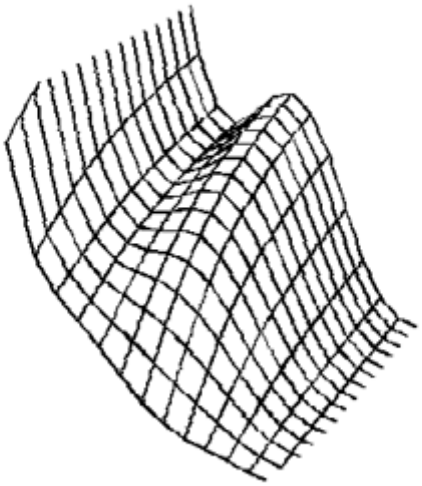
Universitat Autònoma de Barcelona

DESENVOLUPAMENT DE PROCEDIMENTS CINÈTICS PER L'ANÀLISI DE MULTICOMPONENTS



Núria Villegas Forn

DESENVOLUPAMENT DE PROCEDIMENTS CINÈTICS PER L'ANÀLISI DE MULTICOMPONENTS

Memòria presentada per Núria Villegas i Forn
per aspirar al grau de Doctora en Químiques

NÚRIA VILLEGAS FORN

Departament de Química
Unitat de Química Analítica
Edifici Cn
08193 Bellaterra (Barcelona), Spain

Universitat Autònoma de Barcelona

Dr Santiago Maspoch Andrés, Catedrático de Química Analítica de la Facultad de Ciencias de la Universidad Autónoma de Barcelona,

CERTIFICA:

Que el presente trabajo de investigación titulado **"Desenvolupament de procediments cinètics per l'anàlisi de multicomponents",** que constituye la Memoria presentada por Núria Villegas Forn para aspirar al grado de doctor en Químicas, ha sido realizado integramente en los laboratorios de la Unidad de Química Analítica, departamento de Química,de la Universidad Autónoma de Barcelona, bajo mi dirección; reuniendo a mi juicio, las condiciones exigidas en este tipo de trabajo.

Y para que quede constancia, expide y firma el presente certificado en Bellaterra, a 18 de diciembre de 2002.

Dr Santiago Maspoch

ÍNDEX

CAPÍTOL II: DETERMINACIÓ D'ACETAT D'HIDROCORTISONA

CAPÍTOL III: DETERMINACIÓ DE MESCLES DE LEVODOPA I BENSERAZIDA

CAPÍTOL IV: DETERMINACIÓ DE MESCLES TERNÀRIES D'ALDEHIDS PER REACCIÓ AMB MBTH

OBJECTE I CONTINGUT DE LA TESI

PRESENTACIÓ

La resolució d'una mescla a partir de la mesura directa del seu espectre i un adequat tractament quimiomètric és avui en dia una tècnica analítica ben establerta. Aquests mètodes resulten ser una bona eina per la seva gran simplicitat, no obstant això, es troben limitats a causa de la similitud espectral existent entre les espècies químiques d'estructura similar.

És en aquest context, que tenen gran interès els mètodes cinètico-espectrofotomètrics diferencials basats en la velocitat de reacció. Malgrat els mètodes basats en l'equilibri siguin més fàcils d'establir, la selectivitat dels cinètics és molt superior, ja que permeten diferenciar la contribució de cada un dels analits, en base a la diferent velocitat de reacció amb un mateix reactiu. A més, malgrat en molts casos només s'aprofiti la diferent velocitat de reacció, també es pot usar la diferencia espectral si la reacció es registra a varies longituds d'ona, obtenint així més informació discriminant per resoldre el sistema.

Però a pesar de la seva gran potencialitat, els mètodes cinètics de multicomponents no han estat utilitzats de forma important en la resolució de sistemes reals, restringit-se més a l'estudi teòric/acadèmic de sistemes.

Una de les causes que ha provocat això han estat les eines matemàtiques utilitzades pel tractament de les dades fins fa pocs anys. En la gran majoria d'elles és necessari el coneixement

del mecanisme cinètic que segueixen les diferents reaccions implicades així com de les constant de velocitat, les quals s'obtenen a partir de les reaccions de cada analit per separat. Això restringeix la seva aplicació a sistemes cinètics simples, on les reaccions implicades segueixen una cinètica de primer o pseudoprimer ordre, i sempre en absència d'interaccions entre analits.

A més, s'utilitzen poques mesures per a realitzar els càlculs, que en molts casos estan basats en mètodes gràfics. Evidentment, això comporta una imprecisió considerable en la predicció de les concentracions i especialment quan les constants de velocitat són molt pròximes.

Com a resultat, quan s'utilitzen els mètodes de càlcul tradicionals per resoldre els sistemes cinètics, només un nombre molt reduït de reaccions químiques poden ser aplicades amb finalitat analítica, i a més, també es molt reduït el nombre d'espècies que poden ser analitzades atès que es requereixen grans diferències en les velocitats de reacció.

En els últims anys s'han desenvolupat gran quantitat d'eines quimiomètriques per tal d'extreure la màxima informació química a partir de l'anàlisi de les dades experimentals obtingudes per gran quantitat de tècniques analítiques. D'aquesta manera, el desenvolupament dels mètodes quimiomètrics de calibració multivariable en diferents àmbits, està fent possible també el desenvolupament dels mètodes cinètics d'anàlisi. En són un bon exemple els mètodes lineals basats en la reducció de variables, com ara la Regressió Parcial per Mínims Quadrats (PLS), així com mètodes de regressió no lineals com són les Xarxes Neuronals Artificials (ANN), els quals es poden considerar procediments generals de calibració per a la determinació simultània de varis analits, ja que permeten modelar el sistema sense assumir prèviament el model cinètic que segueixen les reaccions químiques, ni conèixer els valors de les constants de velocitat. A més, es poden obtenir bons resultats inclús quan hi ha present una clara no linealitat provinent de la interacció entre els components d'un sistema.

CONTINGUT DE LA TESI

L'objectiu d'aquesta memòria és el desenvolupament de procediments cinètics d'anàlisi de mescles d'espècies químiques, que es troben en mostres comercials de naturalesa i complexitat variada, basats en la utilització de les tècniques de calibració multivariable. La finalitat de la seva aplicació és la obtenció de les concentracions de cada un dels components de la mescla, sense separació

prèvia, demostrant d'aquesta manera l'aplicabilitat real d'aquests procediments.

Enumerant els sistemes estudiats de forma cronològica a la seva realització, així com en complexitat, en els dos primers s'analitzen dos preparats farmacèutics comercials els quals contenen com a principis actius dues espècies amb estructures i propietats molt similars. En el tercer s'analitzen detergents desinfectants comercials, els quals estan composats per mescles ternàries dels aldehids; formaldehid, glioxal glutaraldehid. En tots aquests sistemes, les reaccions seguides presenten mecanismes coneguts.

En el quart i últim s'analitzen diferents preparats farmacèutics utilitzats en la pràctica veterinària, els quals estan constituïts per mescles ternàries de diferents sulfonamides. En aquest darrer sistema, es desconeix el mecanisme cinètic exacte que segueixen les diferents espècies.

La metòdica seguida en tots els treballs realitzats ha estat sempre la mateixa. Primer s'intenten trobar les condicions experimentals més adequades que permetin obtenir una major discriminació cinètico-espectral entre els diferents analits que composen la mostra. Llavors es preparen al laboratori un conjunt de mescles que contenen els analits d'interès en diferents concentracions i es construeixen els models de calibració que ajustin bé la informació proporcionada pel registre cinètic. Per la construcció dels models de calibració s'han utilitzat diferents mètodes d'anàlisis multivariable com són la regressió en Components Principals (PCR), la regressió Parcial per Mínims Quadrats (PLS), la regressió Parcial per Mínims Quadrats Multi Via (nPLS) i les Xarxes Neuronals (ANN). En cada cas s'intenta trobar el mètode de calibració més adient basant-nos d'entrada en la descripció i àmbit d'aplicació de cada mètode, i en l'anàlisi dels resultats obtinguts pel conjunt dels analits.

Un cop construïts els models de calibració, s'analitzen les mostres reals objecte d'anàlisi, la naturalesa de les quals, com ja s'ha esmentat, és força variada.

Malgrat la complexitat dels sistemes químics estudiats, especialment els realitzats en els últims treballs de la tesi, ha estat possible quantificar els diferents analits amb bona exactitud i precisió.

Fent una descripció més en detall dels diferents capítols que constitueixen aquesta tesi, ens trobem primer un exemple on s'aplica un mètode cinètic en la quantificació d'un únic analit, l'acetat de hidrocortisona. El mètode es basa en una reacció de condensació d'aquesta espècie amb àcid isonicotínic hidrazida, que segueix una cinètica de pseudoprimer ordre. En la mostra que s'analitza també hi ha present altres espècies, concretament hi ha la prednisona que presenta una

estructura molt propera a l'espècie que s'analitza. Però en les condicions d'anàlisi escollides aquestes espècies no contribueixen significativament al senyal mesurat. Per tant es tracta doncs d'un cas relativament senzill a primer cop d'ull, i potser si que ho és. Però l'objectiu d'aquest primer treball ha estat posar de manifest la superioritat dels mètodes de calibració multivariables front els mètodes més clàssic d'anàlisi. L'utilització conjunta d'informació cinètica i espectral mesurada per un número gran de longituds d'ona, millora l'exactitud i també la precisió del mètode cinètic ja que es suavitza l'efecte de petites variacions de les constants de velocitat entre experiments, produïdes per petits canvis en les condicions experimentals.

En el capítol 3 s'estudia un sistema més complicat, composat per una barreja de levodopa i benserazida. Inicialment el sistema s'aborda per mitjà d'un mètode espectroscòpic basat en el registre dels seus espectres, el qual permet l'obtenció de bons resultats. Posteriorment es vol veure si és possible resoldre el sistema per mitjà d'un mètode cinètic diferencial d'anàlisi. La reacció escollida és la oxidació dels dos compostos amb un excés de periodat sòdic. En les condicions de treball tenim que la cinètica de reacció no és tant simple com abans, ja que mentre que per un cantó la levodopa s'oxida seguint una cinètica de primer ordre, la benserazida es desvia d'aquest comportament. A més a més, s'ha pogut veure la presència d'una reacció creuada, amb aparició d'espècies no presents en la reacció per separat.

Un altre fet a destacar d'aquest capítol, és que s'inicia l'aplicació d'una nova metodologia de calibració, la regressió parcial per mínims quadrats multi via (nPLS), més adient pel tractament de dades tridimensionals, y es compara amb el PLS convencional. Malgrat la complexitat del sistema plantejat, aquest es resolt molt bé.

A continuació, es passa segons el programa inicial de la tesi, a l'estudi d'un sistema més complex composat per una mescla de tres analits. Es tracta de mescles de formaldehid, glioxal y glutaraldehid, utilitzades en formulacions de detergents comercials per la desinfecció hospitalària. La complexitat és deguda a la gran similitud entre els analits, així com a la matriu de la mostra, la qual conté tensioactius i colorants dels quals es desconeix tant la seva naturalesa com el seu contingut.

L'estudi del sistema es materialitza en dos treballs diferents, basats ambdós en la reacció de condensació d'aquestes especies amb la 3-metil-2-benzotiazolona hidrazona (MBTH) en medi àcid. Els dos procediments cinètics que es plantegen volen ser dues alternatives vàlides en la recerca de les condiciones experimentals més favorables per l'anàlisi d'aquestes mescles per mitjà

d'un mètode cinètico-espectrofotomètric diferencial.

En el primer treball, la mostra es dilueix per evitar interferències per part de les altres especies contingudes en la mostres, y es treballa en lleuger excés de reactiu ja que aquest absorbeix en la regió espectral d'interès. Això fa que el mecanisme cinètic que segueixen les reaccions sigui de segon ordre.

En el segon treball s'analitza la mostra directament, eliminant la dilució, per tal de facilitar al màxim el treball d'anàlisi. En aquesta situació el reactiu es troba en gran defecte en el medi de reacció, cosa que fa que les reaccions segueixin un mecanisme cinètic de pseudoprimer ordre respecte al reactiu. A més, es detecten interferències de la matriu, atribuïdes a la presència del tensioactiu, cosa que fa que per la calibració s'hagi d'utilitzar un sistema mixte, format per mescles dels aldehids per tal de cobrir un interval ampli de concentracions, i per mostres reals per tal d'incorporar l'efecte de la matriu en el calibrat.

El resultat de tot l'estudi, és l'obtenció de dues situacions químiques totalment diferents i pronunciadament no lineals. Per aplicació del mètode de calibració multivariable més adequat en cada cas, s'assoleixen resultats molt bons en les dues situacions.

Arribats a aquest punt, finalment ens trobem en condicions de passar a determinar un sistema d'extrema complexitat. Es tracta de l'anàlisi de mescles de fins a quatre sulfonamides a partir de la reacció de azo-copulació oxidativa amb MBTH en presència de Fe(III). Les sulfonamides escollides són el sulfatiazol, la sulfadiazina, la sulfamerazina i la sulfametazina, totes elles utilitzades conjuntament en preparacions comercials per ús veterinari en el tractament d'infeccions urinàries.

En aquest darrer treball, a part de la gran similitud estructural entre els analits, tenim que el reactiu format in-situ es consumeix en processos laterals a una velocitat molt alta, competitiva amb la reacció analítica d'interès. Això fa que ens trobem lluny de les condicions de pseudoprimer ordre respecte als analits.

El resultat és un mecanisme de reacció complex i molt depenent de les condicions experimentals que ens torna a conduir cap a un sistema subjecte a fortes no-linealitats. A més, tenim que el comportament per parelles de sulfonamides és molt igual, existint poques diferències entre el seu senyal.

Per aquests motiu en una primera etapa del treball es vol aprofundir en l'estudi de la influencia de les variables experimentals. L'objectiu és la recerca de les condicions més idònies,

les quals permetin una major discriminació espectral i cinètica entre les especies. Aquest estudi exigeix la realització d'un disseny experimental per registrar el comportament de cada sulfonamida per separat sota les variables considerades més influents.

La segona part del treball la constitueix el registre cinètic de les mescles sota la situació de màxima discriminació, i el tractament de les dades obtingudes a fi d'obtenir el millors resultats possibles.

Com en els altres treballs, aquest es resolt bé per aplicació dels diferents mètodes d'anàlisi multivariable.

A la taula I es mostren de forma resumida les característiques dels sistemes cinètics estudiats així com els objectius buscats en cada un d'ells. Dins dels objectius, es tracten aspectes pràctics importants associats a la quantificació de les mescles per aplicació dels mètodes multivariables, els quals permeten la correcte aplicació d'aquests algoritmes, contribuint en el seu desenvolupament quan són aplicats en sistemes cinètics reals.

Taula I. Anàlisi de mostres reals. Descripció dels treballs realitzats

Analits	Reacció analítica	Condicions i sistema	Algoritmes utilitzats	Objectius buscats
Acetat d'hidrocortisona	Condensació amb isoniàcid (INH)	•Cinètica 2n ordre, amb excés de reactiu •Presència d'un altre cortisoesteroide que reacciona lentament	PLS	-Quantificació, i demostrar l'aplicabilitat dels mètodes cinètics/calibració multivariable utilitzats
Levodopa i benserazida	Oxidació amb periodat sòdic	•Levodopa: cinètica 1r ordre •Benserazida: cinètica complexa. Formació ràpida d'un producte que desapareix •Formació de productes de reacció creuada	PLS, nPLS	-Quantificació, i demostrar el bon funcionament dels mètodes de calibració multivariables davant de les desviacions de la linealitat presents -Aplicació del nPLS i comparar-lo amb el PLS convencional
Formaldehid, glioxal i glutaraldehid	Formació imina amb 3-metil-2-benzotiazolona hidrazona (MBTH)	•Cinètica 2n ordre. Lleuger excés de reactiu •Matriu complexa: elevada concentració detensioactius i colorants que porta a una gran dilució	PLS	-Recerca del sistema experimental més favorable per l'anàlisis cinètic -Quantificació, i demostrar el bon funcionament dels mètodes de calibració multivariables en els diferents sistemes cinètics no-lineals resultants
		•Anàlisi de la mostra directament •Cinètica de pseudoprimer ordre respecte reactiu •Apareix un efecte matriu que es corregeix amb la construcció d'una calibració mixta	PLS ANN	
Sulfatiazol, Sulfadiazina, Sulfamerazina, Sulfametazina	Azocopulació oxidativa amb MBTH i Fe(III), HCl	•Cinètica complexa a causa de la ràpida descomposició del reactiu, que fa que el sistema estigui subjecte a no-linealitats. L'analit mai arriba a reaccionar completament •Gran similitud en senyal per parelles de sulfonamides	PLS nPLS ANN	-Recerca de les condicions òptimes per aplicació d'un disseny experimental, per obtenir la major discriminació entre els senyals de cada analit -Quantificació, i demostrar el bon funcionament dels mètodes de calibració multivariables en una situació d'elevada complexitat; cinètica desconeguda subjecte a fortes no-linealitats

CAPÍTOL I

INTRODUCCIÓ

En el transcurs d'aquesta tesi es presenten diferents treballs, en els quals s'utilitzen tècniques quimiomètriques de calibració multivariable pel tractament de les dades obtingudes en l'anàlisi de mescles per aplicació de mètodes cinètics. L'evolució dels mètodes cinètics moderns, va molt lligada al desenvolupament dels mètodes quimiomètrics. Aquests últims permeten tractar l'enorme quantitat de dades que són registrades amb pocs segons, i convertir-ho amb informació útil.

Per la importància compartida que tenen per l'èxit dels treballs desenvolupats en aquesta memòria, en aquests primer capítol s'introduiran en primer lloc les tècniques quimiomètriques, fent especial èmfasi a les utilitzades en els diferents treballs, i a continuació es presentaran els mètodes cinètics d'anàlisis.

I.1. CALIBRACIÓ MULTIVARIABLE

La calibració és un dels passos més importants en l'anàlisi químic. Només és possible obtenir una bona exactitud quan s'utilitza un bon procediment de calibració. Excepte en casos molt concrets (p.ex. la gravimetria), la concentració d'un analit no es pot mesurar directament, sinó que es mesura a partir de la magnitud d'un senyal. La condició es que existeixi una relació entre aquesta magnitud del senyal i la concentració de l'analit.

Malauradament, aquests senyals analítics, en la majoria dels casos, no son exclusius d'una única espècie. Per exemple, l'espectre d'una mostra serà la suma de les contribucions a l'absorbància a cada longitud d'ona de cada un dels components que la formen, a part de les interaccions que puguin existir entre ells. A fi que el senyal analític depengui únicament de l'analit es poden separar físicament els components d'una mostra com a pas previ al registre del senyal, o bé emmascarar les substancies interferents que acompanyen a l'analit. Tradicionalment l'anàlisi de mostres complexes només es podia realitzar per aplicació de mètodes potents de separació, i això explica el gran desenvolupament que han experimentat les modernes tècniques cromatogràfiques, mitjançant les quals, un cop separats els diferents components es pot registrar un senyal analític degut a cada component individual. Però un inconvenient d'aquests processos és que són lents i no permeten la realització d'un gran nombre d'anàlisis en un breu període de temps, a més, requereixen el consum de grans quantitats de dissolvents orgànics força cars i a la

vegada contaminants.

Les tècniques instrumentals, i especialment les espectroscòpiques, generen en poc temps una gran quantitat de dades relatives a les mostres analitzades. Però això no vol dir que quantes més dades s'obtinguin més informació es té del sistema. Tal i com van dir Beede i Kowalski al 1987 [Beebe, K.R., 1987], només quan les dades son interpretades i utilitzables es converteixen en valuoses pels químics i per la societat en general; aleshores les dades es converteixen en informació. És en aquest punt on la quimiometria té un paper decisiu.

La quimiometria, la qual es pot definir com la part de la química que, utilitzant mètodes matemàtics, estadístics i de lògica formal; a) dissenya o selecciona procediments de mesura òptims, i b) proporciona la màxima informació rellevant de les dades analítiques [Massart, D.L., 1988], és una disciplina relativament nova en l'àrea de la química, nascuda aproximadament al final dels 70 [Wold, 1972; Wold, 1974; Kowalski, 1975; Kowalski, 1987], que els darrers anys ha penetrat amb força en diversos camps, especialment els de la química analítica [Sharaf, 1986], la química de síntesi [Carlson, 1992], i l'enginyeria química [Kresta, 1991; Wise, 1990]. Han sorgit també disciplines similars a la quimiometria en altres àrees de les ciències socials, com la psicometria [Kroonenberg, 1983] i l'economètria [Christ, 1966]. El desenvolupament ràpid de la quimiometria en química analítica es deu fonamentalment a la utilització massiva dels ordinadors acoblats a la moderna instrumentació. S'ha passat de mesures puntuals univariables d'un sol canal o longitud d'ona, a mesures espectrals multivariables multicanal que tenen un gran contingut informatiu. Els mètodes tradicionals d'anàlisi i procés de dades resulten totalment insuficients i l'embut es troba actualment en l'etapa de l'anàlisi de dades instrumentals.

La utilització dels mètodes quimiomètrics permet, entre altres coses, la identificació de mostres, l'anàlisi de mescles complexes sense la necessitat de separacions prèvies, l'augment de l'interval de concentracions de treball, la possibilitat de determinar simultàniament varis analits, l'augment de la selectivitat respecte als mètodes convencionals, etc. Les principals avantatges que deriven de tot això són un coneixement més ampli del problema i la possibilitat d'una alta velocitat d'anàlisi, cosa que permet reduir costos i temps d'anàlisi de forma important. Tot això ha produït un gran augment dels articles relacionats amb la quimiometria, fins al punt de constituir un apartat propi en les revisions bianuals de *Analytical Chemistry*, i l'aparició de revistes especialitzades en el tema, com *Journal of Chemometrics* o *Chemometriccs and Intelligent Laboratory Systems.*

Entre la gran varietat de tècniques quimiomètriques existents, les aplicades en el desenvolupament de mètodes cinètico-espectrals d'anàlisi portats a terme en aquesta memòria es situen dins de:

- *El disseny d'experiències*. En tots els treballs s'ha aplicat un disseny experimental inicial amb l'objectiu de maximitzar i facilitar l'extracció d'informació significativa de les dades generades.

- *L'analisi quantitatiu*. S'han utilitzat tècniques de calibració diferents, escollint unes o altres segons les necessitats de cada cas. Les tècniques emprades han estat tècniques de reducció de variables, tant algoritmes bilineals com els multi via, i les xarxes neuronals artificials.

I.1.1. ANÀLISI QUANTITATIU. MÈTODES DE CALIBRACIÓ MULTIVARIABLE

La calibració és el procés que permet establir la relació entre la resposta instrumental i la quantitat d'espècie química o propietat que es desitgi quantificar. L'equació matemàtica que relaciona el senyal analític amb la concentració s'anomena model o equació de calibració, i la representació gràfica que els relaciona rep el nom de corba de calibració.

En el procés de calibració es poden distingir les següents etapes:

(a) *Preparació d'un conjunt de calibració o conjunt d'entrenament (training set)*. Obtenció d'un conjunt limitat de mostres de composició coneguda que sigui representatiu de tot l'interval de concentracions en el que es vol treballar, així com de les possibles interferències i altres espècies presents en las mostres, encara que no es vulguin determinar.

(b) *Registre de la informació analítica*. La informació pot provenir de fonts molt diverses, que en el cas particular d'aquesta memòria és espectrofotomètrica, obtinguda a partir del seguiment de l'evolució dels espectres d'absorció UV-Vis. A partir d'aquest senyal instrumental s'obtindrà la informació química desitjada.

(c) *Pretractament de les dades.* En aquesta etapa es minimitzen les contribucions no desitjades presents en el senyal analític, que disminueixen la reproduïbilitat i poden provocar que el sistema presenti comportaments que donarien lloc a estimacions errònies dels paràmetres desitjats.

(d) *Construcció del model.* Selecció del model més senzill possible que estableix la relació entre el senyal analític i la variable resposta desitjada. El model pot tenir una base totalment empírica o bé teòrica que expliqui el fenomen físic o químic responsable del senyal analític. En moltes situacions, implica un estudi complexe de les matrius de dades obtingudes, ja que no sempre és necessària la utilització de tota la informació registrada. La optimització es realitza assajant diferents algoritmes, pretractaments matemàtics, intervals de longitud d'ona, etc.

(e*) Validació del model.* Aplicació del model establert a un número limitat de mostres de les què

es coneix la propietat a determinar, i que no han estat utilitzades en l'etapa de construcció del model. D'aquesta forma es verifica que el model construït constitueix una correcta descripció del sistema en estudi.

(f) *Predicció de noves mostres.* Utilització del model calculat i validat per a predir la propietat a determinar, en el nostre cas la concentració de l'analit, en noves mostres de les què s'ha registrat prèviament el senyal analític.

El tipus de calibració més senzilla es la univariable, en la que es relaciona només una variable dependent i una independent (per exemple, la absorció a una longitud d'ona amb la concentració d'analit); els seus principis son àmpliament coneguts [Miller, 1988], motiu pel qual no es farà aquí cap comentari més al respecte. En la calibració multivariable intervenen més d'una variable dependent i/o independent (p.ex. la absorció a varies longituds d'ona i la concentració de varis analits). La seva utilització és necessària quan es vol quantificar un o varis analits en una mostra i no hi ha variables específiques.

I.1.1.1. CLASSIFICACIÓ DELS MÈTODES DE CALIBRACIÓ

S'han proposat diverses classificacions dels mètodes de calibració. Una d'elles és la presentada per Booksh i Kowalski [Booksh, 1994] en la seva teoria de química analítica: els autors classifiquen els mètodes de calibració des del punt de vista matemàtic, en funció de la dimensió de les dades disponibles per mostra. Així, les dades poden ser classificades com:

(a) Dades univariables:

- *Dades escalars. Instruments d'ordre zero.*

El tipus més senzill de dades en química és de naturalesa univariable, és a dir, mesures en les que únicament s'obté un valor numèric, un escalar. Aquesta situació es dona en molts casos, per exemple, en les mesures potenciomètriques de pH o en les mesures espectrofotòmetriques realitzades a una única longitud d'ona. El mètode matemàtic utilitzat en l'extracció de la informació analítica és la estadística univariable i la regressió lineal univariable. L'aplicació

d'aquests mètodes suposa que la resposta instrumental és conseqüència de la presència d'un únic component químic, és a dir, que no hi ha interferències ni contribucions desconegudes al soroll de fons en el senyal instrumental.

(b) Dades multivariables:

- *Dades vectorials. Instruments de primer odre.*

L'instrument proporciona un vector de dades (p.ex. l'espectre) a l'analitzar cada mostra. Els mètodes matemàtics utilitzats són els derivats de l'anàlisi multivariable de les dades i de l'estadística multivariable [Martens, 1989; Sekulink, 1993]. En aquest cas no és necessari modelar explícitament les interferències ni el soroll de fons abans d'extreure la informació química rellevant, però si que és necessari que els patrons siguin de la mateixa naturalesa que les mostres estudiades. Quan el problema conté interferències no presents en els patrons, els mètodes de calibració multivariable tampoc donen resultats satisfactoris, però si que permeten la detecció d'aquestes interferències. Les no-linealitats produïdes per exemple per efectes de matriu, poden ser corregides escollint el mètode de calibració no-lineal més adequat a les dades disponibles.

- *Matriu de dades. Instruments de segon ordre.*

En aquests cas, per a cada mostra analitzada s'obté una matriu de dades. En aquesta matriu es consideren dues direccions, les files i les columnes, que es corresponen, generalment, a dos tipus diferents de mesura. Aquest és el cas que es dóna quan es combinen les mesures d'emissió fluorescent a diferents longituds d'ona d'excitació; o quan s'efectua un seguiment de resposta de tipus multivariable amb el temps, pH o concentració.

La construcció d'un model de calibració amb dades de segon ordre es pot portar a terme de manera similar al de primer ordre si s'efectua un desdoblament de les dades (*unfolding*), de tal manera que per a cada mostra es tingui un tensor de primer ordre (figura 1.1). En aquestes situacions es mescla la informació d'ambdós ordres i tant sols es poden aprofitar les avantatges de la calibració de primer ordre. Quan es manté l'estructura tridimensional de les dades (per exemple temps - longitud d'ona - mostra) la calibració és de segon ordre. En aquest cas es poden utilitzar els anomenats mètodes d'anàlisi de tres vies (*three-way data analysis*) [Kruskal, 1989].

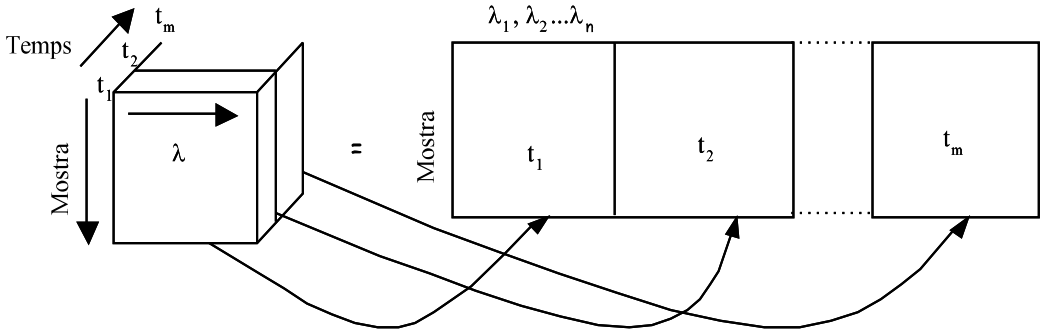


Figura 1.1. Desdoblament d'una estructura tridimensional de dades.

- *Dades d'ordre superior. Instruments d'ordre superior.*

L'extensió dels conceptes anteriors ens condueix als instruments d'ordre superior i a les dades obtingudes amb ells. No hi ha límit en el màxim ordre de les dades que poden ser obtingudes. Els instruments moderns d'espectroscòpia de fluorescència poden proporcionar la variació dels espectres excitació-emissió amb el temps, originant una estructura tridimensional de les dades per mostra. Una avantatge afegida d'utilitzar instruments d'ordre superior és l'augment de la selectivitat. També en aquest cas es poden aplicar els mètodes d'anàlisi de N vies.

La classificació dels mètodes de calibració que s'ha descrit, no és però l'única que s'utilitza. Martens proposa unes altres classificacions basades en conceptes metodològics [Martens, 1989]:

(a) Com s'obtenen els paràmetres de la calibració:

- *Calibració directa*: els paràmetres de la calibració es calculen directament a partir del senyal de cada un dels analits de forma individual.

- *Calibració indirecta*: els paràmetres es calculen a partir dels senyals analítics de mescles dels components. Aquest tipus de calibració s'utilitza quan no és possible obtenir un calibrat amb el senyal analític de l'analit aïllat, degut a la presència d'interferències, o bé quan no és possible registrar el senyal aïllat per problemes d'estabilitat.

(b) Segons les variables que es relacionen:

- *Calibració lineal*: relaciona les variables dependents amb funcions lineals de les variables independents, o bé com funcions polinòmiques que són lineals en els coeficients. Els models lineals són els que es poden descriure com:

$$y = b_0 + \sum_{k=1}^{k} b_k x_k \qquad \textbf{(1)}$$

essent b_0, b_k els paràmetres a determinar, y la variable dependent, i x les variables independents.

- *Calibració no lineal*: models no lineals en els paràmetres.

(c) Magnitud utilitzada com a variable independent:

- *Calibració clàssica*: la variable independent és la concentració del analit.

- *Calibració inversa*: s'utilitza la concentració com a variable dependent i el senyal analític com a variable independent.

(d) Número de variables que utilitza:

- *Calibració amb selecció de variables*: mètodes que només utilitzen un nombre reduït de variables, ignorant la resta, habitualment per exigències matemàtiques.

- *Calibració d'espectre complert*: pot utilitzar totes les variables amb informació rellevant, és a dir, totes les longituds d'ona en el cas dels mètodes espectroscòpics. Si les variables espectrals son molt colineals pot ser més interessant fer una selecció de variables. Dins d'aquest apartat s'han de mencionar els mètodes de compressió de variables, basats en la descomposició de les dades en components principals.

(e) Segons el número de variables utilitzades:

- *Calibració univariable*: en el model es relaciona una única variable dependent amb una variable independent.

- *Calibració multivariable*: intervenen més d'una variable dependent i/o independent. En aquest cas els models poden classificar-se en dos grups:
    - *Models rígids*: necessiten disposar d'informació de totes les fonts que poden contribuir al senyal.
    - *Models flexibles*: només necessiten informació de les fonts que es volen determinar, encara que hagin altres espècies o fenòmens físics que contribueixin al senyal analític enregistrat.

A continuació es descriuen breument, els mètodes de calibració multivariable utilitzats en aquesta memòria.

I.1.1.2. PRETRACTAMENT DE LES DADES REGISTRADES

Juntament amb la contribució de l'analit al senyal, es troben components o efectes no desitjats, els quals, d'una forma genèrica, són anomenats soroll. Aquest pot ser un soroll no estructurat, aleatori, o bé un soroll estructurat fruit d'interferències químiques i/o físiques. Per a millorar la qualitat de les dades, existeixen tractaments matemàtics que es poden aplicar abans d'establir la relació senyal-concentració. En aquesta memòria s'han utilitzat els següents:

(a) *Promitjat d'espectres*. El soroll instrumental és aleatori, pel què promitjant n senyals analítiques obtingudes a partir d'una mateixa mostra, la relació senyal/soroll augmenta en un factor $n^{1/2}$. El promitjat dels espectres és una operació que quasi sempre acompanya al registre del espectre d'una mostra, essent l'espectre final obtingut el resultat del promitjat d'un número definit per l'usuari del espectrofotòmetre.

(b) *Suavitzat espectral*. Promitjar espectres pot no ser suficient per disminuir el soroll d'alta freqüència en aquells casos en què la relació senyal/soroll és petita, pel què és necesari aplicar algoritmes sobre els espectres que minimitzin o eliminin aquest efecte. Per suavitzar els espectres es poden trobar diferents algoritmes, entre els què podem destacar el de Savitzky-Golay [Savitzky,

1964] o el filtrat utilitzant la transformada de Fourier [Horlick, 1972].

(c) *Derivades.* La derivada és un dels pretractaments més utilitzats en espectroscòpia per la seva capacitat de disminuir variacions de la línia base, com les degudes a la turbidesa de les mostres o la presència de bombolles d'aire. La utilització de la primera derivada elimina els termes constants a totes les longituds d'ona, és a dir, desplaçaments de la línia base. La segona derivada elimina els termes que varien linealment amb la longitud d'ona. No es comú l'ús de derivades d'ordre superior. Un dels mètodes més utilitzats pel càlcul de les derivades és el proposat per Savitzky-Golay [Savitzky, 1964]. Degut a què l'ús d'aquest pretracament espectral disminueix sempre la relació senyal/soroll, s'ha de ser cautelós en la seva utilització.

I.1.1.3. AVALUACIÓ DE LA CAPACITAT PREDICTIVA D'UN MODEL

L'objectiu de la calibració és obtenir uns paràmetres de regressió que permetin calcular la concentració en futures mostres de forma que, per a cada mostra *i* i analit *j*, el residual de la concentració, f_{ij}, sigui el més petit possible,

$$f_{ij} = \hat{y}_{ij} - y_{ij} \qquad \textbf{(2)}$$

on $_{ij}$ és la concentració calculada, i y_{ij} la real o experimental.

Es desitja minimitzar algun tipus d'error de predicció mig per a la població sencera a la que s'aplicarà la calibració. Per avaluar aquesta capacitat predictiva, es sol utilitzar el sumatori del quadrat dels residuals, (($-y_{ij})^2$), anomenat habitualment PRESS (*Predicted Residual Error Sum of Squares*) o el seu valor mig, obtingut dividint el PRESS pel número de mostres, MSE (*Mean Squared Error*).

El càlcul de la concentració utilitzant el model construït s'anomena predicció. Així, es pot calcular el MSE de la predicció (MSEP) com

$$MSEP = \frac{\sum_{i=1}^{m_p} (\hat{y}_{ij} - y_{ij})^2}{m_p} \quad \textbf{(3)}$$

on m_p és és el número de mostres de predicció. També és normal utilitzar l'arrel quadrada d'aquest valor, RMSEP (*Root Mean Squared Error of Prediction*), atès que presenta les mateixes unitats en què es mesura la concentració. Quan es pretén comparar errors de calibració i/o predicció entre diferents analits, és útil l'error estàndard relatiu, RSE (*Relative Standard Error*) definit per l'analit *j* com

$$RSE(\%)_j = \sqrt{\frac{\sum_{i=1}^{m} (\hat{y}_{ij} - y_{ij})^2}{\sum_{i=1}^{m} (y_{ij})^2}} \times 100 \quad \textbf{(4)}$$

on el sumatori s'estén a les *m* mostres. Els símbols utilitzats per referir-se a les mostres de calibració o de predicció són RSEC i el RSEP, respectivament. Aquest paràmetre és independent dels intervals de concentració individuals, i per tant permet comparar errors entre analits a diferents concentracions.

També es pot calcular amb la mateixa finalitat, la reproduïbilitat dels diferents models assajats. La reproduïbilitat es defineix [International Standard, ISO 5725-1:1994(E)] com *the closeness of agreement between the results of the same mesurand, where the measurements are carried out under reproducibility conditions*. Estrictament, la reproduïbilitat d'un mètode analític tan sols pot determinar-se a partir d'un exercici de comparació entre laboratoris. Malgrat això, quan els replicats de les diferents mostres s'han preparat i mesurat en diferents dies, utilitzant diferents dissolucions, es pot considerar que les diferències trobades en la predicció dels replicats són una bona mesura de la reproduïbilitat del mètode. Aquesta mesura de la reproduïbilitat ha estat definida por ICH [ICH, 1996] com a precisió intermèdia (*intermediate precision)*.

A partir de la desviació estàndard d'aquests replicats (s_R), la reproduïbilitat límit (R) pot expressar-se per un nivell de confiança del 95%, com la diferència màxima esperable entre dos mesures, la qual pot ser calculada segons l'equació següent:

$$R = 2{,}83 s_R \quad \textbf{(5)}$$

Una manera més convencional d'expressar-la és considerar la desviació estàndard relativa ARSD *(Average Relative Standard Deviation*) definida per mitjà de l'equació:

$$ARSD\% = \frac{s_R}{\bar{c}} \times 100 \quad \textbf{(6)}$$

on *c* és la concentració mitja del conjunt de mostres *m*.

En el cas de tenir les mostres mesurades per duplicat el ARSDBR (*between replicates)* ve definit per l'equació següent:

$$ARSD\% = \sqrt{\frac{\sum_{i=1}^{m/2} (\hat{c}_{(i)1} - \hat{c}_{(i)2})^2}{m}} \times \frac{1}{\bar{c}} \times 100 \quad \textbf{(7)}$$

És la reproduïbilitat calculada a partir de la diferència dels resultats dels replicats de mostres d'idèntica composició però preparades en dies diferents.

I.1.2. MÈTODES LINEALS PEL TRACTAMENT DE DADES DE PRIMER ORDRE BASATS EN LA REDUCCIÓ DE VARIABLES

Els mètodes de reducció de variables es basen en el fet que la informació continguda en les variables mesurades es pot concentrar en un nombre menor de variables sense pèrdua d'informació rellevant. Aleshores, la regressió de les respostes no es fa amb les dades originals sinó es aquestes noves variables, simplificant el model i la interpretació dels resultats.

En aquest tipus de mètodes es poden utilitzar totes les variables a les que s'ha registrat el senyal analític (calibració d'espectre complert) sense necessitat de fer una selecció prèvia. A més, permeten realitzar l'anàlisi de tan sols algun dels components que contribueixen a la mesura sense la necessitat de conèixer res de les altres espècies presents. Per aquest motiu s'anomenen Mètodes Flexibles de Calibració [Vandeginste, 1990].

La importància d'aquests mètodes és deguda a la possibilitat de resoldre alguns dels problemes que es troben moltes vegades quan es vol predir **Y** a partir de **X** com són:

- *Falta de selectivitat*. Les mesures experimentals **X** poden trobar-se afectades per diferents interferents a part de l'analit. Això fa que es necessitin varies variables **X** per a predir **Y**.

- *La colinealitat*. És a dir, pot haver-hi redundància en la informació **X**. Quan més gran és el grau de colinealitat tindrem una major variància en els paràmetres de regressió i això afectarà a la precisió dels valors a determinar.

- *Falta de coneixement del model de **Y** en **X***. A vegades no es coneixen tots els constituents que modifiquen **X**, i si es coneixen, no sempre es coneixen les seves interaccions, o no es pot linealitzar totalment la resposta de l'instrument.

Entre aquests mètodes de reducció de variables destaquen la regressió en components principals (PCR) i la regressió parcial per mínims quadrats (PLS).

I.1.2.1. TRACTAMENT PREVI DE LES DADES

Els procediments de reducció de variables no se solen aplicar directament a les dades originals, sinó que aquestes són centrades o autoescalades. Els efectes d'aquests pretractaments sobre la regressió utilitzant PCR i PLS, s'han discutit a la bibliografia [Haaland, 1988a; Geladi, 1986a; Blanco, 1994].

Considerem una matriu **X** de dades on cada fila és una mostra i cada columna una variable (exemple, espectre o un perfil cinètic). Si anomenem x_{ik} a l'element de la matriu que està a la fila *i* i en la columna *k*, es poden efectuar les següents operacions:

- *Centrat de les dades per columna.* Es calcula el valor mig de cada variable de la matriu i es resta a cada punt de la columna. El valor mig correspon al centre del model, i d'aquesta forma els valors de totes les variables estan referides a aquest centre. Aquest tractament permet mantenir les unitats originals.

$$x_{ik} - \bar{x}_k \qquad \textbf{(8)}$$

- *Autoescalat de les dades*. Després de centrar cada columna, es divideix el resultat per la desviació estàndard de la mateixa, s_k. Així, la variància de cada variable és igual a la unitat.

$$\frac{x_{ik} - \bar{x}_k}{s_k} \qquad \textbf{(9)}$$

Geomètricament, és equivalent a canviar la longitud dels eixos de coordenades. Aquest tractament ha de ser utilitzat quan les variables originals estan expressades en unitats diferents o quan les seves variàncies són molt diferents, perquè en cas contrari les variables amb gran variància quedarien primades; d'aquesta manera, tots els eixos tenen la mateixa longitud i cada variable té la mateixa influència en el càlcul.

Si les dades són espectres de mostres pot ser més interessant no escalar les dades, ja que el fet l'escalar donaria igual importància a les variables que presenten un valor baix d'absorbància, o amb molt de soroll, que als màxims d'absorció. Si les dades provenen del registre de l'absorbància a una única longitud d'ona amb el temps, l'escalat pot ser una bona solució, ja que es dóna igual importància a tots els temps. Això és important quan tenim analits amb velocitats

molt diferents i, on en els primers temps, l'analit que reacciona més ràpidament influeix molt en el senyal.

I.1.2.2. ANÀLISI EN COMPONENTS PRINCIPALS (PCA)

Com a pas previ a la descripció dels mètodes de regressió PCR i PLS es descriu l'Anàlisi en Components Principals o PCA (*Principal Component Analysis*), ja que és el primer pas en moltes de les manipulacions de dades que utilitzen una reducció de variables.

L'anàlisi en Components Principals és una tècnica aplicada en molts camps de la ciència i amb diferents objectius [Jackson, 1991; Wold, 1987]. Entre ells es poden destacar la classificació de mostres, i la reducció de dades, que és el pas previ de diferents mètodes de calibració multivariable.

El punt de partida en tots els anàlisis multivariables és una matriu de dades **X**. Cada fila d'aquesta matriu s'anomena objecte, en l'aplicació química aquest es correspon a una mostra; i cada una de les columnes, anomenades variables, corresponen a les mesures que es realitzen als objectes. En espectroscòpia, aquesta matriu pot correspondre als espectres d'un conjunt de mostres on cada objecte és l'espectre d'una mostra, i les columnes són les longituds d'ona a les que s'ha registrat l'espectre. Si es registren m mostres a k longituds d'ona es tindrà una matriu de dimensions $m \times k$ (m files i k columnes).

En el cas de disposar de dades cinètiques-espectrofotomètriques, és a dir, de dades registrades a l'espectre de cada mostra a diferents temps, cada fila és una mostra i cada columna és una longitud d'ona a un temps fix. Si es registren m mostres a k longituds d'ona i a t temps, es tindran m objectes i $k \times t$ variables. La matriu estarà formada per $m \times (k \times t)$ dades, amb m files i $(k \times t)$ columnes. El tractament d'aquestes dades és el mateix que el que s'utilitza quan es registra l'espectre a un sol temps.

Un espectre pot representar-se com k punts en un espai de 2 dimensions, espectre clàssic, o com un punt en un espai de k dimensions. Si tenim m mostres, cada una es pot representar com un punt en el espai de k dimensions.

L'objectiu del PCA és descriure l'estructura del núvol de punts o agrupacions que formen els m punts en l'espai de k dimensions Si les mostres no tenen res en comú, els m punts estaran

dispersos en l'espai, però si es troben relacionades (com p.ex. en el cas d'espectres de mescles de dos components en diferents proporcions) els *m* punts apareixeran ordenats.

De totes les possibles característiques que poden servir per descriure l'aspecte del núvol de punts, el PCA busca les direccions que expliquen la màxima variabilitat de les mostres, definint uns nous eixos de coordenades que descriuen l'espai en el qual es troben les mostres. Aquests nous eixos de coordenades són els anomenats components principals o PC (*Principal Component*).

Geomètricament el PCA és un canvi d'eixos, representant aquestes mostres en un nou sistema de coordenades amb un número inferior d'eixos a l'utilitzat inicialment (figura 1.2). El primer component principal (definit com una combinació lineal de les *k* variables) és la direcció que explica la màxima variabilitat de les mostres en l'espai dimensional *k*; el segon s'escull de manera que sigui perpendicular al primer i expliqui la màxima variabilitat de les mostres un cop eliminada l'explicada pel primer component, i així successivament. No tots els components principals contenen la mateixa informació: els primers són els que descriuen la major variació de les dades, que s'associa a la informació més rellevant, mestres que els últims descriuen variació en les dades que poden ser degudes a soroll o error experimental, i poden ser descartats.

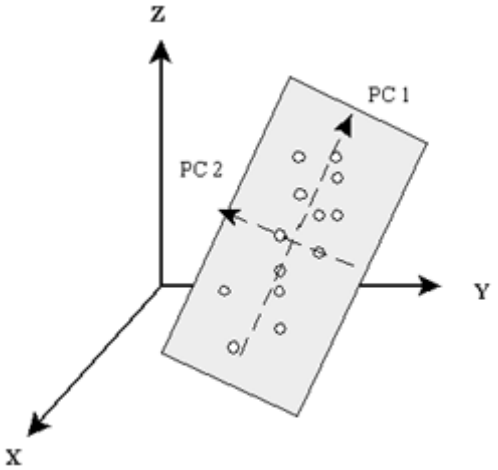


Figura 1.2. Interpretació geomètrica d'un PCA.

Seguint aquesta metodologia s'aconsegueix que la matriu **X** de partida que estava descrita per un elevat número de variables més o menys correlacionades, estigui ara definida pels components principals, que són variables no-correlacionades en un nou sistema de eixos ortogonals. Amb la reducció de la dimensionalitat del sistema inicial, s'aconsegueix eliminar la informació redundant i la variabilitat no desitjada, però es manté la informació rellevant del sistema.

El nou espai trobat tindrà tantes dimensions com fonts de variabilitat en les mostres; així un sistema de tres analits vindrà definit per un espai tridimensional, un de quatre per un sistema tetradimensional, etc. En algunes ocasions poden haver-hi altres espècies presents a les mostres, o pot no complir-se una relació perfectament lineal entre el senyal analític registrat i les concentracions dels analits que la produeixen, com és el cas del no compliment de la llei de Lambert-Beer en tot l'interval de mesura. Fins-hi tot, poden existir altres tipus de no-linealitats en el sistema com poden ser la presència d'interaccions. En aquestes situacions l'espai trobat tindrà més dimensions que analits presents en les mostres.

Matemàticament, els nous eixos es defineixen com els *loadings* (els cosinus dels angles que formen cada un d'aquests nous eixos respecte als antics) i els *scores* (les coordenades de les mostres en aquests nous eixos). La matriu de dades **X** es descompon en el producte de dos matrius, **T** (matriu de scores) i **P** (matriu de loadings), més una matriu **E** de residuals de la matriu **X**. La dimensionalitat de les noves variables és a (on $a < k$).

$$X = TP^T + E = \sum_{a=1}^{a} t_a p_a^T + E \qquad \textbf{(10)}$$

essent t_a i p_a, respectivament, el vector de scores i el vector loadings del component principal a-éssim, i el superíndex T indica matriu trasposada.

Existeixen diferents algoritmes de càlcul per obtenir les matrius **T** i **P**. Un dels més coneguts i utilitzats és l'algoritme NIPALS (*Nonlinear Iterative Partial Least Squares*), desenvolupat per Word [Wold, 1966]. Aquest és l'algoritme que utilitza el software *The Unscrambler* [CAMO AS, 1996] utilitzat per la majoria dels càlculs d'aquesta memòria.

És molt important determinar correctament el número de components que són necessaris per descriure l'estructura de les dades de la matriu **X**. La regla bàsica ha d'ésser la interpretabilitat

química. Els components principals ajuden al químic a la tasca d'extreure la informació útil, però no poden decidir per ell. A l'apartat *I.1.2.5.* es fa una breu descripció dels diferents procediments per estimar el número de components principals [Blanco, 1994].

I.1.2.3. REGRESSIÓ EN COMPONENTS PRINCIPALS (PCR)

La Regressió en Components Principals o PCR (*Principal Component Regression*) aprofita les propietats de la descomposició en components principals, PCA, realitzant una regressió múltiple inversa (ILS, *Inverse Least Squares*) de la propietat a determinar sobre els scores, considerant que aquests contenen la mateixa informació que les dades originals però havent eliminat el soroll.

Ja que es tracta d'un sistema ortogonal, s'eliminen els problemes d'inversió de la matriu originats per la colinealitat de les dades.

Si tenim una mostra amb un conjunt de p espècies que contribueixen al senyal analític, tindrem p variables, $y_1, y_2, y_3, \ldots, y_P$ representant la concentració de cada component, les quals poden ser descrites en forma de vector **y**. L'espectre de la mostra, registrat a k longituds d'ona, constitueix un conjunt de k variables independents $x_1, x_2, x_3, \ldots, x_k$ que poden ser escrites en forma de vector **x.** Si es construeix un conjunt de calibració amb m objectes, es poden agrupar els vectors que descriuen cada un dels components en cada mostra en dues matrius: la matriu **Y**, que conté les concentracions o les propietats a determinar de cada component en cada mostra, de dimensions ($m \times p$), i la matriu **X**, que conté els espectres de cada mostra, de dimensions ($m \times k$). D'aquesta manera, dins de les matrius tant les propietats a determinar com la informació espectral de cada mostra estan descrites en una fila, mentre que cada columna conté la informació d'una variable concreta per totes les mostres presents.

El primer pas és realitzar una descomposició de la matriu **X** en els seus components principals tal com s'ha indicat en l'apartat anterior.

Un cop escollit el número a de components principals òptims per a descriure la matriu **X**, aquesta es pot representar per la seva matriu de scores **T** ,

$$T = XP \qquad \textbf{(11)}$$

Fins aquest punt s'ha realitzat un PCA, obtenint a partir de la matriu de dades **X** la matriu de scores **T** i la de loadings **P**.

Les dades de concentració es relacionen amb els scores segons l'expressió:

$$Y = TB + E \quad \textbf{(12)}$$

on **B** és la matriu de regressors que es troba per mínims quadrats coneixent els valors de **Y** del conjunt de calibració:

$$B = (T^T T)^{-1} T^T Y \quad \textbf{(13)}$$

Destacar que T^TT es pot invertir sense problemes ja que els scores són ortogonals.

Un cop establert el model de calibració, es poden realitzar els càlculs per predir un conjunt de noves mostres. En primer lloc, la matriu de dades espectroscòpiques (cinètico-espectroscòpiques desdoblada) del conjunt de mostres de predicció (o el vector, si únicament es vol predir una mostra), $\mathbf{X}^*$, es centra o autoescala utilitzant els valors calculats a partir de la matriu de dades **X** utilitzada en la calibració (el superíndex* es refereix a les noves mostres per a efectuar la predicció). A partir de la matriu de loadings calculada a la calibració, pel número *a de* components principals òptims, es calculen els scores d'aquestes noves mostres de predicció, $\mathbf{T}^*$

$$T^* = X^* P \quad \textbf{(14)}$$

i, per últim, s'utilitza la matriu de regressors calculada la calibració, juntament amb els scores d'aquestes mostres, pel càlcul de les concentracions o propietat a determinar en les mostres problema.

$$Y^* = T^* B \quad \textbf{(15)}$$

Un dels principals problemes del PCR és que els components principals escollits que millor representen la matriu de dades, **X**, pot no ser l'òptim per a la predicció de les concentracions dels analits que volem determinar [Joliffe, 1982; Sutter 1992]. Per aquest motiu s'ha desenvolupat una altra tècnica de calibració que intenta concentrar el màxim poder predictiu en els primers components principals, és la regressió parcial per mínims quadrats.

I.1.2.4. REGRESSIÓ PARCIAL PER MÍNIMS QUADRATS (PLS)

El mètode de Regressió Parcial per Mínims Quadrats (PLS, *Partial Least-Squares Regression*) va ser desenvolupat per H. Wold en 1975 [Wold, H., 1975]. En aquest mètode de regressió s'estableix una relació entre les variables *x* i les *y* a través d'unes variables auxiliars anomenades variables latents, factors o components, cada una de les quals és una combinació lineal de les dades originals $x_1, x_2, \ldots, x_k$. L'algoritme PLS utilitza tant la informació continguda a la matriu de dades (matriu **X,** p. ex. dades cinètico-espectroscòpiques), com la informació continguda a la matriu de la propietat a determinar (matriu **Y**, p. ex. concentracions), amb l'objectiu que els primers components continguin la màxima informació del sistema per a la predicció de la propietat a determinar de noves mostres.

Abans de realitzar la descomposició en factors, les matrius **X** i **Y** es centren o autoescalen a variància unitat com en el cas de PCA. Aleshores cada una de les matrius es descompon simultàniament en una suma de a factors ($a \quad k$) de manera que:

$$X = TP^T + E = \sum_{i=1}^{a} t_i p_i^T + E \qquad \textbf{(16)}$$

$$Y = UQ^T + F = \sum_{i=1}^{a} u_i q_i^T + F \qquad \textbf{(17)}$$

on **T** és la matriu de scores, **P** la de loadings i **E** la matriu de residuals per la matriu de dades (matriu **X**); **U** és la matriu de scores, **Q** la matriu de loadings i **F** la matriu de residuals per la matriu de la propietat a determinar (matriu **Y**). Si tenim *m* mostres, *a* factors, *k* variables i *p* analits, la dimensionalitat de les matrius és la següent: **T** i **U** ($m \times a$), $\mathbf{P^T}$ ($a \times$ k) i $\mathbf{Q}^T$ ($a \times$ p).

En aquest cas, els loadings no coincideixen exactament amb la direcció de màxima variabilitat de les mostres com en el cas del PCA, ja que estan corregits per tal d'obtenir la màxima capacitat predictiva per la matriu **Y**.

La descomposició d'ambdues matrius no és independent, sinó que es realitza de forma simultània, establint-se una relació interna entre els scores dels blocs **X** i **Y** de manera que per cada component a, es compleix que

$$\hat{u}_a = b_a t_a \quad \textbf{(18)}$$

on el símbol ^ índica que és un valor calculat, i b_a és el coeficient de regressió per cada un dels components.

Un cop establert el model de calibració es pot realitzar la predicció de la propietat modelada en un nou conjunt de mostres, segons la següent expressió,

$$y_i^T = \hat{b}_o^T + x_i^T \hat{B} \quad \textbf{(19)}$$

essent $\mathbf{x_i}$ el vector que defineix el senyal analític de la mostra, $\mathbf{y_i}$ el vector de concentracions o propietat a determinar, **B** és la matriu dels regressors b_a, de dimensions $(a \times a)$ i $\mathbf{b_0}^T$ un vector que permet realitzar la predicció d'una mostra nova sense la necessitat de descompondre.

En el cas de calcular una sola propietat de les presents a la matriu **Y**, l'algoritme rep el nom de PLS1, el qual es pot considerar una simplifacació de l'algoritme global conegut com a PLS2, amb el qual es determinen simultàniament varies propietats.

I.1.2.5. SELECCIÓ DEL NÚMERO ÒPTIM DE COMPONENTS PRINCIPALS/ FACTORS

La selecció del número de components principals o factors que configuren el model òptim, és el punt clau en la utilització de qualsevol tècnica de calibració que realitzi una reducció de variables. S'han proposat diferents tècniques per estimar-lo que es basen, en general, en l'anàlisi de l'error de predicció en utilitzar diferent número de components principals/ factors [Malinowsky, 1991].

Per estimar aquest error i construir el model de calibració, s'utilitzen dues metodologies alternatives ; predicció externa i interna.

Per la predicció externa dels models, s'utilitzen dos grups de mostres, un anomenat pròpiament de calibració, i un altre anomenat conjunt de prova o *test set*. Les mostres del *test set* han d'ésser independents de les del conjunt de calibració però representatives del mateix i de les futures mostres a analitzar. La concentració de les mostres del *test set* és coneguda i, per tant, és possible comprovar com és comporta el model front a mostres diferents a les utilitzades en la construcció del mateix.

La predicció interna, s' utilitza quan el número de mostres disponible és relativament petit. La metòdica que es segueix és l'anomenada validació creuada (*cross validation*) [Geladi, 1986b;

Wold, 1978], que utilitza per comprovar el model mostres del propi conjunt de calibració. Mitjançant aquest mètode, el conjunt de mostres de calibració es divideix en diferents blocs o segments. El model es construeix reservant un dels segments com a conjunt de dades per comprovar els resultats i la resta s'utilitza per construir el model. El procés es repeteix tantes vegades com el número de segments escollits, de forma que cada cop es deixa un segment fora del calibrat i la resta s'utilitza per construir el model. El número de mostres que s'utilitza en cada segment de validació és variable; el cas extrem és utilitzar només una mostra (mètode *leave-one-out*). Al final del procés tots els segments han participat en la construcció i validació del model. En els mètodes de reducció de variables el procés es realitza per a cada factor o component principal calculant el MSE per a cada segment i acumulant-lo, de forma que s'aconsegueix una bona estimació del poder predictiu de les mostres de calibració. El MSECV (*Mean Squared Error of prediction by Cross Validation*) per a cada factor *a* i cada analit *j* és

$$MSECV_{j,a} = \frac{\sum_{i=1}^{m_c} (\hat{y}_{ij} - y_{ij})^2}{m_c} \qquad \textbf{(20)}$$

on m_c és el número de mostres de calibració en el segment de *cross validation*.

Donat que el mètode més habitual per la construcció del model és el de la validació creuada, una forma molt corrent d'escollir el número de components principals/ factors òptims és la suggerida per Wold [Wold, 1978], que consisteix en representar el valor de MSECV vers el número de factors i buscar el mínim. Es parteix de la idea que l'error disminueix en augmentar el nombre de factors, ja que es modela cada cop millor el sistema, fins que s'assoleix un punt en el que els nous factors introduïts tan sols expliquen soroll i aleshores el MSECV torna a augmentar com a conseqüència del sobreajust del model. Si bé és una idea raonable, el fet d'utilitzar únicament un número limitat de mostres (com a màxim totes les presents en el conjunt de calibració) fa que el mètode estigui subjecte a un cert error i presenti el perill de generar un cert sobreajust de les dades [Osten, 1988]. Altres autors prefereixen utilitzar el primer mínim local que apareix en la representació del MSECV vers el nombre de factors [Martens, 1989], encara que pot ser que en aquest cas es produeixi un subajust de les dades. Quan la construcció del model es realitza per *test set*, el número de factors es selecciona en funció de la evolució de l'error per aquest conjunt de mostres.

Un altre mètode és el proposat per Haaland i Thomas [Haaland, 1988b] on s'escull el nombre de components dels quals el PRESS no sigui significativament superior que el mínim PRESS del model, evitant d'aquesta manera el sobreajust. Utilitzant aquest criteri, primer es construeix el model mitjançant la validació creuada i es calcula el valor del PRESS per a cada component principal. El mínim valor del PRESS vindrà donat per un nombre de components que anomenem a^*. Cada valor de PRESS obtingut amb un nombre de components menor al a^* es compara amb el valor de PRESS(a^*) mitjançant un test F. Esquemàticament el procediment seguit és el següent:

Per a cada component $a = 1, 2, ..., a^*$ es calcula

$$F(a) = \frac{PRESS(a)}{PRESS(a^*)} \qquad \textbf{(21)}$$

Aleshores, s'escull com a número de components òptim el menor a de manera que $F(a)<F_{\ ,m,m}$, on $F_{\ ,m,m}$ és el valor tabulat per una prova F unilateral, amb un percentatge de nivell de significació de (1-) i m graus de llibertat. Basant-se en un criteri empírics, el valor de recomanat per Haaland i Thomas és de 0,25.

L'observació visual del loadings pot ser un altre criteri de selecció del número de factors. S'escull aquell número de factors els loadings dels quals encara presentin una estructura definida, és a dir, encara continguin informació rellevant i no únicament soroll.

I.1.3. MÈTODES LINEALS PEL TRACTAMENT DE DADES DE SEGON ORDRE BASATS EN LA REDUCCIÓ DE VARIABLES

I.1.3.1. INTRODUCCIÓ

Tots els mètodes citats fins aquest punt estan dissenyats per tal de tractar dades de primer ordre. Per a poder utilitzar aquests mètodes de calibració en cas que es disposi de dades de segon ordre o superior (multidimensionals), aquestes han d'ésser desdoblades. Aquesta operació resulta en una pèrdua de l'estructura tridimensional i per tant, de la seva informació implícita.

Recentment, s'han desenvolupat tècniques per tractar aquest tipus de dades directament sense la necessitat d'un desdoblament previ. És el cas de la Regressió Parcial per Mínims Quadrats Multi Via o nPLS, emprada diverses vegades en aquesta memòria, o d'altres com ara el PARAFAC, que poden ser més adequades en altres situacions.

Aquests mètodes, de la mateixa manera que els bidimensionals, descomponen les dades en conjunts de scores i loadings per tal de descriure-les de forma més condensada. Ambdós presenten avantatges i inconvenients, per tant, pot ser necessari provar-los a fi de trobar el més adient a cada problema plantejat.

La experiència pràctica existent en la utilització dels mètodes bidimensionals com ara el PCA o el PLS, porta a pensar que aquests últims són més simples que els mètodes multidimensionals, no obstant això, si considerem l'estructura multidimensional de les dades, això no és cert, ja que en aplicar un PCA o PLS les dades han d'ésser prèviament desdoblades en una matriu bidimensional. Les variables apareixen barrejades i, per tant, l'efecte d'una variable no està associat amb un únic element del vector loading, sinó amb varis d'ells. Quan no es té en compte l'estructura tridimensional, el model utilitza per igual totes les dades, utilitzant els graus de llibertat necessaris a fi d'obtenir el millor ajust possible. Així, els mètodes bidimensionals, poden donar lloc a models més complexes i difícils d'interpretar, donat que el nombre de paràmetres que es calculen és superior que en el cas dels mètodes tridimensionals.

I.1.3.2. PLS MULTI VIA (nPLS)

La Regressió Parcial per Mínims Quadrats Multi Via o nPLS és un mètode que realitza la descomposició de les dades en components trilineals, però enlloc d'un vector de scores i un de loadings com en el cas del PCA bilineal, cada component consisteix d'un vector de scores i dos vectors de loadings (figura 1.3). L'estructura cúbica de les dades primàries **X** es descompon en un conjunt de cubs de manera òptima, i on el rang pot ser superior que el número d'espècies absorbents.

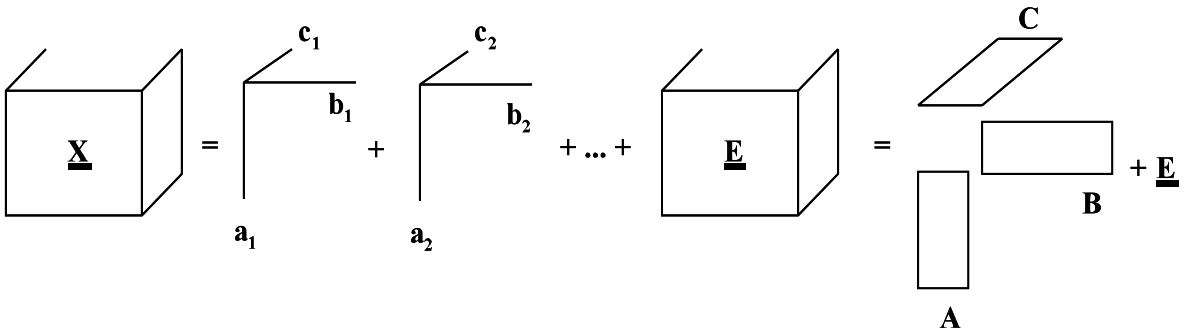


Figura1.3. Descomposició de dades d'estructura tridimensional en components trilineals.

A la pràctica no es distingeix entre scores i loadings, ja que són tractats numèricament per igual. Per tant, un model n-PLS d'una estructura tridimensional ve definit per tres matrius **A** ($m \times F$), **B** ($k \times F$), i **C** ($t \times F$), els elements de les quals són a_{mf}, b_{kf} i c_{tf}, respectivament. F representa el número de factors; m, k i t són el número de mostres, longituds d'ona i temps, respectivament.

Les avantatges del nPLS són varies:

- La solució és fàcil d'interpretar, comparada amb els mètodes *unfolding*, i això és especialment important quan es disposa d'un número molt gran de variables.

- L'algoritme és ràpid degut a la poca quantitat de paràmetres a calcular i al fet que el problema es redueix a una descomposició en valors propis.

- Una altra avantatge d'aquest mètode de calibració front altres mètodes trilineals com el PARAFAC, és la possibilitat d'estabilització de la solució en cas de baixa contribució neta de l'analit, degut a la incorporació de la variable dependent en la descomposició.

Però, evidentment, com en tots els mètodes de calibració, hi ha certes desavantatges. Es perd ajust en el model trilineal comparant-lo amb un bilineal ja que hi ha restriccions més severes.

A la pràctica, durant el procés de calibració no es mira només la falta d'ajust, sinó també, si el model és l'apropiat per el problema en concret. Per tant, s'han de provar diferents mètodes per veure quin és el més apropiat. Si diferents mètodes descriuen més o menys per igual les dades, s'ha d'escollir sempre el més simple i interpretable.

I.1.3.2.1. El model nPLS

A la versió trilineal del PLS les dades amb estructura tridimensional es descomponen seguint la filosofia del PLS, descrivint la covariància de les variables dependents i independents. Si assumim que tant el bloc de variables dependents com el de independents tenen estructura tridimensional, $\underline{\mathbf{X}}$ és el conjunt de variables independents de dimensions $m \times k \times t$ i $\mathbf{X}$ la matriu desdoblada $m \times kt$. $\underline{\mathbf{Y}}$ és el conjunt de variables dependents de dimensions $m \times p \times l$ i $\mathbf{Y}$ la matriu desdoblada $m \times pl$. De forma general el model nPLS pot ser escrit com :

$$\underline{X} = T(W^t \otimes W^k)^T + E_x \qquad \textbf{(22)}$$

$$\underline{Y} = U(Q^l \otimes Q^p)^T + E_y \qquad \textbf{(23)}$$

on $\mathbf{T}$ i $\mathbf{U}$ són les matrius de scores, $\mathbf{W}$ i $\mathbf{Q}$ les de loadings (el superíndex indica l'ordre considerat) i $\mathbf{E}_x$ i $\mathbf{E}_y$ els errors per les variables independents i dependents, respectivament.

El software utilitzat en aquesta memòria per la construcció dels models nPLS, escrit en codi Matlab, està basat en l'algoritme descrit per Bro [Bro, 1996 i 1998]. L'algoritme descompon $\underline{\mathbf{X}}$ (s'assumeix que està centrada en la direcció de m) en conjunts de *triades*. Una *triada* consisteix en un vector $\mathbf{t}$ de scores i dos vectors de pesos, un en el segon ordre anomenat $\mathbf{w}^k$ ($k \times 1$) i un altre en el tercer anomenat $\mathbf{w}^t$ ($t \times 1$).

Per aquest mètode de calibració existeixen els mateixos mètodes per a determinar el número de factors òptims que en el cas del PLS bidimensional.

Una descripció detallada de les propietats del nPLS es pot trobar descrita a la bibliografia [Smilde, 1997; De Jong, 1998; Bro, 1998].

I.1.3.2.2. Pretractament de les dades

El pretractament de les dades tridimensionals és més complicat que en el cas de les bidimensionals, degut a la major dimensionalitat de les mateixes. El centrat d'un mode pot ésser realitzat desdoblant prime1r les dades de calibració en una matriu $m \times kt$ i centrant aquesta matriu com en el PCA ordinari:

$$x_{mkt}^{cent} = x_{mkt} - \bar{x}_{kt} \qquad \textbf{(24)}$$

on

$$\bar{x}_{kt} = \frac{\sum_{im=1}^{m} x_{mkt}}{m} \qquad \textbf{(25)}$$

Això es refereix a un centrat senzill i s'anomena centrat a través del primer mode [Ten Berge, 1989]. El centrat pot ser aplicat a qualsevol dels modes, dependent del problema. Si s'ha de realitzar el centrat a través de més d'un mode s'ha d'efectuar primer el centrat en un mode i, posteriorment, centrar novament el resultat obtingut. A la bibliografia [Ten Berge, 1989; Kruskal, 1983; Harshman, 1984] es descriu l'efecte d'escalar i centrar en dades trilineals. L'escalat també té que ser efectuat tenint en compte l'estructura tridimensional de les dades. Si la variable k del segon mode té que ser escalada (comparada amb la resta de variables en el segon mode), és necessari escalar totes les columnes on intervé la variable k. Això significa que s'han d'escalar matrius senceres en lloc de columnes. Matemàticament l'escalat pot ser descrit com,

$$x_{mkt}^{scal} = \frac{x_{mkt}}{s_m} \qquad \textbf{(26)}$$

on s_m pot ser definit com

$$s_m = \left[\sum_{k=1}^{k} \sum_{t=1}^{t} x_{mkt}^2 \right]^{1/2} \quad \textbf{(27)}$$

Quan es vol realitzar l'escalat dins de diferents modes, la situació es complica, perquè l'escalat d'un mode afecta l'escala dels altres. Si es desitja escalar a variància unitat dins de varis modes, ha d'ésser efectuat simultàniament fins a convergència.

En el software utilitzat, es disposa d'un programa en codi Matlab per a realitzar l'escalat iteratiu i el centrat.

I.1.4. MÈTODES NO-LINEALS: XARXES NEURONALS ALS ARTIFICIALS (ANN)

I.1.4.1. INTRODUCCIÓ

Les Xarxes Neuronals Artificials (ANN, *Artificial Neural Networks*) es poden definir com un sistema iteratiu de càlcul que intenta reproduir, de forma simple i senzilla, el sistema de connexions que existeix entre les neurones del cervell humà.

A les xarxes neuronals artificials, la neurona artificial (anomenada simplement neurona o node) intenta simular la neurona biològica [Wythoff, 1993]. La sinapsi es representa per una connexió entre dues neurones i la força sinàptica per un pes associat a aquesta connexió (un número real). Les senyals d'entrada passen a les neurones, on es realitza la seva suma ponderada. A continuació, es transformen passant a través d'una funció de transferència cap a la sortida. La propagació del senyal ve determinat per les connexions entre les neurones i pels seus pesos associats (figura 1.4).

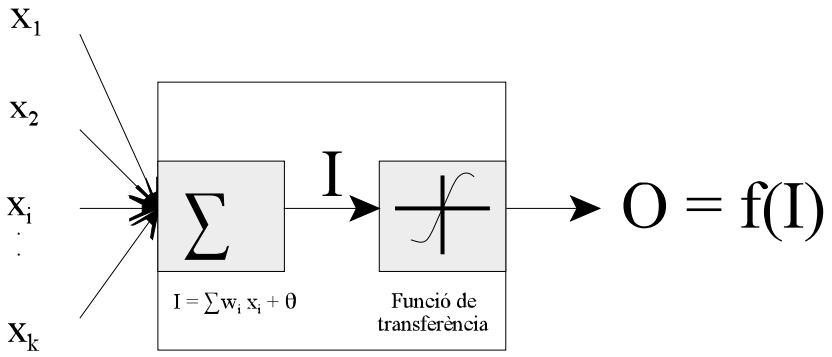


Figura 1.4. Xarxa Neuronal Artificial

La manera com es troben connectades entre si les neurones individuals és el que defineix les xarxes, les quals s'organitzen generalment en una seqüència de capes o nivells. D'aquesta manera, es pot definir una capa d'entrada com aquella en la que les dades es presenten a la xarxa, i una capa de sortida, a partir de la qual s'obtenen les respostes (p.ex. concentracions), anomenant a la resta de capes o nivells situats entre la capa d'entrada i de sortida com capes ocultes (figura 1.5). El número de neurones que hi ha en cada capa defineix l'arquitectura de la xarxa neuronal, representada normalment per (*i, h_p, o*), on *i* és el número de neurones de la capa d'entrada, que

normalment és igual al número de variables; h_i el número de neurones de la capa oculta *i*; *o* el número de neurones de la capa de sortida, que en aquesta memòria correspon a les concentracions dels analits a determinar.

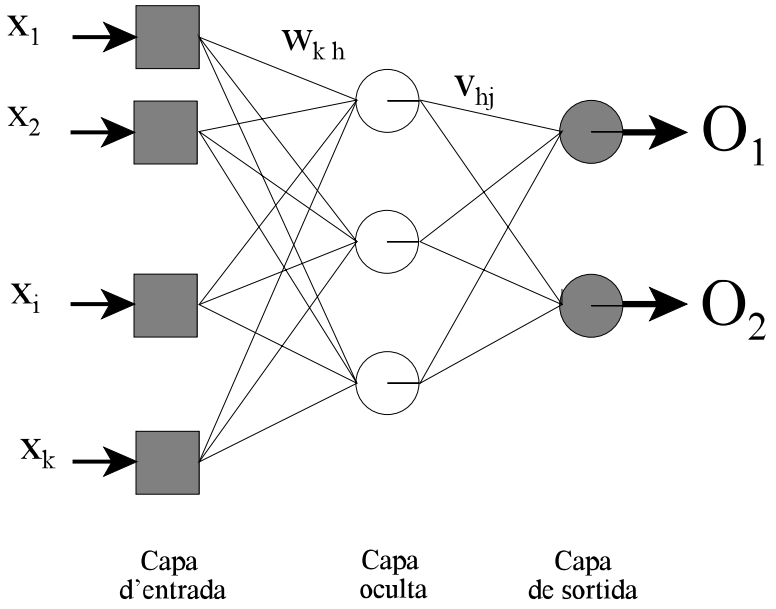


Figura 1.5. Estructura d'una Xarxa MPL

Un procés portat a terme amb ANN consta de dues etapes: l'etapa d'aprenentatge (*learning*) i l'etapa de resposta. Durant l'aprenentatge, la xarxa neuronal aprèn a partir d'exemples que se li presenten, adaptant els pesos de les connexions en resposta als senyals que li arriben de la capa d'entrada i, opcionalment, de la resposta desitjada. L'aprenentatge pot ser supervisat, si existeix una resposta correcta a la què el mètode s'ha d'aproximar; no supervisat, si no es coneix la resposta que el mètode ha de proporcionar; o per reforçament si l'aprenentatge no es basa en una resposta correcta, sinó únicament en una indicació de si és o no bona. En funció d'això últim es produeix un reforçament o inhibició de les connexions.

Per a cada classe d'aprenentatge existeix una regla d'aprenentatge que especifica com han d'adaptar-se els pesos en resposta als exemples dels què està aprenent. En aquesta memòria, s'ha utilitzat la regla delta, basada en reduir l'error entre la sortida obtinguda per la xarxa i la sortida desitjada. És per tant, un esquema d'aprenentatge supervisat.

L'etapa de resposta es refereix a com la xarxa processa globalment els senyals que arriben

al seu nivell d'entrada i proporciona la resposta en el nivell de sortida. La resposta s'integra contínuament en el procés d'aprenentatge comparant-se amb la resposta desitjada, i creant així una funció d'error.

Existeixen diferents tipus de xarxes neuronals i s'utilitzen unes o altres depenent del tipus de problema que es vulgui resoldre. Totes elles es defineixen pels tres elements ja citats: els elements de càlcul (neurones), l'arquitectura de la xarxa (distribució i connexió entre les neurones) i la regla d'entrenament utilitzada. A la bibliografia [Zupan, 1993; Zupan, 1991] es poden trobar diferents tipus de xarxes, i també algunes aplicacions de les xarxes neuronals a l'àmbit de la química.

I.1.4.2. XARXES MULTI-CAPES PERCEPTRÓ (MPLS)

La major part de les aplicacions a l'anàlisi químic de les xarxes neuronals es basen en els sistemes de multi-capes anomenades *multi-layer perceptron (*MLP*)*, també anomenades *multi-layer feed-fordward network (*MLF*)*. Aquestes xarxes utilitzen com algoritme d'aprenentatge l'algoritme de propagació dels errors cap endarrera (*back-propagation algorithn*).

El tipus d'aprenentatge d'aquestes xarxes és supervisat, amb connexions totals entre neurones de capes consecutives i sempre cap endavant, és a dir, els elements de cada nivell es connecten amb tots els elements dels següent nivell (figura 1.5).

Aquestes xarxes es caracteritzen per la seva capacitat de modelar correctament sistemes en què la relació entre dades experimentals i la propietat a determinar és no-lineal [Long, 1990; Despagne, 1998], encara que també han estat aplicades amb èxit a problemes de classificació [McCarrick, 1991].

L'estructura d'una xarxa MLP ve definida inicialment per una capa d'entrada, amb un número de neurones igual al número de variables que defineixen a cada un dels objectes (p.ex. *k* valors d'absorbància). Aquesta capa és la que rep la informació i la traspassa a la següent capa, la capa oculta, que serà la responsable de processar i transmetre la informació a la següent capa o a la capa de sortida. El número de capes ocultes, així com les neurones que la configuren, han d'ésser optimitzades en cada cas particular. La capa de sortida sempre estarà constituïda per tantes neurones com variables de sortida desitgem obtenir (p.ex. concentracions d'un o més analits, *p*).

L'equació general que descriu els valors de sortida o_{ij} d'una xarxa MLP amb una capa oculta, a partir dels valors d'entrada x_{ik} , és:

$$O_{ij} = g\left[v_{oj} + \sum_{k=1}^{ki} v_{kj} f\left[w_{ok} + \sum_{k=1}^{k} x_{ik} w_{kk}\right]\right] \qquad \textbf{(28)}$$

essent $i = 1, ..., m$, mostres o objectes, $k = 1, ..., k$, entrades o variables, $h = 1, ..., h_i$, neurones ocultes i $j = 1, ..., p$, sortides de la xarxa; ***f*** és la funció de transferència corresponent a una neurona de la capa oculta (generalment ***f*** serà la funció lineal o sigmoidal); ***g*** la funció de transferència d'una neurona de la capa de sortida. Ambdues funcions, ***f*** i ***g***, poden ser no-lineals cosa que justifica l'enorme capacitat d'ajust d'aquests sistemes.

Aquesta última expressió és indicativa d'un model paramètric no-lineal, que ens serveix per relacionar **Y** amb totes les variables **X**.

I.1.4.2.1. Aprenentatge per retropropagació

L'algoritme de propagació dels errors cap endarrera o retropropagació (*back-propagation algorithn*) [Zupan, 1993], no és més que un aprenentatge supervisat en què els pesos de les connexions són corregits en funció dels valors de referència de les sortides obtingudes. Aquest procés de correcció de pesos pot fer-se de dues maneres: després de cada nova entrada individual, correcció immediata, o després de què totes les entrades hagin estat comprovades, correcció post-posada. En el primer cas, la correcció es realitza immediatament després de què l'error hagi estat detectat; en el segon, l'error individual per a tots els parells de dades s'acumula i l'error acumulat de tot el conjunt d'aprenentatge s'utilitza per la correcció. La més utilitzada és la correcció immediata. A la bibliografia pot trobar-se més informació sobre les avantatges i inconvenients dels modes de correcció esmentats [Finnoff, 1994; López-Fandiño, 1997].

Durant l'aprenentatge, les dades **X** són presentades a la xarxa neuronal i la matriu de sortida, **O** és immediatament comparada amb la matriu de referència **Y**, que és el valor correcte de sortida per les dades **X.** Un cop es coneix l'error produït per la xarxa, la regla delta generalitzada ho aprofita per a corregir els pesos al llarg de tota la xarxa en la direcció oposada a l'entrada de dades, d'aquí prové el nom de retropropagació.

La regla delta generalitzada proposa que la correcció de pesos $\boldsymbol{W}$ en una capa és proporcional a (funció que depèn de l'error) i a l'entrada **X** que, en aquest cas, és la sortida de la capa anterior $\boldsymbol{O}^{ant}$. Un cop corregit el vector de pesos, el senyal de sortida s'hauria d'apropar més al valor de sortida conegut. La regla delta es pot expressar com:

$$\Delta W = \eta \delta X \qquad \textbf{(29)}$$

o és la constant de proporcionalitat anomenada *learning rate* o velocitat d'aprenentatge, que ens determina la velocitat a la que seran canviats els pesos, i és la constant de correcció (funció d'error) que es pretén trobar.

I.1.4.2.2. Consideracions pràctiques per a la construcció d'una xarxa

A continuació es tracten diferents aspectes pràctics que s'han de tenir en compte quan es desenvolupen mètodes de calibració utilitzant xarxes MLP. Podem distingir dos tipus de qüestions: les relacionades amb les mostres, i aquelles que tenen que veure amb la recerca de la topologia òptima de la xarxa.

I.1.4.2.2.1. Mostres

En principi, les mostres que s'utilitzen en l'aprenentatge han de definir de forma correcte l'espai d'aplicació de la xarxa, ja que l'equació construïda no és una descripció del sistema complet sinó que està restringida pels objectes utilitzats en el procés d'aprenentatge.

(a) *Conjunts de mostres*

Quan es treballa amb xarxes neuronals són necessaris, com a mínim, tres conjunts de mostres. Durant l'etapa de construcció del model s'utilitzen dos dels conjunts d'objectes: el conjunt d'aprenentatge (*training set*) i el de validació (*test set*). Aquest segon conjunt d'objectes és independent del conjunt d'aprenentatge, però ha d'ésser representatiu del mateix. S'utilitza per optimitzar el model i evitar un possible sobreajust. Finalment, per assegurar la capacitat predictiva

real de la xarxa construïda ha d'utilitzar-se un tercer conjunt de mostres, conjunt de predicció extern (*external predicction set*).

La selecció d'aquests conjunts es troba fortament influenciada pel tipus de xarxa a utilitzar però sempre ha de mantenir-se la representativitat del sistema en estudi en els tres conjunts. És important evitar introduir la dependència temporal o d'ordre d'assignació dels objectes, seleccionant-los de forma aleatòria o segons un disseny experimental prèviament definit.

(b) *Número de mostres*

Durant el procés d'aprenentatge s'optimitzen els pesos de la xarxa a partir dels exemples que s'utilitzen en el conjunt d'entrenament. Per tant, serà necessari aconseguir el número d'objectes suficients per a poder modelar la relació entre **X** i **Y**. El número de pesos a optimitzar augmenta, lògicament, al augmentar el número de neurones de les capes, i això determina el número d'objectes a considerar. Per exemple, amb una sola capa oculta, el número de pesos correspon a:

$$n^o de\ pesos = (k + 1) \times h_i + (h_i + 1) \times p \qquad \textbf{(30)}$$

on k és el número de neurones de la capa d'entrada, h_i el número de neurones de la capa oculta i p el número de neurones de la capa de sortida.

S'observa clarament com el número de pesos (és a dir, el número de paràmetres ajustables) creix ràpidament a l'augmentar el número de neurones, fent que es necessiti un gran nombre d'objectes en el conjunt d'aprenentatge a fi d'evitar tenir més paràmetres que dades. En general, es recomana [Zupan, 1993] que el número mínim de mostres en el conjunt d'aprenentatge sigui el doble que el número de pesos totals. Altres autors aconsellen que el número de mostres sigui 10 cops el número de pesos a ajustar [Bos, 1993]. Per altra banda, també seran necessàries més mostres per a la validació del model. La gran quantitat de mostres necessàries és una de les majors limitacions pràctiques en el ús de les xarxes neuronals.

(c) *Detecció d'objectes anòmals*

Aquest és un punt difícil en les xarxes neuronals. Si prèviament a la utilització de les xarxes neuronals s'ha aplicat alguna altra tècnica de calibració multivariable, com és el cas d'un PLS o PCR, pot succeir que algunes mostres hagin estat considerades *outliers* i, no obstant això, siguin modelades correctament per les xarxes neuronals. La manera més habitual i simple de trobar objectes anòmals és per mitjà de representacions gràfiques de les dades **X** i **Y**. La representació pot simplificar-se si es realitza un anàlisi en components principals, PCA.

(d) *Pretractament de les dades*

Cada tipus de variable presentada al model neuronal té un rang d'operació individual. Algunes variables presentaran un valor mig més alt i/o una variància més elevada que la resta.

Si aquestes variables són aplicades directament al model neuronal, és de suposar que tindran uns pesos associats més grans. Com a conseqüència, aquestes variables d'entrada exerciran una gran influència en la resposta del model.

En existir els termes de biax en les capes d'entrada i les ocultes, no és necessari centrar les variables. Tampoc és necessari escalar-les a variància unitat ja que el mètode no es basa en la maximització de la variància-covariància. Normalment les variables s'escalen en un cert interval (*range scaling*) de forma que quedin distribuïdes adequadament segons la funció de transferència escollida. Anàlogament es recomana aquest escalat per les variables **Y**.

(e) *Número de variables d'entrada*

Com ja s'ha esmentat abans, recordar que la relació mostres/pesos ha d'ésser el més gran possible. En general, quan es te un nombre elevat d'objectes, i s'ha controlat correctament el procés de presa de dades, la xarxa pot ser construïda utilitzant dades originals. Tanmateix, aquesta situació no és la més freqüent. Normalment, necessitarem reduir significativament el número de variables d'entrada a fi que el número de pesos a utilitzar estigui en d'acord amb l'equació anterior. Això pot realitzar-se a partir de mètodes que permetin disminuir la dimensionalitat inicial del nostre sistema com pot ser, per exemple, la utilització dels scores obtinguts en un anàlisi en components

principals (PCA) com a variables d'entrada a la xarxa. És una bona estratègia i proporciona certes avantatges addicionals, com és una major velocitat d'aprenentatge, la disminució del risc de sobreajust i l'augment de la robustesa de la xarxa quan el número d'objectes que són utilitzats en el procés d'aprenentatge és petit.

De totes maneres, la utilització de tècniques de reducció de variables han de realitzar-se sempre en el conjunt d'entrenament i, a partir dels valors trobats, calcular els scores dels objectes que configuren els altres dos conjunts de mostres.

I.1.4.2.2.2. Topologia de la xarxa: disseny i optimització

Un cop es disposa de les dades i objectes representatius del problema, i s'han pretractat convenientment, el següent pas és el disseny de la xarxa.

(a) *Arquitectura de la xarxa*

La selecció de l'arquitectura adequada, és a dir, la selecció del número de capes i de neurones en cada una d'elles, és un punt de gran importància per aconseguir modelar la relació existent entre les dades d'entrada i de sortida.

El número de neurones de la capa d'entrada i de sortida ve determinat per la naturalesa del problema. Tindrem tantes neurones d'entrada com variables (originals o procedents d'una reducció prèvia) i tantes neurones en la capa de sortida com paràmetres a estimar (p.ex. concentracions). En general, amb una sola capa oculta és suficient. El número de neurones que conformen aquesta capa oculta ha d'optimitzar-se. La utilització d'un número de neurones inferior a l'adequat provoca que la xarxa no funcioni correctament, mestre que l'ús d'un número massa gran pot crear problemes de sobreajust. Per tant, és convenient assajar un número diferent de neurones en la capa oculta i escollir aquella xarxa que proporcioni un mínim error sense arribar a sobreajustar el sistema.

(b) *La funció de transferència*

A la capa d'entrada, la funció de transferència és una funció de pas, la qual només introdueix les dades originals a la xarxa sense fer-hi cap transformació.

Les funcions que s'utilitzen en les neurones de la capa oculta han de complir una sèrie de condicions per poder aplicar la regla d'aprenentatge:

- Ser monòtona creixent en l'interval d'aplicació (implica que la funció sigui continua en l'interval i es trobi acotada).

-Tenir una derivada continua en aquest interval.

Generalment, en aquest tipus de xarxes, la funció de transferència de les neurones de la capa oculta és del tipus sigmoidal. Una de les més habituals és l'anomenada funció logística:

$$f(x) = \frac{1}{1+e^{-x}} \tag{31}$$

També s'utilitza la funció tangent hiperbòlica:

$$f(x) = \tanh(x) \tag{32}$$

Les funcions de transferència utilitzades en les neurones de la capa de sortida poden ser lineals o no-lineals.

(c) Altres paràmetres optimitzables

Existeixen altres paràmetres a optimitzar que poden influir de manera decisiva a la velocitat d'aprenentatge i al comportament de la xarxa, com poden ser els valors inicials dels pesos de les connexions i la seva distribució (uniforme o gaussiana) que, en general, han d'ésser valors petits i aleatoris; el mètode d'actualització dels pesos, és a dir, correcció immediata o correcció post-posada; els valors dels paràmetres velocitat d'aprenentatge, etc.

I.1.4.3. ENTRENAMENT I AVALUACIÓ D'UNA XARXA NEURONAL

Quan s'està desenvolupant un model neuronal l'objectiu és trobar la solució global en el domini d'un problema, el qual pot contenir una gran quantitat de solucions subòptimes o locals.

Una solució global s'obté amb aquell model que produeix el mínim error possible. La xarxa ideal és aquella que proporciona el mínim error possible. Generalment, trobar aquest model no és possible, encara que hem d'aproximar-nos a ell. Per això, necessitem controlar la variació de l'error comès per les diferents xarxes o, dit d'una altra manera, per les combinacions de pesos que es van obtenint durant l'aprenentatge.

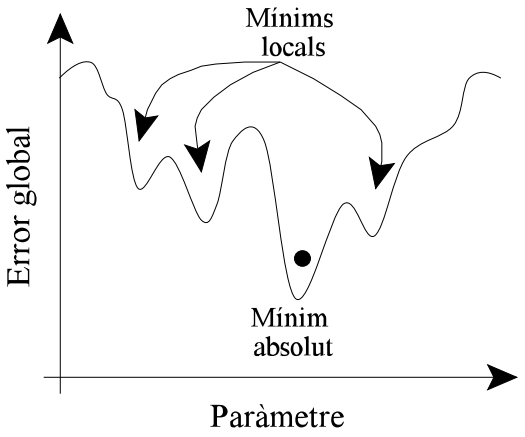


Figura 1.6. Possibles solucions d'una ANN.

La majoria de les xarxes neuronals aprenen per a obtenir sortides que minimitzin l'error entre els valors obtinguts i els de referència. No obstant això, i com s'observa a la figura 1.6, qualsevol problema complex té varies solucions subòptimes. Per evitar que el procés d'aprenentatge es pari en una d'aquestes solucions, és important definir correctament la configuració de la xarxa i també realitzar una adequada divisió de tots els objectes o mostres disponibles en els diferents conjunts amb els que es treballa: entrenament i validació.

El comportament d'una ANN supervisada durant el procés d'aprenentatge normalment es segueix utilitzant l'arrel quadrada del valor mig del sumatori del quadrat dels residuals de predicció RMSEP (*Root mean Square Error of Predicction*),

$$RMSE = \sqrt{\frac{1}{mp}\sum_{i=1}^{m}\sum_{j=1}^{p}(y_{ij} - \hat{y}_{ij})^2} \quad \textbf{(33)}$$

on m és el número de mostres en el conjunt considerat (entrenament, validació o extern), p el número de neurones de la capa de sortida (número d'espècies), y_{ij} el valor de referència de la resposta i $_{ij}$ la resposta calculada per la xarxa. Atès que és una mesura quadràtica de l'error, l'utilització del RMSE tendeix a amplificar les grans diferències i a esmorteir les petites.

El software utilitzat en el desenvolupament d'aquest treball, calcula aquest error individualment per a cada sortida. Per tant, si el què s'està quantificant és la concentració de varis analits es pot observar l'ajust de la xarxa per a cada espècie, independentment, mitjançant l'evolució del $RMSE_j$ per a cada iteració (figura 1.7).

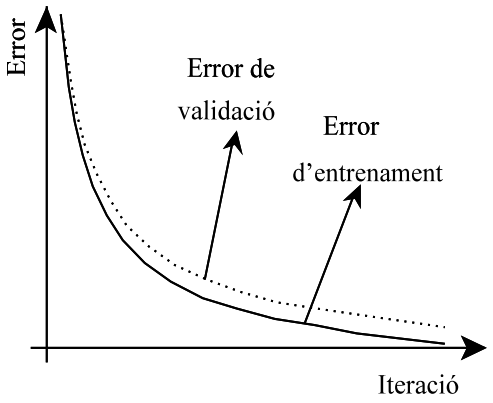


Figura 1.7. Evolució de l'error.

L'evolució del $RMSE_j$ al llarg de tot el procés d'entrenament proporciona informació important. Podem trobar-nos davant situacions molt variades:

(a) Tant el $RMSE_j$ del conjunt d'entrenament com el de validació tendeixen a un mínim (situació òptima).

(b) El $RMSE_j$ convergeix amb dificultat ja que es produeixen grans oscil lacions al voltant de valors mínims. Es pot solucionar agafant un valor elevat al principi de velocitat d'aprenentatge, , i anar disminuint-lo a mesura que transcorre l'entrenament.

(c) L'entrenament pot detenir-se o evolucionar molt lentament degut a que la xarxa ha caigut en un mínim local i no pot sortir d'ell. Per a solucionar-ho s'han d'agafar pesos inicials

aleatoris més petits i/o tornar a escalar les variables d'entrada d'una altra forma. És a dir, es torna a iniciar el càlcul canviant la configuració de la xarxa.

(d) El fenomen de sobreentrenament (*overtraining*) es pot detectar clarament perquè la corba de validació no segueix la mateixa tendència que la d'entrenament. A partir d'un moment determinat es separen i la corba de validació comença a pujar fins arribar a estabilitzar-se. La situació pot ser deguda a dos motius: el número de neurones de la capa oculta és massa gran o s'ha efectuat masses iteracions d'entrenament. En general, la disminució del número de neurones en la capa oculta normalment acaba amb el problema.

(e) L'última situació succeeix quan hi ha una falta de concordança entre el conjunt d'entrenament i el de validació. La corba de validació, encara que segueix la mateixa tendència que la de calibració, es separa d'aquesta després de l'inici i no presenta mínim. No s'han repartit correctament les mostres disponibles entre els dos conjunts, per tant, és necessari distribuir-les.