Universidad Nacional de Rosario

Facultad de Ciencias Bioquímicas y Farmacéuticas

Tesis de Doctorado en Ciencias Químicas

"*Calibración analítica multidimensional: estudio de cifras de mérito y desarrollo de nuevos algoritmos*"

Presentada por Franco Allegrini

Director: Dr. Alejandro C. Olivieri

Rosario, Argentina

-2015-

“*Calibración analítica multidimensional: estudio de cifras de mérito y desarrollo de nuevos algoritmos*”

Franco Allegrini

Licenciado en Química

Universidad Nacional de Rosario

Esta Tesis es presentada como parte de los requisitos para optar al grado académico de Doctor en Ciencias Químicas, de la Universidad Nacional de Rosario y no ha sido presentada previamente para la obtención de otro título en esta u otra Universidad. La misma contiene investigaciones llevadas a cabo en el Instituto de Química de Rosario, en el departamento de Química Analítica, dependiente de la Facultad de Cs. Bioquímicas y Farmacéuticas, durante el período comprendido entre Abril de 2012 y Diciembre de 2015, bajo la dirección del Dr. Alejandro C. Olivieri.

Franco Allegrini

DNI 32.500.806

AGRADECIMIENTOS Y DEDICATORIAS

A mi familia, por su apoyo incondicional.

A mis amigos, por acompañarme en todo momento y ser mi "cable a tierra".

A mis colegas y compañeros de trabajo en Argentina, por hacer agradable y llevadero el día a día.

A mis amigos, compañeros y colegas extranjeros a quienes conocí durante mis viajes al exterior como parte del desarrollo de esta tesis de doctorado, por haberme recibido con gran hospitalidad y calidez, aportando en gran medida a mi crecimiento humano y profesional.

A la Fundación Josefina Prats, por sus premios estímulo que contribuyeron tanto a la realización de mi tesina de grado, como de esta tesis doctoral.

Al consorcio Eurotango II por el otorgamiento la beca de intercambio que me permitió realizar parte de mi doctorado en en la ciudad de Lille, Francia.

A CONICET por el otorgamiento de la beca doctoral que permitió dedicarme en tiempo completo al desarrollo de mi tesis.

A la Universidad Nacional de Rosario y en particular a los docentes de la Facultad de Ciencias Bioquímicas y Farmacéuticas, principales responsables de mi formación profesional, que en última instancia posibilitó abordar la ardua empresa que significa llevar adelante un trabajo de tesis.

Un agradecimiento especial al Dr. Alejandro C. Olivieri, quien desde un principio, más allá de su indiscutida capacidad docente y profesional, supo transmitirme a través de su comportamiento, lecciones de paciencia, tolerancia, humildad y simpleza, mostrando ser un buen líder más que sólo un "buen director".

A mis abuelos Elcio Imbern y María Costa y a mi tío Raúl Imbern, que en paz descansen…

"La lluvia cae, la temperatura decrece, la inspiración se aproxima…" (Autor anónimo, región Nord Pas de Calais, Francia).

El presente trabajo de tesis dio lugar a las siguientes publicaciones científicas en revistas y capítulos de libros:

- Allegrini, F., and Olivieri, A. C. (2012) "Analytical Figures of Merit for Partial Least-Squares Coupled to Residual Multilinearization", *Analytical Chemistry*, 84, 10823-10830.

- Allegrini, F.; and Olivieri, A. C. (2013), "An integrated approach to the simultaneous selection of variables, mathematical pre-processing and calibration samples in partial least squares multivariate calibration" , *Talanta,* 115, 755-760.

- Allegrini, F., and Olivieri, A. C. (2014), "IUPAC-Consistent Approach to the Limit of Detection in Partial Least-Squares Calibration", *Analytical Chemistry*, 86, 7858-7866.

- Olivieri, A. C., Bortolato, S., Allegrini, F. (2015) "Figures of merit in multi-way calibration" in "Fundamentals and Analytical Applications of Multi-way Calibration", *Elsevier Physical Sciences Series*, Elsevier: Amsterdam, Capítulo 13.

- Allegrini, F., Wentzell P. D., Olivieri A. C. (2015) " Generalized error-dependent prediction uncertainty in multivariate calibration", *Analytica Chimica Acta.* En prensa.

También se presentaron los siguientes trabajos en reuniones científicas:

- Allegrini, F., Olivieri, A. C. "Cifras de mérito en calibración analítica multivariada: incertidumbre y límite de detección." 8vo. Congreso Argentino de Química Analítica. La Plata, Argentina. Modalidad de presentación: oral. Noviembre de 2015.

- Allegrini, F., Pisano, P., y Olivieri, A. C. "Estrategias de selección aplicadas a cuantificación y clasificación mediante cuadrados mínimos parciales (PLS)." 3er Congreso Uruguayo de Química Analítica. Montevideo, Uruguay. Modalidad de presentación: póster. Octubre de 2014.

- Allegrini, F., Bauza, C., Ibañez, G. y Olivieri, A. C. "Cifras de mérito en calibración multivariada de orden superior. Estudio teórico y experimental." 5to. Congreso Iberoamericano y 2do. Congreso Uruguayo de Química Analítica. Montevideo, Uruguay Modalidad de presentación: póster. Octubre de 2012.

ÍNDICE

ABREVIATURAS

En general, las convenciones utilizadas en esta tesis son las que se indican a continuación, coincidiendo con las que normalmente se utilizan en publicaciones relacionadas con esta área temática. En caso que la abreviatura coincida con la denominación en inglés, la misma se detalla debidamente en la tabla. Sin embargo, durante el transcurso del texto estas abreviaturas se denominarán de manera extendida en castellano. Al mismo tiempo, los vocablos específicos en inglés que aparecen en el texto, se diferencian utilizando itálicas.

Abreviatura	Denominación extendida en inglés	Denominación extendida en español
ACO	*Ant Colony Optimization*	Optimización por colonias de hormigas
ALS	*Alternating Least Squares*	Cuadrados mínimos alternantes
GA	*Genetic Algorithm*	Algoritmo genético
ACOGASS	*Ant Colony Optimization, Genetic Algorithm and Sample Selection*	Optimización por colonias de hormigas, algoritmo genético y selección de muestras
CLS	*Classical Least Squares*	Cuadrados mínimos clásicos
EEFM	*Excitation Emission Fluorescence Matrix*	Matrices de excitación emisión de fluorescencia
EIV	*Error in Variable*	Error en la variable
EVD	*Eigenvalue Decomposition*	Descomposición en valores singulares
KS	*Kennard Stone*	*Kennard Stone*
LOD	*Limit of Detection*	Límite de detección
MCR-ALS	*Multivariate Curve Resolution coupled to Alternating Least Squares*	Resolución multivariada de curvas acoplada a cuadrados mínimos alternantes
MLR	*Multiple Linear Regression*	Regresión lineal múltiple
MLPCR	*Maximum Likelihood Principal Component Regression*	Regresión por components principales basada en maxima probabilidad
MSC	*Multiplicative Scattering Correction*	Corrección de la dispersion multiplicativa
NIPALS	*Non linear iterative Partial Least Squares*	Método iterativo no lineal de cuadrados mínimos parciales
NIRS	*Near Infrarred Spectroscopy*	Espectroscopía de infrarrojo cercano
NAS	*Net Analyte Signal*	Señal neta del analito

ILS	*Inverse Least Squares*	Cuadrados mínimos inversos
iPLS	*Interval Partial Least Squares*	Cuadrados mínimos parciales por intervalo
PARAFAC	*Parallel Factor Analysis*	Análisis paralelo de factores
PCA	*Principal Component Analysis*	Análisis por componentes principales
PCR	*Principal Component Regression*	Regresión por componentes principales
PLS	*Partial Least Squares*	Cuadrados mínimos parciales
PRESS	*Predicted Error Sum of Squares*	Suma de cuadrados de los errores predichos
PSO	*Particle Swarm Optimization*	Optimización por enjambre de partículas
ISO	*International Standarization Organization*	Organización interna
IUPAC	*International Union of Pure and Applied Chemistry*	Unión internacional de química pura y aplicada
RBL	*Residual Bilinearization*	Bilinelización residual
RQL	*Residual Quadrilinerization*	Cuadrilinealización residual
RML	*Residual Multilinearization*	Multilinealización residual
RTL	*Residual Trilinearization*	Trilinealización residual
RMSEP	*Root Mean Squared Error of Prediction*	Raíz cuadrada del error cuadrado medio de predicción
RMSEP$_{mon}$	*Root Mean Square Error of Prediction in Monitoring set*	Raíz cuadrada del error cuadrado medio de monitoreo
RMSEP$_{cal}$	*Root Mean Square Error of Prediction in Monitoring set*	Raíz cuadrada del error cuadrado medio de monitoreo.
RMSECV	*Root Mean Squared Error of Cross Validation*	Raíz cuadrada del error cuadrado medio de validación cruzada
SEN	*Sensitivity*	Sensibilidad
SEN$_{FO}$	*Faber and Olivieri Sensitivity*	Sensibilidad en PARAFAC según Faber y Olivieri.
SEN$_{MKL}$	*Messik, Kalivas and Lang sensitivity*	Sensibilidad en PARAFAC según Messik, Kalivas y Lang.
SEN$_{HCD}$	*Ho, Christian and Davidson sensitivity*	Sensibilidad según Ho, Christian and Davidson.
SNV	*Standard Normal Variate*	Variación estándar normal
SPXY	*Sample Partitioning based on joint X-Y distances*	Partición de muestras basada en la distancia conjunta X-Y
SR	*Selectivity Radio*	Razón de selectividad
SS	*Sample Selection*	Selección de muestras
SSR	*Sum of Squared Residuals*	Suma de cuadrados residuales
SVD	*Singular Value Decomposition*	Descomposición por valores

		singulares
U-PLS	*Unfolded PLS*	PLS desdoblado
UVE-PLS	*Uninformative Variable Elimination in PLS*	Eliminación de variables que no aportan información en PLS
VIF	*Variance Inflation Factor*	Factor de inflación de la variancia
VIP	*Variable Importance in Projection*	Importancia de la variable en la proyección

LENGUAJE ESPECIAL

Dada la gran cantidad de ecuaciones y la diversidad de modelos de cuantificación abordados durante esta tesis, algunos de los símbolos coinciden para dos temas y/o modelos distintos. Es por este motivo que por cuestiones de mayor claridad se optó por dividir esta tabla de nomenclatura por capítulos. Los símbolos que se encuentran repetidos en varios capítulos se describen únicamente en el capítulo que aparecen inicialmente. En caso que en un nuevo capítulo se use el mismo símbolo que en anteriores, éste se describe nuevamente.

Las matrices se representan como letras mayúsculas en negrita y los vectores columna como letras minúsculas en negrita. Las magnitudes escalares, por otro lado, se representan en minúscula e itálica. Los símbolos que representen valores estimados de una cantidad desconocida, se designan a través de la mascarilla "$\widehat{\ }$". La transpuesta de una matriz se indica con el superíndice "T", y la norma euclidiana de un vector a través de "$\|\ \ \|$". La matriz inversa, por su parte, se indica utilizando el superíndice "-1". El resto de los caracteres modificadores y operadores que aparezcan se discuten oportunamente.

Parámetro	**Tamaño**	**Detalle**
Capítulo 1		
A	Escalar	Número de componentes o rango químico efectivo del sistema.
A^*	Escalar	Número de componentes para el cual el PRESS es mínimo durante la validación cruzada utilizando PLS.
a_{in}	Escalar	Valor de *score* proporcional a la concentración del constituyente n en la muestra i.
$\mathbf{a}_n$	$I\times1$	Vector de *scores* proporcionales a concentración cuando

		se trabaja en datos multi vía, para un componente *n* particular.
$a_{\text{int},n}$	Escalar	Factor de escalado de los perfiles de los interferentes obtenidos durante el procedimento de bilinearización residual para el interferente *n*
B	$J \times A$	Matriz de coeficientes de regresión en ILS.
b	Escalar	Coeficiente de regresión univariado (pendiente de la recta de calibrado).
b_{jn}	Escalar	Valor de respuesta específica dada por el canal instrumental *j* para el constituyente *n*. Definición análoga para c_{kn} en el canal instrumental *k*.
$\hat{b}$	Escalar	Coeficiente de regresión univariado estimado.
$\hat{\mathbf{b}}_{\text{ILS}}$	$K \times 1$	Vector de coeficientes de regresión estimados para un analito particular cuando se utiliza ILS. Definiciones análogas para $\hat{\mathbf{b}}_{\text{PLS}}$ y $\hat{\mathbf{b}}_{\text{PCR}}$, cuando se utilizan los modelos PLS y PCR.
$\mathbf{b}_n$	$J \times 1$	Perfil para un componente particular en el primer modo instrumental. Definición análoga para $\mathbf{c}_n$ en el segundo modo instrumental.
$\mathbf{b}_{\text{int},n}$	$J \times 1$	Perfil resultante de aplicar SVD para segundo orden, o Tucker o PARAFAC para órdenes superiores, durante los procedimientos RBL, RTL y RQL, para el interferente *n* y para el primer modo de medición. Definición análoga para $\mathbf{c}_{\text{int},n}$ y $\mathbf{d}_{\text{int},n}$ $\mathbf{e}_{\text{int},n}$ en el segundo, tercer y cuarto modo de medición (tamaños $K \times 1$, $L \times 1$ y $M \times 1$, respectivamente).
$\mathbf{B}_{\text{int}}$	$J \times A_{int}$	Matriz cuyas columnas están compuestas por los $\mathbf{b}_{\text{int},n}$ para cada uno de los *n* interferentes modelados durante RBL, RTL o RQL. Definiciones análogas para $\mathbf{C}_{\text{int}}$, $\mathbf{D}_{\text{int}}$, $\mathbf{E}_{\text{int}}$ (tamaños K x A_{int}, L x A_{int} y M x A_{int}, respectivamente).
c	$I \times 1$	Vector de concentraciones de calibrado en CLS.
$\hat{\mathbf{c}}_{test}$	$I_{test} \times 1$	Vector de concentraciones estimadas en CLS.
C	$I \times A$	Matriz de concentraciones de cada uno de los analitos que dan respuesta, presentes en la muestra.
e	Escalar	Error en la respuesta observada.
e	$I \times 1$	Vector de errores en concentración de las muestras predichas
$\hat{\mathbf{e}}$	$I \times 1$	Vector de errores estimados en la concentración de las muestras predichas.
$e_{jk:}$	Escalar	Residuos resultantes del modelado bilineal de los elementos x_{jk}
$e_{aug,pk}$	Escalar	Error resultante del modelado bilineal de un elemento

		genérico $x_{\text{aug},pk}$ correspondiente a una matriz aumentada.
E	$I \times K$	Matriz de errores en respuestas instrumentales, o de errores espectrales.
f_{obj}	Escalar	Función objetivo a minimizar en calibración univariada.
I	Escalar	Número de muestras de calibración.
I_{cal}	Escalar	Cantidad de muestras de calibrado cuando se trabaja con datos multi-vía (según el orden se referirá al número de matrices, cubos u otros arreglos de orden superior).
J	Escalar	Cantidad de sensores en el primer canal o modo de medición.
K	Escalar	Cantidad de sensores en el segundo canal o modo de medición.
L	Escalar	Cantidad de sensores en el tercer canal o modo de medición.
M	Escalar	Cantidad de sensores en el cuarto canal o modo de medición.
N	Escalar	Número de constituyentes que están generando respuesta cuando se trabaja con datos multi-vía.
N_{int}	Escalar	Número de inteferentes modelados durante el procedimiento RML
p	Escalar	Índice de variación de los sensores en la dirección aumentada (varía de 1 a IJ o IK dependiendo de la dirección de aumento).
P	$J \times A$	Matriz de *loadings* espectrales en PLS.
$\mathbf{r}_{test}$		Vector de absorbancias o respuestas instrumentales para las muestras de *test* en CLS
R	$I \times J$	Matriz de absorbancias o respuestas instrumentales (sólo en CLS)
s_p	Escalar	Desvío estándar residual resultante de la predicción por U-PLS.
s_u	Escalar	Desvío estándar residual luego de aplicar RML.
s	$J \times 1$	Sensibilidades o absortividades molares normalizadas
S	$I \times J$	Matriz de espectros puros a concentraciones unitarias o de sensibilidades para cada sensor y para cada uno de los componentes presentes
$\hat{\mathbf{S}}$	$I \times J$	Matriz S estimada
t	$1 \times A$	Vector de *scores* para una muestra particular.
$\mathbf{t}_a$	$I \times 1$	Vector de *scores* para un componente particular a.
T	$I \times A$	Matriz de *scores* resultantes de PCA/PLS.
v_a	Escalar	Elemento del vector **v** para un componente particular.
λ_i	Escalar	i-ésimo autovalor resultante de la descomposición por PCA.
v	$A \times 1$	Vector de coeficientes de regresión en el espacio de las

		variables latentes.
$\mathbf{v}_a$	$J\times1$	Vector de *loadings* para un componente particular. Columnas de la matriz **V**.
V	$J\times A$	Matriz de *loadings* obtenida en la descomposición por PCA o por PLS.
$\mathbf{w}_a$	$J\times1$	Vector de pesos de *loadings* para un componente particular. Columnas de la matriz **W**.
W	$J\times A$	Matriz de pesos de los *loadings* en PLS.
x_{jk}	Escalar	Elemento genérico de una matriz de datos (arreglo de dos vías).
$x_{\text{aug},pk}$	Escalar	Elemento genérico de uma matriz aumentada.
x_{ijk}	Escalar	Elemento genérico de un cubo de datos (arreglo de tres vías)
$\hat{x}_{test}$	Escalar	Concentración predicha por modelo de calibración univariada clásica.
x	I × 1	Vector de concentraciones de calibrado para cada muestra en calibración univariada
$\mathbf{x}_{test}$	$J\times1$	Vector de respuestas instrumentales para una muestra de *test* particular.
X	$I\times J$	Matriz de respuestas instrumentales de calibrado en ILS.
$\mathbf{X}_{\text{aug}}$	$IJ\times K$	Matriz aumentada
$\mathbf{X}^{\text{o}}$	$I\times J$	Matriz de verdaderos valores de respuestas instrumentales.
$\mathbf{X}_{test}$	$I\times J$	Matriz de respuestas instrumentales de *test*.
$\underline{\mathbf{X}}$	$I\times J\times K$	Arreglo de tres vías o superior
y_i	Escalar	Señal medida para la *i*-ésima muestra de calibrado.
$\hat{y}_i$	Escalar	Señal estimada estimada para la *i*-ésima muestra de calibrado en calibración univariada.
$\hat{y}_{test}$	Escalar	Concentración predicha por modelos inversos para una muestra de *test* particular.
y	$I\times1$	Vector de señales (respuestas) correspondientes a las muestras de calibración. Vector de concentraciones correspondientes a las muestras de calibración (calibración multivariada inversa).
$\hat{\mathbf{y}}$	$I\times1$	Vector de señales (respuestas) estimadas correspondiente a las muestras de calibrado (calibración univariada). Vector de concentraciones estimadas (calibración multivariada inversa).
Y	$I\times A$	Matriz de concentraciones en ILS.
Capítulo 2		

a	Escalar	Número de hormiga.
bk	$J\times1$	Vector de *background* adicionado a la señal del espectro simulado.
F_i	Escalar	Razón *F* para detección de *outliers*, calculada para la *i*-ésima muestra.
I	Escalar	Cantidad de muestras de calibrado.
J	Escalar	Cantidad de bloques de sensores generados/ Cantidad total de sensores.
p_j	Escalar	Cantidad de feromona asociada a la *j*-ésima variable.
p	$J\times1$	Vector feromona.
$\Delta\mathbf{p}$	$J\times1$	Vector de cambio en la cantidad de feromona.
Δp_a	Escalar	Contribución al cambio en la cantidad de feromona dado por la hormiga *a*.
ρ	Escalar	Parámetro que indica la velocidad de convergencia a cero ("evaporación") de los elementos del vector feromona.
$prob_j$	Escalar	Probabilidad de ser seleccionada asociada a la *j*-ésima variable.
s	Escalar	Cantidad de variables seleccionadas en una determinada iteración.
$\mathbf{s}_1$,$\mathbf{s}_2$, $\mathbf{s}_3$	$J\times1$	Espectros simulados de cada uno de los tres componentes del sistema.
t	Escalar	Número de iteración actual de cálculo.
υ_k	Escalar	Índice que identifica las variables seleccionadas en el vector **v**
v	$s\times1$	Vector de variables seleccionadas.
x	$J\times1$	Espectro simulado para una determinada muestra.
y_i	Escalar	Concentración nominal de la muestra *i*.
$\hat{y}_{\mathrm{cal},i}$	Escalar	Valor de concentración de calibración estimado.
$\hat{y}_{\mathrm{mon},i}$	Escalar	Valor de concentración de monitoreo estimada.
Capítulo 3		
b	2×1 (univariada) $J\times1$ (multivariada)	Vector con parámetros ajustables en calibración univariada (tamaño 2×1 e incluye pendiente (b_1) y ordenada al origen (b_o)) y en calibración multivariada (tamaño $J\times1$ e incluye los coeficientes de regresión en el espacio de las variables latentes). En este último caso equivaldría a mencionar de manera genérica los vectores $\hat{\mathbf{b}}_{\mathrm{PLS}}$ y $\hat{\mathbf{b}}_{\mathrm{PCR}}$ definidos durante el Capítulo 1.
j	$n\times1$	Vector con las derivadas parciales de una variable dependiente respecto a cada una de las variables independientes.

σ_y	Escalar	Variancia del error en señal (calibración univariada).
$\sigma^2_{\hat{y}}$	Escalar	Variancia del error en la concentración estimada (calibración multivariada).
$\sigma^2_{y_o}$	Escalar	Variancia entorno al punto de señal y_o.
$\sigma^2_{x_o}$	Escalar	Variancia entorno al punto de concentración x_o.
$\mathbf{\Sigma_b}$	$P \times P$	Matriz de covariancia del error de los parámetros estimados.
$\mathbf{\Sigma}_y^{-1}$	$I \times I$	Inversa de la matriz de covariancia del error de las concentraciones
$\mathbf{\Sigma}_x$	$J \times J$	Matriz de covariancia del error para la señal de una muestra desconocida.
$\mathbf{\Sigma}_X$	$J \times J$	Matriz de covariancia del error para las señales de calibración.
t	Escalar	Valores de coeficientes t de Student para un conjunto de muestras de predicción.
$\mathbf{x}_o$	1×2	Vector que contiene la concentración correspondiente al punto de la recta en el cual se está calculando la incertidumbre y un elemento igual a 1 que modela la ordenada al origen.
$\mathbf{X}$	$I \times P$ (univariada) $I \times J$ (multivariada)	Matriz de concentraciones de I muestras y P parámetros ajustables (en calibración univariada P = 2) en calibración univariada. En calibración multivariada corresponde a la matriz con las señales de calibrado.
x_o		Valor de concentración en torno al cual se calcula la incertidumbre en regresión univariada.
y_o	Escalar	Valor de respuesta en torno al cual se calcula la incertidumbre en la curva de regresión univariada.
$\mathbf{y}_{cal}$	$I \times 1$	Vector con valores nominales o de referencia de las concentraciones de calibrado.
$\hat{y}$	Escalar	Concentración estimada para una muestra genérica.
Capítulo 4		
f_a	Escalar	a-ésimo elemento diagonal de una matriz cuadrada de a x a que se calcula como $(\mathbf{T}^T \mathbf{T})^{-1/2}$.
h_{0min}	Escalar	Leva mínima obtenida a partir de la proyección de los *scores* de calibrado al plano de concentración 0.
h_{0max}	Escalar	Leva máxima obtenida a partir de la proyección de los scores de calibrado al plano de concentración 0.
$t_{\alpha,v}$	Escalar	Valor de probabilidad asociado a una distribución t de Student con v grados de libertad y probabilidad asociada α (probabilidad de falsos positivos).

$t_{\beta,v}$	Escalar	Valor de probabilidad asociado a una distribución *t* de Student con *v* grados de libertad y probabilidad asociada β (probabilidad de falsos negativos).
$t_{a\mathrm{N}}$	Escalar	Elemento específico del vector de scores normalizados $\mathbf{t}_{\mathrm{N}}$.
$t_{a\mathrm{Ncal}}$	Escalar	Elemento específico del vector $\mathbf{t}_{\mathrm{Ncal}}$.
$t_{a\mathrm{N0cal}}$	Escalar	Elemento específico del vector $\mathbf{t}_{\mathrm{N0cal}}$.
$\mathbf{t}_{\mathrm{N}}$	$A\times1$	Vector de *scores* normalizados para una muestra genérica.
$\mathbf{t}_{\mathrm{Ncal}}$	$A\times1$	Vector de scores normalizados para una muestra de calibración.
$\mathbf{t}_{\mathrm{N0cal}}$	$A\times1$	Vector de scores para la proyección de una muestra de calibración perpendicular al plano π_0 definido por la concentración 0 del analito.
Capítulo 5		
B	$J\times A$	Matriz que contiene los perfiles de todos los componentes para el primer modo instrumental.
$\mathbf{B}_{\mathrm{cal}}$	$J\times N_{\mathrm{cal}}$	Perfiles en el primer modo instrumental de los componentes calibrados obtenidos por PARAFAC. Definiciones análogas para $\mathbf{C}_{\mathrm{cal}}$ $\mathbf{D}_{\mathrm{cal}}$ $\mathbf{E}_{\mathrm{cal}}$ para el segundo, tercer y cuarto modo instrumental (tamaños $K\times N_{\mathrm{cal}}$, $L\times N_{\mathrm{cal}}$, $M\times N_{\mathrm{cal}}$, respectivamente).
$\mathbf{B}_{\mathrm{int}}$	$J\times N_{\mathrm{int}}$	Perfiles en el primer modo instrumental de los componentes inesperados (no incluidos en la calibración) Definiciones análogas para $\mathbf{C}_{\mathrm{int}}$ $\mathbf{D}_{\mathrm{int}}$ $\mathbf{E}_{\mathrm{int}}$ para el segundo, tercer y cuarto modo instrumental (tamaños $K\times N_{\mathrm{int}}$, $L\times N_{\mathrm{int}}$, $M\times N_{\mathrm{int}}$, respectivamente).
C	$K\times A$	Matriz que contiene los perfiles de todos los componentes para el segundo modo instrumental
$\mathbf{E}_{\mathrm{NAS}}$	$J\times K$	Residuo correspondiente al modelado de la matriz de señal neta.
J	$JK\times[J+K)N_{unx}+N_{cal}]$ (PARAFAC) $JK\times[(J+K)N_{\mathrm{int}}+A)]$ (U-PLS/RML)	Matriz Jacobiana asociada a la propagación de errores en PARAFAC y en PLS/RML
N_{cal}	Escalar	Número de componentes de calibrado
P	$JK\times A$	Matriz de *loadings* de calibrado en PLS/RBL
$\mathbf{P}_{\mathrm{B}}$	$J\times J$	Matriz de proyección ortogonal al espacio de los

		interferentes en el primer modo instrumental. Definiciones análogas para $\mathbf{P_C}$, $\mathbf{P_D}$,$\mathbf{P_E}$.
$\mathbf{P}_{\text{eff,NAS}}$	$JK \times A$	Matriz de *loadings* efectiva en desarrollo a partir de NAS
$\mathbf{P}_{\text{eff,EP}}$	$JK \times A$	Matriz de *loadings* efectivos en desarrollo a partir de propagación de errors.
$\mathbf{P}_{\text{Zint}}$	$JK \times JK$	Matriz que describe la proyección ortogonal al espacio definidco por $\mathbf{Z}_{\text{int}}$
$\mathbf{S}$	$K \times N$	Matriz con perfiles para todos los componentes en la dirección no aumentada de MCR
$\mathbf{\Sigma}$	$(J+K+A) \times (J+K+A)$	Matriz de variancia covariancia del conjunto de parámetros ajustables en U-PLS/RML.
$\mathbf{\Sigma}_t$	$A \times A$	Matriz de variancia covariancia de los *scores*
$\mathbf{X}$	$J \times K$	Matriz de datos para una muestra en particular, sin desdoblar.
$\mathbf{X}_{\text{NAS}}$	$J \times K$	Matriz de datos sin desdoblar luego de multiplicarla por las matrices de proyección ortogonal
$\mathbf{Z}_{\text{cal}}$	$JK \times (J+K)N_{cal}$	Bloque correspondiente a los perfiles de los analitos calibrados
$\mathbf{Z}_{\text{int}}$	$JK \times (J+K)N_{int}$	Submatriz de bloque dependiente de las propiedades de los interferentes

SOFTWARE UTILIZADO

Todos los cálculos computacionales realizados durante esta tesis doctoral se hicieron utilizando el software MATLAB 7.10[1] y las correspondientes rutinas se encuentran disponibles por solicitud a los autores.

INTRODUCCIÓN GENERAL

"-¿Por dónde empiezo majestad?

-Empieza por el principio -contestó gravemente el rey-. Y sigue hasta que llegues al final. Entonces te detienes." (Alicia en el País de las Maravillas).

CIFRAS DE MÉRITO: DE LA CALIBRACIÓN UNIVARIADA A LA MULTIVARIADA

Desde su concepción como ciencia de los instrumentos y las mediciones químicas, la química analítica estuvo predestinada a beneficiarse significativamente de la quimiometría, que se ocupa de extraer información de las mediciones químicas utilizando métodos matemáticos, estadísticos y computacionales. Los beneficios de la quimiometría se hacen realmente importantes cuando la información no se encuentra accesible de una manera directa, es decir, cuando está "oculta" por fenómenos de interferencia, tales como especies químicas con señales superpuestas, o el propio ruido instrumental. Mientras que los métodos analíticos univariados requieren de una selectividad total para su correcto funcionamiento, los métodos quimiométricos multivariados son mucho más flexibles, permitiendo obtener buenos resultados analíticos, incluso en condiciones de selectividad parcial. Como resultado de lo anterior, los procedimientos de preparación de muestras se hacen más simples, permitiendo ahorrar tiempo y disminuir costos.

El objetivo fundamental de la química analítica es el desarrollo de sistemas y metodologías que permitan lograr mediciones químicas de calidad. Por lo tanto, la cuantificación criteriosa de esta calidad constituye uno de los cimientos de la disciplina. Es por esto que entre las actividades relevantes en esta área de la química, se encuentra el establecimiento de un conjunto de números llamados "cifras de mérito". Estas últimas permiten calificar los distintos métodos, y comparar unos con otros en cuanto a su eficiencia para lograr el objetivo final de la disciplina. Si bien al término "mérito" se le adjudican habitualmente connotaciones positivas, la etimología latina indica que proviene de "*meritum*", traducido como "acción que hace a una persona digna de galardón o de sanción". Las cifras analíticas de mérito actúan en idéntico sentido, auspiciando o desfavoreciendo una metodología según su eficiencia o calidad.

En términos de estas cifras de mérito, medir y procesar datos multivariados y multi-vía permite a los químicos analíticos acceder a una serie de ventajas como son: (1) mayor sensibilidad, debido a que se promedia el ruido proveniente de una gran cantidad de mediciones, (2) mayor selectividad, ya que cada nuevo sensor y dimensión medidos proveen un grado adicional de selectividad parcial, y (3) sólo en el caso de datos multi- vía, modelado de la contribución del analito y su determinación cuantitativa en presencia de interferentes desconocidos, ausentes en las muestras de calibrado (cualidad conocida como "ventaja de segundo orden").[2]

Teniendo en cuenta los ítems 1 y 2, una pregunta que emerge inmediatamente es: ¿cómo deberían estimarse cifras de mérito como la sensibilidad, la selectividad e incluso el límite de detección, cuando se trabaja con datos multivariados y multi-vía? Como se resaltó anteriormente, la búsqueda de nuevos estimadores que permitan mejorar las cifras de mérito analíticas constituye una importante guía en la evolución de las investigaciones en química analítica moderna, siendo la sensibilidad (SEN) y el límite de detección (LOD) las dos cifras más conocidas y utilizadas. En efecto, encontrar una expresión para calcular estimadores consistentes para estas cifras tiene una influencia relevante en diferentes actividades como: (1) comparación del desempeño de distintos procedimientos experimentales, (2) optimización de una determinada metodología bajo diversas condiciones experimentales, y (3) desarrollo de protocolos oficiales de análisis y validación, tal y como se documentan en los estándares internacionales.[3,4]

Particularmente, el límite de detección se popularizó como un descriptor de la calidad de un determinado método cuando se trabaja en química analítica aplicada. Esto se debe fundamentalmente a dos motivos: (1) se puede expresar en unidades de concentración, permitiendo una comparación directa entre diferentes métodos y (2) es necesario para evaluar capacidades de detección que son de importancia fundamental en ciertas áreas específicas como el control de *doping* en deportes, el monitoreo de trazas de contaminantes en muestras ambientales, etc. Sin embargo, en el núcleo de la definición del LOD se encuentra la sensibilidad, que es también un parámetro importante en la estimación de otras cifras de mérito, tales como: (1) sensibilidad analítica, importante en la comparación de metodologías basadas en señales muy distintas, ya que es independiente del instrumento y la técnica aplicada, (2) selectividad, que permite hacer inferencias acerca de la posibilidad de cuantificar analitos en presencia de interferentes, e (3) incertidumbre en la predicción, que da una idea acerca de cuán precisa es una determinada predicción.

Por los motivos mencionados, encontrar estimadores confiables de las cifras de mérito no es un tema menor, y esta es la razón por la cual este trabajo de tesis tiene como uno de sus ejes fundamentales lograr un panorama más claro en lo que respecta a la estimación de cifras de mérito en calibración multivariada y multidimensional.

Sensibilidad

De acuerdo con la Unión Internacional de Química Pura y Aplicada (IUPAC), en calibración clásica con un único componente o calibración univariada, la sensibilidad se define como "el cambio en la respuesta instrumental dividida por el correspondiente *estímulo* (la concentración del analito de interés)", es decir, la pendiente de la recta de calibrado.[5] En calibración multivariada de primer orden, la situación en lo que respecta a la definición de sensibilidad se vuelve más compleja.[6] Por ejemplo, cuando el solapamiento espectral es muy severo, una señal intensa pierde su utilidad debido a las señales provenientes de otros constituyentes. Lo anterior conduce a otra propiedad importante de la sensibilidad multivariada, dada por su especificidad para cada analito.

Como se explicará con mayor detalle al final del Capítulo 1, los modelos de calibración multi-vía parten de la medición de matrices de datos por cada muestra (e incluso arreglos con 3 o más modos o vías) con el objetivo de confeccionar un modelo que permita medir la concentración de un analito en particular, y constituyen una generalización importante de la calibración multivariada.[7] En este campo, se han propuesto varias expresiones para calcular la sensibilidad, algunas de ellas basadas en la extensión del concepto de señal analítica neta (NAS) cuando se trabaja con varias vías.[8-11] La señal analítica neta es la proporción de la señal total que puede ser adjudicada específicamente al analito de interés. Por lo tanto, la pendiente del gráfico de calibración pseudo-univariada de la NAS en función de las concentraciones, o lo que es lo mismo, la NAS a concentración unitaria, surge como una definición razonable de sensibilidad. De cualquier manera, este tipo de "extensión intuitiva" generó ciertas dificultades, debido al surgimiento de varias definiciones diferentes de NAS, sin una clara relación entre ellas.[12-14] Es más notorio aún que la extrapolación a la calibración de orden superior llevó a una importante subestimación de las sensibilidades.

En el Capítulo 5 se presentará una alternativa a la aproximación NAS, basada en el análisis de cómo la incertidumbre en señales instrumentales se propaga a la incertidumbre en concentraciones predichas.[15-16] Como se explicará más en detalle posteriormente, gracias a los desarrollos en este sentido, hoy en día es posible reunir todas las expresiones existentes para calcular la sensibilidad en una única ecuación matemática general que incluye todos los grados posibles de complejidad de los datos, desde el caso univariado hasta multi-vía, y en el último caso incluyendo a los algoritmos más utilizados.[17]

Esta expresión matemática está en concordancia con el hecho que la sensibilidad multi-vía presenta propiedades incluso más interesantes en comparación con las contrapartes univariada y de primer orden: no es únicamente específica de cada analito, sino también dependiente de la muestra de *test* y del algoritmo de procesamiento de datos empleado. Esto implica que cada muestra, con su composición química cualitativa y específica, lleva a un valor específico de sensibilidad. Igualmente, las herramientas computacionales empleadas por cada algoritmo afectarán la sensibilidad del sistema, y por lo tanto deberían ser consideradas como parte integrante del protocolo analítico multi-vía. La **Figura 1** resume las diferentes maneras en las que puede definirse la sensibilidad, normalmente interpretada como el cociente entre un valor de entrada y otro de salida, dependiendo del orden de los datos que se estén analizando.

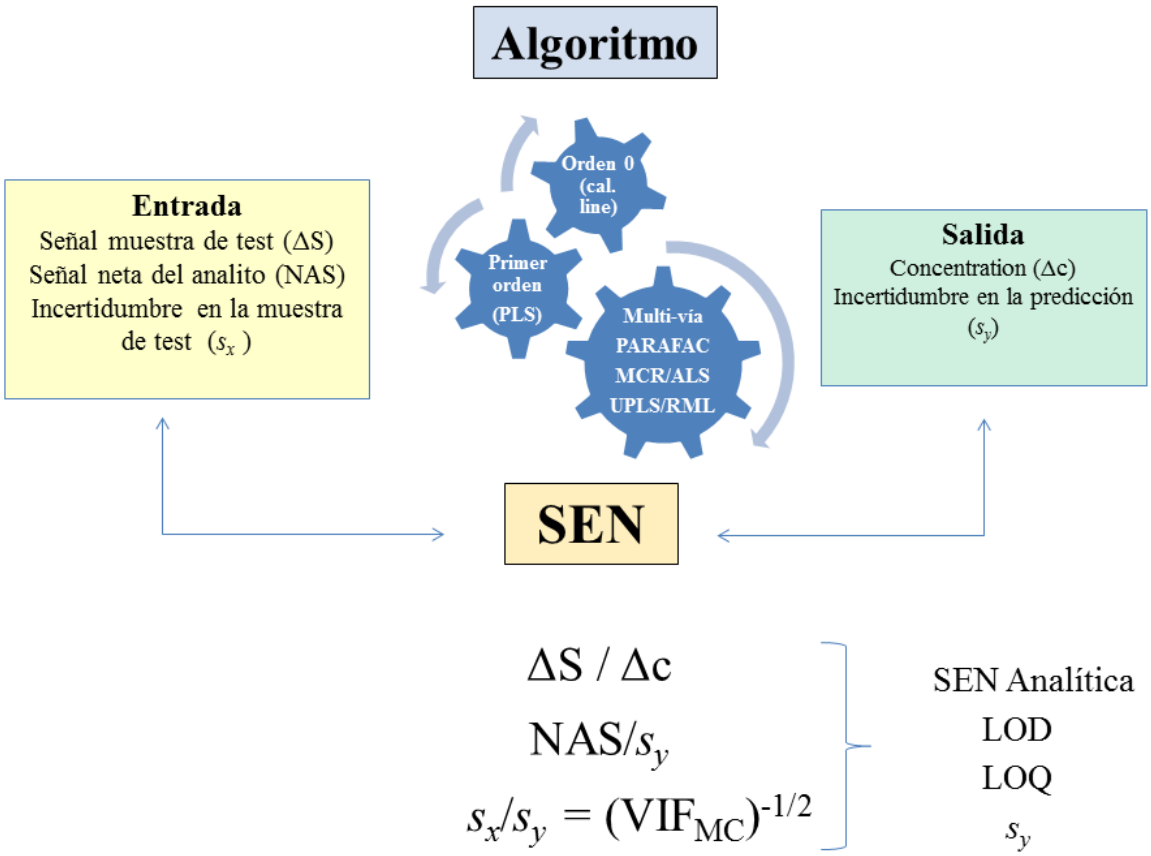


Figura 1. Representación esquemática de las distintas maneras posibles de definir la sensibilidad de acuerdo con el orden de los datos y el algoritmo utilizado.

Límite de detección

En lo que respecta al límite de detección, la transición entre la calibración univariada y la multivariada requiere especial atención, como en el caso de la sensibilidad. Los términos asociados con lo que en general se conoce como capacidad de detección han estado presentes en la literatura científica por más de 100 años. Durante este período de tiempo, se han presentado numerosos términos, definiciones y aproximaciones de cálculo, y un trabajo de revisión recientemente publicado resume adecuadamente la historia en torno a los estimadores de esta cifra de mérito.[18]

Actualmente, la IUPAC adopta la definición propuesta por la Organización Internacional de Estandarización (documento ISO 11843)[19] para el límite de detección, como "la mínima cantidad de sustancia que puede distinguirse de la ausencia de esa sustancia (valor del blanco) con un determinado límite de confianza"[20-22] Esto implica que

el LOD es la mínima cantidad detectable con una determinada probabilidad de falsos positivos (asumir que el analito está presente cuando en realidad no lo está, también conocida como α- o de errores de tipo I) y de falsos negativos (asumir que el analito está ausente cuando en realidad no lo está, también conocido como β- o probabilidad de errores de tipo II).[20-22] Cuando la señal analítica es univariada y la calibración es específica del analito, el estimador se encuentra muy bien definido, y el límite de detección se puede estimar directamente a partir de la recta de calibración. La regla de detección recomendada está basada en el *test* de Neyman-Pearson, que considera los falsos positivos y los falsos negativos para la hipótesis nula "no hay analito" y la alternativa "hay analito".[20]

Sin embargo, cuando se trabaja en calibración multivariada, como en el caso del bien conocido algoritmo de regresión por cuadrados mínimos parciales (PLS),[23-26]es necesario abordar algunos aspectos que quedan por fuera del campo de aplicación de las normas ISO.[27] De hecho, no existe aún un estimador aceptado de manera general en lo que respecta a estudios analíticos mediante PLS, por lo que el interés en la temática continúa vigente.[17] Esto se relaciona con la inclusión del algoritmo PLS en muchos instrumentos comerciales, particularmente en aquellos basados en mediciones por espectroscopía de infrarrojo cercano (NIRS)[28], así como con el surgimiento continuo de técnicas analíticas más novedosas y sensibles, y con el lanzamiento de regulaciones respecto a la exposición humana o del medioambiente a niveles bajos de potenciales riesgos químicos para la salud.

La principal dificultad para estimar el límite de detección multivariado es que las señales instrumentales no son específicas para un analito particular. Por esta razón, cuando se extrapola la concentración del analito de interés al valor cero, la composición de la matriz de cada muestra juega un rol fundamental. Como se desarrollará en el Capítulo 4, es posible proponer un estimador que adopta la forma de un intervalo de límite de detección, lo que constituye una alternativa en respuesta a la dificultad previamente mencionada.

Otras cifras de mérito

La sensibilidad constituye el núcleo a partir del cual se pueden definir otras cifras de mérito. Entre ellas, el límite de detección es una de las más conocidas y ampliamente utilizadas. Sin embargo, en algunos casos puede ser útil definir otras cifras para resaltar ciertas características y diferencias en los datos, las muestras o la metodología que se está analizando. Ejemplos de estas cifras de mérito son la sensibilidad analítica, la

incertidumbre en la predicción y el límite de cuantificación. Entre éstas, la incertidumbre, dada por el desvío estándar en la predicción de una determinada muestra, es de fundamental importancia. Un estimador confiable para esta cifra no sólo es la base para calcular el límite de detección (que en última instancia es una función del desvío estándar del blanco) sino que también permite comprender cómo el error se propaga e influye en un determinado modelo de calibración. Teniendo en cuenta lo anterior, en el Capítulo 2 se abordará una descripción detallada de la aplicación de la propagación de errores para calcular un desvío estándar de predicción dependiente de cada muestra cuando se utilizan PLS y la regresión por componentes principales (PCR).[23-26]

Del ruido homoscedástico (iid) al heteroscedástico y correlacionado (no iid)

Mientras que la mayoría de las herramientas quimiométricas se diseñaron para extraer información de mediciones químicas multivariadas, un punto importante que hasta el momento no se consideró en detalle, es el rol de los errores de medición multivariados en este proceso. Sin embargo, recientemente se popularizó una nueva rama de la quimiometría aplicada a la química analítica, relativa al desarrollo de metodologías para comprender y caracterizar la estructura de error de un determinado conjunto de datos.[29-31]

La mayoría de los modelos y algoritmos para tratar datos multivariados y multi-vía se diseñaron bajo la suposición simplificada de que el error que afecta al sistema en estudio se encuentra distribuido de manera idéntica e independiente (error iid). En muchos casos, especialmente cuando estos errores no son importantes o cuando los supuestos anteriores son aproximadamente válidos, las herramientas quimiométricas tradicionales pueden aplicarse con buenos resultados. Sin embargo, existen casos en que la consideración de la estructura del error puede significar la diferencia entre el éxito o la deficiencia de un determinado análisis.[32,33] Esta misma discusión se podría extender a la estimación de cifras de mérito. Aunque hasta el momento las expresiones para el cálculo de la sensibilidad obtenidas por propagación de errores se definen suponiendo que el ruido es iid, cuando la incertidumbre en la predicción se estima en un contexto de errores correlacionados y/o heteroscedásticos, el escenario cambia y es necesario realizar nuevas consideraciones, tal como se presentará en el Capítulo 3.

DESARROLLO DE NUEVOS ALGORITMOS

Otra línea de investigación importante en lo que respecta al análisis de datos multivariados en la química analítica moderna involucra la optimización de los modelos predictivos que hasta el momento se han establecido y popularizado en la literatura. Entre estos, la regresión por PLS, que se explicará detalladamente durante el Capítulo 1, es la elección más frecuente cuando se trabaja con datos de primer orden (es decir, cuando para cada muestra se obtiene un vector de datos, o tensor de orden 1, como puede ser un espectro). Dado un espectro para una muestra en particular, es probable que algunas de las señales no sean selectivas en lo que respecta al analito que se desea cuantificar, mientras que otras serán parcialmente selectivas. Es por esto que es común someter a las variables predictoras a un procedimiento de selección cuidadoso antes de incorporarlas a la regresión PLS. Esto significa que el modelo multivariado será construido solamente con un determinado número de señales. El propósito de la selección de variables es obtener modelos basados en datos espectrales conteniendo la mayor cantidad de información posible en lo que respecta al analito o propiedad de interés. Esto conlleva a modelos PLS más robustos y exactos, justificando el creciente interés en esta temática, especialmente en espectroscopía de infrarrojo cercano (NIRS).

Utilidad y características generales de los métodos de selección de variables

Como se explicará en mayor detalle durante el Capítulo 1, en el marco de la regresión lineal multivariada (MLR), también denominada cuadrados mínimos inversos (ILS),[25,34,35] suele presentarse un problema conocido como "deficiencia de rango" de la matriz de datos. Esta limitación algebraica surge cuando el número de objetos (como las muestras de calibrado) es menor que el número de variables predictoras (como los sensores o canales de medición), o cuando la colinealidad entre las variables es muy marcada. En estos casos, la reducción del número de variables predictoras se convierte en una condición necesaria para poder aplicar el modelo citado.

Otra alternativa es la de utilizar métodos basados en variables latentes como PLS o PCR, que permiten reducir el número de variables originales, encontrando las direcciones del espacio muestral en las cuales la variancia espectral (PCR) o la covariancia espectros-concentraciones (PLS) es máxima. Sin embargo, incluso en estos casos, la introducción de

un paso previo de selección de variables puede aportar ciertos beneficios.[36] De hecho, puede llevar a una mejora en las predicciones del modelo, permitir una mejor interpretación de los resultados y, en algunas situaciones, ayudar a reducir los costos de medición.

Partiendo de las consideraciones anteriores, no es sorprendente que en los últimos años se hayan propuesto distintas estrategias para elegir un número reducido de variables para construir el modelo de regresión. A la hora de decidir qué metodología utilizar, es importante tener en cuenta la naturaleza del problema que se está abordando, así como las características de los datos que se analizan. Sin embargo, independientemente de la estrategia empleada, hay algunas cuestiones de índole general que deberían tenerse en cuenta.

En particular, una potencial desventaja a considerar es que la selección de variables podría resultar en modelos con una gran tendencia al sobreajuste. Por lo tanto, siempre que se adopte una estrategia de este tipo, es especialmente importante llevar adelante una validación adecuada del grupo de predictores elegidos. Otra característica que es común a todas las estrategias de selección de variables es que son sensibles a los llamados *outliers* o mediciones atípicas: debido a que estos influyen de manera significativa en la confección del modelo, los *outliers* podrían inducir a la selección de subconjuntos de variables que no son óptimas para caracterizar el grueso de los datos.

Como parte de una guía general a la hora de seleccionar variables, también se debe tener en cuenta que la elección del método debería realizarse considerando el tipo de señal con el cual se esté trabajando: por ejemplo, si el procedimiento se aplica a una señal homogénea, como un espectro o un cromatograma, donde las variables contiguas contienen información similar y correlacionada, sería razonable apuntar a una metodología que permita seleccionar intervalos o regiones en lugar de variables independientes. En este sentido, es importante tener precaución a la hora de aplicar métodos de preprocesamiento de señales, ya que algunos pueden producir distintos resultados según se apliquen antes o después del paso de selección de variables.

Estrategias de selección de variables

Utilización de los parámetros del modelo

Siempre que sea posible construir un modelo y validarlo, la inspección de los parámetros del modelo puede ayudar a identificar variables potencialmente relevantes o irrelevantes. De hecho, si el modelo funciona de manera razonablemente correcta, la evaluación de los coeficientes de regresión y, en el caso de los modelos bilineales los *loadings*, pueden proveer información importante respecto de las variables a seleccionar para obtener un buen modelo. Por otro lado, las variables asociadas a *loadings* o coeficientes de regresión cercanos a cero, o regiones de señal que se corresponden con parámetros ruidosos, donde se esperaría muy poca variación, muy probablemente no sean significativas.

Sin embargo, es preciso ser muy cuidadoso a la hora de realizar este tipo de análisis, ya que los coeficientes de regresión y los *loadings* provenientes de métodos basados en variables latentes surgen a partir de combinaciones lineales, cuya relación con las variables originales en muchas casos puede no ser tan directa, haciendo el análisis más complejo de lo que resulta a simple vista.[37] De aquí el surgimiento de metodologías de selección basadas en índices construidos a partir de estos parámetros.

Importancia de las variables en función de índices construidos a partir del modelo

Los parámetros obtenidos a partir del modelo también pueden utilizarse para calcular índices que reflejan la importancia relativa de los predictores a la hora de definir el modelo. En particular, existen dos índices que se utilizan con frecuencia cuando se trabaja con calibraciones basadas en variables latentes: (1) la importancia de las variables en la proyección (VIP)[38] y (2) la razón de selectividad (SR)[39]. VIP es una medida de la proporción en que una determinada variable individual contribuye a la definición del espacio **X** (señales) e **Y** (concentraciones) durante el modelado por PLS. Por otro lado, SR es la razón entre la porción de variancia total explicada por el modelo bilineal y su correspondiente variancia residual. Al igual que para los valores de VIP, mientras mayor sea el valor de SR, más relevante se considerará el correspondiente predictor.

iPLS

Cuando las variables dependientes son homogéneas y se definen como función de un conjunto de variables continuas (como un espectro o un cromatograma), debido a la correlación que existe entre ellas, es más conveniente seleccionar grupos de variables en lugar de una única por vez. Una manera de hacerlo en el contexto de los métodos basados en variables latentes es utilizar un análisis por intervalos.[40]

En general, cuando se utilizan métodos quimiométricos basados en intervalos, se calcula un modelo para cada uno de estos intervalos y el resultado individual se compara con el del modelo global. En el caso particular de PLS, se elabora un modelo con las variables pertenecientes a cada intervalo y de acuerdo con el resultado de la predicción, comparado con el resultado para el modelo con todas las variables, se establece un valor de peso para ese intervalo particular.

Algoritmos estocásticos

Desde un punto de vista matemático, la selección de variables puede interpretarse como un problema de optimización, que puede formularse como "encontrar el subconjunto de predictores que llevan al mejor resultado en términos de algún criterio de ajuste definido por el usuario." Teniendo en cuenta este principio, en los últimos años surgió un conjunto de algoritmos inspirados en el funcionamiento de la naturaleza y que operan siguiendo un mecanismo estocástico. Este término implica que operan siguiendo pasos de aleatorización, evaluación y posterior cálculo de probabilidades, hasta llegar a un punto de convergencia que representa la solución al problema.

Como se verá durante el Capítulo 2, el ejemplo típico y referente de este tipo de metodologías son los algoritmos genéticos (GA), que llevan adelante una optimización numérica basada en la adaptación en términos matemáticos del concepto de "supervivencia del más apto."[41-43] Sin embargo, recientemente se introdujo en química analítica otro tipo de algoritmo relacionado, basado en la filosofía de optimización por colonias de hormigas (ACO)[44,45] Este algoritmo imita el comportamiento de los insectos en la búsqueda de nuevas fuentes de alimento, y se ha demostrado ser igual o más eficiente que los algoritmos genéticos a la hora de elegir variables espectrales.[46]

Optimización integrada de PLS

Teniendo en cuenta lo anterior, otro de los ejes fundamentales de este trabajo de tesis será el de confeccionar y poner a prueba un nuevo algoritmo de selección de variables basado en la filosofía de la computación natural, que busca diseñar algoritmos inspirados en el funcionamiento de la naturaleza. Como se mostrará durante el Capítulo 2, es posible combinar la selección de variables con estrategias de selección de muestras y preprocesamientos matemáticos, intentando arribar a un procedimiento de optimización integrado que tenga en cuenta los principales factores que influyen a la hora de desarrollar una calibración robusta.

OBJETIVOS GENERALES

El título de la tesis junto con los temas desarrollados durante esta introducción general anticipan los objetivos fundamentales que se pueden resumir como:

1) Estudio y desarrollo de algoritmos de optimización de modelos de calibración multivariada.

2) Estudio y validación de estimadores confiables de las cifras de mérito, que permitan darle un marco formal a la utilización de modelos multivariados y multi-vía en química analítica cuantitativa.

RESUMEN DEL CONTENIDO DE CAPÍTULOS DE LA TESIS

Durante esta introducción, se presentaron de manera general los distintos temas que serán abordados durante el trabajo de tesis. Sin embargo, no se hizo referencia a la manera en que se organizaron estos temas. En términos generales, cada capítulo conserva una cierta independencia respecto del resto, aunque el vínculo entre algunos es más marcado que otros.

El Capítulo 1 es básicamente una sección en la que se describen los modelos cuantitativos más utilizados en calibración analítica multivariada y multi-vía. La mayoría de estos modelos se utilizan en los capítulos subsiguientes, mientras que los que no lo hacen sirven como base para explicar el funcionamiento del resto.

El Capítulo 2 se enfoca en el modelo PLS y sus alternativas de optimización por selección de variables. Una vez realizado un pantallazo introductorio acerca de los distintos mecanismos de selección de variables, se presenta un algoritmo basado en una estrategia de optimización integrada que no sólo considera la selección de variables predictoras sino que también tiene en cuenta posibles preprocesamientos, detección de *outliers* y distintas formas de seleccionar muestras para subdividir el conjunto total.

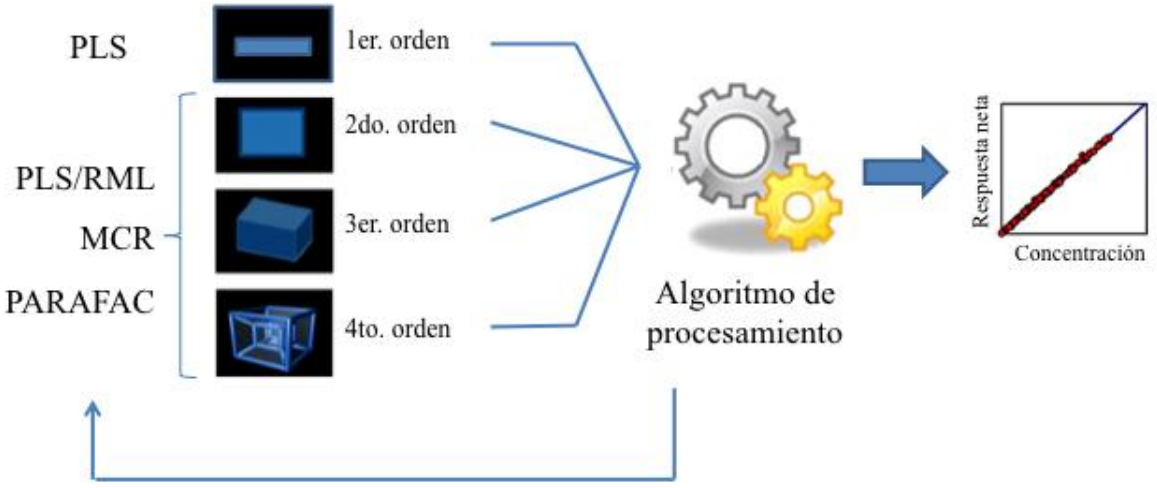
Los Capítulos 3 y 4 también están centrados en PLS, pero a diferencia del Capítulo 2, el objetivo es el estudio y desarrollo de estimadores de cifras de mérito que se utilizan con frecuencia a la hora de evaluar este modelo. En tal sentido, en el Capítulo 3 se estudia el cálculo de la incertidumbre en la predicción, teniendo en cuenta distintas estructuras y fuentes de error. Dado que la estimación en la incertidumbre de las muestras blanco es la base para estimar el límite de detección, el Capítulo 4 está estrechamente relacionado con el 3, y se enfoca en una nueva propuesta para estimar el LOD en el contexto de la calibración multivariada basada en la generación de variables latentes.

El Capítulo 5, al igual que los anteriores, se relaciona con la estimación de cifras de mérito. Sin embargo, a diferencia de éstos, se busca un estimador general de la sensibilidad cuando se trabaja empleando la versión multi-vía de PLS.

Finalmente, en el anexo se presenta una investigación resultante de una beca internacional de intercambio para estudiantes de doctorado, que si bien no forma parte del hilo conductor principal de esta tesis, se enfoca en una temática de gran impacto en la química analítica cualitativa actual, como es la microscopía de fluorescencia de superresolución. En el marco de esta investigación, se desarrolló una nueva metodología de procesamiento de imágenes resultante de la aplicación de técnicas quimiométricas de tratamiento de datos multidimensionales.

CAPÍTULO 1

REVISIÓN DE MODELOS DE CALIBRACIÓN MULTIVARIADA Y MULTI VÍA



"Mientras más lejos puedas mirar hacia atrás, más lejos podrás mirar hacia delante." (Sir Winston Churchill)

1.1 Resumen

Gran parte de este trabajo de tesis se estructura en torno al modelo de calibración multivariada PLS, así como en una de sus versiones multi-vía, U-PLS/RML.[47-50] Estos modelos se encuentran entre los más utilizados y difundidos en la actualidad. Sin embargo, describirlos de manera aislada sin antes detenerse en el funcionamiento de otras metodologías de calibración multivariada, no sería del todo adecuado en el marco de esta tesis, de fuerte sustento quimiométrico. Es por esto, que en las siguientes secciones se irán presentando paso a paso cada uno de los modelos que fueron surgiendo hasta llegar a PLS, partiendo de uno de los pilares fundamentales de la química analítica: la calibración univariada. En el caso de calibración de múltiples vías también se realizará una breve descripción de los otros dos modelos que constituyen la base a la hora de intentar analizar datos de esta naturaleza: análisis paralelo de factores (PARAFAC) [51] y resolución multivariada de curvas acoplada a cuadrados mínimos alternantes (MCR-ALS).[52]

Si bien este capítulo no se referirá a investigaciones o descubrimientos novedosos, constituye una sección clave de descripción de métodos que luego serán aplicados y analizados en las investigaciones que se desarrollarán en los capítulos siguientes.

1.2 Cuadrados mínimos y calibración univariada

Es muy probable que el primer contacto de la mayoría de los químicos con la forma más simple y conocida de calibración, es decir la calibración analítica univariada (o de "orden cero", como se verá más adelante), haya sido a través de la aplicación de la ley de Beer. En regresión univariada y en calibración, esta ley se puede representar de manera por medio de la siguiente notación vectorial:

$$\mathbf{y} = \mathbf{x}b + \mathbf{e} \qquad (1.1)$$

donde **x** es la variable independiente (concentraciones nominales en el escenario de la ley de Beer), **y** es la variable dependiente (lecturas instrumentales), *b*, el coeficiente de regresión (absortividad molar a una determinada longitud de onda y para un paso óptico constante), y **e** el error. Sin perder la generalidad, en lo que sigue de este capítulo se supondrá que la ordenada al origen se eliminó de alguna manera en caso que estuviera presente.

Generar un modelo de trabajo para la relación descripta en la **Ecuación 1.1** requiere estimar el parámetro *b*. Esta estimación se obtiene minimizando la suma de cuadrados de los residuos (SSR) de cada una de las *I* muestras de calibrado, representada a través de la función objetivo:

$$f_{\text{obj}} = \sum_{i=1}^{I} (y_i - x_i b)^2 = (\mathbf{y}-\mathbf{x}b)^{\mathrm{T}}(\mathbf{y}-\mathbf{x}b) \qquad (1.2)$$

que corresponde a resolver la siguiente ecuación lineal para obtener el valor estimado

$$\hat{b} = (\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}\mathbf{x}^{\mathrm{T}}\mathbf{y} \qquad (1.3)$$

El superíndice '+' indica la operación pseudoinversa, denominada de esta forma ya que la inversa de una matriz que no es cuadrada no existe. Cuando $\mathbf{x}^{+} = (\mathbf{x}^{\mathrm{T}}\mathbf{x})^{-1}\mathbf{x}^{\mathrm{T}}$, como se indicó más arriba, se denomina pseudoinversa de Moore-Penrose. Finalmente, en la etapa de predicción, los estimadores de $\hat{\mathbf{y}}$ y $\hat{\mathbf{e}}$ se obtienen del modo que sigue:

$$\hat{\mathbf{y}} = \mathbf{x}\hat{b} \quad (1.4)$$

$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} \quad (1.5)$$

Por otro lado, la estimación de la concentración de una muestra de *test* desconocida $\hat{x}_{test}$ a partir de una medición instrumental y_{test} puede realizarse como:

$$\hat{x}_{test} = \hat{b} y_{test} \quad (1.6)$$

El gráfico de puntos de señal en función de las concentraciones correspondientes, es la manera tradicional de representar los conceptos relacionados a calibración univariada. Esto es lo que se conoce como una representación en el "espacio de las variables". Sin embargo, varias interpretaciones subyacentes a la metodología de calibración univariada en general, que luego podrían ser útiles a la hora de interpretar esquemas multivariados, no son tan claramente visibles. [53]

Teniendo en cuenta lo anterior, una alternativa interesante es la de representar los datos en el "espacio de las muestras", que utiliza a las muestras como los ejes del gráfico, en lugar de las variables.[53] De esta manera, considerando el modelo definido en la **Ecuación 1.1**, el vector correspondiente a las variables **y**, es simplemente un múltiplo escalar del vector **x**, es decir, los dos vectores deberían estar orientados en la misma dirección en el espacio de las muestras. Sin embargo, dado que **y** se encuentra "corrompido" por el error en la medición, presentará una desviación respecto a la colinealidad absoluta con **x**. Lo anterior se ilustra claramente en la **Figura 1.1**.

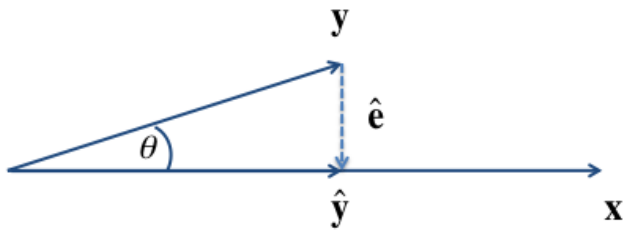


Figura 1.1. Representación vectorial en el espacio de las muestras del modelo de regresión univariada.

Dado que la solución de mínimos cuadrados para b ($\hat{b}$) lleva a un mínimo para la función objetivo de la **Ecuación 1.2**, debería también minimizar el tamaño del vector error estimado, $\hat{\mathbf{e}}$:

$$f_{obj} = \sum_{i=1}^{I} (y_i - \hat{y}_i)^2 = (\mathbf{y}-\hat{\mathbf{y}})^T(\mathbf{y}-\hat{\mathbf{y}}) \quad (1.7)$$

$$f_{obj} = \sum_{i=1}^{I} (\hat{e}_i)^2 = \hat{\mathbf{e}}^T\hat{\mathbf{e}} = \|\hat{\mathbf{e}}\|^2 \quad (1.8)$$

El símbolo $\| \ \|$ representa la norma Euclideana o el tamaño del vector. Por lo tanto, la solución de mínimos cuadrados en el espacio de las muestras es claramente la solución que lleva al menor vector **e** desde el extremo de **y** a un punto de **x**. Esto ocurrirá siempre que este vector sea ortogonal a **x**, haciendo evidente que $\hat{y}$ es la proyección ortogonal de **y** en **x**, y $\hat{b}$ el escalar que satisface la **Ecuación 1.3**. Lo anterior se hace mucho más evidente si se analiza la **Ecuación 1.4**, ya que $\mathbf{x}\mathbf{x}^+$ es una matriz de proyección ortogonal:

$$\hat{\mathbf{y}} = \mathbf{x}\mathbf{x}^+\mathbf{y} \quad (1.9)$$

Mientras que la simplicidad de la calibración univariada es un atributo muy atractivo, tiene algunas limitaciones importantes. Entre ellas, la fundamental es que los métodos univariados requieren de una selectividad absoluta para el analito de interés. En principio, las interferencias sólo podrían modelarse en caso que la cantidad de interferente fuera constante tanto en muestras de calibrado como en muestras de predicción. Esta severa limitante, impide desde el punto de vista matemático, realizar calibraciones en presencia de interferentes así como un análisis simultáneo de múltiples componentes. Es por esto, que el movimiento desde el mundo de la calibración univariada a la multivariada, se justifica por la aparición de numerosas ventajas que resultan muy importantes desde el punto de vista de la química analítica cuantitativa.

1.3 Modelos clásicos vs. Modelos inversos

En la sección anterior, el supuesto básico fue que el vector de mediciones instrumentales, **y**, se proyecta en el espacio de las concentraciones, **x**, para poder obtener la

solución de mínimos cuadrados para el coeficiente de regresión. Esta proyección se realiza de esta manera ya que se supone que los errores en las medidas de absorbancia, **y**, son sustancialmente mayores que los errores en concentraciones, **x**. Es posible invertir la representación clásica transformando las medidas de absorbancia en **x**, y las concentraciones en **y**, lo cual implicaría que los errores en valores de concentración serían significativamente mayores que los errores en respuesta. En la literatura quimiométrica, esta forma de modelar los datos se conoce como calibración inversa. En contraste con la manera clásica de calibrar presentada en la sección anterior, la calibración univariada inversa estaría utilizando mínimos cuadrados para proyectar el vector de concentraciones (ahora denominado **y**) en el vector de las respuestas (ahora **x**). Más allá de la factibilidad de esta posibilidad, en calibración univariada sólo existen diferencias menores entre la calibración clásica y la inversa. Sin embargo, en calibración multivariada, esta distinción pasa a ser mucho más importante tanto desde el punto de vista práctico como teórico.

1.4 Modelos multivariados de primer orden

En calibración multivariada de primer orden, se mide un vector (tensor de primer orden) con una cantidad *J* de sensores o elementos predictores, para cada una de las *I* muestras. Además, este tipo de calibración, incluye la situación en la que para cada muestra se realizan medidas de concentración de *A* especies individuales contenidas en esa muestra.

Como ya se anticipó, la utilización de métodos de calibración de primer orden en química analítica se justifica por una serie de ventajas respecto a los de orden cero (univariados) tales como: (1) determinar de manera simultánea múltiples componentes de una mezcla, (2) determinar la concentración de un analito en presencia de especies interferentes, en caso de ser incluidas en la calibración, (3) detectar la presencia de interferentes en caso que no sean contemplados durante la calibración, aunque no pueda corregirse (lo que se conoce normalmente como “ventaja de primer orden”) y (4) mejorar la precisión en las determinaciones debido al uso de múltiples respuestas.

En las secciones siguientes se presentarán en orden de aparición, complejidad y eficiencia, los métodos de calibración analítica multivariada más utilizados en la actualidad.

1.4.1 Regresión por cuadrados mínimos clásicos (CLS)

Así como la **Ecuación 1.1** es una extensión de la ley de Beer cuando se incluyen muchas muestras, también es posible expresar esta ley para cuando se miden múltiples longitudes de onda. Para un sistema de un único componente con *I* muestras, esta relación se convierte en

$$\mathbf{R}=\mathbf{c}\mathbf{s}^{T} \qquad (1.10)$$

donde **s** es un vector de $J\times1$ de absortividades molares normalizadas para el mismo paso óptico en cada una de las J longitudes de onda, y **R** es una matriz de absorbancia de $I\times J$, con cada una de las filas correspondientes a la medición espectral para cada muestra distinta. La matriz de espectros queda expresada como el simple producto externo entre un vector de concentraciones y un vector con el espectro del componente a cuantificar puro. A pesar de que **R** simboliza una matriz de espectros y **c** un vector de concentraciones, esto no es excluyente. **R** bien podría referirse a voltamperogramas y **c** a índices de refracción. Sin embargo, para evitar abstracciones innecesarias, se seguirán utilizando los términos "concentraciones" y "espectros" sin pérdida de generalidad.

Si las mezclas analizadas contienen un conjunto de *A* componentes ópticamente activos, la **Ecuación 1.10** se puede extender de manera trivial como:

$$\mathbf{R}=\mathbf{c}_1\mathbf{s}_1^{T}+\mathbf{c}_2\mathbf{s}_2^{T}+\ldots+\mathbf{c}_A\mathbf{s}_A^{T}=\mathbf{C}\mathbf{S}^{T} \qquad (1.11)$$

donde **C** ($I\times A$) es la matriz de concentraciones (donde las columnas representan las concentraciones de cada uno de los *A* componentes), y **S** la matriz de los espectros puros a concentraciones unitarias. La matriz de respuestas, seguirá teniendo el tamaño $I\times J$, aunque resultará de las contribuciones espectrales de cada uno de los *A* componentes.

En el escenario de calibración clásico, se conocen tanto las concentraciones para cada uno de los componentes espectralmente activos presentes en la mezcla así como la respuesta total medida. En consecuencia, el modelo de calibrado en este caso se basa en resolver la **Ecuación 1.11**, para obtener un estimador de **S**, ($\hat{\mathbf{S}}$) lo que se conoce normalmente como calibración indirecta:

$$\hat{\mathbf{S}} = \mathbf{R}^{T}\mathbf{C}(\mathbf{C}^{T}\mathbf{C})^{-1} \qquad (1.12)$$

En caso que se conozca **S**, la calibración puede llevarse adelante sin estimación, situación conocida como calibración directa.

En la etapa de predicción, al igual que en el caso univariado, se medirán las señales instrumentales para una muestra incógnita, obteniéndose un vector de señales instrumentales $\mathbf{y}_{test,}$ que junto con $\hat{\mathbf{S}}$ permitirá calcular la concentración de la muestra incógnita o de *test* como:

$$\hat{\mathbf{c}}_{test} = (\hat{\mathbf{S}}^{T}\hat{\mathbf{S}})^{-1}\,\hat{\mathbf{S}}^{T}\,\mathbf{r}_{test}^{T} = \mathbf{S}^{+}\mathbf{r}_{test}^{T} \qquad (1.13)$$

La forma de calibrar descripta en el párrafo anterior se denomina regresión por cuadrados mínimos clásicos (CLS) y es muy útil en caso que se disponga de la información necesaria para desarrollarla (concentraciones de los componentes que puedan llegar a influir en la señal final y espectros medidos para cada muestra). Este escenario de calibración se resume en el diagrama de bloques de la **Figura 1.2**.

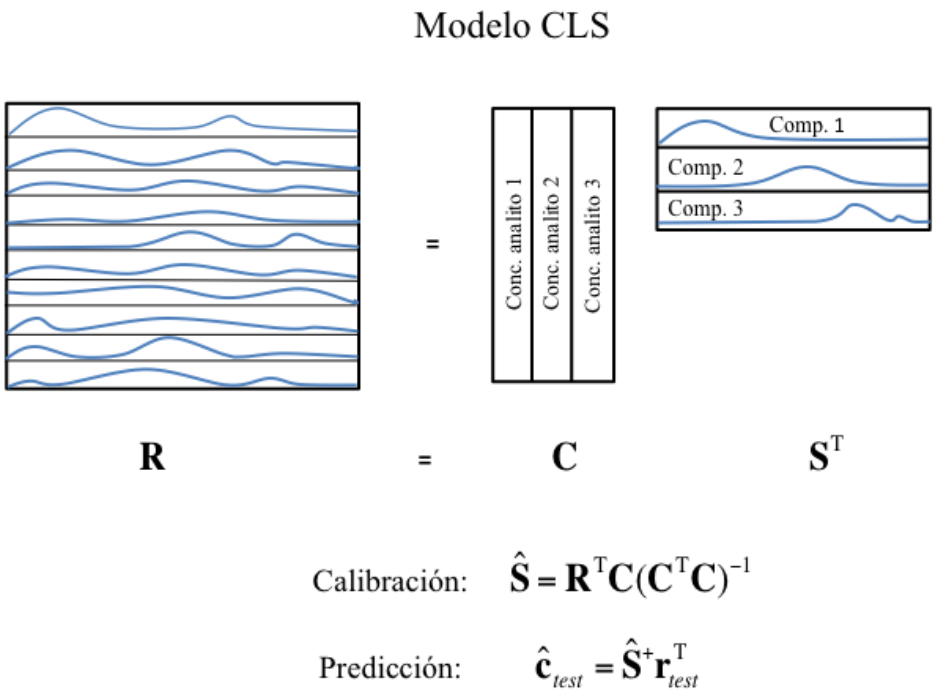


Figura 1.2. Representación en bloques del modelo de calibración multivariada CLS.

Comparada con la regresión univariada, la regresión CLS tiene las siguientes ventajas: (1) permite la determinación de múltiples analitos simultáneamente, (2) los analitos pueden determinarse en presencia de interferentes conocidos, (3) es posible

detectar *outliers* e interferentes no modelados, y (4) utiliza toda la información medida para obtener una precisión máxima. Al mismo tiempo, podrían resaltarse la siguientes desventajas: (1) requiere medir un vector de datos para cada muestra y (2) no es tan intuitiva o fácilmente visualizable y la matemática es más compleja.

Por otro lado, comparada con el resto de los métodos de calibración de primer orden que serán presentados más adelante la principal ventaja de CLS es que se utiliza un modelo muy bien definido que incluye a todos los componentes presentes característica que le da a este modelo simpleza y confiabilidad. Sin embargo, como ya se había dado a entender, la principal desventaja es que requiere conocer exactamente las concentraciones de cada uno de los constituyentes de la mezcla. Otra desventaja adicional desde el punto de vista matemático, es que las inversas de ($\mathbf{C}^T\mathbf{C}$) y ($\mathbf{S}^T\mathbf{S}$) deben poder calcularse, lo cual en caso de significativa colinealidad espectral entre componentes resulta una complicación ya que estas matrices se vuelven singulares o casi singulares, su determinante se acerca a 0 y por lo tanto resulta difícil su inversión.

1.4.2 Regresión por cuadrados mínimos inversos (ILS)

En muchas situaciones, resulta útil calibrar un determinado analito sin conocer las concentraciones del resto de los componentes de la muestra. Como se mencionó anteriormente, en esta situación, se pone de manifiesto una importante limitante del modelo multivariado CLS. La manera de resolver este inconveniente surge de considerar que la **Ecuación 1.11** puede invertirse y expresarse como:

$$\mathbf{C} = \mathbf{RS} \qquad (1.14)$$

O de la manera más general propuesta para el caso univariado

$$\mathbf{Y} = \mathbf{XB} \qquad (1.15)$$

Esta manera de expresar la relación entre la variable a predecir y la predictora se conoce como calibración inversa y se encuentra representada a través del diagrama de bloques de la **Figura 1.3**. En este caso, **Y** es la matriz de concentraciones de tamaño $I \times A$, **X** una matriz de $I \times J$ de respuestas instrumentales, y **B** una matriz de $J \times A$ de coeficientes de regresión que relacionan **X** con **Y**. Una característica interesante de este modelo de regresión inversa es que ya no es necesario conocer todos los componentes del sistema en estudio. Esto quiere decir que el modelo se puede "desacoplar", es decir, restringir el

análisis para el analito que se desea cuantificar sin considerar el resto de los analitos activos en la mezcla. De esta forma, es posible calibrar y cuantificar en presencia de especies interferentes, sin que se necesiten conocer los espectros de los componentes puros ni las concentraciones de las especies que interfieren en la señal del compuesto a determinar.

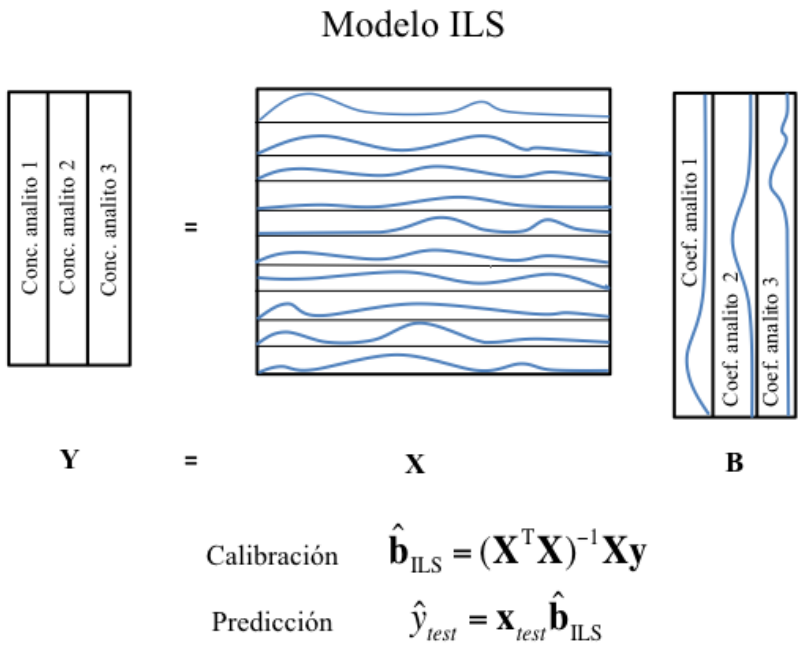


Figura 1.3. Representación en bloques del modelo de calibración multivariada ILS.

En la regresión por cuadrados mínimos inversos (ILS) o calibración multivariada inversa (muchas veces también denominada con la sigla MLR), la solución por mínimos cuadrados para la calibración de un único analito de interés en una mezcla con otras especies interferentes está dada por:

$$\hat{\mathbf{b}}_{\mathrm{ILS}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \qquad (1.16)$$

donde como ya se indicó previamente, **y** es el vector con las concentraciones del analito, y las filas de **X** contienen los espectros medidos. En analogía con la expresión univariada complementaria, la **Ecuación 1.16** a veces se expresa como:

$$\hat{\mathbf{b}}_{\mathrm{ILS}} = \mathbf{X}^{+}\mathbf{y} \qquad (1.17)$$

Teniendo en cuenta esta analogía, es posible visualizar geométricamente la operación algebraica en múltiples dimensiones definida en esta ecuación: el vector de concentraciones estimadas, $\hat{\mathbf{y}}$, es la proyección ortogonal de **y** al subespacio S_x determinado por los vectores columna de **X**, al igual que lo que ocurría en la **Ecuación 1.9**[53]

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{X}\mathbf{X}^+\mathbf{y} \quad (1.18)$$

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}}_{ILS} \quad (1.19)$$

En la **Figura 1.4** se ilustra la solución multivariada por cuadrados mínimos, suponiendo que hay únicamente dos componentes espectralmente activos (un analito y un interferente) en cada una de las muestras.

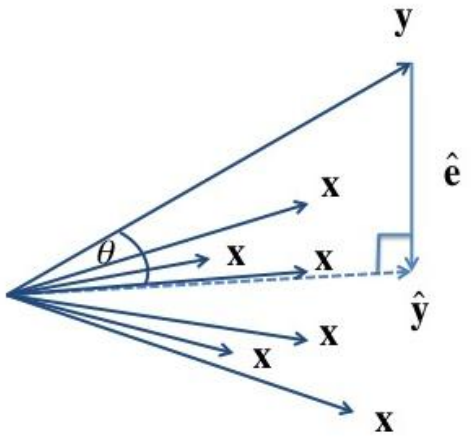


Figura 1.4. Representación vectorial de la solución por cuadrados mínimos en un sistema de dos componentes espectralmente activos. Esta solución corresponde a la proyección ortogonal de **y** al subespacio S_x definido por los vectores columna de **X, x.**

Una vez que los parámetros de la regresión se estiman a través del procedimiento de calibración correspondiente a este modelo, las predicciones pueden realizarse fácilmente por medio de la ecuación

$$\hat{y}_{test} = \mathbf{x}_{test}\hat{\mathbf{b}}_{ILS} \quad (1.20)$$

donde $\mathbf{x}_{test}$ es el vector de respuesta instrumental (espectro), de tamaño $J \times 1$. Para múltiples muestras

$$\hat{\mathbf{y}}_{test} = \mathbf{X}_{test}\hat{\mathbf{b}}_{\mathrm{ILS}} \qquad (1.21)$$

donde $\hat{\mathbf{b}}_{\mathrm{ILS}}$ (tamaño $J\times1$) es un vector de regresión estimado por el modelo ILS para el analito que se desea cuantificar.

Mientras que la calibración inversa logra sortear las dificultades asociadas con la calibración clásica, tiene una dificultad importante relacionada a la inversión de la matriz ($\mathbf{X}^{\mathrm{T}}\mathbf{X}$) en la **Ecuación 1.18**. Para que esta matriz pueda invertirse, el número de sensores (canales de medición o longitudes de onda) en el domino de los espectros debe ser menor que el número de muestras utilizado para construir el modelo de calibración ($J < I$). Dado que la mayoría de los instrumentos y espectrómetros actuales disponen de cientos e incluso miles de canales de medición, la demanda que plantea el modelo ILS es una limitante importante. Otro factor importante que puede generar dificultades a la hora de invertir $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ es la colinealidad que normalmente se observa en **X.**

Hasta cierto punto, la dificultad anterior puede resolverse utilizando métodos de selección de variables en los cuales se selecciona un número de sensores espectrales que hacen posible la inversión $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ y permiten llevar a cabo la calibración. En relación a esta dificultad, en el siguiente capítulo se presentará una discusión más detallada de algunos de los posibles métodos de selección de variables. Si bien estos métodos se utilizarán en el marco de la optimización de modelos multivariados más eficientes que ILS, se podrían aplicar sin ningún problema a este último.[54,55]

Como síntesis de este apartado podría decirse que la posibilidad de desacoplar componentes, es decir de poder estudiar mezclas complejas mediante un proceso de calibración en el que se conoce sólo la concentración del componente de interés, origina la principal ventaja ILS. Por otro lado, la desventaja más importante tiene que ver con que este método es sensible a las colinealidades espectrales, originando las dificultades matemáticas discutidas anteriormente, que en última instancia llevan a que se tenga que utilizar un número reducido de sensores, con la consecuente pérdida de información, y por ende de sensibilidad.

1.4.4 Análisis por componentes principales

PCA[26,56,57] es una técnica de descomposición/proyección de tipo bilineal, capaz de condensar grandes cantidades de datos en unos pocos parámetros, conocidos como

componentes principales, variables latentes o factores, que capturan niveles, diferencias y similitudes entre las muestras y las variables que constituyen los datos modelados. Esto se logra por medio de una transformación lineal que se realiza bajo la restricción de preservar las mayores fuentes de variación de los datos y la ortogonalidad de las variables latentes.

Este tipo de análisis constituye una herramienta sumamente versátil para simplificar la complejidad intrínseca de los datos multidimensionales. Algunos ejemplos clásicos de aplicación en química incluyen:

- Análisis de mezclas: puede ayudar a determinar el número, identidad y concentraciones de los componentes presentes en mezclas desconocidas.

- Análisis exploratorio de datos y reconocimiento de patrones: permite la identificación de distintas clases de muestras en un juegos de datos, el reconocimiento de *outliers*, y la caracterización de los datos.

- Modelado: puede ayudar a determinar si un determinado sistema responde a un modelo físico particular, revelando relaciones ocultas entre las variables.

- Calibración multivariada: provee una alternativa muy interesante para resolver las dificultades previamente mencionadas para la regresión por cuadrados mínimos parciales a través de la regresión por componentes principales. Esta aplicación es la que más interesa a los químicos analíticos cuantitativos y por lo tanto la que será abordada con mayor detalle en el marco de este capítulo.

La suposición implícita de este modelo es que los sistemas en estudio pueden observarse indirectamente en el sentido que los fenómenos que son responsables de la variación y los patrones que se pueden distinguir en los datos se encuentran de alguna forma “ocultos” y por lo tanto no se pueden medir u observar de manera directa.

Una característica única de PCA, relacionada estrictamente con las técnicas de proyección es que permite una visión simultánea e interrelaciona tanto del espacio de las muestras como el de las variables.

A pesar de que la idea es fundamentalmente simple, la dificultad para entender esta técnica se vincula con la gran cantidad de aplicaciones, terminología y descripciones matemáticas que se utilizan en distintas áreas variando desde la matemática, pasando por la física y hasta la economía.

En términos generales, PCA descompone la matriz de datos o respuestas **X** (tamaño $I \times J$) de la siguiente manera:

$$\mathbf{X} = \mathbf{T}\mathbf{V}^{\mathrm{T}} + \mathbf{E} \qquad (1.22)$$

Los vectores de *scores*, $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_A]$ (tamaño $I \times A$) contienen las coordenadas de las muestras en el espacio de los componentes principales del sistema y por lo tanto permiten, por medio de un gráfico de puntos, analizar la similitud o disimilitud entre las muestras. Por otro lado, los vectores de *loadings*, $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_A]$ (tamaño $J \times A$), representan los pesos con los que cada una de las variables originales contribuyen a los componentes principales. **E** es la matriz de residuos, es decir, la parte de los datos que el modelo no llega a explicar. Esta última matriz tiene las mismas dimensiones que **X** y normalmente se utiliza como herramienta de diagnóstico para identificar *outliers* de variables o de muestras. Esta propiedad del modelo PCA se conoce como "ventaja de primer orden".

Es importante destacar que en este caso A hace referencia al número de componentes o rango químico efectivo de la matriz de datos, que coincide con la cantidad de factores o variables latentes generadas por PCA. En otras palabras, A corresponde al número de fuentes de variación significativas que afectan al sistema en estudio.

En este punto es importante realizar una aclaración respecto de un concepto muy utilizado a la hora de analizar las propiedades de los datos multivariados, como es el concepto de bilinealidad. En general, cuando una matriz **X** puede descomponerse como el producto de otras dos matrices, tal y como muestra la **Ecuación 1.22**, se dice que la misma es bilineal, ya que es lineal tanto en **T** como en **V**. Sin embargo, esto no es completamente verdadero. La clave de la bilinealidad de **X**, es que el número de columnas de **T** y **V** sea igual al número de constituyentes que están generando la variación en la medición de **X**. De hecho, cualquier matriz de datos se puede descomponer en el producto de dos matrices. En matrices no bilineales, sin embargo, el número de columnas de **T** y **V** puede exceder significativamente el número de constituyentes químicos o fuentes de variación principales. En este sentido, un concepto que condensa apropiadamente la discusión anterior es el de rango de una matriz. Sin ir a mayores detalles matemáticos, el rango es el número de términos bilineales necesarios para reproducir una matriz de datos: cuando el rango se encuentra cercano al número de constituyentes la matriz es bilineal, mientras que en cualquier otro caso es no bilineal. Por lo tanto, la forma correcta de expresar esta

condición es como “bilineal de rango bajo”. Sin embargo, en lo que resta de esta tesis, se utilizará simplemente la denominación bilineal en lugar de rango, ya que es una terminología más acorde al lenguaje utilizado por los químicos.

Desde un punto de vista geométrico, PCA es una proyección ortogonal de **X** en el sistema de coordenadas definido por los vectores **V**. La **Figura 1.6** muestra un ejemplo de un juego de muestras caracterizadas por dos variables $\mathbf{x}_1$ y $\mathbf{x}_2$, proyectadas en las líneas rectas definidas por los vectores de *loadings* $\mathbf{v}_1$ y $\mathbf{v}_2$. Para cada una de las *I* muestras, se obtiene un vector de *scores* $\mathbf{t}_a$ conteniendo los *scores*, es decir las coordenadas en el espacio de los componentes principales.

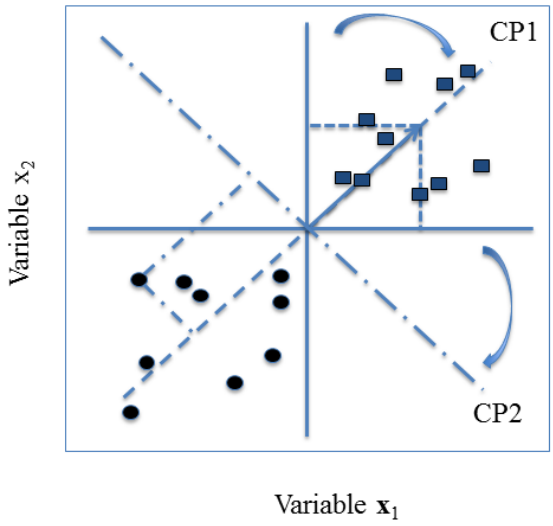


Figura 1.6. Ilustración de la operación de rotación realizada por el modelo PCA en un sistema de 20 muestras caracterizadas por dos variables. Las muestras están graficadas en el espacio de las variables originales $\mathbf{x}_1$ y $\mathbf{x}_2$. Las líneas punteadas representan las direcciones de los ejes correspondientes al primer y segundo componente principal.

Considerando la proyección de estas muestras en el espacio de componentes principales (**Figura 1.6**) puede verse que las dos categorías se encuentran bien separadas en el primer componente principal, mientras que el segundo componente describe fundamentalmente una variabilidad sistemática. Por lo tanto, en este caso el primer componente, es suficiente para retener la información en este conjunto de datos.

De esta manera, PCA opera reduciendo las dimensiones desde un número de variables *I* en **X** a *A* variables latentes que describen la estructura fundamental de los datos. La representación de los *scores* utilizando un gráfico de puntos de dos o tres

dimensiones, permite una visualización intermedia del posicionamiento de las muestras en el espacio de los componentes principales y hace que puedan identificarse con mayor facilidad grupos o tendencias. Los *loadings* representan el peso de cada una delas variables originales a la hora de determinar la dirección de los componentes principales. Dicho en otras palabras, dado que los componentes principales se definen como las direcciones de máxima variancia, los *loadings* estarían indicando cuáles de las variables originales varían en mayor medida para muestras con distintos valores de *scores* en cada componente.

Desde el punto de vista algebraico, PCA puede ser formulado como un problema de maximización matemática con ciertas restricciones. Como ya se indicó previamente los componentes principales son combinaciones lineales de las variables originales:

$$\mathbf{t}_a = \mathbf{X}\mathbf{v}_a \qquad (1.23)$$

donde $\mathbf{v}_a$ y $\mathbf{t}_a$ son los vectores de *loadings y scores* para un componente particular (por eso se utiliza la notación de vector), cumplen con la condición $\mathbf{v}_{ai}^{\mathrm{T}}\mathbf{v}_{ai}=1$ (normalización), $\mathbf{v}_{ai}^{\mathrm{T}}\mathbf{v}_{aj} = 0$ para $j \neq i$ (ortogonalización) y maximización de la var($\mathbf{t}_a$). Por lo tanto, la expresión a ser maximizada, para $a = 1,\ldots A$ es:

$$(\mathbf{X}\mathbf{v}_a)^{\mathrm{T}}(\mathbf{X}\mathbf{v}_a) = \mathbf{v}_a^{\mathbf{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{v}_a = \mathbf{v}_a^{\mathbf{T}}\mathrm{cov}(\mathbf{X})\mathbf{v}_a \qquad (1.24)$$

y la solución puede ser formulada a partir de un problema de autovectores y autovalores, ya bien conocido en el campo del álgebra:

$$\mathrm{cov}(\mathbf{X})\mathbf{v}_{\boldsymbol{a}} = \lambda_a \mathbf{v}_{\boldsymbol{a}} \qquad (1.25)$$

Esto significa en última instancia que los valores de los *loadings* corresponden a los autovectores de la matriz de covariancia de **X** (simbolizada como cov(**X**)) y λ_a son los correspondientes autovalores.

En otras palabras, el cálculo de los componentes principales lleva a una diagonalización de la matriz de covariancia de **X**, cuando **X** se encuentra centrada. En caso que **X** se haya autoescalado lleva a una diagonalización de la matriz de correlación de **X**.

Usualmente, los componentes principales se ordenan en orden decreciente de variancia. Además, considerando la propiedad algebraica de la conservación de la traza, la suma de los autovalores es igual a la variancia total de la matriz **X** (simbolizada como var(**X**)):

$$\sum_a \lambda_a = \mathrm{var}(\mathbf{X}) \qquad (1.26)$$

En caso que **X** se haya autoescalado y sea de rango completo, los autovalores suman el número de variables, *J*.

Teniendo en cuenta lo planteado anteriormente, los *loadings* se pueden obtener a partir de cualquier método que permita el cálculo de autovalores y autovectores. Seguidamente, los vectores de *scores* se pueden calcular a partir de la expresión

$$\mathbf{T}=\mathbf{XV} \qquad (1.27)$$

Los principales algoritmos utilizados para calcular autovectores y autovalores no se describirán en detalle (recurrir a Referencia 58), aunque resulta importante destacar que la manera en la que operan difiere fundamentalmente en dos aspectos:

- La matriz sobre la cual operan: en el caso de la descomposición en autovalores (EVD) y el método POWER, lo hacen sobre $\mathbf{X}^T\mathbf{X}$, y en el caso de la descomposición en valores singulares (SVD)[59] y el método iterativo no lineal de cuadrados mínimos parciales (NIPALS),[60] sobre **X.** Sin embargo, SVD también podría operar sobre $\mathbf{X}^T\mathbf{X}$ dando los mismos resultados que la descomposición en autovalores.

- La manera en que se obtienen los componentes principales: en EVD y SVD simultáneamente y en POWER y NIPALS secuencialmente.

En todos los casos para los cuales la dimensión de las columnas es mucho menor que la de las filas, se puede operar sobre $\mathbf{X}\mathbf{X}^T$ (EVD, POWER, SVD) y sobre $\mathbf{X}^T$ (NIPALS).

La principal ventaja de NIPALS es que es secuencial, de manera que puede detenerse luego que se calculó un determinado número de componentes. Para esto suele utilizarse un criterio de detención, como por ejemplo, un porcentaje deseado de variancia explicada.

1.4.5 Regresión por componentes principales (PCR)

Como el nombre lo sugiere, PCR está basado en el uso del análisis por componentes principales descripto con anterioridad para generar una descripción parsimoniosa (es decir, con la menor cantidad posible de variables latentes) de la matriz

independiente **X**.[25] De hecho, como la proyección de las muestras sobre los primeros componentes principales constituye la mejor aproximación de los datos originales en un espacio de pocas dimensiones, resulta bastante intuitivo pensar que los *scores* de PCA pueden ser utilizados como predictores en ILS (o MLR), de manera de poder superar las limitaciones del método cuando se trabaja con matrices experimentales deficientes de rango. De esta manera, el modelado por PCR es un proceso de dos pasos que en primera instancia realiza la descomposición por PCA del bloque predictor y seguidamente construye un modelo MLR con los *scores*. Este es uno de los motivos por los cuales previamente se había mencionado la importancia del análisis secuencial de los modelos de multivariados de predicción lineales que preceden a PCR y PLS, a fines de lograr una mejor interpretación de estos últimos.

En términos matemáticos, la matriz **X** se describe por medio del modelo bilineal presentado en la **Ecuación 1.22**. Basado en esta descomposición y como ya se mencionó previamente, PCR opera construyendo un modelo MLR sobre los *scores* computados a través de PCA. De esta forma, la ecuación de regresión pasa a expresarse como

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e} = \mathbf{T}\mathbf{v} + \mathbf{e} \qquad (1.28)$$

donde **e** es un vector con los residuos de la regresión de la matriz residual permite diferenciarla del bloque **X**, y **v** (tamaño $1 \times A$) es la matriz de coeficientes de regresión para el modelo MLR que relaciona las variables dependientes **Y** con la matriz de variables latentes **T.** En analogía con la descripción que se presentó para ILS, **v** puede estimarse como:

$$\mathbf{v} = \mathbf{T}^{+}\mathbf{y} = (\mathbf{T}^{T}\mathbf{T})^{-1}\mathbf{T}^{T}\mathbf{y} \qquad (1.29)$$

Analizando la **Ecuación 1.29**, resulta evidente que el vector de coeficientes de regresión **v** relaciona el vector dependiente **y** con los *scores* **T** de **X**. Sin embargo, para lograr una interpretación más simple de los resultados, es más directo expresar la relación de regresión directamente en términos de las variables originales de la siguiente manera:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}}_{PCR} \qquad (1.30)$$

En particular, $\hat{\mathbf{b}}_{PCR}$ es un vector de coeficientes de regresión espectrales estimados, que se reconstruye a partir de **v** por medio de la matriz de *loadings*:

$$\hat{\mathbf{b}}_{PCR}=\mathbf{Vv} \quad (1.31)$$

De esta forma, es posible predecir los valores de las respuestas de nuevas muestras de una manera completamente idéntica a la que ya se presentó en la sección en la que se describió ILS con la única excepción que los coeficientes se obtienen a partir de variables latentes que se generan definiendo un nuevo espacio de proyección por medio de combinaciones lineales de las variables originales y de acuerdo con un criterio específico que en este caso es el de máxima variancia (aunque bien podrían utilizarse otros dependiendo de la necesidad).

Comparado con MLR, como PCR incluye un paso de proyección donde los datos se representan en un espacio de variables latentes de dimensión menor, es necesario decidir cuál debe ser la complejidad de este espacio o, en otros términos, cuántos componentes principales se necesitan. En general, se presenta la problemática de alcanzar un balance a la hora de seleccionar el número óptimo de componentes ya que incluir muy pocos factores puede llegar a modelos que no son capaces de ajustar **X** de una manera lo suficientemente correcta como para predecir **y** de manera precisa, a la vez que la utilización de un exceso de componentes puede resultar en un sobreajuste de **y** y **X**. Por lo tanto, que elegir el número de factores que se utilizarán para modelar los datos requerirá algún procedimiento de validación, en el que el número óptimo de componentes principales se selecciona como aquel que lleva al menor error de predicción entre las estimaciones de la validación.

Una posible desventaja del modelado por PCR es que los componentes principales no se relacionan necesariamente con **y.** De hecho, la principal característica de PCA es que extrae aquellas fuentes que capturan lo máximo posible la variación en **X**: sin embargo, en caso que haya muchas fuentes de variación que no aportan información o cuando los niveles de ruido son muy elevados, seguramente la correlación con **y** será muy pobre (y, por lo tanto, no predictiva). Para solucionar este tipo de problemas, algunos autores sugirieron seleccionar únicamente las variables latentes que más se correlacionan con las respuestas. Sin embargo, la manera más común de lidiar con este tipo de limitaciones es utilizar algún criterio diferente para proyectar los datos en un espacio de pocas dimensiones, lo cual explícitamente exige tener en cuenta las concentraciones o valores de referencia del componente de interés a la hora de ejecutar la descomposición bilineal. El ejemplo más popularizado de este tipo de métodos es la regresión por PLS.

1.4.6 Regresión por cuadrados mínimos parciales (PLS)

Como se describió en la sección anterior, PCR consiste en un modelo que opera básicamente en dos pasos en los cuales la etapa de proyección se encuentra separada de la de regresión. Esto último tiene la desventaja de que en algunas condiciones, los componentes que se extraen en el paso de descomposición, basados únicamente en la información contenida en la matriz **X**, pueden resultar poco predictivos en relación con el vector de valores de referencia **y**. Teniendo en cuenta esta consideración, se propuso la metodología de cuadrados mínimos parciales, en la que la información en **y** se utiliza activamente en la definición del espacio de variables latentes. PLS busca aquellos componentes que logren el máximo compromiso entre explicar la variación en el bloque **X** y predecir las respuestas en **y**. Esto corresponde a un modelo bilineal que puede resumirse matemáticamente como:

$$\mathbf{X}=\mathbf{T}\mathbf{P}^{T}+\mathbf{E} \tag{1.32}$$

$$\mathbf{y} = \hat{\mathbf{y}}+ \mathbf{e} = \mathbf{T}\mathbf{V} + \mathbf{e} \tag{1.33}$$

que es formalmente idéntico a lo que se presentó en PCR, con la diferencia que los componentes calculados y los coeficientes del modelo no tienen el mismo valor, debido a que las proyecciones están guiadas por criterios diferentes. En particular, como ya se mencionó, los *scores* de PLS **T** se definen de tal manera que sean relevantes tanto desde el punto de vista interpretativo como predictivo, lo cual se traduce matemáticamente por medio del concepto estadístico de covariancia. De hecho, dado que no sólo se tiene en cuenta la correlación entre las variables, sino también el grado de variación en cada una, la covariancia representa la medida justa de interrelación permitiendo formular el criterio que permite definir las variables latentes del modelo PLS, criterio que se aplica componente a componente y que por lo tanto no puede traducirse en una única función global de optimización. De esta forma, PLS consiste en un algoritmo secuencial en el que las variables latentes se van computando de tal manera que el primer componente de PLS es la dirección de máxima covariancia respecto a la variable dependiente, el segundo componente es ortogonal al primero y contiene la máxima variancia residual, y así sucesivamente.

De acuerdo con el criterio descripto previamente, el cálculo del primer vector latente de PLS corresponde a identificar una dirección en el espacio multivariado, definida

por el vector de pesos unitario $\mathbf{w}_1$, de manera tal que los *scores* en la dirección $\mathbf{t}_1$ tengan máxima covariancia con **y:**

$$\max_{\mathbf{w1}}[\mathrm{cov}(\mathbf{t_1},\mathbf{y})] = \max_{\mathbf{w1}}\left(\mathbf{t}_1^{\mathrm{T}}\mathbf{y}\right) \quad (1.34)$$

donde

$$\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1 \quad (1.35)$$

$$\mathbf{w}_1 = \mathbf{X}\mathbf{y}/(\mathbf{y}^{\mathrm{T}}\mathbf{y}) \quad (1.36)$$

y

$$\|\mathbf{w}_1\|^2 = 1 \quad (1.37)$$

Existen varias maneras de calcular los parámetros del modelo PLS de manera que cumplan con la condición establecida en la **Ecuación 1.34**.[61] Uno de los algoritmos más conocidos debido a su gran estabilidad numérica es NIPALS. Este algoritmo calcula los *scores* **T**, los *loadings* **P** (de tamaño $J \times A$ y similares a los utilizados en PCR) y un conjunto adicional de vectores conocidos como *loadings* de peso, que se localizan en una matriz normalmente denominada como **W** (con la misma dimensionalidad que los *loadings* **P**). La adición de estos pesos en PLS, resulta importante para mantener a los *scores* ortogonales entre sí. Es importante aclarar que las columnas de **W** no son autovalores. Son sólo factores que se obtienen por medio de una metodología distinta de la utilizada en PCR, y cuyos elementos dependen de la concentración de calibración del analito de interés. Para mayor especificidad, en la Referencia 61 se muestra el código detallado del algoritmo NIPALS para PLS, así como del resto de los posibles algoritmos que permiten aplicar este modelo. Respecto de NIPALS, que es el algoritmo que se utilizó durante el desarrollo de esta tesis, es interesante remarcar que el mismo opera a través de pasos de deflación, en los cuales la matriz **X** reconstruida a partir de los valores calculados de **T** y **P** se resta a la matriz **X** original, o del paso anterior en las siguientes iteraciones.

La etapa de calibrado, requiere un primer paso que consiste en la estimación del número óptimo de factores *A*, que normalmente se realiza por medio de una técnica conocida como validación cruzada dejando de lado una muestra por vez.[24] En este procedimiento, cada muestra se quita del conjunto de calibración, y su concentración se predice usando un modelo construido con los espectros de las muestras restantes y un número de prueba de factores. El cuadrado del error para la predicción de cada muestra que

fue dejada de lado se acumula en un parámetro llamado PRESS, que es una función de *A*. El número óptimo de factores se estima computando las razones $F(A) = \text{PRESS}(A<A^*)/\text{PRESS}(A^*)$, (donde $\text{PRESS} = \Sigma(y_{i,} - \hat{y}_i)^2$, A es el número de factores de prueba, A^* corresponde al PRESS mínimo, $y_{i,}$ es la concentración nominal del analito de la muestra *i*, y $\hat{y}_i$ el correspondiente valor de concentración predicha), seleccionando el número de factores que corresponden a una probabilidad menor al 75% para $F>1$.

El resultado principal de la calibración es el vector de coeficientes de regresión **v** cuyos elementos (v_1,.., v_A) se calculan en cada una de las iteraciones del algoritmo. Luego, en la etapa de predicción, estos coeficientes se utilizan para estimar las concentraciones del analito en la muestra.

Es posible formular una relación matemática para obtener **T** a partir de **X** teniendo en cuenta los distintos pasos de deflación como:

$$\mathbf{T}=\mathbf{XV} \qquad (1.38)$$

En particular se puede demostrar que los valores de la matriz **V** se pueden obtener de:

$$\mathbf{V}=\mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1} \qquad (1.40)$$

Dado que los valores de respuesta se pueden predecir de los *scores* de PLS como:

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e} = \mathbf{Tv} + \mathbf{e} \qquad (1.41)$$

es posible expresar el modelo lineal en términos de los predictores originales combinando las **Ecuaciones 1.38-1.41**:

$$\hat{\mathbf{y}} = \mathbf{Tv} = \mathbf{XVv} = \mathbf{XW}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{v} = \mathbf{X}\hat{\mathbf{b}}_{PLS} \qquad (1.42)$$

En esta ecuación, el vector $\hat{\mathbf{b}}_{PLS}$ contiene los coeficientes de regresión para el modelo PLS expresados en términos de las variables originales y se define como:

$$\hat{\mathbf{b}}_{PLS}= \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{v} \qquad (1.43)$$

De manera análoga a lo descripto para MLR este conjunto de coeficientes permite predecir los valores de las respuestas para una nueva muestra $\hat{y}_{test}$ dado un conjunto de predictores medidos $\mathbf{x}_{test}$, de acuerdo con:

$$\hat{y}_{test} = \mathbf{x}_{test}\hat{\mathbf{b}}_{\mathrm{PLS}} \qquad (1.44)$$

1.5 Métodos multivariados de orden superior

Conocer las propiedades y la estructura de los datos multi-vía que se están midiendo, proporciona el criterio para seleccionar un determinado modelo o herramienta de procesamiento de datos. Como ya se mencionó durante la introducción de esta tesis, esta elección de modelo afectará de manera importante el valor de las cifras de mérito, llevando a otra característica peculiar de la calibración multi-vía: la especificidad de estas cifras de acuerdo con el algoritmo utilizado. Es por esto que en esta sección se hará referencia a una de las principales características de los datos multi-vía para luego describir más detalladamente a los tres más utilizados en la actualidad.

1.5.1 Modelos bilineales y trilineales

El arreglo multi-vía más simple que puede encontrarse es una matriz para cada muestra individual, lo cual genera datos de segundo orden. Si en la muestra hay N constituyentes que están generando respuesta, cada uno de los elementos x_{jk} de estas matrices de datos se puede escribir como:

$$x_{jk} = \sum_{n=1}^{N} b_{jn}c_{kn} + e_{jk} \qquad (1.45)$$

donde b_{jn} y c_{kn} son los valores de las respuestas específicas dadas por los canales instrumentales j y k para el constituyente n, y e_{jk} es un término que contiene los residuos del modelado. Como ya se mencionó previamente durante este capítulo, cuando la matriz no se puede expresar como la suma de unos pocos términos bilineales, se la considera como no bilineal. En general, la bilinealidad se pierde cuando los fenómenos que se dan en los dos modos instrumentales son mutuamente dependientes [62].

Los datos de segundo orden, pueden considerarse como el "ingrediente" a partir del cual se generan arreglos de tres vías, que son los datos multi-vía más simples. Un arreglo de tres vías será trilineal de bajo rango si puede expresarse como la suma de pocos

componentes trilineales cuando la mezcla contiene sólo algunos componentes. Las matrices de excitación y emisión que se obtienen por fluorescencia (EEFM) son un ejemplo típico para el cual la condición de trilinealidad se cumple. Si un cierto número de EEFM (*I*) se apila en la dirección de las muestras, generando un arreglo de tres vías **X**, y las muestras son mezclas de *N* constituyentes fluorescentes, la señal específica x_{ijk} de la muestra *i*, a la longitud de onda de emisión *j*, y la longitud de onda de excitación *k* pueden escribirse como:

$$x_{ijk} = \sum_{n=1}^{N} a_{in} b_{jn} c_{kn} + e_{ijk} \qquad (1.46)$$

donde a_{in} es proporcional a la concentración del constituyente *n* en la muestra *i*, b_{jn} al rendimiento cuántico de emisión a la longitud de onda *j*, y c_{kn} al coeficiente de absorción a la longitud de onda de excitación *k*.

1.5.2 Modelos que permiten lidiar con desviaciones de la trilinealidad

Aunque en la **Ecuación 1.46** no sea completamente evidente, para fines prácticos resulta útil considerar las siguientes condiciones que deben cumplirse para que un conjunto de datos sea trilineal: (1) las matrices de datos individuales deberían ser bilineales, es decir, los perfiles $\mathbf{b}_n$ y $\mathbf{c}_n$ no deben depender uno de otro, y (2) b y c no deben depender de la muestra, es decir, deberían existir vectores únicos $\mathbf{b}_n$ y $\mathbf{c}_n$, para todos los modos instrumentales y para todas las muestras. En los datos cromatográfico-espectrales, los perfiles de elución no son exactamente reproducibles de muestra a muestra. Debido a esto, un arreglo de tres vías compuesto por matrices de este tipo en general no será trilineal. Sin embargo, dado que las matrices individuales son bilineales, se pueden aumentar los datos en el sentido de los tiempos de elución. Esto lleva a una única matriz aumentada en la dirección del tiempo ($\mathbf{X}_{aug}$) que también es bilineal y puede expresarse como:

$$x_{aug,pk} = \sum_{n=1}^{N} b_{aug,pn} c_{kn} + e_{aug,pk} \qquad (1.47)$$

con el índice *p* variando de 1 a *IJ*, debido a que el tamaño de la matriz aumentada es de $IJ \times K$ (*I* = número de muestras, *J* = número de tiempos de elución, *K*= número de longitudes de onda y otros sensores espectrales). El perfil espectral $\mathbf{c}_n$ (en el modo no aumentado) es único para cada constituyente y común para todas las muestras, mientras

que $\mathbf{b}_{aug,n,}$ es el perfil aumentado en el tiempo y está compuesto de *I* subperfiles sucesivos a *J* tiempos cada uno.

Como consecuencia de la discusión anterior, es posible clasificar los datos de tres vías como: (1) trilineales, (2) no trilineales con un único modo en el cual se pierde la linealidad y por lo tanto es posible desplegarlo y procesar los datos como una matriz aumentada, y (3) otros no trilineales. Este último tipo se refiere a estructuras de datos en los que la trilinealidad no se cumple en dos de los modos instrumentales de medición o casos en los que las matrices individuales no son bilineales.

Teniendo en cuenta los algoritmos disponibles, éstos se pueden clasificar en tres grupos dependiendo de la conexión entre el modelo matemático subyacente y los distintos tipos de datos que pueden presentarse : (1) modelo multilineal, (2) modelo bilineal para una matriz aumentada y (3) modelo basado en variables latentes. El grupo 1 incluye a PARAFAC[51], el grupo 2 la resolución multivariada de curvas acoplada a cuadrados mínimos a (MCR-ALS)[52] particularmente en su versión extendida[63], y el grupo 3 tanto la versión desplegada como la multi-vía de cuadrados mínimos parciales (U-PLS y N-PLS).[64,65]

Tanto PARAFAC como MCR-ALS alcanzan la ventaja de segundo orden procesando simultáneamente las muestras de calibrado y las de *test*, debido a que el funcionamiento interno del algoritmo permite descomponer la contribución de los potenciales agentes interferentes y la de los analitos de la señal total. Sin embargo, en el caso de las metodologías tipo PLS, la ventaja de segundo orden se alcanza en una etapa pos-calibración basada en un procedimiento conocido como multilinealización residual (RML).[47-50]

Como nota final de esta sección es importante considerar que en ocasiones entre estos algoritmos, la diferenciación en términos de aplicación, no está completamente definida, y podría darse un solapamiento. De cualquier modo, es probable que los desarrollos futuros tengan en cuenta consideraciones acerca de la sensibilidad y el LOD, como herramientas para tomar decisiones en este sentido.

1.5.3 Análisis paralelo de factores (PARAFAC)

PARAFAC es en esencia un método de descomposición que puede interpretarse conceptualmente como un análisis de componentes principales (PCA) para datos multi-

vía.[51] Este método fue desarrollado originalmente por Harshman[66] para analizar datos provenientes del campo de la psicología, y por Carrol and Chang con el nombre de CADECOMP.[67]En ambos casos los autores se basaron en el principio de perfiles proporcionales sugerido por Cattell.[68]

Entre los métodos de descomposición trilineales, PARAFAC es uno de los más utilizados en el ámbito de la química analítica, siendo el ejemplo típico de aplicación el análisis de conjuntos de matrices de excitación-emisión de fluorescencia. Las medidas de EEFM son rápidas y normalmente no requieren de un paso previo de preparación. La estructura de los datos (dos conjuntos independientes de variables como son los espectros de excitación y emisión, y otro conjunto de variables dependiendo de ambos perfiles como son las concentraciones), como se mencionó previamente, hace que las matrices de fluorescencia cumplan con el requisito de trilinealidad (en caso que no surjan artefactos dispersivos). En consecuencia, la combinación de EEFM y PARAFAC se convirtió en una herramienta analítica muy popular para analizar muestras con matrices muy variadas, especialmente en la ciencia de los alimentos.

A lo descripto en el párrafo anterior y a la posibilidad de alcanzar la ventaja de segundo orden, se suma una serie de propiedades que convierten a PARAFAC en una técnica atractiva desde un punto de vista analítico. La primera es que, a diferencia de PCA, no presenta la necesidad de ortogonalidad a la hora de computar los factores para construir el modelo. En otras palabras, bajo las restricciones adecuadas, los *loadings* de PARAFAC reflejarán el comportamiento fisicoquímico real de los analitos que influyen en la variabilidad de la señal. Por lo tanto, si los datos son aproximadamente trilineales, utilizando un número adecuado de componentes y bajo una relación señal-ruido apropiada, será posible llegar a una comprensión del fenómeno subyacente al sistema en estudio. La segunda es la capacidad del modelo PARAFAC de converger a una solución única. Esto quiere decir que, en la mayoría de las circunstancias, partiendo de una determinada estructura de datos, se llega a un único modelo que no requiere de ningún tipo de procesamiento posterior. Esta es una diferencia fundamental respecto de MCR-ALS, que si bien presenta un grado de versatilidad mayor que PARAFAC, para poder converger a una solución única requiere de la aplicación de restricciones adecuadas que permitan llegar a la solución más conveniente entre varias posibles (problema conocido como de “ambigüedad rotacional”).

En la calibración de tres vías utilizando el modelo trilineal PARAFAC se mide una matriz de datos para cada muestra, y las matrices de calibración se combinan con las de la muestra incógnita generando un arreglo de tres vías **X** (**Figura 1.7**) que puede descomponerse de acuerdo con la **Ecuación 1.46**.

Por medio del algoritmo ALS, de este arreglo pueden obtenerse los perfiles para cada uno de los componentes, designados como $\mathbf{a}_n$, $\mathbf{b}_n$, y $\mathbf{c}_n$, que recolectarán, respectivamente, las concentraciones relativas para el componente *n*, así como sus perfiles en ambos modos de datos.

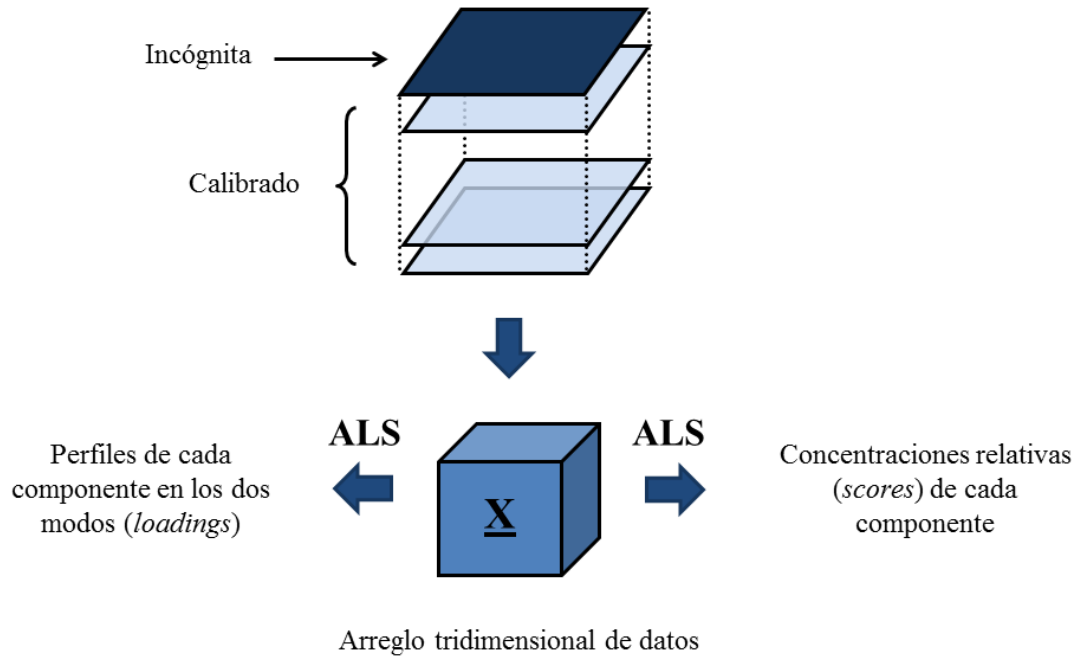


Figura 1.7. Representación gráfica del modelo PARAFAC para datos de segundo orden. Las matrices de datos para cada una de las muestras de calibración se ubican junto con la muestra de *test* que se desea determinar formando un cubo o arreglo de tres vías, que se analiza manteniendo la estructura original de los datos.

1.5.4 Resolución multivariada de curvas (MCR)

Este modelo, en su modo aumentado, se utiliza con frecuencia en química analítica cuantitativa. Como muestra la **Figura 1.8**, ubica las matrices (o arreglos multi-vía desdoblados en matrices) de un grupo de muestras, cada una contigua a la otra, formando una matriz aumentada que se supone sigue un modelo bilineal. Esto implica que un elemento de matriz (*p*,*k*) es la suma de las contribuciones de la forma ($b_{aug,pn} \times c_{kn}$), donde

$b_{aug,pn}$ describe los elementos de los perfiles para cada muestra en la dirección aumentada, y c_{kn} en el modo común (p va de 1 a IJ).

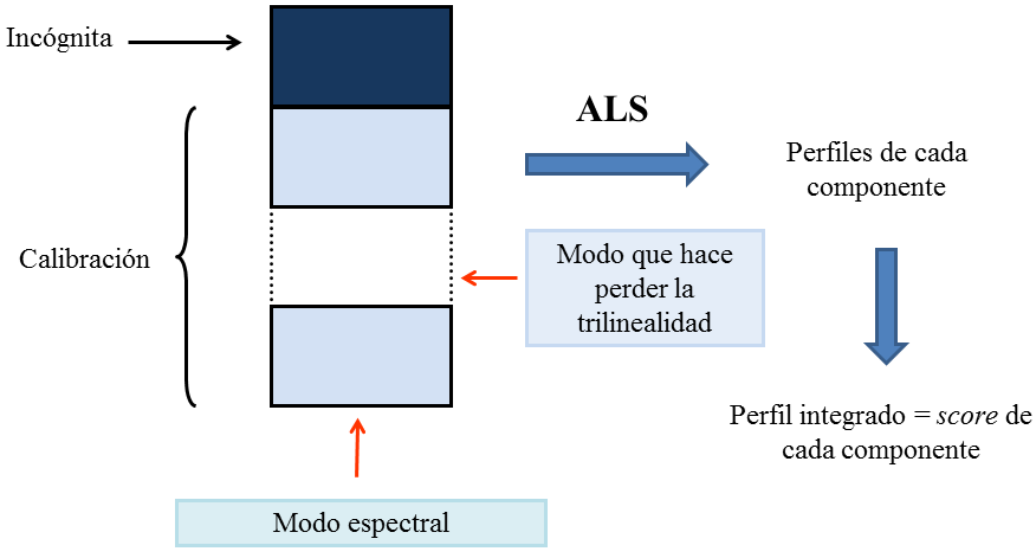


Figura 1.8. Representación gráfica del modelo MCR. Las matrices de datos obtenidas para cada muestra se ubican de manera contigua generando una matriz de datos aumentada (**X***aug*). Al igual que en PARAFAC, las muestras de calibración se procesan junto con la incógnita.

MCR puede ser interpretado como un desdoblamiento de los datos no trilineales de tres vías en una matriz bilineal que se ajusta a la **Ecuación 1.46**.

Durante la calibración, equivale a descomponer una matriz de datos aumentada y generada a partir de las matrices de las muestras de calibración y de muestras desconocidas. Al igual que en PARAFAC, la descomposición mencionada se logra por medio del algoritmo ALS, en conjunto con una serie de restricciones, que permiten que la solución sea interpretable físicamente, y ayuda a limitar el número de posibles soluciones (disminuyendo la ambigüedad rotacional). MCR-ALS necesita estimaciones iniciales de los perfiles de los componentes, aunque estos últimos pueden obtenerse eficientemente mediante una variedad de métodos.[69] Para la cuantificación de analitos, se computan las áreas debajo de cada uno de los perfiles de las muestras en el modo aumentado y se utilizan en un gráfico de calibración pseudo-univariada.

1.5.5 Cuadrados mínimos parciales desdoblados acoplados a multilinealización residual (U-PLS/RML)

En el modelo de cuadrados mínimos parciales desdoblados, los datos originales son vectores que se desdoblan antes de aplicar PLS, tal como se muestra en la **Figura 1.9**. Esto se realiza concatenando para cada muestra las dos dimensiones instrumentales (en caso que se trate de un sistema de segundo orden), de tal manera que cada matriz de muestras de $J \times K$ genere un vector de $JK \times 1$.Seguidamente, en la fase de calibración, se emplea la información de concentración sin incluir información de la muestra incógnita. Con los I_{cal} vectores de calibración y el vector de concentraciones **y** (de tamaño $I_{cal} \times 1$) se construye el modelo PLS convencional. Este genera un conjunto de *loadings* **P** y de *weight loadings* **W** (ambos de tamaño $JK \times A$, donde A es el número de factores latentes), así como también los coeficientes de regresión **v** (de tamaño $A \times 1$).

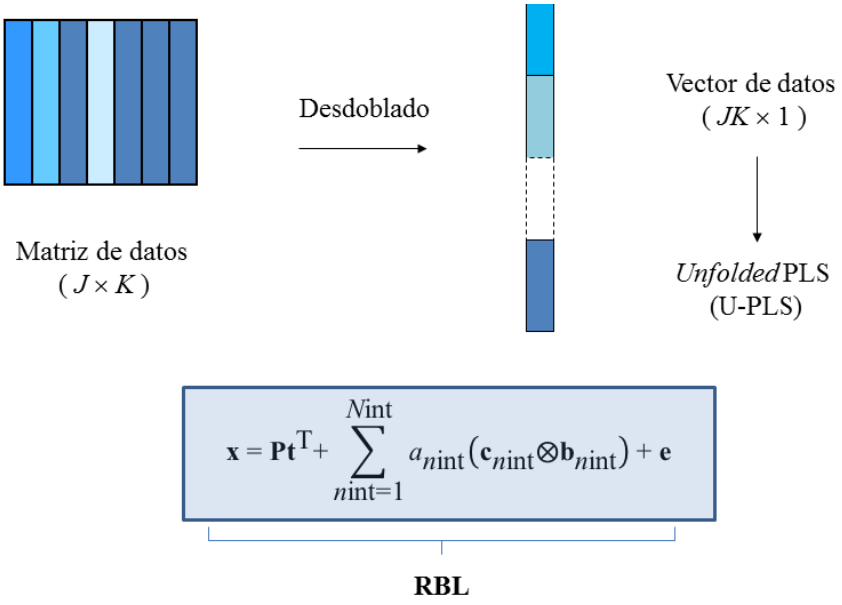


Figura 1.9. Representación gráfica del modelo UPLS/RBL. Desdoblamiento de la matriz de datos y obtención de un vector sobre el que luego se aplica PLS de manera tradicional. Los *scores* resultantes se someten a un procedimiento de bilinealización residual para modelar los componentes inesperados y evitar que estos generen un sesgo en la predicción.

El parámetro A puede seleccionarse por medio de técnicas como validación cruzada dejando de lado una muestra por vez.[24]

Si no aparecen componentes inesperados en la muestra de *test*, **v** se puede utilizar para estimar la concentración del analito de acuerdo con:

$$\hat{y}_{test} = \mathbf{tv} \qquad (1.47)$$

donde **t** son los *scores* de la muestra de *test*, obtenido al proyectar los datos vectorizados provenientes de la muestra de *test* $\mathbf{x} = \mathrm{vec}(\mathbf{X}_{test})$ en el espacio de los A factores latentes de acuerdo con:

$$\mathbf{t} = \mathbf{x}^{\mathrm{T}} (\mathbf{P}^{\mathrm{T}} \mathbf{W})^{-1} \mathbf{W} \qquad (1.48)$$

donde vec(.) es el operador de vectorización.

Cuando aparecen constituyentes inesperados en **x**, los *scores* de la muestra dados por la **Ecuación 1.48** no son adecuados para la predicción de analitos a través de la **Ecuación 1.47**. En este caso, los residuos de la predicción por U-PLS (s_p, ver **Ecuación 1.49** abajo) serán anormalmente mayores que el ruido instrumental típico:

$$s_{\mathrm{p}} = \| \mathbf{x} - \mathbf{P}\mathbf{t}^{\mathrm{T}} \| / (JK - A)^{1/2} \qquad (1.49)$$

donde $\| \cdot \|$ indica la norma Euclidiana.

Esta situación puede ser abordada por un procedimiento posterior a la calibración llamado bilinealización residual (RBL), que se basa en un modelo bilineal para las señales interferentes.[70] Este último normalmente consiste en un análisis de componentes principales (PCA) basado en descomposición de valores singulares (SVD). El objetivo del procedimiento RBL es minimizar la norma del vector residual **e**, que se computa mientras se ajustan los datos de la muestra a la suma de las contribuciones relevantes:

$$\mathbf{x} = \mathbf{P}\mathbf{t}^{\mathrm{T}} + \sum_{n_{\mathrm{int}}=1}^{N_{\mathrm{int}}} \mathrm{a}_{n\mathrm{int}} (\mathbf{c}_{n\mathrm{int}} \otimes \mathbf{b}_{n\mathrm{int}}) + \mathbf{e} = \mathbf{P}\mathbf{t}^{\mathrm{T}} + \mathrm{vec}(\mathbf{B}_{\mathrm{int}} \mathbf{A}_{\mathrm{int}} \mathbf{C}_{\mathrm{int}}^{\mathrm{T}}) + \mathbf{e}$$

$$= \mathbf{P}\mathbf{t}^{\mathrm{T}} + \mathrm{vec}\{\mathrm{SVD}[\mathrm{reshape}(\mathbf{x} - \mathbf{P}\mathbf{t}^{\mathrm{T}})]\} + \mathbf{e} \qquad (1.50)$$

donde $\mathbf{B}_{\mathrm{int}}$ y $\mathbf{C}_{\mathrm{int}}$ son matrices cuyas columnas son $\mathbf{b}_{\mathrm{int}n}$ y $\mathbf{c}_{\mathrm{int}n}$ que a la vez son los primeros autovectores izquierdo y derecho de la matriz ('reshape' indica la transformación de un vector de $JK \times 1$ en una matriz de $J \times K$), $\otimes$ es el operador matemático correspondiente al producto de Kronecker y SVD implica la descomposición por valores singulares con los primeros N_{int} componentes principales. La matriz diagonal $\mathbf{A}_{\mathrm{int}}$ contiene los N_{int} valores

singulares obtenidos del análisis por SVD. Los detalles para estimar el valor óptimo de N_{int} ya fueron mencionados anteriormente.

Durante RBL, **P** se mantiene constante a los valores de calibración, y **t** se varía hasta que $\| \mathbf{e} \|$ se minimiza en la **Ecuación 1.50**, empleando normalmente un procedimiento de Gauss-Newton. Es importante tener en cuenta que en algunos casos este esquema converge a un mínimo erróneo desde el punto de vista químico.[71] Para resolver este problema, se ha propuesto preceder RBL con un paso de optimización por enjambre de partículas (PSO), un método estocástico para encontrar el mínimo global basado en computación natural. Una vez que $\| \mathbf{e} \|$ se minimiza, las concentraciones del analito se obtienen utilizando la **Ecuación 1.47**, pero con la introducción de un vector **t** "depurado" del efecto de los componentes inesperados por medio del procedimiento RBL.

Para un único componente inesperado (es decir, N_{int} =1), el análisis es directo y proporciona los correspondientes perfiles de interferente puro. Sin embargo, para componentes inesperados adicionales (N_{int} >1), los perfiles $\mathbf{B}_{int}$ y $\mathbf{C}_{int}$ devueltos no se corresponden con perfiles verdaderos. En lo que se refiere a la estimación de N_{int}, debería notarse que el objetivo que guía RBL es la minimización de los errores residuales s_u a un nivel compatible con el valor de los residuos en las señales medidas, donde s_u se calcula como:[72]

$$s_u = \| \mathbf{e} \| / [(J - N_{int})(K - N_{int}) - A]^{1/2} \qquad (1.51)$$

De esta forma, si se considera más de un componente inesperado, RBL debería seleccionar el modelo más simple que lleve a un valor residual que no sea estadísticamente diferente del mínimo. Es importante notar que en la discusión anterior aparecen dos parámetros residuales diferentes que no deberían confundirse: s_p (**Ecuación 1.49**) corresponde a la diferencia entre la señal de la muestra de *test* y aquella modelada por U-PLS antes de aplicar el procedimiento RBL, mientras que s_u (**Ecuación 1.51**) surge de la diferencia luego de que se efectúa el modelado del interferente por RBL. Es esta última la que debería ser comparable con el ruido instrumental si RBL resulta exitoso.

Para datos de tercer y cuarto orden, se utiliza el algoritmo U-PLS combinado con trilinealización residual (RTL)[49] o con cuadrilinelización residual (RQL).[50] Estos son formalmente análogos al modelo previo de U-PLS/RBL, excepto que los datos originales para cada muestra son arreglos de tres vías de tamaño $J \times K \times L$ en RTL y de cuatro vías con tamaño $J \times K \times L \times M$ en RQL. El cambio más importante teniendo en cuenta este último

hecho se aplica a la **Ecuación 1.50**, que emplea SVD para modelar las señales interferentes. Esta se reemplaza por modelos de Tucker, como extensiones apropiadas de los procedimientos de tres y cuatro vías de datos instrumentales por cada muestra. De aquí que la **Ecuación 1.50** del modelo RTL se reemplaza por:

$$\mathbf{x} = \mathbf{P}\mathbf{t}^T + \text{vec}\{\text{Tucker3}[\textit{reshape}(\mathbf{x} - \mathbf{P}\mathbf{t}^T)]\} + \mathbf{e} \quad (1.52)$$

y en RQL por:

$$\mathbf{x} = \mathbf{P}\mathbf{t}^T + \text{vec}\{\text{Tucker4}\,[\textit{reshape}(\mathbf{x} - \mathbf{P}\mathbf{t}^T)]\} + \mathbf{e} \quad (1.53)$$

donde los modelos de Tucker3 y Tucker4 se construyen con un máximo de *N* componentes en cada dimensión instrumental (ver arriba). Tener en cuenta que *'reshape'* indica la transformación de un vector de $JKL\times1$ en un arreglo de tres vías de $J\times K\times L$ en la **Ecuación 1.52** y un vector de $JKLM\times1$ en un arreglo de cuatro vías de $J\times K\times L\times M$ en la **Ecuación 1.53**.

Ambos métodos apuntan a minimizar los residuos **e**, manteniendo los *loadings* **P** constantes a los valores de calibración, y variando **t** hasta que el valor de || **e** || se minimiza utilizando un procedimiento de Gauss-Newton. Luego de completar el proceso, RTL permite conocer las contribuciones de los interferentes en la forma de los *loadings* $\mathbf{B}_{int}$, $\mathbf{C}_{int}$ and $\mathbf{D}_{int}$ definidos para *N* componentes en tres modos de datos instrumentales. La estimación de *N* sigue principios análogos a aquellos discutidos anteriormente para RBL. En lo que respecta a RQL, los resultados son análogos a RTL, con excepción de que se obtiene la matriz de *loadings* $\mathbf{E}_{int}$, correspondiente a la cuarta dimensión instrumental.

Los modelos de Tucker de las **Ecuaciones 1.52** y **1.53** normalmente se construyen restringiendo los *loadings* a la ortogonalidad, como una extensión lógica de PCA a arreglos de tres y cuatro vías. Para un único componente inesperado, el modelo de Tucker se construye con un único componente en todos los modos, lo cual proporciona los perfiles interferentes correspondientes. Para componentes inesperados adicionales, sin embargo, los perfiles devueltos no se parecen, en general, a perfiles reales. Más aún, en este último caso, en principio se podrían construir diferentes modelos de Tucker, ya que el número de *loadings* podría ser diferente en cada modo. De esta forma, al considerarse dos componentes inesperados, por ejemplo, se deberían explorar los posibles modelos de Tucker teniendo uno o dos *loadings* en cada modo, y seleccionar el modelo más simple que

de un valor residual estadísticamente indistinguible del mínimo. Para más de un componente inesperado se recomienda un procedimiento similar.

El uso de modelos de Tucker en RTL y RQL es la forma más general de modelar arreglos de múltiples vías describiendo la contribución de los interferentes a la señal total. De cualquier manera, en casos prácticos donde las señales provienen de algún fenómeno químico conocido, se pueden emplear modelos más simples. Si se sospechan que estas señales son multilineales, por ejemplo, se podría utilizar el análisis paralelo de factores (PARAFAC) para modelar estas contribuciones. En tal caso, las **Ecuaciones 1.52** y **1.53** se podrían reemplazar respectivamente por:

$$\mathbf{x} = \mathbf{P}\mathbf{t}^{\mathrm{T}} + \sum_{n=1}^{N} a_{\mathrm{int}n}(\mathbf{d}_{\mathrm{int}n} \otimes \mathbf{c}_{\mathrm{int}n} \otimes \mathbf{b}_{\mathrm{int}n}) + \mathbf{e} =$$

$$= \mathbf{P}\mathbf{t}^{\mathrm{T}} + \mathrm{vec}\{\mathrm{PARAFAC}[\mathit{reshape}(\mathbf{x} - \mathbf{P}\mathbf{t}^{\mathrm{T}})]\} + \mathbf{e} \qquad (1.54)$$

$$\mathbf{x} = \mathbf{P}\mathbf{t}^{\mathrm{T}} + \sum_{n=1}^{N} a_{\mathrm{int}n}(\mathbf{e}_{\mathrm{int}n} \otimes \mathbf{d}_{\mathrm{int}n} \otimes \mathbf{c}_{\mathrm{int}n} \otimes \mathbf{b}_{\mathrm{int}n}) + \mathbf{e} =$$

$$= \mathbf{P}\mathbf{t} + \mathrm{vec}\{\mathrm{PARAFAC}[\mathit{reshape}(\mathbf{x} - \mathbf{P}\mathbf{t}^{\mathrm{T}})]\} + \mathbf{e} \qquad (1.55)$$

donde $\mathbf{b}_{\mathrm{int}n}$, $\mathbf{c}_{\mathrm{int}n}$, $\mathbf{d}_{\mathrm{int}n}$ y $\mathbf{e}_{\mathrm{int}n}$ son los perfiles de los interferentes, es decir, columnas de las matrices $\mathbf{B}_{\mathrm{int}}$, $\mathbf{C}_{\mathrm{int}}$, $\mathbf{D}_{\mathrm{int}}$ and $\mathbf{E}_{\mathrm{int}}$ respectivamente. Si los perfiles se normalizan a longitud unitaria, se requieren los factores de escalado $a_{\mathrm{int}n}$, como aparecen en las **Ecuaciones 1.54** y **1.55**.

Finalmente, los nuevos *scores* se obtienen modelando el arreglo residual [*reshape*(**x** – **Pt**)], utilizando en cada caso modelos PARAFAC de *N* componentes para tres y cuatro vías, donde *'reshape'* significa la transformación de un vector en arreglos de tres o cuatro vías en las **Ecuaciones 1.54** y **1.55** respectivamente.

Los modelos de PARAFAC de tres y cuatro vías presentados son casos especiales de los modelos de Tucker, con la restricción que: (1) el número de *loadings* en cada modo instrumental son los mismos en PARAFAC pero pueden ser diferentes en los modelos de Tucker, y (2) no se permiten interacciones entre los perfiles de PARAFAC, en contraste a los modelos de Tucker. En el caso de los datos analizados en esta tesis (tanto simulados como experimentales), las señales interferentes se pueden modelar en principio utilizando PARAFAC, ya que son multilineales. De cualquier modo, en todos los casos se utilizó la forma más general de RML, es decir, se emplearon modelos de Tucker para tener en

cuenta las contribuciones de los interferentes. Esto se hizo con el propósito de chequear si la expresión de sensibilidad derivada en la sección siguiente es aplicable a un marco más general posible en el contexto del modelo U-PLS/RML.

1.6 Síntesis final del capítulo

Las características de los modelos descriptos durante este capítulo pueden resumirse por medio del siguiente cuadro:

Modelo	**Tipo de datos al que se aplica**	**Ventajas**	**Limitaciones**	**Ecuación de calibración y predicción**
Calibración univariada	Escalar	- Simplicidad. - Universalidad	- No admite presencia de interferentes.	$\hat{x}_{test} = \hat{b} y_{test}$ $\hat{b} = (\mathbf{x}^T\mathbf{x})^{-1}\mathbf{x}^T\mathbf{y}$
CLS	Vector (datos de primer orden)	- Simplicidad (extensión directa de calibración univariada). - Utiliza variables originales. - Se puede calibrar utilizando el espectro completo.	-Sensible a colinealidad espectral. - Necesidad de conocer todos los componentes espectralmente activod de mezclas incógnita.	$\hat{\mathbf{c}}_{test}=\mathbf{S}^+\mathbf{r}_{\boldsymbol{test}}^{\mathbf{T}}$ $\hat{\mathbf{S}} = \mathbf{R}^T\mathbf{C}(\mathbf{C}^T\mathbf{C})^{-1}$
ILS		- Calibración directa (posibilidad de desacoplar componentes)	- Sensible a colinealidad espectral. -Requiere más muestras que sensores de medición.	$\hat{y}_{test}= \mathbf{x}_{test}\hat{\mathbf{b}}_{\mathrm{ILS}}$ $\hat{\mathbf{b}}_{\mathrm{ILS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
PCR		- Calibración	- Utiliza factores	$\hat{y}_{test}= \mathbf{x}_{test}\hat{\mathbf{b}}_{\mathrm{PCR}}$ $\hat{\mathbf{b}}_{\mathrm{PCR}}=\mathbf{V}\mathbf{v}$

		directa. - Elimina problemas asociados con la colinealidad espectral. - Posibilidad de detectar la presencia de potenciales interferentes (ventaja de primer orden).	calculados únicamente a partir de los espectros sin considerar las concentraciones. - No permite lidiar con desviaciones severas de la linealidad.	$\mathbf{v} = \mathbf{T}^{+}\mathbf{y} = (\mathbf{T}^{T}\mathbf{T})^{-1}\mathbf{T}^{T}\mathbf{y}$
PLS			- No permite lidiar con desviaciones severas de la linealidad.	$\hat{y}_{test} = \mathbf{x}_{test}\hat{\mathbf{b}}_{PLS}$ $\hat{\mathbf{b}}_{PLS} = \mathbf{V}\mathbf{v}$ $\mathbf{V} = \mathbf{W}(\mathbf{P}^{T}\mathbf{W})^{-1}$ $\mathbf{v} = \mathbf{T}^{+}\mathbf{y} = (\mathbf{T}^{T}\mathbf{T})^{-1}\mathbf{T}^{T}\mathbf{y}$
PARAFAC	Matrices, arreglos de tres vías y superiores	- Ventaja de segundo orden. - Solución única - Se obtienen perfiles con sentido fisicoquímico. - Se mantiene la estructura original de los datos	-La calidad predictiva disminuye significativamente frente a desviaciones de la trilinealidad	$\hat{y}_{test} = a_{test,n}/k$ *k*: pendiente de la recta de calibración pseudo-univariada. $a_{test,n}$: *scores* obtenidos por descomposición trilineal, para el component *n* en la muestra de *test*.
MCR-ALS		- Ventaja de segundo orden, permite tratar desviaciones de la trilinealidad - Utilizando las restricciones adecuadas se obtienen perfiles con sentido fisicoquímico	-Soluciones variables de acuerdo al tipo de restricciones aplicadas - Ambigüedad rotacional	$\hat{y}_{test} = a_{test,n}/k$ *k*: pendiente de la recta de calibración pseudo-univariada. $a_{test,n}$: *scores* obtenidos a partir del área bajo la curva del perfil resuelto por descomposición bilineal, para el componente de interés *n*

				en la muestra de *test*.
UPLS/RML UPLS/RML		-Ventaja de segundo orden. - Admite ciertas desviaciones de la trilinealidad.	- Trabaja con variables latentes y por lo tanto los perfiles obtenidos no son fáciles de interpretar.	$\hat{y}_{test} = \mathbf{v}^{T}\mathbf{t}_{RML}$ $\mathbf{t}_{RBL}$: *scores* obtenidos por el procedimiento de multilinealización residual

CAPÍTULO 2

OPTIMIZACIÓN DEL MODELO PLS

"Inténtalo otra vez. Falla otra vez. Falla mejor." (Samuel Beckett).

2.1 Resumen

En este capítulo se describirá una nueva estrategia para optimizar el método multivariado de regresión por cuadrados mínimos parciales (PLS) presentado en el capítulo anterior. Ésta fue diseñada integrando tres recursos de optimización que hasta el momento demostraron ser eficientes a la hora de mejorar los modelos de calibración PLS: (1) la selección de variables basada en la optimización por colonias de hormigas, (2) la aplicación de distintas combinaciones de preprocesamientos matemáticos utilizando un algoritmo genético y (3) la selección de muestras basada en metodologías de medida de distancia. Como parte de la optimización de los modelos también se incluyó la detección de *outliers*. Para lograr una metodología integrada, todos los caminos de optimización mencionados se combinaron en un único algoritmo, cuyo objetivo fue el de encontrar el mejor modelo de calibración PLS por medio de una secuencia iterativa de tipo Monte Carlo. Finalmente, para demostrar la eficiencia de la estrategia propuesta, esta se evaluó en juegos de datos simulados y experimentales.

2.2 Introducción

El origen de la calibración multivariada de primer orden se remonta a los años 60. Hoy en día, se ha establecido como una metodología robusta y confiable para el análisis de materiales industriales, con el paradigmático ejemplo de la fuerte unión existente entre la espectroscopía de infrarrojo cercano y la regresión por PLS, dando lugar a una exitosa combinación entre técnicas instrumentales y quimiométricas.[28] Como se ha mencionado en el capítulo anterior, actualmente PLS es considerado como una referencia en la mayoría de las aplicaciones de primer orden, existiendo numerosos libros y artículos referidos a este modelo.[23-26] Luego del surgimiento y establecimiento del modelo PLS de calibración, las principales investigaciones se orientaron a optimizar su funcionamiento. Como se hizo referencia previamente, existen tres maneras principales de llevar a cabo esta optimización: (1) seleccionando las variables (longitudes de onda o sensores en general) más relevantes, (2) aplicando métodos de preprocesamiento matemático de espectros y (3) seleccionando un conjunto de muestras de calibración que sean lo más representativas posibles.

Durante la introducción general de esta tesis, se anticipó que en calibración espectroscópica multivariada la selección de variables apunta a escoger racionalmente, a partir de un conjunto de espectros, aquellas longitudes de onda donde las señales tienen un máximo de información en lo que respecta al analito de interés, descartando al mismo tiempo aquellas que poseen información irrelevante (ruido o regiones de saturación), o aquellas fuertemente solapadas con otros componentes de la muestra que no son de interés analítico.[73,74] Aunque el enfoque está dirigido principalmente a la información espectral, la selección de variables también puede aplicarse a cualquier técnica multivariada en la que algunos sensores podrían ser, en principio, más selectivos en lo que respecta al analito o propiedad de interés, al tiempo que otros podrían generar señales omisibles. Existen publicaciones que muestran la eficacia de la selección de variables en la optimización del rendimiento de PLS, respaldando el continuo interés en esta actividad quimiométrica. [75-77]

Las técnicas de preprocesamiento matemático tienen como principal objetivo eliminar las variaciones espectrales que se producen de una medición a otra y que no tienen relación con los cambios en la concentración del analito a cuantificar.[78,79] La eliminación de efectos poco deseados, como la dispersión que se genera en materiales sólidos y semisólidos en NIRS, conduce a modelos PLS más parsimoniosos (es decir, que

requieren menos variables latentes) y con mejores indicadores estadísticos que aquellos que utilizan datos crudos.

En tercer lugar, la selección de muestras es una actividad importante en el análisis por regresión PLS de muestras complejas (ya sea resultantes de la manufactura industrial o provenientes de la naturaleza) y busca dar representatividad al juego de datos utilizado para construir un modelo particular.[80] Esto significa que sus correspondientes espectros deberían incluir la mayor proporción posible de la variabilidad potencial de muestras futuras.

La detección de *outliers* se ha discutido extensamente en la literatura y se han propuesto varios métodos diagnósticos.[26] Desde un punto de vista formal, un *outlier* es un valor que no es representativo del resto de los datos. En el contexto de la calibración PLS, el objetivo principal es el de identificar muestras con características que las hagan significativamente diferentes de las restantes.[26]

Las actividades mencionadas en los párrafos anteriores están mutuamente conectadas. El preprocesamiento espectral modifica las características del espacio espectral, llevando a la selección de diferentes muestras de entrenamiento y de diferentes longitudes de onda. De la misma manera, modificar las regiones espectrales utilizadas tiene una importante influencia en el método de preprocesamiento elegido. Las selección de muestras, por otro lado, es también importante en la optimización: si las muestras que son realmente representativas se incluyen sólo en el juego de muestras de monitoreo (utilizado para evaluar la calibración) y no en el de entrenamiento (utilizado para efectuar la calibración propiamente dicha), la elección de los parámetros del modelo podría estar mal direccionada. Los *outliers* (muestras con concentraciones nominales o propiedades de referencia erróneas), también pueden afectar potencialmente a la calibración y por esto deberían ser identificados y excluidos de la misma.

Teniendo en cuenta lo expuesto anteriormente, los distintos procesos de selección podrían, en principio, llevarse adelante sobre la base de un mecanismo de prueba y error hasta alcanzar la convergencia, aunque sería mucho más conveniente disponer de una metodología que simultáneamente seleccione variables, preprocesamientos, muestras y *outliers*. Un primer paso respecto de esta integración resultó de combinar la selección de variables y diversos métodos de preprocesamiento en un único algoritmo genético (AG).[81]

En este capítulo, se presentará una nueva estrategia integrada de optimización de calibraciones PLS, que combina las actividades mencionadas en los párrafos anteriores en un único algoritmo que utiliza procedimientos específicos y óptimos para cada uno de los tipos de selección (preprocesamientos, muestras y variables). A este algoritmo se lo denominó ACOGASS debido a que integra tres algoritmos conocidos como: ACO para selección de variables, GA para seleccionar preprocesamientos matemáticos y diversas estrategias de selección de muestras (SS).

2.3 Objetivos específicos

1) Combinar estrategias de optimización del modelo de regresión multivariada PLS que normalmente se aplican de manera separada, en un único algoritmo iterativo e integrado.

2) Evaluar el rendimiento del algoritmo desarrollado en distintos conjuntos de muestras medidas por NIRS, analizando por medio de indicadores estadísticos la calidad predictiva del modelo PLS optimizado a través del mismo.

2.4 Estrategias generales de optimización el algoritmo PLS

Como se había mencionado durante la introducción general de esta tesis, existen distintos métodos de selección de variables que pueden utilizarse para optimizar PLS. Dependiendo de cómo se realice esta selección, estos métodos pueden identificarse de manera general en tres categorías esquematizadas en la **Figura 2.1**:[77]

1) Selección por filtros: utilizan directamente los parámetros de salida del algoritmo PLS para identificar un subconjunto de variables de mayor relevancia. Es decir, se seleccionan las variables en dos pasos: en el primero se ajustan los datos utilizando la regresión por PLS y en el segundo, se lleva adelante la selección de variables determinando algún límite o medida de significancia obtenida a partir del modelo PLS ajustado. La principal ventaja de este tipo de métodos es que requieren poco esfuerzo desde el punto de vista del tiempo de cálculo. Sin embargo tienen como desventaja la dificultad para establecer un valor límite confiable que permita determinar si la variable será seleccionada o descartada. Ejemplos clásicos de este tipo de métodos lo constituyen el análisis de los coeficientes de regresión y/o de los *loadings*, y la medida de importancia de las variables en la proyección.

2) Selección iterativa: las variables identificadas por el método de filtrado se vuelven a introducir en el modelo PLS para realizar un nuevo ajuste y así sucesivamente una determinada cantidad de veces. Estos métodos se distinguen según la elección de la metodología de filtrado y según cómo se lleve adelante el proceso de iteración (por lo general se alterna el ajuste con la selección de variables). Habitualmente funcionan a través de algún método de aprendizaje supervisado orientado por el valor obtenido de una determinada función objetivo. El algoritmo de búsqueda extrae un subconjunto de variables relevantes y evalúa cada subconjunto ajustando el modelo según ese conjunto de variables seleccionadas. En teoría, lo ideal sería evaluar todos los posibles conjuntos de datos que se puedan generar. Sin embargo, para un conjunto elevado de variables esto se torna sumamente impráctico y demandante desde el punto de vista computacional. La solución a este problema es la utilización de algoritmos de búsqueda que sean capaces de evitar mínimos (o máximos según el caso) locales. Estos algoritmos interactúan con el modelo con cierto riesgo de sobreajuste y tiempo de cálculo computacional intensivo.

Teniendo en cuenta lo anterior, los métodos de selección iterativa se pueden categorizar dependiendo del algoritmo de búsqueda utilizado, que puede ser determinista o estocástico. Los algoritmos de búsqueda estocásticos utilizan algún tipo de aleatorización en la selección de subconjuntos mientras que los deterministas no. Ejemplos de metodologías de selección de tipo deterministas son la selección de variables no informativas (UVE-PLS),[82] eliminación de variables "hacia atrás",[83,84] y eliminación regularizada.[82] Ejemplos de metodologías estocásticas son las que se utilizaron en este trabajo, es decir, los algoritmos genéticos y la optimización por colonias de hormigas.

Como última nota respecto de este tipo de métodos, es importante destacar que los métodos de selección iterativa de tipo deterministas son en general más simples y requieren menos cómputos, tienen menor riesgo de sobreajuste y necesitan un número menor de parámetros a ajustar que los métodos estocásticos. Sin embargo, tienen una probabilidad más alta de caer en mínimos locales comparados con las metodologías estocásticas.

3) Selección embebida: la selección de variables se integra a los pasos normales de PLS generando una versión de PLS modificado. Es decir, combinan la selección de variables y el modelado en un procedimiento de un único paso. La búsqueda de un subconjunto óptimo de variables se realiza sobre cada uno de los componentes del modelo PLS. Estas metodologías anidan la selección de variables en el mismo algoritmo PLS, y por lo tanto funcionan por medio de un único mecanismo iterativo siguiendo la forma del algoritmo PLS estándar, a diferencia de las metodologías de selección descriptas en el punto anterior en las que las iteraciones son dobles: una se utiliza para buscar variables y la otra para realizar el ajuste PLS. Por lo tanto, los métodos embebidos normalmente llevan menos tiempo que los basados en una selección con doble iteración.

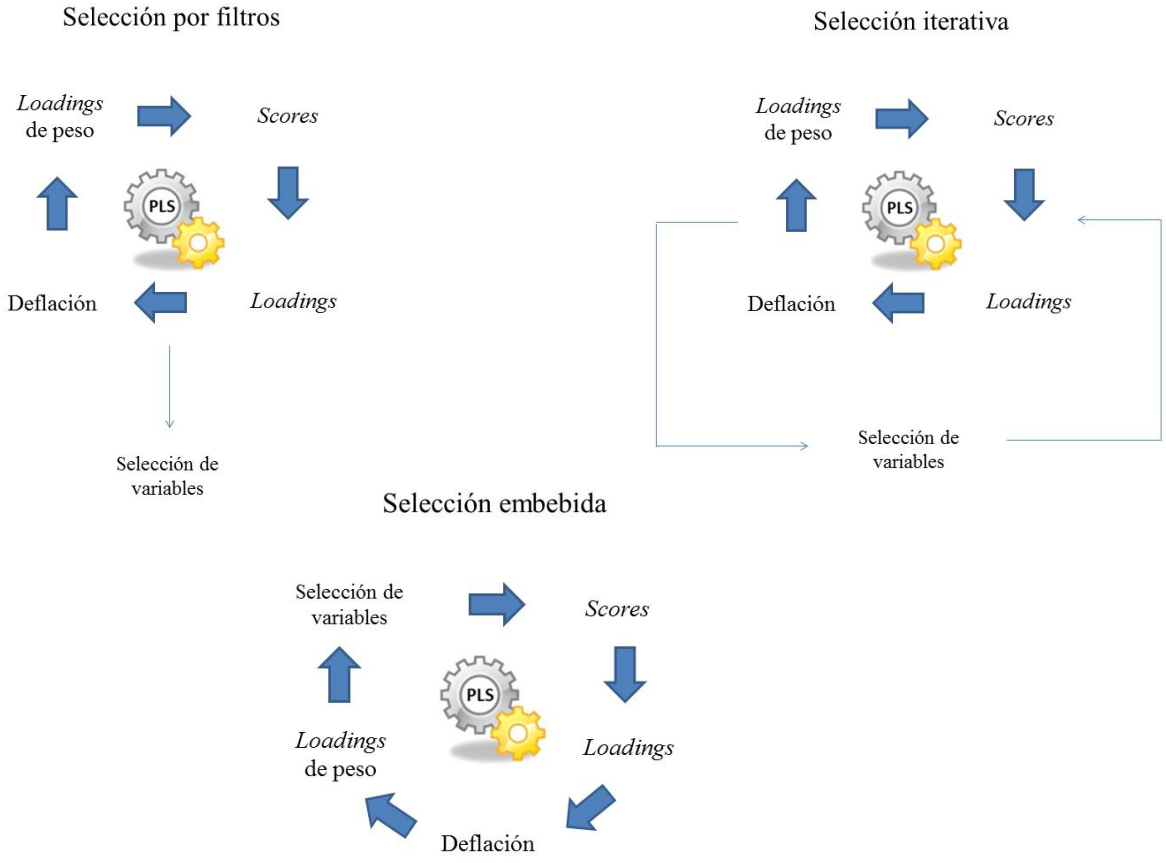


Figura 2.1. Esquemas generales de funcionamiento de los métodos de selección por filtro, iterativos y embebidos, aplicados normalmente a PLS.

2.5 Nuevo método estocástico integrado de optimización

Selección de variables utilizando un algoritmo de optimización por colonia de hormigas (ACO)

Para llevar a cabo la selección de variables, se utilizó la estrategia de optimización por colonias de hormigas (ACO), en lugar de los algoritmos genéticos. Como se anticipó, el funcionamiento de ACO está inspirado en el comportamiento de las hormigas en la búsqueda del mejor camino hacia las fuentes de alimento, por medio de un mecanismo de señalización a través de feromonas.

Al igual que en otros métodos de selección de variables, en el algoritmo ACO propuesto, cada variable a ser seleccionada se define como un rango o bloque de sensores con un ancho espectral predefinido. De esta manera, esta metodología selecciona variables una a una hasta que se haya elegido un cierto número máximo establecido por el usuario.

Es importante aclarar que la estrategia diseñada no sigue exactamente la formulación del algoritmo ACO original.[85] Esto se debe principalmente a que el objetivo principal fue el desarrollo de un algoritmo simple, inspirado en la filosofía de optimización por colonias de hormigas y capaz de llegar a resultados de selección aceptables cuando se acopla a cálculos de tipo Monte Carlo. De cualquier manera, el algoritmo básico descripto en este informe tiene cierta similaridad con el "sistema de hormigas" desarrollado por Dorigo.[44]

El diagrama de flujo de la **Figura 2.2** ilustra de manera compacta el funcionamiento de ACO. Inicialmente se genera un vector **p** de tamaño $J \times 1$, donde J es el número total de variables disponibles (es decir, bloques de sensores individuales). Un elemento de vector genérico del vector **p**, p_j recolecta la cantidad de feromona asociada a la j-ésima variable en cada paso. Al comenzar, todos los elementos de **p** son iguales a 1, lo cual significa que todas las variables tienen la misma probabilidad de ser seleccionadas.

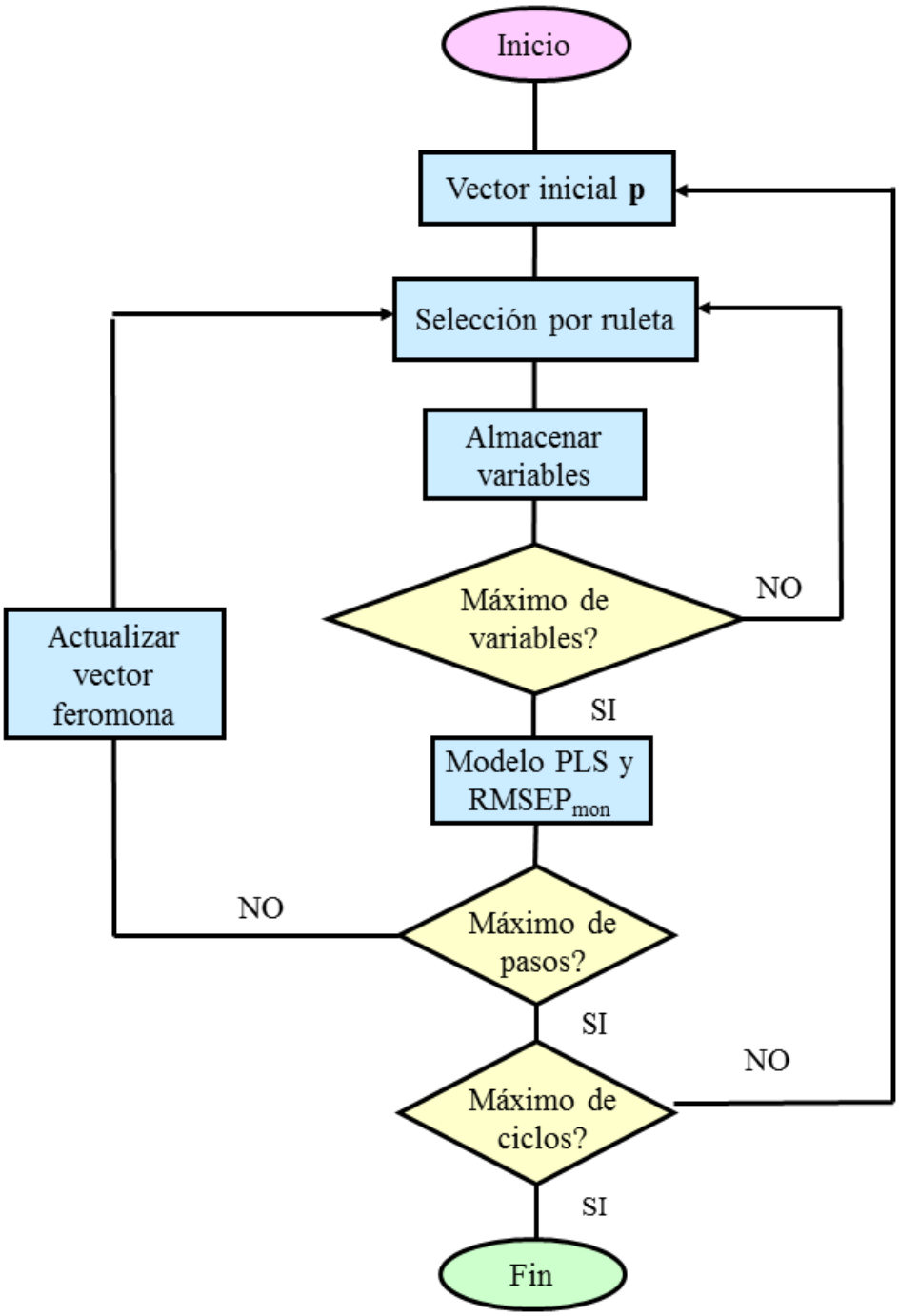


Figura 2.2. Diagrama de flujo del algoritmo ACO utilizado para selección de variables en PLS.

Seguidamente se elige un determinado número de variables (*s*) de las *J* variables disponibles de acuerdo con el contenido de feromona en el elemento correspondiente del vector **p** utilizando un modelo de selección por "rueda de ruleta" (**Figura 2.2**). En esta metodología de selección, se le da un valor de ajuste a cada una de las variables

participantes, relacionado con la probabilidad de selección. Si se le asigna p_j (elemento j del vector **p**) al ajuste de la j-ésima variable, su probabilidad de ser seleccionada será:

$$prob_j = p_j / \sum_{j=1}^{J} p_j \qquad (2.1)$$

La implementación del modelo de rueda de ruleta en un casino se lleva a cabo de la siguiente manera: una proporción de la rueda se asigna a cada uno de los posibles candidatos de acuerdo con un valor de ajuste obtenido (normalmente la raíz cuadrada de error cuadrado medio de predicción o RMSEP), de manera que los candidatos que mejor ajusten abarquen una mayor proporción de la rueda. Esto puede lograrse dividiendo el valor del ajuste cuando se selecciona un conjunto determinado de variables por el que se obtiene cuando se seleccionan todas las variables, normalizando de esta manera a 1. El valor de ajuste es normalmente el RMSEP obtenido al evaluar el modelo predictivo generado con las variables seleccionadas, sobre un conjunto de muestras de monitoreo ($RMSEP_{mon}$). Seguidamente, se realiza una selección semi-aleatoria similar a la rotación de una rueda de ruleta. En este caso el grado de ajuste de cada variable está dado por los elementos de un vector **p,** que a su vez tienen más probabilidad de seleccionar una variable determinada ($prob_j$). Luego de la selección, al valor $prob_j$ de la variable recién seleccionada se le asigna 0 para evitar la duplicación, y la selección comienza nuevamente siguiendo el mismo esquema de ruleta hasta que un conjunto de s variables se haya seleccionado. Esto da lugar a un vector **v** de tamaño $s \times 1$ de variables seleccionadas. Nótese que en el primer paso, todas las variables tienen la misma probabilidad de ser seleccionadas, pero en la medida que **p** se actualiza en los sucesivos pasos, esta probabilidad cambiará.

Con las variables que fueron seleccionadas, se estima el $RMSEP_{mon}$. Como se mencionó previamente, este parámetro se computa construyendo un modelo PLS que correlacione la propiedad o concentración de interés con las señales generadas por las variables seleccionadas por cada hormiga artificial o ente algorítmico de búsqueda. Es importante notar que se debe definir un número máximo de factores latentes de PLS antes que el programa comience a funcionar. Luego, el número óptimo de factores se estima en cada paso de selección como aquel que lleva a un valor de $RMSE_{mon}$ que no es estadísticamente diferente del mínimo, evitando de esta forma un sobreajuste.

En cada uno de los pasos sucesivos, el vector **p** se actualiza de acuerdo con:

$$\mathbf{p}(t) = (1-\rho)\ \mathbf{p}\ (t-1) + \Delta\mathbf{p} \qquad (2.2)$$

donde t se refiere a la iteración actual, ρ la velocidad de evaporación de feromonas ($\rho<1$) y $\Delta\mathbf{p}$ es el vector de cambio de feromona. Estos cambios se dan sólo en algunas variables, debido a que cada hormiga deposita feromonas en los elementos del vector correspondientes a las variables que seleccionó. Específicamente, si el vector **v** es el vector de variables seleccionadas, la contribución a $\Delta\mathbf{p}$ de una hormiga determinada se da sobre el vector $\Delta\mathbf{p}$ con un índice υ_k de forma que:

$$\Delta p_a[\upsilon_k] = -\log(\mathrm{RMSE}_{\mathrm{mon}})_a \qquad (2.3)$$

donde a identifica a una hormiga en particular.

Una vez que las s variables son seleccionadas por cada hormiga, los valores de Δp_a para todas las variables y todas las hormigas son sumados en las posiciones apropiadas de los vectores, para obtener el vector $\Delta\mathbf{p}$ requerido en la **Ecuación 2.2**.

El esquema anterior muestra que varias hormigas pueden contribuir al mismo elemento del vector $\Delta\mathbf{p}$**,** mostrando un comportamiento cooperativo ausente en AG.

El algoritmo descripto ha sido recientemente aplicado con éxito en nuestro laboratorio para seleccionar sensores espectrales en distintos juegos de datos NIRS, optimizando calibraciones PLS y mostrando mejores rendimientos en este sentido con respecto a otras alternativas como son los algoritmos genéticos y la optimización por enjambre de partículas (PSO). La mejoría en la performance se vincula fundamentalmente con dos razones complementarias: (1) la efectividad de la colonia de hormigas artificiales en la búsqueda de mejores soluciones, y (2) el acoplamiento de ACO con una estrategia de tipo Monte Carlo que da mayor repetitividad y confiabilidad a la selección de variables. Por estos motivos, se eligió ACO como algoritmo para seleccionar variables en ACOGASS. La **Figura 2.3** muestra un diagrama de flujo de la fracción del algoritmo ACO descripto que se integró a ACOGASS.

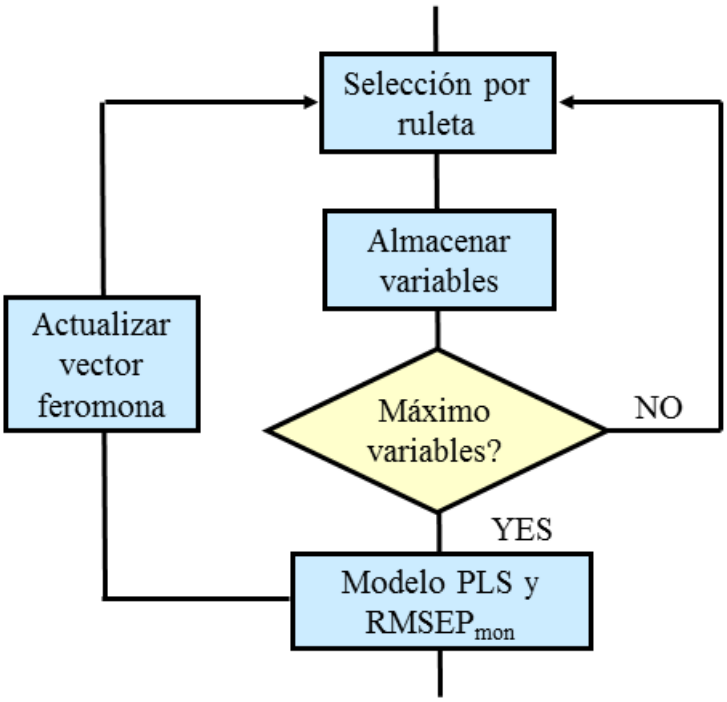


Figura 2.3. Diagrama de flujo de la sección del algoritmo ACO para selección de variables que se utilizó para construir ACOGASS.

Selección de preprocesamientos utilizando un algoritmo genético (GA)

La elección de un método de preprocesamiento adecuado o de la combinación de un conjunto de métodos puede consumir una gran cantidad de tiempo si se lleva adelante sobre la base de prueba y error. Es por esto que se decidió implementar esta actividad por medio de un AG.[41, 42]

El funcionamiento de los algoritmos genéticos está inspirado en el proceso de entrecruzamiento y mutación de los cromosomas, y al igual que ACO son de naturaleza estocástica. En el caso particular de ACOGASS, cada cromosoma consiste en una combinación particular de métodos de preprocesamiento.

Se define en primer lugar una población de *N* cromosomas, cada uno conteniendo un número de genes igual al número total de métodos de preprocesamiento considerados, inicializados con dígitos binarios al azar con una probabilidad de 50% para valores de 1 y una de 50% para valores de 0. Un valor de 1 implica la inclusión del gen (y del preprocesamiento asociado) en el modelo, mientras que el 0 indica su exclusión. Una vez

que se construyeron *N* modelos iniciales, se ordenan de acuerdo con su desempeño en la minimización de la función objetivo ($RMSEP_{mon}$). A la mitad de los cromosomas con los mejores valores de las funciones de ajuste se les permite sobrevivir, mutar y recombinarse para producir descendencia. Para esto, se emplea el llamado entrecruzamiento "de sitio único", en el que se selecciona un punto al azar a lo largo de un par de cromosomas padres, y toda la "información genética" codificada en uno de los padres hasta ese punto se transfiere a la descendencia, mientras que los genes restantes se toman del otro padre. En este proceso, la probabilidad de recombinación es del 50%, y en la descendencia la probabilidad de mutaciones puede ser del 5% o del 10%. Es importante tener en cuenta que en cada iteración se elige el mejor cromosoma de la mitad que se había seleccionado para generar la futura descendencia, y se guarda sin ser modificado, para asegurar que los individuos de la generación posterior sean iguales o mejores que los de la presente. Esta característica del funcionamiento del algoritmo se conoce como "conservación de élite".

El ciclo anterior se repite durante una determinada cantidad de generaciones. Luego, como ya se describió para el caso del algoritmo ACO, se aplica una metodología de tipo Monte Carlo. Si un determinado preprocesamiento se selecciona más veces de las que se rechaza a través de los distintos ciclos de Monte Carlo, llevando a errores de predicción significativamente más bajos, se considera como útil para el juego particular que se está estudiando y se incluye en el modelo final PLS.

La **Figura 2.4** muestra un diagrama de flujo con la fracción del AG original que fue integrada a ACOGASS.

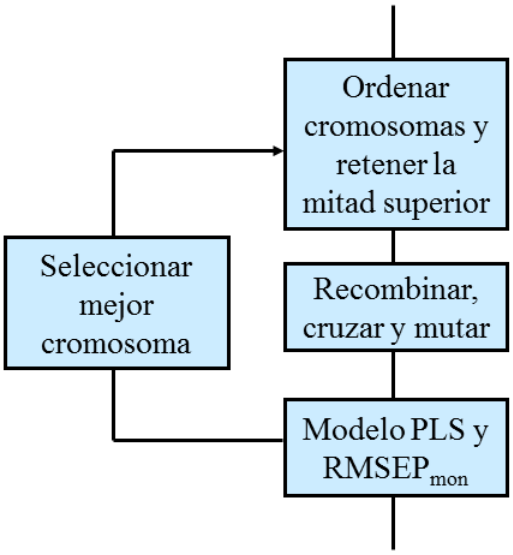


Figura 2.4. Diagrama de flujo de la sección del algoritmo genético que se utilizó para seleccionar la combinación óptima de preprocesamientos en ACOGASS.

Selección de muestras a través de métodos basados en distancias

La selección de muestras durante la optimización de un modelo puede realizarse utilizando varios métodos como aquellos basados en intercambios,[86] proyecciones sucesivas,[87] o distancia entre muestras.[88,89] Todos estos métodos son normalmente muy efectivos y permiten obtener conjuntos de muestras representativos.

En la estrategia integrada ACOGASS que se propuso en este trabajo, se implementaron dos métodos basados en distancias: Kennard-Stone (KS)[88] y partición de muestras basada en la distancia conjunta *X*-*Y* (SPXY).[89] La diferencia fundamental entre estos es que KS opera calculando distancias entre muestras en un espacio de variables latentes (generadas ya sea a través de PCR o de PLS), mientras que SPXY lo hace empleando la matriz de datos instrumentales y el vector de valores de referencia simultáneamente y sin modificarlos.

Detección de *outliers*

El criterio utilizado para detectar *outliers* fue la comparación del cociente estadístico *F* con valores críticos, tanto para muestras de entrenamiento como de monitoreo.[24] El valor experimental de *F* podría estar basado tanto en concentraciones como en residuos espectrales, y se computa como el cociente entre el error cuadrado para una muestra particular y el error cuadrado medio para el resto de las muestras. En este informe,

se utilizaron los residuos de las concentraciones para detectar *outliers* debido a que: (1) se conocen las concentraciones nominales para las muestras de entrenamiento y monitoreo y (2) el objetivo del algoritmo es generar un modelo cuya principal ventaja sea una habilidad predictiva mejorada.

2.6 Configuración de los parámetros del algoritmo

En las calibraciones PLS normalmente se utilizan dos juegos de datos: un conjunto de calibración, normalmente empleado para construir el modelo de regresión, y un conjunto de validación o *test* , a partir del cual se calcula un $RMSEP_{test}$ para controlar la capacidad de predicción del modelo PLS luego que los parámetros de calibración han sido optimizados. Para optimizar el modelo, por otro lado, los datos de calibración se dividen en un grupo de entrenamiento y otro de monitoreo. El propósito del conjunto de monitoreo es el de guiar las selecciones durante la fase de optimización. En estos tres conjuntos de datos (entrenamiento, monitoreo y validación), los valores de referencia (ya sean concentraciones del analito o propiedades de referencia) deben conocerse. Cuando se lleva a cabo la selección de muestras, el juego de entrenamiento y el de monitoreo se unen formando un único juego, que luego se divide nuevamente de acuerdo con el método de selección de muestras empleado y con las variables y preprocesamientos que se hayan empleado previamente. Como se hizo referencia, se implementaron dos posibles estrategias para llevar adelante esta última tarea: (1) el algoritmo de Kennard-Stone basado en *scores* obtenidos por PLS o por Análisis por Componentes Principales (PCA)[90] y (2) selección basada en distancias X-Y.[89] Por otro lado, si no se proporciona ningún juego de monitoreo, el conjunto completo se divide aleatoriamente para crear uno.

Si el cociente F_i para la *i*-ésima muestra excede el valor crítico los *outliers* son marcados.[89] En el caso de las muestras de calibración, F_i está dado por:

$$F_i = \frac{(I\text{-}1)(\hat{y}_{\text{cal},i} - y_i)^2}{\sum_{i' \neq i}(\hat{y}_{\text{cal},i'} - y_{i'})} \qquad (2.4)$$

donde y_i es la concentración nominal de la muestra *i*, $\hat{y}_{\text{cal},i}$ es el correspondiente valor de concentración de calibración estimado por el modelo de regresión e *I* es el número de muestras de calibración. En el caso de las muestras de monitoreo, se computa el siguiente radio: [86]

$$F_i = \frac{(I\text{-}1)(\hat{y}_{\text{mon},i} - y_i)^2}{\sum_{i'=1}^{I}(\hat{y}_{\text{cal},i'} - y_{i'})} \quad (2.5)$$

donde *i'* corresponde a las muestras de calibración e *i* a las de monitoreo.

En lo que respecta a la selección de métodos de preprocesamiento matemático, el algoritmo utiliza un AG adaptado para elegir entre uno o más preprocesamientos entre los siguientes: (1) corrección por *scattering* multiplicativo (MSC) [78] (2) SNV [79] (3) *detrend,* (4) primera derivada (5) segunda derivada (en los últimos dos casos las derivadas se computaron empleando la aproximación de Savitzky-Golay).[91] Estas cuatro metodologías son las que se usan con más frecuencia en aplicaciones de NIR/PLS.[74] La implementación del algoritmo genético requiere que se fije el número de cromosomas y de generaciones. Es importante aclarar que el centrado se aplicó a todos los conjuntos de datos como un preprocesamiento por *default*, como se hace normalmente en la mayoría de las aplicaciones NIR/PLS.

Finalmente, la actividad más importante es la selección de variables relevantes (longitudes de onda en el caso de estudios NIR/PLS). Como ya se mencionó, esto se realizó utilizando una estrategia de selección por colonias de hormigas, debido al éxito de esta última técnica en aplicaciones relacionadas.[46] La implementación de ACO requiere la elección de un determinado número de hormigas, que como ya se explicó con anterioridad son los entes artificiales de selección, así como también del número de generaciones o épocas de evolución durante las cuales las hormigas buscarán la mejor combinación de variables. Casualmente, en la metodología propuesta, el número de épocas en ACO es igual a la cantidad de generaciones en AG.

Un factor importante con el que se debe ser cuidadoso a la hora de modificarlo, es el ancho de la ventana, es decir, el número de sensores individuales a ser incluidos en cada uno de los bloques de sensores o variables a ser seleccionadas. La ventana elegida debería reflejar el ancho típico de una banda espectral. Por ejemplo, si una banda típica tiene un ancho de 50 nm, y el espectro se registra en intervalos de 2 nm, un valor razonable para el ancho de ventana es 25 (ancho de banda/intervalo de medición del equipo utilizado). Durante la ejecución del algoritmo, se varía el número de variables seleccionadas en un rango determinado (es decir, entre un mínimo y un máximo, ambos seleccionados por el usuario).

El parámetro que guía la selección de muestras y de métodos de preprocesamiento realizados por ACO y GA es el $RMSEP_{mon}$. Es por esto que un parámetro importante en este sentido es el número de factores de PLS utilizados para construir los modelos en cada uno de los pasos del algoritmo. En primera instancia, al igual que para ACO, se debe estimar un valor inicial que el propio usuario introduce en ACOGASS por medio de una metodología conocida como validación cruzada utilizando una muestra por vez. Esta validación se realiza sobre los datos crudos, es decir, con los espectros completos y sin preprocesamiento.[86] De cualquier forma, durante las iteraciones del programa, el número de variables latentes se readapta en cada paso examinando los valores de $RMSEP_{mon}$ en función del número de factores PLS, y seleccionando el aquel número para el cual no se dan mayores cambios en el valor de $RMSEP_{mon}$. No se utiliza la validación cruzada de una muestra por vez en cada paso, debido a que incrementa significativamente el tiempo de operación del algoritmo.

El diagrama de flujo que se presenta en la **Figura 2.5**, resume adecuadamente los pasos mencionados previamente. Como puede observarse, todas las actividades anteriores se repiten un cierto número de veces permitiendo obtener resultados más confiables a través de una metodología de tipo Monte Carlo.[88] Como es usual, una vez terminado el proceso de selección, se construye un histograma reflejando la frecuencia relativa de selección de cada variable. Aquellas que se encuentran por encima de un cierto nivel de tolerancia, son finalmente elegidas para construir el modelo PLS utilizando las muestras de entrenamiento y los preprocesamientos seleccionados. El modelo óptimo puede aplicarse, en caso de ser necesario, a las muestras de *test* para chequear la habilidad predictiva.

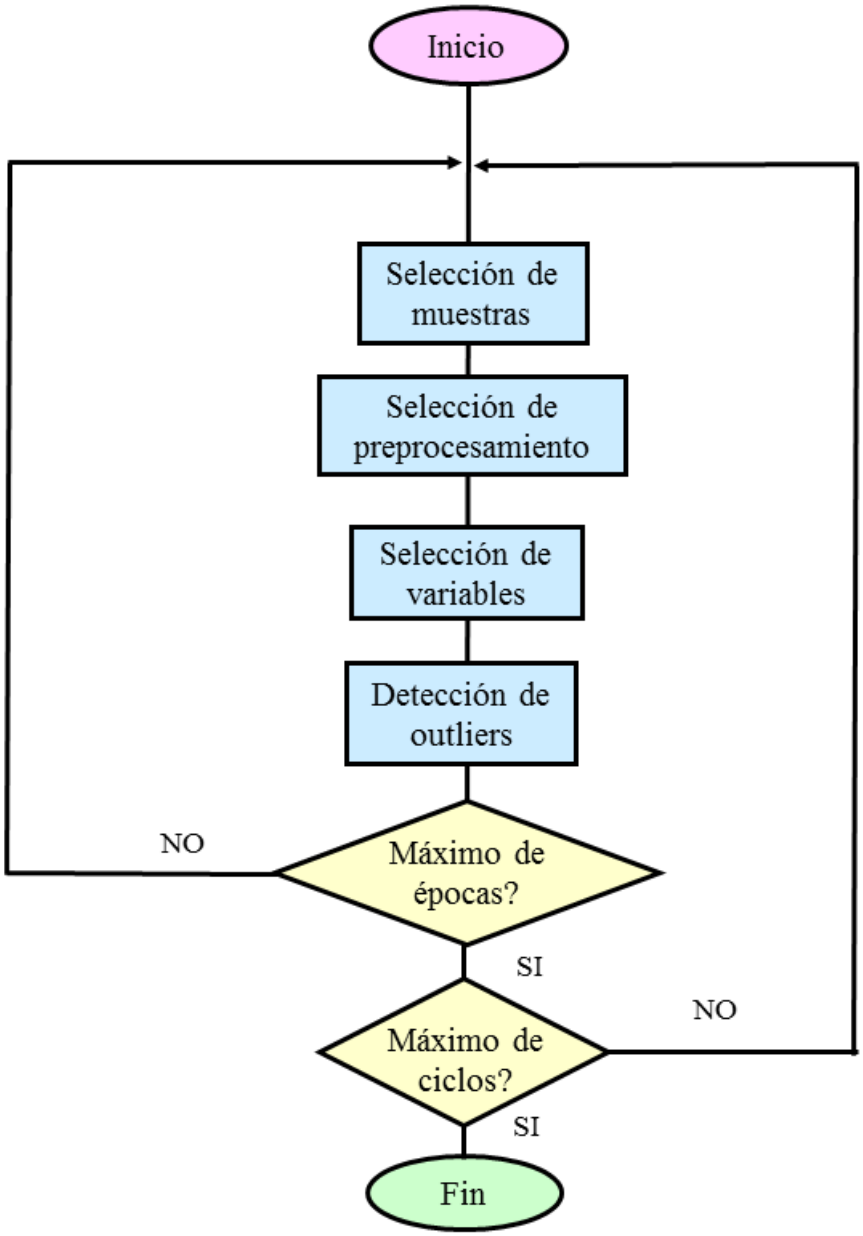


Figura 2.5 Diagrama de flujo del algoritmo ACOGASS que implementa simultáneamente selección de muestras, preprocesamiento, selección de variables y detección de variables en una estrategia de tipo Monte Carlo.

Resulta importante realizar una aclaración final respecto de las actividades que se describieron. Es probable que un usuario experimentado en NIR/PLS sea capaz de eliminar los rangos de longitudes de onda que no aportan información relevante, mediante una simple inspección visual del espectro (como pueden ser en los casos de regiones de saturación de señal o de alto nivel de ruido), como así también de determinar qué preprocesamiento es el más adecuado para un determinado conjunto de datos si el material

que se está analizando es sólido o semisólido. Estas formas intuitivas de selección de variables y preprocesamientos podrían mejorar la *performance* en la predicción de los modelos PLS. Sin embargo, la intención de este trabajo fue la de desarrollar una metodología completamente automatizada, que pueda ser incorporada a un *software* NIR/PLS en el futuro, y operada por cualquier operario, aunque no esté completamente capacitado en materia de calibraciones.

2.7 Datos simulados

Se construyó un juego de datos sintéticos, simulando el espectro de tres componentes y un *background* no lineal variable con cada muestra, siendo el componente 1 el analito de interés. Todos los constituyentes estuvieron presentes en 70 muestras de entrenamiento, 30 muestras de monitoreo y 100 muestras de *test*, en concentraciones aleatorias variando entre 0 y 1 para los constituyentes 1 y 2, y entre 5 y 10 para el componente 3 (en este caso el objetivo fue el de asegurar altas concentraciones relativas de este último componente). La **Figura 2.6 A** muestra los espectros de los componentes puros a concentraciones unitarias, así como una señal de *background* típica, definida en un rango espectral de 100 sensores. A partir de estos perfiles sin adición de ruido, se construyeron muestras de entrenamiento, *test* y monitoreo. Específicamente, cada espectro **x**, ya sea de entrenamiento, monitoreo y *test* fue creado utilizando la siguiente expresión:

$$\mathbf{x}=y_1\mathbf{s}_1+y_2\mathbf{s}_2+y_3\mathbf{s}_3+\mathbf{bk} \qquad (2.6)$$

donde , y son los espectros de los componentes puros a concentraciones unitarias, y_1, y_2 e y_3 son las concentraciones de los componentes en una muestra específica y **bk** es la señal de *background*. Se adicionó ruido Gaussiano con una desviación estándar de 0.01 unidades a todas las concentraciones, antes de insertar los valores en la **Ecuación 2.6**. Seguidamente se adicionó también un vector de ruido en señal a cada **x** luego de aplicar la **Ecuación 2.6**. Las señales mayores a 5 unidades se cortaron en este último valor, adicionándose ruido con una unidad de desviación estándar (esto simula la saturación del detector a altas absorbancias en un experimento real). En la **Figura 2.6** se grafica la matriz resultante para las señales de entrenamiento. Se pueden observar las variaciones y la naturaleza no lineal de la señal de *background* adicionada, que hace necesaria la aplicación de métodos de preprocesamiento para eliminar este efecto.

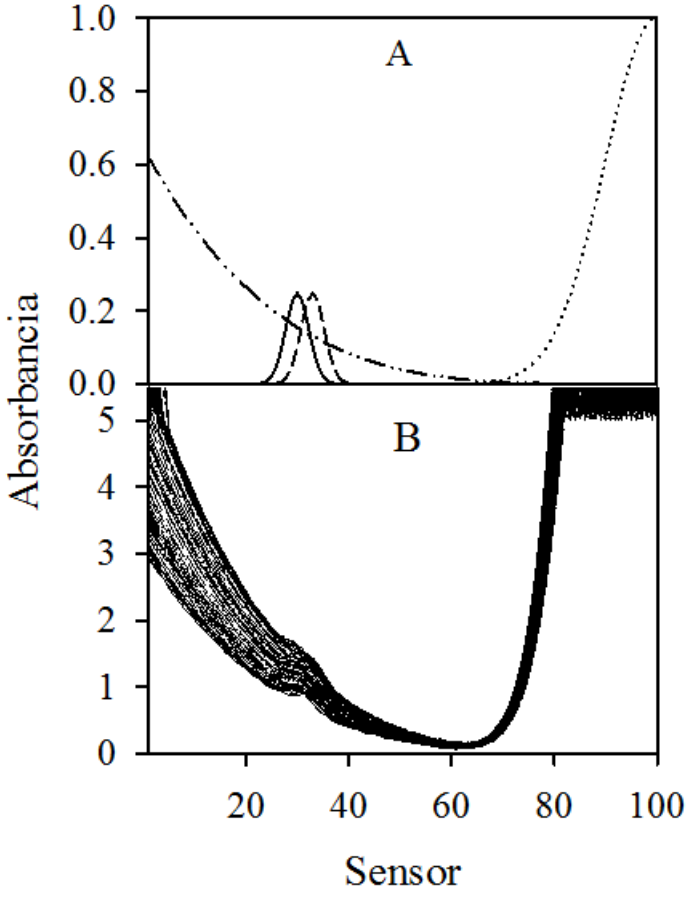


Figura 2.6. (A) Gráfico de los espectros de los componentes puros (analito 1, línea sólida, componente 2, línea rayada, componente 3, línea de puntos) y señal de fondo (línea de puntos y rayas), utilizados para construir el conjunto de datos simulados. (B) Gráfico de los 70 espectros simulados. Los espectros para las muestras de *test* y de monitoreo son similares.

2.8 Datos experimentales

Muestras de jugo de caña de azúcar (conjunto BRIX)

Para construir este juego de datos, se midieron espectros NIR para una serie de muestras de jugos de caña de azúcar utilizando un espectrofotómetro FOSS NIRSystems 6500, equipado con una celda de 1 mm de paso óptico. Los espectros se obtuvieron por medio del software ISISCAN del espectrofotómetro y luego se convirtieron al formato correspondiente (ASCII) para poder procesarlos. Los datos de referencia, en grados Brix, se midieron con un refractómetro Leica AR600. Los jugos de caña de azúcar se analizaron en la estación Obispo Colombres, Tucumán, Argentina. Este laboratorio recibe muestras de diferentes productores azucareros de la provincia de Tucumán. Las muestras de caña son

primeramente procesadas en los molinos azucareros, donde se extrae el jugo (65 % de la caña) y luego son enviadas al laboratorio. Para el juego de calibración, se seleccionaron 59 muestras de manera aleatoria, con valores de grados Brix entre 11.76 y 23.15. El juego de monitoreo estuvo compuesto de 23 muestras y el de *test* por otras 23 muestras con valores diferentes a aquellos empleados en la calibración. La medición de los espectros se realizó en un rango de longitudes de onda comprendido entre 400 y 2498 nm cada 2 nm (es decir, 1050 puntos).

Muestras de maíz (conjunto CORN)

Este es un conjunto de datos de acceso libre medidos por la empresa Cargill (http://www.eigenvector.com/data/Corn), que consiste en los espectros NIR de 80 muestras de maíz medidas en un rango de longitudes de onda que va desde 1100 a 2498 nm en intervalos de 2 nm (700 canales). Para este conjunto de espectros se midieron varios parámetros de referencia entre los cuales se ha seleccionado el contenido de almidón, con valores entre 62.83 y 66.47.

2.9 Software

El algoritmo integrado descripto, se implementó en forma de interface gráfica que puede utilizarse a través de la versión de MATLAB 7.4.0 (R2007a) o superiores. Para detalles acerca del instructivo de uso se puede recurrir al archivo 'ACOGASS_manual.pdf', provisto por el software. Los códigos de MATLAB, el manual y el conjunto de datos con el ejemplo simulado discutido durante este capítulo, se pueden descargar libremente de www.iquir-conicet.gov.ar/descargas/acogass.zip.

2.10 Resultados

2.10.1 Simulaciones

En este juego de datos, como ya se explicó durante su descripción, aparecen tres componentes, siendo uno de ellos el analito de interés, junto con una señal adicional de *background*. Uno de los componentes genera una señal intensa con saturación en los sensores 80-100, mientras que también aparece una señal no lineal y dependiente de cada muestra en los sensores 1-50 (Figura 2.7 A). Lo que se esperaría de la estrategia ACOGASS sería la obtención de valores razonablemente bajos de RMSEP (tanto para los juegos de monitoreo como los de *test*), seleccionando aquellas regiones espectrales

aparentemente útiles ubicadas en los sensores 25-40, aplicando un método de preprocesamiento adecuado para aliviar el efecto del *background* variable no lineal, y optimizar el número de variables latentes PLS a 2 ó 3 como máximo.

El algoritmo ACOGASS se aplicó sobre este conjunto de datos utilizando los parámetros que se muestran en la **Tabla 2.1**. Nótese que cada variable comprende dos sensores individuales, lo que significaría aproximadamente la mitad del ancho de banda de los picos de los analitos individuales (**Figura 2.7 A**). Inicialmente se fijó el número de variables latentes en 4 (**Tabla 2.1**), debido a que existen cuatro fuentes de variación activas en este conjunto de datos.

Tabla 2.1. Valores específicos de los parámetros.

Parámetro	Simulados	BRIX	CORN
Número de hormigas	20	20	20
Proporción de hormigas ciegas [a]	0.3	0.3	0.3
Número mínimo de variables	4	4	4
Número máximo de variables	8	8	8
Número de cromosomas	20	20	20
Frecuencia de mutaciones [a]	0.1	0.1	0.1
Ciclos	10	10	10
Épocas	50	50	50
Ventana de sensores	2	20	20
Tolerancia	0.3	0.3	0.3
Variables latentes [b]	4	12	17

[a] La proporción de hormigas "ciegas" y la frecuencia de mutaciones son parámetros que introducen aleatoriedad en la búsqueda del error de monitoreo mínimo.

[b] Estimado utilizando validación cruzada dejando de lado una muestra por vez sin aplicar preprocesamiento en el rango espectral completo.

De acuerdo con los resultados que se presentan en la **Tabla 2.2** para el caso de las cifras de mérito computadas para las muestras de validación (distintas de las de entrenamiento y monitoreo), se puede observar claramente que ACOGASS encuentra la respuesta adecuada. Por el contrario, cuando se utiliza el espectro completo sin preprocesamiento, se obtiene un error de predicción mayor. Mientras tanto, ACOGASS seleccionó el *detrend* como método de preprocesado, lo cual resulta razonable teniendo en cuenta que este método permite eliminar efectivamente las señales de *background* variable y no lineal. El número óptimo de variables latentes necesarios fue 2, como era de esperarse. El $RMSE_{test}$ de 0.03 unidades, calculado luego de la selección efectuada por ACOGASS, es también un resultado razonable. Para comparar los valores de RMSEP (obtenidos antes y después de la selección) se utilizó el test de aleatorización sugerido por van der Voet. El resultado de este *test* indica que el RMSEP obtenido mediante ACOGASS es significativamente menor que el que se obtiene sin selección, debido a que el valor de probabilidad (*p*) es significativamente menor que el nivel crítico de 0.05. Otros indicadores estadísticos que muestran la mejora del modelo son el error relativo de predicción REP%=5.7% (calculado con respecto al valor promedio del entrenamiento), y un coeficiente de correlación R^2=0.9900 (**Tabla 2.2**).

Tabla 2.2. Cifras de mérito y otros resultados de ACOGASS

	Simulados	***BRIX***	***CORN***
Espectro completo			
$RMSEP_{test}$	0.28	0.75	0.39
Nro. de variables latentes	4	12	17
Preprocesamiento	Ninguno	Ninguno	Ninguno
Outliers en calibración	1	0	0
Outliers en *test*	0	0	0
Luego de la selección con ACOGASS			
$RMSEP_{test}$	0.03	0.25	0.11
Nro. de variables latentes	2	9	14
Preprocesamiento	*Detrend*	Ninguno	*MSC*
Outliers en calibración	0	0	0
Outliers en *test*	0	0	0

Comparando con los resultados obtenidos utilizando el espectro completo, la mejora en la habilidad predictiva utilizando selección de variables y de preprocesamientos resultó ser muy significativa (**Figura 2.7 B**).

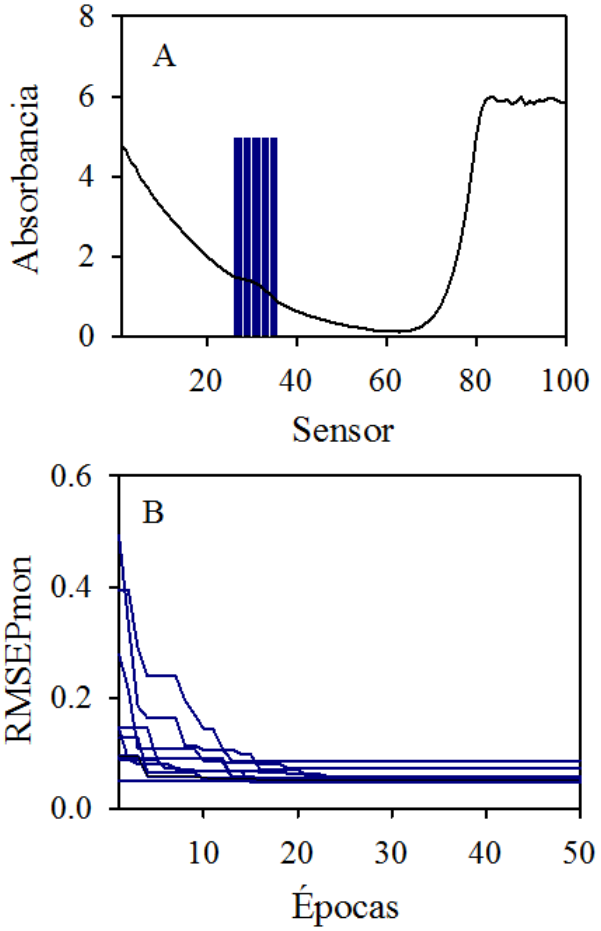


Figura 2.7. (A) Barras azules mostrando las variables seleccionadas (bloques de sensores) en el juego de datos simulados. La línea sólida negra corresponde al espectro medio calculado a partir de las muestras de entrenamiento. (B) Evolución del error de monitoreo ($RMSEP_{mon}$) como función de las épocas en el juego de datos simulados.

2.10.2 Datos experimentales BRIX

Las principales características espectrales del juego de datos Brix involucran una zona de señal de alta absorbancia debido al agua (alrededor de 1950 nm), regiones con señales significativas entre 1450 y 2500 nm, y otra en la que prevalece el ruido por debajo de los 1300 nm. Como se describió anteriormente, el juego de 105 muestras se dividió aleatoriamente en entrenamiento, monitoreo y *test* con 59, 23 y 23 muestras respectivamente. En la validación cruzada utilizando el espectro completo, se requirieron 12 variables latentes PLS, que luego fueron utilizadas como el máximo número de factores en ACOGASS (**Tabla 2.1**). Debido a que el valor de la ventana de sensores utilizado fue 20, el mínimo número de sensores seleccionables fue de 40 nm, ya que el intervalo de

medición en este caso es de 2 nm. Esto es razonable si se tiene en cuenta el ancho de las bandas espectrales a media altura (**Figura 2.8 A**). El resto de los parámetros ACOGASS se muestran en la **Tabla 2.1**.

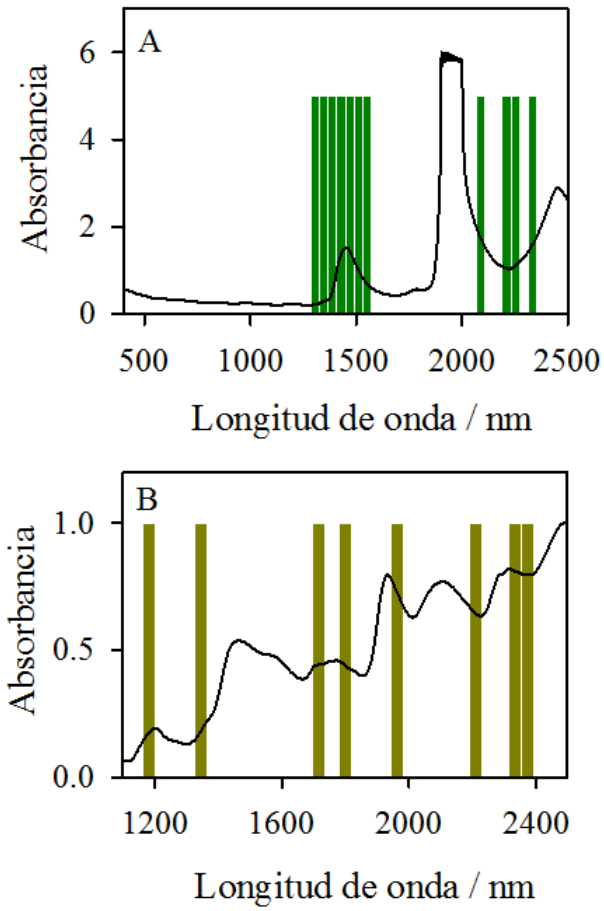


Figura 2.8. (A) Variables seleccionadas (bloques de sensores) en el conjunto de datos BRIX. La línea negra sólida corresponde al espectro medio calculado a partir de las muestras de entrenamiento. (B) Lo mismo que (A) para el juego de datos CORN.

Como puede verse en la **Tabla 2.2**, las cifras de mérito obtenidas muestran una mejora considerable en la predicción luego de que se seleccionan las regiones espectrales indicadas en la **Figura 2.8 A**. El RMSEP desciende significativamente desde 0.75 a 0.25 unidades Brix en comparación con el valor obtenido al aplicar el proceso de selección. Esto corresponde a una disminución del error relativo de predicción porcentual (REP%) de un 4.2% a 1.4%. Esta mejora mostró ser significativa luego de aplicar el *test* de aleatorización para comparar RMSEPs (es decir $p<0.05$, ver **Tabla 2.2**).

Otra observación importante es que el número óptimo de variables latentes para ACOGASS es menor que cuando se aplica el modelo con el espectro completo, lo cual era de esperarse debido a la reducción de regiones espectrales utilizadas durante el entrenamiento y la eliminación de características espectrales que no están correlacionadas con los valores de referencia Brix. Además, a pesar que se probaron muchas combinaciones de métodos de preprocesamiento, ninguna fue elegida. Este hecho se encuentra en concordancia con las características de estas muestras que, al ser líquidas y utilizarse el modo de medición por transmitancia, en principio no deberían ser afectadas por el fenómeno de *scattering* que normalmente causa desviaciones en la línea de base.

2.10.3 Datos experimentales CORN

Este juego de datos se encuentra disponible en internet y tiene como objetivo la calibración de almidón y otros parámetros de relevancia en semillas de maíz. El juego de datos constituido por 80 muestras se dividió, al igual que en los casos anteriores, en entrenamiento (40 muestras), monitoreo (20 muestras) y *test* (20 muestras). En lo que respecta a la determinación de almidón, la validación cruzada indicó una cantidad óptima de 17 factores al utilizarse el rango espectral completo. Este número se redujo significativamente luego de la selección de variables, con la correspondiente mejora en las cifras de mérito (**Tabla 2.2**). La **Figura 2.8 B** muestra las regiones seleccionadas por ACOGASS cuando se utilizan los parámetros indicados en la **Tabla 2.1**. Al igual que en el caso de los datos BRIX, la reducción en el RMSE también fue significativa (p<0.05), de 0.23 a 0.11, lo cual corresponde a valores de REP% de 0.60 y 0.17 respectivamente.

La selección de MSC como único preprocesamiento es razonable si se tiene en cuenta que se trata de una muestra sólida (maíz molido) medida por reflectancia, y por lo tanto es de esperar que exista una fuerte dispersión de la radiación llevando a efectos de *scattering*.

Si se procesa el espectro completo aplicando MSC, se llega a un modelo PLS con 14 variables latentes y con un RMSEP de 0.21 para el juego de *test*. Esto implica una cierta mejora respecto del valor que aparece en la **Tabla 2.2**, aunque representa un rsultado subóptimo si se compara con el obtenido a través de ACOGASS.

2.11 Conclusión

Durante este capítulo, se investigó una nueva estrategia para implementar de manera combinada de tres de los principales formas de optimización de PLS: selección de variables, preprocesamientos y muestras. Esta estrategia está basada en una metodología de tipo Monte Carlo que integra el algoritmo ACO para selección de variables, AG para seleccionar métodos de preprocesamiento, y dos de los métodos de selección de muestras más ampliamente difundidos. El algoritmo fue evaluado utilizando juegos de muestras NIRS de distinta naturaleza y los resultados fueron satisfactorios. Todas estas características constituyen una metodología innovadora basada en el uso de métodos combinados para obtener una calibración PLS completamente optimizada.

2.12 Perspectivas

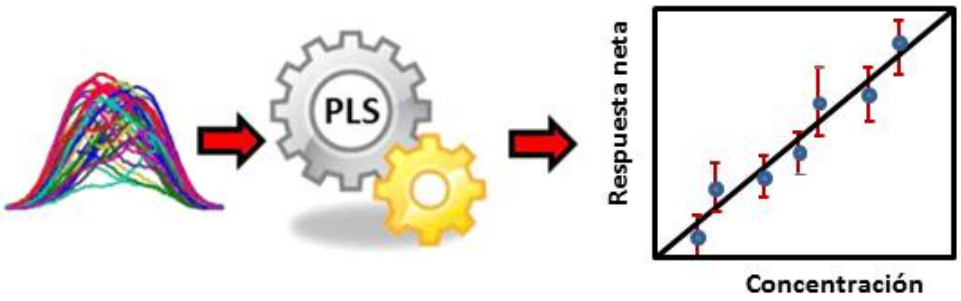
Si bien la implementación del algoritmo ACOGASS implica un paso más en la automatización del proceso de optimización del algoritmo PLS, tal como se describió previamente, existe un grupo de parámetros que el usuario debe fijar "manualmente" (ver **Tabla 2.1**) y que requieren de un cierto conocimiento acerca del funcionamiento de las estrategias de selección y del tipo de datos analizados. Entre estos parámetros, los que más influencia tienen en la optimización son el número mínimo y máximo de variables a seleccionar por cada hormiga y el ancho espectral que incluye cada una de estas variables. De esta manera, el siguiente paso hacia la automatización completa de la optimización de PLS por medio de algoritmos estocásticos podría ser la de utilizar los mismos recursos del algoritmo para autoajustar sus metaparámetros (parámetros que se definen antes de inicializar el algoritmo). Es decir, los posibles números de variables, así como el ancho espectral que abarca cada variable, podrían incluirse como elementos de un vector feromona (en el caso de ACO) cuyo valor óptimo se vaya actualizando de acuerdo a los resultados de la función objetivo que se van obteniendo en las sucesivas iteraciones, así como lo hacen los preprocesamientos y las distintas combinaciones de variables que se van evaluando.

Otra posibilidad de modificación del algoritmo ACOGASS, particularmente en lo que respecta a la selección de sensores espectrales, consistiría en utilizar una versión reversa o "*backward*" de ACO. Es decir, en lugar que las variables seleccionadas por cada hormiga sean las que se incluyen en el modelo, hacer que estas sean las que se eliminan. Es

decir, las posiciones del vector feromona con mayores valores pasarían a ser las posiciones que más afectan negativamente la calidad del modelo y de las predicciones. Si bien en principio se podría intuir que esta forma de proceder podría conducir a resultados similares que la propuesta original, la filosofía de funcionamiento es opuesta, y hasta el momento no hay estudios que demuestren una equivalencia en los resultados.

CAPÍTULO 3

ESQUEMA GENERALIZADO PARA EL CÁLCULO DEL ERROR ESTÁNDAR DE PREDICCIÓN EN CALIBRACIÓN MULTIVARIADA



"Denme siempre el fructífero error, lleno de semillas, desbordado por sus propias correcciones. Ustedes pueden mantener su verdad, estéril por sí misma." (Vilfredo Pareto).

3.1 Resumen

La mayoría de las expresiones que se utilizan actualmente para calcular cifras de mérito en calibración multivariada suponen que el error instrumental tanto en la muestra incógnita como en las de calibrado, se encuentra distribuido de manera idéntica e independiente (ruido iid). Sin embargo, se sabe que esta condición no siempre es alcanzada por sistemas experimentales reales, donde la existencia de numerosos factores externos puede llevar a estructuras de ruido heteroscedásticas y/o correlacionadas. En este tercer capítulo de tesis, se analiza la influencia de las desviaciones del paradigma iid clásico en un contexto de propagación de errores basado en los principios de la teoría EIV. Se presentarán nuevas expresiones derivadas con el objetivo de calcular el error estándar de predicción bajo diversos escenarios. Estas expresiones permiten un estudio cuantitativo de la influencia de las diferentes fuentes de error instrumental que afectan al sistema que se está analizando. Como se mostrará más adelante, se observan diferencias significativas cuando el error de predicción se estima en cada uno de los escenarios estudiados,

utilizando los algoritmos más populares para análisis de datos multivariados de primer orden, tanto en condiciones simuladas como experimentales.

3.2 Introducción

A pesar del uso expandido en química analítica de los modelos de regresión basados en variables latentes (PLS y PCR), una característica importante, aunque hasta el momento ignorada, es que estos modelos se diseñaron para trabajar de manera óptima cuando los errores de medición se encuentran distribuidos de manera idéntica e independiente (iid). La misma situación se presenta a la hora de tener en cuenta la estimación de las cifras de mérito analíticas, la mayoría de las cuales se definieron bajo el mismo supuesto de ruido. [92-98]

Las desviaciones de la condición iid se pueden modelar adecuadamente a través de métodos tradicionales de calibración, a pesar de que esto lleva a mayores errores de predicción.[33] Cuando la estructura del error se desvía significativamente de la situación ideal, se requieren acciones específicas. En este sentido hay dos alternativas. Una es la de aplicar un método de preprocesamiento acorde, que modifique la estructura del error para aproximarla al caso iid (es decir, "sintonizar" el error con el modelo). Sin embargo, esta manera de proceder sólo es posible para ciertas estructuras de error con el riesgo que el preprocesamiento aplicado lleve a resultados subóptimos si se aplica en un contexto erróneo.[33] La segunda alternativa es la de utilizar algoritmos basados en principios de máxima probabilidad (ML) como MLPCR "sintonizando" el modelo a la estructura del error). Esta última alternativa requiere la estimación de la matriz de covariancia del error asociada a la estructura del ruido, ya sea por medio de replicado o por técnicas específicas de modelado empírico.[32,33]

Sin embargo, la incertidumbre en la predicción y otras cifras de mérito relevantes que dependen de aquella, son afectadas significativamente por la estructura del ruido, como se mostrará más adelante. Algunas razones importantes por las cuales es importante profundizar en los estudios en esta temática son: (1) todos los procedimientos de validación requieren, como buena práctica analítica, informar los resultados junto con una estimación confiable de su incertidumbre,[12] y (2) la estimación de la incertidumbre es un paso fundamental en el cálculo de otras cifras de mérito importantes como el límite de detección.[17] Incluso cuando el análisis de replicados de muestras permitiría una estimación

experimental de la incertidumbre en la predicción, estudios como los que se presentarán en este capítulo permiten una mayor comprensión de las distintas fuentes de error que afectan a esta incertidumbre así como la proporción en que lo hacen. Esto es importante teniendo en cuenta la optimización de métodos para mejorar la precisión, lo cual podría alcanzarse incluso en ausencia de réplicas.[29]

La matriz de covariancia del error cumple un rol central en la propagación de errores para estimar la incertidumbre en calibración multivariada de primer orden. Sin embargo, aunque se la mencionó durante el desarrollo de expresiones basadas en la suposición iid, en ningún caso se indagó más profundamente sobre su rol en casos en los que esta situación no se cumple.[99] Por otro lado, Wentzell y colaboradores resaltaron en varias publicaciones la necesidad de estimar la estructura del ruido en datos multivariados, proponiendo y probando diversas estrategias para modelar la matriz variancia-covariancia.[29] Aún en ausencia de réplicas, el ruido heteroscedástico puede caracterizarse por medio de filtros digitales.[31] Este es un paso importante en la identificación de juegos de datos que no cumplen con la condición iid, pero no permite arribar a conclusiones certeras cuando los errores están correlacionados. Aunque estas dos líneas de trabajo (la estimación de la incertidumbre en la predicción y de la matriz de covariancia del error) son complementarias, hasta el momento no se han realizado mayores esfuerzos en combinarlas.

En este capítulo se presentará un esquema general para estimar la incertidumbre en la predicción específica para cada muestra cuando se calibra utilizando datos multivariados de primer orden. Este esquema está basado en una metodología de propagación de errores, y requiere de una estimación adecuada de la matriz de covariancia del error que es la que caracteriza a la estructura de error multivariado. Como se mostrará más adelante, se pueden describir tres posibles situaciones, dependiendo del tipo de ruido que presenten las muestras analizadas. Teniendo en cuenta este panorama, se presentarán expresiones que luego se contrastarán y validarán por medio de estudios de adición de ruido. La estrategia presentada se desarrolló y evaluó en modelos de calibración multivariada tradicionales como son PCR y PLS. Un supuesto importante que subyace a esta investigación, es que la naturaleza del ruido no afecta seriamente al error de predicción hasta el punto de requerir la utilización de estrategias de máxima probabilidad para procesar los datos. En cualquier caso, las expresiones propuestas tienen el potencial de poder extenderse a otros modelos multivariados sin pérdida de generalidad.

3.3 Objetivos específicos

1) Investigar los posibles escenarios que pueden presentarse cuando la estructura de error de las muestras analizadas difiere del supuesto iid clásico.

2) Diseñar un esquema de fórmulas para calcular el desvío estándar de predicción por muestra, teniendo en cuenta los posibles escenarios planteados.

3.4 Desviaciones del comportamiento iid y tipos de errores multivariados

El ruido en las señales instrumentales se puede clasificar de acuerdo con numerosos criterios siendo los más comunes: (1) la fuente de la cual proviene, (2) la distribución que adopta, (3) las características en el dominio de frecuencias, y (4) las características en el dominio del tiempo. Desafortunadamente, las clasificaciones basadas en estos métodos no son mutuamente exclusivas. Sin embargo, en química analítica, y más específicamente en calibración multivariada una manera relativamente simple de clasificarlos es de acuerdo a si están distribuidos de manera idéntica e independiente (iid) o no.

La denominación "ruido iid" contiene una gran cantidad de información. El concepto de independencia en lo que respecta a los errores de medición, implica que el error observado para un determinado sensor no se encuentra relacionado con el error de otro sensor diferente. La independencia en los errores de medición también implica que los errores de medición no están correlacionados. Idénticamente distribuido, por su parte, hace referencia a la homogeneidad en la variancia del error, considerada a través de todos los canales que conforman el vector señal; es decir, la variancia del error en todos los canales del vector de señal es la misma. Los términos homoscedástico y heteroscedástico también se utilizan para indicar si los errores de medición se encuentran distribuidos idénticamente o no. Finalmente, es importante aclarar que la denominación iid lleva implícita la condición de normal, en el sentido que se supone una distribución normal del ruido para muchas mediciones repetidas. Por lo tanto, se dice que los errores son iid si cumplen con todas las condiciones mencionadas anteriormente, y no iid si alguna de las condiciones no se respeta.

En la **Tabla 3.1** se muestra una lista con algunos tipos de ruidos y su descripción de acuerdo con el criterio de clasificación utilizado. Como se detallará más adelante, en este

trabajo se utilizaron dos tipos de ruidos que constituyen ejemplos típicos de dos desviaciones diferentes del supuesto iid: (1) el ruido llamado "rosa", un tipo de ruido parcialmente correlacionado y que por lo tanto se desvía del supuesto de independencia, y (2) un ruido cuya variancia es proporcional a la señal y en consecuencia constituye una desviación del supuesto de idéntico.

Tabla 3.1. Algunas clasificaciones de ruidos de medición experimentales (se puede agregar una columna mostrando la ecuación para generarlo)

Tipo de error	Descripción
ERRORES INDEPENDIENTES / NO CORRELACIONADO	Errores para los cuales la covariancia del error es cero
ERRORES HOMOSCEDÁSTICOS	Errores que tienen la misma variancia (variancia uniforme)
Ruido blanco	Errores de medición no correlacionados. Puede implicar error homoscedástico con distribución normal.
ERRORES CORRELACIONADOS	Errores para los cuales la covariancia del error no es cero
Ruido aditivo de *offset*	Tipo de ruido correlacionado que desplaza de manera aleatoria la señal completa hacia arriba o hacia abajo una cantidad fija (es decir, desplaza la posición de la línea de base)
Ruido multiplicativo	Tipo de ruido correlacionado que desplaza la señal completa hacia arriba o abajo una cantidad proporcional a la magnitud de la señal. Normalmente asociado a un ruido de oscilación de la fuente.
Ruido rosa o ruido *1/f*	Tipo de ruido parcialmente correlacionado y de baja frecuencia, para el cual los errores de medición en medidas adyacentes están más correlacionados que las medidas que están muy alejadas. Caracterizado por una variación aleatoria que se da lentamente. También incluye lo que se conoce como ruido

	"marrón" ($1/f^2$)
ERRORES HETEROSCEDÁSTICOS	Errores con variancia distinta (variancia no uniforme)
Ruido proporcional	Tipo de ruido heteroscedástico en el que la desviación estándar del error es proporcional a la magnitud de la señal. Normalmente asociado al ruido de fluctuación de la fuente emisora.
Shot noise	Tipo de ruido heteroscedástico para el cual la desviación estándar de la señal es proporcional a la raíz cuadrada de la señal. Proviene de una distribución de Poisson y se asocia por ejemplo, al ruido del fotomultiplicador

3.5 La matriz de covariancia del error

Para comprender el concepto y la utilidad de la matriz de covariancia del error, es importante sentar la diferencia entre "error" e "incertidumbre".[30] Aunque frecuentemente se usen como sinónimos, la incertidumbre es una caracterización estadística del error en las medidas de las réplicas, que puede expresarse como una diferencia entre un valor medido y un valor real, y puede ser positiva como negativa.

Como muestra la **Figura 3.1**, la matriz de covariancia del error consiste en una matriz cuadrada simétrica que contiene, como elementos diagonales, la variancia del error asociada a cada canal de medición, y como elementos fuera de la diagonal, todas las covariancias entre los errores de medición en canales distintos. Esta matriz permite la visualización de la relación estadística entre los elementos de error de un predictor típico en calibración de primer orden, a diferencia de la calibración univariada donde se trabaja con escalares simples.[30] Respecto del análisis de la estructura del error, la diagonal de la matriz de covariancia brinda información acerca de la heteroscedasticidad del error, mientras que los elementos fuera de la diagonal describen la naturaleza del error correlacionado.

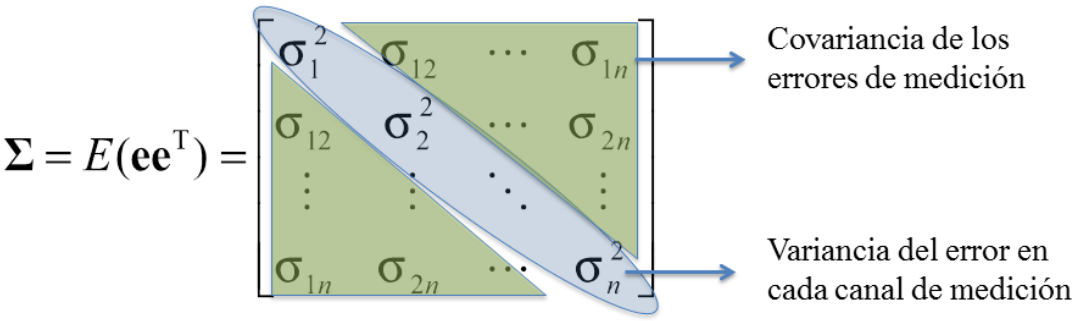


Figura 3.1. Representación esquemática general de la matriz variancia covariancia del error. Σ simboliza la matriz de covariancia del error, **e** el vector de error y *E*() se refiere a un valor esperado.

3.5.1 Estimación de la matriz de covariancia del error

Como se discutió en la sección anterior, la matriz de covariancia del error juega un rol fundamental a la hora de estimar la incertidumbre en la predicción en presencia de ruido no iid. Por este motivo, resulta interesante analizar de qué manera se puede estimar esta matriz. Hay tres alternativas principales: (1) réplicas experimentales, (2) predicción teórica y (3) modelado empírico.[29,30]

Las réplicas experimentales consisten en registrar *N* vectores $\mathbf{x}_j$ correspondientes a las distintas réplicas y luego calcular la matriz de covariancia del error como:

$$\boldsymbol{\Sigma}_x = \frac{1}{N-1}\sum_{j=1}^{N}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})^{\mathrm{T}} \qquad (3.1)$$

Hay dos puntos cruciales en este tipo de procedimientos. Uno es la forma en que se definen las réplicas. Por ejemplo, una réplica se puede referir a barridos repetidos, sin extraer la muestra del instrumento, extrayéndola o a muestras preparadas y medidas separadamente. Otro factor importante a considerar es el número de réplicas que se requiere para evitar una incertidumbre excesiva en los valores estimados de la variancia del error, es decir, para minimizar el "ruido en el ruido". Se conoce que para alcanzar un valor de desvío estándar razonable en la variancia, se requiere un número muy grande de replicados (más de 100).[30] Sin embargo, como este número no es realista en términos prácticos, la mejor opción consiste en colectar las estimaciones de las matrices de

covariancia del error obtenidas para un número razonable de replicados para cada muestra, y luego promediarlas.[29] Lo anterior sólo es posible en la medida en que los vectores medidos no cambien significativamente entre muestras, una situación típica en muchos instrumentos analíticos. Esta estrategia fue la que se utilizó en el caso de los datos experimentales que se procesaron posteriormente en este trabajo.

La predicción teórica de la matriz de covariancia del error también es posible para datos simulados y para ciertos datos experimentales en los que las fuentes de error están bien caracterizadas. Esta es la estrategia que se empleó para los datos simulados utilizados en este trabajo. La **Figura 3.2** muestra secuencias de error típicas y las correspondientes matrices de covariancia (presentadas como mapas de color) para cada una de las estructuras de error discutidas anteriormente (iid, correlacionada, heteroscedástica), así como sus correspondientes combinaciones.

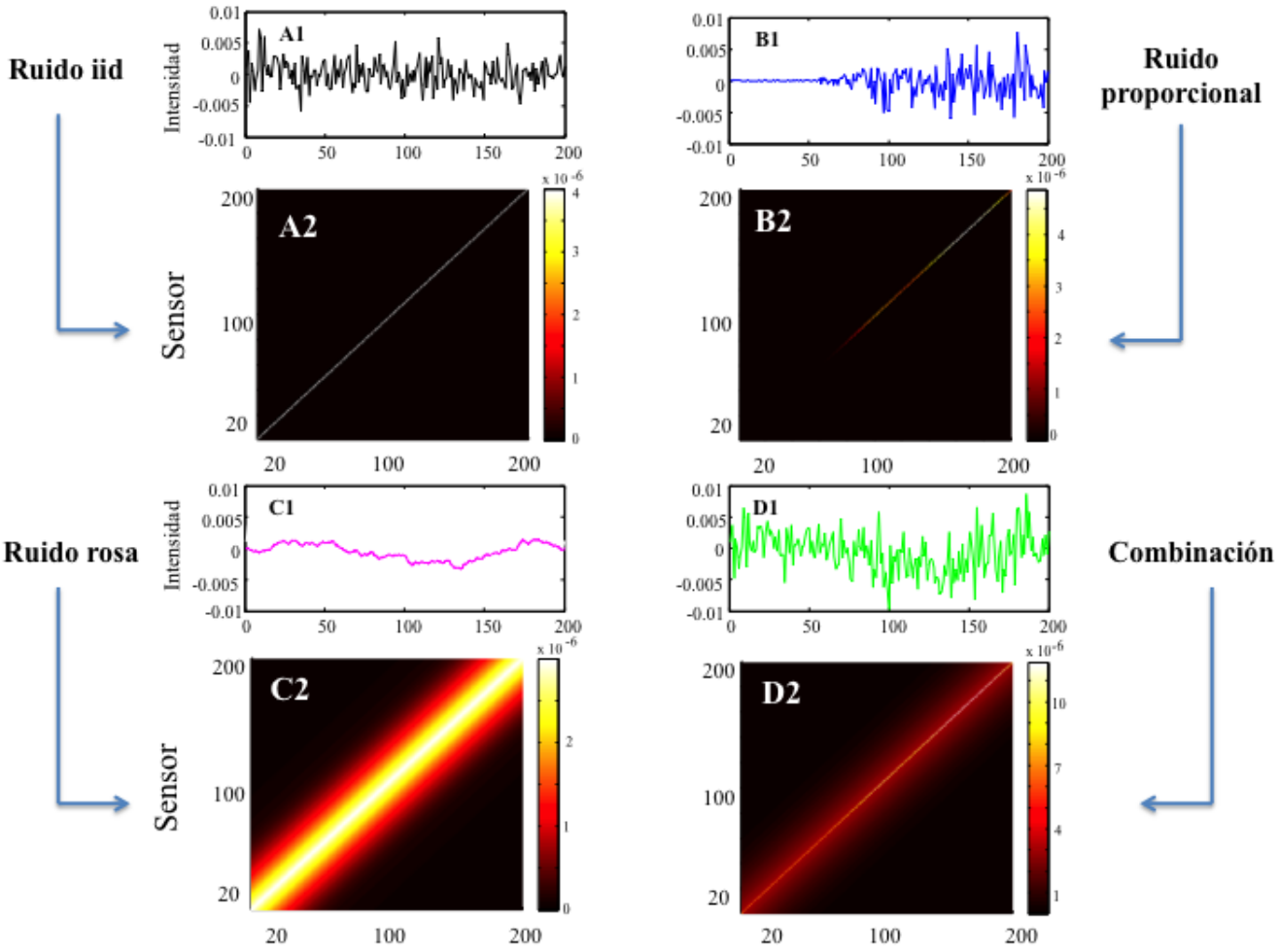


Figura 3.2. Secuencias de ruido típicas (A1, B1, C1 y D1) y matrices de covariancia del error teóricas (A2, B2, C2 y D2) utilizada en las simulaciones. Las matrices se ilustran a partir de las imágenes resultantes de escalar los datos al rango completo del correspondiente mapa de colores. Los diferentes sub paneles ilustran ruido iid (A1 y A2), ruido proporcional (B1 y B2) ruido rosa (C1 y C2) y la suma de las tres secuencias de ruido (D1 y D2).

Finalmente, el modelado empírico representa una alternativa intermedia entre las opciones que se presentaron en los párrafos anteriores. Apunta fundamentalmente a encontrar un modelo capaz de proporcionar una estimación confiable de la matriz de covariancia del error.[29] Aunque esta opción no es tan simple y directa como la de los replicados experimentales y requiere experiencia por parte del analista, tiene varias ventajas, tales como una mejor comprensión del tipo de error de medición que está actuando como limitante en el sistema en estudio, una reducción de la necesidad de réplicas, y el suavizado de las variaciones estocásticas inherentes a la estimación de la

matriz de covariancia del error experimentales. Aunque esta aproximación no fue utilizada en este trabajo, podría convertirse en una alternativa muy práctica para futuras aplicaciones, al mismo tiempo que resalta la importancia y el potencial de las expresiones propuestas en aplicaciones reales. Una vez que se logra un modelo confiable de la matriz de covariancia del error a través de las réplicas de las muestras de calibrado, este puede aplicarse a la estimación de la incertidumbre en nuevas muestras de *test*, incluso sin la necesidad de nuevas réplicas.

Un concepto complementario, utilizado para analizar la estructura del error, es la matriz de correlaciones, obtenida dividiendo cada elemento de la matriz de covariancia del error por los dos desvíos estándares que están contribuyendo.[29,30] Los elementos de la diagonal de la matriz de correlación son todos 1, mientras que los que se encuentran fuera de la diagonal indican el grado de correlación del error espectral.

3.6 Propagación de errores

El error de predicción es una función de los datos de entrada del modelo de calibración utilizado. Toda función diferenciable se puede aproximar por medio de una expansión en forma de serie de Taylor truncada luego del término lineal. Esta aproximación se conoce como linealización local en estadística y como propagación de errores en química y física.[100] Considerando una variable calculada *z* (como la concentración predicha de un analito en una muestra de *test*), que es una función de múltiples variables de manera que $z=f(x_1,x_2,\ldots)$, la fórmula general para la propagación del error se encuentra representada por la siguiente ecuación:[30]

$$\sigma_z^2 = \sum_i \left(\frac{\partial z}{\partial x_i} \right) \sigma_i^2 + 2 \sum_i \sum_{j>i} \left(\frac{\partial z}{\partial x_i} \right) \left(\frac{\partial z}{\partial x_j} \right) \sigma_{ij} \qquad (3.2)$$

donde σ_z^2 es la variancia de los errores en x_i y σ_{ij}^2 la covariancia de los errores en x_i y x_j. Para simplificar la notación, resulta conveniente representar la **Ecuación 3.2** en forma matricial. Esto puede lograrse definiendo el Jacobiano **j** como un vector columna que contiene las derivadas parciales de *z* respecto de *x*:

$$\mathbf{j} = \begin{bmatrix} \frac{\partial z}{\partial x_1} \\ \frac{\partial z}{\partial x_2} \\ \cdots \\ \frac{\partial z}{\partial x_n} \end{bmatrix} \quad (3.3)$$

de esta manera, la variancia en *z* se puede expresar como:

$$\sigma_z^2 = \mathbf{j}^T \mathbf{\Sigma}_x \mathbf{j} \quad (3.4)$$

donde $\mathbf{\Sigma}_x$ (tamaño $n \times n$) es la matriz de covariancia del error para el vector **x**. La **Ecuación 3.4** resulta útil para describir los cambios en la incertidumbre que se dan cuando se aplica una transformación al vector de medidas **x**, produciendo un nuevo escalar *y*. Por ejemplo, si $y = \mathbf{c}^T \mathbf{x}$, siendo **c** un vector de transformación genérico, la matriz de covariancia del error para *y* es:

$$\sigma_y^2 = \mathbf{c}^T \mathbf{\Sigma}_x \mathbf{c} \quad (3.5)$$

Una característica interesante de la expresión anterior es que se puede aplicar a una amplia variedad de situaciones, incluyendo suavizado, diferenciación, proyecciones en subespacios, transformaciones por *wavelet,* y como se mostrará a continuación para estimar la incertidumbre en la predicción en calibración univariada y multivariada.

Aplicación a calibración univariada

El desarrollo anterior puede aplicarse para calcular la variancia en el modelo de regresión lineal. La ecuación básica de este modelo, se puede expresar convenientemente como:

$$y_o = \mathbf{x}_o \mathbf{b} \quad (3.6)$$

donde **b** es un vector columna que incluye a los dos parámetros ajustables del modelo (pendiente (b_1) y ordenada al origen (b_o)) y $\mathbf{x}_o$ un vector fila de dos elementos que contiene la concentración correspondiente al punto de la recta en el cual se está calculando la incertidumbre y un elemento igual a 1 que modela la ordenada al origen (a diferencia de la **Ecuación 1.1**, para dar mayor generalidad al desarrollo, en este caso se considera la ordenada al origen como un parámetro ajustable más. De esta manera, aplicando la **Ecuación 3.4** se llega a que el valor de la incertidumbre en torno a y_o estará dada por:

$$\sigma^2_{y_o} = \mathbf{x}_o \mathbf{\Sigma}_b \mathbf{x}_o^T \quad (3.7)$$

Teniendo en cuenta que para varias muestras la **Ecuación 3.6** se puede expresar de modo aún más general como:

$$\mathbf{y} = \mathbf{Xb} \quad (3.8)$$

donde **X** es una matriz de *I* muestras y *P* parámetros ajustables (en calibración univariada *P* = 2). La matriz de variancia covariancia de **b** puede calcularse teniendo en cuenta que estos parámetros surgen de proyectar el vector de respuestas **y** en el espacio definido por **x**. Es decir,

$$\hat{\mathbf{b}} = \mathbf{X}^+ \mathbf{y} \quad (3.9)$$

de manera que

$$\mathbf{\Sigma}_{\hat{\mathbf{b}}} = (\mathbf{X}^T \mathbf{\Sigma}_y^{-1} \mathbf{X})^{-1} \quad (3.10)$$

donde $\mathbf{\Sigma}_{\hat{\mathbf{b}}}$ tiene un tamaño de $P \times P$ y $\mathbf{\Sigma}_y^{-1}$ (tamaño $I \times I$) es la inversa de la matriz de covariancia de las respuestas. Para errores distribuidos de manera idéntica e independiente (iid) esta matriz se puede aproximar como una matriz identidad multiplicada por la variancia en los residuos de la regresión s^2_{res}.

De esta forma, la variancia entorno a $\mathbf{y}_o$ se calcula a partir de la siguiente expresión de carácter general:

$$\sigma^2_{y_o} = \mathbf{x}_o^T (\mathbf{X}^T \mathbf{\Sigma}_y^{-1} \mathbf{X})^{-1} \mathbf{x}_o \quad (3.11)$$

que para una regresión no ponderada (errores iid), se transforma en:

$$\sigma^2_{y_o} = \sigma^2_y \, \mathbf{x}_o^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_o \approx s^2_{res} \, \mathbf{x}_o^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_o \quad (3.12)$$

y por manipulación algebraica simple se puede expresar como:

$$\sigma^2_{y_o} = s^2_{res} \left(\frac{1}{n} + \frac{(x_o - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right) \quad (3.13)$$

Esta ecuación es la que permite calcular el intervalo de confianza como una región hiperbólica entorno a la recta de regresión.

El resultado anterior puede utilizarse, por ejemplo, para estimar la incertidumbre en la concentración predicha de una muestra desconocida (x_o), a partir del gráfico de calibración univariada, partiendo de la observación de una respuesta (y_o). Esto se ilustra de manera muy aproximada en la **Figura 3.3**, donde puede verse que esta incertidumbre total dependerá tanto de la incertidumbre en la respuesta medida, como de la incertidumbre dada por la recta de regresión. Es decir, existirá una fuente de incertidumbre debida exclusivamente a la señal y otra debida al calibrado.

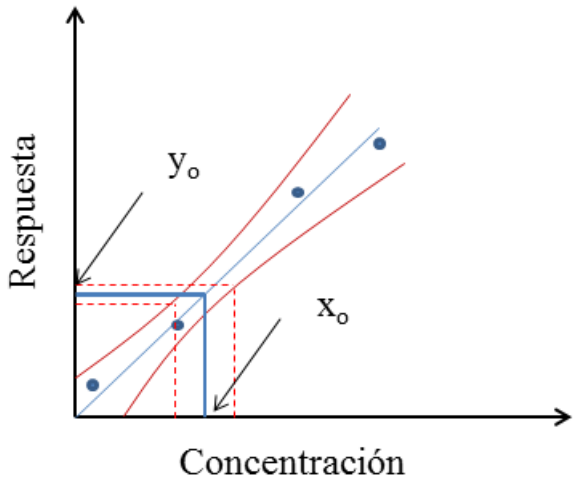


Figura 3.3. Curva de calibración univariada (línea sólida azul) con las correspondientes bandas de incertidumbre estimadas a partir de la **Ecuación 3.13** (línea sólida roja). Para un determinado valor de respuesta (y_o) su incertidumbre se propaga a través del modelo de calibrado generando una determinada incertidumbre en concentración.

El cálculo de la incertidumbre mencionada en el párrafo anterior, sólo tienen sentido cuando la ecuación del modelo que se está analizando es invertible. En el caso de la regresión lineal esto puede hacerse fácilmente expresando x_o como función de y_o ($x_o = f(y_o)$). Es decir:

$$x_o = \frac{y_o - b_o}{b_1} = \frac{y_o}{b_1} - \frac{b_o}{b_1} \qquad (3.14)$$

De esta manera, el vector **j** en este caso está definido:

$$\mathbf{j} = \begin{bmatrix} \frac{\partial f(y_o)}{\partial y_o} \\ \frac{\partial f(y_o)}{\partial b_o} \\ \frac{\partial f(y_o)}{\partial b_1} \end{bmatrix} \tag{3.15}$$

por lo que la variancia en x_o puede expresarse en forma matricial como:

$$\sigma^2_{x_o} = \mathbf{j}^T \begin{bmatrix} \sigma^2_{y_o} & 0 \\ 0 & (\mathbf{X}^T \mathbf{\Sigma}_y^{-1} \mathbf{X})^{-1} \end{bmatrix} \mathbf{j} \tag{3.16}$$

En el caso que las incertidumbres en y sean uniformes (error homoscedástico), y si se midieron m réplicas de y_o la **Ecuación 3.16** se reduce a:

$$\sigma^2_{y_o} = \frac{\sigma^2_y}{m}$$

$$\sigma^2_{x_o} = \mathbf{j}^T \left\{ \sigma^2_y \begin{bmatrix} \sigma^2_{y_o} & 0 \\ 0 & (\mathbf{X}^T \mathbf{X})^{-1} \end{bmatrix} \right\} \mathbf{j} \tag{3.17}$$

Calculando los elementos del vector **j** a partir de las correspondientes derivadas de la **Ecuación 3.15** se llega a:

$$\frac{\partial x_o}{\partial y_o} = \frac{1}{b_1}, \quad \frac{\partial x_o}{\partial b_o} = -\frac{1}{b_1}, \quad \frac{\partial x_o}{\partial b_1} = -\frac{(y_o - b_o)}{b_1^2} \tag{3.18}$$

y consecuentemente **j** puede expresarse como:

$$\mathbf{j} = \begin{bmatrix} \frac{1}{b_1} \\ -\frac{1}{b_1} \\ -\frac{(y_o - b_o)}{b_1^2} \end{bmatrix} \tag{3.19}$$

que luego de cierta manipulación algebraica utilizando la definición:

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \tag{3.20}$$

permite llegar a la ecuación para calcular la incertidumbre en regresión lineal no ponderada:

$$\sigma_{x_o}^2 = \frac{\sigma_y^2}{b_1^2}\left(\frac{1}{m}+\frac{1}{n}+\frac{(y_o-\bar{x})^2}{b_1\sum(x_i-\bar{x})^2}\right) \quad (3.21)$$

donde al igual que en la **Ecuación 3.13** σ_y^2 puede estimarse a partir de los residuos de la regresión s_{res}^2.

Aplicación a calibración multivariada

La misma base de razonamiento que se utilizó para llegar a la **Ecuación 3.11**, puede utilizarse para seguir cómo los errores en las mediciones originales se propagan a través de los diferentes pasos en los análisis por PCR o PLS. En este caso, sin embargo, al tratarse de una calibración de tipo inversa, se podrá arribar a la incertidumbre en la concentración predicha suponiendo que la principal fuente de error es la señal en en la muestra cuya incertidumbre se desea determinar. En este caso, la ecuación de predicción se puede expresarse convenientemente como:

$$\hat{y} = \mathbf{t}\,\mathbf{v} \quad (3.22)$$

donde **v** es el vector de los coeficientes de regresión en el espacio de las variables latentes (tamaño $A\times1$). Dado que **t** es la proyección del vector de la muestra desconocida **x** en el subespacio definido por los *loadings* **V**, la **Ecuación 3.22** puede escribirse como:

$$\hat{y} = \mathbf{v}^T\mathbf{V}^+\,\mathbf{x} = \mathbf{b}^T\,\mathbf{x} \quad (3.23)$$

donde **b** , a diferencia de las secciones anteriores es el vector de regresión de PLS o PCR en el espacio de las variables originales, tal y como se definió en el Capítulo 1. Si se lleva adelante la propagación del error en la ecuación anterior, por analogía con la **Ecuación 3.4**, la incertidumbre en la predicción debido a la incertidumbre en la señal instrumental de la muestra desconocida estará dada por:

$$\sigma_{\hat{y}}^2 = \mathbf{b}^T\,\boldsymbol{\Sigma}_x\mathbf{b} \quad (3.24)$$

3.7 Esquema general para la determinación de la incertidumbre en la predicción

En trabajos previos de la literatura se demostró que la incertidumbre global en la predicción en modelos multivariados inversos se puede estimar como la suma de tres términos independientes. Estos términos tienen en cuenta la propagación de incertidumbre

derivada de: (1) señales instrumentales en las muestras de *test* (σ_1), (2) señales instrumentales en los datos de calibrado (σ_2), y (3) concentraciones de calibración (σ_3).[12] Esto puede expresarse de manera general como una expresión de propagación de la variancia dada por:

$$\sigma_{\hat{y}}^2 = \sigma_1^2 + \sigma_2^2 + \sigma_3^2 \qquad (3.25)$$

Tomando este punto de partida, la propagación del error puede realizarse teniendo en cuenta los supuestos iniciales sobre la estructura de error del conjunto de datos que se esté analizando. Como resume la **Tabla 3.2**, se pueden dar tres posibles escenarios en lo que respecta al error instrumental: (1) ruido iid (caso 1), (2) ruido no iid con todas las muestras de *test* y de calibración teniendo misma estructura de error (caso 2), (3) error no iid en muestras de calibración y de *test* con diferentes estructuras de error según la muestra (caso 3). La **Tabla 3.3** muestra las expresiones finales obtenidas para cada uno de los escenarios descriptos previamente. Los principios para llegar a estas expresiones son los que se presentaron de manera resumida en la sección anterior.

Como se esperaba para el caso 1, donde la estructura del error es iid (**Tabla 3.2**), la matriz de covariancia del error es una matriz identidad multiplicada por la desviación estándar, y la ecuación para $\sigma_{\hat{y}}^2$ coincide con la propuesta por Faber y Kowalski.[12] De cualquier manera, la suposición iid no es necesaria para derivar una expresión más general basada en la estimación de la matriz de covariancia del error, tal como como se propone para los casos 2 y 3 en la **Tabla 3.2**.

Tabla 3.2 Expresiones obtenidas por propagación de errors para cada uno de los posibles casos de estructura de error.[a]

	$\sigma_{\hat{y}}^2$	σ_1^2	σ_2^2	σ_3^2
Caso 1		$\mathbf{b}^{\mathrm{T}}\mathbf{b}\sigma_x^2 = \frac{\sigma_x^2}{\mathrm{SEN}^2}$	$h\mathbf{b}^{\mathrm{T}}\mathbf{b}\sigma_X^2 = h\frac{\sigma_X^2}{\mathrm{SEN}^2}$	$h\sigma_{y_{\mathrm{cal}}}^2$
Caso 2		$\mathbf{b}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathrm{x}}\mathbf{b}$	$h\mathbf{b}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathrm{X}}\mathbf{b}$	$h\sigma_{y_{\mathrm{cal}}}^2$
Caso 3		$\mathbf{b}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathrm{x}}\mathbf{b}$	$h\mathbf{b}^{\mathbf{T}}\boldsymbol{\Sigma}_{\mathrm{X,eff}}\mathbf{b}$	$h\sigma_{y_{\mathrm{cal}}}^2$

[a] SEN = sensibilidad del analito, h = leva de la muestra, σ_x^2 = variancia del error en las señales de la muestra de *test*, σ_X^2 = variancia del error en las señales de calibración, $\sigma_{y_{\mathrm{cal}}}^2$ = variancia del error en las concentraciones de calibración, **V**= matriz de *loadings* de calibrado, **b** = vector de coeficientes de regresión, $\mathbf{y}_{\mathrm{cal}}$= vector de concentraciones de calibrado, $\boldsymbol{\Sigma}_{\mathrm{x}}, \boldsymbol{\Sigma}_{\mathrm{X}}$= matrices de covariancia del error para las señales de la muestra y las señales de calibración respectivamente, y $\boldsymbol{\Sigma}_{\mathrm{X,eff}}$, matriz de covariancia efectiva del error para el conjunto de calibrado.

Las expresiones para los términos 1 y 2 en el caso 2 (**Tabla 3.2**) son extensiones naturales del caso 1. Es importante mencionar que las matrices de covariancia del error están representadas por diferentes símbolos: el subíndice 'x' en $\boldsymbol{\Sigma}_{\mathrm{x}}$ se refiere a la matriz de covariancia del error para la muestra de *test*, mientras que 'X' en $\boldsymbol{\Sigma}_{\mathrm{X}}$ se reserva para los datos de calibrado. De cualquier manera, en el caso 2 se supone que todas las muestras tienen la misma estructura de error y por lo tanto las matrices de covariancia del error para las muestras de *test* y de calibrado serán iguales. La nomenclatura de la **Tabla 3.2** intenta distinguir la independencia que, en principio, existiría entre los términos 1 y 2.

El caso 3 de la **Tabla 3.2** corresponde a la más general de las situaciones analizadas hasta el momento. La expresión del término 2 en el caso 3 resulta de particular interés debido a su complejidad. En este caso, la estructura del error varía de muestra a muestra en el conjunto de calibrado, haciendo necesaria la inclusión de matrices variancia covariancia del error individuales asociadas a cada una de las muestras de calibración (se desarrollará con mayor detalle en la información suplementaria). El resultado es una expresión relacionada con la análoga para el caso 2, reemplazando la matriz de covariancia del error clásica por una efectiva. Esta última consiste en un promedio pesado de todas las matrices de covariancia, es decir:

$$\sum_{X,eff} = \frac{1}{h}\sum_{X,1} h_1^2 + \sum_{X,2} h_2^2 + \ldots + \sum_{X,i} h_i^2 \qquad (3.26)$$

donde $\mathbf{\Sigma}_{X,1}$, $\mathbf{\Sigma}_{X,2}$,… son las matrices de covariancia del error para cada muestra de calibración, h_1, h_2,…son los elementos del vector $\mathbf{h} = \mathbf{tT}^+$ de tamaño 1 x *I*, y *h* es la leva de la muestra.

3.8 Resultados

3.8.1 Simulaciones

Para validar las ecuaciones mostradas en la **Tabla 3.2**, se aplicó una metodología iterativa de adición de ruido y posterior calibración/predicción (**Figura 3.3**). Para cada estructura de error de la señal (iid,correlacinada,heteroscedástica), lo anterior se realizó de la manera ya discutida, en cuatro situaciones diferentes, incluyendo sólo error en concentraciones de calibración, sólo en señales de calibración, sólo en las muestras de *test*, y en todas ellas en conjunto. Para testear la adecuación de las ecuaciones de un modo confiable, se utilizó la matriz de covariancia del error teórica (ver **Figura 3.2**).

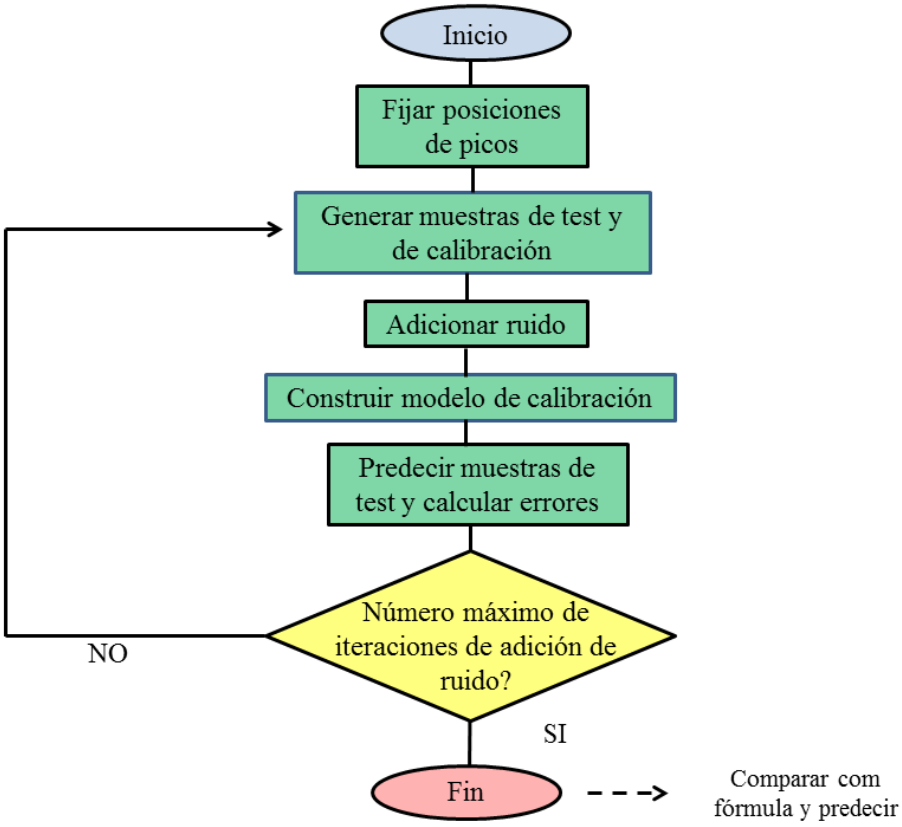


Figura 3.4. Diagrama de flujo resumiendo la rutina utilizada para realizar las simulaciones de adición de ruido.

Debido al número de sistemas estudiados, una manera conveniente de resumir los datos es graficando las incertidumbres obtenidas por Monte Carlo respecto de aquellas estimadas a través de la expresión correspondiente en la **Tabla 3.2**, identificando tres fuentes diferentes de incertidumbre utilizando símbolos específicos. Esta estrategia se ha empleado previamente para probar expresiones para el cálculo de la sensibilidad en sistemas multi-vía.[15-16] La **Figura 3.5** muestra los resultados obtenidos para los juegos de datos simulados en este trabajo bajo el efecto de los distintos tipos de ruido descriptos con anterioridad.

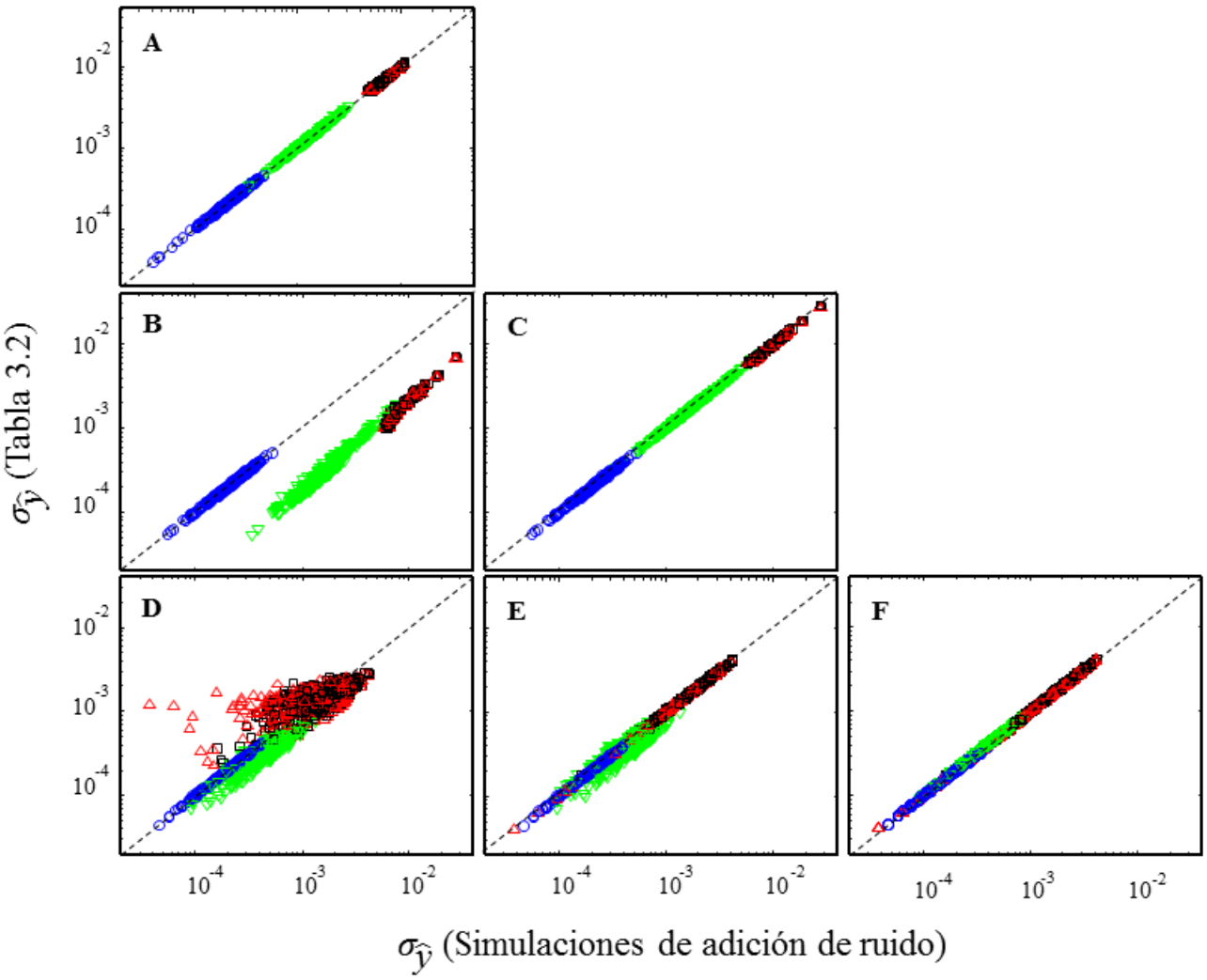


Figura 3.5.Gráficos de las incertidumbres en las concentraciones predichas calculadas por medio de las expresiones propuestas, como función de los resultados de las simulaciones de adición de ruido. Los diferentes paneles muestran los resultados de adición de: ruido iid (A), ruido rosa (B y C), y ruido proporcional (D,E y F). Los valores teóricos fueron estimados utilizando la siguiente información proveniente de la **Tabla 3.2**: (A), (B) y (D), expresiones para el caso1, (C) y (E), expresiones para el caso 2 y (F), expresiones para el caso 3. En el gráfico (E), $\mathbf{\Sigma}_X$ es la matriz de covariancia del error media obtenida sobre el conjunto de calibración. En todos los gráficos, los símbolos identifican los siguientes casos: círculos azules, ruido en las concentraciones de calibración, triángulos verdes hacia abajo ruido sólo en las señales de calibración, triángulos rojos hacia arriba, ruido en las señales de la muestra de *test* y cuadrados negros, ruido en concentraciones y señales. La línea de puntos indica correlación perfecta. Todos los ejes se encuentran en escala logarítmica.

Los resultados presentados en la **Figura 3.5** permiten concluir que la propagación de errores basada en la inclusión de la matriz de covariancia del error es apropiada, debido a que las incertidumbres de Monte Carlo coinciden con los correspondientes valores de incertidumbre derivados por las ecuaciones propuestas (**Figuras 3.5 A, 3.5 C** y **3.5 F**).

Como es de esperarse, cuando se supone una estructura de ruido de tipo iid en sistemas con una estructura de error real que difiere de esta situación, la expresión clásica para el caso 1 falla de manera notoria en la estimación de la incertidumbre (lo cual puede apreciarse fácilmente en las **Figuras 3.5 B** y **3.5 D**). En estos casos, las diferencias con respecto a las expresiones que tienen en cuenta la correlación y la heteroscedasticidad de los datos son, de al menos un orden de magnitud.

El análisis de los efectos del ruido proporcional requiere una atención especial. La expresión que idealmente debería utilizarse cuando aparece este tipo de ruido se muestra en la **Tabla 3.2** para el caso 3, e incluye un segundo término que tiene en cuenta la variación en la matriz de covariancia del error entre las distintas muestras de calibrado. Una alternativa más simple sería utilizar para el caso 2, un término que depende de la matriz media de covariancia del error obtenida a partir de todas las muestras de calibrado. Esto representa una aproximación al caso 2 cuando los espectros de calibración son similares y resulta en un cálculo mucho más directo. Sin embargo, en la medida que las matrices de covariancia del error individuales se desvían de la media, la exactitud en las incertidumbres predichas disminuye, como se muestra en la **Figura 3.5 E**, y por lo tanto deberían utilizarse las expresiones para el caso 3.

Es interesante notar que, para las condiciones de simulación empleadas, el cálculo de la incertidumbre en la predicción suponiendo errores iid (Caso 1) lleva a una subestimación del verdadero valor de incertidumbre ante la presencia de error correlacionado (**Figura 3.5 B**), mientras que para el ruido proporcional esta relación no es directa (**Figura 3.5 D**). La primera observación puede ser entendida cualitativamente en relación con la Ecuación para el caso 1 en la **Tabla 3.2**, donde los términos para iid excluyen la contribución de la covariancia del error, que en este caso es positiva. En otras palabras, la correlación significa que los errores aleatorios no se cancelan de la manera que lo hacen en el caso de una estructura de error no correlacionada. La incertidumbre que surge del ruido proporcional, por otro lado, puede interpretarse teniendo en cuenta que el vector de regresión pesará de manera diferente a las distintas regiones espectrales en la ecuación que se utiliza para predecir la incertidumbre (**Ecuación 3.24**). En consecuencia, si en las regiones espectrales donde el vector de regresión es significativo hay señales con mayor intensidad que en otras regiones, las incertidumbres calculadas utilizando una incertidumbre de medición promedio tipo iid serán menores. Ocurrirá lo opuesto para altas

intensidades espectrales en regiones donde el vector de regresión es pequeño, debido a la presencia en esas zonas de componentes de la muestra distintos del analito.

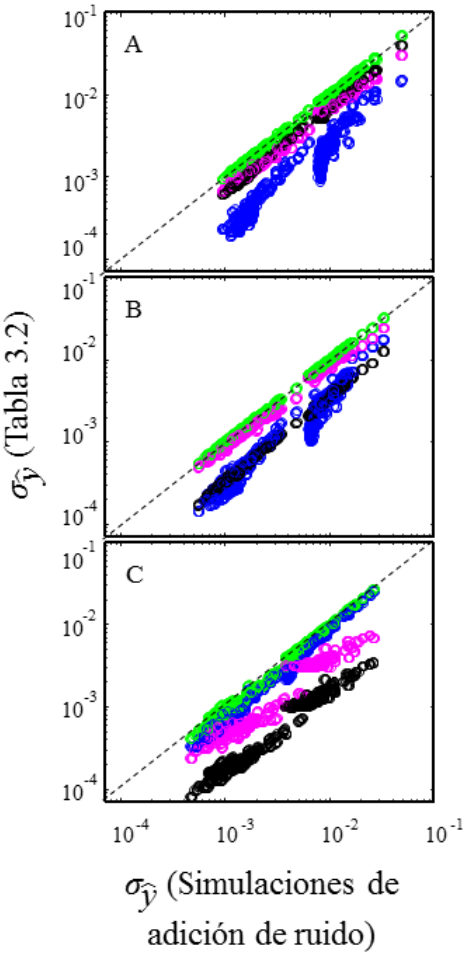


Figura 3.6 Gráfico de incertidumbres en concentraciones predichas en función de los resultados obtenidos en las simulaciones de adición de ruido, cuando se utiliza una combinación de ruido iid, rosa y proporcional en distintas proporciones relativas. Las principales fuentes de error son: (A) iid, (B) rosa, y (C) proporcional. Los círculos verdes corresponden a la desviación estándar calculada utilizando las expresiones de la **Tabla 3.2**, con una matriz de covariancia del error construida a partir de la suma de las tres fuentes de error individuales. Los círculos negros corresponden a incertidumbres calculadas cuando en la matriz de covariancia sólo se considera el ruido iid, los círculos rosas cuando sólo se considera el ruido rosa, y los círculos azules cuando sólo se tiene en cuenta el ruido proporcional.

Finalmente, se estudiaron sistemas de ruido combinados, que se construyeron adicionando todas las fuentes de error mencionadas, de manera de analizar el efecto integrado. Con este propósito, se fue modificando la proporción relativa de cada tipo de ruido dando lugar a tres casos distintos: (1) ruido iid dominante (**Figura 3.6**) (2) ruido parcialmente correlacionado dominante y (3) ruido proporcional dominante. En cada uno

de estos casos, la incertidumbre en la predicción se estimó a partir de 4 matrices de covariancia del error diferentes: una incluyendo todos los tipos de error (círculos verdes), y el resto considerando separadamente cada una de las posibles fuentes d error estudiadas (círculos azules, negros y rosas respectivamente). En la **Figura 3.6** es evidente que las ecuaciones que utilizan las últimas tres matrices de covariancia del error mencionadas, subestiman la incertidumbre en la predicción, tal y como era de esperarse. La mayor desviación de la recta identidad, que marca la situación ideal, corresponde a las estimaciones basadas en la matriz de covariancia para la fuente de error menos influyente en proporción relativa. Por otro lado, si sólo se tienen en cuenta la fuente de error predominante, la estimación sigue siendo buena a pesar del hecho que otros tipos de ruidos se encuentren presentes en cantidades menores.

3.8.2 Datos experimentales

El impacto de la estructura del error en la estimación de la incertidumbre tendría que poder evidenciarse también cuando se trabaja con datos experimentales, en los que el ruido rara vez es homoscedástico. En estos sistemas, la estrategia de adición de ruido y predicción no es viable, ya que no se dispone de los datos sin error. Por lo tanto, es necesario encontrar una manera confiable de evaluar cuál de las expresiones tratadas es más realista.

Una alternativa interesante, es la aproximación propuesta por Faber y Bro, que se basa en calcular el siguiente valor de *t* para cada una de las muestras predichas:[100]

$$t = \frac{\hat{y} - y_{\text{ref}}}{\sigma_{\hat{y}}} \qquad (3.27)$$

donde y_{ref} e $\hat{y}$ son los valores de concentración predichos y de referencia y $\sigma_{\hat{y}}$ la desviación estándar estimada por la correspondiente ecuación utilizada para estimar la incertidumbre en la predicción. En caso que se obtengan estimaciones válidas de la incertidumbre, los valores de *t* para un grupo de muestras de predicción deberían estar aproximadamente distribuidos como una *t* de Student, que, para un gran número de grados de libertad, tiene un valor de desvío estándar cercano a 1.[100] En consecuencia, si el verdadero error de predicción se encuentra correctamente estimado, el desvío estándar de la *t* calculados para cada una de las muestras de *test* predichas utilizando una estrategia de tipo *leave-one-out*, debería aproximarse a la unidad.

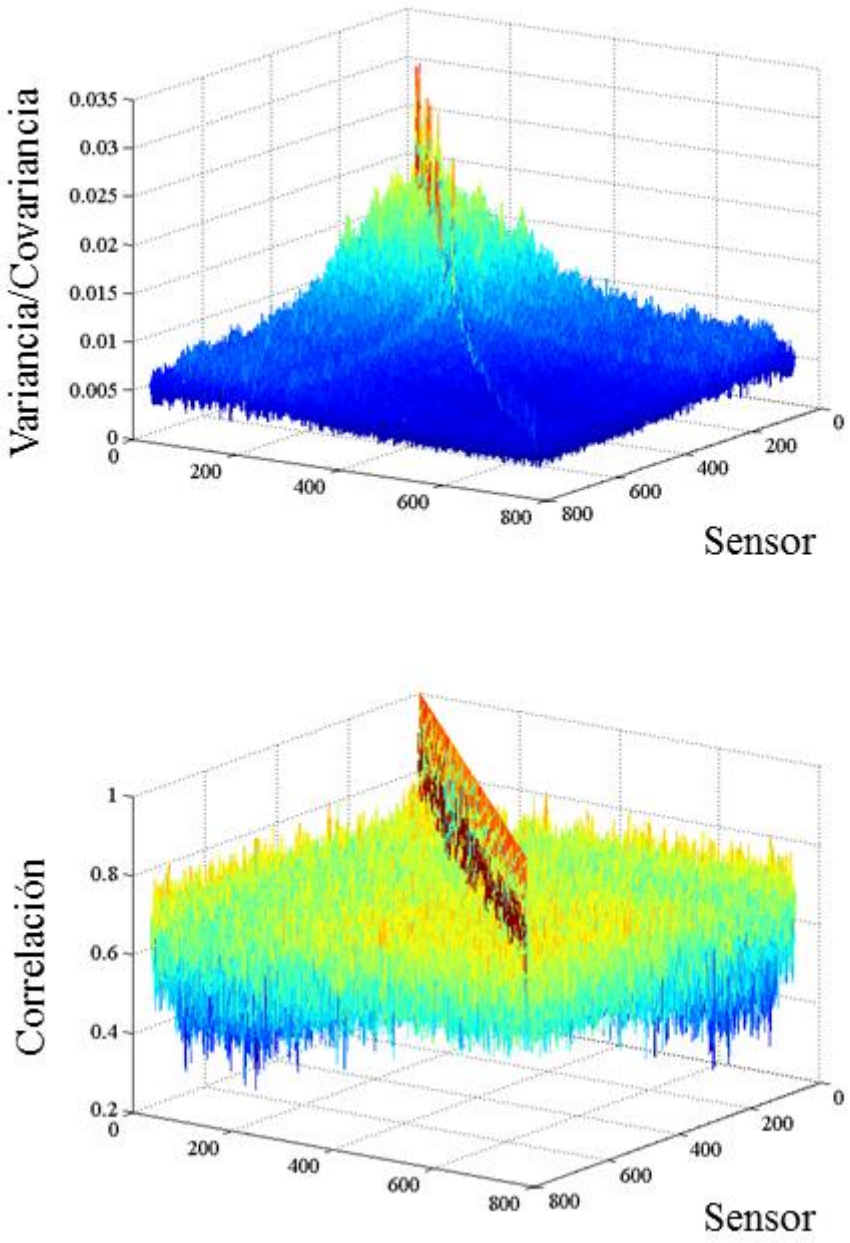


Figura 3.6. Matriz de covariancia del error (A) y matriz de correlación (B) para datos de emisión de fluorescencia.

En la **Figura 3.6** se observa la matriz de covariancia del error correspondiente al juego de datos de emisión de fluorescencia analizado, así como la matriz de correlación. Estas matrices ya han sido caracterizadas y modeladas por Wentzell y colaboradores,[29] utilizando una estrategia de *target testing* para identificar las fuentes de error más

importantes, en conjunto con análisis por componentes principales para calcular las contribuciones correspondientes. Las conclusiones más importantes fueron que las contribuciones principales a la estructura del error son: (1) ruido proporcional relacionado al ya bien conocido en microscopía de fluorescencia *shot noise*, (2) ruido de *offset* constante proveniente del posicionamiento de la celda de medición o de la medida del blanco, y (3) ruido de *offset* de tipo multiplicativo proporcional a la raíz cuadrada del espectro medio.

Para cada una de las muestras del conjunto completo, se quitaron las cinco réplicas y luego se confeccionó el modelo de calibración junto con la estimación de la matriz de covariancia del error, utilizando el resto de las muestras con sus respectivas réplicas. Las predicciones se realizaron utilizando PCR y PLS con cuatro variables latentes obteniéndose resultados muy similares para ambos modelos. A pesar de que se podría esperar un óptimo de tres variables latentes en relación a la composición de la muestra, otros factores como el corrimiento de la línea de base podrían incrementar este número.[101] La **Tabla 3.3** muestra el RMSECV y la desviación estándar media (MSD) calculada utilizando las ecuaciones en la **Tabla 3.2**, para diferentes supuestos de estructura de ruido. Como puede apreciarse, el error estándar de predicción se subestima de manera notoria cuando se asume que el ruido es iid (lo cual es consistente con la **Figura 3.5 B** y algunos casos de la 3.5 D en las simulaciones). Sin embargo, si se supone una estructura de ruido de tipo heteroscedástica, es decir, considerando sólo los elementos diagonales de la matriz de covariancia del error, el valor de MSD se incrementa y aproxima al RMSECV experimental. El cálculo de la incertidumbre se hace incluso más certero cuando se considera la matriz de covariancia del error completa, llegando a valores que concuerdan muy bien con los observados experimentalmente.

Tabla 3.3. Verificación de la estimación de la incertidumbre para los modelos construídos para cada uno de los compuestos del conjunto de datos reales correspondientes a espectros de emisión de fluorescencia.[a]

	A	RMSECV	MSD			σ_t		
			iid	het	niid	iid	het	niid
Acenapththylene	4	0.0081	0.0036	0.0072	0.0075	2.27	1.14	1.10
Naphthalene	4	0.0011	0.0004	0.0008	0.0011	2.63	1.46	1.07
Phenanthrene	4	0.0003	0.0001	0.0002	0.0003	2.46	1.83	1.02

[a] *A* = número de variables latentes, RMSEP = raíz cuadrada del error cuadrado medio de predicción, MSD = desvío estándar medio, σ_t, desvío estándar de los valores de *t*, het = heteroscedástico y no correlacionado, niid = no-iid, (heteroscedástico y correlacionado).

La **Tabla 3.3** también muestra el desvío estándar de los valores de $t(\sigma_t)$ dados por la **Ecuación 3.27**. Un valor de σ_t menor a la unidad implica una sobrestimación del error estándar de predicción y viceversa. Los resultados coinciden con los discutidos anteriormente al comparar los valores de RMSECV y MSD. El valor de σ_t calculado para los tres analitos bajo el supuesto iid se encuentra alrededor de 2 para los tres analitos, implicando una subestimación significativa del error estándar de predicción. Sin embargo, σ_t pasa a estar más cercano a la unidad cuando se tiene en cuenta la heteroscedasticidad del ruido del sistema en estudio, a la vez que se aproxima mucho al valor ideal para los tres compuestos cuando también se tiene en cuenta la influencia del ruido correlacionado. Al igual que para los valores de MSD, la presencia de errores correlacionados parece tener el mayor efecto para el naftaleno y el fenantreno. En los tres casos, la incertidumbre en la predicción está levemente subestimada, y esto podría deberse al uso de una matriz de covariancia del error promedio en lugar de una efectiva que tenga en cuenta las influencias de cada muestra de calibrado, como muestra la **Figura 3.6.**

Las tendencias discutidas muestran claramente la importancia de prestar atención a las diferentes contribuciones que afectan la estructura del error de los datos que se estén analizando para poder estimar correctamente los valores de incertidumbre. Es importante notar que las expresiones propuestas permiten la estimación de desvíos en las predicciones para distintas fuentes de error, sin depender del valor de RMSEP, que no es específico para cada muestra. Aunque en este caso se utilizó una estrategia de combinación de replicados

de las distintas muestras, los efectos sobre la matriz de covariancia del error sólo fueron analizados en términos de los errores heteroscedásticos (elementos diagonales) o correlacionados (elementos no diagonales). Sin embargo, si se utilizara una estrategia de modelado empírico, sería posible hacer un análisis de la influencia de cada una de las fuentes de error identificadas, como ya se realizó en datos simulados (**Figura 3.6**).

3.9 Conclusión

Desde el comienzo de la quimiometría como disciplina para asistir a la química analítica, sus pioneros han insistido en que los desarrollos realizados en este campo deberían funcionar como una guía de ayuda para los fabricantes de equipos a la hora de tomar decisiones en lo que respecta a la mejora en la calidad de las mediciones. Con este objetivo, en este capítulo se investigó un estimador para calcular la incertidumbre en la predicción, basado en la obtención previa de un modelo de la matriz de covariancia del error. Para cuantificar el impacto de las distintas fuentes de error, se realizó una propagación de la estructura de error multivariada al resultado final. La comparación de las expresiones propuestas con la fórmula clásica utilizada para el caso del ruido iid, muestra diferencias significativas en la incertidumbre de predicción, dependiendo de la variación de la estructura del error en muestras de *test* y de calibrado. Finalmente, una conclusión importante respecto de los datos instrumentales, es que la medición de una cantidad razonable de réplicas durante la etapa de calibrado, combinada con una estrategia de agrupación de replicados o de modelado empírico por PCA, permite estimar la incertidumbre en la predicción de muestras futuras, incluso si estas se midieron en ausencia de replicados.

3.10 Apéndice

En este apéndice se presentará una deducción detallada del segundo término de la expresión que permite calcular la variancia del error en la predicción debido a la contribución del calibrado (σ_2^2) cuando la matriz de covariancia del error varía de manera significativa de muestra a muestra. Como se explicó en la Sección 3.7, la **Tabla 3.2** resume el esquema generalizado propuesto para el cálculo de la incertidumbre en calibración multivariada, teniendo en cuenta las características de las distintas estructuras de error que pueden afectar al sistema en estudio. Faber y Kowalski[12] desarrollaron la deducción de los tres términos para el caso 1 (ruido iid). Para los términos 1 y 2 en el caso 2 y para el

término 1 en el caso 3 (matriz de covariancia del error común para todas las muestras) la extensión es casi directa, tal como fue descripto en la Sección 3.6, por medio de una metodología de propagación de errores. Sin embargo, para el término 2 en el caso 3, es decir, cuando cada muestra tiene una estructura de error específica, como por ejemplo en presencia de ruido proporcional, la deducción merece una atención especial.

La concentración predicha de un analito utilizando PLS o PCR ($\hat{y}$) es un escalar que puede expresarse como:

$$\hat{y} = \mathbf{t}\mathbf{T}^{+}\mathbf{y}_{\text{cal}} \qquad \text{(A3-1)}$$

donde **t** y **T** corresponden al vector de *scores* para la muestra de *test* y a la matriz de *scores* de calibración, $\mathbf{y}_{\text{cal}}$ es el vector de concentraciones de calibrado del analito de interés y '+' implica la operación pseudoinversa. Para llegar a las expresiones de interés es necesario diferenciar la **Ecuación A3-1**. Es importante tener en cuenta que en el contexto del término 2 en la expresión para el cálculo de incertidumbre, se supone que las concentraciones de calibración, al igual que las señales de la muestra de *test* no contienen ruido, mientras que éste afecta solamente a las señales de calibrado. Sin embargo, es lógico esperar que la propagación de este ruido afecte la descomposición por PLS de la matriz de datos **X**. Esto generaría ruido en: (1) los *scores* de calibrado **T**, (2) los *loadings* de calibrado **V**, y (3) los *scores* de la muestra de *test* **t**, debido a que estos últimos son el resultado de proyectar la señal de la muestra de *test* sobre los *loadings* de calibrado.

De esta manera, la diferenciación de la **Ecuación A3-1** lleva a:

$$d\hat{y} = \mathbf{t}d(\mathbf{T}^{+})\mathbf{y}_{\text{cal}} + (d\mathbf{t})\mathbf{T}^{+}\mathbf{y}_{\text{cal}} \qquad \text{(A3-2)}$$

donde $d\hat{y}$ es un escalar que representa un cambio diferencial en la concentración predicha de $\hat{y}$ debido al error que afecta a las señales de calibración. La correspondiente variancia en la predicción se puede expresar como:

$$\sigma_2^2 = E(d\hat{y}^2) \qquad \text{(A3-3)}$$

donde E() significa valores esperados.

Para calcular el cambio en los valores de *scores* de la muestra *d***t** en la **Ecuación A3-2**, hay que considerar la relación que existe entre **t**, el vector señal de la muestra de *test* **x** y los *loadings* de calibración P:

$$\mathbf{x} = \mathbf{P}\,\mathbf{t}^{T} \qquad \text{(A3-4)}$$

y por lo tanto se obtiene la siguiente expresión ya que se supone que **x** no propaga ruido:

$$0 = \mathbf{P}\,d\mathbf{t}^{T} + d\mathbf{P}\mathbf{t}^{T} \qquad \text{(A3-5)}$$

de donde *d***t** se puede obtener como:

$$d\mathbf{t} = -\,\mathbf{t}\,d\mathbf{P}^{T}\,\mathbf{V} \qquad \text{(A3-6)}$$

donde **V** es el espacio generalizado de *loadings*, es decir, en PCR **V**=**P** y en PLS $\mathbf{V}=(\mathbf{W}^{T}\mathbf{P})^{-1}\mathbf{W}$. En estos modelos, $\mathbf{V}^{T}\mathbf{V}$ y $\mathbf{P}^{T}\mathbf{V}$ corresponden a una matriz de tamaño *A x A*, donde *A* es el número de variables latentes.

Reemplazando este valor de *d***t** en la **Ecuación A3-2** se obtiene:

$$d\hat{y} = \mathbf{t}d(\mathbf{T}^{+})\mathbf{y}_{cal} - \mathbf{t}(d\mathbf{P}^{T})\mathbf{V}\mathbf{T}^{+}\mathbf{y}_{cal} = \mathbf{t}d(\mathbf{T}^{+})\mathbf{y}_{cal} - \mathbf{t}(d\mathbf{P}^{T})\mathbf{V}\mathbf{v} \qquad \text{(A3-7)}$$

Para poder expresar esta última ecuación como una función del ruido en señales de calibración (d**X**), se debe analizar la relación $\mathbf{X} = \mathbf{T}\mathbf{P}^{T}$ del modo siguiente:

$$\mathbf{P}^{T} = \mathbf{T}^{+}\mathbf{X} \qquad \text{(A3-8)}$$

$$d\mathbf{P}^{T} = \mathbf{T}^{+}d\mathbf{X} + d(\mathbf{T}^{+})\,\mathbf{X} \qquad \text{(A3-9)}$$

De aquí, la **Ecuación A3-7** se transforma en:

$$d\hat{y} = \mathbf{t}d(\mathbf{T}^{+})\mathbf{y}_{cal} - \mathbf{t}\mathbf{T}^{+}(d\mathbf{X})\mathbf{V}\mathbf{v} - \mathbf{t}(d\mathbf{T}^{+})\mathbf{X}\mathbf{V}\mathbf{v} \qquad \text{(A3-10)}$$

En este punto es interesante considerar el último término de la **Ecuación A3-10**. El producto (**XVv**= **Tv**) es el vector de concentraciones predichas en el conjunto de calibración, y con un buen grado de aproximación, es equivalente a $\mathbf{y}_{cal}$. Por lo tanto, el primer y último término de la **Ecuación A3** se cancelan aproximadamente uno a otro dejando un único término que tiene en cuenta $d\hat{y}$:

$$d\hat{y} = -\mathbf{t}\mathbf{T}^{+}(d\mathbf{X})\mathbf{V}\mathbf{v} \qquad \text{(A3-11)}$$

A partir de este resultado, es posible llegar a una expresión para calcular la variancia en las concentraciones predichas.

De las **Ecuaciones A3-3** y **A3-11**:

$$\sigma_2^2 = E[\mathbf{v}^T\mathbf{V}^T(d\mathbf{X}^T)\mathbf{T}^{+T}\mathbf{t}^T\mathbf{t}\mathbf{T}^+(d\mathbf{X})\mathbf{V}\mathbf{v}] \quad \text{(A3-12)}$$

En esta última ecuación, el vector $\mathbf{t}\mathbf{T}^+$ (de tamaño $1\times J$) se puede interpretar como un vector leva **h,** que da lugar a la leva multivariada (h) correspondiente a la muestra de *test*, un parámetro adimensional que posiciona a la muestra de *test* en relación al espacio de calibrado:

$$h = \| \mathbf{h} \|^2 = \mathbf{h}\,\mathbf{h}^T = \mathbf{t}\mathbf{T}^+\,\mathbf{T}^{+T}\mathbf{t}^T \quad \text{(A3-13)}$$

En la **Ecuación A3-12**, sin embargo, el factor más relevante es $\mathbf{T}^{+T}\mathbf{t}^T\mathbf{t}\mathbf{T}^+$, que corresponde a la siguiente matriz **H** de $I\times I$:

$$\mathbf{H}= \mathbf{h}^T\mathbf{h} = \begin{bmatrix} h_1^2 & \dots & h_1h_I \\ \dots & \dots & \dots \\ h_Ih_1 & \dots & h_I^2 \end{bmatrix} \quad \text{(A3-14)}$$

Para poder expresar la **Ecuación A3-12** como una función de la matriz de covariancia del error para las señales de calibración, en primer lugar se la expande en términos de los valores específicos del producto $(d\mathbf{X}^T)\mathbf{H}(d\mathbf{X})$:

$$(d\mathbf{X}^T)\mathbf{H}(d\mathbf{X}) = \begin{bmatrix} dx_{11} & \dots & dx_{1I} \\ \dots & \dots & \dots \\ dx_{J1} & \dots & dx_{JI} \end{bmatrix}\begin{bmatrix} h_1^2 & \dots & h_1h_I \\ \dots & \dots & \dots \\ h_Ih_1 & \dots & h_I^2 \end{bmatrix}\begin{bmatrix} dx_{11} & \dots & dx_{J1} \\ \dots & \dots & \dots \\ dx_{1I} & \dots & dx_{IJ} \end{bmatrix} \quad \text{(A3-15)}$$

En esta última ecuación, sólo se pueden conservar los productos correspondientes a *dx* de la misma muestra, ya que los valores esperados del producto cruzado para muestras diferentes pueden considerarse como 0. En este sentido, la **Ecuación A3-15** lleva a una expresión particularmente simple:

$$E[(\mathrm{d}\mathbf{X}^T)\mathbf{H}(\mathrm{d}\mathbf{X})\,] = \sum_{\mathrm{X},1} h_1^2 + \sum_{\mathrm{X},2} h_2^2 + \ldots + \sum_{\mathrm{X},i} h_i^2 \quad \text{(A3-16)}$$

donde $\mathbf{\Sigma}_{\mathrm{X},i}$ es la matriz de covariancia del error para la muestra de calibración i. La **Ecuación A3-16** permite definir una matriz de covariancia del error efectiva $\mathbf{\Sigma}_{\mathrm{X,eff}}$, como el promedio pesado de todas las matrices de covariancia del error para el conjunto de muestras de calibrado:

$$\sum_{X,eff} = \frac{1}{h}\sum_{X,1} h_1^2 + \sum_{X,2} h_2^2 + \ldots + \sum_{X,i} h_i^2 \qquad \text{(A3-17)}$$

En esta ecuación, el peso corresponde al cuadrado de h_i dividido por la leva h. Finalmente, partiendo de los resultados anteriores, se puede obtener una expresión definida y compacta para el término 2 en el caso 3:

$$\sigma_2^2 = h\mathbf{v}^T\mathbf{V}^T\Sigma_{X,eff}\mathbf{V}\mathbf{v} = h\mathbf{b}^T\Sigma_{X,eff}\mathbf{b} \qquad \text{(A3-18)}$$

donde **b** es el vector de coeficientes de regresión generado por el modelo multivariado.

Consistencia de la ecuación para el caso 3 con las del caso 1 y el caso 2

Un paso necesario para validar la **Ecuación A3-18** es mostrar la consistencia entre los casos 1 y 2. Suponiendo que todas las matrices de covariancia del error son idénticas (caso 2), la **Ecuación A3-19** lleva a:

$$\sum_{X,eff} = \frac{1}{h}\sum_X (h_1^2 + h_1^2 + \ldots + h_i^2) = \Sigma_X \qquad \text{(A3-19)}$$

Si se inserta este resultado en la **Ecuación A3-18**:

$$\sigma_2^2 = h\mathbf{b}^T\mathbf{\Sigma}_X\mathbf{b} \qquad \text{(A3-20)}$$

Esta última ecuación muestra la consistencia de la **Ecuación A3-18** con el caso 2.

Finalmente, bajo el supuesto iid, $\mathbf{\Sigma}_X = \sigma_X\mathbf{I}_I$ ($\mathbf{I}_I$ es una matriz identidad de tamaño $I \times I$), generando:

$$\sigma_2^2 = h\sigma_X^2\mathbf{b}^T\mathbf{b} = h\frac{\sigma_X^2}{SEN^2} \qquad \text{(A3-21)}$$

donde SEN se refiere a la sensibilidad para el analito que se está cuantificando, dada por:

$$SEN = \frac{1}{\sqrt{\mathbf{b}^T\mathbf{b}}} \qquad \text{(A3-22)}$$

Teniendo en cuenta estos últimos resultados, la consistencia de la **Ecuación A3-18** con el caso 1 queda demostrada.

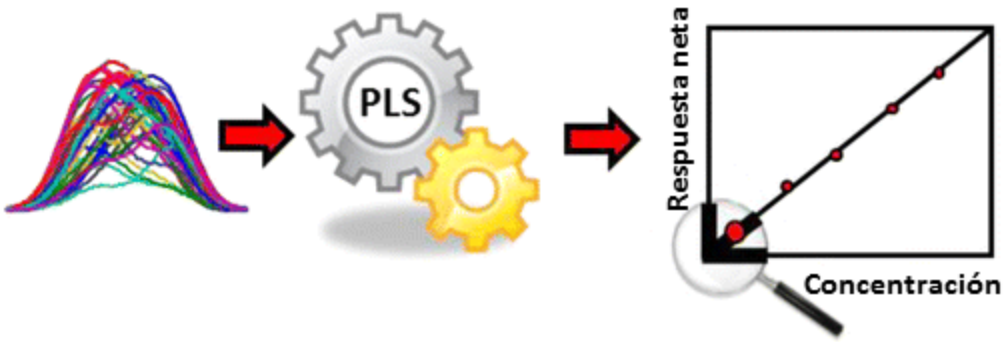
3.11 Perspectivas

Futuras investigaciones en la temática desarrollada durante este capítulo, podrían incluir la extensión del análisis actual a metodologías de máxima probabilidad como MLPCR, que tienen en cuenta la estructura del error a la hora de construir modelos multivariados.

El resto de las perspectivas de este capítulo serán presentadas junto con las del Capítulo 4 en la Sección 4.13, debido a la estrecha relación que existe entre el cálculo del desvío estándar de predicción y el del límite de detección, cifra de mérito que será tratada en el próximo capítulo.

CAPÍTULO 4

LÍMITE DE DETECCIÓN EN CALIBRACIÓN MULTIVARIADA DE PRIMER ORDEN POR PCR Y PLS



"La diferencia entre la estupidez y el genio es que el genio tiene sus límites." (Albert Einstein).

4.1 Resumen

En calibración multivariada, no existe actualmente ningún procedimiento estandarizado que permita calcular el límite de detección. Esto se debe fundamentalmente a que, en este escenario, la definición de un estimador consistente resulta más compleja que una simple extensión de la expresión utilizada en calibración univariada. Por este motivo, aunque se realizaron algunos intentos, se necesita de un esfuerzo adicional para llegar a un acuerdo total entre las definiciones del LOD en los campos univariado y multivariado.

Teniendo en cuenta lo anterior, en este capítulo se describirá una nueva metodología para estimar el LOD cuando se trabaja en calibración por cuadrados mínimos parciales. En lugar de un único valor, este método se basa en proponer un intervalo de límites de detección, que dependerá de la influencia de la variación de la composición de las muestras en el espacio de calibrado. Esta manera de interpretar el LOD contrasta de algún modo con otras definiciones basadas en una extensión de la fórmula de cálculo

univariada. Con la nueva definición propuesta, el intervalo de límites de detección se convierte en un parámetro que caracteriza a la calibración PLS completa, y no a cada una de las muestras en particular (como se había propuesto hasta el momento). La nueva metodología tiene en cuenta las recomendaciones oficiales de la IUPAC, así como también los últimos desarrollos en la teoría llamada EIV[12] para calibración PLS. Para dar cuenta de las características del nuevo concepto de LOD, este se estudió tanto en juegos de datos simulados como reales.

4.2 Introducción

El LOD se encuentra entre las cifras de mérito más controvertidas.[93,102,103] Posiblemente esta controversia tenga que ver con la importancia que esta cifra de mérito tiene en química analítica, que a su vez radica en su carácter de excelente medida de la calidad de un modelo de calibración, ya que su definición abarca en conjunto dos conceptos analíticos de gran importancia: la sensibilidad y la precisión en las determinaciones analíticas.

Cuando la señal analítica es univariada y específica para cada muestra, la regla de detección está basada en el *test* de Neyman-Pearson, que considera las probabilidades de error de falsos negativos y falsos positivos para la hipótesis nula "no hay analito" y la hipótesis alternativa "hay analito". La estimación puede realizarse directamente a partir de la curva de calibración univariada, como una alternativa simple a la recomendación original, en la que el LOD se estima a partir del nivel medio de la señal y el desvío estándar para un conjunto de medidas realizadas sobre muestras a concentraciones cercanas al límite de detección esperado (1 a 5 veces).[104]

Sin embargo, como se mencionó durante la introducción general, cuando se trabaja en calibración multivariada, surge la dificultad que las señales instrumentales no son específicas para un analito en particular. En respuesta a esto, Lorber y colaboradores desarrollaron una aproximación para el cálculo del LOD, basada en el concepto de NAS.[12] Sin embargo, la dificultad de este estimador reside en que sólo considera la incertidumbre en la medida de la señal, haciendo que la aplicación real sea bastante limitada, debido a que otras fuentes importantes de incertidumbre son las concentraciones y las señales de calibración. Otras estrategias similares se basan en calcular la desviación estándar del blanco a partir de los residuos espectrales y tienen la misma desventaja.[105]

Por su parte, Rius y colaboradores sugirieron un límite de detección multivariado basado en el cálculo de lo que denominaron "respuesta de detección", que es específica para el analito de interés y evalúa probabilidades de error tanto de tipo I como de tipo II.[106] Este trabajo sugiere que el valor del LOD debería calcularse para cada muestra de *test*, indicando implícitamente que el LOD podría interpretarse como un rango de concentraciones en lugar de un único valor de concentración. Es decir, sugiere que el LOD multivariado no sólo sería específico para cada analito sino también para cada muestra. Sin embargo, los autores expusieron la necesidad de llevar a cabo más estudios para calcular una respuesta de detección no ambigua. En varios trabajos de la literatura,[107,108] se propuso un método similar basado en una fórmula simplificada para calcular un error estándar en concentraciones calculadas por PLS, específico para cada muestra.[100] De cualquier manera, en todos estos trabajos, la leva (parámetro adimensional utilizado para medir la posición de la muestra en el espacio de calibrado) de cada muestra a concentración 0 del analito es sólo una aproximación, y no existe un procedimiento adecuadamente establecido para calcularla.

Finalmente, Ortiz y colaboradores propusieron una metodología para calcular el LOD que surge directamente de extender las recomendaciones IUPAC para métodos univariados a calibración multivariada.[27,109] Esta generalización está basada en la prueba matemática de que el límite de detección, tal como está definido en las normas ISO y por la IUPAC para calibración univariada, es invariable para transformaciones lineales de la respuesta. Como consecuencia, se obtiene el mismo límite de detección utilizando la regresión de concentraciones estimadas versus concentraciones de calibración. Estos últimos valores pueden medirse por una técnica de referencia, o bien asignarse nominalmente cuando se preparan en el laboratorio a partir de estándares del analito de interés. Aunque esta propuesta "pseudounivariada" aparece en principio como válida, no es completa dese el punto de vista de su correspondencia con los últimos avances realizados en propagación de errores en calibración PLS, basados en los modelos EIV.[12] En particular, no es consistente con la idea de un límite de detección dependiente de cada muestra.[110,92]

La metodología para el cálculo del límite de detección multivariado que se describirá en este trabajo de tesis está basada en una serie de ideas complementarias tales como: (1) cada muestra de *test* tiene en principio un valor de LOD específicamente asociado, (2) el universo de muestras de *test* se encuentra bien representado por el conjunto

de muestras de calibración, (3) las levas para las muestras de calibración pueden ser extrapoladas a concentración cero de los analitos, y (4) se puede estimar un rango de valores de LOD para el modelo PLS en conjunto. Los límites mínimo y máximo del intervalo (LOD_{min} y LOD_{max} respectivamente) corresponden a las muestras de calibración con los valores mínimo y máximo de leva, resultantes de una extrapolación a concentración cero del analito de interés. Los resultados obtenidos permiten establecer una conexión entre el LOD_{min}, LOD_{max} y el límite de detección pseudounivariado (LOD_{pu}). Finalmente, la propuesta será testeada en algunos sistemas simulados y experimentales.

4.3 Objetivos específicos

1) Llevar a cabo un análisis de las propuestas para el cálculo del LOD que se realizaron desde el surgimiento de la calibración multivariada (primer orden) como alternativa analítica a la calibración univariada (orden 0).

2) Desarrollar una metodología para calcular el LOD, consistente con los principios de funcionamiento del modelo de regresión PLS.

3) Comparar el método propuesto con otro de gran difusión en la actualidad, tanto en datos simulados como reales.

4.4 Cálculo del error estándar de predicción por muestra bajo el supuesto de ruido idéntico e independientemente distribuido (iid)

Como se mencionó durante la Sección 3.7, cuando se trabaja bajo el supuesto de que el ruido en la señal se distribuye de manera idéntica e independiente (supuesto *iid*), la variancia en las concentraciones predichas del analito por medio del modelo PLS está dada por la siguiente expresión, bien conocida y presentada como "caso 1" durante la Sección 3.7 del Capítulo 3:

$$\sigma_{\hat{y}}^2 = \text{SEN}^{-2}\sigma_x^2 + h\text{SEN}^{-2}\sigma_X^2 + h\sigma_y^2 \qquad (4.1)$$

Como se indicó en la sección citada, el segundo y el tercer término de la **Ecuación 4.1** provienen de las incertidumbres en la calibración y están escalados por la leva de la muestra. Esta última es proporcional a la distancia de Mahalanobis de una muestra al centro del espacio de calibrado (para datos que se encuentran centrados), y puede

expresarse como función de concentraciones, variables instrumentales o variables latentes. En esta última situación, una expresión apropiada para *h* sería:[99]

$$h = \mathbf{t}\,(\mathbf{T}^{T}\,\mathbf{T})^{-1}\,\mathbf{t}^{T} \qquad (4.2)$$

donde **T** es la matriz de *scores* (tamaño *I*×*A*) para las muestras de calibrado, que se obtiene proyectando la matriz de señales de calibración **X** sobre los *loadings* de PLS y **t** un vector de *scores* de la muestra a la cual se desea determinar la leva (tamaño *1*×*A*). Los valores apropiados de σ_x^2 (variancia del error en señal) y σ_y^2 (variancia del error en concentración) provienen normalmente de un análisis de réplicas o se estiman a partir de otras fuentes.[100]

Nótese que cuando tanto señales como concentraciones están centradas antes del modelado por PLS, se requieren dos términos adicionales en el lado derecho de la **Ecuación 4.1**, que tienen la misma forma que los últimos dos términos de esta expresión, pero con la leva reemplazada por 1/*I*, donde *I* es el número de muestras de calibración.[99] Una manera simple de tener en cuenta este hecho es definir una "nueva leva efectiva" como (*h* +1/*I*) para ser utilizada en la **Ecuación 4.2** y en todas las ecuaciones que requieren que se estime $\sigma_{\hat{y}}^2$ en datos centrados.

4.5 Fundamento del concepto de intervalo de LOD

De acuerdo con las últimas recomendaciones de la IUPAC, la estimación del límite de detección debe cumplir con dos condiciones: (1) estar basada en la teoría de prueba de hipótesis, teniendo en cuenta las probabilidades de falsos positivos y falsos negativos e (2) incluir todas las posibles fuentes de error, tanto en los pasos de calibrado como de predicción, que puedan afectar el resultado final.

Considerando la primera condición, el LOD multivariado debería estar basado en la misma expresión que se utiliza en calibración univariada:[110]

$$\mathrm{LOD} = (t_{\alpha,\nu} + t_{\beta,\nu})\,\sigma_{y_o} \qquad (4.3)$$

donde σ_{y_o} es el desvío estándar en la concentración para la muestra blanco, y $t_{\alpha,\upsilon}$y $t_{\beta,\upsilon}$ los coeficientes de una distribución *t* de *Student* con υ grados de libertad. Estos últimos dos parámetros tienen en cuenta la probabilidad de cometer errores tipo I (suponer que el analito está presente cuando en realidad está ausente) con una probabilidad α, y errores tipo II (suponer que el analito está ausente cuando en realidad está presente) con una

probabilidad β. Normalmente a α y β se les asigna un valor de 0.05 (es decir, un intervalo de confianza del 95%), υ es normalmente grande para un conjunto de calibración con múltiples muestras y, por lo tanto, el factor $\left(t_{\alpha,\upsilon}+t_{\beta,\upsilon}\right)$ toma el valor aproximado de 3.3.

Es importante notar que en la **Ecuación 4.3** la distancia entre el blanco y el LOD es aproximada por la suma de dos intervalos de confianza. Una aproximación más rigurosa sugiere el uso de un parámetro de no centralidad para una distribución *t* no centrada, en lugar de una suma de coeficientes *t*.[111] De cualquier modo, los valores obtenidos por medio de estos métodos estadísticos alternativos no difieren de manera significativa.[112] Igualmente, en caso de requerirse mayores detalles, se puede encontrar en la literatura un análisis riguroso de los estimadores del límite de detección basados en intervalos de predicción.[113,114]

Un punto de importancia fundamental en la **Ecuación 4.3** es el criterio adoptado para estimar la variancia en las concentraciones predichas, que estaría relacionado con punto 2 de los requisitos presentados al comienzo de esta sección. En este sentido, para poder estimar el LOD, la **Ecuación 4.3** requiere que se conozca el valor de $\sigma_{y_o}^2$, es decir, la variancia en la concentración para una muestra blanco (el valor de $\sigma_{\hat{y}}^2$ cuando y=0), que en principio podría obtenerse a partir de la **Ecuación 4.1**. Por lo tanto, la leva cuando la concentración del analito es 0 (h_0) juega un rol fundamental. Sorpresivamente, hasta el momento no han surgido propuestas para estimar este parámetro. Sólo se han sugerido aproximaciones a h_0, que tienen en cuenta muestras que se supone se encuentran cercanas al límite de detección.[107,108]

Una extensión directa e intuitiva del concepto univariado de LOD lleva a pensar en un único valor de límite de detección para el caso multivariado, aunque un análisis más profundo indica que este no sería el caso. En calibración univariada, existe un único valor de h_0, que en principio se puede estimar de manera confiable a partir de los parámetros de la calibración. Sin embargo, en calibración multivariada, h_0 tomará diferentes valores dependiendo de la composición de cada muestra. De acuerdo con la **Ecuación 4.2**, cada muestra de *test* a concentración cero del analito, pero conteniendo diferentes niveles de otros componentes, todos contribuyendo al espectro de la muestra, generarán un conjunto de *scores*, y por lo tanto un valor específico de la leva h_0.[92] Es por esto, que en el contexto de la calibración PLS, es más razonable considerar la existencia de un intervalo de límites

de detección, cuyos valores dependerán de la variabilidad de la composición de fondo de las muestras, en lugar de fijar un único valor de LOD.

4.6 Cálculo del intervalo de LOD

Si se piensa en el caso simple de un sistema ternario formado por un analito a ser cuantificado, en presencia de dos componentes adicionales, el número de variables latentes de calibrado para construir un modelo PLS sería de 3. Esto significa que cada muestra tiene asociado un vector de *scores* **t** de tamaño *1×3*, y por lo tanto se puede graficar como un punto en un espacio tridimensional. La **Figura 4.1** muestra la ubicación de un conjunto de muestras donde se puede ver que: (1) las muestras a concentración cero del analito (círculos rojos) se encuentran en una región definida H_0 del plano π_0, y (2) las proyecciones de las posiciones de las demás muestras de *test* (círculos amarillos), perpendiculares a π_0, también se encuentran en H_0. Esto sugiere que esta última región comprende todas las muestras blanco posibles (tomando como referencia el analito 1 como componente de interés) y representadas por el conjunto de calibración elegido. La idea fundamental del presente trabajo es encontrar los límites de $\mathbf{H}_0$ en el espacio de los *scores*, incluso si las muestras blanco no se incluyeron en el conjunto de calibración.

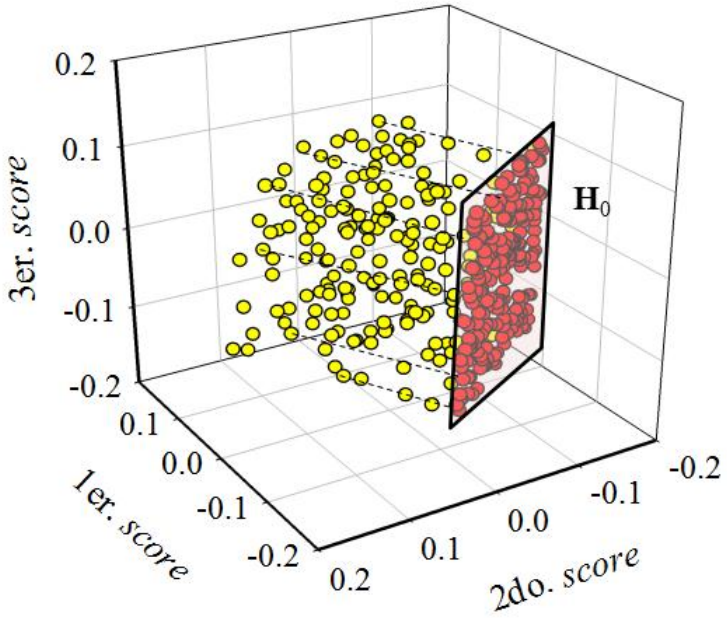


Figura 4.1 Localización de las muestras en el espacio de los *scores* obtenidos por PLS para un conjunto de datos simulados con tres componentes: en círculos amarillos muestras a concentraciones aleatorias de los tres componentes (en un rango de 0 a 1); en círculos rojos, muestras a concentración cero del analito a cuantificar y concentraciones aleatorias de los dos componentes adicionales (en el mismo rango de valores).

En general, para todos los conjuntos de calibración existe un plano π_0 que representa los *scores* de las muestras para las cuales la concentración del analito de interés se encuentra ausente (es decir, el *background* específico de cada muestra). Teniendo en cuenta la ecuación de predicción en el espacio de las variables latentes para PLS y PCR, el hiperplano *A*-dimensional del espacio de *scores* se puede definir utilizando la siguiente ecuación (suponiendo centrado de señales y concentraciones):

$$\pi_0: \mathbf{tv} + \bar{y}_{cal} = 0 \qquad (4.4)$$

Dado que el LOD es una función de la variancia en las concentraciones predichas del analito para una muestra blanco, que es a la vez función de h_0, estimar el límite de detección consistiría en encontrar el mínimo (h_{0min}) y el máximo (h_{0max}) valor de este parámetro para un determinado juego de calibrado. Desde un punto de vista geométrico, h_{0min} es la mínima distancia entre π_0 y el centro del espacio de *scores* normalizado (ver Apéndice), esto es, la distancia perpendicular desde π_0 al centro. Como se demuestra en el Apéndice, h_{0min} está dado por

$$h_{0\min} = \frac{\bar{y}_{\text{cal}}^2}{\sum_{i=1}^{I} y_i^2} \qquad (4.5)$$

donde y_i es la concentración centrada para la *i*-ésima muestra de calibrado. La leva calculada por medio de la **Ecuación 4.5** corresponde al valor obtenido en calibración univariada para un dado conjunto de calibración, en caso que el resto de los componentes estuviera ausente.[5] Por otro lado, el límite superior $h_{0\max}$ se puede estimar calculando las levas de las proyecciones de todas las muestras de calibración en π_0 (ver apéndice):

$$h_{0\text{cal}} = h_{\text{cal}} + h_{0\min}\left[1 - \left(\frac{y_{\text{cal}}}{\bar{y}_{\text{cal}}}\right)^2\right] \qquad (4.6)$$

donde h_{cal} e y_{cal} son la leva y la concentración centrada de un analito para una muestra de calibración genérica. De esta manera, se puede encontrar el máximo de todos los $h_{0\text{cal}}$ posibles:

$$h_{0\max} = \max(h_{0\text{cal}}) \qquad (4.7)$$

Los valores de $h_{0\min}$ y $h_{0\max}$ (o las "levas efectivas", ($h_{0\min}$ + 1/*I*) y ($h_{0\max}$+1/*I*) para datos centrados) se pueden introducir en las **Ecuaciones 4.8** y **4.9** para obtener los valores mínimo y máximo del intervalo de límites de detección:

$$\text{LOD}_{\min} = 3.3\,[\text{SEN}^{-2}\sigma_x^2 + h_{0\min}\text{SEN}^{-2}\sigma_X^2 + h_{0\min}\,\sigma_{y_{\text{cal}}}^2]^{1/2} \qquad (4.8)$$

$$\text{LOD}_{\max} = 3.3\,[\text{SEN}^{-2}\sigma_x^2 + h_{0\max}\text{SEN}^{-2}\sigma_X^2 + h_{0\max}\sigma_{y_{\text{cal}}}^2]^{1/2} \qquad (4.9)$$

Estos límites pueden informarse para una calibración PLS basada en un determinado conjunto de muestras y caracterizar el modelo completo en lugar de una muestra específica de *test*.

Es importante resaltar que tanto $\text{LOD}_{\min}$ como $\text{LOD}_{\max}$ dependen de la leva, que a su vez es función de la matriz de *scores* **T**. Dado que esta matriz depende del diseño de calibrado, esto es, el conjunto de muestras seleccionados para la calibración y el número de variables latentes de calibración, los límites del intervalo LOD también dependerán de estos dos factores. La importancia de las metodologías para determinar un número de factores que impidan el sobreajuste, así como para seleccionar un conjunto de muestras que incluyan la mayor variabilidad posible de futuras muestras en el espacio de calibrado, ha

sido tratada en detalle en la literatura.[88,24] Esto implica que un supuesto subyacente importante de este trabajo es que el diseño correcto de la calibración debería dar a una predicción no sesgada del límite de detección.

4.7 Regla de decisión para la detección

Una vez que se establecen los límites del intervalo de LOD, el analista puede concluir que el analito no fue detectado en una muestra determinada si la concentración predicha se encuentra por debajo del LOD_{min}, o que está presente si la concentración predicha está por encima del LOD_{max}. Por lo tanto, en principio, la pregunta quedaría sin respuesta para muestras cuya concentración predicha de analito se encuentre entre los límites del intervalo de LOD. La **Figura 4.1** muestra una representación esquemática de las tres situaciones que se podrían presentar.

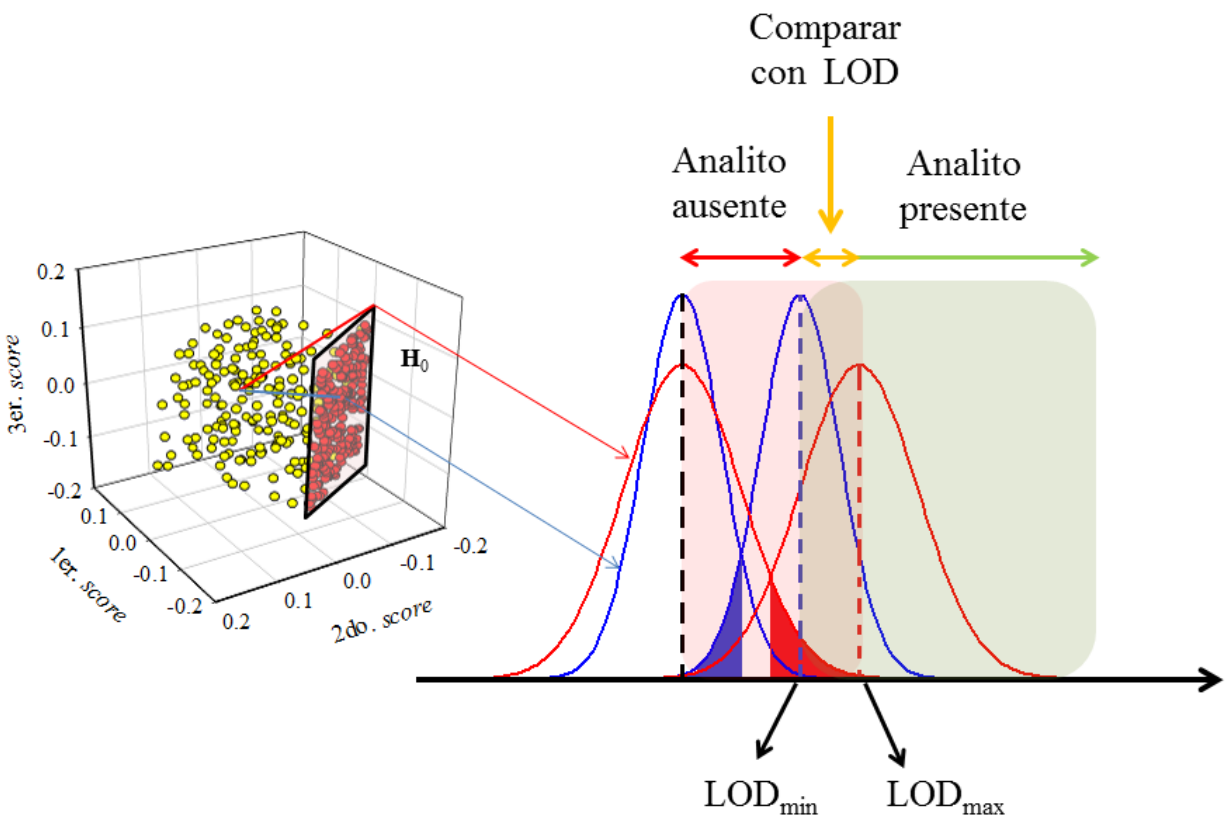


Figura 4.2. Representación esquemática de la metodología de intervalos de límite de detección para calcular el LOD en calibración por PLS. En líneas sólidas rojas se muestra la leva y la correspondiente desviación estándar utilizada para calcular el LOD_{min}, en líneas sólidas azules se muestra la leva y el correspondiente desvío estándar utilizado para calcular el LOD_{max}.

En el rango de concentraciones $LOD_{min}<y<LOD_{max}$, la pregunta puede ser respondida estimando un valor específico de LOD para la muestra de *test*, aproximando su leva real a h_0, lo que correspondería a los componentes del *background* (es decir, en ausencia de analito). Lo anterior sería equivalente a suponer que la muestra es un blanco, lo cual es en principio lógico, dado que la concentración del analito en una muestra de este tipo probablemente sea muy baja. El valor de LOD obtenido puede ser utilizado para controlar si la concentración predicha se encuentra por debajo (analito ausente) o por encima (analito presente) de este valor de LOD específico para la muestra en análisis.

4.8 LOD pseudounivariado (LODpu)

En esta estrategia, las concentraciones estimadas del conjunto de calibrado se grafican en función de las respectivas concentraciones nominales o medidas.[27] El resultado es un gráfico de calibración pseudounivariado en el cual la escala vertical es la concentración estimada del analito en lugar de las variables instrumentales o latentes. El gráfico se procesa como en el caso de la calibración univariada, suponiendo que la forma de la fórmula utilizada para calcular el límite de detección no se modifica ante cualquier tipo de transformación lineal aplicada a la señal. Esto lleva a un valor de LOD_{pu}, que se estima utilizando la expresión univariada clásica a través de la ecuación:[93]

$$LOD_{pu} = 3.3\, s_{pu}^{-1} [(1 + h_{0min} + 1/I)\, \sigma_{pu}^2]^{1/2} \qquad (4.10)$$

donde s_{pu} es la pendiente de la curva de calibración pseudounivariada and σ_{pu}^2 es la variancia de los residuos de la regresión. La **Ecuación 4.10** no incluye ningún término que tenga en cuenta la incertidumbre en las concentraciones de calibrado, como se acostumbra en calibración multivariada.

El parámetro LOD_{pu} tiene la ventaja de ser una única cifra de mérito que caracteriza a la calibración PLS global. Sin embargo, la idea subyacente no es consistente con la del intervalo LOD descripto anteriormente, al igual que la relación entre el LOD_{pu} y los valores mínimo y máximo de los valores del intervalo LOD_{min} y LOD_{max}. Como ya se mencionó anteriormente, uno de los objetivos de este trabajo es el de poder explicar esta relación, como será desarrollado en secciones posteriores.

4.9 Datos

4.9.1 Simulados

Los datos sintéticos fueron generados imitando un sistema de tres componentes, siendo el componente 1 el analito de interés. Cada espectro del conjunto de calibración y de *test* (**x**) se construyó utilizando la siguiente expresión:

$$\mathbf{x} = y_1\mathbf{s}_1 + y_2\mathbf{s}_2 + y_3\mathbf{s}_3 \qquad (4.11)$$

donde $\mathbf{s}_1$, $\mathbf{s}_2$, y $\mathbf{s}_3$ son los espectros de los componentes puros a concentración unitaria definidos en un rango de 100 sensores (**Figura 4.3**), e y_1, y_2 e y_3 son las concentraciones de los componentes para una muestra determinada. Las señales de los componentes puros $\mathbf{s}_1$, $\mathbf{s}_2$, y $\mathbf{s}_3$ son funciones de forma Gaussiana, centradas en los sensores 50, 40 y 20, respectivamente, con anchos de banda a mitad de altura de 24 sensores en los tres casos. Todos los constituyentes presentes en el conjunto de calibración están compuestos por 100 muestras a concentraciones aleatorias entre 0 y 1. Se crearon dos tipos de muestras de *test*, donde: (1) todos los componentes tienen concentraciones entre 0 y 1 en 100 muestras diferentes, y (2) el analito de interés (componente 1) se encuentra ausente, y los dos componentes restantes en concentraciones aleatorias en el rango de 0 a 1 en 100 muestras adicionales diferentes.

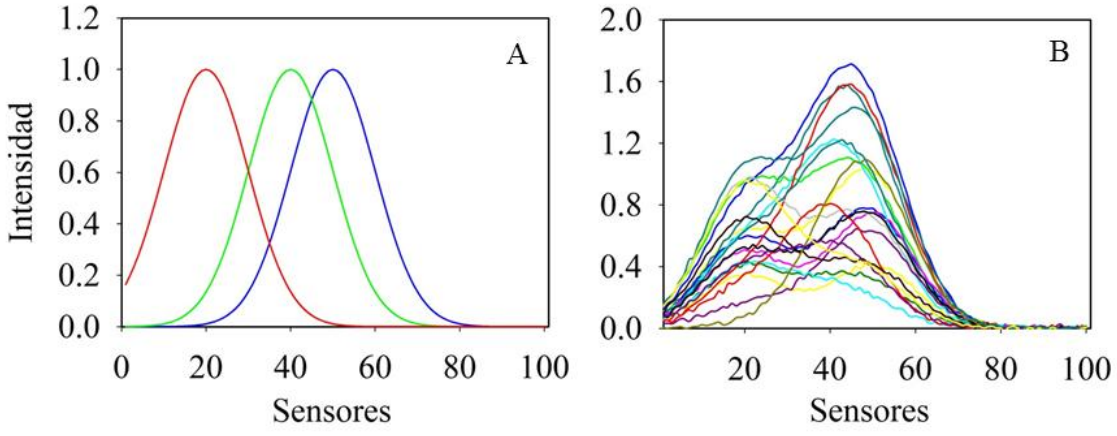


Figura 4.3.(A) Espectros de los componentes puros utilizados para construir los conjuntos de datos simulados: línea azul, analito de interés; líneas verdes y rojas, componentes adicionales de las muestras. (B) Espectros de calibración representativos creados a partir de los perfiles sin ruido mostrados en (A), incluyendo ruido instrumental aleatorio.

Se adicionó ruido Gaussiano iid de tres maneras diferentes: (1) sólo en concentraciones de calibración, (2) sólo en señales de calibración y (3) tanto en concentraciones como en señales. La **Figura 4.3 B** muestra algunas señales de calibración incluyendo ruido en señal. Para cada una de estas formas de adición de ruido, el procedimiento de calibración y predicción se repitió 1000 veces centrando tanto las señales como las concentraciones de calibrado, y seguidamente se calculó la curva de calibración pseudounivariada por medio de la regresión de los valores predichos de concentración del analito respecto a las concentraciones nominales para el conjunto de calibrado. Los parámetros estadísitcos de la curva de calibración se utilizaron para estimar el LOD_{pu} en cada ciclo de adición, del mismo modo propuesto por Ortiz y colaboradores. Luego se comparó el valor medio del LOD_{pu} con los extremos del LOD propuesto en este trabajo, estimado a partir de las **Ecuaciones 4.8** y **4.9** utilizando las levas efectivas (h_{0min} +1/*I*) y (h_{0max} + 1/*I*).

4.9.2 Experimentales

Se emplearon varios juegos de datos experimentales que habían sido previamente analizados por PLS, determinando el LOD utilizando las dos mismas metodologías que en el caso de los datos simulados. Estos datos corresponden a los siguientes analitos y tipos de muestras: (1) ion fluoruro en aguas naturales conteniendo sulfato como potencial interferente, [115] (2) 2-s-butil-4,6-dinitrofenol (DINOSEB) en una mezcla de reacción compleja conteniendo hidrocarburos aromáticos,[116] (3) bromhexina en jarabes antitusivos,[117] (4) antibiótico tetraciclina en suero humano,[118] (5) biodiesel en muestras de combustible,[119] y (6) humedad en semillas de maíz. Los datos espectrales para estos sistemas son los siguientes: (1), (2), y (3), espectros UV-visible, (4) espectro de fluorescencia sincrónica, y (5) y (6), espectros NIR. Los detalles experimentales sobre la preparación de los estándares de calibración y las muestras de *test*, la medición de las señales experimentales y el modelado por PLS pueden encontrarse en las referencias 115 a 119. El juego de datos nro. (6) se encuentra disponible en Internet en el sitio http://www.eigenvector.com/data/Corn/. En todos los casos, tanto la señal como la concentración se centraron antes del modelado por PLS.

4.10 Resultados

Datos simulados

El conjunto de datos simulados se utilizó para calcular y comparar el LOD_{pu} con el intervalo de límites de detección propuesto en este trabajo (desde LOD_{min} a LOD_{max}). Las simulaciones de adición de ruido permitieron estudiar el comportamiento de ambas formas de estimar el límite de detección bajo el efecto de distintos niveles de ruido. Las simulaciones se llevaron a cabo de la siguiente manera: luego de crear un conjunto de datos con una sensibilidad determinada dada por la posición relativa de los picos respecto a los interferentes, se adicionó ruido de acuerdo a como fue presentado en la sección de Datos. Se construyó un modelo PLS a partir de espectros y concentraciones centradas utilizando tres variables latentes de calibración y las concentraciones se predijeron tanto en las muestras de calibración como en las de *test*. El proceso de calibración/predicción se repitió 1000 veces utilizando diferentes "semillas de aleatoriedad", dependiendo de la manera en que el ruido fue adicionado a los datos sintéticos. En cada uno de estos ciclos, las concentraciones predichas de los analitos en las muestras de calibración fueron contrastadas por medio de una regresión respecto a los valores nominales de concentración, estimando el valor de LOD_{pu} a través de la **Ecuación 4.10**.

Si bien los valores de LOD_{min} y LOD_{max} no cambian de manera significativa de un ciclo de cálculo a otro, los valores de LOD_{pu} obtenidos en cada ciclo de adición de ruido siguen un comportamiento Gaussiano, como muestra la **Figura 4.4** en dos casos típicos. Las medias de las distribuciones de LOD_{pu} se comparan en la **Tabla 4.1** con el valor mínimo y máximo del intervalo LOD (LOD_{min} y LOD_{max}) en varios casos diferentes. Resulta interesante notar que la distribución del LOD_{pu} se encuentra centrada en el límite inferior LOD_{min} del intervalo LOD propuesto, cuando el ruido en las concentraciones de calibrado es despreciable comparado al nivel de ruido en señales instrumentales (**Tabla 4.1** y **Figura 4.4 A**). Este resultado se puede explicar por medio de los siguientes hechos teniendo en cuenta la estimación del LOD_{pu}: (1) la variancia de los residuos de regresión pseudounivariados σ^2_{pu} aproxima $(SEN^{-2}\sigma^2_x)$[13] y (2) se estima que la pendiente de la regresión s_{pu} sea cercana a 1. La introducción de estos parámetros en la **Ecuación 4.10** lleva a un LOD_{pu} idéntico al LOD_{min} (dado por la **Ecuación 4.8** con $\sigma^2_{y_{cal}}=0$ y leva efectiva $(h_{0min} + 1/I)$).

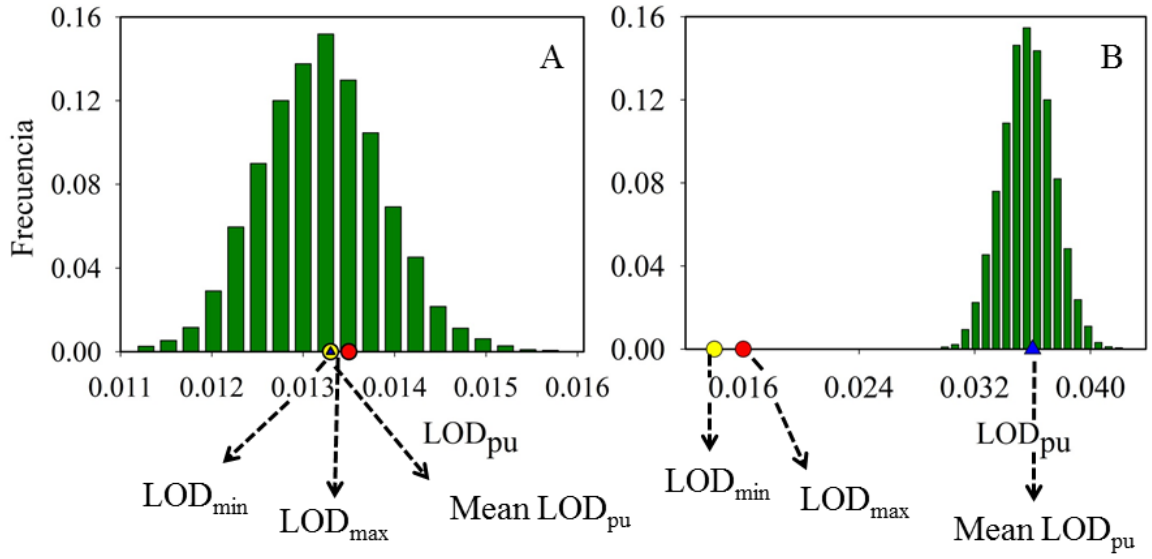


Figura 4.4 Distribución de los histogramas de valores de LOD_{pu} luego de realizar varios cálculos de adición de ruido en un conjunto de datos simulados típico, para incertidumbres despreciables (A) y finitas (B) en concentraciones de calibración. Se muestran también las relaciones entre el valor medio de los distintos LOD_{pu}, el LOD_{min} y el LOD_{max}. Las incertidumbres utilizadas en (A) y (B) son: en concentración 0 y 0.01 y en señal, 0.01 en ambos casos, respectivamente.

Tabla 4.1. Comparación de los valores de LOD en un sistema simulado.[a]

Incertidumbre en señales instrumentales	Incertidumbre en concentraciones de calibración	LOD_{pu}promedio	LOD_{min}/LOD_{max}
0.005	0	0.0067	0.0067/0.0069
0	0.005	0.017	0.0033/0.0052
0.005	0.005	0.018	0.0075/0.0086
0.01	0	0.013	0.013/0.014
0	0.01	0.033	0.0047/0.0073
0.01	0.01	0.036	0.014/0.016
0.008	0.001	0.0106	0.0106/0.0108

[a] Todos los valores están dados en unidades arbitrarias de concentración y senal

Por otro lado, cuando la incertidumbre en concentración compite con el ruido instrumental en valores relativos, la relación mutua entre LOD_{pu}, LOD_{min}, y LOD_{max} es menos clara. Como se muestra en la **Tabla 4.1** y se ilustra en la Figura 4.4 B, el valor de LOD_{pu} pude ser incluso mayor que el límite superior LOD_{max}. Esto puede explicarse sobre

la base de cómo los errores en las concentraciones de calibración se incorporan a la definición de LOD. En la estimación tanto de LOD_{min} como de LOD_{max}, el término que tiene en cuenta la contribución de los errores en concentraciones se encuentra escalado por la leva, pero en el caso de LOD_{pu} se incorpora directamente en el primer término dependiente de la muestra de *test* de la expresión para el cálculo del LOD. En este último caso, la señal se reemplaza por las concentraciones estimadas y, por lo tanto, los errores se propagan directamente a la desviación estándar en las concentraciones predichas. En cualquier caso, la aproximación conceptual al LOD_{pu} es radicalmente distinta a la del rango de valores de LOD que, en principio, debería llevar a un mejor entendimiento en lo que respecta a la capacidad de detección de PLS.

Datos experimentales

En todos los sistemas experimentales, los modelos PLS se construyeron tal como se informa en la literatura,[115-119] utilizando un número de muestras de calibración y variables latentes como los que se indican en la **Tabla 4.2**. Los valores de LOD_{pu} se estimaron tal como se describió más arriba, a partir del gráfico pseudounivariado de concentraciones estimadas vs. nominales (o medidas, dependiendo del sistema) en el conjunto de muestras de calibración. Para la estimación del intervalo LOD propuesto en el presente trabajo, se utilizaron las **Ecuaciones 4.8** y **4.9**, insertando los valores apropiados de los siguientes parámetros: (1) sensibilidad, como la inversa de la norma de los coeficientes de regresión obtenidos por el modelo PLS, (2) los valores mínimo y máximo de las levas efectivas ($h_{0min} + 1/I$) y ($h_{0max} + 1/I$), dado que los datos se centraron. La variancia del erro en la señal espectral se estimó a partir de los residuos espectrales promedio obtenidos cuando se modelan las muestras de *test* (**Tabla 4.2**).

Teniendo en cuenta las variancias del error en las concentraciones, cuando las muestras de calibración se preparan a partir de los estándares de analito, las incertidumbres normalmente son conocidas por parte del analista por medio de un análisis de propagación de errores. Esto se da en las primeros 5 conjuntos de muestras de la (**Tabla 4.2**). En la última entrada de la tabla, por otra parte, los valores de humedad se midieron por una técnica de referencia, y la incertidumbre en principio debería estimarse por medio de un análisis de réplicas. En ausencia de información, se puede utilizar la incertidumbre media a la hora de predecir las concentraciones de calibración utilizando el modelo PLS. Esta discusión resalta la necesidad de estimar las incertidumbres en las concentraciones

utilizadas en la calibración de una manera confiable (tanto por medida de réplicas como por propagación de errores), ya que constituyen un aspecto fundamental en los cálculos de LOD actuales.

Como puede verse en los primeros 5 casos de la **Tabla 4.2**, los valores de LOD_{pu} son mayores que el valor máximo LOD_{max} del rango de LOD propuesto. Esto probablemente se deba a que en estos casos los errores en concentraciones de calibración son relevantes, al igual que en la mayoría de los sistemas analíticos, y coincide con las conclusiones alcanzadas durante el estudio de simulación. En el caso de la calibración para determinar humedad en semillas de maíz (última entrada de la **Tabla 4.2**), los valores de referencia fueron medidos por un método gravimétrico muy preciso. En situaciones de incertidumbres en concentración pequeñas, el LOD_{pu} aproxima al LOD_{min}, en concordancia con los resultados de las simulaciones.

Tabla 4.2. Comparación de los valores de LOD en sistemas experimentales.[a]

Sistema	Fluoruro en aguas naturales	DINOSEB en una mezcla de reacción	Bromhexina en suero	Tetraciclina en suero	Biodiesel en combustible	Humedad en maíz
Espectro	UV-visible	UV-visible	UV-visible	Fluorescencia sincrónica	NIR	NIR
Rango de concent.	0-1.4 mg L^{-1}	0-261 mg L^{-1}	1.55-2.66×10^{-4} mol L^{-1}	0-4 mg L^{-1}	0-20 %	9.4-10.9 %
I	36	10	12	50	48	50
A	4	2	3	4	11	13
σ_X	0.001	0.001	0.006	3	0.001	0.001
σ_{ycal}	0.01	0.3	1×10^{-6}	0.1	0.01	0.005
LOD_{pu}	0.18	1.7	0.065	0.30	2.8	0.080
LOD_{min}	0.028	0.47	0.053	0.13	0.74	0.080
LOD_{max}	0.040	0.77	0.057	0.20	1.1	0.081

[a] I = número de muestras de calibración. A = número de variables latentes de PLS. Todos los valores de LOD están dados en las mismas unidades que el rango correspondiente de concentraciones.La incertidumbre en la señal σ_x está dada en unidades de absorbancia, excepto por la tetraciclina en suero, que se encuentra dada en unidades arbitrarias de fluorescencia. La incertidumbre en concentraciones se encuentra en las mismas unidades que los rangos de concentración correspondientes.

El caso correspondiente a la tetraciclina detectada en suero humano (**Tabla 4.2**) merece especial atención. En la Referencia 46 se utilizó un procedimiento experimental bastante tedioso, preparando un juego de muestras experimentales con varias

concentraciones de analitos cercanas al valor de LOD esperado. Se llevó a cabo un análisis estadístico detallado para detectar las concentraciones de analitos para las cuales las concentraciones predichas eran estadísticamente diferentes de 0. El valor informado fue de ca. 0.30 mg L^{-1},[118] el cual puede ser comparado favorablemente con los límites del intervalo LOD que se muestran en la **Tabla 4.2**. Esto implicaría que el LOD para este modelo PLS puede estimarse adecuadamente partiendo del conjunto de calibrado, sin la necesidad de preparar un conjunto de muestras adicional con bajas concentraciones del analito de interés.

4.11 Conclusión

Durante este capítulo se estudió una nueva manera de calcular el límite de detección cuando se utiliza el modelo PLS, analizando resultados obtenidos tanto con juegos de datos simulados como experimentales. El método propuesto está basado en un análisis geométrico de la definición de leva multivariada en el espacio de las variables latentes, y combina criterios matemáticos y analíticos, generando una nueva forma de estimar el límite de detección como un intervalo de detección. Esta propuesta representa un balance razonable entre las dos tendencias actuales respecto del cálculo del límite de detección multivariado: una que intenta calcular un límite de detección dependiente de cada muestra basada en el modelo EIV, y la otra que busca extender la definición univariada ISO/IUPAC proponiendo un único valor de LOD para un determinado modelo de calibración. El estimador propuesto, en principio, podría extenderse fácilmente a otros métodos inversos multivariados, aunque todavía son necesarios más estudios para aplicarlos a datos multi-vía más complejos.

4.12 Apéndice

En este apéndice, se realizará la derivación de algunos de los resultados importantes en lo que respecta a la propuesta de intervalos de límite de detección realizada en la presente sección. En primer lugar, es importante resaltar que las levas son distancias al cuadrado en un espacio de *scores* en el que estos últimos se encuentran normalizados. Esto implica que cada elemento de score t_a se multiplica por un factor f_a, que es el a-ésimo elemento diagonal de una matriz cuadrada de $A \times A$ que se calcula como $(\mathbf{T}^T \mathbf{T})^{-1/2}$ (donde T es la matriz de *scores* de calibración). En lo que sigue, se llamarán a los vectores de *scores* normalizados como $\mathbf{t}_N$ en el caso de una muestra genérica, $\mathbf{t}_{Ncal}$ para muestras de

calibración, y $\mathbf{t}_{N0cal}$ para la proyección de una muestra de calibración perpendicular al plano π_0 definido por la concentración 0 del analito. Los elementos específicos de este vector se denominarán $t_{a\mathrm{N}}$, $t_{a\mathrm{Ncal}}$, y $t_{a\mathrm{N0cal}}$, respectivamente.

La expresión que define π_0 en el espacio de los *scores*, suponiendo que tanto las señales como las concentraciones están centradas, es:

$$\pi_0: \mathbf{tv} + \bar{y}_{\mathrm{cal}} = 0 \qquad \text{(A4-1)}$$

que puede escribirse en términos de los *scores* normalizados de la siguiente manera:

$$\sum_{a=1}^{A} \frac{v_a t_{a\mathrm{N}}}{f_a \bar{y}_{\mathrm{cal}}} = -1 \qquad \text{(A4-2)}$$

Una muestra de calibrado localizada en $\mathbf{t}_{Ncal}$ puede proyectarse perpendicularmente a π_0 a través de la recta paramétrica:

$$t_{a\mathrm{N}} = -\frac{v_a}{f_a \bar{y}_{\mathrm{cal}}} k + t_{a\mathrm{Ncal}} \qquad \text{(A4-3)}$$

donde k es un parámetro variable. La intersección de esta última línea con π_0 se da en el siguiente punto:

$$\sum_{a=1}^{A} k \left(\frac{v_a}{f_a \bar{y}_{\mathrm{cal}}} \right)^2 - \frac{v_a t_{a\mathrm{Ncal}}}{f_a \bar{y}_{\mathrm{cal}}} = 1 \qquad \text{(A4-4)}$$

de donde k puede calcularse como:

$$k = \frac{\bar{y}_{\mathrm{cal}}^2 + \bar{y}_{\mathrm{cal}} \sum_{a=1}^{A} \frac{v_a t_{a\mathrm{Ncal}}}{f_a}}{\sum_{a=1}^{A} \left(\frac{v_a}{f_a} \right)^2} \qquad \text{(A4-5)}$$

Por lo tanto, una coordenada genérica para el punto de intersección es:

$$t_{a\mathrm{N0cal}} = -\left(\frac{v_a}{f_a} \right) \frac{\bar{y}_{\mathrm{cal}} + \sum_{a=1}^{A} \frac{v_a t_{a\mathrm{Ncal}}}{f_a}}{\sum_{a=1}^{A} \left(\frac{v_a}{f_a} \right)^2} + t_{a\mathrm{Ncal}} \qquad \text{(A4-6)}$$

Dado que el valor de $\left(\sum_{a=1}^{A}\frac{v_a t_{a\mathrm{Ncal}}}{f_a}\right)$ es igual a la concentración centrada de una determinada muestra de calibrado, la **Ecuación A4-6** se puede reordenar de la siguiente manera:

$$t_{a\mathrm{N0cal}} = -\frac{v_a(\bar{y}_{\mathrm{cal}} + y_{\mathrm{cal}})}{f_a\sum_{a=1}^{A}\left(\frac{v_a}{f_a}\right)^2} + t_{a\mathrm{Ncal}} \qquad \text{(A4-7)}$$

En la **Ecuación A4-7**, $\sum_{a=1}^{A}\left(\frac{v_a}{f_a}\right)^2$ se puede expresar en función de las concentraciones de calibración si se tiene en cuenta que las columnas $\mathbf{t}_a$ de la matriz **T** son ortogonales ($\mathbf{t}_a^{\mathrm{T}}\mathbf{t}_{a'} = \sum_{i=1}^{I} t_{ia}t_{ia'} = 0$):

$$\sum_{a=1}^{A}\left(\frac{v_a}{f_a}\right)^2 = \sum_{a=1}^{A} v_a^2\mathbf{t}_a^{\mathrm{T}}\mathbf{t}_a = \sum_{a=1}^{A}(v_a\sum_{i=1}^{I}t_{ia}^2) = \sum_{i=1}^{I}\left(\sum_{a=1}^{A}v_a t_{ia}\right)^2 \approx \sum_{i=1}^{I} y_i^2 \qquad \text{(A4-8)}$$

donde y_i es la concentración centrada para la i-ésima muestra de calibración, estimada a partir del producto de los coeficientes de regresión v_a y los *scores* de la muestra.

Si se define la mínima leva proyectada $h_{0\mathrm{min}}$ como la expresión conocida de la leva pseudounivariada para una muestra blanco,

$$h_{0\mathrm{min}} \approx \frac{\bar{y}_{\mathrm{cal}}^2}{\sum_{i=1}^{I} y_i^2} \qquad \text{(A4-9)}$$

se puede transformar la expresión A4-7 en la siguiente expresión simplificada:

$$t_{a\mathrm{N0cal}} = -\frac{v_a(\bar{y}_{\mathrm{cal}} + y_{\mathrm{cal}})}{f_a\bar{y}_{\mathrm{cal}}^2}h_{0\mathrm{min}} + t_{a\mathrm{Ncal}} \qquad \text{(A4-10)}$$

La raíz cuadrada de la norma del vector $\mathbf{t}_{\mathrm{N0cal}}$ (con coordenadas dadas en la **Ecuación A4-10**) es la leva de la muestra a concentración 0 del analito, proyectada hipotéticamente perpendicular a π_0. De las expresiones anteriores se puede demostrar que:

$$h_{0\text{cal}} = h_{\text{cal}} + h_{0\text{min}}\left[1 - \left(\frac{y_{\text{cal}}}{\bar{y}_{\text{cal}}}\right)^2\right] \qquad \text{(A4-11)}$$

donde h_{cal} es la leva de la muestra de calibración e y_{cal} se encuentra centrada. Se puede ver fácilmente que en el centro de la calibración, donde tanto h_{cal} e y_{cal} son cero, se obtiene la mínima proyección a π_0 (es decir, $h_{0\text{cal}}$=$h_{0\text{min}}$), lo que explica el nombre $h_{0\text{min}}$ en la **Ecuación A4-9**.

Es interesante notar que la **Ecuación A4-11** puede derivarse de argumentos trigonométricos simples:

$$h_{0\text{cal}} = h_{0\text{min}} + Q^2 = h_{0\text{min}} + (h_{\text{cal}} - M^2) \qquad \text{(A4-12)}$$

donde los segmentos M y Q se definen en la **Figura A4**. En esta figura, las levas se interpretan como distancias al cuadrado proporcionales a la concentración, por lo que puede demostrarse que $M^2 = h_{0\,\text{min}}\left(\frac{y_{\text{cal}} - \bar{y}_{\text{cal}}}{\bar{y}_{\text{cal}}}\right)^2$, y la **Ecuación A4-11** sigue inmediatamente a la A4-12.

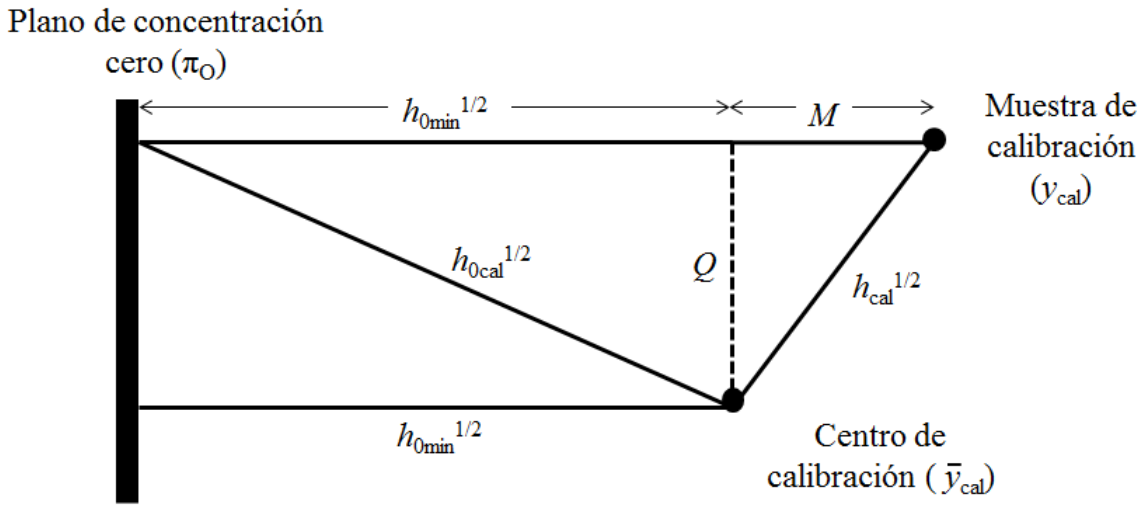


Figura A4. Representación esquemática de los parámetros relativos a la leva que se utilizan en este trabajo. La línea gruesa negra representa la proyección del plano π_0, los círculos negros indican la localización del centro de calibrado (concentración del analito = $\bar{y}_{cal}$) y de una determinada muestra de calibración (concentración del analito = y_{cal}). También se muestran las distancias adicionales (raíz cuadrada de los valores de leva) en el espacio de los *scores*.

Es importante tener en cuenta que todas las expresiones para el cálculo de la leva discutidas con anterioridad corresponden a datos centrados por la media (tanto en señales como en concentraciones). Antes de insertar cualquiera de estas levas, en particular los valores mínimos y máximos h_{0min} y h_{0max}, en la expresión correspondiente para calcular la incertidumbre en la concentración, deben convertirse en levas efectivas, es decir, (h_{0min}+1/I) y h_{0max} + 1/I).

4.13 Perspectivas

En la Sección 4.4 de este capítulo se mencionó que el estimador para el cálculo del LOD se basa en suponer que la estructura de error que afecta al sistema en estudio es de tipo iid. En caso que este supuesto se cumpla, la suma de cuadrados de los residuos espectrales, constituye una buena aproximación de la variancia del error en señal y por lo tanto su valor puede utilizarse en la **Ecuaciones 4.8** y **4.9** para calcular el LOD_{min} y el LOD_{max} a partir de las levas mínima y máxima extrapoladas a concentración 0. Sin embargo, como se vio durante el Capítulo 3, el supuesto iid no siempre se cumple. Por lo tanto, lo anterior conduce a la necesidad y consecuente posibilidad de extender las ecuaciones para el estimador propuesto considerando los distintos casos tratados durante el

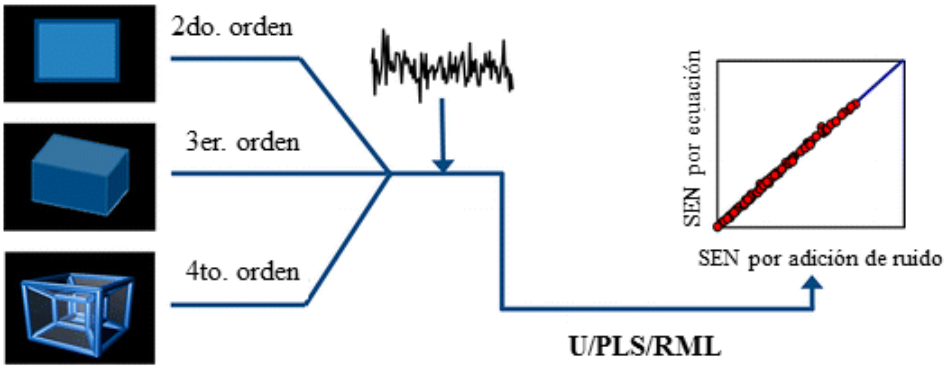
Capítulo 3. Esto implicaría de alguna manera, extrapolar las matrices de covariancia del error cuando la concentración del analito de interés es igual a 0.

En la situación descripta para el caso 2 de la **Tabla 3.2** del Capítulo 3, es decir cuando todas las muestras tienen una misma estructura de error y por lo tanto pueden representarse por la misma matriz de covariancia del error, esta extensión sería relativamente sencilla. Consistiría simplemente en multiplicar h_{0min} y h_{0max} en el término 2, manteniendo la misma matriz de covariancia del error, bajo el supuesto lógico de que ésta no se modifica cunado se realiza la extrapolación a concentración 0. Es decir, en este caso, las muestras del conjunto de calibración que darían lugar al LOD_{min} y al LOD_{max} serían las mismas que en el caso 1.

Sin embargo para el caso 3 la situación es más compleja ya que, por ejemplo, si el ruido es proporcional la extrapolación a concentración 0 de la matriz variancia covariancia del error que dará lugar al menor y mayor desvío estándar del blanco no se corresponde necesariamente con las muestras que tienen las levas mínimas o máximas a concentración cero, sino con aquellas cuyos *scores* a concentración cero dan lugar a una reconstrucción espectral tal que la matriz de covariancia del error efectiva genere un mínimo o un máximo en el segundo término de la **Tabla 3.2**.

CAPÍTULO 5

CÁLCULO DE LA SENSIBILIDAD EN CALIBRACIÓN MULTI VÍA CUANDO SE UTILIZA EL MODELO U-PLS/RML



"Dónde hay más sensibilidad, allí es más fuerte el martirio." (Leonardo Da Vinci).

5.1 Resumen

En este capítulo se presentará el desarrollo de una nueva expresión que permite calcular la sensibilidad cuando se utiliza el algoritmo PLS desdoblado y acoplado a cuadrados mínimos parciales U-PLS/RML. Este estudio, sumado a los desarrollos realizados sobre PARAFAC y MCR-ALS, completa el esquema de expresiones para el cálculo de sensibilidad de los algoritmos de calibración multivariada más utilizados. En un contexto de ruido homoscedástico, la sensibilidad puede ser utilizada para calcular otras cifras de mérito relevantes como la sensibilidad analítica, el límite de detección, el límite de cuantificación y la incertidumbre en la concentración predicha. Los resultados se sustentan en simulaciones de adición de ruido que tienen en cuenta una extensa variedad de sistemas con distinto número de analitos y agentes interferentes, diferentes grados de solapamiento entre los perfiles de los componentes, y diferente número de modos de datos instrumentales por muestra, de manera tal que en todos los casos se necesita de los beneficios de la ventaja de segundo orden. También se discutirá la conexión entre la presente propuesta basada en propagación de errores y el concepto intuitivo de señal analítica neta. Además, se incluye un ejemplo experimental para el cual se estudian datos

de segundo, tercer y cuarto orden, teniendo en cuenta el mejoramiento en las cifras de mérito al incrementar el orden de los datos, lo cual es consistente con una disminución en el error de predicción promedio.

5.2 Introducción

Como se evidenció en los capítulos anteriores, cuando PLS se aplica a calibración multivariada de primer orden, se requiere de una cantidad de muestras de calibración lo suficientemente grande como para construir un modelo que tenga en cuenta los potenciales interferentes que pueden aparecer en nuevas muestras. En este sentido, resulta interesante la extensión de PLS a calibración multivariada de orden superior, es decir cuando los datos corresponden a arreglos con dos o más modos instrumentales o vías para cada muestra. Esto se debe a que en esta situación es posible determinar el analito de interés utilizando un conjunto de calibrado en el que sólo el analito puro se encuentra presente, incluso cuando las nuevas muestras contengan agentes interferentes no calibrados. Esta virtud que poseen algunos modelos multi-vía se presentó durante el primer capítulo, y normalmente se conoce como "ventaja de segundo orden". [2,120-126]

En principio, el algoritmo PLS puede aplicarse a datos de orden superior de dos maneras distintas: (1) desplegando los datos originales y generando vectores sobre los cuales se aplica un análisis PLS clásico, dando lugar a lo que se conoce como modelo U-PLS,[64] (2) procesando los arreglos de datos originales utilizando una versión multidimensional de PLS llamada N-PLS.[64] Sin embargo, ninguna de estas estrategias permite alcanzar la ventaja de segundo orden por sí sola. Para alcanzar este último objetivo, en 1990 se desarrolló la bilinealización residual (RBL) y se combinó con PLS para analizar datos de segundo orden.[127] Luego de un lapso de aproximadamente 15 años, RBL fue redescubierto,[128] y aplicado a una amplia variedad de sistemas analíticos, mostrando una gran potencialidad debido a la flexibilidad de su modelo basado en la generación de variables latentes.[122] Seguidamente, se desarrollaron extensiones de RBL para analizar datos de tercer y cuarto orden por medio de los procedimientos de trilinealización residual (RTL)[49] y cuadrilinealización residual (RQL).[50] Todos estos métodos son miembros de una familia colectivamente conocida como multilinealización residual (RML), que se combinaron con U-PLS, N-PLS y otros métodos multivariados. Existen aplicaciones a datos de varios órdenes, medidos para muestras de composición compleja y diversos orígenes, que se revisaron y resumieron en varios trabajos.[120,122-127]

Como se mostró en detalle durante los capítulos anteriores de esta tesis, las cifras de mérito pueden estimarse de manera confiable tanto en calibración de orden cero como en calibración de primer orden. Incluso gran parte de estos estimadores se encuentran documentados en informes técnicos de la IUPAC.[5,110] Más específicamente en el caso de PLS, como fue detallado durante el segundo capítulo, muchos trabajos se focalizaron en desarrollar expresiones que permitieran calcular la incertidumbre en concentraciones predichas de manera específica para cada muestra.[99,129-134] En estas expresiones, una cifra de mérito de importancia fundamental es la sensibilidad, que a su vez constituye un elemento crucial a la hora de estimar otros parámetros importantes como la sensibilidad analítica, la selectividad, el límite de detección y el límite de cuantificación.[135] De manera general, la sensibilidad en química analítica puede definirse como el cambio en la respuesta neta para un determinado cambio en concentración. De acuerdo con lo tratado en capítulos anteriores, en calibración univariada esta sensibilidad es numéricamente igual a la pendiente de la curva de calibrado[5]. Por otro lado, en calibración multivariada de primer orden, normalmente se define la sensibilidad como la pendiente calibración pseudounivariada basada en lo que se conoce como señal analítica neta (NAS), que es la porción de la señal total que corresponde únicamente al analito de interés.[136] Lo anterior se representa esquemáticamente en la **Figura 5.1**. Esta interpretación, si bien surge a partir del modelo CLS es perfectamente compatible con la definición de sensibilidad dada para PLS y PCR en capítulos anteriores, como la inversa del cuadrado de la norma de los coeficientes de regresión obtenidos por el modelo.

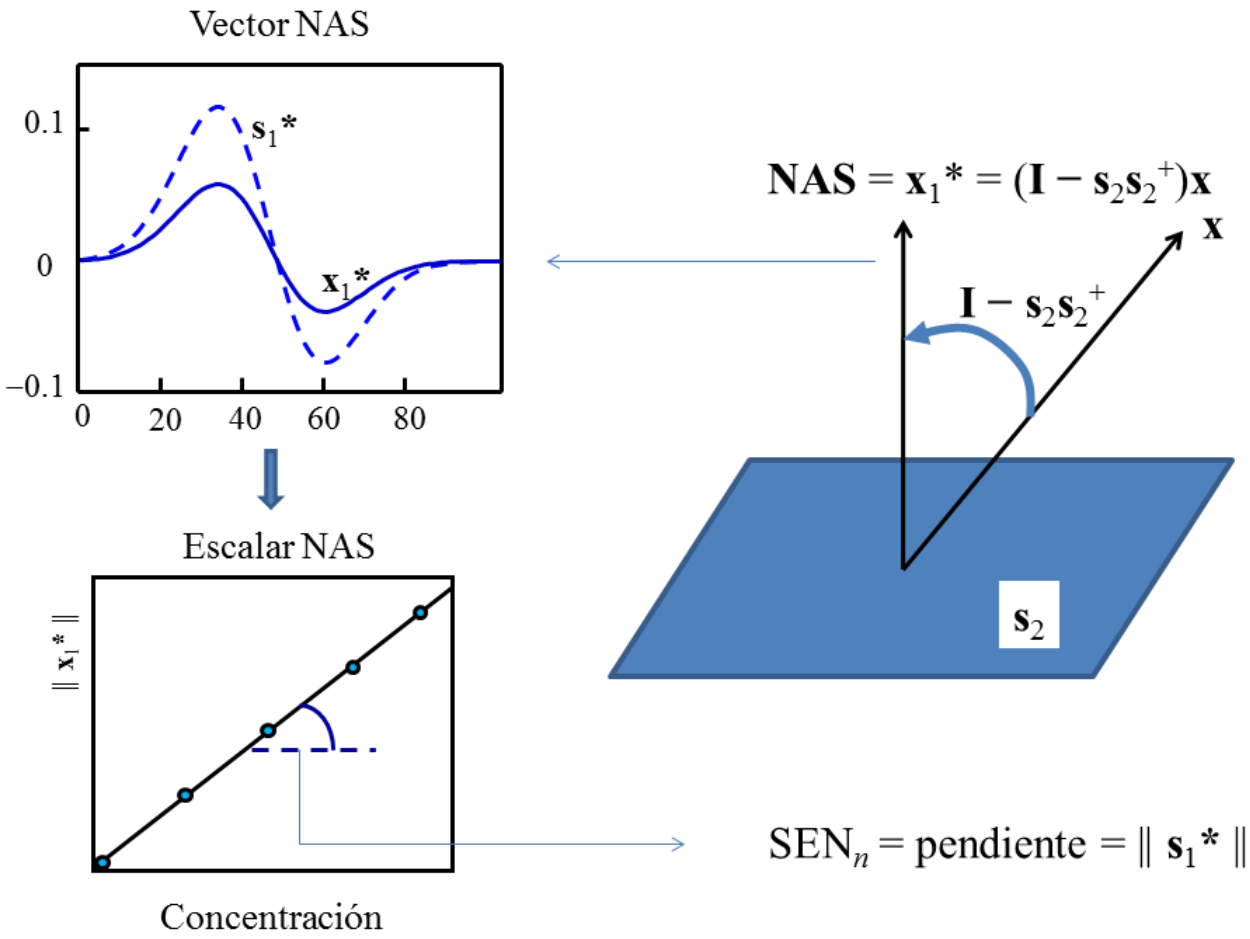


Figura 5.1. Representación gráfica del concepto de señal analítica neta. El espectro de una muestra (**x**) se proyecta ortogonalmente al espacio definido por el resto de los componentes de la muestra ($\mathbf{s}_2$), dando un valor para la señal neta del analito ($\mathbf{x}_1^*$). La norma de este vector da lugar a un escalar NAS que puede graficarse en función de la concentración para llegar a un valor de sensibilidad (SEN).

Hasta el momento, en calibración de segundo orden, se desarrollaron diferentes definiciones de NAS en el marco de uno de los modelos más utilizados en este marco, como lo es PARAFAC. [8,9] Estas definiciones alternativas iniciales, mostraron ser casos especiales de otra expresión más general que reveló las dificultades del concepto de NAS en el escenario de segundo orden.[11] La extensión de la aproximación NAS a datos de tercer orden es inclusive más problemática,[13] aunque recientemente se desarrolló una expresión mejorada para el cálculo de la sensibilidad en PARAFAC. Como se detallará más adelante, en este último caso, se utilizó una nueva aproximación basada en principios de

propagación de errores que no incluye, al menos de manera explícita, argumentos basados en el concepto de NAS. Estos mismos principios, también se utilizaron para llegar a una fórmula para el cálculo de la sensibilidad en el caso de MCR-ALS, que es otro de los modelos más utilizados en calibración analítica multivariada.

En lo que respecta a las metodologías de tipo PLS/RML, hasta ahora sólo se conoce una expresión provisoria de sensibilidad para el caso de PLS/RBL,[48] basada en el concepto controvertido de NAS. Sin embargo, no existen expresiones para el resto de las cifras de mérito en PLS/RBL. Lo mismo ocurre con las extensiones de tercer y cuarto orden PLS/RTL y PLS/RQL. Como respuesta a este panorama incompleto, este capítulo se centrará en el desarrollo de expresiones para el cálculo de la sensibilidad en toda la familia de métodos de calibración PLS/RML. El propósito de la temática de este capítulo es doble: por un lado proveer a los químicos analíticos un conjunto completo de cifras de mérito para informar resultados significativos derivados de los análisis por medio de algoritmos PLS de orden superior, y por otro lado, obtener una única expresión matemática que se pueda aplicar a todas las metodologías PLS/RML.

Para completar el capítulo se presentará un ejemplo experimental que se puede adaptar para generar datos de segundo, tercer y cuarto orden, permitiendo ilustrar aplicaciones reales de las expresiones desarrolladas y demostrar la mejora en las cifras de mérito que se produce cuando se incrementa el orden de los datos.

5.3 Objetivos específicos

1) Analizar los avances realizados en la determinación de la sensibilidad para sistemas de más de dos vías, cuando se utiliza cada uno de los tres algoritmos fundamentales de la calibración analítica multivariada.

2) Generar simulaciones de adición de ruido para estudiar exhaustivamente la factibilidad de las nuevas expresiones propuestas.

3) Partiendo de las expresiones propuestas recientemente en PARAFAC y MCR-ALS para calcular la sensibilidad, proponer una nueva aproximación que permita obtener la sensibilidad cuando se emplea el modelo U-PLS/RML.

5.4 Antecedentes

5.4.1 Cálculo de la sensibilidad en PARAFAC

Aunque en el campo de los datos de tres vías se utilizó el concepto de NAS desde un principio, pronto se encontraron dificultades en su interpretación, ya que surgieron definiciones conflictivas. En el marco del modelo PARAFAC, se postularon dos definiciones rivales de sensibilidad. Una de ellas, postulada por Messick, Kalivas y Lang (MKL)[9] y otra por Ho, Christian y Davidson (HCD).[8] Estas expresiones llevaban a distintos valores, sin que se pudiera establecer el verdadero motivo. Este tema fue resuelto demostrando que tanto MKL como HCD son casos especiales de una definición más general de la sensibilidad en PARAFAC (FO, Faber y Olivieri).

Para la discusión que sigue, es importante resaltar que la descomposición de un conjunto de datos de tres vías típico por medio de PARAFAC, genera los perfiles de los componentes a partir de dos modos instrumentales que se almacenan en las matrices **B** y **C**, así como también a partir de los *scores* del analito, los cuales se utilizan luego en un gráfico de calibración pseudo-univariada para predecir las concentraciones. En este contexto, la sensibilidad FO está dada por:

$$\mathrm{SEN_{FO}}=s_n\left\{\left[\left(\mathbf{B}_{\mathrm{cal}}^{\mathrm{T}}\mathbf{P}_{\mathrm{B,int}}\mathbf{B}_{\mathrm{cal}}\right)^{*}\left(\mathbf{C}_{\mathrm{cal}}^{\mathrm{T}}\mathbf{P}_{\mathrm{C,int}}\mathbf{C}_{\mathrm{cal}}\right)\right]^{-1}\right\}_{nn}^{-1/2} \qquad (5.1)$$

donde s_n es la señal del analito puro a concentración unitaria (es decir, la pendiente de la curva de calibración pseudo-univariada), $\mathbf{B}_{\mathrm{cal}}$ y $\mathbf{C}_{\mathrm{cal}}$son las submatrices de **B** y **C** que contienen los *loadings* para los analitos calibrados en cada uno de los modos instrumentales, '*´ es el producto matricial "elemento a elemento" o producto de Hadamard, el subíndice *'nn´* se utiliza para denotar el *n*-ésimo elemento diagonal de la matriz, y $\mathbf{P}_{\mathrm{B,unx}}$ y $\mathbf{P}_{\mathrm{C,unx}}$ son matrices de proyección dadas por:

$$\mathbf{P}_{\mathrm{B,int}}=\mathbf{I}-\mathbf{B}_{\mathrm{int}}\mathbf{B}_{\mathrm{int}}^{+} \qquad (5.2)$$

$$\mathbf{P}_{\mathrm{C,int}}=\mathbf{I}-\mathbf{C}_{\mathrm{int}}\mathbf{C}_{\mathrm{int}}^{+} \qquad (5.3)$$

donde **I** representa matrices identidad dimensionadas adecuadamente, $\mathbf{B}_{\mathrm{int}}$ y $\mathbf{C}_{\mathrm{int}}$ recogen los *loadings* para los interferentes potenciales, y el supraíndice '+´ indica la operación inversa generalizada. En ausencia de interferentes potenciales, las matrices de proyección

de las **Ecuaciones 5.2** y **5.3** son simplemente matrices unitarias, y la **Ecuación 5.4** se reduce a la ecuación MKL:

$$\mathrm{SEN}_{\mathrm{MKL}}=s_n\left\{\left[\left(\mathbf{B}_{\mathrm{cal}}^{\mathrm{T}}\mathbf{B}_{\mathrm{cal}}\right)^{*}\left(\mathbf{C}_{\mathrm{cal}}^{\mathrm{T}}\mathbf{C}_{\mathrm{cal}}\right)\right]^{-1}\right\}_{nn}^{-1/2} \qquad (5.4)$$

Por otro lado, cuando se calibra un único analito en presencia de potenciales interferentes, la **Ecuación 5.1** se reduce a la expresión HCD:

$$\mathrm{SEN}_{\mathrm{HCD}}=s_n\left\{\left[\left(\mathbf{B}^{\mathrm{T}}\mathbf{B}\right)^{-1}\right]_{nn}\left[\left(\mathbf{C}^{\mathrm{T}}\mathbf{C}\right)^{-1}\right]_{nn}\right\}^{-1/2} \qquad (5.5)$$

El ajuste de la **Ecuación 5.4** se corroboró a través de numerosas simulaciones de adición de ruido, obteniéndose valores realistas en sistemas experimentales.

Hasta el momento, no se conocían expresiones generales para calcular la incertidumbre en la concentración predicha del analito en PARAFAC de cuatro vías. Sin embargo, se habían desarrollado varias conjeturas en este contexto: una de ellas implica una extrapolación directa de los resultados obtenidos en PARAFAC de tres vías. En el caso general, esto lleva a la siguiente expresión:

$$\mathrm{SEN}_{\mathrm{FO4}}=s_n\left\{\left[\left(\mathbf{B}_{\mathrm{cal}}^{\mathrm{T}}\mathbf{P}_{\mathrm{B,int}}\mathbf{B}_{\mathrm{cal}}\right)^{*}\left(\mathbf{C}_{\mathrm{cal}}^{\mathrm{T}}\mathbf{P}_{\mathrm{C,int}}\mathbf{C}_{\mathrm{cal}}\right)^{*}\left(\mathbf{D}_{\mathrm{cal}}^{\mathrm{T}}\mathbf{P}_{\mathrm{D,int}}\mathbf{D}_{\mathrm{cal}}\right)\right]^{-1}\right\}_{nn}^{-1/2} \qquad (5.6)$$

donde $\mathbf{P}_{\mathrm{D,int}}$ es análoga a $\mathbf{P}_{\mathrm{B,int}}$ y $\mathbf{P}_{\mathrm{C,int}}$ en las **Ecuaciones 5.2** y **5.3**, extendidas a un número mayor de modos de medición.

La **Ecuación 5.1** se reduce tanto a la expresión de MKL4 como a la de HCD4 bajo las mismas circunstancias detalladas anteriormente en relación al análisis de PARAFAC de tres vías. En el caso de la expresión MKL4, se pudo confirmar que es correcta en ausencia de interferentes no modelados, utilizando simulaciones de adición de ruido.[13] Por su parte, HCD4 no ha mostrado correspondencia con resultados simulados[13] al igual que FO4.[15] Esto sugiere que la extensión intuitiva de la aproximación utilizada en la calibración con PARAFAC de tres vías, a PARAFAC de cuatro vías, no es tan directa como se podría anticipar.

Para resolver la problemática planteada en el párrafo anterior, Olivieri y Faber desarrollaron un procedimiento basado en un análisis detallado de la matriz Jacobiana, que

se emplea normalmente durante el procedimiento de propagación de errores tradicional.[15] Esta matriz contiene todas las derivadas parciales de la señal de la muestra de *test* desdoblada con respecto a los parámetros de PARAFAC que se determinan durante el ajuste, cuya conexión con la incertidumbre en los parámetros modelados es bien conocida. Este enfoque tiene algunas ventajas interesantes: (1) coincide completamente con la expresión general FO3 para datos de tres vías, (2) puede ser adaptado directamente a arreglos de más vías, y (3) da una explicación razonable a por qué la forma intuitiva de la **Ecuación 5.6** no es apropiada para datos de cuatro vías.

La clave de esta aproximación está en el reconocimiento de dos bloques diferentes de submatrices en la matriz Jacobiana **J**: (1) un bloque denominado $\mathbf{Z}_{int}$ correspondiente únicamente a los perfiles de los componentes inesperados, y (2) un bloque $\mathbf{Z}_{cal}$ correspondiente a los perfiles de los analitos calibrados. De esta manera:

$$\mathbf{J} = [Z_{int}|Z_{cal}] \qquad (5.7)$$

Luego de un desarrollo algebraico,[15] la sensibilidad para el analito *n* queda dada por una expresión que se parece en gran medida a la que se obtiene cuando se usa del concepto de señal neta del analito (NAS):[11]

$$SEN_{J4} = s_n \left\| n\text{-ésima fila de} (\mathbf{P}_{Zint}\mathbf{Z}_{cal})^{+} \right\|^{-1} \qquad (5.8)$$

donde $\mathbf{P}_{Zunx}$ es una matriz que describe la proyección ortogonal al espacio definido por $\mathbf{Z}_{unx}$ (el subíndice J4 indica que la sensibilidad en sistemas de cuatro vías es obtenida a través de una aproximación por Jacobiano), es decir:

$$\mathbf{P}_{Zint} = \mathbf{I} - \mathbf{Z}_{int}\mathbf{Z}_{int}^{+} \qquad (5.9)$$

5.4.2 Cálculo de la sensibilidad en MCR

En el caso de MCR-ALS, se han realizado algunos intentos para definir las cifras de mérito, basados en técnicas de remuestreo en aproximaciones por adición de ruido en simulaciones de Monte Carlo.[137] También es posible usar una aproximación experimental considerando el gráfico de calibración pseudo-univariada *scores*-concentración generado por MCR-ALS y definir parámetros análogos tomando como referencia la calibración univariada.[138] Esta última estrategia permite estimar el límite de detección y de cuantificación. Sin embargo, aunque las concentraciones analíticas son proporcionales a

los *scores* recuperados por medio del algoritmo MCR-ALS, estos últimos tienen unidades arbitrarias y no reflejan, en general, el efecto del solapamiento de los perfiles de los analitos con los de otros componentes. De tal modo, la sensibilidad no puede ser definida como la pendiente del gráfico de calibración pseudo-univariada de MCR-ALS.

En un trabajo reciente, se utilizó una definición de sensibilidad basada en un análisis de propagación de errores.[16] Esta última consiste en la inversa de la relación entre la incertidumbre en la concentración predicha respecto a la incertidumbre en la señal. [11,15] La expresión para estimar la sensibilidad en este caso está dada por:

$$\mathrm{SEN_{MCR}} = m_n \left[IJ \left(\mathbf{S}^{\mathrm{T}} \mathbf{S} \right)^{-1}_{nn} \right]^{-1/2} \qquad (5.10)$$

donde *n* es el índice para el analito de interés en una mezcla con varios componentes, m_n es la pendiente del gráfico de calibración pseudo-univariada, $\mathbf{S}^{\mathrm{T}}$ es una matriz que contiene los perfiles para todos los componentes en la dirección no aumentada de MCR, e *IJ* es el número de sensores en la matriz de datos de la muestra de *test* en la dirección aumentada de MCR.

Una conclusión interesante que se puede desprender de esta fórmula es que la sensibilidad es menor que la pendiente de la curva de calibración pseudo-univariada, y disminuye al aumentar el solapamiento entre los perfiles de los componentes de la muestra (perfiles en el modo no aumentado). Esto se mide a partir del factor de solapamiento $\left[\left(\mathbf{S}^{\mathrm{T}} \mathbf{S} \right)^{-1}_{nn} \right]^{-1/2}$.

Detalles acerca del tratamiento algebraico y estadístico necesario para arribar a las expresiones antes mencionadas pueden encontrarse en la literatura.[16]

5.5 Cálculo de la sensibilidad en U-PLS/RML

5.5.1 Aproximación por propagación de errores

Para todos los modelos U-PLS/RML presentados más arriba, la concentración del analito se predice utilizando la expresión clásica de U-PLS:

$$y = \mathbf{tv} \qquad (5.11)$$

donde **v** es el vector de los coeficientes de regresión de calibración en el espacio latente, y **t** es el vector de *scores* de la muestra, una vez que se tuvieron en cuenta las contribuciones de las interferencias por el procedimiento RML. Suponiendo que **v** es preciso, es decir, que la principal fuente de error corresponde a los elementos de **t,** el error en las concentraciones predichas estará dado por:

$$\sigma_{\hat{y}}^2 = \mathbf{v}\boldsymbol{\Sigma}_t\,\mathbf{v}^T \qquad (5.12)$$

donde $\boldsymbol{\Sigma}_t$ es la matriz de variancia-covarianciaa para los elementos del vector **t**. Esta matriz se puede estimar considerando en primer lugar un ejemplo simple en el cual hay sólo un interferente en la **Ecuación 1.50** del modelo RBL, escrita de la siguiente manera:

$$\mathbf{x} = \mathbf{P}\mathbf{t}^T + \mathbf{c}_{int1}\otimes\mathbf{b}_{int1} + \mathbf{e} \qquad (5.13)$$

Nótese que si se compara con la **Ecuación 1.50** a_{int1} no se consideró, sin que esto signifique pérdida de generalidad. Los parámetros a ajustar en la **Ecuación 5.13** son los *J* elementos de $\mathbf{b}_{int1}$, los *K* elementos de $\mathbf{c}_{int1}$ y los *A* elementos de **t**. Así, el Jacobiano asociado a estos parámetros es:

$$\mathbf{J} = [\mathbf{I}_c\otimes\mathbf{b}_{int1} \mid \mathbf{c}_{int1}\otimes\mathbf{I}_b \mid \mathbf{P}] = [\mathbf{Z}_{int} \mid \mathbf{P}] \qquad (5.14)$$

donde $\mathbf{I}_c$ y $\mathbf{I}_b$ son matrices identidad de $K\times K$ y $J\times J$ respectivamente, y el bloque submatriz correspondiente al interferente es:

$$\mathbf{Z}_{int} = [\mathbf{I}_c\otimes\mathbf{b}_{int1} \mid \mathbf{c}_{int1}\otimes\mathbf{I}_b] \qquad (5.15)$$

Esta matriz $\mathbf{Z}_{int}$ depende únicamente de las propiedades de los interferentes. De la **Ecuación 5.15**, la matriz de variancia-covariancia de $(J+K+A)\times(J+K+A)$ es, para todos los parámetros estimados:

$$\boldsymbol{\Sigma} = \sigma_x^2\,(\mathbf{J}^T\mathbf{J})^{-1} \qquad (5.16)$$

Una manera conveniente de obtener la submatriz $\boldsymbol{\Sigma}_t$ de $\boldsymbol{\Sigma}$, correspondiente a los *scores* del analito 1, incluye los siguientes pasos. En primer lugar, el último bloque de **J** (la matriz **P** en la **Ecuación 5.14**) se proyecta ortogonalmente a $\mathbf{Z}_{int}$ para dar $\mathbf{J}_t^+$ de la siguiente forma:

$$\mathbf{J}_t^+ = [(\mathbf{I} - \mathbf{Z}_{int}\mathbf{Z}_{int}^+)\,\mathbf{P}]^+ \qquad (5.17)$$

donde el supraíndice '+' indica la inversa generalizada de una matriz. La **Ecuación 5.17** proviene de las propiedades bien conocidas de la operación de inversa generalizada para matrices de bloque.[139]

Nótese que $\mathbf{J}_t^+$ es de hecho un sub-bloque de la matriz $\mathbf{J}^+$ completa.

Finalmente, la matriz de variancia covariancia estará dada por:

$$\mathbf{\Sigma}_t = \sigma_x^2 \, (\mathbf{J}_t^+)^T \mathbf{J}_t^+ = \sigma_x^2 \, [\mathbf{P}^T \, (\mathbf{I} - \mathbf{Z}_{int}\mathbf{Z}_{int}^+) \, \mathbf{P}]^{-1} \qquad (5.18)$$

que puede escribirse de esta manera ya que la matriz de proyección $(\mathbf{I} - \mathbf{Z}_{int}\mathbf{Z}_{int}^+)$ es idempotente.

Reemplazando este resultado en la **Ecuación 5.12** se llega a:

$$\sigma_{\hat{y}}^2 = \sigma_x^2 \mathbf{v}^T \, [\mathbf{P}^T \, (\mathbf{I} - \mathbf{Z}_{int}\mathbf{Z}_{int}^+) \, \mathbf{P}]^{-1} \mathbf{v} \qquad (5.19)$$

Dado que la sensibilidad puede definirse como la relación entre el error en la señal y el error en la concentración, finalmente se obtiene el resultado:

$$SEN_J = [\sigma_x^2 / \, \sigma_y^2]^{1/2} = \{\mathbf{v}^T \, [\mathbf{P}^T \, (\mathbf{I} - \mathbf{Z}_{int}\mathbf{Z}_{int}^+) \, \mathbf{P}]^{-1} \, \mathbf{v}\}^{-1/2} \qquad (5.20)$$

donde el subíndice 'J' indica la aproximación por Jacobiano. Como se muestra más abajo, esta es una ecuación completamente general, aplicable a cualquier número de interferentes y de modos instrumentales. A pesar de que la **Ecuación 5.20** se ha obtenido partiendo de un sistema simple, donde solamente aparece un único interferente en las muestras de *test*, puede extenderse fácilmente a más interferentes, adaptando acordemente la matriz $\mathbf{Z}_{int}$, como se ha mostrado recientemente.[15] Las matrices generales $\mathbf{Z}_{int}$, se encuentran en la **Tabla 5.1** para cualquier número de fuentes de interferencia. Al construir estas matrices $\mathbf{Z}_{int}$, no se necesita tener acceso a los perfiles reales de los interferentes: sus combinaciones lineales, es decir, *loadings* provistos por PCA, son suficientes, dado que estos cubren adecuadamente el espacio completo de los interferentes.

Tabla 5.1. Expresiones para el cálculo de la sensibilidad en U-PLS/RML para datos de orden creciente, derivadas de las aproximaciones por Jacobiano y por NAS.

Aproximación por Jacobiano[a]	
Expresión general SEN_J	$SEN_J = \{\mathbf{v}^T [\mathbf{P}^T (\mathbf{I} - \mathbf{Z}_{int}\mathbf{Z}_{int}^{+}) \mathbf{P}]^{-1} \mathbf{v}\}^{-1/2}$ $\mathbf{Z}_{int} = [\mathbf{Z}_{int1} \mid \mathbf{Z}_{int2} \mid \ldots \mid \mathbf{Z}_{intN}]$
Orden de los datos	Repetición del bloque $\mathbf{Z}_{intn}$ en $\mathbf{Z}_{int}$
2	$[\mathbf{I}_c \otimes \mathbf{b}_{intn} \mid \mathbf{c}_{intn} \otimes \mathbf{I}_b]$
3	$[\mathbf{I}_d \otimes \mathbf{c}_{intn} \otimes \mathbf{b}_{intn} \mid \mathbf{d}_{intn} \otimes \mathbf{I}_c \otimes \mathbf{b}_{intn} \mid \mathbf{d}_{intn} \otimes \mathbf{c}_{intn} \otimes \mathbf{I}_b]$
4	$[\mathbf{I}_e \otimes \mathbf{d}_{intn} \otimes \mathbf{c}_{intn} \otimes \mathbf{b}_{intn} \mid \mathbf{e}_{intn} \otimes \mathbf{I}_d \otimes \mathbf{c}_{intn} \otimes \mathbf{b}_{intn} \mid \mathbf{e}_{intn} \otimes \mathbf{d}_{intn} \otimes \mathbf{I}_c \otimes \mathbf{b}_{intn} \mid \mathbf{e}_{intn} \otimes \mathbf{d}_{intn} \otimes \mathbf{c}_{intn} \otimes \mathbf{I}_b]$
Aproximación por señal neta del analito[b]	
Orden de los datos	Expresión específica
2	$SEN_{NAS2} = [\mathbf{v}^T (\mathbf{P}^T(\mathbf{P}_C \otimes \mathbf{P}_B)\mathbf{P})^{-1} \mathbf{v}]^{-1/2}$
3	$SEN_{NAS3} = [\mathbf{v}^T (\mathbf{P}^T(\mathbf{P}_D \otimes \mathbf{P}_C \otimes \mathbf{P}_B)\mathbf{P})^{-1} \mathbf{v}]^{-1/2}$
4	$SEN_{NAS4} = [\mathbf{v}^T (\mathbf{P}^T(\mathbf{P}_E \otimes \mathbf{P}_D \otimes \mathbf{P}_C \otimes \mathbf{P}_B)\mathbf{P})^{-1} \mathbf{v}]^{-1/2}$

[a] Los subíndices 'int1', 'int2', 'int$_n$', 'int$N$' indican la numeración de las fuentes de interferencia, con los perfiles en cada modo de datos obtenidos durante RML como b, c, d, e, etc., dependiendo del orden de los datos y Ib, Ic, Id y Iem son matrices unitarias de tamaño $J \times J$, $K \times K$, $L \times L$ y $M \times M$ respectivamente.

[b] $\mathbf{P}_B = \mathbf{I} - \mathbf{B}_{int}\mathbf{B}_{int}^{+}$, con la matriz de *loadings* $\mathbf{B}_{int}$ conteniendo los perfiles de los interferentes en el primer modo instrumental. Las definiciones de $\mathbf{P}_C$, $\mathbf{P}_D$ y $\mathbf{P}_E$ son análogas a aquellas para $\mathbf{P}_B$.

Al generalizar la **Ecuación 5.20** para datos de un número mayor de modos instrumentales, el procedimiento descripto más arriba puede aplicarse de la misma manera, modificando únicamente $\mathbf{Z}_{int}$ para acomodar los casos correspondientes, como muestra la **Tabla 5.1**. La base para derivar estas expresiones se puede encontrar en el trabajo reciente de sensibilidad en PARAFAC aplicado a calibración de cuarto orden. En este sentido, la

metodología U-PLS/RML se puede aplicar a cualquier número de modos de datos instrumentales, con una expresión completamente general que posibilita una medida adecuada de la sensibilidad. Esta última, a su vez, permite obtener las restantes cifras de mérito.

5.5.2 Aproximación por NAS

Una aproximación previa que se ha empleado para derivar una expresión para la sensibilidad en U-PLS/RBL se basa en el concepto de señal neta del analito (NAS). Como se explicó, apunta básicamente a eliminar, de la expresión para las señales de las muestras de *test*, la contribución de las interferencias, utilizando proyecciones ortogonales apropiadas, respecto del espacio delimitado por los interferentes. Por ejemplo, en la **Ecuación 5.13** para el modelo RBL, los interferentes se podrían eliminar en principio de dos maneras diferentes. En una se proyecta el vector **x** desdoblado, ortogonalmente al espacio de los vectores desdoblados de los interferentes, mientras que la segunda incluye dos proyecciones ortogonales (izquierda y derecha) de las señales de la muestra de *test*, en forma de matriz **X,** a los espacios individuales delimitados por los perfiles de los interferentes en ambos modos instrumentales. De acuerdo con la referencia 48, únicamente la última aproximación parece arrojar resultados correctos. Esta implica definir dos matrices de proyección ortogonal:

$$\mathbf{P}_B = \mathbf{I} - \mathbf{b}_{int1}\mathbf{b}_{int1}^{+} \quad (5.21)$$

$$\mathbf{P}_C = \mathbf{I} - \mathbf{c}_{int1}\mathbf{c}_{int1}^{+} \quad (5.22)$$

y luego aplicarlas para eliminar la contribución de los interferentes de la siguientes manera:

$$\mathbf{P}_B\mathbf{X}\mathbf{P}_C = \mathbf{X}_{NAS} = \mathbf{P}_B\text{reshape}(\mathbf{P}\,\mathbf{t}^{\mathbf{T}})\,\mathbf{P}_C + \mathbf{E}_{NAS} \quad (5.23)$$

La **Ecuación 5.23** puede utilizarse para definir un modelo PLS efectivo, en el cual los interferentes se encuentran ausentes, y donde las nuevas señales no son las señales **X** puras, sino las correspondientes señales netas del analito $\mathbf{X}_{NAS}$. A partir de los argumentos ya discutidos en la Referencia 84, esta aproximación lleva a la definición de sensibilidad como:

$$SEN_{NAS2} = \{\mathbf{v}^T\,[\mathbf{P}^T\,(\mathbf{P}_C \otimes \mathbf{P}_B)\,\mathbf{P}]^{-1}\mathbf{v}\}^{-1/2} \quad (5.24)$$

donde ‘NAS2’ significa aproximación NAS a datos de segundo orden.

Resulta interesante que esta aproximación es, en principio, incorrecta, ya que ignora los errores en los coeficientes de regresión propagados a las concentraciones predichas. Utilizando la **Ecuación 5.21**, se ha demostrado que la concentración predicha del analito se encuentra dada por un modelo PLS "efectivo", es decir:

$$y = \mathbf{tv} = \mathbf{x}^{\mathbf{T}}\mathbf{P}_{\text{eff,NAS}}{}^{+}\mathbf{v} \qquad (5.25)$$

donde se considera una matriz de *loadings* efectiva $\mathbf{P}_{\text{eff,NAS}} = (\mathbf{P}_C \otimes \mathbf{P}_B)\ \mathbf{P}$. Suponiendo que tanto $\mathbf{P}_{\text{eff}}{}^{+}$ como $\mathbf{v}$ son precisos en las **Ecuación 5.23**, la teoría de propagación de errores lleva a:

$$\sigma_{\hat{y}}^2 = \sigma_x^2\ \mathbf{v}^{\text{T}}\ (\mathbf{P}_{\text{eff,NAS}}{}^{\text{T}}\mathbf{P}_{\text{eff,NAS}})^{-1}\mathbf{v} \qquad (5.26)$$

y finalmente a la sensibilidad como:

$$\text{SEN}_{\text{NAS2}} = [\mathbf{v}^{\text{T}}\ (\mathbf{P}_{\text{eff,NAS}}{}^{\text{T}}\mathbf{P}_{\text{eff,NAS}})^{-1}\mathbf{v}]^{-1/2} \qquad (5.27)$$

que es idéntico a la **Ecuación 5.24**, y (ver más abajo) es también igual a SEN_J en la **Ecuación 5.20** cuando se consideran datos de segundo orden.

El principal error en la inferencia anterior se encuentra en asumir que $\mathbf{P}_{\text{eff}}$ es preciso, debido a que su cálculo incluye $\mathbf{P}$, el cual sí es preciso porque se construye a través del modelo de calibración. Sin embargo, los perfiles $\mathbf{b}_{\text{int1}}$ y $\mathbf{c}_{\text{int1}}$ no lo son, ya que acarrean incertidumbre transmitida a través del modelo RBL.

La **Ecuación 5.24** podría, en principio, extenderse intuitivamente a datos de tercer y cuarto orden dando:

$$\text{SEN}_{\text{NAS3}} = \{\mathbf{v}^{\text{T}}\ [\mathbf{P}^{\text{T}}\ (\mathbf{P}_D \otimes \mathbf{P}_C \otimes \mathbf{P}_B)\ \mathbf{P}]^{-1}\mathbf{v}\}^{-1/2} \qquad (5.28)$$

$$\text{SEN}_{\text{NAS4}} = \{\mathbf{v}^{\text{T}}\ [\mathbf{P}^{\text{T}}\ (\mathbf{P}_E \otimes \mathbf{P}_D \otimes \mathbf{P}_C \otimes \mathbf{P}_B)\ \mathbf{P}]^{-1}\mathbf{v}\}^{-1/2} \qquad (5.29)$$

donde $\mathbf{P}_D$ y $\mathbf{P}_E$ son matrices de proyección ortogonal, análogas a $\mathbf{P}_B$ y $\mathbf{P}_C$, definidas en la tercera y cuarta dimensión instrumental de las señales de la muestra de *test*, respectivamente. Estas son una extensión del concepto de NAS a la tercera y cuarta dimensión instrumental. La **Tabla 5.1** contiene estas expresiones basadas en NAS para diferentes órdenes de datos y en presencia de varias fuentes de interferencia.

Sin embargo, el ejemplo anterior muestra que, en principio, todas las inferencias basadas en el concepto de señal analítica neta para datos de segundo, tercer y cuarto orden, son incorrectas. De cualquier manera, en el caso de los datos de segundo orden la

Ecuación 5.24 sería válida, como se muestra en la Referencia 4. Dado que la relación matemática entre *y* en la **Ecuación 5.25** y los perfiles de los interferentes que contienen error es muy compleja, es difícil establecer la verdadera razón por la cual 5.24 da resultados correctos, (es decir, iguales a los de la **Ecuación 5.20**, mientras que las **Ecuaciones 5.28** y **5.29** no son correctas y difieren de la **Ecuación 5.20** para datos de tercer y cuarto orden. Probablemente el motivo tenga que ver con la manera en que NAS debería definirse para datos de mayor orden, un tema que definitivamente requiere de futuras investigaciones.

En la **Figura 5.2** se muestra una comparación de los valores de sensibilidad obtenidos con las diferentes expresiones. En ella, se considera una mezcla binaria que contiene un único analito calibrado así como un único interferente no calibrado, ambos con perfiles gaussianos en todas los modos instrumentales. El eje vertical corresponde a: 1) círculos negros, en el caso de los valores de sensibilidad dados por la **Ecuación 5.24**, es decir, sensibilidad basada en NAS para datos de segundo orden, SEN_{NAS2} ,2) triángulos rojos, en el caso de los valores de la **Ecuación 5.28** para datos de tercer orden, SEN_{NAS3}, y 3) cuadrados azules, para valores de la **Ecuación 5.29** en datos de cuarto orden, SEN_{NAS4}. En todos los casos, el eje horizontal corresponde a valores dados por la ecuación general del Jacobiano (5.20), SEN_J. Como puede verse, solo en el caso de los datos de segundo orden la sensibilidad NAS coincide con la ecuación general, mientras que los datos de tercer y cuarto orden los valores de NAS conducen a una subestimación.

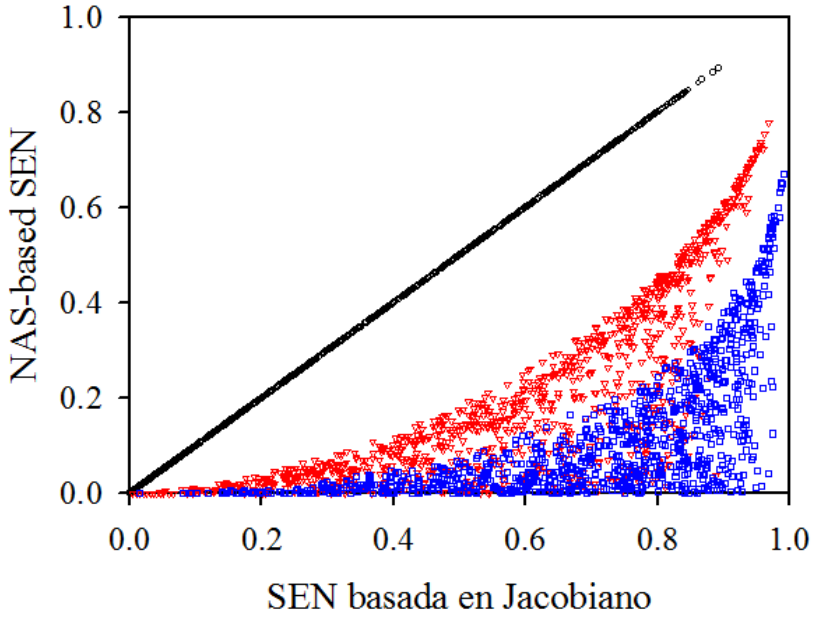


Figura 5.2. Sensibilidades en U-PLS/RML calculadas para datos de segundo, tercero y cuarto orden utilizando expresiones basadas en el concepto de NAS como funciones de las estimaciones obtenidas utilizando la expresión general obtenida por propagación de errores a partir del Jacobiano. Las sensibilidades corresponden a un sistema de un analito con un interferentes con perfiles gaussianos para ambos componentes en todos los modos instrumentales, localizados en 1000 posiciones aleatorias en cada uno de los modos instrumentales. En círculos negros, datos de segundo orden, en triángulos rojos datos de tercer orden y en círculos azules datos de cuarto orden.

5.6 Datos

5.6.1 Simulados

Para cada muestra simulada, se generó un arreglo de datos que consistió en las siguientes dimensiones: $J \times K$ para datos de segundo orden, $J \times K \times L$, para datos de tercer orden y $J \times K \times L \times M$ para datos de cuarto orden, donde *J, K, L y M* son el número de puntos o sensores en cada modo instrumental. El número de muestras analizadas por PLS/RML en todos los casos fue $I=I_{cal}+1$, es decir, el número de muestras de calibración más la muestra de *test*.

Para representar las medidas instrumentales, se definieron perfiles de forma Gaussiana para cuatro componentes distintos (1, 2, 3 y 4), en cada uno de los modos instrumentales. Los perfiles que se muestran en la **Figura 5.3** corresponden a los componentes a

concentración unitaria que se normalizaron de manera que el área debajo de cada uno sea igual a 1.

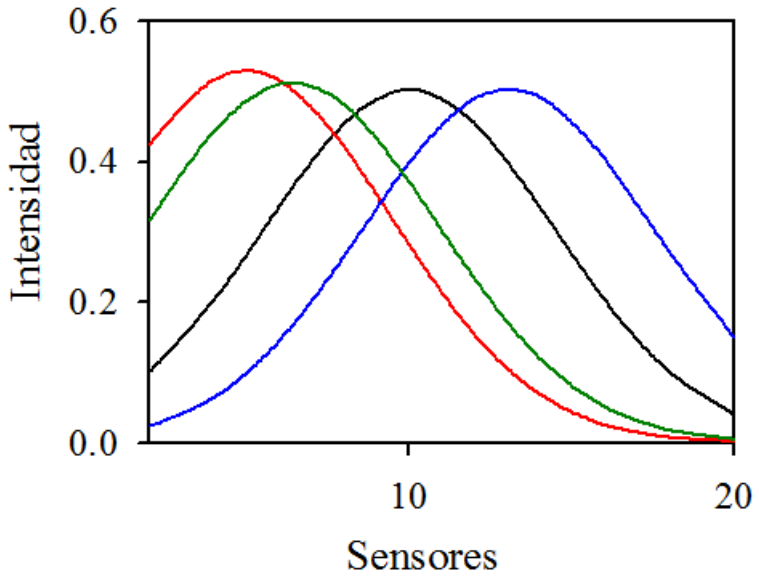


Figura 5.3. Perfiles sin ruido de los componentes utilizados para construir el primer modo de los datos simulados de segundo orden. La línea negra identifica el analito de interés; los colores restantes corresponden al resto de los componentes de la muestra. Los perfiles y el solapamiento relativo en el resto de los modos de datos y órdenes son similares a los que se muestran en esta figura.

El número de puntos de datos utilizados en cada uno de los modos fueron: J=20, K=15, L=15, y M=10. En todos los casos, el máximo del pico para el perfil Gaussiano del componente 1 (analito de interés) se fijó en el centro de cada uno de los rangos de datos. Por su parte, el resto de se ubicó aleatoriamente evitando superposición absoluta, y dando lugar de esta manera a 10 grados distintos de superposición espectral. La figura muestra una situación particular para los cuatro componentes posibles.

En todos los juegos de datos simulados, los conjuntos de calibración fueron creados con concentraciones de analitos tomadas aleatoriamente e uniformemente distribuidas en el rango 0-1, con el siguiente número de muestras: 10 cuando se incluye en la calibración un único analito, 20 para 2 y 30 para 3. Las cuatro muestras de *test* también se generaron a través de concentraciones tomadas aleatoriamente de entre 0 y 1. Las mismas se analizaron uniéndolas, una a una, al conjunto de muestras de calibrado.

Los juegos de datos se definieron de acuerdo con el número total de componentes (B para sistemas binarios, T para ternarios y Q para cuaternarios) y con tres números identificando el orden de los datos, el número de analitos y el número de agentes interferentes. La lista de los sistemas simulados estudiados se muestra en la **Tabla 5.2**. En total, se analizaron 960 sistemas diferentes, correspondientes a 3 órdenes de datos distintos, 6 combinaciones diferentes de número de analitos y de interferentes, 4 concentraciones distintas de los componentes y 10 sensibilidades diferentes variando de acuerdo con el grado de solapamiento espectral.

Tabla 5.2. Sistemas simulados, nomenclatura, orden de los datos y numeración de los componentes

Sistema[a]	Orden de los datos	Componente(s) en la calibración	Agentes interferentes
B2_11	2	1	2
T2_12	2	1	2 y 3
T2_21	2	1 y 2	3
C2_13	2	1	2,3 y 4
C2_22	2	1 y 2	3 y 4
C2_31	2	1,2 y 3	4
B3_11	3	1	2
T3_12	3	1	2 y 3
T3_21	3	1 y 2	3
C3_13	3	1	2,3, y, 4
C3_22	3	1 y 2	3 y 4
C3_31	3	1,2, y 3	4
B4_11	4	1	2
T4_12	4	1	2 y 3
T4_21	4	1 y 2	3
C4_13	4	1	2,3, y 4
C4_22	4	1 y 2	3 y 4
C4_31	4	1, 2, 3	4

[a]La primera letra identifica el número total de componentes (B, binario; T, ternario; C, cuaternario), el primer número el orden de los datos y los dos números finales el número de analitos calibrados y el número de agentes interferentes respectivamente.

5.6.2 Adición de ruido

Un postulado fundamental para poder validar las expresiones para el cálculo de sensibilidad es que el ruido que afecta a los sistemas en estudio es de tipo iid. Por lo tanto, de acuerdo con el esquema general para el cálculo de la incertidumbre que se presentó en el Capítulo 3, el error estándar en la predicción de la concentración del analito utilizando un modelo PLS puede calcularse a través de la fórmula para el caso 1 en la **Tabla 3.2**.

En cada uno de los juegos de datos simulados, el valor en la incertidumbre, fue de 0.002 unidades de desvío estándar en señal, y de 0.001 unidades de desvío estándar en concentración.

Luego de generar cada juego de datos, se adicionó ruido utilizando las diferentes maneras mencionadas anteriormente, y para cada ciclo de adición de ruido, la muestra se sometió a U-PLS/RBL. El proceso de calibración y predicción se llevó a cabo siguiendo los pasos detallados durante la descripción del modelo U-PLS/RBL en el Capítulo 1. Se realizaron 1000 repeticiones siguiendo el mismo esquema que el del diagrama de flujo de la **Figura 3.2**, utilizando una inicialización aleatoria para generar la incertidumbre en la señal, en la concentración o en ambos.

5.7 Resultados

5.7.1 Análisis por simulaciones

En cada sistema de datos simulados, se utilizaron los datos de calibración para construir un modelo U-PLS basado en el analito de interés (componente 1 en todos los casos). Dependiendo del número de modos instrumentales por muestra, se aplicó RBL, RTL o RQL para obtener los *scores* adecuados que permitan la predicción del analito de interés. Repitiendo el proceso de calibración y predicción un determinado número de veces y reiniciando aleatoriamente para adicionar ruido a los datos se obtiene la incertidumbre en la concentración predicha. Como se discutió previamente, se tuvieron en cuenta 4 situaciones diferentes incluyendo ruido en las concentraciones de calibración, en las señales de calibración, en la señal de la muestra de *test* y en todas las situaciones mencionadas anteriormente. Esto permitió chequear el funcionamiento de la fórmula para el caso 1 en la **Tabla 3.2** en conjunto con cada uno de los tres términos por separado en presencia de tres fuentes diferentes de incertidumbre.

Al igual que en capítulos anteriores, debido al gran número de sistemas estudiados, una manera conveniente de analizar los resultados consiste en graficar el valor de incertidumbre obtenido mediante las simulaciones de adición de ruido y compararlo con el obtenido mediante la expresión para el cálculo de la incertidumbre cuando el ruido es iid (caso 1, **Tabla 3.2**) identificando las diferentes fuentes de error por medio de la utilización de diversos símbolos. La **Figura 5.3 A** incluye todos los sistemas de segundo orden, la **Figura 5.3 B** los de tercer orden y la 5.3 C los de cuarto orden estudiados.

Los resultados que se muestran en las **Figuras 5.3 A-C** sugieren que la aproximación presentada para calcular la sensibilidad en U-PLS/RML y las correspondientes incertidumbres en las concentraciones predichas es apropiada, ya que la incertidumbre en las predicciones obtenidas por adición de ruido se correlacionan con las predichas a partir de insertar la sensibilidad calculada a partir de la **Ecuación 5.20**.

Una inspección visual de esta gráfica indica inmediatamente que la incertidumbre transmitida por la calibración (círculos azules) es menor que la que se propaga a través de la muestra de *test*, debido a que las primeras están escaladas por la leva de la muestra.

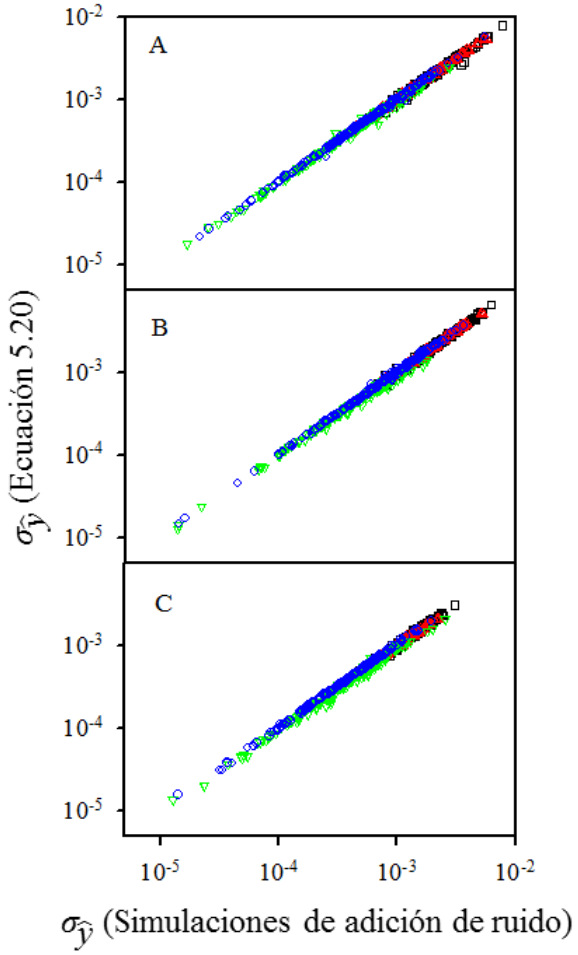


Figura 5.3 Gráfico de incertidumbres en las concentraciones predichas luego de las simulaciones de adición de ruido, como función de las estimaciones basadas en la **Ecuación 5.20** (A) Resultados para todos los sistemas de segundo orden (B2_11, T2_12, T2_21, Q2_13, Q2_22, y Q2_31, ver **Tabla 5.2** con la descripción de los símbolos). (B) Resultados para todos los sistemas de tercer orden (B3_11, T3_12, T3_21, Q3_13, Q3_22, y Q3_31). (C) Resultados para todos los sistemas de cuarto orden (B4_11, T4_12, T4_21, Q4_13, Q4_22, y Q4_31). En los tres gráficos, los símbolos identifican los siguientes casos: círculos azules, ruido sólo en concentraciones de calibración; triángulos verdes orientados hacia abajo, ruido sólo en señales de calibración; triángulos rojos hacia arriba, ruido sólo en señales de las muestras de *test*; cuadrados negros, ruido en concentraciones y en señales.

Resulta interesante observar que la aproximación intuitiva dada por el concepto de NAS no es adecuada, en términos generales, para cubrir las sensibilidades esperadas. Solo en el caso de U-PLS/RBL aplicado a señales de segundo orden las aproximaciones por NAS y por Jacobiano coinciden generando valores idénticos de sensibilidad. La **Figura 5.4** muestra los resultados obtenidos cuanto se contrastaron los valores de desvío estándar

obtenidos por simulaciones de adición de ruido con los calculados por la expresión que resulta de la extensión intuitiva del concepto de NAS a datos de tercer orden.

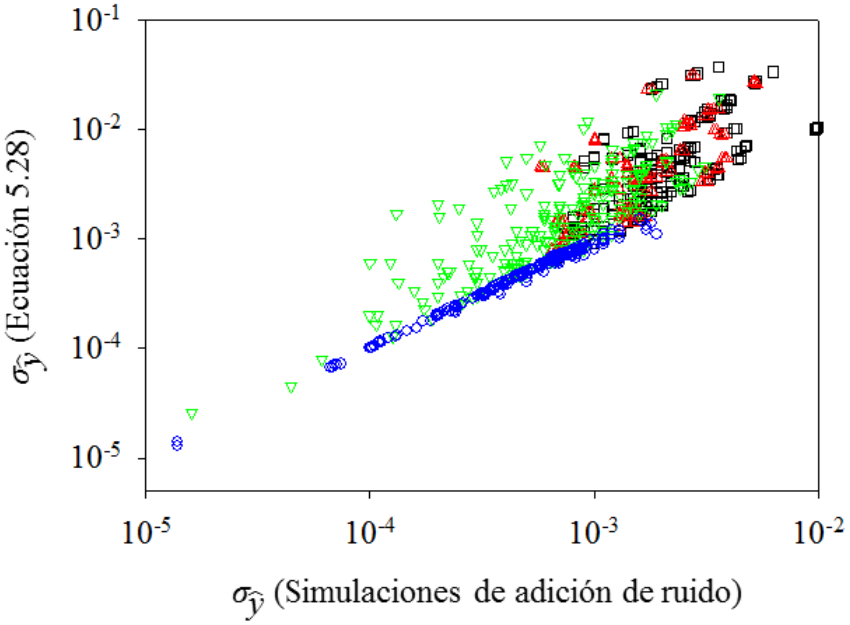


Figura 5.4 Gráfico de comparación entre las incertidumbres en concentraciones predichas obtenidas por medio de ciclos de adición de ruido respecto a las calculadas utilizando la **Ecuación 5.28**, en sistemas simulados de tercer orden (ver **Tabla 5.2**). La simbología coincide con la que se presentó en la **Figura 5.3**.

En MCR de orden 2 y PARAFAC de orden 2 a 4, se analizaron los mismos tipos de sistemas que los que se mencionaron anteriormente utilizando la aproximación por Jacobiano. Si bien la fórmula para estimar el desvío estándar de predicción por muestra cuando el ruido es de tipo iid, se desarrolló inicialmente para PLS, debido a su carácter general (dado que incluye parámetros y fuentes de error que aparecen en la mayoría de los modelos de calibrado), en principio sería también válido utilizarla para analizar las diferentes fuentes de ruido en MCR y PARAFAC. La única diferencia es que en estos casos, el cálculo de la leva se realiza a partir de las concentraciones reales y no de los *scores*, ya que estos modelos no emplean variables latentes como PLS. De esta manera, la correspondiente fórmula en estos casos está dada por:

$$h_i = \hat{y}^2_{test,i} / \|\mathbf{y}_{cal}\|^2 \qquad (5.30)$$

donde $\hat{y}_{test,i}$ es la concentración predicha para la muestra incógnita e $\mathbf{y}_{cal}$ es un vector conteniendo las concentraciones de las muestras de calibración.

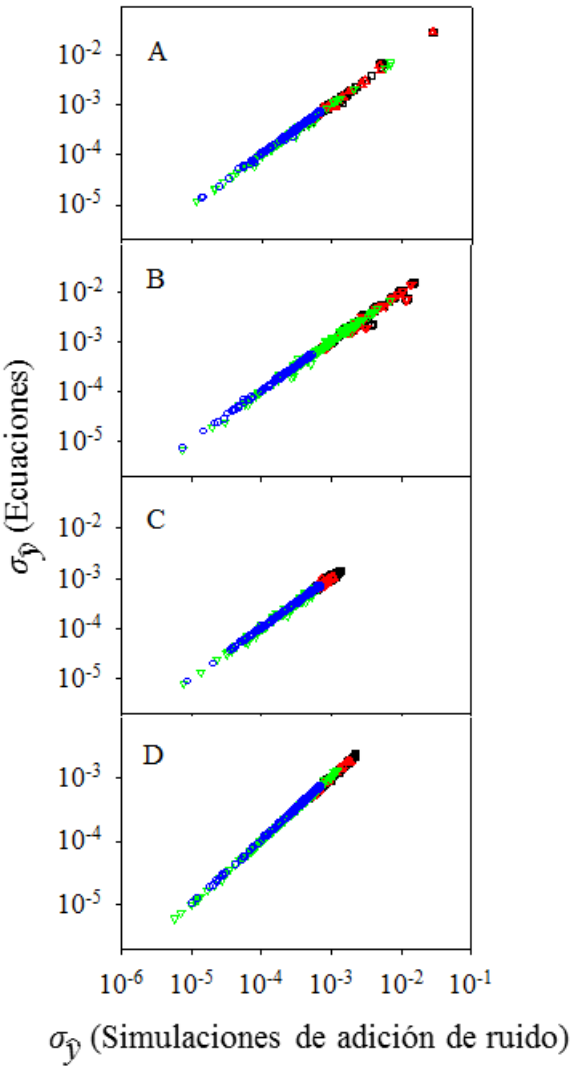


Figura 5.5 Gráficos de comparación entre incertidumbre en concentraciones predichas calculadas utilizando simulaciones de adición de ruido respecto a las obtenidas utilizando las expresiones 5.8 en el caso de MCR (B) y 5.10 en el caso de PARAFAC de dos (A), tres (C) y cuatro (D) vías.

Como muestran las **Figuras 5.5 A-D** al igual que en U-PLS/RML, en estos modelos la expresión para cálculo de sensibilidad propuesta permite calcular valores de incertidumbre que se condicen perfectamente con los resultados generados por medio de simulaciones de adición de ruido, para numerosos sistemas. Sin embargo, dado que esta expresión de carácter general se dedujo a partir de propagar errores a través del modelo

PLS, su extensión a PARAFAC o MCR-ALS es sólo una aproximación. Esto lleva a que, para casos en los que hay muchos interferentes y un fuerte solapamiento de picos, la aproximación pierda validez, especialmente en los términos que incluyen a la leva (errores en señal de calibración y en concentraciones). Para resolver esta situación, habría que deducir una fórmula para el cálculo de incertidumbre bajo supuesto iid considerando más detallada y específicamente la mecánica de funcionamiento de PARAFAC y MCR-ALS. De cualquier modo, dados los fines de este capítulo, que discute fundamentalmente el cálculo de la sensibilidad, en U-PLS/RML lo anterior escapa el alcance de esta tesis y constituye una posible perspectiva.

5.7.2 Análisis de un sistema experimental

Los datos experimentales que se analizaron corresponden a la cuantificación del pesticida fluorescente carbaril, que se hidroliza en medio alcalino dando lugar al compuesto fluorescente 1-naftol. La cinética de esta reacción depende del pH y da la oportunidad de medir matrices de excitación emisión como función del tiempo y del pH, es decir, de generar datos de cuarto orden. Las muestras de calibración contienen únicamente el analito carbaril, mientras que las muestras de *test* tienen, junto con el analito de interés, otro pesticida fluorescente que funciona como interferente (ya sea tiabendazol o fuberidazol). En consecuencia, se requiere la ventaja de segundo orden para poder lograr una determinación exitosa.

Dado que se dispone de datos de cuarto orden, es posible adaptarlos para explorar las posibilidades que ofrecen las calibraciones de segundo, tercero y cuarto orden. El primer paso, consistió en seleccionar, para los datos de cuarto orden completos correspondientes a una determinada muestra, un sub-arreglo de segundo orden que correspondería a medir matrices de excitación-emisión de fluorescencia a valores fijos de tiempo de reacción y pH. Estos datos se analizaron utilizando U-PLS/RBL. Cuando se utilizó validación cruzada para obtener el número óptimo de variables latentes de calibración se obtuvo un valor de A=1. Aunque aparezcan dos componentes en la calibración, los mismos están mutuamente correlacionados, ya que un componente se hidroliza para dar lugar al segundo, hecho por el cual una única variable latente en U-PLS es suficiente para modelar los datos. La **Figura 5.6** muestra los espectros de emisión y de excitación que se obtuvieron utilizando RBL en dos muestras de *test* típicas (una conteniendo fuberidazol como agente interferente y la otra tiabendazol), en comparación

con los perfiles conocidos para carbaril puro (analito de interés) y el 1-naftol (producto de hidrólisis). La posibilidad de obtener estos perfiles de manera satisfactoria, está directamente relacionada con la ventaja de segundo orden, permitiendo a RBL quitar la contribución de los agentes interferentes de la señal de *test* total de cada muestra. Los resultados de la predicción se muestran en la **Tabla 5.3**, junto con las correspondientes cifras de mérito.

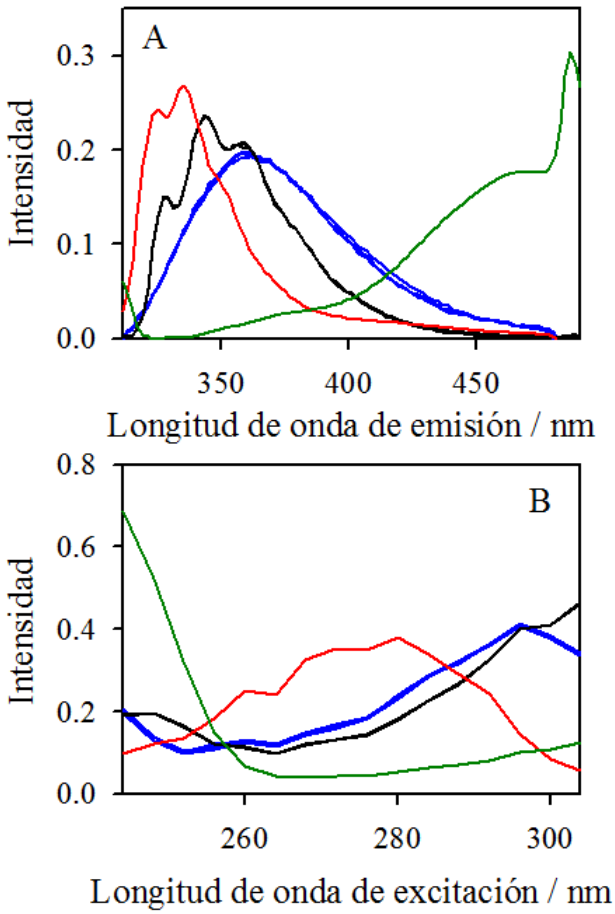


Figura 5.6. Perfiles de excitación (A) y emisión (B) para los componentes de la muestra. Las líneas verdes y rojas corresponden a los espectros experimentales para el analito carbaril y su producto de hidrólisis 1-naftol, respectivamente. Las líneas azules y negras (similares entre sí) indican que los perfiles para los interferentes fuberidazol y tiabendazol, tal y como se obtienen para las muestras 1 y 6 respectivamente, por medio de un análisis U-PLS/RML de los datos de segundo, tercero y cuarto orden.

La sensibilidad debería mejorar al obtener y procesar datos de tercer orden correspondientes a la medida de las matrices mencionadas previamente como función del tiempo. Esto puede estudiarse seleccionando, de los datos de cuarto orden de cada muestra,

datos de tercer orden a un valor de pH fijo (10). Cuando estos datos se calibraron utilizando U-PLS, la validación cruzada sugirió nuevamente un valor de *A=1*, incluso cuando en los datos de calibrado hay dos componentes que generan una respuesta. De la misma manera que el caso anterior, este resultado es comprensible en vistas de la correlación mutua de estos dos componentes debido a la cinética de reacción que se monitorea. Para cada una de las muestras de *test* que contienen interferencias, RTL permitió modelar el correspondiente agente interferente en cada muestra. En la **Figura 5.6** se muestra una comparación de los perfiles de excitación y emisión, donde se observa una excelente correspondencia entre los resultados provenientes de RBL aplicado a datos de segundo orden. El perfil de tiempo (no se muestra) que devuelve RTL del conjunto de datos de tercer orden, es un perfil contante (como era de esperarse teniendo en cuenta que el agente interferente es estable con el pH). Los resultados específicos de la predicción y las cifras de mérito se muestran en la **Tabla 5.3**.

Finalmente, el conjunto completo de datos de cuarto orden se sometió a U-PLS/RQL, con resultados cualitativos similares en comparación con los análisis realizados anteriormente. Los perfiles de excitación y emisión de fluorescencia que se obtienen por medio de este algoritmo son comparables con los de U-PLS/RBL y U-PLS/RTL tal como se muestra en la **Figura 5.6**. Al igual que en los casos anteriores, la determinación cuantitativa del analito carbaril se muestra en la **Tabla 5.3** permitiendo una comparación entre las distintas formas de analizar los datos.

Tabla 5.3. Resultados analíticos y cifras de mérito para el ejemplo experimental utilizando U-PLS/RBL, U-PLS/RTL, y U-PLS/RQL.

Muestra	Nominal	U-PLS/RBL	U-PLS/RTL	U-PLS/RQL
		Resultados Analíticos[a]		
1	100	86 (1.6)	85 (0.6)	98 (0.5)
2	125	108 (1.5)	107 (0.5)	120 (0.4)
3	150	134 (1.5)	136 (0.5)	149 (0.4)
4	200	172 (1.6)	171 (0.5)	186 (0.4)
5	250	218 (1.6)	223 (0.6)	245 (0.6)
6	100	85 (1.6)	85 (0.6)	89 (0.5)
7	100	91 (1.6)	92 (0.5)	94 (0.5)
8	200	184 (1.5)	185 (0.5)	197 (0.4)
9	250	202 (1.6)	253 (0.7)	241 (0.6)
		Cifras de Mérito[b]		
RMSEP / μg L^{-1}		25	18	7.3
REP / %		16	12	4.9
SEN / AFU L μg^{-1}		1.3	5.5	12
γ / L μg^{-1}		0.7	3.1	6.7
LOD / μg L^{-1}		5.3	2	1.5
LOQ / μg L^{-1}		16	6	4.5

[a] Todas las concentraciones están en unidades de μg L^{-1}. Los correspondientes desvíos estándar en paréntesis. En todos los casos los datos se procesaron luego de aplicar centrado, con una variable latente para U-PLS y un componente RML. [b]REP = Error Relativo de Predicción basado en la concentración de calibración media, SEN calculada a partir de la **Ecuación 5.20**, y γ = SEN/[σ_x]$^{1/2}$, LOD = 3.3 σ_{y_o}y y LOQ = 10 σ_{y_o} , donde σ_{y_o} se calcula a partir de la **Ecuación 4.1** completa con valores de σ_x^2= 2 unidades arbitrarias de fluorescencia (AFU) y σ_y^2 = 1 μg L^{-1}.

La comparación de las cifras de mérito presentadas en la **Tabla 5.3** para datos de segundo, tercero y cuarto orden en los sistemas experimentales que se estudiaron, muestran un claro incremento en la sensibilidad y la sensibilidad analítica en la medida que se incrementa el orden de los datos. También puede observarse una mejoría en los indicadores de error promedio en las concentraciones (RMSE y REP), al igual que la incertidumbre en las concentraciones predichas y la capacidad de detección (LOD y LOQ). Sin embargo, la mejora en SD, LOD y LOQ no es directamente proporcional a la ganancia en sensibilidad. Esto es esperable si se inspecciona la **Ecuación 4.1**, donde 2 de los tres términos están afectados por el parámetro de sensibilidad. El último término, sin embargo, depende de la leva de la muestra así como de la incertidumbre en las concentraciones de calibración pero no de la sensibilidad. Esto implica un efecto más suave a altas sensibilidades ya que la incertidumbre estaría controlada principalmente por la incertidumbre en las concentraciones de calibrado, que es constante con el orden de los datos. La **Figura 5.7** compara los valores de LOD con una aproximación que ignora la leva de la muestra y considera únicamente el primer término de la **Ecuación 4.1** al estimar $\sigma(y_o)$.

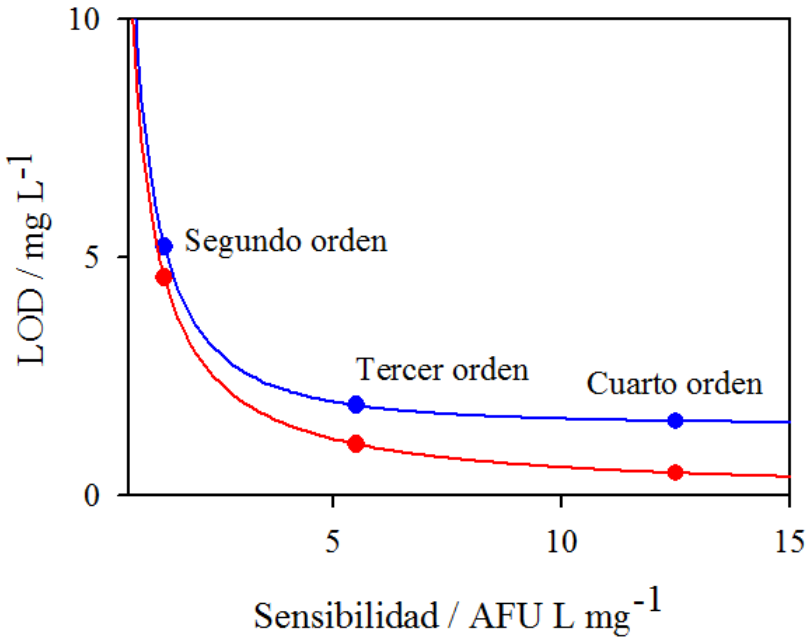


Figura 5.7. Límites de detección para la determinación del analito carbaril en el ejemplo experimental. En círculos rojos, valores del límite de detección obtenidos por U-PLS/RML para los diferentes órdenes de datos experimentales a partir de la expresión aproximada (sólo primer término de la **Ecuación 4.1**), con la línea sólida roja mostrando la variación del LOD como una función de SEN_J. En círculos azules, valores de LOD obtenidos de la expresión recomendada por la IUPAC (**Ecuación 4.1**), insertando la ecuación para el cálculo de σ (y_o) el correspondiente valor de

h para la muestra de *test* 1: para los datos de segundo orden, *h*=0.23, para tercer orden, *h*=0.24 y para cuarto orden, *h*=0.25. La línea sólida azul corresponde a la **Ecuación 4.1** asumiendo *h*=0.24.

Es evidente que esta última aproximación sobrestima notoriamente el límite de detección, mientras que los valores mostrados en la **Tabla 5.3**, basado en las **Ecuación 4.1** completa con valores aproximados para la leva del blanco, da lugar a una estimación más realista.

5.8 Conclusión

Por medio de pruebas realizadas a través de simulaciones de adición de ruido aplicadas sobre sistemas simulados de orden 2 y superiores con cantidades variables de componentes y de interferentes, se pudo demostrar la utilidad de nuevas expresiones para calcular la sensibilidad cuando se aplican los modelos MCR, PARAFAC y U-PLS/RML. Estas expresiones tienen como factor común la utilización de aproximaciones basadas en el empleo de la matriz Jacobiana que surge de una metodología de propagación de errores, que considera todas las fuentes de variación de cada modelo, ya que se construye teniendo en cuenta todos aquellos parámetros ajustables que participan en cada modelo.

Los resultados obtenidos por esta aproximación y su comparación con los que se obtuvieron empleando la aproximación NAS, constituyen un fuerte indicador de que el concepto intuitivo de NAS para órdenes superiores debería ser revisado para hacerlo consistente con la correspondiente aproximación por propagación de errores.

Un punto importante a tener en cuenta, es que a partir del cuerpo de trabajo definido en este capítulo, es posible escribir una expresión que incluya todas las ecuaciones de sensibilidad en un único esquema de trabajo unificado. Es decir, una ecuación que tenga en cuenta desde orden 0 (calibración univariada) hasta modelos de calibración basados en datos de cualquier orden y número de modos instrumentales. El resultado principal se puede condensar apropiadamente en la siguiente expresión:[17]

$$\mathrm{SEN}_n = \left\{ \mathbf{g}_n^{\mathrm{T}} \left[\mathbf{Z}_{\mathrm{cal}}^{\mathrm{T}} \left(\mathbf{I} - \mathbf{Z}_{\mathrm{int}} \mathbf{Z}_{\mathrm{int}}^{+} \right) \mathbf{Z}_{\mathrm{cal}} \right]^{-1} \mathbf{g}_n \right\}^{-1/2} \qquad (5.31)$$

Los parámetros que aparecen en esta ecuación pueden definirse específicamente para cada escenario de calibración, aunque también se puede realizar una descripción general de tipo cualitativa. Tanto la matriz $\mathbf{Z}_{\mathrm{cal}}$ como el vector que marca la especificidad

para cada analito g_n se refieren a la fase de calibración. La matriz $\mathbf{Z}_{cal}$ recolecta los perfiles (ya sea en su forma pura o como en combinaciones lineales) para los componentes esperados presentes en el conjunto de calibrado, mientras que $\mathbf{g_n}$, selecciona y combina esta información, haciéndola específica para el analito de interés *n*. El otro conjunto de parámetros a considerar se encuentra en la matriz $(\mathbf{I} - \mathbf{Z}_{int}\mathbf{Z}_{int}^{+})$, que es la manifestación matemática de la ventaja de segundo orden, y por lo tanto sólo aparece en metodologías de calibración de órdenes superiores (tres vías y más). El propósito de esta matriz es corregir los perfiles de los componentes esperados del efecto de solapamiento de los perfiles de los componentes inesperados, o potenciales agentes interferentes. Específicamente, la matriz $(\mathbf{I} - \mathbf{Z}_{int}\mathbf{Z}_{int}^{+})$ depende de los perfiles de los componentes inesperados que podrían aparecer en una determinada muestra de *test* y sólo se incluyen en caso que se pueda alcanzar la ventaja de segundo orden, debido a que esta información sólo se encuentra disponible en ese caso. Los perfiles para los constituyentes inesperados podrían ser: (1) perfiles de los componentes puros (o aproximaciones de los mismos) dados por MCR-ALS o PARAFAC y todas sus variantes de descomposición multilineal o (2) perfiles latentes (combinaciones lineales de las variables originales o *loadings*) devueltos por el procedimiento RML. En este punto resulta interesante observar que $(\mathbf{I} - \mathbf{Z}_{int}\mathbf{Z}_{int}^{+})$ define una proyección ortogonal al espacio definido por los componentes inesperados, debido a que $\mathbf{Z}_{int}$ sólo contiene información relativa a las señales generadas por estos compuestos. Aunque la forma específica de $\mathbf{Z}_{int}$ difiera de lo esperado intuitivamente a partir de la extensión simple del concepto de NAS desde primer orden a órdenes superiores, se podía anticipar un principio de la **Ecuación 5.29** por inspección de la ecuación FO, la cual no se sustenta en los principios de propagación de errores.

Finalmente es importante resaltar las propiedades de la sensibilidad multi-vía que se reflejan claramente en los parámetros de la **Ecuación 5.31**: (1) es específica para cada analito, ya que el factor $\mathbf{g}_n$ depende del analito de interés, (2) es específica para cada muestra, ya que la composición de cada muestra de *test* es única en lo que respecta a componentes inesperados, generando una única matriz $\mathbf{Z}_{int}$ y (3) es específica para cada algoritmo, porque cada metodología de procesamiento da lugar al conjunto específico de parámetros $\mathbf{Z}_{exp}$, $\mathbf{g}_n$ y $(\mathbf{I} - \mathbf{Z}_{int}\mathbf{Z}_{int}^{+})$.

5.9 Perspectivas

Si se tiene en cuenta el desarrollo realizado durante la Sección 5.5.1 de este capítulo para obtener la expresión para el cálculo de la sensibilidad en U-PLS/RML, se realizan dos suposiciones importantes: (1) el error es iid y sólo se emplea como una perturbación mínima del sistema y (2) los coeficientes de regresión **v** obtenidos a partir de la calibración son precisos, es decir la fuente principal de error proviene de los *scores* **t** que resultan de la proyección de la señal de la muestra de *test* en el espacio de variables latentes generado a partir de las muestras de calibración. Por lo tanto sólo se está teniendo en cuenta la propagación del error en la señal de la muestra de *test*.

Sin embargo, considerando el planteo del Capítulo 3 de esta tesis, si se desea calcular la incertidumbre en la predicción cuando se emplea U-PLS/RBL, es necesario considerar todas las fuentes de error presentes en el sistema, al igual que la estructura del error que lo está afectando. Teniendo en cuenta que en la medida que se adicionan nuevos modos instrumentales a la calibración la diversidad de los datos en cuanto a su origen será mayor, y también se incrementará la posibilidad de que la estructura del error se aleje del supuesto iid clásico. Por lo tanto, extender el esquema empleado para calcular el error de predicción cuando se calibra sobre datos de múltiples vías sería sumamente útil.

Como fue descripto durante secciones previas, el modelo U-PLS/RBL intenta modelar la señal desdoblada correspondiente a una determinada muestra de *test* como la suma de dos contribuciones: (1) la porción de la señal de *test* modelada por la calibración, y (2) la señal de los interferentes modelados por RBL:

$$\mathbf{x} = \text{Modelo de calibración para } \mathbf{x} + \text{Modelo RBL para } \mathbf{x} + \mathbf{e} = \qquad \text{(P5-1)}$$

$$= \mathbf{P}\mathbf{t}^{\mathrm{T}} + \sum_{n=1}^{N\,\mathrm{int}} \mathbf{c}_{\mathrm{int},n} \otimes \mathbf{b}_{\mathrm{int},n} + \mathbf{e}$$

En esta ecuación, el producto $\mathbf{P}\mathbf{t}^{\mathbf{T}}$ representa la parte de **x** que se puede modelar a partir de los parámetros de la calibración, mientras que la sumatoria de los productos de Kronecker representa la contribución de los interferentes.

La diferenciación considerando ruido tanto en calibración como en las señales de la muestra de *test* lleva a la siguiente expresión:

$$d\mathbf{x} = d\mathbf{P}\,\mathbf{t}^{\mathrm{T}} + \mathbf{P}\,d\mathbf{t}^{\mathrm{T}} + d\left(\sum_{n=1}^{N\,\mathrm{int}} \mathbf{c}_{\mathrm{int},n} \otimes \mathbf{b}_{\mathrm{int},n}\right) \qquad \text{(P5-2)}$$

Como se demostró el último término se puede expresar como el producto de la matriz $\mathbf{Z}_{\mathrm{int}}$ que representa el espacio definido por los interferentes y los vectores de los diferenciales de $\mathbf{c}_{\mathrm{int}}$ y $\mathbf{b}_{\mathrm{int}}$:

$$d\mathbf{x} = d\mathbf{P}\,\mathbf{t}^{\mathrm{T}} + \mathbf{P}\,d\mathbf{t}^{\mathrm{T}} + \mathbf{Z}_{\mathrm{int}}\,[d\mathbf{b}_{\mathrm{int},1}\,;\, d\mathbf{c}_{\mathrm{int},1}\,;\, d\mathbf{b}_{\mathrm{int},2}\,;\, d\mathbf{c}_{\mathrm{int},2}\,;\,\ldots] \qquad \text{(P5-3)}$$

donde ';' indica notación de MATLAB, y $\mathbf{Z}_{\mathrm{int}}$ se puede escribir como:

$$\mathbf{Z}_{\mathrm{int}} = [\mathbf{c}_{\mathrm{int},1}\otimes\mathbf{I}_{\mathrm{b}} \mid \mathbf{I}_{\mathrm{c}}\otimes\mathbf{b}_{\mathrm{int},1} \mid \mathbf{c}_{\mathrm{int},2}\otimes\mathbf{I}_{\mathrm{b}} \mid \mathbf{I}_{\mathrm{c}}\otimes\mathbf{b}_{\mathrm{int},2} \mid \ldots] \qquad \text{(P5-4)}$$

donde $\mathbf{I}_{\mathrm{b}}$ e $\mathbf{I}_{\mathrm{c}}$ son matrices identidad de un tamaño adecuado ($J \times J$ para $\mathbf{I}_{\mathrm{b}}$ y $K \times K$ para $\mathbf{I}_{\mathrm{c}}$). Para más detalles acerca de las expresiones A5-3 y A5-4 ver Apéndice.

El último término es proporcional a $\mathbf{Z}_{\mathrm{int}}$ y se puede eliminar por medio de una proyección ortogonal :

$$\mathbf{P}_{\mathrm{Zint}}\,d\mathbf{x} = \mathbf{P}_{\mathrm{Zint}}d\mathbf{P}\,\mathbf{t}^{\mathrm{T}} + \mathbf{P}_{\mathrm{Zint}}\,\mathbf{P}\,d\mathbf{t}^{\mathrm{T}} \qquad \text{(P5-5)}$$

donde

$$\mathbf{P}_{\mathrm{Zint}} = (\mathbf{I} - \mathbf{Z}_{\mathrm{int}}\,\mathbf{Z}_{\mathrm{int}}^{+}) \qquad \text{(P5-6)}$$

por lo que

$$\mathbf{P}_{\mathrm{Zint}}\,\mathbf{P}\,d\mathbf{t}^{\mathrm{T}} = \mathbf{P}_{\mathrm{Zint}}\,d\mathbf{x} - \mathbf{P}_{\mathrm{Zint}}\,d\mathbf{P}\,\mathbf{t}^{\mathrm{T}} \qquad \text{(P5-7)}$$

de donde $d\mathbf{t}^{\mathrm{T}}$ se puede despejar por multiplicación de ($\mathbf{P}_{\mathrm{Zint}}\,\mathbf{P}$):

$$d\mathbf{t}^{\mathrm{T}} = (\mathbf{P}_{\mathrm{Zint}}\,\mathbf{P})^{+}\,\mathbf{P}_{\mathrm{Zint}}[d\mathbf{x} - d\mathbf{P}\,\mathbf{t}^{\mathrm{T}}] \qquad \text{(P5-8)}$$

Transponiendo:

$$d\mathbf{t} = [d\mathbf{x}^{\mathrm{T}} - \mathbf{t}\,d\mathbf{P}^{\mathrm{T}}]\mathbf{P}_{\mathrm{Zint}}\,(\mathbf{P}_{\mathrm{Zint}}\,\mathbf{P})^{+\mathrm{T}} \qquad \text{(P5-9)}$$

En esta operación es importante tener en cuenta que $\mathbf{P}_{\mathrm{Zint}}$ es simétrica e idempotente.

Dado que $\mathbf{P}^{\mathrm{T}} = \mathbf{T}^{+}\mathbf{X}$:

$$d\mathbf{P}^{\mathrm{T}} = \mathbf{T}^{+}d\mathbf{X} + d(\mathbf{T}^{+})\,\mathbf{X} \qquad \text{(P5-10)}$$

y reemplazando este resultado en la última expresión finalmente se obtiene :

$$d\mathbf{t} = \{d\mathbf{x}^{\mathrm{T}} - \mathbf{t}\,[\mathbf{T}^{+}d\mathbf{X} + d(\mathbf{T}^{+})\,\mathbf{X}]\}\;\mathbf{P}_{\mathrm{Zint}}\,(\mathbf{P}_{\mathrm{Zint}}\,\mathbf{P})^{+\mathrm{T}} \qquad \text{(P5-11)}$$

Reemplazando este resultado en la expresión para el cambio diferencial en la concentración predicha:

$$d\hat{y} = \mathbf{t}d(\mathbf{T}^{+})\mathbf{y}_{\mathrm{cal}} + (d\mathbf{t})\mathbf{T}^{+}\mathbf{y}_{\mathrm{cal}} + \mathbf{t}\mathbf{T}^{+}d\mathbf{y}_{\mathrm{cal}} \qquad \text{(P5-12)}$$

El resultado es:

$$d\hat{y} = \mathbf{t}d(\mathbf{T}^{+})\mathbf{y}_{\mathrm{cal}} + (d\mathbf{x}^{\mathrm{T}})\mathbf{P}_{\mathrm{Zint}}(\mathbf{P}_{\mathrm{Zint}}\mathbf{P})^{+\mathrm{T}}\mathbf{T}^{+}\mathbf{y}_{\mathrm{cal}} - \mathbf{t}\mathbf{T}^{+}d\mathbf{X}\mathbf{P}_{\mathrm{Zint}}(\mathbf{P}_{\mathrm{Zint}}\mathbf{P})^{+\mathrm{T}}\mathbf{T}^{+}\mathbf{y}_{\mathrm{cal}} +$$
$$- \mathbf{t}d\mathbf{T}^{+}\mathbf{X}\mathbf{P}_{\mathrm{Zint}}(\mathbf{P}_{\mathrm{Zint}}\mathbf{P})^{+\mathrm{T}}\mathbf{T}^{+}\mathbf{y}_{\mathrm{cal}} + \mathbf{t}\mathbf{T}^{+}d\mathbf{y}_{\mathrm{cal}} \qquad \text{(P5-13)}$$

El factor ($\mathbf{P}_{\mathrm{Zint}}\,\mathbf{P}$) se puede definir como una matriz de *loadings* efectivos $\mathbf{P}_{\mathrm{eff,EP}}$:

$$\mathbf{P}_{\mathrm{Zint}}\,\mathbf{P}_{\mathrm{eff,EP}}{}^{+\mathrm{T}} = \mathbf{P}_{\mathrm{Zint}}\,(\mathbf{P}_{\mathrm{Zint}}\,\mathbf{P})^{+\mathrm{T}} = \mathbf{P}_{\mathrm{Zint}}\,[(\mathbf{P}^{\mathrm{T}}\mathbf{P}_{\mathrm{Zint}}\,\mathbf{P}_{\mathrm{Zint}}\,\mathbf{P})^{-1}\mathbf{P}^{\mathrm{T}}\,\mathbf{P}_{\mathrm{Zint}}]^{\mathrm{T}} =$$
$$= \mathbf{P}_{\mathrm{Zint}}\mathbf{P}_{\mathrm{Zint}}\mathbf{P}\,(\mathbf{P}^{\mathrm{T}}\mathbf{P}_{\mathrm{Zint}}\,\mathbf{P}_{\mathrm{Zint}}\,\mathbf{P})^{-1} = \mathbf{P}_{\mathrm{Zint}}\mathbf{P}\,(\mathbf{P}^{\mathrm{T}}\mathbf{P}_{\mathrm{Zint}}\,\mathbf{P}_{\mathrm{Zint}}\,\mathbf{P})^{-1} =$$
$$= \mathbf{P}_{\mathrm{eff,EP}}\,(\mathbf{P}_{\mathrm{eff,EP}}{}^{\mathrm{T}}\mathbf{P}_{\mathrm{eff,EP}})^{-1} = \mathbf{P}_{\mathrm{eff,EP}}{}^{+\mathrm{T}} \qquad \text{(P5-14)}$$

Siguiendo el mismo razonamiento, el producto [$\mathbf{P}_{\mathrm{Zint}}\;\mathbf{P}_{\mathrm{eff,EP}}{}^{+\mathrm{T}}\;\mathbf{T}^{+}\mathbf{y}_{\mathrm{cal}}$] puede interpretarse a su vez como el vector de coeficientes de regresión efectivo $\mathbf{b}_{\mathrm{eff}}$. De esto último surge la expresión que se muestra a continuación:

$$\mathbf{X}\,\mathbf{P}_{\mathrm{Zint}}\,(\mathbf{P}_{\mathrm{Zint}}\,\mathbf{P})^{+\mathrm{T}}\,\mathbf{T}^{+}\mathbf{y}_{\mathrm{cal}} = \mathbf{X}\,\mathbf{P}_{\mathrm{eff,EP}}{}^{+\mathrm{T}}\,\mathbf{T}^{+}\,\mathbf{y}_{\mathrm{cal}} = \mathbf{X}\,\mathbf{b}_{\mathrm{eff}} = \hat{\mathbf{y}}_{\mathrm{cal}} \qquad \text{(P5-15)}$$

donde $\hat{\mathbf{y}}_{\mathrm{cal}}$ es el vector de concentraciones predichas de calibración. Dado que los dos vectores $\hat{\mathbf{y}}_{\mathrm{cal}}$ e $\mathbf{y}_{\mathrm{cal}}$, aproximadamente se cancelan

$$d\hat{y} = (d\mathbf{x}^{\mathrm{T}})\mathbf{P}_{\mathrm{Zint}}(\mathbf{P}_{\mathrm{Zint}}\mathbf{P})^{+\mathrm{T}}\mathbf{T}^{+}\mathbf{y}_{\mathrm{cal}} - \mathbf{t}\mathbf{T}^{+}d\mathbf{X}\mathbf{P}_{\mathrm{Zint}}(\mathbf{P}_{\mathrm{Zint}}\mathbf{P})^{+\mathrm{T}}\mathbf{T}^{+}\mathbf{y}_{\mathrm{cal}} + \mathbf{t}\mathbf{T}^{+}d\mathbf{y}_{\mathrm{cal}} \qquad \text{(P5-16)}$$

$$d\hat{y} = (d\mathbf{x}^{\mathrm{T}})(\mathbf{P}_{\mathrm{eff,EP}})^{+\mathrm{T}}\mathbf{v} - \mathbf{h}d\mathbf{X}(\mathbf{P}_{\mathrm{eff,EP}})^{+\mathrm{T}}\mathbf{v} + \mathbf{h}d\mathbf{y}_{\mathrm{cal}} \qquad \text{(P5-17)}$$

donde $\mathbf{h} = \mathbf{t}\mathbf{T}^{+}$ es el vector leva de la muestra.

Estos términos no están correlacionados unos con otros, por lo que contribuirán de manera independiente a la variancia en la predicción:

$$E(d\hat{y}^2) = \mathbf{v}^{\mathrm{T}}(\mathbf{P}_{\mathrm{eff,EP}})^{+}(d\mathbf{x}d\mathbf{x}^{\mathrm{T}})(\mathbf{P}_{\mathrm{eff,EP}})^{+\mathrm{T}}\mathbf{v} - \mathbf{v}^{\mathrm{T}}(\mathbf{P}_{\mathrm{eff,EP}})^{+}d\mathbf{X}^{\mathrm{T}}\mathbf{h}^{\mathrm{T}}\mathbf{h}d\mathbf{X}(\mathbf{P}_{\mathrm{eff,EP}})^{+\mathrm{T}}\mathbf{v}$$

$$+ \mathbf{h}d\mathbf{y}_{cal}d\mathbf{y}_{cal}{}^{T}\mathbf{h}^{T} \qquad \text{(P5-18)}$$

Se supone que los términos cruzados entre $d\mathbf{x}$ y $d\mathbf{X}$ tendrán contribuciones despreciables, debido a que el ruido de las distintas muestras no se encuentra correlacionado.

En el primer término, el valor esperado del producto ($d\mathbf{x}d\mathbf{x}^{T}$) es la matriz de covariancia del error de las señales de la muestra de *test*. En el segundo término, el producto ($d\mathbf{X}^{T}\mathbf{h}^{T}\mathbf{h}d\mathbf{X}$) genera una matriz de covariancia del error efectiva para las señales de la muestra de calibración.

Definiendo la matriz **H:**

$$\mathbf{H}= \mathbf{h}^{T}\mathbf{h} = \begin{bmatrix} h_1^2 & \dots & h_1 h_I \\ \dots & \dots & \dots \\ h_I h_1 & \dots & h_I^2 \end{bmatrix} \qquad \text{(P5-16)}$$

Expandiendo ($d\mathbf{X}^{T}\mathbf{h}^{T}\mathbf{h}d\mathbf{X}$) en término de los valores específicos

$$(d\mathbf{X}^{T})\mathbf{H}(d\mathbf{X}) = \begin{bmatrix} dx_{11} & \dots & dx_{1I} \\ \dots & \dots & \dots \\ dx_{J1} & \dots & dx_{JI} \end{bmatrix} \begin{bmatrix} h_1^2 & \dots & h_1 h_I \\ \dots & \dots & \dots \\ h_I h_1 & \dots & h_I^2 \end{bmatrix} \begin{bmatrix} dx_{11} & \dots & dx_{J1} \\ \dots & \dots & \dots \\ dx_{1I} & \dots & dx_{IJ} \end{bmatrix} \qquad \text{(P5-17)}$$

En la última ecuación, sólo se consideran los productos que contienen productos de valores de *dX* correspondientes a la misma muestra, dado que los valores esperados de los productos cruzados para diferentes muestras se pueden considerar iguales a 0. Esto lleva a la expresión particularmente simple:

$$E[(\mathrm{d}\mathbf{X}^{T})\mathbf{H}(\mathrm{d}\mathbf{X})\,] = \sum_{\mathrm{X},1} h_1^2 + \sum_{\mathrm{X},2} h_2^2 + \ldots + \sum_{\mathrm{X},\mathrm{i}} h_i^2 \qquad \text{(P5-18)}$$

donde $\mathbf{\Sigma}_{\mathrm{X},i}$ es la matriz de covariancia del error para la muestra de calibración *m*. Lo anterior permite definir, al igual que para primer orden, una matriz de covariancia del error efectiva $\mathbf{\Sigma}_{\mathrm{X,eff}}$, como el promedio pesado de todas las matrices de covariancia del error para las muestras de calibración:

$$\sum_{\mathrm{X,eff}} = \frac{1}{h}\sum_{\mathrm{X},1} h_1^2 + \sum_{\mathrm{X},2} h_2^2 + \ldots + \sum_{\mathrm{X},i} h_{\mathrm{i}}^2 \qquad \text{(P5-19)}$$

A partir de aquí resulta fácil mostrar que la variancia en la predicción está dada por la siguiente forma simple de expresión final:

$$\sigma_{\hat{y}}^2 = \mathbf{b}_{\text{eff}}^{\text{T}} \boldsymbol{\Sigma}_{\text{x}}^2 \mathbf{b}_{\text{eff}} + h\mathbf{b}_{\text{eff}}^{\text{T}} \boldsymbol{\Sigma}_{\text{X,eff}}^2 \mathbf{b}_{\text{eff}} + h\sigma_{\text{ycal}}^2 \quad \text{(P5-20)}$$

Esta es la expresión más completa que permite incluir todos los tipos de estructuras de ruido, ya que $\boldsymbol{\Sigma}_{\text{X,eff}}$ genera $\boldsymbol{\Sigma}_{\text{X}}$ cuando todas las muestras de calibración tienen la misma matriz de covariancia del error y a $\sigma_{\text{X}}^2\mathbf{I}_n$ en caso que el ruido sea iid. Es interesante notar que la leva se calcula a partir de los *scores* de la muestra de *test* luego de haber pasado por el ajuste RBL, es decir, luego del modelado de señales interferentes. Por otro lado, el efecto de los interferentes se hace presente en la ecuación anterior a través del vector de coeficientes de regresión efectivo $\mathbf{b}_{\text{eff}}$. Cualquier señal interferente que se solape con la de los analitos llevará a una menor sensibilidad en la determinación modificando el valor de $\mathbf{b}_{\text{eff}}$.

Si bien esta deducción es un punto de partida para generar un esquema general para el cálculo de incertidumbre en PLS/RML, al igual que el que se presentó en el Capítulo 3 para calibración de primer orden, todavía es necesario realizar las simulaciones de adición de ruido correspondientes para validarlo y aplicarlo sobre datos experimentales reales en presencia de réplicas. Además, como se mencionó durante el análisis de datos simulados, se podría pensar en una extensión de este procedimiento de propagación de errores para llegar a una fórmula específica en PARAFAC y en MCR-ALS.

5.10 Apéndice

Para llegar a la **Ecuación P5-4**, se puede considerar el caso más sencillo posible como son dos interferentes modelados por RBL ($\mathbf{c}_{\text{int,1}}\otimes\mathbf{b}_{\text{int,1}} + \mathbf{c}_{\text{int,2}}\otimes\mathbf{b}_{\text{int,2}}$), con dos sensores en un modo y tres en el otro:

$$\mathbf{c}_{\text{int,1}}\otimes\mathbf{b}_{\text{int,1}} + \mathbf{c}_{\text{int,2}}\otimes\mathbf{b}_{\text{int,2}} = \begin{bmatrix} c_{11}b_{11} + c_{11}b_{12} \\ c_{11}b_{21} + c_{11}b_{22} \\ c_{21}b_{11} + c_{21}b_{12} \\ c_{21}b_{21} + c_{21}b_{22} \\ c_{31}b_{11} + c_{31}b_{12} \\ c_{31}b_{21} + c_{31}b_{22} \end{bmatrix} \quad \text{(A5-1)}$$

donde el primer subíndice representa el número de sensor y el segundo el índice del interferente. El subíndice ‘int’ no se escribió en el lado derecho de la ecuación por cuestiones de simplicidad. La diferenciación de A5-1 conduce a:

$$d(\mathbf{c}_{int,1}\otimes\mathbf{b}_{int,1} + \mathbf{c}_{int,2}\otimes\mathbf{b}_{int,2}) = \begin{bmatrix} c_{11}db_{11} + dc_{11}b_{11} + c_{12}db_{12} + dc_{12}b_{12} \\ c_{11}db_{21} + dc_{11}b_{11} + c_{12}db_{22} + dc_{12}b_{22} \\ c_{211}db_{11} + dc_{21}b_{11} + c_{22}db_{12} + dc_{22}b_{12} \\ c_{21}db_{21} + dc_{21}b_{11} + c_{22}db_{22} + dc_{22}b_{22} \\ c_{31}db_{11} + dc_{31}b_{11} + c_{32}db_{12} + dc_{32}b_{12} \\ c_{31}db_{21} + dc_{31}b_{11} + c_{32}db_{22} + dc_{32}b_{22} \end{bmatrix} \qquad \text{(A5-2)}$$

Esta última ecuación se puede escribir como el producto de una matriz y un vector de diferenciales:

$d(\mathbf{c}_{int,1}\otimes\mathbf{b}_{int,1} + \mathbf{c}_{int,2}\otimes\mathbf{b}_{int,2}) =$

$$= \begin{bmatrix} c_{11} & 0 & b_{11} & 0 & 0 & c_{12} & 0 & b_{12} & 0 & 0 \\ 0 & c_{11} & b_{21} & 0 & 0 & 0 & c_{12} & b_{22} & 0 & 0 \\ c_{21} & 0 & 0 & b_{11} & 0 & c_{22} & 0 & 0 & b_{12} & 0 \\ 0 & c_{21} & 0 & b_{21} & 0 & 0 & c_{22} & 0 & b_{22} & 0 \\ c_{31} & 0 & 0 & 0 & b_{11} & c_{32} & 0 & 0 & 0 & b_{12} \\ 0 & c_{31} & 0 & 0 & b_{21} & 0 & c_{32} & 0 & 0 & b_{22} \end{bmatrix} \begin{bmatrix} db_{11} \\ db_{12} \\ dc_{11} \\ dc_{21} \\ dc_{31} \\ db_{12} \\ db_{22} \\ dc_{12} \\ dc_{22} \\ dc_{32} \end{bmatrix} =$$

$= \mathbf{Z}_{int}\,[d\mathbf{b}_{int,1}\,;\, d\mathbf{c}_{int,1}\,;\, d\mathbf{b}_{int,2}\,;\, d\mathbf{c}_{int,2}] =$

$= [\mathbf{c}_{int,1}\otimes\mathbf{I}_b\,,\, \mathbf{I}_c\otimes\mathbf{b}_{int,1}\,,\, \mathbf{c}_{int,2}\otimes\mathbf{I}_b\,,\, \mathbf{I}_c\otimes\mathbf{b}_{int,2}]\,[d\mathbf{b}_{int,1}\,;\, d\mathbf{c}_{int,1}\,;\, d\mathbf{b}_{int,2}\,;\, d\mathbf{c}_{int,2}]$

(A5-3)

6. CONCLUSIÓN GENERAL

Claramente, los Capítulos 3, 4 y 5 de esta tesis, constituyen su hilo conductor. Durante estos capítulos, se mostró que el cálculo de cifras de mérito en calibración multivariada y multi-vía no resulta de una extensión simple y directa de las expresiones univariadas, por lo que el tema requiere una atención especial.

Al realizar un análisis de las distintas estructuras y formas de procesar los datos, tal como se presentó en el primer capítulo, surge la necesidad de estudiar y racionalizar la definición de estimadores, dependiendo no sólo de la muestra que se esté analizando, sino también de la mecánica de funcionamiento del algoritmo que se esté utilizando y de la estructura de error que esté afectando al sistema en estudio. En este sentido, el desarrollo presentado en la sección teórica del Capítulo 5, que constituye uno de los últimos eslabones en la obtención de una fórmula general para el cálculo de la sensibilidad, que se puede aplicar a todos los órdenes de datos y a las herramientas de procesamiento más utilizadas. En un contexto más amplio, una expresión de tal generalidad, constituye un paso importante hacia un mejor entendimiento de la información necesaria para desarrollar metodologías de validación multi-vía confiables.

La definición del límite de detección, abordada durante el Capítulo 4, también constituye un punto importante de analizar ya que esta cifra de mérito combina dos conceptos analíticos importantes: la sensibilidad y la precisión de la determinación analítica. En este contexto, en esta tesis se realizaron avances para poder llegar a un estimador confiable del LOD basado en criterios no sólo matemáticos sino también analíticos. De cualquier modo para extender estos conceptos a datos multi-vía, deberían realizarse estudios más detallados.

Otra cifra de mérito que se estudió en esta tesis es la incertidumbre en la predicción. Es notorio que en muchos trabajos de la literatura y tesis vinculadas a la química analítica, esta cifra muchas veces no se informa, ya sea por desconocimiento o porque se subestima su importancia. Este no es un hecho menor, especialmente si se tiene en cuenta que esta cifra es la base sobre la cual se estima el límite de detección, y sobre la que es posible evaluar la calidad de una medición. Por lo tanto, la extensión de las expresiones que se habían propuesto hasta el momento, a casos en los que la estructura del error no es iid, es una temática de fundamental importancia en lo que respecta a los cimientos de la química analítica como ciencia que tiene como uno de sus objetivos primordiales mejorar la calidad

de las mediciones químicas. Para esto, se incursionó detenidamente en los distintos tipos de ruido instrumental que pueden llegar a influir en las mediciones, para luego, por medio de una metodología de propagación de errores que tenga en cuenta los pasos del modelo matemático de regresión que se esté utilizando, llegar a una estimación confiable de la incertidumbre en la concentración estimada.

El Capítulo 2, si bien no está estrictamente vinculado a la estimación de cifras de mérito, hace uso de una de ellas (el RMSEP) para poder optimizar uno de los modelos de calibración multivariada de primer orden más utilizados en la actualidad, como lo es PLS. Este capítulo, se aparta de la dinámica de estudio de cifras de mérito para evaluar los modelos de predicción, que domina los capítulos siguientes. Lo anterior se realizó con el fin de desarrollar un algoritmo de optimización de estos modelos. Para desarrollar este algoritmo se indagó sobre uno de los principios básicos a la hora de poner a prueba y validar un modelo de calibración multivariados: la selección de muestras y variables modificará el tipo de preprocesamiento necesario, a la vez que estos modificarán las variables significativas. Esta mecánica intuitiva que generalmente se ejecuta por medio de un mecanismo de prueba se automatizó utilizando algoritmos estocásticos basados en computación natural y guiados por una función objetivo.

En términos generales, a lo largo de esta tesis se intentó mostrar que el empleo criterioso de herramientas matemáticas, estadísticas y computacionales permite un enriquecimiento constructivo para el desarrollo sólido y bien fundamentado de una de las ramas de la química con mayor diversidad de aplicaciones, como lo es la química analítica.

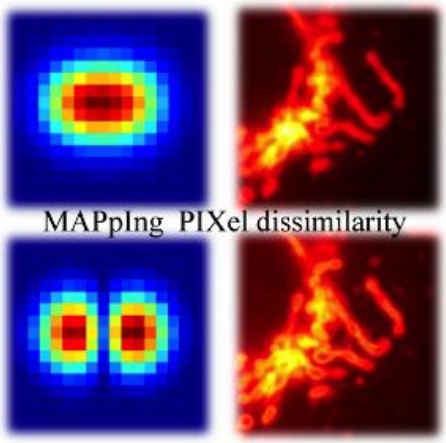
"*La originalidad consiste en el retorno al origen. Así pues, original es aquello que vuelve a la simplicidad de las primeras soluciones*." (Antoni Gaudi).

7. ANEXO

El tema que se desarrollará a continuación se incorporó como un anexo de esta tesis ya que la temática abordada, si bien forma parte del campo de estudio de la química analítica y utiliza recursos provenientes de la quimiometría, se aparta significativamente de las líneas que se trataron durante los capítulos anteriores. Es por este motivo que las abreviaturas y el lenguaje especial de símbolos se tratan aparte y no se mezclan con los del resto de la tesis

Los resultados que se presentarán son parte de un trabajo realizado durante una estadía en la Universidad de Lille, Francia en el marco del programa de becas para estudiantes de doctorado denominado Eurotango II. El objetivo principal de este trabajo fue el desarrollo de un modelo de post-procesamiento de imágenes obtenidas por microscopía de fluorescencia, con el objeto de superar el límite físico impuesto por el fenómeno óptico de difracción y extraer mayor información visual que la que se obtienen de las imágenes “en crudo”.

MAPAS DE DISIMILITUD APLICADOS A LA MEJORA DE IMÁGENES EN MICROSCOPÍA DE FLUORESCENCIA DE ALTA RESOLUCIÓN



"Lo escencial es invisible a los ojos…o casi."

7.1 Resumen

Los últimos avances en la obtención de imágenes biológicas por microscopía de fluorescencia, con sensibilidades que llegan al nivel de las moléculas individuales, dependen del análisis y la visualización de datos obtenidos a partir de fluoróforos cuya intensidad oscila entre estados brillantes y oscuros en la medida que pasa el tiempo. En este capítulo se describirá una técnica conocida como mapeo de la disimilitud entre pixeles (Mappix) demostrando que es posible mejorar esta visualización representando a los pixeles como disimilitudes entre las señales de las fluctuaciones fluorescentes. Esta disimilitud se calcula respecto de la media de las señales a través de todos los pixeles. Mappix resulta en una mayor extracción de información visual de las imágenes obtenidas. Más aún, permite evitar un sesgo en la distribución espacial del brillo fluorescente a diferencia de otras técnicas similares. Esto posibilita que las grandes diferencias de brillo entre los distintos fluoróforos se puedan controlar adecuadamente, lo que resulta crítico para lograr una buena fidelidad en la imagen final. El método Mappix se pondrá a prueba con datos tanto simulados como reales. Los resultados obtenidos para células HEK que expresan la proteína fluorescente y fotoconvertible Dronpa, muestran que para muestras con una gran densidad de emisores, se pueden obtener mejoras en las imágenes posibilitando un nivel de detalle mayor a la hora de extraer información estructural. A

pesar de algunas limitaciones, la comparación con los métodos desarrollados hasta el momento revela que Mappix puede ser un complemento muy útil en aplicaciones relacionadas con la obtención de imágenes fluorescentes para aplicaciones biológicas.

7.2 Introducción

La microscopía de fluorescencia es una técnica analítica cualitativa que tradicionalmente tuvo gran aplicación en biología permitiendo visualizar estructuras y subestructuras celulares. Sin embargo, la resolución de los microscopios convencionales está limitada por el límite óptico del fenómeno de difracción. La microscopía funcional de subdifracción se basa en el uso de sondas brillantes capaces de emitir radiación fluorescente y que pueden alternar entre dos estados diferentes (un estado de emisión o de "*on*" y otro de no emisión o de "*off*"). Por otro lado, la superresolución funcional se logra a través de distintas técnicas de obtención y análisis de imágenes, que permitien una inspección más profunda de las distintas estructuras celulares. [140,141] Entre estas técnicas se encuentran las comúnmente denominadas "deterministas" como son STED (*STimulated Emission Depletion*),[142]RESOLFT (*Reversible Optically Linear Fluorescence Transitions Microscopy*),[143,144,145] NSIM (*Non-linear Structured Illumination Microscopy*),[146] e ISM (*Image Scanning Microscopy*).[147] Estas técnicas combinan innovaciones en óptica e iluminación, así como en las propiedades de los fluoróforos para alcanzar una resolución espacial por debajo de límite de difracción.

Otro grupo de técnicas se conoce como estocásticas de super-resolución. Estas tienen un campo de aplicación más amplio que las anteriores y sacan provecho de la naturaleza de la emisión fluorescente individual de cada molécula. De esta manera, las señales surgen de la activación estocástica y separada de fluoróforos individuales en una muestra determinada, y de su observación a través de miles de cuadros de imágenes. La estrategia más directa para el análisis de estos datos consiste en localizar cada uno de los emisores individuales, ajustando una función Gaussiana en dos dimensiones al cono de emisión de cada fluoróforo (comúnmente denominado como PSF (*Point Spread Function*), en caso que la distribución de los emisores esté lo suficientemente separada en los distintos cuadros de imágenes.[148] El proceso de localización se repite a través de las sucesivas imágenes que se van obteniendo en el tiempo y las posiciones resultantes se suman para generar una imagen de superresolución. La localización de la posición de fluoróforos individuales es la base sobre la cual se diseñaron la mayoría de los métodos

estocásticos como STORM (*STochastical Optical Reconstruction Microscopy*)[149] y PALM (PhotoActivation Localization Microscopy).[140,141,150] Normalmente, cuando se logra resolver cada uno de los emisores individualmente, se pueden alcanzar resoluciones que se encuentran en el orden de los 10 nm. Sin embargo, el requerimiento subyacente de una baja densidad de fluoróforos, tiene algunas limitaciones en lo que respecta a las aplicaciones vinculadas a la visualización de células vivas.

Recientemente se publicaron una serie de métodos basados en el ajuste de funciones Gaussianas múltiples.[151,152] Sin embargo, en este sentido, se pueden cuestionar la confiabilidad y el costo computacional comparado con las rutinas de localización de emisores de manera individual. También se desarrollaron una serie de estrategias basadas en la resta de imágenes sucesivas,[153]que si bien tienen sus limitaciones (como por ejemplo para emisores que se mantienen activos durante una serie de varias imágenes), muestran la utilidad del principio de aplicar estrategias de localización sobre datos previamente procesados en lugar de hacerlo directamente sobre los datos crudos. Alternativamente, se propuso una metodología de tipo Bayesiana.[154] Sin embargo, el desarrollo de estrategias capaces de manipular imágenes con una densidad de emisores lo suficientemente alta como para llegar a obtener información visual confiable, todavía constituye uno de los problemas relevantes de la microscopía de fluorescencia de alta resolución.[155]

Como alternativa a los procedimientos de localización, algunos métodos como STICS (*Spatiotemporal Image Correlation Spectroscopy*)[156] dependen del análisis estadístico de las fluctuaciones de intensidad fluorescente en cada pixel para extraer información de los procesos dinámicos que se dan como consecuencia de la irradiación y la emisión fluorescente en cada muestra. Otro ejemplo de esta forma de analizar los datos, es la técnica SOFI (*Super-resolution Optical Fluctuation Imaging*)[157,158] que permite alcanzar una resolución del nivel de la subdifracción para estructuras que contienen una gran densidad de emisores adosados. Esta última, se basa en la integración de lo que se conoce como función de autocorrelación en el tiempo de la señal para cada pixel, lo que genera un nuevo conjunto de valores que permiten incrementar la resolución en la nueva imagen generada. Hoy en día, SOFI es el método de superresolución basado en análisis de correlación más conocido y utilizado. Una ventaja importante es su compatibilidad con una amplia gama de modalidades de obtención de imágenes y de condiciones de emisión y desactivación, así como el hecho que, al menos teóricamente, la resolución que puede alcanzar es ilimitada. Sin embargo, dado que SOFI opera reduciendo el ancho de la PSF de

manera más significativa en la medida que se incrementa el orden del cálculo de la función de autocorrelación, el brillo de cada emisor se presentará en la imagen final elevado a la *n*-ésima potencia. Si los emisores presentan un rango de brillos diferentes, esto resultará en un sesgo de las contribuciones de los emisores, lo cual en última instancia complica la representación de la imagen y obstaculiza la interpretación de las muestras.

En este anexo, se propondrá una alternativa para procesar las fluctuaciones fluorescentes presentes en imágenes obtenidas por microscopía de fluorescencia con una sensibilidad que puede llegar al orden de las moléculas individuales. La idea de esta metodología se basa en evaluar la disimilitud de las señales de los pixeles con respecto a la media de la señal calculada a través de todos los pixeles, resultando en una imagen que consiste en los valores de disimilitud calculados para cada uno de los pixeles de la imagen. Intuitivamente, la disimilitud puede entenderse como una medida de la correlación y la distancia entre dos vectores de señal. En la literatura, se utilizaron diferentes criterios para evaluar la disimilitud entre señales[159] con la idea común y general de resaltar características que son difíciles de evaluar por medio de aproximaciones estadísticas, como pueden ser la forma o la semejanza. En este trabajo, la disimilitud se estimó adaptando un criterio quimiométrico que originalmente se diseñó para distinguir los espectros más puros en cuanto al contenido de un determinado analito de un conjunto de espectros similares en cuanto a su forma.[160] La ventaja de este procedimiento es que la intensidad de los pixeles detectores que contienen información generada por solapamiento de las señales de las fuentes emisoras, se atenuarán de manera significativa permitiendo distinguir entre fuentes emisoras adyacentes. Por lo tanto Mappix (*Mapping Pixel Dissimilarity*) lleva a un aumento del contraste que permite resolver emisores que se encuentran muy cercanos, superando el límite de difracción. Mappix es una herramienta que sirve para aumentar los detalles a la hora de visualizar imágenes pero no puede considerarse estrictamente como una técnica que opere a nivel subdifracción, al mismo tiempo que todavía no es posible proporcionar una descripción matemática rigurosa de la imagen. En el transcurso de este anexo se realizará una comparación entre SOFI y Mappix sobre datos simulados y sobre células vivas que expresan la proteína fotoconvertible fluorescente Dronpa. El método propuesto tiene una relevancia particular en la obtención de imágenes de alta sensibilidad para mejorar la visualización de estructuras biológicas por microscopía de fluorescencia debido a que los resultados son más robustos a la escala de brillos y a la densidad de emisores.

7.3 Disimilitud entre pixeles

Los datos provenientes de la microscopía de fluorescencia sensible a moléculas individuales consisten normalmente en un apilamiento de *K* cuadros de imágenes donde *K* puede superar varios de miles. Cada cuadro contiene *n* x *n* valores de intensidad de pixeles. La seña asociada a cada pixel *i* corresponde a un vector $\mathbf{p}_i$ de *K* elementos que contiene las intensidades de fluorescencia obtenidas a través de un determinado tiempo de medición. La media de todos los vectores asociados a pixel $\mathbf{p}_i$ se normaliza a longitud unitaria.

La disimilitud entre dos objetos mide la independencia entre la secuencia de medidas que representan a estos objetos. Esta disimilitud se puede estimar de varias maneras basadas en correlación y/o distancia. En este trabajo, la disimilitud d_i se define como la magnitud del producto vectorial entre un vector asociado a un determinado pixel $\mathbf{p}_i$ y la media generada sobre todos los vectores de pixeles, **r**, tal como muestra la **Ecuación 7.1**

$$d_i = \mathbf{p}_i \times \mathbf{r} = \|\mathbf{p}_i\|\|\mathbf{r}\| \operatorname{sen} \alpha_i \qquad (7.1)$$

donde la notación $\|\ \|$ corresponde a la norma de un vector tomada en la dimensión del tiempo de la señal (recordar que $\|\mathbf{r}\|=1$) y α_i es el ángulo entre el vector asociado a un determinado pixel y el vector medio. La magnitud del producto vectorial entre dos vectores se puede calcular fácilmente como el determinante de una matriz cuadrada que a su vez se genera a partir del producto de una matriz **Y** compuesta por el vector de referencia y el vector que contiene las señales en el tiempo del pixel a comparar, tal y como se muestra en la **Figura 7.1**.

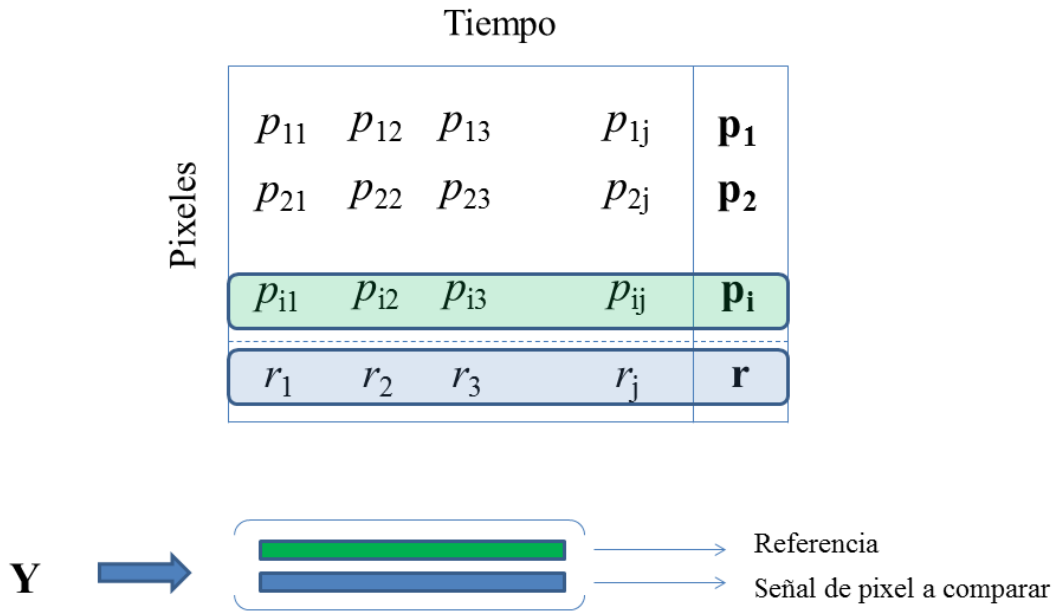


Figura 7.1. Representación de los datos analizados. En primer lugar se desdoblan las imágenes generando una matriz de datos conteniendo la señal fluorescente de cada pixel a cada uno de los tiempos de adquisición de las imágenes.

La manera en que se evalúa la disimilitud se puede entender intuitivamente a partir de la interpretación geométrica del producto cruzado. La magnitud del producto cruzado se puede interpretar como el área positiva del paralelogramo que tiene como lados a **p** y a **r** (**Figura 7.2**). Este valor tiende a 0 en la medida que dos vectores tienen la misma dirección o cuando la longitud de uno de los vectores se aproxima a 0. Mientras mayor sea la diferencia angular entre los vectores **p** y **r**, y mayor sea la norma de la señal del vector (pixel más brillante), mayor será el valor de disimilitud. Otro punto a tener en cuenta y que se puede derivar directamente de la **Ecuación 7.1** es que suponiendo un valor constante del ángulo α entre dos vectores, el valor de intensidad de brillo fluorescente para un determinado pixel tiene una relación lineal con el valor de disimilitud, lo cual resulta importante a la hora de conservar la relación de brillo original de la imagen.

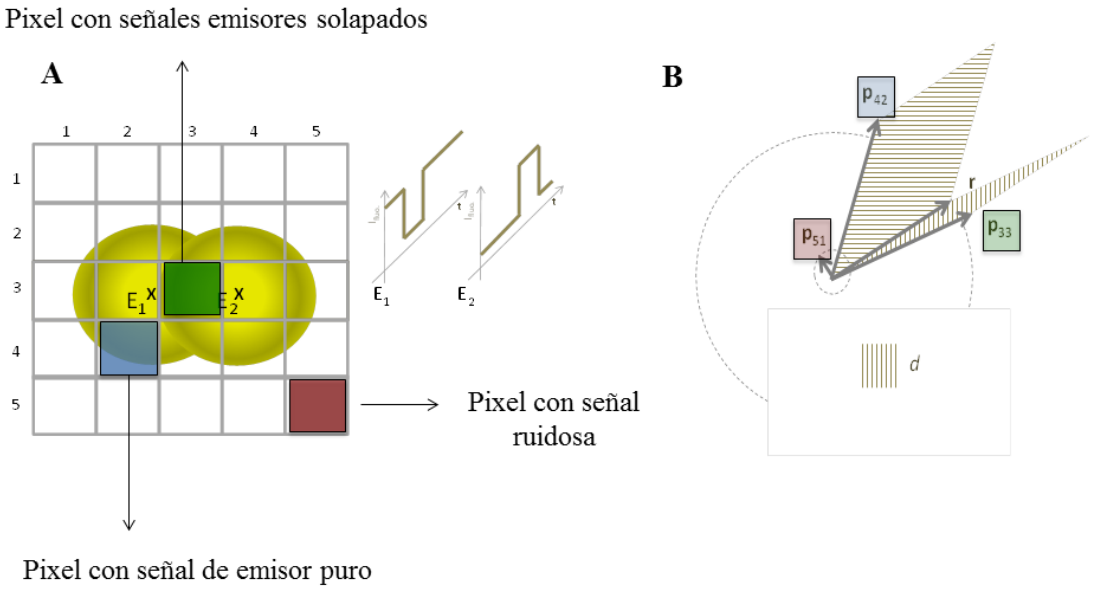


Figura 7.2. Interpretación geométrica del concepto de disimilitud. (A) Ilustración esquemática de los distintos tipos de pixeles que pueden presentarse en una imagen. (B) Representación de la magnitud del producto vectorial entre los vectores correspondientes a las señales de cada uno de los pixeles resaltados en la cuadrilla.

Una característica particularmente interesante en lo que respecta a la propuesta de este anexo es que se pueden obtener valores bajos de disimilitud para dos situaciones diferentes. La primera es para pixeles con valores muy bajos de señal (es decir pixeles en los que sólo hay ruido del detector). La segunda, de gran interés en lo que respecta a potenciar los detalles de la imagen final en el contexto de la microscopía de fluorescencia, se relaciona con valores similares entre la señal de un determinado pixel y el vector de señal media **r**. Esto puede evidenciarse en el caso de pixeles que contienen contribuciones de múltiples emisores, que son pixeles que se aproximan al vector medio. Como consecuencia de la característica anterior, la intensidad de Mappix se reducirá para pixeles que detecten fluorescencia solapada de distintos emisores.

7.4 Imágenes obtenidas a partir de Mappix

La **Figura 7.3** muestra una representación esquemática del método Mappix a través de un experimento conceptual que incluye a dos fluoróforos poco espaciados cuya intensidad fluctúa en el tiempo. Debido a la naturaleza fluctuante de los emisores, cada pixel del detector muestra un nivel de fluorescencia que varía en el tiempo (**Figura 7.3 A**).

Estos valores son las componentes del vector **p** correspondiente a un pixel determinado. La magnitud de estos vectores se corresponde con la intensidad media calculada a través de cada uno de los pixeles. La imagen original, correspondiente al promedio de la película de imágenes se muestra en la **Figura 7.3 B**. Los resultados obtenidos cuando se calcula la disimilitud entre pixeles se muestran en la **Figura 7.3 C**. Para pixeles brillantes en la **Figura 7.3 B,** es decir cuando las emisiones de dos emisores se solapan, en la **Figura 7.3 C** se observa la separación de dos focos de emisión que originalmente se encontraban muy solapados.

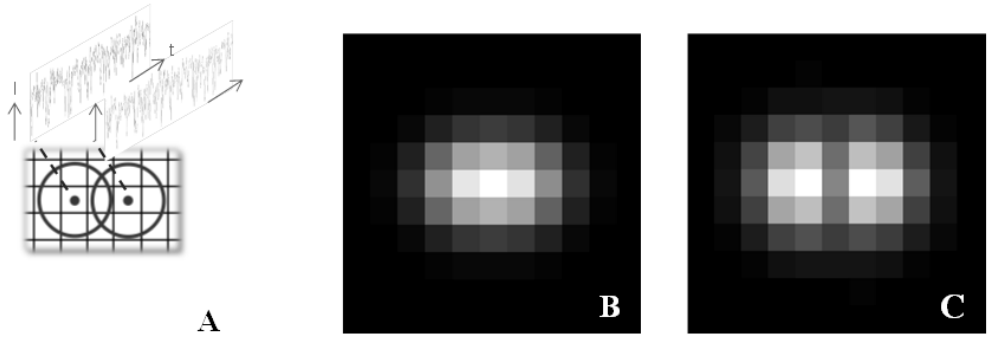


Figura 7.3. Principio de Mappix, basado en la evaluación y el mapeo de los valores disimilitud para imágenes obtenidas por microscopía de fluorescencia. (A) Señales de dos emisores vecinos separados por una distancia de dos pixeles obtenidas para pixeles de una cámara CCD. Cada pixel contiene un determinado patrón de fluctuación fluorescente en el tiempo. (B) Imagen original. Los pixeles blancos corresponden a señales más brillantes. (C) Imagen obtenida a partir de los valores de disimilitud. Los pixeles oscuros corresponden a valores bajos de disimilitud. Su puede observar una separación vertical entre los dos emisores, resultando en la separación de las dos fuentes de emisión. Los dos emisores se pueden detectar en la medida que la distancia entre ellos es mayor que 1.5 veces el tamaño de un pixel.

Mappix es un método que actúa a nivel de los pixeles y, como tal, está limitado por el tamaño óptico finito de los pixeles de la cámara. Dos emisores podrán distinguirse entre sí si están separados a una distancia de al menos 1.5 pixeles incluso en situaciones en las que las emisiones individuales son difíciles de observar selectivamente en el tiempo. Sin embargo, es importante notar que donde las señales se solapan, la forma de la señal procesada resultante cambia y ya no puede aproximarse como una suma de Gaussianas.

En principio, existen dos alternativas para obtener las imágenes finales por Mappix, como se muestra en la **Figura 7.4**. Una es considerar el conjunto completo de cuadros de imágenes (**Figura 7.4 A**) y la otra es repetir el procedimiento para secuencias de imágenes más cortas (**Figura 7.4 B**). En el último caso, la salida consiste en múltiples imágenes y los resultados se promedian para generar la imagen final de Mappix. Para una secuencia de imágenes corta, la disimilitud se calcula respecto de la media calculada a partir de las imágenes que componen la secuencia. Esto permite corregir la variación dinámica de la señal que podría darse en una escala de tiempo más larga que la fluctuación de cada uno de los emisores individualmente. La elección del número de cuadros por secuencia depende del tipo de datos. Normalmente se obtienen buenos resultados cuando se utilizan alrededor de 10 imágenes por secuencia para un total de alrededor de 1000 imágenes.

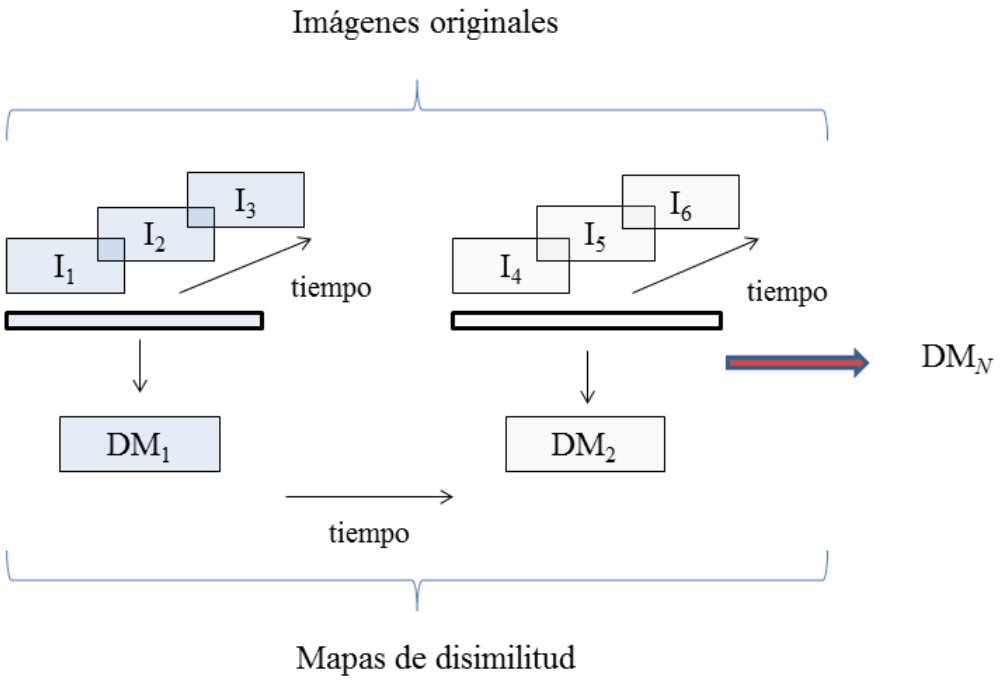


Figura 7.4. Esquema representativo de la estrategia utilizada para construir los mapas de disimilitud (DM) agrupando imágenes sucesivas (I).

Otra característica de Mappix, es que se puede calcular para la imagen completa (todos los pixeles) o localmente para una determinada zona de la imagen. Sin embargo, se debe tener en cuenta que en caso de proceder como se mencionó anteriormente el vector medio para cada zona de la imagen será distinto, por lo cual la reconstrucción de la

imagen global a partir de análisis locales no es una buena alternativa y todavía no se exploró en más detalle.

7.5 Software utilizado

Las rutinas de Mappix se escribieron utilizando el software MATLAB versión 7.4.0 (R2007a) o superior. Estas rutinas están disponibles por solicitud a los autores. Las imágenes de SOFI se obtuvieron por medio del *pack* Localizer [161], disponible de forma libre y gratuita, y que implementa las técnicas computacionales de procesamiento de datos más empleadas para los distintos tipos de imágenes de microscopía de fluorescencia de superresolución.

7.6 Generación de datos simulados

Para evaluar la metodología que se presentó en las secciones 7.2 y 7.3 se simularon datos de microscopía de fluorescencia estocástica. Para esto se posicionaron los supuestos emisores en una grilla regular que representa los pixeles de la cámara del microscopio. La intensidad relativa de cada fluoróforo en cada pixel se obtiene por integración numérica de una distribución normal acumulativa con un ancho a media altura (FWHM) de 270 nm. Los emisores alternan entre distintos niveles de brillo y oscuridad y se pueden describir a través de un modelo de dos estados en el que $k_{on} = 1/\tau_{on}$ y $k_{off} = 1/\tau_{off}$, son las constantes de velocidad desde el estado de "*on*" a "*off*" y de "*off*" a "*on*", respectivamente. En el caso de las simulaciones que se presentarán más adelante, el tiempo esperado de actividad (τ_{on}), se fijó en torno a 2 s. y el tiempo esperado de oscuridad (τ_{off}) en torno a 0.5 s. Estos tiempos característicos se muestrearon a partir de una distribución exponencial con tiempos de decaimiento que no se sincronizaron con los de adquisición. Se asumió una emisión de 10000 fotones por segundo para cada fluoróforo, y una duración por cuadro de imagen de 1 s. En consecuencia, para obtener el número de fotones emitidos por un fluoróforo en un determinado cuadro de adquisición, la velocidad de emisión se multiplicó por el tiempo que el fluoróforo permaneció en este cuadro. El ruido electrónico se adicionó utilizando una distribución Gaussiana. Se generaron distintos juegos de datos generando 1000 cuadros de imágenes, en una grilla de pixeles de 40×40 suponiendo un tamaño de pixel de 30 nm. Para las simulaciones con dos emisores, estos se posicionaron de tal manera que la distancia entre ellos fuera de 210 nm. Por otro lado, en el caso del sistema de cuatro emisores igualmente espaciados en una línea, éstos se posicionaron de tal manera que la

distancia entre ellos fuera de 210 nm. En esta disposición, se reportaron los resultados obtenidos en dos situaciones. En el primer caso, la intensidad de los emisores solapados es la misma, mientras que para el otro caso se cambió la intensidad entre emisores sucesivos, estableciendo un límite en la cantidad máxima de fotones que puede generar un emisor.

7.7 Generación de imágenes reales por microscopía de fluorescencia

Los experimentos de imágenes sobre células vivas se realizaron en laboratorios pertenecientes a la Universidad de Leuven (Bélgica), utilizando un sistema Olympus Cell*TIRF que funciona por el principio de reflexión total interna y está equipado con fuentes de laser de 488 y 405 nm, un objetivo Olympus 150x NA (*Numerical Aperture*) 1.45, y una cámara Hamamatsu ImageEM con sensor CCD. El tamaño óptico de los pixeles es de 100 nm.

Las células utilizadas fueron células embrionarias de hígado (HEK), cultivadas en un medio con los nutrientes y condiciones necesarias para mantenerlas vivas. 24 horas previas a la visualización, las células se transfectaron con un plásmido que codifica una proteína fluorescente (DAKAP-Dronpa). Finalmente, las células se lavaron 3 veces con una solución de sales balanceadas y se llevó adelante la adqusición a temperatura ambiente, utilizando un láser de 488 nm.

Para cada célula se obtuvieron aproximadamente 1000 imágenes utilizando una exposición por cuadro de 33 ms. La ganancia del multiplicador se fijó aproximadamente en 300 de acuerdo con las indicaciones de los fabricantes.

7.8 Resultados obtenidos en simulaciones

La habilidad de los métodos estocásticos de superresolución para resolver características estructurales depende de la densidad de etiquetas fluorescentes y del potencial para separar la emisión de fluoróforos individuales en el tiempo.[162] Un valor de τ_{off} mucho mayor que τ_{on}, es crucial para lograr una localización exitosa. Como se informó en la literatura, se requiere un cociente de alrededor de 1000 para separar dos filamentos adyacentes ubicados a una distancia de 30 nm., utilizando la metodología de localización STORM (en fluoróforos ubicados cada 8.5 nm) [163] Teniendo en cuenta este límite, en este capítulo anexo, se investigó la posibilidad de detectar y visualizar estructuras marcadas para las cuales este cociente es mucho menor. Es importante destacar que los

procedimientos de localización de moléculas individuales no se pueden aplicar en estas condiciones debido a la existencia de más de un emisor activo al mismo tiempo en un área limitada por el solapamiento generado por el límite óptico de difracción.

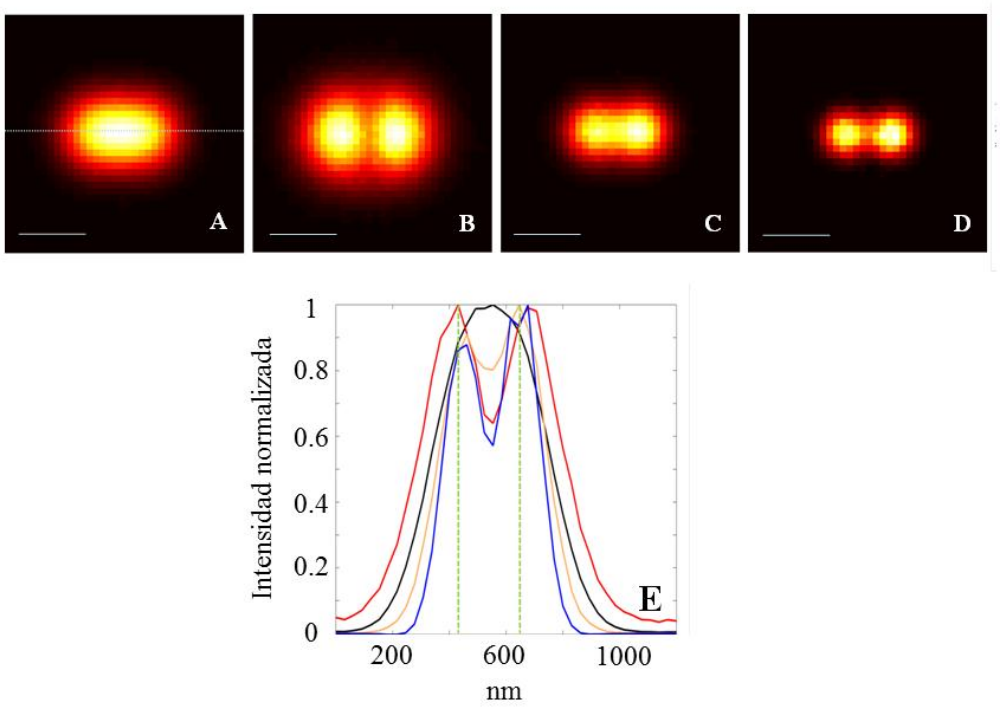


Figura 7.5. Comparación de los resultados obtenidos por Mappix y SOFI en datos simulados. (A) Imagen original correspondiente a dos emisores cercanos separados por 210 nm para una PSF de 270 nm (FWHM). (B-D) Imágenes obtenidas aplicando Mappix, SOFI de 2do. orden y SOFI de 3er. orden respectivamente. (E) Perfiles de intensidad normalizados y extraídos a lo largo de la línea de puntos dibujada en la parte A. La línea sólida negra corresponde a la señal de la imagen original, la línea sólida roja a Mappix, y las líneas sólidas de color anaranjado y azul a SOFI de 2do. y 3er. orden respectivamente. Barra de escala: 300 nm.

La **Figura 7.5** muestra los resultados que se obtuvieron al simular dos emisores individuales separados por una distancia de 210 nm, que es una distancia menor al límite de difracción. Estos emisores tienen un radio de *on* a *off* de 0.25 con un valor de τ_{off} dos veces menor que el tiempo de adquisición. Los resultados que se obtienen al aplicar Mappix (**Figura 7.5 B**) aparecen junto a las imágenes resultantes del tratamiento por medio de SOFI, utilizando funciones de autocorrelación de segundo y tercer orden (**Figura 7.5 C** y **7.5 D**, respectivamente). En primer lugar, se puede notar que todas las imágenes muestran una ganancia en detalles e información estructural respecto de la imagen

promedio. Claramente, cuando se utiliza SOFI se observa un incremento en la resolución, siendo este más notorio cuando se utiliza SOFI de tercer orden que es la técnica con la que, en este caso, se obtuvieron mejores resultados. A partir de los perfiles en la **Figura 7.5 E**, también se puede ver que el contraste para la imagen Mappix es el mismo que para la correspondiente a SOFI de 3er. orden, mostrando una buena separación entre las dos fuentes.

A partir de los resultados en la **Figura 7.5**, puede realizarse una discusión acerca de la diferencia conceptual entre Mappix y SOFI. Mappix es una estrategia de procesamiento de la señal que puede ayudar a revelar características estructurales (separando dos emisores vecinos). Sin embargo, al contrario de SOFI, Mappix no proporciona ningún tipo de mejora en cuanto a la reducción de la PSF. En consecuencia, al utilizar Mappix no se obtiene ningún tipo de mejora teórica en la resolución, como se puede ver al comparar los tamaños de las respectivas imágenes en las **Figuras 7.5 A** y **7.5 B**. Sin embargo, el método sigue teniendo su utilidad a la hora de revelar información estructural, debido a que los pixeles en los cuales hay solapamiento entre las señales de las dos fuentes emisoras tienen valores de disimilitud muy bajos. Mappix será capaz de separar y detectar los emisores en la medida que existan pixeles intermedios entre los puntos de emisión. Comparando con SOFI, el cálculo de la *n*-ésima función de autocorrelación genera una imagen en la que resolución se incrementa en un orden dado por un factor igual a la raíz de *n* para la PSF Gaussiana.[157]

En la **Figura 7.6**, se muestran los resultados que se obtienen para una simulación de 4 fluoróforos igualmente espaciados y separados por una distancia de 210 nm. Para los datos simulados que se muestran en la fila que comienza con la **Figura 7.6 A**, la intensidad de los emisores es la misma, mientras que en la fila que se inicia en la **Figura 7.6 F**, se varió la intensidad entre emisores sucesivos de manera decreciente de izquierda a derecha. Al igual que para las simulaciones anteriores, se informan los resultados que se obtienen al aplicar Mappix (**Figuras 7.6 B** y **7.6 G**) y SOFI en sus dos variantes (**Figuras 7.6 C** y **7.6 H** para SOFI de segundo orden y 4D y 4I para SOFI de tercer orden). En primer lugar, se puede ver que todos los métodos permiten alcanzar una mejora en las imágenes. Para realizar una comparación más exhaustiva, las **Figuras 7.6 E** y **7.6 J** muestran las líneas de perfiles a través de las PSF de cada emisor para las dos situaciones analizadas. Los perfiles de la **Figura 7.6 E** resaltan la habilidad de los diferentes métodos para separar emisores solapados con la misma intensidad. Como puede observarse, Mappix genera imágenes con

una ganancia de información comparable a SOFI de segundo orden en términos de separación de las fuentes de emisión. Por su parte, SOFI de tercer orden posibilita una mejora aún mayor en la resolución, al igual que en el ejemplo anterior.

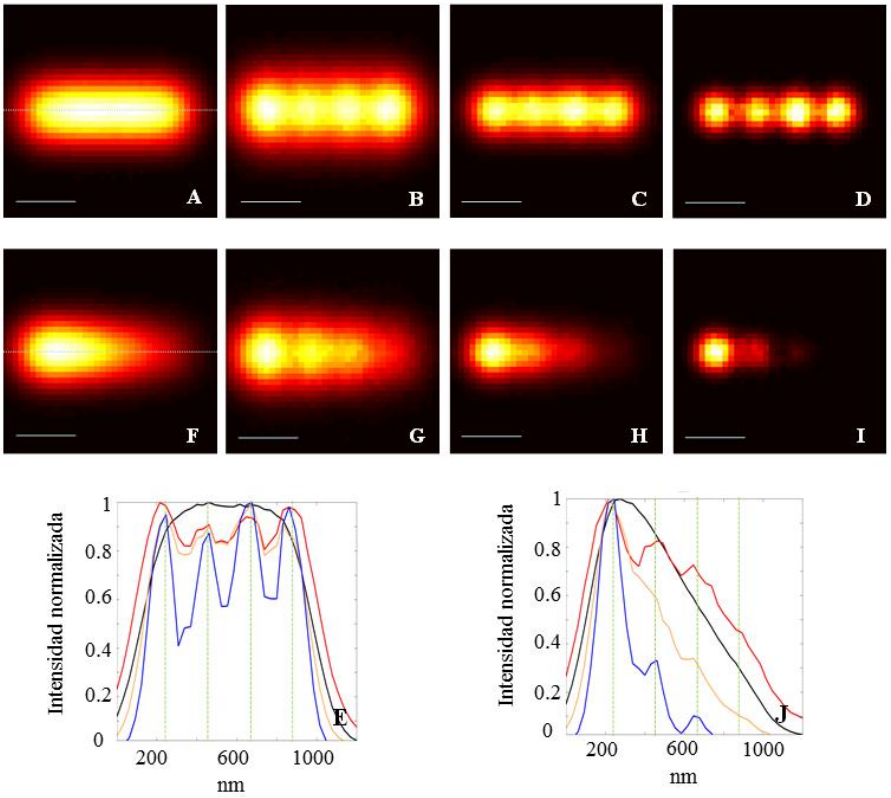


Figura 7.5. Comparación de resultados obtenidos por Mappix y SOFI en datos simulados. (A) Imagen original correspondiente a cuatro emisores vecinos con la misma intensidad separados por 210 nm y una PSF de 270 nm (FWHM). (B-D) Imágenes resultantes de aplicar Mappix, SOFI de 2do. orden y SOFI de 3er. orden respectivamente. (E) Perfiles de intensidad normalizados obtenidos a partir de la línea de puntos en la parte A. (F) Imagen original correspondiente a 4 emisores vecinos con radios de intensidad de 1,0.75, 0.5, y 0.25. (G-I) Mappix, SOFI de 2do. orden y SOFI de 3er. orden, respectivamente. (J) Perfiles de intensidad normalizados extraídos a lo largo de la línea punteada en la parte F. Código de colores en los gráficos E y J: la línea sólida negra corresponde a la señal de la imagen original, la línea sólida roja a Mappix y las líneas anaranjadas y azules a SOFI de 2do. y 3er. orden, respectivamente. Barra de escala: 300 nm.

Cuando los distintos emisores tienen diferentes intensidades, la situación anterior se modifica, tal como muestran los perfiles de intensidad de la **Figura 7.6 J**. En los dos casos de SOFI, puede observarse una alteración importante en el brillo de los distintos emisores debido a que la intensidad se eleva a la potencia de *n* para un cálculo de SOFI de orden *n*.

Esto puede llevar a una distorsión de los resultados, lo cual puede observarse en las **Figuras 7.6 H** y **7.6 I** donde las contribuciones de los emisores más débiles están enmascaradas por las de los más brillantes. En SOFI de segundo orden, la presencia de los emisores ubicados hacia la derecha, casi no puede detectarse porque el contraste es muy bajo. En SOFI de tercer orden, a pesar de cierta mejora teórica en la resolución espacial, la pérdida de información es incluso mayor. Por el contrario, los perfiles en rojo de la **Figura 7.6 J** basados en la intensidad de la imagen Mappix de la **Figura 7.6 G**, permiten detectar los cuatro emisores. Estos resultados demuestran que en el caso de Mappix la distribución relativa de intensidad en la imagen final aproxima mejor a la de los datos sin procesar. En presencia de una escala de brillo muy extensa, como es el caso de esta simulación, esto podría resultar en una mayor eficiencia para detectar emisores solapados y discriminarlos.

7.9 Resultados obtenidos en imágenes reales

En la **Figura 7.7** aparecen los resultados que se obtienen al analizar células HEK293-T expresando la proteína fluorescente Dronpa, que en este caso está unida a la membrana plasmática. Las muestras se excitaron utilizando una línea láser de 488 nm. Se adquirieron 2000 imágenes en una sucesión rápida, utilizando un tiempo de exposición de 10 ms por imagen. Las **Figuras 7.7 B** y **7.7 C** muestran los resultados para SOFI de segundo y tercer orden respectivamente. Estas imágenes se comparan con las que se obtienen a partir de Mappix en la **Figura 7.7 D**. En SOFI de tercer orden, se puede observar una cierta ganancia en la resolución espacial aunque el desvío en los niveles de brillo resulta en un oscurecimiento de ciertas regiones de la imagen. Para SOFI de 2do orden, la situación es menos crítica. La ganancia en información visual y la mejora en la resolución aparente, son bastante similares para SOFI de segundo orden (**Figura S2B**) y para Mappix (**Figura 7.7 D**). Sin embargo, tal como se resalta en los recuadros 1 a 4 de la **Figura 7.7**, para la imagen Mappix, la distribución de intensidad de la imagen original se encuentra menos distorsionada a pesar del amplio rango dinámico de intensidades. De esta forma, utilizando Mappix se puede obtener una imagen adecuadamente balanceada para el conjunto de la estructura.

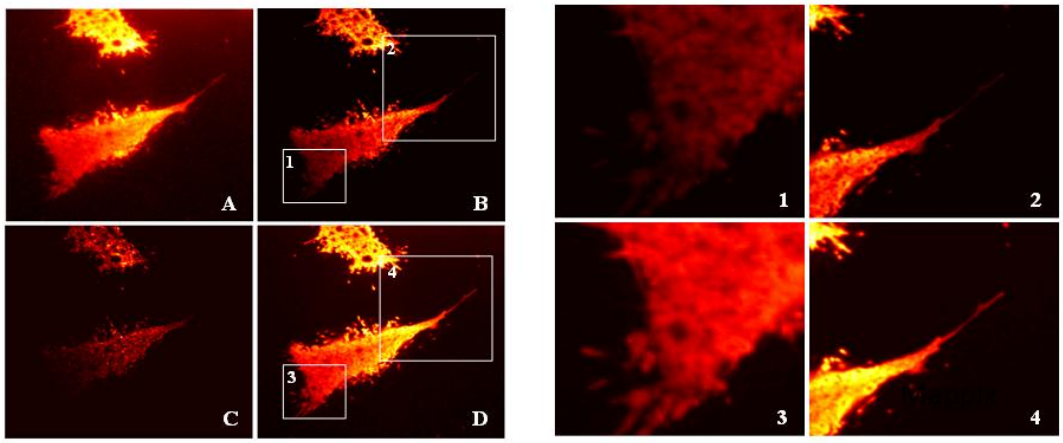


Figura 7.7. Imagen de una célula HEK293T marcada con la proteína Dronpa. (A) Imagen original generada promediando todos los cuadros de la película (2000 cuadros, con tiempos de adquisición de 10 ms. por cuadro). (B) y (C) Imágenes de SOFI de 2do. y 3er. orden respectivamente. (D) Imagen Mappix. (1-4) Vista ampliada de los recuadros blancos en las imágenes A-D.

7.10 Conclusiones

A través del trabajo presentado en este anexo, se demostró que una metodología fundamentada en el cálculo de la disimilitud entre las señales de distintos pixeles, para la cual se utilizó el nombre de Mappix, constituye una alternativa poderosa a la hora de procesar datos de microscopía de fluorescencia con una gran densidad de etiquetas fluorescentes.

El modelo propuesto es muy simple desde el punto de vista matemático y poco exigente en lo que se refiere al costo computacional. Además, el método puede aplicarse a un gran número de modalidades de obtención de imágenes y para datos con distintas características. Sin embargo, en condiciones normales de fluctuación de la señal fluorescente y de densidad de emisores, sólo produce una mejora moderada que no es superior a la que pueden alcanzar otras metodologías que se vienen utilizando hasta la actualidad.

La principal fortaleza de Mappix es su robustez frente a la densidad de fluoróforos, las características de las fluctuaciones y las variaciones de brillo en la imagen original. En estas situaciones, a diferencia de otras técnicas, Mappix sigue posibilitando una mejora en la información visual, resaltando el contenido estructural de la imagen en situaciones en que otras técnicas se ven limitadas severamente. En consecuencia, esta metodología

debería consolidarse como una alternativa de utilidad en algunas aplicaciones de obtención de imágenes biológicas, en las que se requiere resaltar señales para realizar un análisis estructural al tiempo que es necesario preservar el contraste original para un análisis funcional o cuantitativo.

Como nota final, es recomendable utilizar Mappix como una técnica de potenciamiento de imágenes de microscopía de fluorescencia con objetivos de inspección visual y semicuantitativa, combinado con SOFI como técnica de análisis a nivel subdifracción.

Como resultado de las investigaciones presentadas en este anexo se publicó el siguiente trabajo:

- Ruckebusch C.; Bernex R.; Allegrini F.; Sliwa M.; Hofkens J.; Dedecker P. (2015) "Mapping pixel dissimilarity in wide field super-resolution fluorescence microscopy". *Analytical Chemistry*, 87 (9), 4675-4682.

También dio lugar a los siguientes trabajos publicados en reuniones científicas:

- Bernex, R.; Sliwa, M.; Allegrini, F.; Ruckebusch C.; and P. Dedecker. *Signal dissimilarity for functionalsuperresolution. Can it be useful?* 20th International Workshop on "Single Molecule Spectroscopy and Ultrasensitive Analysis in the Life Sciences". Berlín, Alemania. Modalidad de presentación: póster. Septiembre de 2014.

- Ruckebusch C.; Bernex, R.; Sliwa, M.; Allegrini, F.; de Rooi, J.J.; Eilers, P. H. C. *"Strategies for single-molecule fluorescence imaging data analysis"* Chemometrics in Analytical Chemistry. Richmond, Virginia, USA. Modalidad de presentación: oral y poster. Junio de 2014.

- Allegrini F.; Sliwa M.; Ruckebusch C. *"Fluorescence imaging single molecule localization based on dissimilarity maps"*. 13th Scandinavian Symposium on Chemometrics. Estocolmo, Suecia. Modalidad de presentación: póster. Junio de 2013.

BIBLIOGRAFÍA

1 MATLAB 7.10; *The MathWorks Inc*.: Natick, MA, 2010.

2 Booksh, K. S.; Kowalski, B. R.; Theory of analytical chemistry. *Analytical Chemistry,66* (1994), 782A–791A.

3 ISO 5725:1996. (1996) *Accuracy (trueness and precision) of measurement methods and results*; International Organization of Standardization: Geneva, Switzerland.

4 Document No. SANCO/12495/2011.(2012). *Method validation and quality control procedures for pesticide residues analysis in food and feed*, Directorate General for Health and Consumer Affairs (SANCO), European Commission.

5 Danzer, K.; Currie, L. A. (1998). Guidelines for calibration in analytical chemistry. Part 1. Fundamentals and single component calibration. *Pure and Applied Chem*istry, 70, 993–1014.

6 Olivieri, A. C; N. M. Faber; Validation and error, In: S. Brown, R. Tauler, B. Walczak (Eds.). *Comprehensive Chemometrics*, Elsevier, Amsterdam, 2009, Vol. 3, p 91.

7 Sanchez E.; Kowalski B. R. (1988). Tensorial Calibration: I. First-order calibration. *Journal of Chemometrics*. 2, 247–263.

8 Ho, C. N.; Christian, G. D.; DavidsonE. R. (1980). Application of the method of rank annihilation to fluorescent multicomponent mixtures of polynuclear aromatic hydrocarbons. *Analytical Chemistry*, 52, 1071–1079.

9 Messick N. J.; Kalivas J. H.; .Lang P. M. (1996). Selectivity and related measures for *n*th-order data. *Analytical Chemistry*, 68, 1572–1579.

10 Olivieri, A. C.; Faber, N. M. (2004). Standard error of prediction in parallel factor (PARAFAC) analysis of three-way data. *Chemometrics and Intelligent Laboratory Systems*, 70, 75–82.

11 Olivieri, A. C.; Faber N. M. (2005). A closed-form expression for computing the sensitivity in second-order bilinear calibration. *Journal of Chemometrics*, 19, 583–592.

12 Faber, K.; Lorber, A.; Kowalski, B. R. (1997). Analytical figures of merit for tensorial calibration. *Journal of Chemometrics*, 11,419–461.

13 Olivieri A. C. (2005). Computing sensitivity and selectivity in parallel factor analysis and related multi-way techniques: the need for further developments in net analyte signal theory. *Analytical Chemistry*, 77, 4936–4946.

14 A. C. Olivieri. (2008). Analytical advantages of multivariate data processing.One, two, three, infinity? *Analytical Chemistry*, 80, 5713–5720.

15 Olivieri A. C.; Faber K. (2012). New developments for the sensitivity estimation in four-way calibration with the quadrilinear parallel factor model. *Analytical Chemistry*, 84, 186–193.

16 Bauza M. C.; Ibañez G. A.; Tauler R.; Olivieri A. C. (2012). Sensitivity equation for quantitative analysis with multivariate curve resolution-alternating least-squares: theoretical and experimental approach. *Analytical Chemistry*, 84, 8697–8706.

17 Olivieri A. C. (2014). Analytical figures of merit: from univariate to multiway calibration. *Chemical Reviews*, 114, 5358–5378.

18 Belter M.; Sajnóg A.; Barałkiewicz D. (2014). Over a century of detection and quantification capabilities in analytical chemistry–Historical overview and trends. *Talanta*, 129, 606–616.

19 International Organization for Standardization (ISO).(1997). *Capability of detection; Report No. ISO 11843-1*; ISO: Genève, Switzerland.

20 International Organization for Standardization (ISO). (2000). *Capability of detection; Report No. ISO 11843-2*; ISO: Genève, Switzerland.

21 McNaught A. D.; Wilkinson A. (1997). IUPAC. *Compendium of chemical terminology, 2nd ed.*; Blackwell Scientific Publications: Oxford.

22 Inczédy, J.; Lengyel, T.; Ure, A. M.; Gelencsér, A.; Hulanicki, A. (1998). IUPAC Analytical Chemistry Division, *Compendium of Analytical Nomenclature, 3rd. ed.*, Blackwell, Oxford.

23 Wold, S.; Sjöström, M.; Eriksson, L. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems,* 58(2), 109-130.

24 Haaland, D. M.; Thomas, E. V. (1988). Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical Chemistry,* 60(11), 1193-1202.

25 Martens, H.; Næs, T. (1989). Multivariate Calibration; *John Wiley and Sons*, Chichester, U.K.

26 Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; De Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. (1997). *Handbook of Chemometrics and Qualimetrics*; Elsevier: Amsterdam.

27 Ortiz M. C.; Sarabia L. A.; Herrero A.; Sánchez M. S.; Sanz M. B, Rueda M. E.; Giménez D., Meléndez M. E. (2003). Capability of detection of an analytical method evaluating false positive and false negative (ISO 11843) with partial least squares. *Chemometrics and Intelligent Laboratory Systems*, 69, 21–33.

28 Burns D. A.; Ciurczak E. W. (2008). *Handbook of Near-Infrared Analysis, 3rd. ed.*, Practical Spectroscopy Series; CRC Press: Boca Raton, Vol. 35, FL.

29 Leger M. N.; Vega-Montoto L.; Wentzell, P. D. (2005), Methods for systematic investigation of measurement error covariance matrices. *Chemometrics and Intelligent Laboratory Systems,*77, 181–205.

30 Wentzell, P. D. (2014). Measurement errors in multivariate chemical data. *Journal of Brazilian Chemical Society*, 25, 183–196.

31 Wentzell, P. D.; Tarasuk, A. C. (2014). Characterization of heteroscedastic measurement noise in the absence of replicates. *Analytica Chimica Acta*, 847 16–28.

32 Wentzell, P. D.; Andrews, D. T.; Hamilton, D. C.; Faber, K.; Kowalski B. R. (1997). Maximum likelihood principal component analysis. *Journal of Chemometrics*, 11 339–366.

33 Wentzell, P. D., Andrews, D. T., Kowalski, B. R. (1997). Maximum likelihood multivariate calibration. *Analytical Chemistry*, 69, 2299–2311.

34 Draper, N. R.; Smith, H. (1998*). Applied Regression Analysis. 3rd ed*. Wiley-Interscience, New York.

35 Krzanowski, W. J. (2000). *Principles of multivariate analysis: a user´s perspective*. Oxford University Press, New York.

36 Andersen, C.M.; Bro, R. (2010). Variable selection in regression-a tutorial. *Journal of Chemometrics,* 24, 728–737.

37 Brown, C. D.; Green, R. L. (2009). Critical factors limiting the interpretation of regression vectors in multivariate calibration. *TrAC - Trends in Analytical Chemistry,* 28(4), 506-514.

38 Wold S, Johansson E, Cocchi M. PLS: partial least squares projections to latent structures. In: Kubinyi H, editor. (1993). *3D QSAR in drug design: theory, methods and applications*. Leiden, The Netherlands: ESCOM Science Publishers; 523-50.

39 Rajalahti, T., Arneberg, R., Berven, F.S., Myhr, K.-M., Ulvik, R.J., & Kvalheim, O.M. (2009). Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. *Chemometrics and. Intelligent. Laboratory Sysems,* 95, 35–48.

40 Norgard L.; Saudland A.; Wagner J.; Nielsen JP.; Munck L.; Engelsen SB.(2000). Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. *Applied Spectroscopy,* 54, 413-419.

41 Leardi, R.; Seasholtz, M. B.; Pell, R. J. (2002) Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data. *Analytica Chimica Acta,* 461 (2), 189-200.

42 Leardi, R.; Lupiáñez González, A. (1998).Genetic algorithms applied to feature selection in PLS regression: How and when to use them. *Chemometrics and Intelligent Laboratory Systems,* 41 (2), 195-207.

43 Leardi, R. (2003). Genetic algorithm-PLS as a tool for wavelength selection in spectral data sets. *Data Handling in Science and Technology,* 23, 169–196.

44 Dorigo, M. (1992). *Optimization, Learning and Natural Algorithms*; PhD Thesis, Politecnico di Milano, Milan, Italy.

45 Dorigo, M.; Stutzle T. (2004). *Ant colony optimization*; The MIT Press, Cambridge, MA, USA.

46 Allegrini, F.; Olivieri, A. C. (2011), "A new and efficient variable selection algorithm based on ant colony optimization. Applications to near infrared spectroscopy/ partial least-squares analysis". *Analytica Chimica Acta*, 699, 18-25.

47 M. Linder, R. Sundberg, Precision of prediction in second order calibration, with focus on bilinear regression methods, *Journal of Chemometrics*, 16 (2002) 12–27.

48 Olivieri, A. C. (2005). A combined artificial neural network/residual bilinearization approach for obtaining the second-order advantage from three-way non-linear data. *Journal of Chemometrics*, 19, 615-624.

49 Arancibia, J. A.; Olivieri, A. C.; Gil, D. B.; Mansilla, A. E.; Durán-Merás, I.; De La Peña, A. M. (2006). Trilinear least-squares and unfolded-PLS coupled to residual trilinearization: New chemometric tools for the analysis of four-way instrumental data. *Chemometrics and Intelligent Laboratory Systems*, 80(1), 77-86.

50 Maggio, R. M.; Muñoz De La Peña, A.; Olivieri, A. C. (2011). Unfolded partial least-squares with residual quadrilinearization: A new multivariate algorithm for processing five-way data achieving the second-order advantage. Application to fourth-order excitation-emission-kinetic-pH fluorescence analytical data. *Chemometrics and Intelligent Laboratory Systems*, 109(2), 178-185.

51 Bro, R. (1997). PARAFAC. Tutorial and applications. *Chemometrics and Intelligent Laboratory Systems*, 38(2), 149-171.

52 Tauler, R. (1995). Multivariate curve resolution applied to second order data. *Chemometrics and Intelligent Laboratory Systems*, 30(1), 133-146.

53 Brown, C. D. (2000). *Rational Approaches to Data Preprocessing in Multivariate Calibration.*; PhD Thesis, Dalhousie University, Halifax, Canada.

54 Kawakami Harrop Galvão, R.; Fernanda Pimentel, M.; Cesar Ugulino Araujo, M.; Yoneyama, T.; Visani, V. (2001). Aspects of the successive projections algorithm for variable selection in multivariate calibration applied to plasma emission spectrometry. *Analytica Chimica Acta*, 443(1), 107-115.

55 Araújo, M. C. U.; Saldanha, T. C. B.; Galvão, R. K. H.; Yoneyama, T.; Chame, H. C.; Visani, V. (2001). The successive projections algorithm for variable selection in spectroscopic multicomponent analysis. *Chemometrics and Intelligent Laboratory Systems*, 57(2), 65-73.

56 Jolliffe, I.T. (2002) *Principal component analysis, 2nd ed.*, Sprnger-Verlag, New York, USA.

57 Jackson J. E. (2003). *A user´s guide to principal components*. John Wiley and Sons, Hoboken, USA.

58 Wu, W.; Massart, D. L.; De Jong, S. (1997). The kernel PCA algorithms for wide data. Part I: Theory and algorithms. *Chemometrics and Intelligent Laboratory Systems*, 36(2), 165-172.

59 Golub G. H.; Reinsch, C. (1970). Singular value descomposition and least squares solutions. *Numer Math*, 14, 403-420.

60 Wold H. (1975) *Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. In: Gani J, editor. Perspectives in probability and statistics*. Applied probability trust, Sheffield, England.

61 Andersson, M. (2009). A comparison of nine PLS1 algorithms. *Journal of Chemometrics*, 23(10), 518-529.

62 Escandar, G. M.; Olivieri, A.C. (2014). *Practical three-way calibration, 1st. ed.*, Elsevier, Waltham, USA.

63 Tauler, R.; Maeder, M.; de Juan, A. Multiset data analysis: extended multivariate curve resolution, in: Brown, R. Tauler, B. Walczak (Eds.), *Comprehensive Chemometrics*; Elsevier, Amsterdam, 2009, Vol. 2, p 473.

64 Wold, S.; Geladi, P.; Esbensen, K.; and Øhman, J. (1987). Multiway principal components and PLS analysis. *Journal of Chemometrics*, 1, 41-56.

65 Bro, R. (1996). Multiway calibration. Multilinear PLS. *Journal of Chemometrics*, 10(1), 47-61.

66 Harshman, R. A. (1970). Foundations of the PARAFAC procedure, model and conditions for an "explanatory" multi-mode factor analysis. *UCLA Working Papers Phonetics*, 16,1.

67 Carrol, J. D.; Chang, J. (1970). Analysis of individual differences in multidimensional scaling via an N-way gemeralization of "Eckart-Young" decomposition. *Psycometrika*, 35,283.

68 Cattel, R. B. (1944). "Parallel proportional profiles" and other principles for determinating the choice of factors by rotation. *Psycometrika*, 9, 267.

69 Windig, W.; Guilment, J. (1991). Interactive self-modeling mixture analysis. *Analytical Chemistry*, 63, 1425-1432.

70 Öhman, J.; Geladi, P.; and Wold, S. (1990). Residual bilinearization. Part I. Theory and algorithms. *Journal of Chemometrics*, 4, 79-90.

71 Bortolato, S. A.; Arancibia, J. A.; Escandar, G. M.; and Olivieri, A. C. (2007) Improvement of residual bilinearization by particle swarm optimization for achieving the second-order advantage with unfolded partial least-squares. *Journal of Chemometrics*, 21, 557-566.

72 Bortolato, S. A.; Arancibia, J. A.; and Escandar, G. M. (2008) Chemometrics-Assisted Excitation-Emission Fluorescence Spectroscopy on Nylon Membranes. Simultaneous Determination of Benzo a pyrene and Dibenz a,h anthracene at Parts-Per-Trillion Levels in the Presence of the Remaining EPA PAH Priority Pollutants As Interferences. *Analytical Chemistry*, 80, 8276-8286.

73 Galvão, R. K. H.; Araujo, M. C. U. (2009). In: Brown S. D.; Tauler R.; Walczak (Eds.). *Comprehensive Chemometrics*, vol. 3, Elsevier, Amsterdam p. 2333.

74 Siesler, H. W.; Ozaki, Y.; Kawata, S.; Heise (Eds.). (2002). *Near Infrarred Spectroscopy: Principles, Instruments, Applications*. Wiley-VCH, Weinheim, Germany.

75 Broadhurst, D.; Goodacre, R.; Jones, A.; Rowland, J. J.; Kell, D. B. (1997). Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with applications to pyrolysis mass spectrometry. *Analytica Chimica Acta*, 348 (1-3), 71-86.

76 Gauchi, J. P.; Chagnon, P., (2001). Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. *Chemometrics and Intelligent Laboratory Systems*, 58 (2), 171-193.

77 Mehmood, T.; Liland, K. H.; Snipen, L.; Sæbø, S. (2012). A review of variable selection methods in Partial Least Squares Regression. *Chemometrics and Intelligent Laboratory Systems*, 118, 62-69.

78 Geladi, P.; MacDougall, D.; Martens, H., (1985). Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Applied Spectroscopy*, 39 (3), 491-500.

79 Barnes, R. J.; Dhanoa, M. S.; Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Applied Spectroscopy*, 43 (5), 772-777.

80 Lorber, A.; Kowalski, B. R. (1988). The effect of interferences and calibration design on accuracy: Implications for sensor and sample selection. *Journal of Chemometrics*, 2 (1), 67-79.

81 Devos, O.; Duponchel, L. (2011). Parallel genetic algorithm co-optimization of spectral pre-processing and wavelength selection for PLS regression. *Chemometrics and Intelligent Laboratory Systems*, 107 (1), 50-58.

82 Centner, V.; Massart, D. L.; De Noord, O. E.; De Jong, S.; Vandeginste, B. M.; Sterna, C. (1996). Elimination of Uninformative Variables for Multivariate Calibration. *Analytical Chemistry*, 68(21), 3851-3858.

83 Frank, I. E. (1987). Intermediate least squares regression method. *Chemometrics and Intelligent Laboratory Systems*, 1(3), 233-242.

84 Fernández Pierna, J. A.; Abbas, O.; Baeten, V.; Dardenne, P. (2009). A Backward Variable Selection method for PLS regression (BVSPLS). *Analytica Chimica Acta*, 642(1-2), 89-93.

85 Dorigo, M.; Stutzle T. (2004). *Ant colony optimization*; The MIT Press, Cambridge, MA, USA.

86 Ferré, J.; Rius, F. X. (1996). Selection of the Best Calibration Sample Subset for Multivariate Regression. *Analytical Chemistry*, 68 (9), 1565-1571.

87 Dantas Filho, H. A.; Harrop Galvão, R. K.; Ugulino Araújo, M. C.; Da Silva, E. C.; Bezerra Saldanha, T. C.; José, G. E.; Pasquini, C.; Raimundo Jr, I. M.; Rodrigues Rohwedder, J. J. (2004). A strategy for selecting calibration samples for multivariate modelling. *Chemometrics and Intelligent Laboratory Systems,* 72 (1), 83-91.

88 Kennard, R. W.; Stone, L. A. (1969). Computer Aided Design of Experiments. *Technometrics,* 11(1), 137-148.

89 Galvão, R. K. H.; Araujo, M. C. U.; José, G. E.; Pontes, M. J. C.; Silva, E. C.; Saldanha, T. C. B. (2005) A method for calibration and validation subset partitioning. *Talanta,* 67 (4), 736-740.

90 MATLAB.The Mathworks Inc., Natick, Massachusetts, USA.

91 Savitzky, A.; Golay, M. J. E. (1964). Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry,* 36 (8), 1627-1639.

92 Olivieri, A. C. (2014). Analytical figures of merit: From univariate to multiway calibration. *Chemical Reviews,* 114(10), 5358-5378.

93 Currie, L.A. (1999). Detection and quantification limits: origins and historical overview. *Analytica Chimica Acta,* 391, 127-134.

94 Ferrús, R.; Egea, M.R. (1994). Limit of discrimination, limit of detection and sensitivity in analytical systems. *Analytica Chimica Acta,* 287, 119-145.

95 Boqué R.; Faber N.M.; Rius F.X. (2000). Detection limits in classical multivariate calibration models, *Analytica Chimica Acta,* 423, 41-49.

96 Singh, A. (1993). Multivariate decision and detection limits. *Analytica Chimica Acta,* 277, 205-214.

97 Ferré, J.; Boqué, R.; Fernandez Band, B.; Larrechi, M.S.; Rius, F.X. (1997). Figures of merit in multivariate calibration. Determination of four pesticides in water by flow injection analysis and spectrophotometric detection. *Analytica Chimica Acta,* 348, 167-175.

98 Sanz, M.B.; Sarabia, L.A.; Herrero, A.; Ortiz, M.C. (2001). Capability of discrimination: application to soft calibration methods. *Analytica Chimica Acta,* 446 297-311.

99 Faber, K.; Kowalski, B.R. (1997). Propagation of measurement errors for the validation of predictions obtained by principal component regression and partial least squares. *Journal of Chemometrics*, 11, 181-238.

100 Faber, N. M.; Bro, R. (2002). Standard error of prediction for multiway PLS: 1. Background and a simulation study. *Chemometrics and Intelligent Laboratory Systems,* 61(1–2), 133-149.

101 Schreyer, S.K.; Bidinosti, M.; Wentzell, P.D. (2002). Application of maximum likelihood principal components regression to fluorescence emission spectra. *Applied Spectroscopy,* 56, 789-796.

102 Loock, H.-P.; Wentzell, P. D. (2012). Detection limits of chemical sensors: Applications and misapplications. *Sensors and Actuators B: Chemical,* 173(0), 157-163.

103 Boqué, R.; Rius, F. X. (1996). Multivariate detection limits estimators. *Chemometrics and Intelligent Laboratory Systems,* 32(1), 11-23.

104 MacDougall, D.; Crummett, W. B. (1980). Guidelines for data acquisition and data quality evaluation in environmental chemistry. *Analytical Chemistry*, 52(14), 2242-2249.

105 Ostra, M.; Ubide, C.; Vidal, M.; Zuriarrain, J. (2008). Detection limit estimator for multivariate calibration by an extension of the IUPAC recommendations for univariate methods. *Analyst,* 133(4), 532-539.

106 Boqué, R.; Larrechi, M. S.; Rius, F. X. (1999). Multivariate detection limits with fixed probabilities of error. *Chemometrics and Intelligent Laboratory Systems,* 45(1–2), 397-408.

107 Blanco, M.; Castillo, M.; Peinado, A.; Beneyto, R. (2007). Determination of low analyte concentrations by near-infrared spectroscopy: Effect of spectral pretreatments and estimation of multivariate detection limits. *Analytica Chimica Acta,* 581(2), 318-323.

108 Wu, Z.; Sui, C.; Xu, B.; Ai, L.; Ma, Q.; Shi, X.; Qiao, Y. (2013). Multivariate detection limits of on-line NIR model for extraction process of chlorogenic acid from Lonicera japonica. *Journal of Pharmaceutical and Biomedical Analysis,* 77(0), 16-20.

109 Ortiz, M. C.; Sarabia, L. A.; Sánchez, M. S. (2010). Tutorial on evaluation of type I and type II errors in chemical analyses: From the analytical detection to authentication of products and process control. *Analytica Chimica Acta,* 674(2), 123-142.

110 Olivieri, A. C.; Faber, N. M.; Ferré, J.; Boqué, R.; Kalivas, J. H.; Mark, H. (2006). Uncertainty estimation and figures of merit for multivariate calibration: (IUPAC technical report). *Pure and Applied Chemistry,* 78(3), 633-661.

111 Clayton, C. A.; Hines, J. W.; Elkins, P. D. (1987). Detection limits with specified assurance probabilities. *Analytical Chemistry, 59*(20), 2506-2514.

112 Del Río Bocio; F. J., Riu, J.; Boqué, R.; Rius, F. X. (2003). Limits of detection in linear regression with errors in the concentration.*Journal of Chemometrics, 17*(7), 413-421.

113 Voigtman; E. (2008). Limits of detection and decision.Part 2. *Spectrochimica Acta - Part B Atomic Spectroscopy, 63*(2), 129-141.

114 Voigtman; E. (2008). Limits of detection and decision.Part 1.*SpectrochimicaActa - Part B Atomic Spectroscopy, 63*(2), 115-128.

115 Arancibia, J. A.; Rullo, A.; Olivieri, A. C.; Di Nezio; S., Pistonesi; M., Lista, A.; Band, B. S. F. (2004). Fast spectrophotometric determination of fluoride in ground waters by flow injection using partial least-squares calibration. *Analytica Chimica Acta, 512*(1), 157-163.

116 Arancibia, J. A.; Delfa, G. M.; Boschetti, C. E.; Escandar, G. M.; Olivieri, A. C. (2005). Application of partial least-squares spectrophotometric-multivariate calibration to the determination of 2-sec-butyl-4,6-dinitrophenol (dinoseb) and 2,6-dinitro-p-cresol in industrial and water samples containing hydrocarbons. *Analytica Chimica Acta, 553*(1-2), 141-147.

117 Goicoechea, H. C.; Olivieri, A. C. (1999). Determination of bromhexine in cough-cold syrups by absorption spectrophotometry and multivariate calibration using partial least-squares and hybrid linear analyses. Application of a novel method of wavelength selection. *Talanta,* 49(4), 793-800.

118 Goicoechea, H. C.; Olivieri, A. C. (1999). Enhanced synchronous spectrofluorometric determination of tetracycline in blood serum by chemometric analysis. Comparison of partial least-squares and hybrid linear analysis calibrations. *Analytical Chemistry,* 71(19), 4361-4368.

119 Pedrido, M. L.; Bortolato, S.; González Sierra, M.; Olivieri, A.C.; Boschetti, C. E. (2008) *Lab Ciencia*, *3*(39), 14−18.

120 Escandar, G. M.; Olivieri, A. C.; Faber, N. M.; Goicoechea, H. C.; Muñoz de la Peña A.; Poppi, R. J. (2007). Second- and third-order multivariate calibration: data, algorithms and applications. *TrAC Trends in Analytical Chemistry*, 26(7), 752-765.

121 Olivieri, A. C. (2008). Analytical advantages of multivariate data processing. One, two, three, infinity? *Analytical Chemistry,* 80(15), 5713-5720.

122 Olivieri, A. C.; Escandar, G. M.; Peña, A. M. d. l. (2011). Second-order and higher-order multivariate calibration methods applied to non-multilinear data using different algorithms. *TrAC Trends in Analytical Chemistry,* 30(4), 607-617.

123 Bro, R. (2006). Review on Multiway Analysis in Chemistry—2000–2005. *Critical Reviews in Analytical Chemistry,* 36(3-4), 279-293.

124 Ni, Y.; Gu, Y.; Kokot, S. (2012). Interpreting Analytical Chemistry Data: Recent Advances in Curve Resolution with the Aid of Chemometrics. *Analytical Letters,* 45(8), 933-948.

125 Arancibia, J. A.; Damiani, P. C.; Escandar, G. M.; Ibañez, G. A.; Olivieri, A. C. (2012). A review on second- and third-order multivariate calibration applied to chromatographic data. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences,* 910, 22-30.

126 Olivieri, A. C. (2012). Recent advances in analytical calibration with multi-way data. *Analytical Methods,* 4(7), 1876-1886.

127 Öhman, J.; Geladi, P.; Wold, S. (1990). Residual bilinearization. Part I. Theory and algorithms.*J. Chemometrics*, 4, 79-90.

128 Olivieri, A. C. (2005). On a versatile second-order multivariate calibration method based on partial least-squares and residual bilinearization: Second-order advantage and precision properties. *Journal of Chemometrics,* 19(4), 253-265.

129 Denham, M. C. (1997). Prediction intervals in partial least squares. *Journal of Chemometrics,* 11(1), 39-52.

130 Faber, N. M. (2000). Comparison of two recently proposed expressions for partial least squares regression prediction error. *Chemometrics and Intelligent Laboratory Systems,* 52(2), 123-134.

131 Serneels, S.; Lemberge, P.; Van Espen, P. J. (2004). Calculation of PLS prediction intervals using efficient recursive relations for the Jacobian matrix. *Journal of Chemometrics,* 18(2), 76-80.

132 Fernández Pierna, J. A.; Jin, L.; Wahl, F.; Faber, N. M.; Massart, D. L. (2003). Estimation of partial least squares regression prediction uncertainty when the reference values carry a sizeable measurement error. *Chemometrics and Intelligent Laboratory Systems,* 65(2), 281-291.

133 Faber, N. M.; Song, X. H.; Hopke, P. K. (2003). Sample-specific standard error of prediction for partial least squares regression. *TrAC - Trends in Analytical Chemistry,* 22(5), 330-334.

134 Phatak, A.; Reilly, P. M.; Penlidis, A. (1993). An approach to interval estimation in partial least squares regression. *Analytica Chimica Acta, 277*(2), 495-501.

135 Olivieri, A. C.; Faber, N. M. (2009). *Validation and Error, in Comprehensive Chemometrics* (Vol. 3). Amsterdam.

136 Lorber, A. (1986). Error propagation and figures of merit for quantification by solving matrix equations. *Analytical Chemistry,* 58(6), 1167-1172.

137 Jaumot, J.; Gargallo, R.; and Tauler, R. (2004). Noise propagation and error estimations in multivariate curve resolution alternating least squares using resampling methods, *Journal of Chemometrics,* 18, 327-340.

138 Saurina, J.; Leal, C.; Compano, R.; Granados, M.; Prat, M. D.; and Tauler, R. (2001) . Estimation of figures of merit using univariate statistics for quantitative second-order multivariate curve resolution, *Analytica Chimica Acta,* 432, 241-251.

139 Bernstein, D. S. *Matrix Mathematics: Theory, Facts, and Formulas*, 2nd Ed., Princeton University Press, Princeton, NJ, 2009.

140 Klein, T.; Proppert, S.; Sauer, M. (2014). Eight years of single-molecule localization microscopy. *Histochemistry and Cell Biology,* 141(6), 561-575.

141 Sengupta, P.; Van Engelenburg, S. B.; Lippincott-Schwartz, J. (2014). Superresolution imaging of biological systems using photoactivated localization microscopy. *Chemical Reviews,* 114(6), 3189-3202.

142 Hell, S. W., & Wichmann, J. (1994). Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy. *Optics Letters,* 19(11), 780-782.

143 Hell, S. W. (2003). Toward fluorescence nanoscopy. *Nature Biotechnology*, *21*(11), 1347-1355.

144 Dedecker, P., Hotta, J. I., Flors, C., Sliwa, M., Uji-i, H., Roeffaers, M. B. J., Ando, R., Mizuno, H., Miyawaki, A., & Hofkens, J. (2007). Subdiffraction imaging through the selective donut-mode depletion of thermally stable photoswitchable fluorophores: Numerical analysis and application to the fluorescent protein dronpa. *Journal of the American Chemical Society,* 129(51), 16132-16141.

145 Hofmann, M., Eggeling, C., Jakobs, S., Hell, S. W. (2005). Breaking the diffraction barrier in fluorescence microscopy at low light intensities by using reversibly photoswitchable proteins. *Proceedings of the National Academy of Sciences of the United States of America,* 102(49), 17565-17569.

146 Rego, E. H.; Shao, L.; Macklin, J. J.; Winoto, L.; Johansson, G. A.; Kamps-Hughes, N.; Davidson, M. W.; Gustafsson, M. G. L. (2012). Nonlinear structured-illumination microscopy with a photoswitchable protein reveals cellular structures at 50-nm resolution. *Proceedings of the National Academy of Sciences of the United States of America*, 109(3), E135-E143.

147 Muller, C. B.; Enderlein, J. (2010). Image Scanning Microscopy. *Physical Review Letters,* 104(19).

148 Sibarita, J. B. (2014). High-density single-particle tracking: Quantifying molecule organization and dynamics at the nanoscale. *Histochemistry and Cell Biology,* 141(6), 587-595.

149 Rust, M. J.; Bates, M.; Zhuang, X. W. (2006). Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM). *Nature Methods, 3*(10), 793-795.

150 Betzig, E.; Patterson, G. H.; Sougrat, R.; Lindwasser, O. W.; Olenych, S.; Bonifacino, J. S.; Davidson, M. W.; Lippincott-Schwartz, J.; Hess, H. F. (2006). Imaging intracellular fluorescent proteins at nanometer resolution. *Science, 313*(5793), 1642-1645.

151 Huang, F.; Schwartz, S. L.; Byars, J. M.; Lidke, K. A. (2011). Simultaneous multiple-emitter fitting for single molecule super-resolution imaging. *Biomedical Optics Express,* 2(5), 1377-1393.

152 Simonson, P. D.; Rothenberg, E.; Selvin, P. R. (2011). Single-Molecule-Based Super-Resolution Images in the Presence of Multiple Fluorophores. *Nano Letters*, 11(11), 5090-5096.

153 Burnette, D. T.; Sengupta, P.; Dai, Y. H.; Lippincott-Schwartz, J.; Kachar, B. (2011). Bleaching/blinking assisted localization microscopy for superresolution imaging using standard fluorescent molecules. *Proceedings of the National Academy of Sciences of the United States of America,* 108(52), 21081-21086.

154 Cox, S., Rosten, E., Monypenny, J., Jovanovic-Talisman, T., Burnette, D. T., Lippincott-Schwartz, J., Jones, G. E., & Heintzmann, R. (2012). Bayesian localization microscopy reveals nanoscale podosome dynamics. *Nature Methods,* 9(2), 195-200.

155 Holden, S. J.; Uphoff, S.; Kapanidis, A. N. (2011). DAOSTORM: an algorithm for high-density super-resolution microscopy. *Nature Methods,* 8(4), 279-280.

156 Hebert, B.; Costantino, S.; Wiseman, P. W. (2005). Spatiotemporal image correlation spectroscopy (STICS) theory, verification, and application to protein velocity mapping in living CHO cells. *Biophysical Journal,* 88(5), 3601-3614.

157 Dertinger, T.; Colyer, R.; Iyer, G.; Weiss, S.; Enderlein, J. (2009). Fast, background-free, 3D super-resolution optical fluctuation imaging (SOFI). *Proceedings of the National Academy of Sciences of the United States of America,* 106(52), 22287-22292.

158 Mukamel, Eran A.; Babcock, H.; Zhuang, X. (2012). Statistical Deconvolution for Superresolution Fluorescence Microscopy. *Biophysical Journal,* 102(10), 2391-2400.

159 Paclík, P., & Duin, R. P. W. (2003). Dissimilarity-based classification of spectra: Computational issues. *Real-Time Imaging,* 9(4), 237-244.

160 Sanchez, F. C., Toft, J., vandenBogaert, B., & Massart, D. L. (1996). Orthogonal projection approach applied to peak purity assessment. *Analytical Chemistry,* 68(1), 79-85.

161 Dedecker, P.; Duwé, S.; Neely, R. K.; Zhang, J. (2012). Localizer: fast, accurate, open-source, and modular software package for superresolution microscopy. *Journal of Biomedical Optics,* 17(12), 126008-126008.

162 Van de Linde, S.; Wolter, S.; Heilemann, M.; Sauer, M. (2010). The effect of photoswitching kinetics and labeling densities on super-resolution fluorescence imaging. *Journal of Biotechnology,* 149(4), 260-266.

163 Lord, S. J.; Lee, H. L. D.; Moerner, W. E. (2010). Single-molecule spectroscopy and imaging of biomolecules in living cells. *Analytical Chemistry,* 82(6), 2192-2203.