APARTAT II

PRETRACTAMENTS ESPECTRALS

ESTUDI DE LA CORRECCIÓ ORTOGONAL DEL SENYAL (OSC). CONTROL DEL SOBREAJUST[A]

En aquesta part del treball es va decidir fer èmfasi en l'estudi de la correcció ortogonal del senyal. Vistos els resultats obtinguts per diferents autors, i tenint en compte els resultats obtinguts en el treball previ, es va decidir estudiar la manera de controlar amb un criteri objectiu un dels principals inconvenients d'aquest tractament: la seva marcada tendència al sobreajust.

La correcció ortogonal del senyal (OSC) busca la reducció de la variabilitat de la matriu de variables predictores (**X**) a través d'un procés en el que s'intenta eliminar la informació no correlacionada amb la propietat analítica a determinar (**Y**). Com a hipòtesi de base es considera que aquesta informació és ortogonal a la concentració. Es tracta d'una correcció que es troba en fase d'experimentació. De fet, està encara en fase de desenvolupament i sembla que encara estigui buscant el seu lloc. S'han publicat pocs articles sobre el tema[1,2,3]. Tots ells s'orienten cap a l'aplicació de OSC en processos de transferència de calibració i en determinacions quantitatives, ambdues basades en dades espectrals NIR.

En els dos primers articles publicats per Wold i Sjöblom[1,2], es suggereix una estratègia de càlcul simil·lar a **NIPALS** (*Non Iterative Partial Least Squares*) per a l'eliminació de la informació ortogonal a la propietat analítica a determinar. El procés s'inicia amb el càlcul del vector de *scores* **t**, corresponent al primer component d'un anàlisi en

[A] Treball en procés de publicació

components principals (PCA) de les dades **X**. L'obtenció de la component ortogonal a aquest vector de *scores* (**t***) s'aconsegueix a partir de les possibilitats que ofereix l'algoritme NIPALS al construir un model de regressió parcial per mínims quadrats (*Partial Least Squares Regression*, PLSR). Per norma general, a mesura que el procés iteratiu de NIPALS avança, és possible modificar el vector de pesos **w** de manera que aquest compleixi amb certes restriccions. D'aquesta manera, mentre a PLS es busca la maximització de la covariància entre les matrius **X** i **Y**, a OSC es busca la seva minimització, o el que és el mateix, que **t** i **Y** siguin el més ortogonals possible.

Un cop el procés iteratiu ha assolit la convergència desitjada, es pot considerar que els *scores* ortogonals (**t***) que s'han obtingut són bons descriptors de la component de **X** que és ortogonal a **Y**. És a partir d'aquest vector de *scores* **t*** que es resol l'equació **Xw**=**t*** fent servir PLSR com a tècnica d'ajust. D'aquesta manera s'aprofita la capacitat d'aquest mètode numèric per a modelar la informació ortogonal continguda a la matriu de variables predictores **X**. Seguidament, aquesta informació ortogonal és restada de la matriu **X**. Aquest procés de modelat i resta d'informació proporciona, per una banda, una nova matriu **X***, de la que s'ha extret tota la informació ortogonal a la direcció de màxima variabilitat de la variable resposta. Per l'altra, proporciona els vectors necessaris per a la extracció d'aquesta informació d'altres conjunts externs de dades, especialment importants en processos de predicció/validació. Tot aquest procés comporta l'eliminació del primer **factor OSC**. Per a extreure'n un segon, es parteix de les dades ja tractades pel primer factor (**X***) i es repeteix el procés des del començament.

Als articles de Wold i Sjöblom també es suggeria una estratègia alternativa pel càlcul dels *scores* ortogonals, basada en la ortogonalització inicial de les matrius **X** i **Y** i la posterior desomposició de la matriu resultant mitjançant PCA. Tot i la viabilitat d'aquesta proposta, davant la impossibilitat de trobar una solució igualment viable per la correcció de mostres externes (altrament indispensable per a tot procés de calibració/validació), van preferir la utilització de l'estratègia NIPALS ja esmentada.

A l'article publicat per T.Fearn[3] en canvi, es segueix la filosofia de l'**ortogonalització inicial** de les matrius **X** i **Y** i es proposa una rutina que supera els obstacles en l'ortogonalització de dades externes a l'etapa de validació, permetent l'obtenció dels *scores* ortogonals d'aquestes dades. L'autor suggereix un enfoc diferent pel procés d'ortogonalització, que imposa una sèrie de restriccions en el càlcul del vector de pesos **w** i que permet calcular el valor dels *scores* ortogonals (**t***) i dels *loadings* (**p**). Aquests vectors, a més de permetre la correcció de la matriu de variables predictores inicial, permeten la correcció de matrius de dades externes, la qual cosa afavoreix la

retirada de la informació ortogonal de les possibles matrius externes (predicció/validació). En aquest cas, el concepte de factor OSC és diferent, i ve a ser equivalent al component principal obtingut al fer la descomposició en components principals dins l'algoritme anterior.

En ambdós casos, un dels principals inconvenients que pot presentar OSC és la seva tendència al sobreajust, especialment important en el cas d'aplicacions quantitatives. Les diferències existents entre ambdós tipus de rutines (NIPALS/ortogonalització inicial) fan que el camí a recórrer per a evitar problemes d'aquest tipus sigui diferenciat.

Per a evitar els problemes de sobreajust és necessari un control adequat del número de factors OSC. L'avantatge que presenta la via **NIPALS** envers l'**ortogonalització inicial** és que l'efecte d'aquest factor OSC pot arribar a ser matisat a través d'una adeqüada selecció del número de components PLS[B] que modelen la informació ortogonal[C]. Aquest efecte matisant pot tenir diferents conseqüències. Si el número de variables latents és massa baix, faran falta diversos components OSC per a retirar la informació no correlacionada amb la concentració. En canvi, si el nombre de variables latents és massa alt, la solució obtinguda per a la regressió de **X** sobre **t**** s'aproximarà a una solució MLR. En aquest cas, en un únic pas s'aconseguirà un nivell de correcció excessiu, provocant el sobreajust de les dades.

En aquest treball es proposen dues estratègies per a evitar sobreajustos a la correcció OSC (via **NIPALS**), basades en la validació de la regressió de **X** vs. **t****. Aquest procés de validació 'intern' de OSC, hauria de permetre limitar l'extensió del procés de modelat de la informació ortogonal. Amb aquesta intenció s'estudien adaptacions de les metodologies de *test set* i de validació creuada, per a la selecció adeqüada del número de variables latents. Els efectes d'ambdues propostes s'estudien amb la seva aplicació a dos conjunts d'espectres NIR obtinguts mitjançant tècniques de mesura diferents (Reflectància i Transmitància). Ha estat així perquè la naturalesa de la component ortogonal no té perquè ser la mateixa en cada cas i l'efectivitat del pretractament tampoc.

[B] Per a evitar possibles confusions, aquests components PLS, han estat anomenats a partir d'aquest punt mitjançant el sinònim **variables latents**.

[C] Necessàries per a la solució del sistema **Xw**=**t****.

Metodologia Experimental

El conjunt d'espectres NIR en mode reflectància conté els espectres de 80 mostres de blat de moro enregistrades en l'interval 1100-2500 nm amb una resolució espectral de 2 nm. Aquest conjunt rep el nom de *corn.mat*[4] i va ser proporcionat per M. Blackburn i distribuït a través d'internet per Eigenvector Research Ltd. Aquest conjunt d'espectres incorpora, per a cada mostra, els valors de referència individuals de diferents paràmetres: l'oli, la proteïna total, el midó i la humitat, essent aquesta última la que va ser utilitzada pel nostre treball.

Amb la intenció de comprovar si les estratègies de selecció del número de variables latents eren igualment vàlides per a dades espectrals de característiques i orígens diferents, les dues es van aplicar a un segon conjunt de dades, aquest cop enregistrades per transmitància. Aquestes dades corresponien als espectres NIR de 48 mostres d'oli verge d'oliva, enregistrats per duplicat fent servir una cubeta de 4mm de camí òptic. Cadascuna de les mesures individuals va ser el resultat de l'acumulació de 32 escombrats en l'interval 1100-2500 nm, fent servir una resolució espectral de 2 nm. Totes aquestes mesures es van realitzar en un instrument NIRSystems 6500 (FOSS), disposant d'un mòdul de cubeta com a suport i el dispositiu NR6513-A com a detector de transmissió. A cadascuna de les mostres se'ls va determinar el percentatge d'àcid oleic (C18:1) sobre el total d'àcids grassos a la mostra, fent servir el procediment descrit a la legislació europea[5].

Organització de les dades

Els 80 espectres de reflectància es van dividir en dos subconjunts: un de 41 mostres, com a conjunt de dades espectrals de calibració i un altre de 39, que conformava el conjunt de dades de predicció. Ambdós conjunts van ser seleccionats de manera que l'interval de concentracions del paràmetre a determinar (humitat) estigués cobert de la forma més uniforme possible. Mantenint el mateix criteri de distribució uniforme, el conjunt de predicció va ser subdividit en dos subconjunts: un de validació, amb 19 espectres i un altre de predicció, integrat per 20 espectres més. La finalitat d'aquesta subdivisió no va ser una altra que la de poder disposar d'un conjunt de dades independent que permetés avaluar la capacitat predictiva del model quantitatiu, un cop aquest hagués estat definit.

Del total de 48 espectres de transmitància, 28 van passar al conjunt de dades de calibració i 20 al de predicció. De la mateixa manera que en el cas anterior, els espectres per a ambdós subconjunts van ser seleccionats de forma que l'interval de

concentracions del paràmetre analític quedés cobert de forma uniforme per cadascun dels dos subconjunts. Per les mateixes raons que en el cas dels espectres de reflectància, el conjunt de predicció va ser subdividit en dos, originant els subconjunts de validació i de predicció (de 10 espectres cadascun).

Efecte de les variables latents en el sobreajust

Amb la finalitat de comprovar quin era l'efecte de les variables latents del procés de modelat de la informació ortogonal en una determinació quantitativa, es van tractar dels dos conjunts (calibració i validació) amb un únic factor OSC i fent servir un número de variables latents creixent. Seguidament, es van calcular models PLSR amb els diferents conjunts de dades tractades, fent servir diferent número de components PLS. Per a avaluar els resultats quantitatius obtinguts es va fer servir el valor de RMSE de predicció per a cadascun dels dos conjunts (calibració i validació) i se'n va fer la representació gràfica.

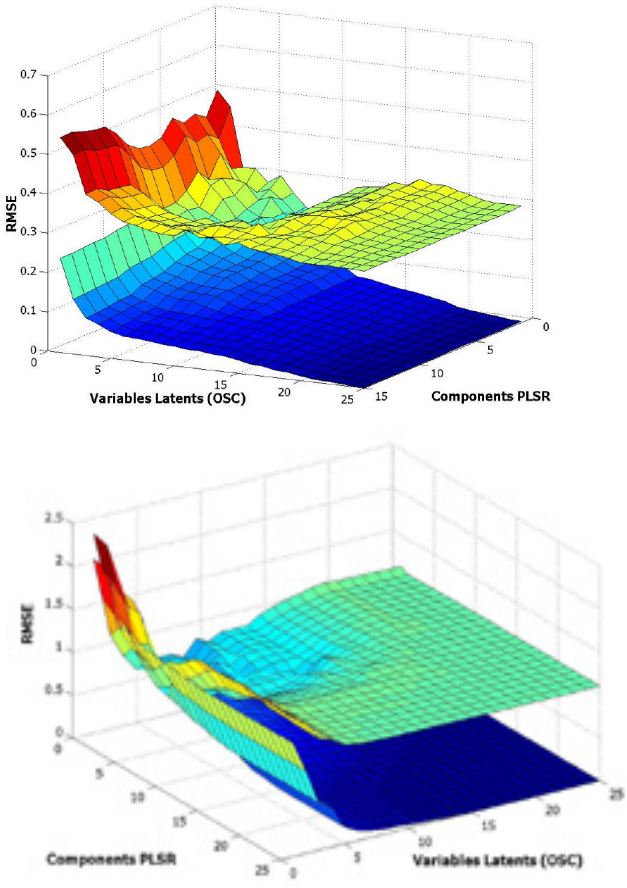

Figura AII.1. Superfícies d'error (RMSE) en calibració i predicció per a dades NIR de reflectància (superior) i de transmitància (inferior) fent servir diferent número de variables latents i un sol factor OSC.

Com a un primer acostament al problema, tant el conjunt de dades de reflectància com el de transmitància va ser utilitzat de forma directa, sense portar a terme cap selecció de variables prèvia. Tal i com es veu a la figura AII.1, les superfícies d'error calculades, tant per les dades de reflectància com per les d'absorbància demostren un perfil simil·lar. En ambdós casos s'observa l'existència d'una sèrie de condicions de número de variables latents i de components PLS del model

quantitatiu final en les que l'error de predicció del conjunt de validació divergeix clarament de l'error del conjunt de calibració. De la mateix manera, també es comprova l'existència d'un interval de condicions on la diferència entre els errors es fa mínima. En una primera aproximació, el que es veu fa pensar en l'existència d'una certa relació entre el número de components PLS i el de variables latents.

Quan el número de variables latents utilitzat per a tractar les dades mitjançant OSC és petit, hi ha una relació clara entre la capacitat predictiva i el número de components PLS utilitzats a l'equació quantitativa final. Aquesta relació sembla lògica, donat que la informació ortogonal no retirada ha de ser tinguda en compte per l'equació de calibració i es necessita augmentar el número de components principals per a fer-ho. Si s'incrementa aquest número de components PLS de forma excessiva, apareix el sobreajust de les dades degut a l'equació de calibració. Quan el número de variables latents seleccionat és l'adequat, s'elimina un percentatge alt de la informació ortogonal. Aquesta eliminació permet obtenir uns resultats quantitatius força bons fent servir un número baix de components PLS a l'equació quantitativa. Si el número de variables latents és clarament excessiu, el sobreajust és pràcticament independent del número de components PLS utilitzats. En aquests dos darrers casos, en els que la selecció del número de variables latents és propera a l'adequada o lleugerament excessiva, el possible sobreajust queda lligat al número de variables latents utilitzades a la correcció.

Per tant, s'observa com hi ha zones de les superfícies d'error amb diferents condicions on l'efecte de sobreajust pot considerar-se molt reduït i dependent del número de variables latents del procés de filtrat. Aquest conjunt d'observacions fa palesa la necessitat de disposar d'alguna estratègia que permeti determinar aquestes condicions de forma ràpida i de tal manera que s'evitin els sobreajustos deguts a una selecció incorrecta del número de variables latents.

Test Set

En un apartat del rerafons teòric s'ha fet referència a aquest tipus de metodologia per a determinar les condicions predictives idònies d'una equació de calibració. A títol de recordatori, és una estratègia de validació externa que avalua la capacitat predictiva de l'equació calculada en determinades condicions, aplicada a un conjunt de dades externes '*test set*'. Aquestes dades, tot i que no participen en el procés de calibració són les responsables de les seves condicions finals (número de components PLS, interval espectral).

A l'aplicació d'aquest concepte a OSC, el primer pas comporta el centrat d'un únic conjunt de variables predictores (dades espectrals) **X** i de variables resposta (concentració) **Y**, que conté dos subconjunts $\mathbf{X}_c$ i $\mathbf{X}_{ts}$, i d'un vector de concentracions únic (**Y**), que es pot desglosar en $\mathbf{Y}_c$ i $\mathbf{Y}_{ts}$. A partir de les dades centrades del conjunt de variables predictores es calcula el primer component principal, via SVD, la qual cosa permet obtenir els *scores* **t** d'aquest a partir dels vectors de *loading* calculats.

Seguidament, es centren els *scores* segons

$$\mathbf{t} = \mathbf{t} - \bar{\mathbf{t}} \qquad [1]$$

i s'ortogonalitzen envers les seves concentracions (ja centrades)

$$\mathbf{t}' = \mathbf{t} - ((\mathbf{y}^T\mathbf{t})/(\mathbf{y}^T\mathbf{y})) * \mathbf{y} \qquad [2]$$

Els *scores* ortogonals obtinguts **t**', es reescalen segons

$$\mathbf{t}^* = \mathbf{t}' + \bar{\mathbf{t}} \qquad [3]$$

En el següent pas, es calcula el valor del vector de *loading* ortogonal i s'escala a longitud unitat, segons

$$\mathbf{p}^* = (\mathbf{x}^T\mathbf{t}^*)/(\mathbf{t}^{*T}\mathbf{t}^*) \qquad [4]$$
$$\mathbf{p}^* = \mathbf{p}^*/\sqrt{\mathbf{p}^{*T}\mathbf{p}^*}$$

A partir del valor de **p*** obtingut es calcula un nou valor per **t** i es repeteix el procés iteratiu des de (1) a (4) fins que el valor de la diferència entre el vector de *scores* ortogonals (**t***), obtingut al final de cada procés iteratiu, i el valor de **t** calculat no presenta diferències apreciables respecte el calculat en el cicle anterior. Al final d'aquest procés iteratiu, els *scores* **t*** que s'obtenen resulten ser bons descriptors de la informació ortogonal continguda en el conjunt de dades inicial.

Aleshores, desglosant la matriu de *scores* ortogonals **t*** en els dos subconjunts incials ($\mathbf{t}_c^*$ i $\mathbf{t}_{ts}^*$), es fan servir $\mathbf{X}_c$ i $\mathbf{t}_c$* com a matrius de calibració en el procés d'obtenció de la equació PLSR que relaciona la matriu de variables predictores amb els *scores* ortogonals. Les equacions obtingudes mitjançant PLS, fent servir diferent número de variables latents, s'utilitzen per a calcular novament els valors de *score* del conjunt de test ($\hat{\mathbf{t}}_{\mathbf{ts}}^*$). Establint una comparació entre aquests valors calculats i els valors obtinguts inicialment es pot fer una estimació del número de variables latents

necessari. Per a fer-ho, es fa servir el valor de PRESS per als valors de *score* del conjunt de test

$$\mathbf{PRESS} = \sum_{i=1}^{N} (\mathbf{t}^{*}_{\mathbf{ts}_i} - \hat{\mathbf{t}}^{*}_{\mathbf{ts}_i})^2$$

On $\hat{\mathbf{t}}^{*}_{\mathbf{ts}_i}$ són cadascun dels *i scores* ortogonals calculats via PLS pel conjunt de test i $\mathbf{t}^{*}_{\mathbf{ts}_i}$ són els valors calculats inicialment al tractar tot el conjunt d'espectres (calibració i test) alhora, fent servir el mateix número de variables latents.

A partir d'aquest estadístic global de validació, es pot seleccionar el número de variables latents tenint en compte el valor del mínim de PRESS per les variables latents calculades[D]. Conegut el número de variables latents seleccionat (n), es calcula el vector de regressors PLS (**rv**), cosa que permet el recàlcul d'un nou vector de *scores* ortogonal **t****. Aquest nou vector conté la informació ortogonal modelada per n variables latents. Es calcula segons

$$\mathbf{t}^{**} = \mathbf{x} * \mathbf{rv} / (\mathbf{rv}^{\mathbf{T}} * \mathbf{rv})$$

El fet de disposar d'aquests *scores* **t****, permet el càlcul d'un nou vector de *loading* (**p**) que respon a la informació ortogonal modelada. Aquest, conjuntament amb el vector de regressors PLS, és necessari per a la correcció de conjunts de dades externs. Es calcula segons

$$\mathbf{p} = \mathbf{x}^{\mathbf{T}} * \mathbf{t}^{**} / (\mathbf{t}^{**^{\mathbf{T}}} * \mathbf{t}^{**})$$

A partir d'aquests descriptors de la informació ortogonal a la concentració, aquesta ja pot ser retirada de forma efectiva de la matriu de variables predictores (**X**)

$$\mathbf{X}_{\mathbf{OSC}} = \mathbf{X} - \mathbf{t}^{**} * \mathbf{p}^{\mathbf{T}}$$

Quan es vol retirar un segon factor OSC, es repeteixen tota aquesta sèrie de passos, però partint de $\mathbf{X}_{OSC}$ en lloc de la matriu **X**.

Validació Creuada

La validació creuada segueix la mateixa metòdica exposada en el rerafons teòric. Resumint, es fan servir de forma seqüencial les variables espectrals corresponents als registres de diferents mostres de calibració per a definir les condicions predictives

[D] Veure l'apartat 1.4.1.5. del rerafons teòric: *Avaluació de resultats. Determinació de les condicions finals de les equacions de calibració.*

idònies de l'equació final. La utilització de les pròpies dades de la calibració la fa especialment indicada per aquells casos en els que la quantitat de dades disponibles és limitada.

Un cop s'han obtingut els *scores* ortogonals **t*** pel procediment iteratiu assenyalat a l'apartat anterior (des de (1) fins a (4)), es calcula la regressió de **X** vs. **t***. D'aquesta manera es calcula el vector de regressors que relaciona **X** amb **t*** i que respon a la informació ortogonal que teòricament ha de ser retirada. En aquest punt es fa servir la metodologia de validació creuada coneguda com a *'leave one out'* per a determinar les condicions finals de l'equació de modelat de **X** vs. **t***. A cada pas de la iteració es retira l'espectre d'una mostra *i* diferent del conjunt de dades de calibració i es calcula el valor del seu *score* ortogonal $\mathbf{t}_i$* a partir de l'equació provisional obtinguda amb les dades de la resta de mostres. Per a calcular el valor de PRESS es compara aquest valor de $\mathbf{t}_i$* amb el que s'obté per a la mateixa mostra *i* quan es tracta tot el conjunt de mostres alhora, fent servir el mateix número de variables latents. Per a seleccionar les condicions òptimes de la regressió es segueix l'evolució del PRESS dels *scores* $\mathbf{t}_i$* envers el número de variables latents utilitzades. Es selecciona com a número de variables latents adequat aquell en el que el valor del PRESS presenta un primer mínim. Un cop seleccionat aquest número es prepara el vector de regressors final que permet extreure la informació ortogonal.

Aplicació

Els algoritmes proposats per la validació de les condicions de filtrat OSC es van aplicar als dos conjunts de dades NIR ja esmentats anteriorment. Així, en el primer cas es va aplicar a la determinació d'humitat en un conjunt de mostres de blat de moro. En el segon, les dades de transmitància es van fer servir per a la determinació, en oli d'oliva verge, del percentatge d'àcid oleic sobre el total d'àcids grassos lliures.

L'equació de calibració per a relacionar $\mathbf{X}_{OSC}$ amb la variable resposta (paràmetre a determinar) es va calcular per PLSR, fent servir els conjunts de dades espectrals de calibració per a la determinació dels paràmetres *i* de validació per a determinar les condicions de número de components PLS. Aquesta determinació es va fer a partir del valor de PRESS, i per a establir comparacions entre les capacitats predictives de les diferents equacions de calibració es va fer servir el valor de RMSE ('*Root Mean Standard Error*') que en aquest cas es va expressar

$$\mathbf{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (\mathbf{y}_{\mathbf{ts}_i} - \hat{\mathbf{y}}_{\mathbf{ts}_i})^2}{\mathbf{n}}}$$

On $\mathbf{y}_{tsi}$ era el valor de la variable resposta per la mostra *i* del conjunt de validació i $\hat{\mathbf{y}}_{\mathbf{ts_i}}$ era el valor de la variable resposta calculat per l'equació de calibració per a aquella mostra.

Contingut d'humitat en mostres de blat de moro

En la comprovació de l'estratègia de *test set* per a la determinació de les condicions òptimes del pretractament, es va aplicar OSC al conjunt de dades centrades de calibració, seguint el valor de PRESS vs. el número de variables latents implicades en el procés de calibració de $\mathbf{X}$ vs $\mathbf{t}_c$*.

A la figura AII.2a es pot veure gràficament el resultat de representar el valor de PRESS del conjunt de validació envers el número de variables latents. Es detecta un mínim de PRESS a 9 variables latents. Paral·lelament, a la taula AII.1 es mostren els valors numèrics de RMSE pels tres conjunts de dades (calibració, validació i predicció), quan es fan servir diferent número de variables latents. Es comprova com les dades tractades fent servir el número seleccionat proporcionen els valors de RMSE més baixos i, alhora, són resultats independents del nombre de components PLS seleccionats per l'equació de calibració final.

A l'aplicació de la metodologia basada en **validació creuada**, en la seva variant '*leave one out*', per a determinar el número òptim de variables latents també es va fer servir el valor de PRESS calculat en el procés de validació. Per a això es va dibuixar el valor de PRESS vs. el número de variables latents implicades en el procés de modelat de la informació ortogonal i com a criteri de selecció també es va agafar el mínim valor de PRESS. Tal i com es veu a la figura AII.2b, aquest valor mínim es presenta a 10 variables latents. Aquest valor no difereix pràcticament de l'obtingut en la validació del procés d'ortogonalització per *test set*. En ambdós casos els nivells de correcció són molt simil·lars, la qual cosa fa suposar que les capacitats predictives d'ambdues equacions de calibració haurien de ser molt simil·lars. Aquesta suposició es confirma amb els resultats de la taula AII.1.

Valors de RMSE per la Humitat									
Nº de Variables Latents	*Components PLS*								
	RMSEC			*RMSETS*			*RMSEP*		
	1	*2*	*3*	*1*	*2*	*3*	*1*	*2*	*3*
1	0.34	0.31	0.30	0.50	0.44	0.46	0.43	0.40	0.39
2	0.28	0.24	0.22	0.46	0.39	0.42	0.38	0.31	0.30
3	0.22	0.20	0.18	0.31	0.32	0.37	0.25	0.25	0.26
4	0.19	0.18	0.18	0.31	0.28	0.31	0.23	0.24	0.24
5	0.16	0.15	0.15	0.28	0.27	0.30	0.22	0.22	0.21
6	0.14	0.14	0.13	0.30	0.28	0.26	0.19	0.20	0.19
7	0.13	0.13	0.13	0.26	0.26	0.28	0.18	0.19	0.19
8	0.12	0.12	0.11	0.23	0.23	0.24	0.17	0.17	0.16
9	0.10	0.10	0.10	0.21	0.21	0.22	0.13	0.13	0.13
10	0.10	0.10	0.10	0.23	0.23	0.23	0.14	0.14	0.14
11	0.09	0.09	0.09	0.27	0.26	0.25	0.18	0.18	0.15
12	0.07	0.07	0.07	0.30	0.29	0.29	0.18	0.18	0.18

Taula AII.1. Valors de RMSE per la variable resposta humitat. Efecte del número de variables latents utilitzades a la correcció sobre els resultats predictius. En blanc, resultats utilitzant la metodologia de ***test set***. En gris, utilitzant la **validació creuada**.

Valors de RMSE pel percentatge d'àcid Oleic									
Numero de Variables Latents	*Components PLS*								
	RMSEC			*RMSETS*			*RMSEP*		
	1	*2*	*3*	*1*	*2*	*3*	*1*	*2*	*3*
1	2.39	2.04	1.94	2.10	1.70	1.33	2.35	1.55	1.77
2	2.23	1.83	1.72	1.96	1.46	1.06	2.38	1.44	1.63
3	1.66	1.54	1.16	1.34	1.03	0.86	1.38	1.53	0.77
4	1.51	1.41	1.12	0.87	1.11	1.02	1.56	1.06	0.95
5	0.98	0.94	0.78	0.78	0.78	0.76	0.79	0.72	0.71
6	0.75	0.72	0.66	0.77	0.82	0.68	0.84	0.86	0.91
7	0.64	0.62	0.61	0.77	0.80	0.74	1.01	1.00	1.02
8	0.54	0.53	0.52	0.71	0.71	0.71	1.30	1.27	1.26
9	0.50	0.48	0.48	0.88	0.90	0.88	1.24	1.24	1.24
10	0.45	0.43	0.42	0.82	0.83	0.82	1.28	1.28	1.28
11	0.37	0.34	0.34	0.89	0.90	0.89	1.52	1.55	1.55
12	0.36	0.33	0.33	0.89	0.90	0.89	1.54	1.58	1.58

Taula AII.2. Valors de RMSE pel percentatge d'àcid oleic sobre el total d'àcids grassos. Efecte del número de variables latents utilitzades a la correcció sobre els resultats predictius. En blanc, resultats utilitzant la metodologia de ***test set***. En gris, utilitzant la **validació creuada**.

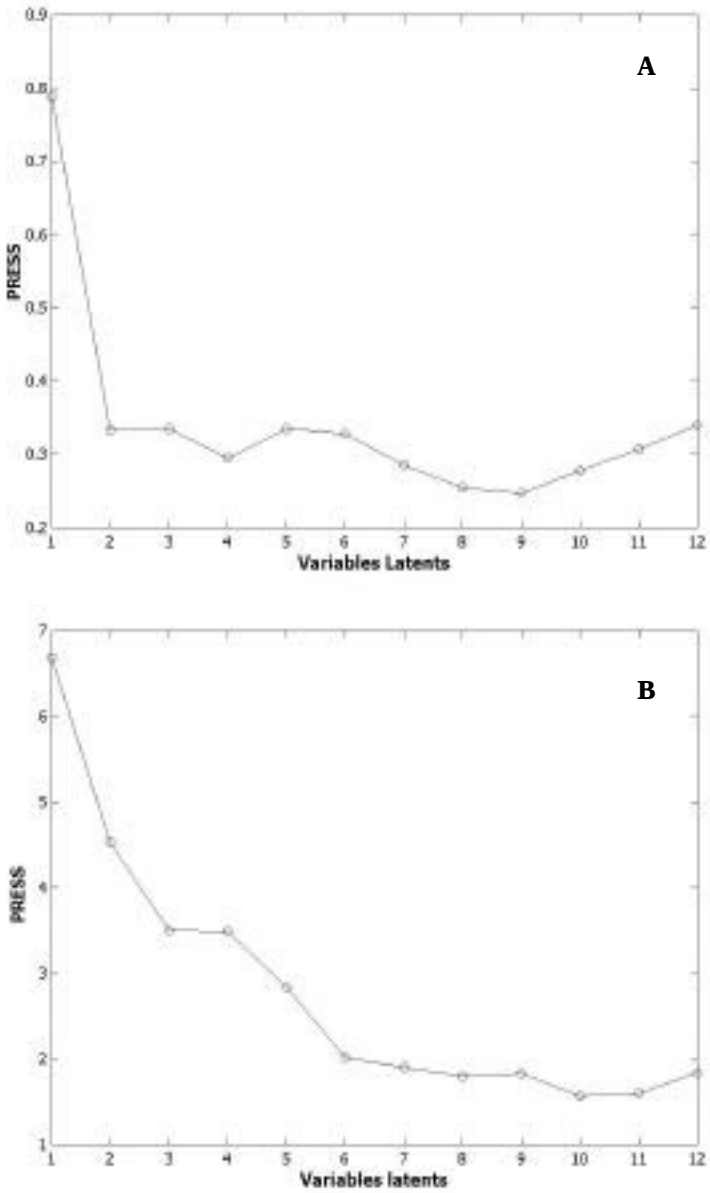

Figura AII.2. Valors de PRESS calculats fent servir la variant de Test Set (A) i de Validació Creuada (B) per a dades de reflectància.

De la comparació entre els resultats de calibració i predicció es desprèn que els errors de predicció són independents del número de components PLS de l'equació de calibració en aquells casos en els que el número de variables latents s'acosta a l'òptim i que no presenten un nivell de sobreajust important. Per altra banda, la capacitat predictiva fent servir dades filtrades amb el número de variables latents seleccionat per *test set* o validació creuada és pràcticament la mateixa.

Contingut d'àcid oleic en mostres d'oli d'oliva verge

Es va seguir exactament el mateix procediment que en el cas anterior, aplicant primerament la metodologia de *test set* a un conjunt de dades espectrals centrades, amb la intenció de determinar el número de variables latents necessàries per l'aplicació òptima de OSC.

En aquest cas, tal i com es pot veure a la figura AII.3a, el valor mínim de PRESS es va obtenir per a 6 variables latents. A la taula AII.2 es presenten els resultats predictius calculats fent servir diferent número de variables latents seleccionades. Es pot veure com quan el número de variables latents de la correcció és inferior al seleccionat, els resultats no són independents del número de components PLS. En canvi, quan el número de variables latents s'acosta al determinat com a òptim, els resultats sí que ho són.

En el segon pas, en el que es va aplicar la metodologia de **validació creuada** en la seva variant *'leave one out'*, al buscar el mínim valor de PRESS es va veure que aquest es trobava a 9 variables latents. De tota manera, i tal i com es pot veure a la figura AII.3b, els resultats presenten poques diferències entre 6 i 9 variables latents. Degut a això, es poden considerar 6 variables latents com a suficients per a fer la correcció. Tal i com es comprova a la taula AII.2, aquesta selecció va permetre obtenir nivells d'error satisfactoris, sense caure en sobreajustos no desitjats.

Efecte d'altres factors OSC

El fet d'aplicar OSC correctament fa suposar que, si s'ha retirat suficientment la informació ortogonal, hauria de ser innecessaria l'addició d'un segon factor OSC, perquè es pot considerar que la major part de la informació ortogonal a la concentració ja ha estat retirada amb el primer. D'aquesta manera, s'ha volgut comprovar com l'ús d'un únic factor OSC calculat a partir d'un número adequat de variables latents és suficient per a extreure la major part de la informació ortogonal continguda al senyal,

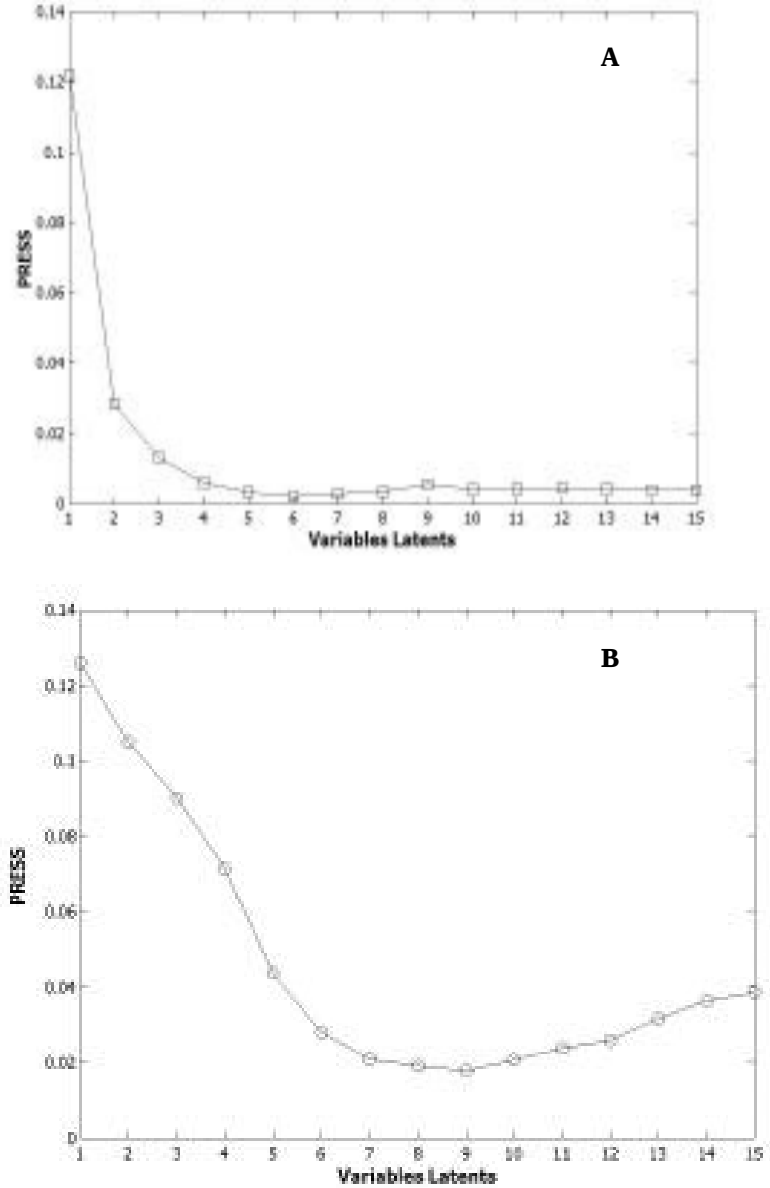

Figura AII.3. Valors de PRESS calculats fent servir la variant de *test set* (A) i de validació creuada (B) per a dades de transmitància

sense presentar sobreajusts a la determinació quantitativa. Amb tal finalitat, partint del conjunt de dades d'absorbància, es va aplicar la correcció OSC amb un factor, determinant el número de variables latents per *test set*. Un cop tractades aquestes dades, se'ls va aplicar un segon factor OSC seleccionant el número de variables latents de la mateixa manera. A l'observar detingudament les corbes de PRESS per a cadascun dels dos factors OSC, es pot veure com el segon factor no retira gaire informació addicional a la ja retirada pel primer factor en el seu mínim de PRESS (6 variables latents). Gràficament, es pot veure com els valors de PRESS pel segon factor quan es fan servir dues variables latents són pràcticament equiparables als obtinguts amb un únic factor i 6 variables latents.

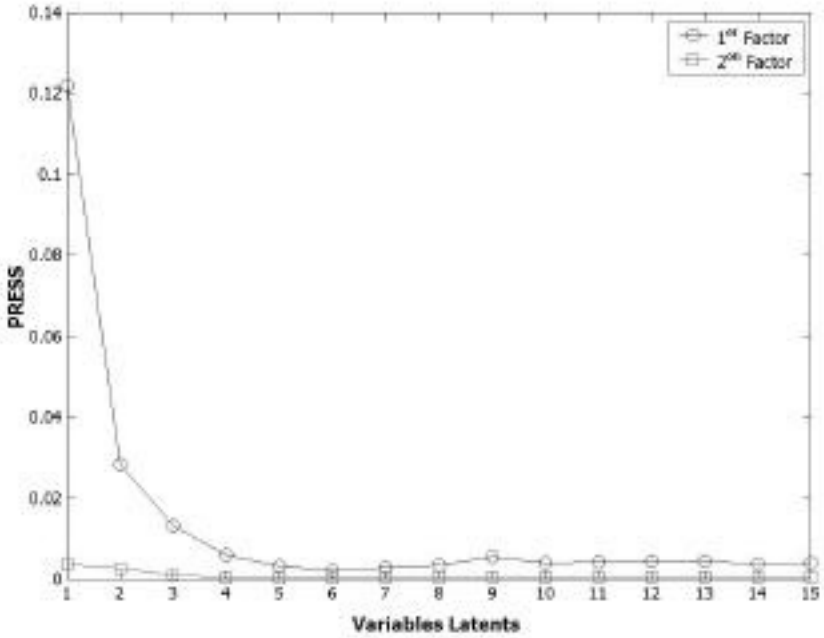

Figura AII.4. Comparació dels gràfics de PRESS fent servir dos factors OSC.

Per a comprovar la capacitat predictiva en models quantitatius que fan servir espectres tractats amb dos factors OSC, s'han calculat equacions de predicció fent servir PLSR. S'ha partit de dades tractades amb un factor OSC i 6 variables latents i, números creixents de variables latents pel segon factor, fins un total de 10. Per a la determinació del número de components PLS òptims per les equacions de calibració s'ha pres com a criteri el mínim valor de RMSE del conjunt de validació (*test set*).

Els resultats quantitatius obtinguts s'adjunten a la taula 3 i es comparen amb els obtinguts fent servir un únic factor. Es pot veure com l'efecte del sobreajust de les

dades de calibració és creixent a mesura que augmenta el número de variables latents i com els errors calculats fent servir dos factors amb el mateix número de variables latents són del mateix ordre que els obtinguts amb un únic factor, tot i que lleugerament més elevats.

Variables Latents		**Comp. PLS**	***RMSEC***	***RMSEP***
OSC 1	OSC 2			
6	-	3	0.66	0.91
6	1	6	1.80	1.38
6	2	2	0.70	0.87
6	3	2	0.69	0.88
6	4	2	0.69	0.88
6	5	2	0.66	0.91
6	6	2	0.65	0.95
6	7	2	0.61	1.01
6	8	2	0.59	1.02
6	9	2	0.55	1.06
6	10	2	0.50	1.24

Taula AII.3. Valors de RMSE a l'utilitzar 2 factors OSC.

Conclusions

En aquest treball es va estudiar l'efecte de les variables latents de la regressió de **X** vs. $\mathbf{t}_c$* a la utilització de OSC lligada a aplicacions quantitatives. Per a fer-ho, es va partir de dos conjunts de dades NIR de diferents orígens.

En els casos estudiats, es va observar un comportament semblant en quant als errors quantitatius calculats. Per les dades de calibració, es donava una disminució constant de l'error de calibració a l'augmentar tant el número de variables latents de la correcció com el número de components PLS de l'equació de calibració quantitativa. En canvi, pels errors de predicció es presentava una zona de variables latents i de components PLS més o menys àmplia, en la que l'error es feia mínim. Fora d'aquests límits, l'error creixia donant lloc a un clar efecte de sobreajust. Aquest comportament es va observar tant en dades de reflectància com en dades de transmitància, la qual cosa va fer suposar que ambdós tipus de senyal contenien informació ortogonal que podia ser modelada en el procés de correcció.

Les estratègies de *test set* i de validació creuada van semblar ser bons mètodes per a la determinació de les condicions òptimes de filtrat. En el procés de selecció del

número de variables latents ambdues estratègies van portar a resultats molt semblants, la qual cosa les fa pràcticament equivalents. La principal diferència entre elles és de tipus general i està relacionada amb el número de mostres necessàries per aplicar a cadascun dels procediments. Des d'aquest punt de vista, la utilització de la validació creuada permetria seleccionar el número de variables latents en aquells casos en els que el nombre de mostres no fos gaire elevat. En canvi, en aquells casos en els que aquest no fós problema, i es pogués disposar d'un conjunt de validació, el procediment de *test set* seria perfectament vàlid.

Es va comprovar que un únic factor OSC, fent servir un nombre adequat de variables latents, era suficient per a extreure la major part de la informació ortogonal continguda a la matriu espectral. El fet d'afegir un segon factor no proporcionava resultats predictius significativament diferents i sí una lleugera tendència al sobreajust.

Finalment, la simetria de les superfícies d'error calculades va fer pensar que quan s'utilitza PLS com a tècnica de calibració conjuntament amb OSC, el seu principal efecte és el de la reducció del número de components PLS necessaris per l'equació final mantenint la capacitat predictiva. Tenint en compte els resultats, la complexitat d'aquesta correcció no justifica, avui per avui, la seva utilització en la majoria de problemes quantitatius resolts mitjançant PLS. De tota manera, altres camps com els de les transferències de calibracions entre instruments poden beneficiar-se de la seva aplicació. En aquests casos concrets, les estratègies proposades per a prevenir la sobrecorrecció de les dades poden arribar a ser força útils.

[1] J. Sjöblom, O. Svensson, M. Josefson, H. Kullberg i S. Wold, *Chemometrics and Intelligent Laboratory Systems*, **44**, 229, (1998).

[2] S. Wold, H. Antti, F. Lindgren i J. Öhman, *Chemometrics and Intelligent Laboratory Systems*, **44**, 175, (1998).

[3] T. Fearn, *Chemometrics and Intelligent Laboratory Systems*, **50**, 47, (2000).

[4] http://www.eigenvector.com . Darrera Consulta: 11.07.2001.

[5] *Directiva UE*, **2568**/**91**, No L 248 (1991).