Oil Chem. Soc.*, **71**, 921, (1994).
 - ⁹ F.R. van de Voort, J. Sedman, G. Emo i A.A. Ismail, *J. Am. Oil Chem. Soc.*, **75**, 780, (1992).
 - ¹⁰ R. Briandet, E.K. Kemsley i R.H. Wilson, *J. Sci. Food Agric.*, **71**, 359, (1996).
 - ¹¹ R. Briandet, E.K. Kemsley i R.H. Wilson, *J. Agric. Food Chem.*, **44**, 170, (1996).
 - ¹² L.H. Ali, G. Angyal, C.M. Weaver, J.I. Rader i M.M. Mossoba, *J. Am. Oil Chem. Soc.*, **73**, 1699 (1996).
 - ¹³ N. Nathier-Dofour, J. Sedman i F.R. van de Voort, *Milchwissenschaft*, **50**, 462, (1995).
 - ¹⁴ V. Bellon-Maurel, C. Vallat i D. Goffinet, *Appl. Spectrosc.*, **49**, 556, (1995).
 - ¹⁵ V. Bellon-Maurel, C. Vallat i D. Goffinet, *Appl. Spectrosc.*, **49**, 563, (1995).
 - ¹⁶ V. Bellon, *Sensors Actuators*, **B12**, 57, (1993).
 - ¹⁷ F. De Lene Mirouze, J.C. Boulou, N. Dupuy, M. Meurens, J.P. Huvenne i P. Legrand, *Appl. Spectrosc.*, **47**, 1187, (1993).
 - ¹⁸ N. Dupuy, M. Meurens, B. Sombret, P. Legrand i J.P. Huvenne, *Appl. Spectrosc.*, **47**, 452, (1993).
 - ¹⁹ M. Defernez i R.H. Wilson, *J. Sci. Food Agric.*, **67**, 461, (1995).
 - ²⁰ M. Defernez, E.K. Kemsley i R.H. Wilson, *J. Agric. Food Chem.*, **43**, 109, (1995).
 - ²¹ M. Defernez, E.K. Kemsley i R.H. Wilson, *Mikrochim. Acta*, **14**, 181, (1997).
 - ²² E. K. Kemsley, J.K. Holland, M. Defernez i R.H. Wilson, *J. Agric. Food Chem.*, **44**, 3864, (1996).
 - ²³ O. Al-Jowder, E.K. Kemsley i R.H. Wilson, *Food Chem.*, **59**, 195, (1997).

B.2. Espectroscopia d'Infrarroig Proper (NIR)

B.2.1 Introducció

La radiació en la zona de l'infrarroig proper va ser descrita a inicis del segle XIX per William Herschel¹. Tot i fer força anys d'aquest fet, l'aplicació d'aquesta espectroscopia a l'àmbit de la química analítica ha estat recent. De fet, el primer registre fotogràfic documentat de l'espectre de líquids orgànics a la zona de 700 a 1200 nm té data del 1881 i va ser fet per Abney i Festing². Les especials característiques de la àmplia zona de l'infrarroig va fer que durant la primera meitat del segle XX, els espectroscopistes s'interessessin més per la zona mitjana de l'espectre a l'hora de buscar aplicacions analítiques i així la zona de l'infrarroig proper va experimentar un cert oblit.

Les primeres aplicacions analítiques basades en espectrofotometria NIR es desenvolupen als anys 50, de forma paral·lela a l'aparició dels primers espectrofotòmetres comercials que poden treballar en aquesta zona espectral equipats amb detector d'efecte fotoelèctric. El 1954, Wilbur Kaye va presentar un treball³ en el que s'enregistraven espectres de diferents productes orgànics entre 700 i 3500 nm i s'assignaven les bandes dels diferents grups funcionals. Es proposen inclús aplicacions analítiques, com la determinació d'alcohols i d'aigua i l'anàlisi d'amines i hidrocarburs. A finals de la dècada dels 50, els treballs documentats d'aplicacions de l'espectroscopia NIR ja són al voltant de 40⁴.

El primer impuls important experimentat per l'espectroscopia NIR és a la dècada dels 60. En aquesta època, Karl Norris, líder d'un grup d'investigació de la USDA (*United States Department of Agriculture*), comença a experimentar amb les possibilitats de la tècnica en l'estudi de matrius complexes d'origen natural^{5,6}. Orienta els seus treballs en el camp de l'anàlisi agroalimentària i fa que, des d'aquest moment, l'interès per l'espectroscopia NIR creixi notablement.

Fins a finals dels anys 70, els instruments comercials són bàsicament instruments de filtres, dissenyats per a aplicacions molt específiques, la qual cosa limita l'expansió de les aplicacions de la tècnica. És a partir d'aquest moment quan comencen a desenvolupar-se nous equips amb dissenys millorats i majors prestacions. En aquest desenvolupament influeixen els avenços en l'òptica i les millores en l'estabilitat dels components electrònics. Així, fan aparició els primers instruments d'espectre complet (escombrat i interferòmetres) i es construeixen els primers espectrofotòmetres que permeten l'enregistrament d'espectres de forma ràpida i altament reproducible^{7,8}.

L'altre punt fort en el que es fonamenta l'expansió de la tècnica prové dels avenços que experimenta la informàtica amb la generalització de l'ús dels ordenadors personals (*PC, Personal Computer*). Aquest fet permet posar a l'abast dels usuaris una capacitat de càlcul de la que no es disposava fins aleshores i que fa que, a partir d'aquest moment, es puguin fer servir algorismes relativament complexos per a superar un dels principals problemes de la tècnica: la no especificitat de les bandes d'absorció NIR. La generalització d'aquestes tècniques matemàtiques i la seva posterior implementació als softwares de control dels instruments permet tractar la complexa senyal NIR per a destriar-ne la informació útil per a l'aplicació a desenvolupar^{9,10}.

Aquests avenços es poden considerar els responsables que el nombre d'aplicacions NIR en els diferents camps hagi crescut enormement en els darrers anys, tal i com es reflecteix a l'article de McLure del 1994¹¹.

Al marge del camp agroalimentari, que pot considerar-se pioner, l'àmbit industrial ha incrementat el seu interès per la tècnica^{12,13}. Els resulta atractiva tant pel fet de poder mesurar mostres sòlides de forma directa, com per la rapidesa de la mesura. A més, el desenvolupament d'instruments de planta, més robustos, permet realitzar mesures *on line*, el que obre les portes al control de processos.

B.2.2. Absorció de la radiació NIR

L'anharmonicitat del moviment vibratori de l'enllaç químic fa que no es compleixin estrictament les regles de selecció. Tot i que en teoria són possibles transicions a qualsevol nivell, en realitat no és així. Com a efecte derivat de la llei de Maxwell-Boltzmann, la probabilitat que es doni una transició a nivells superiors a ± 1 disminueix ràpidament i, a efectes pràctics, només les transicions amb ± 1 , ± 2 i ± 3 són observables. Mentre la primera és coneguda com l'absorció fonamental, les altres dues reben el nom de primer (2₁) i segon (3₁) sobretons. Per a definir la regió espectral del NIR es pot recórrer a la transició IR més energètica. Aquesta es dona a 4000 cm^{-1} , amb el que la regió NIR queda establerta en l'interval $4000\text{-}14300 \text{ cm}^{-1}$ ($700\text{-}2500 \text{ nm}$). És en aquesta zona on es donen les absorcions corresponents al primer i segon sobretons, a més de l'absorció fonamental.

Cal tenir en compte, però, que alhora també són possibles els moviments vibratoris de combinació i de diferència, que presenten les mol·lècules en estat excitat i per tant són molt poc probables a temperatura ambient.

Interpretació de l'espectre NIR

L'estudi de la regió espectral NIR va ser objecte d'un treball exhaustiu durant el període 1945-1965^{14,15,16,17,18}. D'aquella època vénen la majoria dels estudis que buscaven una assignació de bandes en aquesta zona espectral. El posicionament de les bandes MIR va ser un bon punt de partida per tal d'assignar les bandes en la regió NIR.

Tal i com s'ha vist en l'apartat anterior, l'absorció NIR es posiciona en una freqüència que acostuma a ser un múltiple sencer de la banda fonamental IR. Aquesta aproximació és certa, però no té en compte l'anharmonicitat. Resulta més adequat utilitzar

$$\bar{\nu} = n_0^{-1}(1 - nx)$$

On $\bar{\nu}$ i $\bar{\nu}_0$ són els números d'ona de l'absorció NIR i fonamental, respectivament. I x , n són els termes d'anharmonicitat (habitualment 0.01) i sobretot ($n=2,3$ primer i segon).

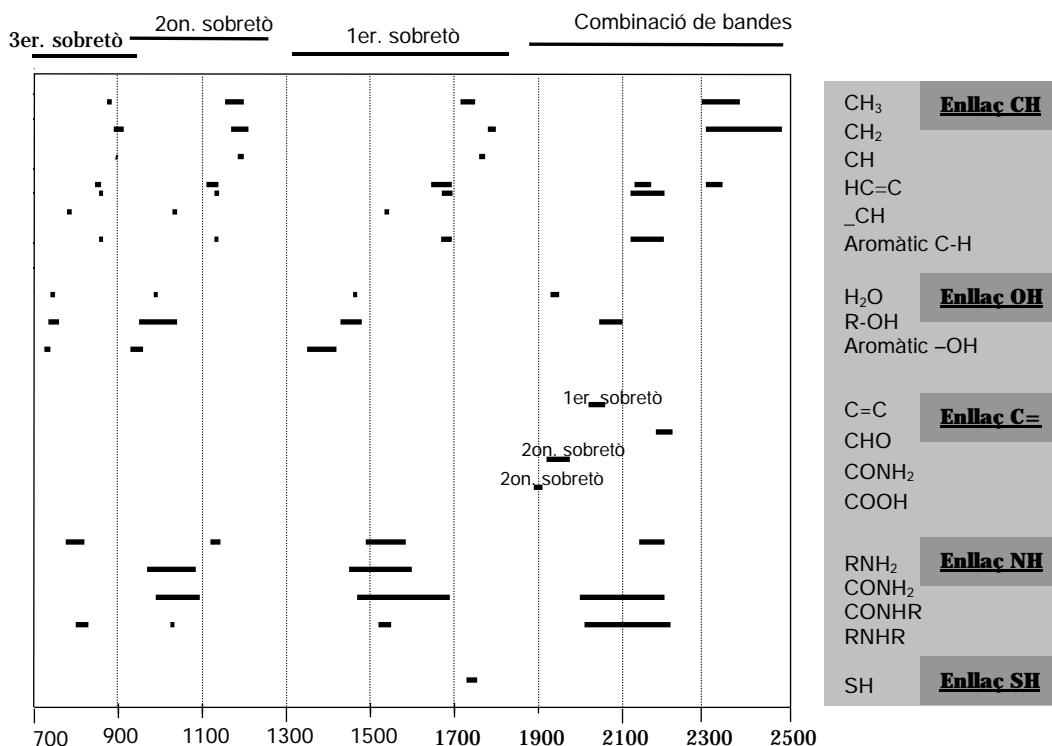
De l'exposat a l'apartat anterior es podria suposar que el nombre d'enllaços que presenten absorció al NIR és potencialment molt elevat. En realitat no és així. Només els enllaços que presenten una anharmonicitat important són susceptibles de presentar una intensitat d'absorció significativa. Entre aquests, els enllaços del tipus AH_n són els que presenten una major anharmonicitat i, per tant, una major absorció. És per això que la majoria de les absorcions a la regió NIR provenen d'enllaços que contenen un àtom d'hidrògen: CH, NH i OH. Altres enllaços que també presenten absorció en aquesta zona de l'espectre són els grups carbonil (C=O) d'aldehids, cetones, ésters i àcids carboxílics, tot i que acostumen a donar bandes relativament febles.

Les correlacions entre espectres i estructura estan documentades a diferents publicacions^{19,20}. A títol de petit resum, les més importants en anàlisi agroalimentària es donen a la Taula B.2.1.

B.2.3. Instrumentació

Els primers espectrofotòmetres comercials NIR apareixen a la dècada dels 50. Es tracta d'instruments que basen la seva arquitectura en la dels espectrofotòmetres de l'època (UV-Vis i MIR) i modifiquen només els materials de certes parts òptiques per tal d'optimitzar la resposta de l'instrument.

Així, en general, aquests instruments fan servir com a font d'emissió policromàtica una làmpada de tungstè i recorren a prismes de quarz o a xarxes de difracció com a sistemes de monocromació. Com a material per a la construcció d'altres elements òptics com les cel·les o les finestres de l'instrument, es recorre també al quarz,



Taula B.2.1. Correspondència entre els diferents enllaços i les bandes d'absorció NIR.

element transparent a la zona de l'espectre IR de longitud d'ona curta. Com a element de detecció es fan servir semiconductors com el PbS.

A partir de la configuració inicial, la instrumentació ha evolucionat en diferents direccions, tant en fonts de radiació policromàtica, com en sistemes de selecció de la longitud d'ona, en mètodes de detecció alternatius i en nous sistemes mesura.

B.2.3.1 Fonts emissores de radiació policromàtica

Actualment hi ha bàsicament dos tipus de fonts, que podem classificar com d'espectre **complet** i d'interval **reduït**.

En el primer cas, la font més utilitzada segueix sent la làmpada de Tungstè. Es tracta d'una font d'elevada intensitat d'emissió i que cobreix àmpliament la zona NIR de

l'espectre electromagnètic. Proporciona una radiació contínua en l'interval 300-2500 nm.

En el segon grup podem trobar els diodes emissors o LEDs (*Light Emission Diodes*)^{21,22,23}. Es tracta de dispositius semiconductors que emeten radiació NIR en un interval estret de longituds d'ona, el que els fa especialment aptes per a la construcció d'instrumentació portàtil per a fer mesures 'in situ'. Entre els semiconductors més utilitzats es troben les combinacions de GaAs, que emeten a l'interval espectral 900-970 nm, i les derivades de la combinació InGaAs, que permeten arribar a longituds d'ona més llargues (1600 nm).

B.2.3.2. Sistemes Monocromadors

A excepció dels instruments basats en LEDs com a font emissora, la resta necessiten de sistemes que permetin la descomposició del feix policromàtic en radiació de longitud d'ona discreta, de forma més o menys acurada. De la mateixa manera que en el cas de l'espectrofotometria MIR, aquests sistemes poden ser classificats com a **dispersius** i com a **no dispersius**.

Dins els sistemes **dispersius** es troben els monocromadors. Estan constituïts per un conjunt de colimadors dels feixos d'entrada-sortida i per un element dispersant. Aquesta és la part fonamental del sistema, ja que permet la descomposició del feix incident per l'efecte d'interaccions constructives i destructives. En instruments NIR, el més habitual és fer servir xarxes de difracció²⁴ com a elements dispersants. Per a seleccionar la longitud d'ona s'utilitza un motor pas a pas que mou la xarxa posicionant-la amb l'escaleta de sortida.

El conjunt de sistemes **no dispersius** és més ampli. Hom disposa d'instruments de filtres convencionals, filtres acustoòptics (AOTF, *Acusto-Optic Tunable Filter*) i instruments de transformada de Fourier (FT), amb característiques força diferents entre sí.

Els sistemes de filtres convencionals es basen en la interposició d'un element semitransparent entre la font de radiació policromàtica i la mostra que permet el pas de radiació d'unes longituds d'ona determinades. Es poden classificar en dos tipus: els filtres d'**absorció** i els filtres d'**interferència**. Els primers són elements elaborats a partir de materials que són transparents a determinades longituds d'ona^{25,26}; els segons són els també anomenats filtres de Fabry-Perot i es basen en la interferència òptica que experimenta el feix de llum al travessar determinats materials amb diferents índexs de refracció i gruix^{27,28,29}.

Els filtres acustoòptics (AOTF)^{30,31} aprofiten l'efecte que experimenta l'índex de refracció d'un material transparent al servir de medi de propagació d'un senyal sonor. Quan es vol preparar un filtre d'aquest tipus s'acoblen dos transductors piezoelèctrics a les cares d'un cristall birrefringent de TeO₂. Els transductors són els responsables de transformar un senyal de radiofreqüència en un senyal acústic. Quan un feix policromàtic incideix en el cristall, a l'aplicar el senyal acústic es generen dos feixos addicionals polaritzats a més del feix original. Els dos feixos polaritzats són monocromàtics i un d'ells és el que s'utilitza amb finalitats analítiques. Aquesta radiació monocromàtica depèn de la velocitat i freqüència de l'ona sonora, i de la birrefringència i mesures del cristall. Per a generar radiacions monocromàtiques de diferent longitud d'ona es fa variar el valor de la radiofreqüència inicial.

Aquests sistemes són ràpids, amb una alta reproduïbilitat de la longitud d'ona i robustos, el que els fa especialment útils en el desenvolupament d'instrumentació de planta industrial. La seva utilització no s'ha estès, principalment pel seu cost econòmic.

Els instruments NIR de **transformada de Fourier** funcionen de forma anàloga als desenvolupats per MIR, on un interferòmetre és el responsable de modular la informació espectral. Les principals diferències instrumentals provenen del propi interferòmetre. Així, a part del conegut interferòmetre de Michelson³², es troben altres sistemes que permeten la generació de l'interferograma amb dissenys òptics totalment diferents.

Un d'aquests sistemes és l'interferòmetre de prismes, on dos prismes de quars de diferent tamany estan confrontats. Es tracta d'un sistema òptic on un dels dos prismes és fix i l'altre mòbil. La radiació és polaritzada a 45° abans de passar pels prismes. En travessar-los, el feix es polaritza a 90°, amb diferents velocitats de propagació. Al trobar-se els dos prismes en moviment, el recorregut del feix de radiació és lleugerament diferent i es produeix un desfasatge entre ambdues components de polarització. En recombinar-se posteriorment, l'interferograma que s'obté dona la relació entre la intensitat radiant i la distància recorreguda pel prisma mòbil.

B.2.3.3 Detectores

Els detectors fotoconductius més emprats a l'infraroig proper són els basats en semiconductors (InGaAs, PbS, InAs, InSb, Si). Tots ells presenten intervals de resposta diferenciats. Així, per a mesures en l'interval 600-1100 nm, el detector més utilitzat és el de silici. Per a mesures en l'interval 1100-2500nm, el més emprat és el detector de PbS. La seva resposta presenta un màxim a 2000 nm que decau ràpidament fins arribar als 2700 nm, on deixa de funcionar.

En espectrofotometria NIR, un aspecte especialment important és la disposició dels sistemes detectors. Així, per a mesures de transmitància, és suficient amb situar el detector en línia amb la mostra i el feix incident. En canvi, per a mesures de reflectància, especialment en sòlids, el que es vol és captar la radiació reflectida per la mostra. En aquests casos s'acostuma a treballar amb més d'un detector i es situen aquests en posicions determinades, no alineades amb el feix incident. Alguns dels dissenys emprats són els de 0-45°, proposat per Pacific Scientific on quatre detectors es situen sobre la mostra formant un angle de 45° amb la superfície. Per a afavorir el procés d'integració del senyal, la mostra es manté en un sistema de cubeta rotatòria que gira durant la mesura. Altres dissenys, on s'augmenta el nombre de detectors i s'elimina la rotació de la cel·la amb la mostra, són petites modificacions d'aquest.

Al marge d'aquests detectors i configuracions, els detectors NIR més moderns són els Focal Point Array (FPA), que vénen a ser l'equivalent a la zona espectral NIR dels sensors CCD (*Charged Coupled Devices*)^{33,34,35}. Són detectors multicanal que, com a tals, permeten una major rapidesa en el registre i una millor relació senyal/soroll quan es comparen amb els detectors monocanal tradicionals. Aquest tipus de detectors obre les portes a l'espectroscopia d'imatges, amb la que es pot obtenir la composició de la mostra en diferents punts a través de l'espectre NIR.

B.2.3.4. Presentació de la mostra

L'espectrofotometria NIR és una tècnica espectrofotomètrica que treballa amb un tractament pràcticament nul de la mostra. Per això, donada la varietat d'estats en els que es poden presentar les mostres, ja siguin líquids de diferents viscositats o sòlids



Figura B.3. Cel·la de Transflectància acoblada a una sonda de fibra òptica.

amb diferents nivells d'aglomeració i tamany, ha estat necessari dissenyar diferents sistemes d'adquisició del senyal³⁶. Així, per a l'anàlisi de sòlids per reflectància, s'han desenvolupat sondes de fibra òptica, dispositius de cubeta (rotatòria i lineal), dispositius estàtics multidetector^{37,38}, etc...

Per a mesures de transmitància, al marge dels dissenys tradicionals, han estat desenvolupats sistemes de mesura per transflectància. Aquests dissenys situen la mostra en una cubeta de quars i posicionen una superfície reflectant a una de les cares. El feix de radiació incident entra per la cara transparent de la cubeta, travessa la mostra

i es reflecteix a l'altra cara de la cubeta. Així retorna el feix travessant de nou la mostra i la radiació reflectida és captada pel detector. El camí òptic resultant és, per tant, 2 vegades l'amplada de la cubeta de quars. Altres dissenys alternatius, com el de Bühler AG, basats en el mateix principi, fan servir sondes de fibra òptica com a elements transmissors dels feixos incident i reflectit. La reflexió es dona per una placa reflectant situada a una certa distància de la sortida de la sonda. Aquesta es posiciona de forma reproducible mitjançant un capçal especial. Aquesta configuració permet fer mesures en líquids de forma ràpida i directa.

B.2.4. Aplicacions a l'anàlisi alimentària

Com ja s'ha esmentat anteriorment, les aplicacions pioneres de l'espectrofotometria NIR van recalar en l'anàlisi de cereals, bàsicament en la determinació d'humitat, proteïna i greix. Això ha fet que, des de fa anys, hagi estat una tècnica amb una forta presència en aquest sector i que el nombre d'aplicacions desenvolupades hagi estat elevat.

Cereals i derivats

En l'anàlisi de cereals, l'espectrofotometria NIR s'utilitza en control de qualitat des de finals dels 70. Es determinen de forma rutinària propietats com la proteïna total i la humitat. Aquest fet ha incrementat els estudis sobre la viabilitat d'altres tipus de determinacions directes, com el rendiment en farina, el midó, i l'absorció d'aigua. L'expansió de l'ús d'aquestes aplicacions topa amb la sensibilitat d'aquestes calibracions a d'altres variables com la varietat del gra, la zona de conreu i la temporada de recollida. Aquests factors obliguen a freqüents recalibracions.

S'han desenvolupat altres aplicacions relacionades amb el conreu de cereals. Així, a Austràlia, la tècnica NIR s'utilitza per a determinar les necessitats de fertilitzants³⁹ dels diferents conreus a través del contingut total en Nitrogen i carbohidrats en mostres del teixit de la planta.

En algunes zones, l'ús de la tècnica per part dels propis pagesos ha proporcionat rendiments econòmics favorables. El fet de poder conèixer el contingut de proteïna en el gra de forma ràpida els ha permès separar o bé fer mescles entre grans provinents de diferents zones i que presenten diferents continguts en proteïna total. D'aquesta manera disposen d'un major nombre de tones de gra que poden ser comercialitzades amb una categoria determinada.

Diferents països productors, com Canadà i els EUA han fet de la tècnica NIR el seu mètode oficial de determinació de diferents paràmetres en gra. D'altres països Europeus han creat xarxes d'instruments que han interconnectat entre sí amb

l'objectiu de proporcionar mesures fiables, que corregeixin la calibració de les petites desviacions entre instruments i permetin l'adequada qualificació dels productes dels pagesos. Aquest procés es va iniciar el 1995 amb l'objectiu de crear la Xarxa Europea del Gra i d'alguna forma estandaritzar els sistemes de determinació per NIR.

Altres aplicacions de la tècnica relacionades amb el món dels cereals han estat desenvolupades per a molins fariners i panificadores industrials⁴⁰.

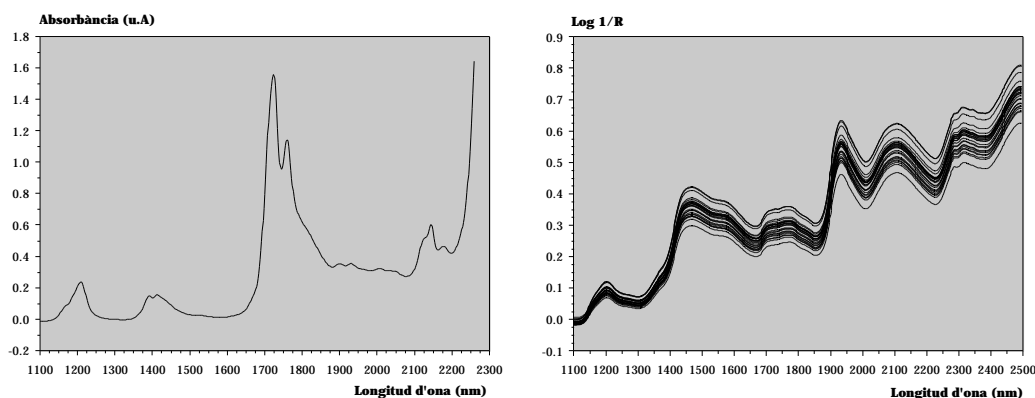


Figura B.4. Espectres NIR d'una mostra d'oli d'oliva verge enregistrada per transmitància i espectres NIR de mostres de blat de moro enregistrades per reflectància.

Productes Làctics

Les aplicacions de l'espectrofotometria NIR en aquest camp són recents. Han estat desenvolupades principalment cap el control de processos de productes com les llets líquides i en pols, les cremes de llet i els formatges⁴¹. D'aquesta manera es determinen proteïna, greix, lactosa i humitat^{42,43}.

Generalment, aquests productes són emulsions i fer-ne un mostratge adequat és complicat. D'aquí que una de les principals vies d'aplicació de la tècnica hagi estat la possibilitat de fer control 'in line' mitjançant sondes de fibra òptica⁴⁴. La utilització d'aquests sistemes de mesura, que permeten un nombre elevat de registres simultanis ha fet que sigui innecessari un mostreig previ i que l'espectre obtingut sigui igualment representatiu de la mostra.

Un bon exemple de l'expansió de la tècnica en aquest camp és la comercialització d'instruments precalibrats per a determinacions en formatges. De tota manera es tracta d'aplicacions molt específiques.

Altres

La tècnica NIR s'ha aplicat a l'anàlisi de **productes càrnics**⁴⁵ per a determinar proteïna, humitat i contingut en greixos. La seva utilització ha portat al

desenvolupament de dispositius específics d'interactància per la determinació del greix. Aquests dispositius permeten determinar aquests paràmetres tant en carns sense processar com en productes càrnics elaborats, com salsitxes, pernills, etc.

L'anàlisi del **peix** (bàsicament de salmons de granja) també presenta possibilitats favorables per a la determinació de forma no destructiva de proteïna, humitat i oli⁴⁶.

En el cas de les **fruites**, la tècnica NIR ha estat orientada cap a la determinació del contingut en sucres, com a indicador del grau de maduresa del fruit. Investigadors japonesos^{47,48} han desenvolupat equips que, basats en mesures d'interactància (variant de reflectància) a través d'una fibra òptica, discriminen de forma automàtica entre fruites de diferents nivells de maduració. Ho han aplicat a mandarines i préssecs. Seguint la mateixa línia, científics Australians han extès aquest concepte cap a les fruites tropicals com les pinyes, els mangos i els melons.

Una de les principals aplicacions de la tècnica NIR en el camp de les **begudes** ha estat la determinació del contingut alcohòlic de cerveses, amb l'ús de cel·les de transmissió o de transmissió. A més, també s'ha utilitzat en el control de matèries primeres com l'ordi, per tal de poder predir-ne el seu rendiment en malta. De tota manera hi ha dubtes de si la mesura directa sobre el gra no germinat té o no en compte tots els possibles processos enzimàtics que porten a l'obtenció de la malta. És per això que, per a augmentar la precisió de les determinacions, s'ha repartit el procés de mesura en les diferents etapes de la seva elaboració.

A part de la indústria **cervesera**, també s'utilitza en la indústria **vitivinícola**, amb la intenció de controlar el contingut en alcohol i el nivell de maduració del raïm; en la indústria de **sucs de fruita** per a determinar-ne el contingut en sucres i en **tès** i **cafès** per a determinar-ne el contingut en humitat i certes característiques organolèptiques.

Olis d'origen vegetal

Les aplicacions quantitatives de determinació de diferents paràmetres analítics basades en mesures NIR ha experimentat un creixement paral·lel al de l'espectrofotometria IR en l'anàlisi d'aquest tipus de mostres. D'aquesta forma s'ha determinat el contingut en àcids grassos *trans*⁴⁹; l'índex de Peròxids^{50,51}; l'índex de Iode⁵², i l'índex de saponificació⁵³ en diferents olis d'origen vegetal.

Autenticació

L'autenticació de productes naturals, ja sigui per assegurar-ne unes certes característiques de qualitat derivades del seu origen, com la prevenció de possibles adulteracions ha estat un altre dels focus de desenvolupament d'aplicacions NIR. El

recull de R. Downey⁵⁴ reflecteix aplicacions NIR per a l'autenticació de cafès, polpes de fruita, llets en pols, suc de taronja, sucres, arrossos, blat i farines de blat. En sí, es cobreix la necessitat genèrica de conèixer amb certesa l'origen dels productes naturals semielaborats. S'han proposat mètodes que permetin l'autenticació de productes càrnics^{55,56,57} i d'olis vegetals^{58,59,60}. En el cas particular d'aquests últims es discrimina entre olis de llavors, olis d'origens diferents i possibles adulteracions⁶¹.

- ¹ W. Herschel, *Philosophical Transactions of the Royal Society*, **90**, 255, (1800).
- ² W. Abney i E.R. Festing, *Philosophical Transactions of the Royal Society*, **172**, 887, (1881).
- ³ W. Kaye, *Spectrochimica Acta*, **6**, 257, (1954).
- ⁴ R.F. Goddu, *Advances in Analytical Chemistry and Instrumentation*, ed. Reilly, Ch. N. Interscience, New York, 374-424 (1960).
- ⁵ K.H. Norris, *Trans. Am. Soc. Agric. Eng.*, **7**, 240, (1964).
- ⁶ I. Ben-Gera i K.H. Norris, *J. Food Sci.*, **7**, 240, (1968).
- ⁷ D. Noble, *Anal. Chem.*, **67**, (1995).
- ⁸ C.M. Henry, *Anal. Chem.*, **71**, 625A, (1999).
- ⁹ H. Mark, *Analytica Chimica Acta*, **223**, 75, (1989).
- ¹⁰ K.R. Beebe i B.R. Kowalski, *Anal. Chem.*, **59**, 1007, (1987).
- ¹¹ W.F. McLure, *Anal. Chem.*, **66**, 43A, (1994).
- ¹² J. Rottetele, W. Kohn i H. Bartels, *NIR News*, **4**, 12, (1993).
- ¹³ I.B. Benson, *Spectroscopy Europe*, **7**, 18, (1995).
- ¹⁴ J.L. Lauer, E.J. Rosenbaum, *Appl. Spectrosc.*, **6**, 29 (1952).
- ¹⁵ W. Kaye, *Spectrochimica Acta*, **6**, 257, (1954).
- ¹⁶ W. Kaye, *Spectrochimica Acta*, **7**, 181, (1955).
- ¹⁷ O.H. Wheeler, *Chemical Reviews*, **59**, 629, (1959).
- ¹⁸ R.F. Goddu, *Advances in Analytical Chemistry and Instrumentation*, **1**, 347, (1960).
- ¹⁹ B.G. Osborne i T. Fearn, *Near Infrared Spectroscopy in Food Analysis*, Longman Scientific & Technical, Harlow, Essex, UK, Third Impression (2000).
- ²⁰ D.A. Burns, E.W. Ciurczak i Marcel Dekker, *Handbook of Near Infrared Analysis*, ed. D.A. Burns, E.W. Ciurczak. Marcel Dekker Inc., New York, (1992).
- ²¹ United States Patent, Pat.no. 4.286.327, (1981).
- ²² A.S. Bonano i P.R. Griffiths, *Appl. Spectrosc.*, **49**, 1590, (1995).
- ²³ A.S. Bonano i P.R. Griffiths, *Appl. Spectrosc.*, **49**, 1598, (1995).
- ²⁴ H.H. Willard, L.L. Merritt Jr., J.A. Dean i F.A. Settle Jr., *Métodos Instrumentales de Análisis*, Grupo Editorial Iberoamérica, México D.F., (1991).
- ²⁵ J.S. Schenk, J. J. Workman Jr. i M.O. Westerhaus, a *Handbook of Near-Infrared Analysis*, ed. D.A. Burns, E.W. Ciurczak. Marcel Dekker Inc., New York, cap. 15, (1992).
- ²⁶ J.T. Diffe, a *Handbook of Near-Infrared Analysis*, Marcel Dekker Inc., New York, vol. 13, cap. 16, pgs. 433-473 (1992).
- ²⁷ D.A. Skoog i J.J. Leary, *Análisis Instrumental*, 4ª ed., McGraw-Hill, Madrid, cap.6 pgs. 101-102, (1994).
- ²⁸ United States Patent, Pat. No. 3.861.788, (1973).
- ²⁹ United States Patent, Pat. No. 4.082.464, (1976).
- ³⁰ C.D. Tran, *Anal. Chem.*, **64**, 971A, (1992).
- ³¹ P.J. Treado, I. W. Levin i E.N. Lewis, *Appl. Spectrosc.*, **46**, 553, (1992).
- ³² H.H. Willard, L.L. Merritt Jr., J.A. Dean i F.A. Settle Jr., *Métodos Instrumentales de Análisis*, Grupo Editorial Iberoamérica, México D.F., (1991).
- ³³ J.V. Sweedler, R.B. Bilhorn, P.M. Epperson, G.R. Sims i M. B. Denton, *Anal. Chem.*, **60**, 282A, (1988).
- ³⁴ P.M. Epperson, J.V. Sweedler, R.B. Bilhorn, G.R. Sims i M. B. Denton, *Anal. Chem.*, **60**, 327A, (1988).
- ³⁵ Q.S. Hanley, C.W. Earle, F.M. Pennebaker, S.P. Madden i M.B. Denton, *Anal. Chem.*, **68**, 661A, (1996).
- ³⁶ *Journal of Near Infrared Spectroscopy. The proceedings of NIR-97*. NIR Publications, vol.6, nºs 1-4, Chichester, (1998).
- ³⁷ *Installation Manual for NIRSystems Scanning Spectrophotometers* (v.6.2), FOSS NIRSystems, (1996).
- ³⁸ J.R. Ferraro, K. Martin i R.J. Jarnutowski, *Spectroscopy*, **12** (8), 23, (1997).
- ³⁹ V.B. McGrath, A.B. Blakeney i G. D. Batten, *New Phytol.*, **136**, 145, (1997).
- ⁴⁰ B.G. Osborne, T. Fearn, A.R. Miller i S. Douglas, *J. Sci. Food Agric.*, **35**, 99, (1984).
- ⁴¹ J. L. Rodriguez-Otero, M. Hermida i J. Centeno, *J. Agr. Food Chem.*, **45**, 2815 (1997).
- ⁴² M. F. Laporte i P. Paquin, *J. of Agr. Food Chem.*, **47**, 2600, (1999).
- ⁴³ M. J. Adams, K. Latham, N. W. Barnett i A. J. Poynton, *J. Sci. Food Agr.*, **79**, 1232, (1999).
- ⁴⁴ J. Y. Chen, C. Iyo i S. Kawano, *J. Near Infr. Spectrosc.*, **7**, 265, (1999).
- ⁴⁵ T. Isaksson, B. N. Nilsen, G. Togersen, R. P. Hammond i K. I. Hildrum, *Meat Sci.*, **43**, 245, (1996).
- ⁴⁶ J. P. Wold i T. Isaksson, *J. Food Sci.*, **62**, 734, (1997).

-
- ⁴⁷ S. Kawano, H. Watanabe i M. Iwamoto, *J. Japan Soc. Hort. Sci.*, **61**, 445, (1992).
- ⁴⁸ S. Kawano, T. Fujiwara i M. Iwamoto, *J. Japan Soc. Hort. Sci.*, **62**, 465, (1993).
- ⁴⁹ H. Li, F. R. van de Voort, A. A. Ismail, J. Sedman i R. Cox, *J. Am. Oil Chem. Soc.*, **77**, 1061, (2000).
- ⁵⁰ H. Li, F. R. van de Voort, A. A. Ismail i R. Cox, *J. Am. Oil Chem. Soc.*, **77**, 137, (2000).
- ⁵¹ M. H. Moh, Y. B. Che Man, F. R. van de Voort i W. J. W. Abdullah, *J. Am. Oil Chem. Soc.*, **76**, 19 (1999).
- ⁵² H. Li, F. R. van de Voort, A. A. Ismail, J. Sedman, R. Cox, C. Simard i H. Buijs, *J. Am. Oil Chem. Soc.*, **77**, 29, (2000).
- ⁵³ H. Li, F. R. van de Voort, J. Sedman i A. A. Ismail, *J. Am. Oil Chem. Soc.*, **76** (4), 491, (1999).
- ⁵⁴ G. Downey, *J. Near Infrared Spectrosc.*, **4**, 47, (1996).
- ⁵⁵ H. B. Ding, R. J. Xu, *J. Agr. Food Chem.*, **48**, 2193 (2000).
- ⁵⁶ J. McElhinney, G. Downey i C. O'Donnell, *J. Food Sci.*, **64**, 587 (1999).
- ⁵⁷ H. Rannou i G. Downey, *Anal. Comm.*, **34**, 401 (1997).
- ⁵⁸ K. M. Bewig, A. D. Clarke, C. Roberts i N. E. Unklesbay, *J. Am. Oil Chem. Soc.*, **71**, 195, (1994).
- ⁵⁹ G. Downey, *The Analyst*, **119**, 2367, (1994).
- ⁶⁰ T. Sato, *J. Am. Oil Chem. Soc.*, **71**, 293, (1994).
- ⁶¹ I. J. Wesley, R. J. Barnes i A. E. J. Mc Gill, *J. Am. Oil Chem. Soc.*, **72**, 289, (1995).