



MÁSTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

PNT_MTEQ002

Rev0

19/10/2009

Pag. 1 de 15

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos.

PRÁCTICA: Casos prácticos de calibración multivariante

CASOS PRÁCTICOS DE CALIBRACIÓN MULTIVARIANTE

ÍNDICE	Pag.
1.- Objetivo	2
2.- Fundamento de la técnica/metodología	2
3.- Instrumentación y protocolo de utilización	3
4.- Material y Reactivos	3
5.- Metodología experimental	3
6.- Registro de datos	3
7.- Obtención de resultados	3
8.- Elaboración del informe	4
9.- Prevención de riesgos	4
10.- Referencias	4
11.- Anexos	5

Elaborado por: Prof. José Ramón Torres Dpto. de Química Analítica	Revisado por: Prof. Salvador Sagrado Dpto. de Química Analítica	VºBº Prof. Amparo Salvador Directora del Máster
Fecha / firma:	Fecha / Firma:	Fecha / Firma:



MÁSTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

PNT_MTEQ002

Rev0

19/10/2009

Pag. 2 de 15

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos.

PRÁCTICA: Casos prácticos de calibración multivariante

1.- Objetivo

Aplicar los conocimientos sobre calibración multivariante a la resolución de situaciones que puedan plantearse en el desarrollo de futuras actividades profesionales o investigadoras, seleccionando las metodologías y herramientas adecuadas, accediendo con criterio a la información necesaria para llevarlas a cabo, planificando en función de los recursos disponibles, tomando las decisiones oportunas, colaborando en equipo eficazmente y emitiendo un informe claro y conciso, argumentando las conclusiones obtenidas.

En particular se desarrollarán los contenidos que figuran en la Guía Docente del Máster, Modulo II, correspondientes a la parte 1 del programa cuyo título es Laboratorio de Calibración y tratamiento de datos y dentro de esta la sesión práctica dedicada a Casos prácticos de calibración multivariante. Dichos contenidos son: Comparación de modelos de regresión multivariante: problemas, limitaciones, ventajas y alcance. Validación de modelos.

2.- Fundamento de la técnica/metodología

En Química Analítica, se entiende por calibración multivariante de métodos establecer la relación entre la concentración de un analito (variable o vector respuesta \mathbf{y}) o varios analitos (análisis multicomponente; matriz \mathbf{Y}) y la señal multivariante (más de 1 variable predictora, matriz \mathbf{X}) de un equipo (en el caso más simple unidimensional, ej. espectro), mediante un modelo de regresión multivariante. En la etapa de calibración (aprendizaje) del modelo \mathbf{y} o \mathbf{Y} son conocidas, vía empleo de patrones o muestras de referencia, y suele incluir la validación del modelo.

Existen diferentes modelos de calibración multivariante, si bien los más empleados son la regresión lineal múltiple (MLR) y la regresión por mínimos cuadrados parciales (PLS, o también PLS1). Mientras que MLR optimiza la correlación entre \mathbf{X} e \mathbf{y} , PLS optimiza la covarianza entre dichas variables. Se trata de modelos predictivos de \mathbf{y} a partir de \mathbf{X} , simplificando $\mathbf{y} = f(\mathbf{X})$. En el caso de MLR las variables \mathbf{X} están explícitas en el modelo, mientras que inicialmente no ocurre esto en PLS, que emplea un escalado previo de los datos de \mathbf{y} y \mathbf{X} y variables transformadas denominadas variables latentes, pudiendo crear modelos con un número inferior (k) de variables latentes a N_V (el número de variables de \mathbf{X}). Existe la posibilidad de reescribir en modelo PLS final en un formato similar al de MLR. El denominado algoritmo PLS2 habilita la posibilidad de llevar a cabo un modelo multicomponente, esto es predictivo de las concentraciones de varios analitos simultáneamente, simplificando $\mathbf{Y} = f(\mathbf{X})$. Ello puede ser interesante si existe correlación entre las variables \mathbf{Y} .

Desde el punto de vista matemático, hay que notar que la calibración es 'inversa' respecto a la que se efectúa en calibración univariante, siendo ahora \mathbf{y} (variable dependiente, la concentración) y \mathbf{X} (variable independiente, las señales). La razón es que ahora el empleo de múltiples sensores de señal, hace que el error quede amortiguado (compensado). Desde el punto de vista práctico, en general, hay que considerar los mismos aspectos de calidad comentados en calibración univariante (PNT_MTEQ001). Por otro lado, el diseño de los patrones está condicionado por el número de analitos en la muestra.

Desde un punto de vista quimiométrico, la calibración multivariante implica el empleo de regresión multivariante, bien sobre variables originales o transformadas, en general empleando algoritmos lineales, para llevar a cabo la obtención (calibración) del modelo y posteriormente la utilización del modelo, para realizar predicciones. El modelo más simple (un analito) puede escribirse de forma matricial como: $\mathbf{y} = \mathbf{X} \mathbf{B} + \mathbf{E}$, donde \mathbf{X} es la matriz que contiene la información relacionada con la señal y \mathbf{E} es la matriz de residuales estimada (información de concentración no explicada por el modelo ajustado). Un análisis de los estadísticos asociados al modelo y a los parámetros, permiten evaluar la calidad del modelo.



MÁSTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

PNT_MTEQ002
Rev0
19/10/2009
Pag. 3 de 15

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos.
PRÁCTICA: Casos prácticos de calibración multivariante

3.- Instrumentación y protocolo de utilización

Para el desarrollo de esta sesión de prácticas tan sólo es necesario disponer de ordenadores (aula de informática), provistos de software ajustado a propósito para la introducción de datos, la estimación de parámetros y estadísticos de calibración y el análisis de resultados. Entre otros, se emplearán los programas comerciales STATGRAPHICS®, SPSS® y UNSCRAMBLER®. También se hará uso de algoritmos escritos en MATLAB®.

El protocolo general para abordar los casos prácticos es el siguiente:

- 1.- Objetivos del estudio
- 2.- Organizar y caracterizar datos
- 3.- Planteamiento
- 4.- Exploración de datos (opcional)
- 5.- Estimación de parámetros y estadísticos
- 6.- Validación del modelo
- 7.- Predicciones y caracterización (si ha lugar)
- 8.- Empleo de los resultados de calibración (toma de decisiones, validación de métodos, comparación...)
- 9.- Informe (analítico, técnico, estadístico...; conclusiones, sugerencias)

4.- Material y Reactivos

Para el desarrollo de esta sesión de prácticas no es necesario disponer de ningún tipo de material de laboratorio ni reactivo. Los datos y otros ficheros se transferirán vía intranet o mediante memorias USB. Será necesario contar con una memoria USB por pareja.

5.- Metodología experimental

Para el desarrollo de esta sesión de prácticas no es necesario ningún tipo de metodología experimental. Cualquier información necesaria se obtiene de datos experimentales obtenidos en otras sesiones prácticas o mediante simulación.

6.- Registro de datos

En general los datos multivariantes pueden abarcar un espacio demasiado grande para ser incluidos en un registro, por lo que se suelen conservar en ficheros cuyo formato depende del tipo de programa que lo ha generado. También es importante saber transferir los datos de unos a otros formatos. Cuando el tamaño lo permita el registro de datos se realizará en hojas de cálculo EXCEL®. Los alumnos desarrollarán croquis con los datos y los parámetros que caracterizan los mismos.

7.- Obtención de resultados

Para el desarrollo de esta sesión de prácticas no es necesario ningún tipo de procedimiento general para la obtención de resultados, más allá del seguimiento del protocolo general del apartado 3. Cada caso particular, dependiendo del objetivo que se persiga, supondrá un tipo de resultados distinto.



MÁSTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

PNT_MTEQ002

Rev0

19/10/2009

Pag. 4 de 15

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos.

PRÁCTICA: Casos prácticos de calibración multivariante

8.- Elaboración del informe

Se realizará un informe para algunos de los casos estudiados, que incluirá la exposición y justificación de los aspectos del protocolo del apartado 3 (ver **Anexos IV y V**; casos prácticos resueltos en el documento [PNT_MTEQ001](#)). Será necesario ajustar el título del informe (Sesión, Caso práctico) y el contenido en cada caso. Se empleará la nomenclatura introducida en el apartado 1, así como las variables del **Anexo I** y las que figuran en el documento [PNT_MTEQ001](#). Al informe se adjuntará un anexo incluyendo cuanta información (tablas, gráficos, volcado de pantallas...) se considere oportuna, convenientemente identificada en la columna de comentarios.

9.- Prevención de riesgos

Para el desarrollo de esta sesión de prácticas no es necesario ningún tipo de adopción de prevención de riesgos particular. No se generan residuos químicos.

10.- Referencias

- [1] D.L. Massart, B.G.M. Vandeginste, L.M.C. Buydens, S. De Jong, P.J. Lewi y J. Smeyers-Verbeke. Handbook of Chemometrics and Qualimetrics: Part A y B, Elsevier, Amsterdam, 1997.
- [2] J.C. Miller y J.N. Miller. Estadística y Quimiometría para Química Analítica, Pearson Education S.A. Madrid, 2002.
- [3] C. Molins-Legua, S. Meseguer-Lloret, Y. Moliner-Martínez y P. Campíns-Falcó. TRAC 2006 25, 282-290.
- [4] R. O. Kuehl, Diseño de experimentos. Principios para el diseño de análisis de investigaciones. 2ª ed. Thomson, México, 2001.
- [5] G. Ramis y M. C. García, Quimiometría, Síntesis, Madrid, 2001.
- [6] S. Sagrado, E. Bonet, M.J. Medina y Y. Martín, Manual Práctico de Calidad en los Laboratorios. Enfoque ISO 17025. AENOR 2005.



MÁSTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

PNT_MTEQ002
Rev0
19/10/2009
Pag. 5 de 15

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos.
PRÁCTICA: Casos prácticos de calibración multivariante

11.- Anexos

Anexo I. Algunas variables a emplear en el informe, adicionales a las indicadas en el Anexo II del documento PNT_MTEQ001 (calibración univariante).

X	Matriz predictora (señal; Variables independientes). x correspondería a cada vector de X
y, Y	Vector/Matriz respuesta (ej. concentración; variable(s) dependiente(s))
E	Matriz de de residuales. e correspondería a cada vector de E
Xt	Matriz de señales de objetos test. No se dispone de yt o Yt Se emplea en la etapa de predicción (empleando el modelo ya calibrado y validado)
Xr	Matriz de señales de objetos de referencia. Se dispone de yr o Yr
yr, Yr	Vector/Matriz de concentraciones de objetos de referencia
k	Número de variables latentes empleadas en un modelo PLS
PRESS_k	SC en predicción de Y . $PRESS_k = \sum (\hat{y}_i - y_i)^2$ (para un modelo con <i>k</i> variables latentes).
RMSE_k	'Root mean square error of prediction' = $\left(\sqrt{PRESS_k} \right) / n$. NOTA se pueden calcular: RMSEC_k (\hat{y} estimadas para X comparadas con y o Y). RMSECV_k (\hat{y} estimadas para X en validación cruzada ^a comparadas con Y). RMSEP_k (\hat{y} estimadas para Xr comparadas con Yr).
EV_k	Varianza explicada (por defecto de y o Y) por el modelo. Puede calcularse en calibración (todos los objetos) o en validación ^a (ej. validación cruzada ^a)

Símbolos: El símbolo '^' indica un valor estimado por el modelo.

^a Detalles sobre validación de modelos: (i) Empleando objetos de referencia: Se emplea **Xr** (como si fuera **Xt**), estimando **Ŷr**, que se compara con **Yr**. (ii) Validación cruzada ('Cross-validation') = división de objetos de calibración en 'g' grupos; cada uno de los cuales se emplea como conjunto de validación, por turno, construyendo un modelo con los restantes objetos, repitiendo el proceso 'g' veces). En cada turno se estima la variable respuesta (**Ŷ_{cv}**) de los objetos del conjunto de validación.



MÁSTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

PNT_MTEQ002

Rev0

19/10/2009

Pag. 6 de 15

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos.

PRÁCTICA: Casos prácticos de calibración multivariante

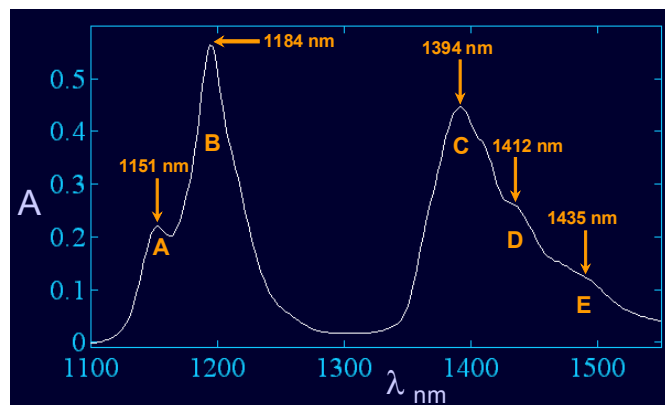
Anexo II. Caso práctico 2a (resuelto)

Estudio de la estructura latente de espectros NIR de gasolinas y predicción del índice de octanos

En este ejemplo realizaremos estudios de reconocimiento de pautas para explorar la estructura (relaciones de asociación) de un conjunto de muestras de gasolina a partir de datos espectrales en el infrarrojo cercano (NIR). Utilizaremos esta información para realizar la predicción de una propiedad asociada al nivel de ramificaciones de la mezcla: el índice de octanos (IO). El IO es una escala que mide el rendimiento de un combustible, expresado como la resistencia a la inflamabilidad por compresión. Durante el funcionamiento de un motor de gasolina, los combustibles con bajo IO tienden a detonar prematuramente durante la fase de compresión (antes de la fase expansiva), lo que se traduce en una importante pérdida de rendimiento. Existen medidas similares para motores diesel (índice de cetanos).

En realidad se han propuesto diversas escalas de índice de octanos. La más utilizada es el IO de investigación o RON (Research Octane Number, propuesto por la American Society for Testing Materials, ASTM), que se mide comparando las detonaciones (ruido) que produce la gasolina respecto del producido por mezclas de isoctano y *n*-heptano, en un motor estándar que funciona a 900 rpm, cuando se inflaman a 149°C. El porcentaje de isoctano en la mezcla que produce el mismo sonido es el RON.

En principio, el IO es un número entre 0 y 100. Sin embargo, algunos combustibles (como el gas licuado de petróleo, etanol y metanol, entre otros) poseen un IO superior a 100. De este modo, es posible incrementar el octanaje añadiéndole algunos compuestos (aditivos). Antiguamente se utilizaba plomo tetraetilo, que si bien mejoraba sus propiedades antidetonantes, era demasiado nocivo. Por esta razón, en la actualidad ha sido reemplazado por etanol, benceno o éteres ramificados (MTBE, ETBE y TAME), en las llamadas "gasolinas ecológicas" (sic). Las gasolinas usuales tienen un IO entre 85 y 100, pero es frecuente encontrar gasolinas con IOs superiores. Un combustible con un IO alto o bajo no es necesariamente malo o bueno, simplemente se ha de usar un motor apropiado para aprovechar adecuadamente sus características.



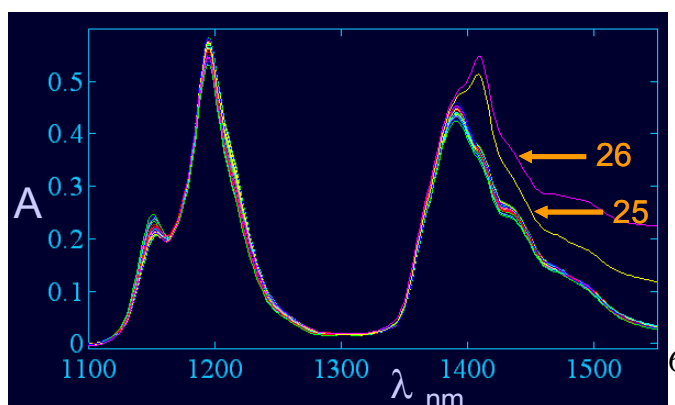
En ausencia de aditivos, el nivel de ramificaciones en las gasolinas favorece el poder antidetonante. A mayor nivel de ramificaciones, mayor es el IO. En la figura puede verse un espectro NIR típico sobre el que se representan las posiciones características de distintas transiciones energéticas moleculares:

- A -Tensión, CH₃ y aromáticos
- B -Tensión, CH₃
- C -Combinación, CH₂
- D -Combinación, CH₂+aromáticos
- E -Combinación, CH₂+aromáticos

A la vista de esta información, comprobamos que los espectros parecen contener información estructural potencialmente correlacionable con el nivel de ramificaciones. Nuestro objetivo va a ser reemplazar el lento y caro método de medida de IO (RON), que requiere calibrados costosos, por medidas instrumentales NIR.

Material: Fichero de datos, que deberá ser importado y manipulado en Excel, y transferido a Unscrambler para su procesamiento. Además de las aulas de Informática de la UV, que disponen del software Unscrambler instalado, cabe la posibilidad de descargar una copia de Unscrambler de activación limitada desde el siguiente enlace: <http://www.camo.com/products/downloads.html>.

Conjunto de entrenamiento ('training set'): 26 muestras de gasolinas sin plomo medidas por NIR a 226 longitudes de onda (la matriz $X_{26 \times 226}$ contiene 26 filas y 226 columnas). Dos de las muestras han sido aditivadas con etanol y presentan un espectro ligeramente alterado (marcadas como 25 y 26 en la figura). Habrá que investigar su estructura latente y decidir si deben excluirse del modelo. El índice de octanos es un vector y de dimensiones 26×1 (El IO ha sido determinado mediante el método ASTM).





MÁSTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

PNT_MTEQ002

Rev0

19/10/2009

Pag. 7 de 15

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos.

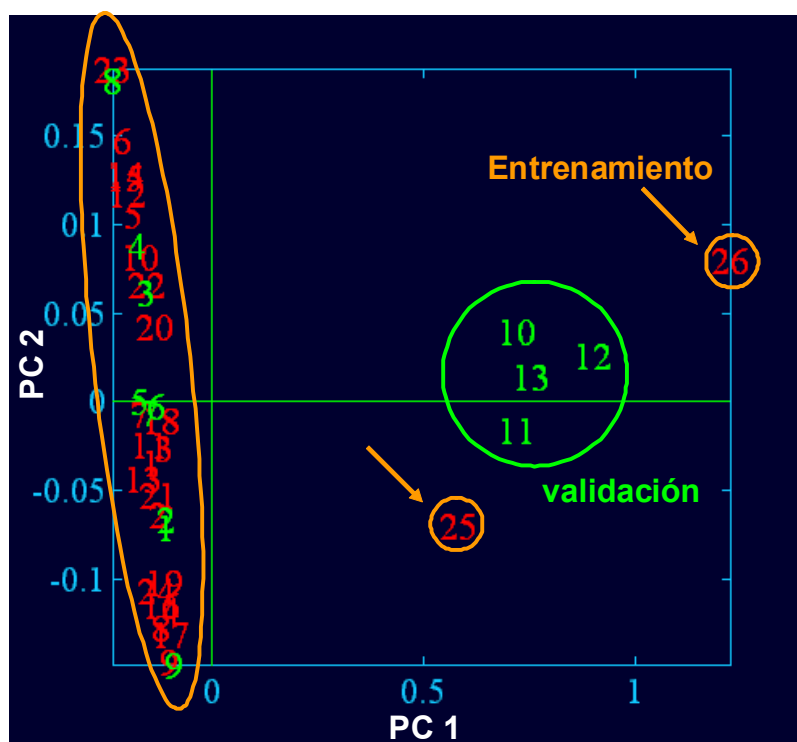
PRÁCTICA: Casos prácticos de calibración multivariante

Tarea: Dibujar los espectros de **X** con Unscrambler

Muestras de referencia o validación ('reference o validation set'): Otras 13 muestras de gasolinas sin plomo. También en este caso, algunas de ellas (4 de las 13) fueron aditivadas con etanol y presentan un espectro de aspecto distinto al resto (**Xr**=13×226). Para estas muestras, el índice de octanos es igualmente conocido: **yr** (13×1).

Tarea: Puesto que ambos conjuntos son equivalentes, forma una matriz única uniéndolos (39 filas y 226 columnas, anexando una matriz bajo la otra). Realiza una rotación propia (vía PCA) con las medidas sólo centradas (no autoescaladas) y proyecta las puntuaciones (scores) sobre el plano PC1 y PC2 para visualizar la estructura. Comenta los resultados encontrados.

Cualquier calibración multivariante debe comenzar examinando la estructura latente de los datos. De existir subestructuras o "clases" (i.e., objetos claramente asociados y diferenciados de los otros), los modelos de calibración deberán particularizarse a estas clases, y las muestras problema referirse al modelo de calibración correspondiente a la clase en la que quedarían asociadas. Invertir esfuerzo en una descripción simultánea de todo el conjunto de entrenamiento cuando existen clases bien diferenciadas empobrecería la calidad de las predicciones y obligaría a incluir un número superior de PCs en el modelo final, con una peor separación de la parte estructural y de la parte irrelevante (ruido). Además, en la etapa de exploración podemos detectar posibles muestras anómalas, si bien sus consecuencias en las predicciones (y eliminación definitiva) se verán posteriormente al construir los modelos. Ésta es una gran ventaja de los modelos de calibración "blandos" (basados en variables latentes): no sólo podemos predecir nuevas muestras, sino que además podemos explorar su estructura y detectar anomalías.



En la figura adjunta se visualizan de inmediato las muestras anómalas o "outliers". Como se observa, las muestras típicas se distribuyen en una gran banda más o menos paralela a PC2. Vemos que PC1 separa los outliers de las muestras típicas, mientras que PC2 parece clasificarlas respecto del IO.

Tarea: Discute la ordenación de las muestras respecto al IO.

¿Qué consecuencias va a tener esta disposición si nuestro objetivo final va a ser construir un modelo multivariante para predecir IO?

¿Por qué no debemos descartar las muestras anómalas en este momento?

Estrategia: para que nos sirva de referencia, veremos las predicciones que se pueden alcanzar con el mejor modelo univariante posible (i.e., deberemos encontrar primero la mejor longitud de onda). Aquí únicamente estamos seleccionando

una variable, pero a menudo, la eliminación de zonas no significativas y selección de variables ("feature selection") conduce a mejoras apreciables en la calidad de los modelos multivariantes.

Localización de la longitud de onda más informativa en la predicción de IO: No sabemos todavía la región del espectro más adecuada para predecir el IO. Lo descubriremos correlacionando la absorbancia de las muestras con su correspondiente índice de octanos para cada una de las longitudes de onda: aquella longitud de onda que de lugar a una máxima correlación será la elegida.

Tarea: Construye una matriz 26×227 anexando la columna de índice de octanos (sitúa **y** a la izquierda de **X** por ejemplo). Calcula su matriz de correlaciones y dibuja R^2 frente a la longitud de onda. Aquella longitud de onda más correlacionada con IO será la



MÁSTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

PNT_MTEQ002

Rev0

19/10/2009

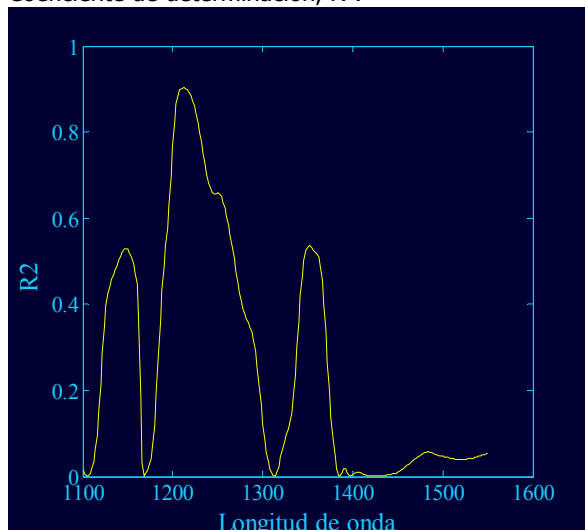
Pag. 8 de 15

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos.

PRÁCTICA: Casos prácticos de calibración multivariante

que proporcione un mayor valor de R . La mejor longitud de onda resulta ser 1206 nm. Confírmalo y estudia la zona espectral involucrada.

Coefficiente de determinación, R^2 :

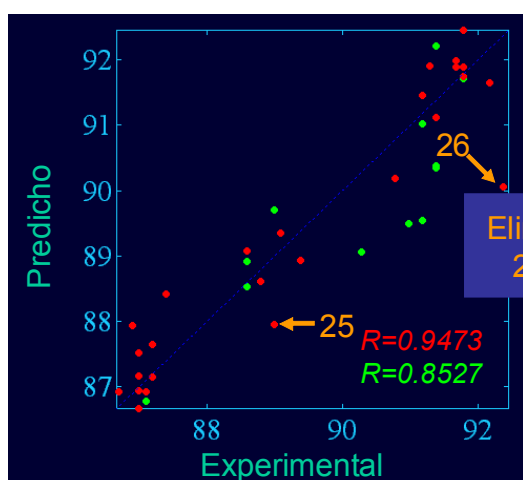


Con la ayuda de la figura adjunta, discute la idoneidad de las diferentes regiones espectrales para predecir el índice de octanos.

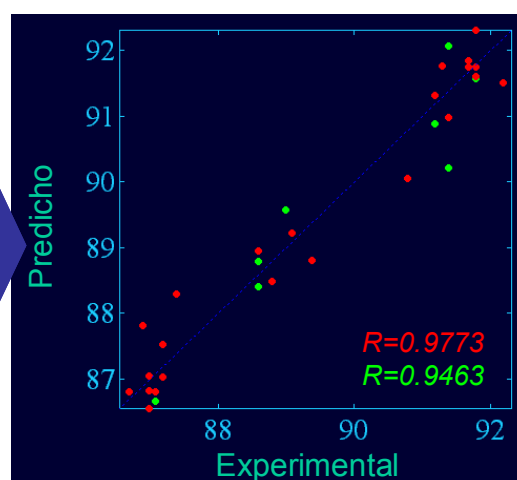
Calibración univariante: Podemos ver que la banda hacia 1200 nm incluye zonas correlacionables con el IO. Así pues, podríamos construir un calibrado univariante eligiendo la mejor longitud de onda.

Tarea: Construye un calibrado univariante, utilizando la columna asociada a 1206 nm y el índice de octanos. Examina la calidad del modelo prediciendo el IO de las muestras del conjunto de validación. Repite el proceso eliminando ahora las dos muestras anómalas (aditivadas con etanol) del conjunto de calibración, y de nuevo predice el IO de las muestras de validación con el nuevo modelo mejorado. Discute las diferencias entre las dos aproximaciones y sus consecuencias.

Calibración multivariante: Como observamos, eliminemos o no las dos muestras anómalas, el error de calibración da lugar a una estimación siempre optimista de la calidad predictiva del modelo, que es considerablemente más pobre. Sin embargo, estamos haciendo uso de la señal correspondiente a una única longitud de onda; cabe esperar que la calidad de las predicciones mejore de forma ostensible cuando seamos capaces de utilizar espectros completos. Pero para ello, necesitaremos técnicas de regresión especiales, capaces de resolver el problema de la colinealidad. Esto es: dos longitudes de onda cualesquiera vecinas proporcionan información similar (i.e., están correlacionadas), y este fenómeno plantea dificultades matemáticas de dependencia lineal (específicamente, inversión de matrices), que han conducido al desarrollo de nuevas formas de regresión, capaces de convertir a la colinealidad en una ventaja. Con Unscrambler podemos elegir usar regresión lineal múltiple (MLR), regresión en componentes principales (PCR), y regresión de mínimos cuadrados parciales -o proyección hacia estructuras latentes- (PLS1 para la predicción una variable, y PLS2 para la predicción simultánea de diversas variables correlacionadas entre sí). De entre estos sistemas, por brevedad, veremos el más útil en la práctica, PLS1, desarrollado por químicos para problemas químicos.



Eliminando
25 y 26





MÁSTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

PNT_MTEQ002

Rev0

19/10/2009

Pag. 9 de 15

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos.

PRÁCTICA: Casos prácticos de calibración multivariante

Regresión de mínimos cuadrados parciales

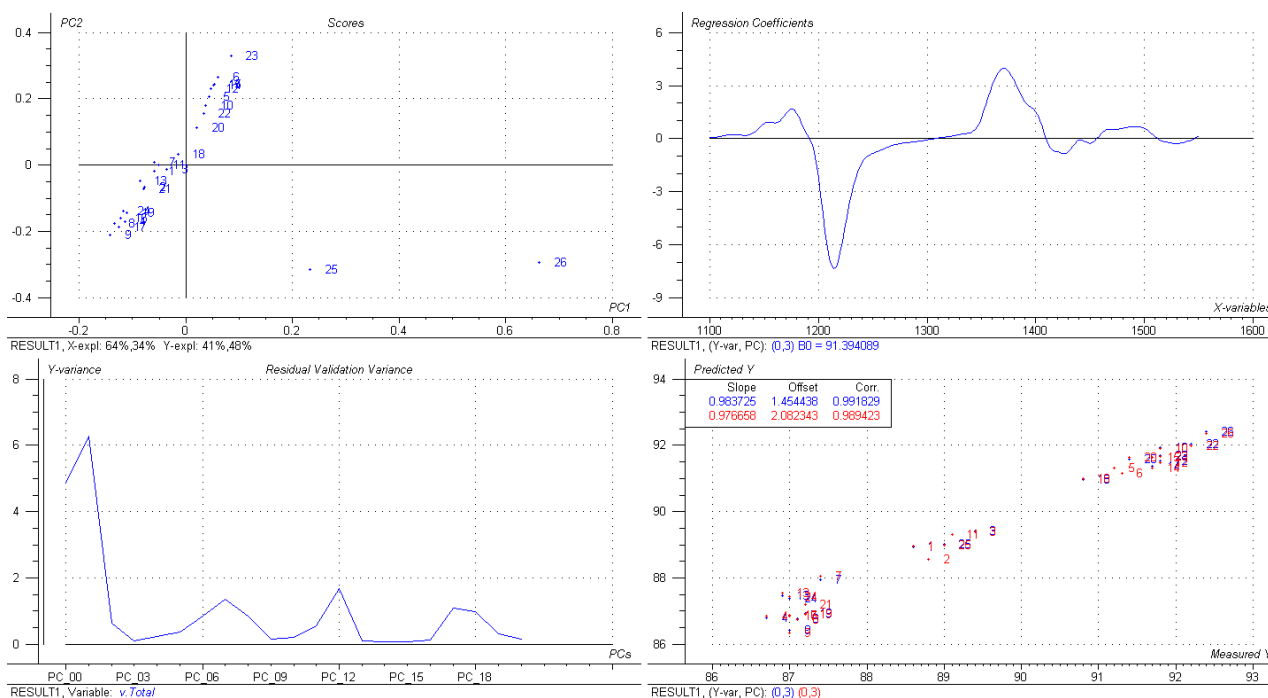
Tarea: Construye un modelo PLS1 con todas las muestras del conjunto de entrenamiento (**sin descartar muestras anómalas**). ¿Qué tipo de preprocesado (e.g., estandarización, centrado, autoescalado) sería el más correcto para este problema? Analiza los cuatro gráficos-resumen que Unscrambler proporciona (opción View – regression overview; selecciona construir el modelo con 3 variables latentes, proyectando sobre PC₁ y PC₂)

(a) **Gráfico de sedimentación** (o scree plot) (debajo, a la izquierda)- muestra la varianza residual de calibración o validación en función del número de componentes principales retenidos en el modelo. **Tarea:** ¿Por qué esta gráfica sugiere que el modelo es incorrecto? ¿Cuántos PCs debiéramos tomar a la vista de estos resultados? ¿Puedes identificar los outliers a partir de la tabla de diagnóstico (actívala desde el menú desplegable: view – warning list): cómo debiéramos usar esa información?

(b) **Gráfico de puntuaciones** (scores) sobre PC1-PC2 (arriba, a la izquierda)- Proyecta las objetos (i.e., muestras) sobre el sistema de componentes principales. **Tarea:** Las puntuaciones representan las coordenadas de los objetos en el sistema de ejes formado por los componentes principales. Discute la información que puedes extraer de este gráfico. Sustitúyelo por un biplot y analiza los resultados.

(c) **Gráfico de coeficientes de regresión** (arriba, a la derecha) – Representa la importancia de cada variable en la predicción de la propiedad escogida, en este caso IO. **Tarea:** ¿Qué zonas del espectro son más informativas para predecir el índice de octanos? ¿Cuáles contribuyen a incrementar el índice de octanos y cuáles contribuyen negativamente? ¿Hay alguna zona del espectro que no aporte información en la predicción del índice de octanos pero que sin embargo absorba significativamente?

(d) **Gráfico de correlación** entre respuestas predichas y reales (debajo, a la derecha)– Indica la calidad del ajuste; una predicción perfecta debería proyectar todos los puntos sobre una recta de ordenada en el origen cero y pendiente uno ("target line" o línea de diana, que podemos activar desde el menú contextual, picando con el botón derecho del ratón). **Tarea:** ¿Cómo es el modelo respecto del encontrado a una longitud de onda única? ¿Cómo sería si reducimos o aumentamos el número de variables latentes, respecto de las tres que el programa sugiere? Discute las ventajas e inconvenientes de hacerlo.



Otros gráficos auxiliares útiles: Puedes picar sobre cualquiera de los gráficos de la vista de resumen y reemplazarlo por otros. Explora las diferentes posibilidades que el programa te ofrece y completa tu análisis a la vista de estas representaciones



MÁSTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

PNT_MTEQ002

Rev0

19/10/2009

Pag. 10 de 15

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos.

PRÁCTICA: Casos prácticos de calibración multivariante

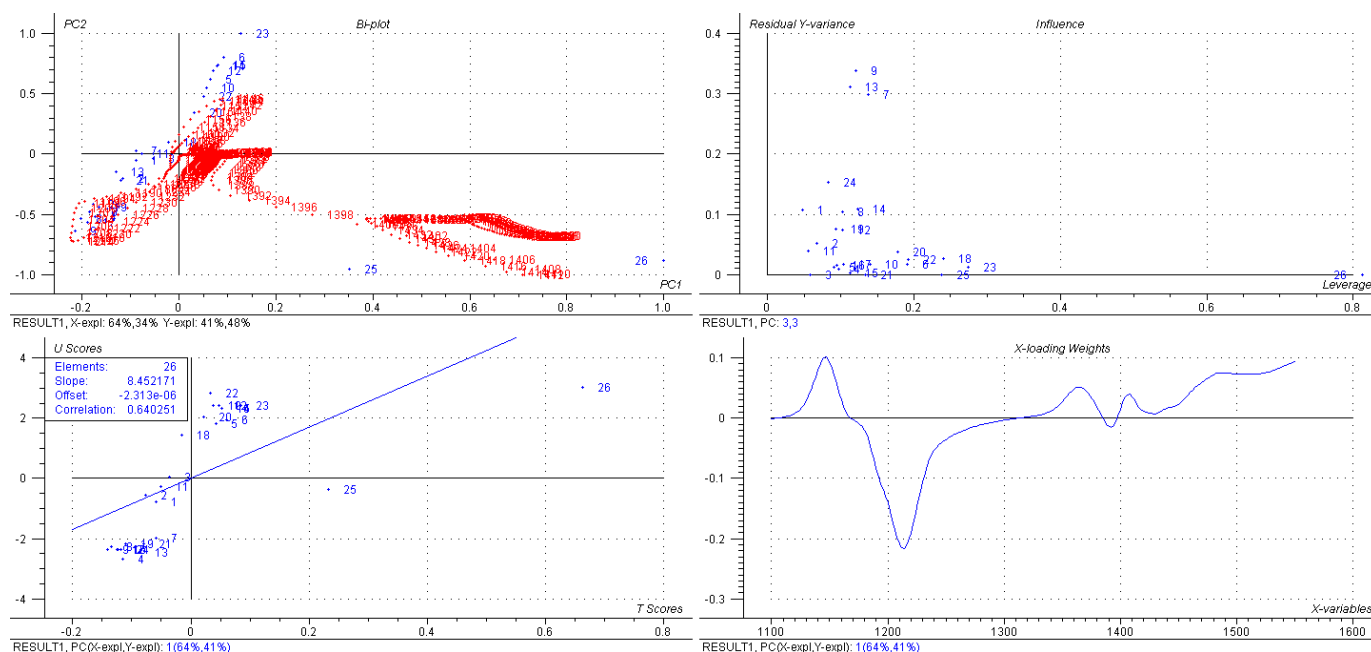
(solo se muestran algunos ejemplos; la posición es arbitraria; puedes explorar otras posibilidades y hacer uso de la ayuda del programa).

(a) **Biplot** o gráfico doble de cargas y puntuaciones (arriba, a la izquierda)- Permite asociar muestras y variables, y detectar relaciones de asociación. En este caso, se puede visualizar que longitudes de onda participan más en la descripción particular de cada muestra. *Tarea:* ¿Qué longitudes de onda contribuirían a incrementar y decrementar el índice de octanos? ¿Cuáles serían irrelevantes o no asociadas a esta propiedad?

(b) **Gráfico de influencia o capacidad de nivelación** (influence Plot, arriba, a la derecha)- Es extremadamente útil para medir la capacidad deformante sobre el modelo de las distintas muestras, y localizar outliers y puntuar su peligrosidad. En este gráfico se representa el residual de cada muestra (eje Y) frente a la distancia al centro del modelo o influencia (leverage). Las muestras más peligrosas son las que tienen simultáneamente un error alto y una gran distancia al centro del modelo, puesto que "atraen el modelo" para conseguir describirlas bien. *Tarea:* ¿Cómo son las muestras en este caso? ¿Hay outliers peligrosos?

(c) **Gráfico de influencia frente a puntuación Y (U-score)** (debajo a la izquierda) – Activar Plot - X-Y regression outliers. En esta gráfica se muestran las distancias al centro del modelo frente a las puntuaciones en el bloque Y. Es muy útil para la detección de puntos anómalos y para evidenciar la condición de correlación lineal entre las puntuaciones de los bloques X e Y en que se basa PLS1. Las muestras deben aparecer formando una banda y los outliers aparecerán aislados, destacando. Si se observan subgrupos, puede ser necesario formar un modelo PLS para cada clase. *Tarea:* Analiza los resultados y discute qué debiéramos hacer en este caso. ¿Hay contradicciones respecto de lo observado en las otras gráficas?

(d) **Gráfico de pesos de las cargas** (loading weights) – Los pesos de las cargas (W) muestra en cuánto cada variable predictora contribuye a la explicación de la respuesta a lo largo de cada componente principal. No confundir con el gráfico de coeficientes del modelo, que da una respuesta global y no está asociado a un componente principal determinado: las contribuciones cambian al considerar los diferentes componentes principales. Ni confundir tampoco con la matriz de cargas de X (matriz **P**), que mide el peso de cada variable predictora en cada componente principal. *Tarea:* Discute las contribuciones a lo largo de PC1, PC2 y PC3 y compara los resultados con los coeficientes de regresión, que vimos anteriormente.



Eliminación de outliers: Los valores anómalos deben ser eliminados UNO a UNO, reconstruyendo el modelo tras cada eliminación y los realizando de nuevo los diagnósticos correspondientes (en este caso, primero eliminamos la muestra 26 y después de repetir el análisis, la 25). La eliminación implica empobrecer la información disponible y perder potencialmente fenómenos que podrían estar presentes en nuevas muestras. Debe hacerse con precaución y solo en caso necesario.



MÁSTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

PNT_MTEQ002

Rev0

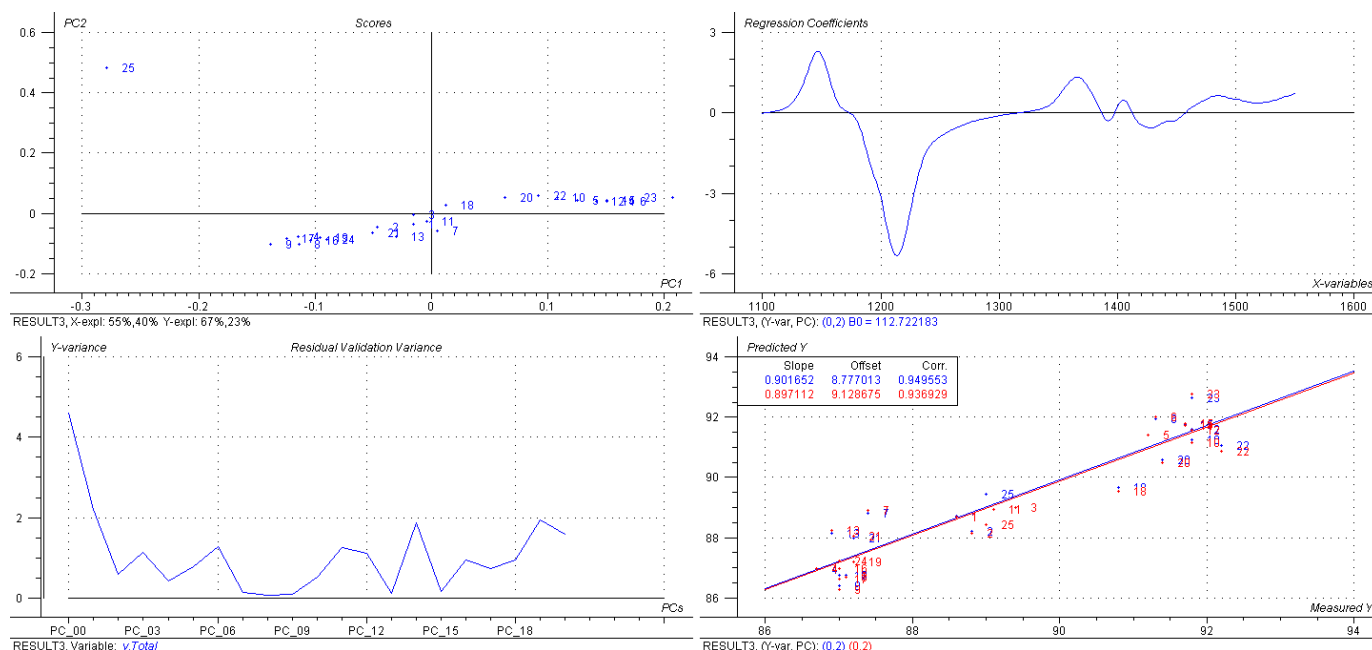
19/10/2009

Pag. 11 de 15

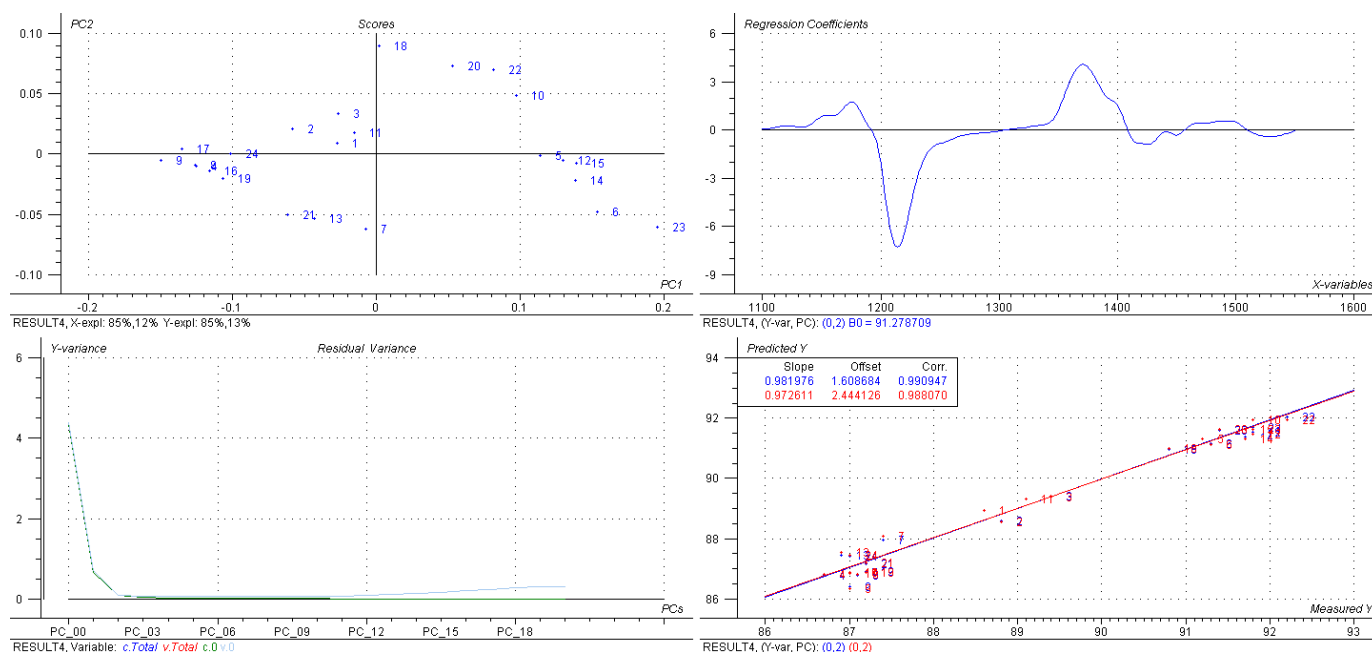
MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos.

PRÁCTICA: Casos prácticos de calibración multivariante

Sin la muestra 26: Fíjate en el cambio que se observa en el gráfico de sedimentación. *Tarea:* ¿Cuántos PC tendremos que tomar ahora para realizar el análisis detallado? ¿Podemos considerar que el calibrado es ya correcto? Completa estos gráficos con otros y discute la necesidad o no de continuar con la detección de valores anómalos.



Sin la muestra 25: Tras la eliminación de la muestra 25 el calibrado es ya completamente correcto. *Examina los diferentes gráficos y toma las conclusiones finales, antes de validar el conjunto de entrenamiento con las 13 muestras que dejamos aparte. ¿Debiéramos eliminar alguna otra muestra? ¿Cuál sería la siguiente? Compara el calibrado univariante con el multivariante.*





MÁSTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

PNT_MTEQ002

Rev0

19/10/2009

Pag. 12 de 15

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos.

PRÁCTICA: Casos prácticos de calibración multivariante

Construcción del modelo PLS1 final y análisis del mismo

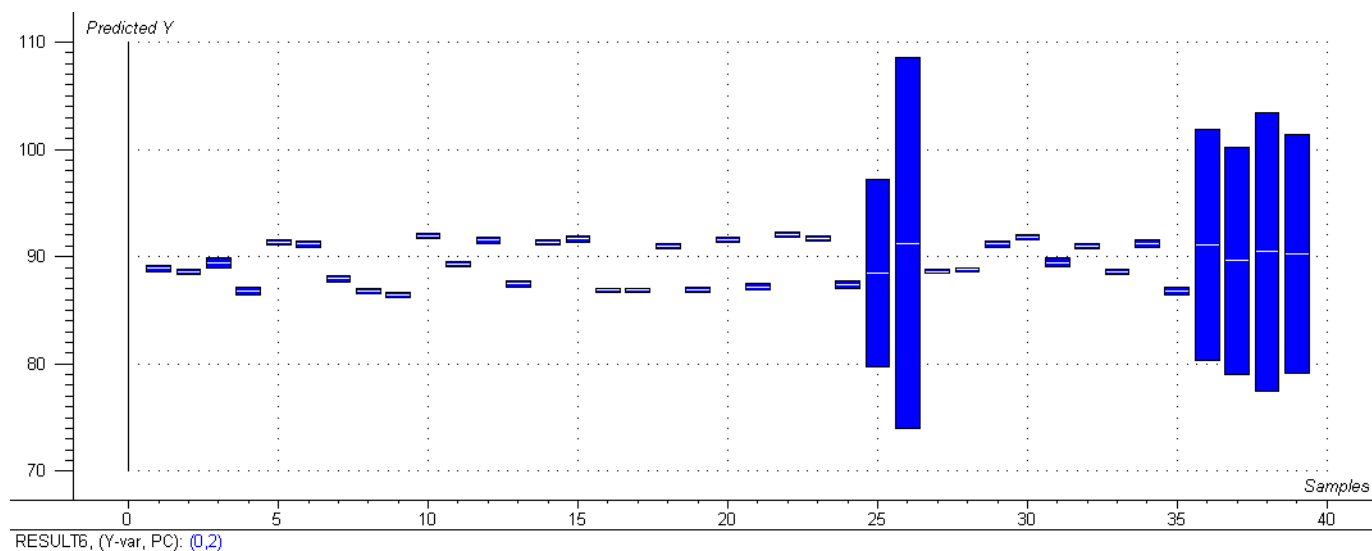
Los modelos multivariantes "blandos" no son únicos: podemos construir diversos modelos válidos, cada uno de los cuales separa estructura y ruido en las señales de forma diferente. Cada uno de estos modelos se caracteriza por un determinado número de componentes principales que hemos conservado: a menor número de componentes retenidos, más seguridad y robustez del modelo pero inferior capacidad predictiva. Si conservamos demasiados componentes, el modelo aparentemente ajusta mejor pero le estaremos dando más importancia a las variables irrelevantes y la incertidumbre experimental: se dice que el modelo será menos robusto, y "sobreajusta". Es decir, con un número de componentes excesivos, no se separa apropiadamente la parte estructural o relevante de los datos, y el modelo invierte esfuerzo en describir aspectos irrelevantes, pudiendo dar lugar a grandes errores cuando predigamos muestras externas. Hay que determinar la cantidad apropiada de componentes, y la forma de hacerlo es mediante validaciones. En la fase de construcción, usaremos únicamente el training set (**X** e **y**; las 26 muestras de calibración).

Tarea: Realiza diversos tipos de validación (*leave-one-out*, *crossvalidation*, *leverage correction*) y compara los gráficos de sedimentación que obtienes. ¿Cuántos componentes principales conviene retener en el modelo final (e.g., 1, 2...)? Discute las ventajas e inconvenientes de cada aproximación y haz tu decisión. También puedes validar con las 13 muestras que dejamos a parte.

Tarea: Realiza el análisis final de los resultados de PLS acuerdo a la decisión que hayas tomado.

Validación externa con las 13 muestras de referencia

Tarea: Salva en disco duro el modelo para poder aplicarlo (cierra la ventana y cuando te pregunte si quieres guardar los resultados, acepta salvar cambios). A continuación cierra sin salvar el resto de ventanas, dejando únicamente abierta la ventana inicial con la tabla de datos. Después, ve a "Task – Predict" y carga el modelo. Predice todas las muestras disponibles con el modelo final. Discute los resultados encontrados.





MÁSTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

PNT_MTEQ002

Rev0

19/10/2009

Pag. 13 de 15

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos.

PRÁCTICA: Casos prácticos de calibración multivariante

Anexo III. Protocolo para empleo de UNSCRAMBLER en validación de modelos de calibración multivariante. Análisis multicomponente (PLS2)

Datos: I+D de un método colorimétrico portátil para determinar in situ y simultáneamente el contenido en Ca y Mg (multicomponente o multianálisis) por formación de complejos (con azul de bromotimol, pH=11, NH₃) y medida del espectro visible. Las concentraciones de Ca y Mg han sido obtenidas por ICP.

Nº muestras = 130

Nº muestras Calibración: 68 (34 muestras x 2 réplicas)

Nº muestras de referencia (Validación): 62 (31 muestras x 2 réplicas)

Nº variables = 151 (espectro 400-700 nm cada 2 nm, restados de un blanco de reactivos)

NOTA-1: Los objetos de validación se procesaron 1 mes después de los de calibración.

NOTA-2: Los 8 primeros objetos de calibración (1-8) y de referencia (69-76) son patrones sintéticos, el resto (9-68, 77-130) muestras reales.

Objetivos: 1) Creación del modelo PLS2 predictivo de Ca y Mg simultáneamente, optimizado/validado con el bloque de validación 2) estudiar muestras y variables espectrales anómalas (en función del pretratamiento), 3) Ver si la matriz **X** (espectros) tienen una estructura relacionada con la distribución de Ca y Mg en las muestras. 4) Significado de las VLs. Los datos se hallan en el fichero **CaMg.00D** (\09MASTER\DatosM). Antes del análisis PLS se grafica el mapa de concentraciones de Ca y Mg en las muestras (**Y**).

a) Resolución vía software UNSCRAMBLER. Protocolo

NOTA: Las instrucciones sombreadas corresponden al menú

	COMENTARIOS	PROTOCOLO
1	Cargar datos	File/open / Ej2_PLS.00D
2	Diferenciar grupos/tipos de variables y muestras (a X se le asigna tipo espectro)	Modify/Edit set / Variable sets: [Add]/ Name: Y / Set interval: 1-2 / [OK] / [Add] Name: X/ Data type: Spectra / Set interval: 3-153/ [OK] Show set of type: sample set/ [Add]/ Name: cal/ Set interval: 1-68/ [OK] [Add]/ Name: Valid/ Set interval: 69-130/ [OK]/[OK]
3.a	Gráfico 2D (visualizar Y).... <i>Opcional</i> Gráfico 4.2.1 (RESULTADOS)	Pinchar en columna 1 y sin soltar arrastrar hasta la 2 Plot/2D Scatter / [OK]
3.b	Desmarcar la selección (que desaparezca la selección de las dos columnas).... <i>Opcional</i>	Window/ Ej2_PLS/ clic en algún dato no seleccionado
4	Modelo PLS2 (más de 1 variable Y) Basado en los datos "cal" pero validado con "Valid" (un conjunto de validación) Datos centrados: Comprobar que Weights = All 1, para X e Y (si no [Weights]/ All/ 1/ update/OK)	Task/Regression Method: PLS2/ [solapa Samples]: All samples[130]/ Validación Method: Test set/ [Setup]/ Test samples: Manual selection: 69-130 / [OK] [X variables]/ Variable set: X[151] / [Y variables]/ Variable set: Y [2] <i>Number of Components: 7/ [OK]/ [View]</i>
5.a	Gráfico de Scores (diferenciar "sets" de calibración y validación)	Clic en ventana sup.izq Edit/Options / sample grouping/ Colors/Calibration and Validation/ [OK]
5.b	Gráfico de coeficientes de regresión para Ca (modelo de 3VLs; si no, cambiar a 3) Gráfico 4.2.2 (RESULTADOS)	Clic en ventana sup.der. Edit/ Options /Curve/ [OK]
5.c	Derivar datos (esta es una de las muchas opciones/condiciones de pretratamiento que se podrían optimizar hasta eliminar los problemas de ruido, etc)	Window/Ej2_PLS Modify/Transform/Derivatives/Norris / Samples: All samples/ Variables: X[151]/ Segment size: 25/ Difference: 24/ [OK]



MÁSTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

PNT_MTEQ002

Rev0

19/10/2009

Pag. 14 de 15

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos.

PRÁCTICA: Casos prácticos de calibración multivariante

6	Recalcular el modelo. Comprobar que los problemas han desaparecido (ej: los coeficientes para Ca y Mg, son continuos)	Task/ Regression / Keep Out of Calculation: 23,33 / [OK] / [View] Plot/ Regresion overview / Y variable : (Ca) / [OK] Plot/ Regresion overview / Y variable : (Mg) / [OK]
7	Salvar modelo	File/ Save/ ModeloPLS2/ [Save]

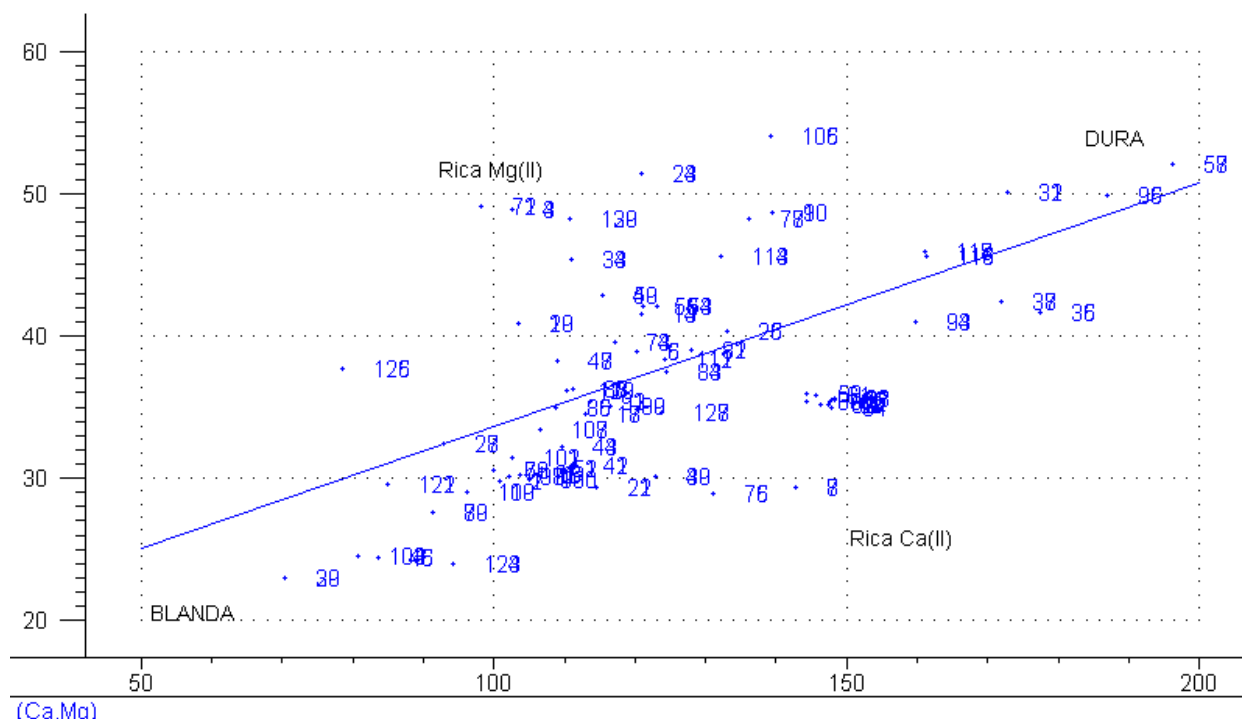
Objetivo: 5) Predicción con el modelo de regresión PLS2 de las concentraciones de Ca y Mg de las muestras de validación

NOTA: Las instrucciones sombreadas corresponden al menú

	COMENTARIOS	PROTOCOLO
1	Prededir las muestra de validación, de las que disponemos de valores de referencia	Window/ Ej2_PLS Task/ Predict/ Sample set: Valid [62]/ Model name: [Find] modeloPLS2 [Select] / [Y-reference]/ Include Y-reference / [OK] / [View]
2.a	Predicciones para Ca	Clic en ventana sup. Plot/prediction/ Predicted vs Reference/ Y-variable: Ca / [OK] View/Trend lines/ Target line
2.b	Predicciones para Mg Ver Gráfico 4.2.3 (RESULTADOS)	Clic en ventana inf. Plot/prediction/ Predicted vs Reference/ Quitar include table/ Y-variable: Mg View/Trend lines/ Target line
3	Cerrar	Cerrar todas las ventanas (x parte sup.der.) sin salvar nada

b) RESULTADOS (Comentados):

Gráfico 4.2.1. Distribución de concentraciones de Ca y Mg en las muestras que permite clasificarlas en agua blanda (ej. 29,30), dura (ej 57,58), rica en Ca (relación Ca/Mg elevada, ej 7,8) o rica en Mg (ej 71,72).





MÁSTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

PNT_MTEQ002

Rev0

19/10/2009

Pag. 15 de 15

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos.

PRÁCTICA: Casos prácticos de calibración multivariante

Gráfico 4.2.2. Los Scores (matriz espectral X) muestran como el eje VL1 es el eje de dureza (desde muestra 29 hasta muestra 58) mientras que el eje VL2 es el eje de riqueza relativa en Ca/Mg (muestras 7 y 72). Se observa que el conjunto de validación es bastante "similar" al conjunto de calibración. Los puntos 33 y 23 son posibles anómalos (ya que sus réplicas, 24 y 34) no presentan valores tan bajos en VL2. Los coeficientes de regresión sugieren problemas (ruido instrumental) o incluso fallo en sensor (error en los diodos; zona 128). Provisionalmente, 3 VLs parece adecuado (según la varianza en Y). El gráfico de Validación confirma el carácter anómalo del punto 33. Sugerencia: eliminar puntos (réplicas) 23 y 33. Probar derivar y/o suavizar (smoothing) los datos X.

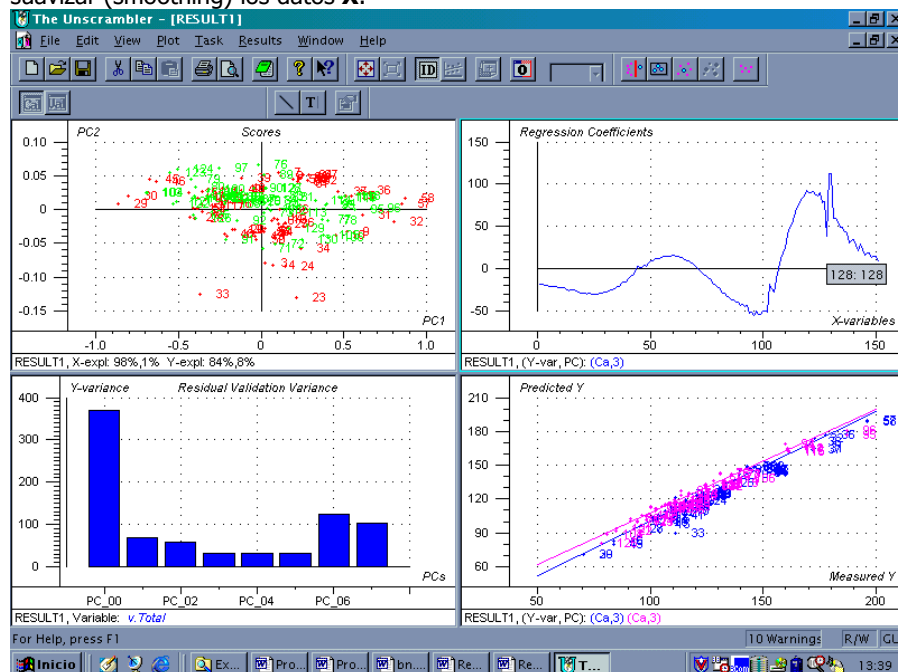


Gráfico 4.2.3. La adecuación del modelo para la predicción simultanea de Ca y Mg en las muestras de referencia parece adecuada comparada con la línea de referencia (target line; $b_0 = 0$; $b_1 = 1$).

