

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Descomposición de una matriz de rango k(3)

De modo que la relación entre ambos espacios: $T = U \cdot \Lambda = X \cdot V$

equivale a...
$$\begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \\ t_{31} & t_{32} \end{pmatrix} = \begin{pmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ u_{31} & u_{32} \end{pmatrix} \cdot \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{pmatrix} \cdot \begin{pmatrix} v_{11} & v_{12} \\ v_{21} & v_{22} \end{pmatrix}$$

y representa para el objeto 2...
$$\begin{matrix} t_{21} = & u_{21} \cdot \lambda_1 & = & x_{21} \cdot v_{11} + x_{22} \cdot v_{21} \\ t_{22} = & u_{22} \cdot \lambda_2 & = & x_{21} \cdot v_{12} + x_{22} \cdot v_{22} \end{matrix}$$

4

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Descomposición de una matriz de rango k(4)

- Las proyecciones de las muestras sobre ese sistema de ejes son las **puntuaciones o scores**
- Las proyecciones de los ejes originales con longitud=1 sobre ese sistema de ejes son las **cargas o loadings**
- El gráfico combinado o doble puntuaciones-cargas se denomina **biplot**

5

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Descomposición de una matriz de rango k(5)

- La suma de los elementos de la diagonal (**traza**) de la matriz Λ da **bajo ciertas condiciones** la **varianza total** de los datos :
$$\text{traza}(\Lambda) = \sum_{p=1}^m \lambda_p$$

Para que los autovalores correspondan a varianzas debe procesarse no X , sino la matriz de varianzas-covarianzas de $\text{Cov}(X) = (X_c^T \cdot X_c) / (n-1)$

- El porcentaje de varianza asociado a un determinado autovalor viene dado por :
$$V_p(\%) = 100 \frac{\lambda_p}{\text{traza}(\Lambda)}$$

Descomposición en valores singulares: el sistema de ecuaciones

$$U \cdot \Lambda = X \cdot V$$
 suele escribirse:
$$X = [U \cdot \Lambda \cdot V^T] = T \cdot P^T$$

- La descomposición de una determinada matriz es única excepto en el signos de U y V , que pueden aparecer invertidos (son direcciones de eje)

6

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Descomposición de una matriz de rango k

(9)

Ejemplo:

$X = \begin{bmatrix} 2 & -1 \\ -5 & 4 \\ 1 & -2 \\ 3 & 0 \end{bmatrix}$

$Xc = \begin{bmatrix} 1.75 & -1.25 \\ -5.25 & 3.75 \\ 0.75 & -2.25 \\ 2.75 & -0.25 \end{bmatrix}$

$||Xc|| = 6.22 \quad 4.56$

Tomamos x_1 como t
y repetimos (2) - (5)

Puntuaciones

Cargas

Iteracion 1

2.2140 0.8229
-6.3873 -0.5682
1.9593
2.4687

Iteracion 2

2.2130 0.8211
-6.3886 -0.5708
1.9627
2.4633

Iteracion 3

2.2129 0.8210
-6.3887 -0.5710
1.9629
2.4629

Iteracion 4

2.2129 0.8210
-6.3887 -0.5710
1.9629
2.4629

Ahora sustraemos a Xc la
contribución del primer factor

$Xc' = Xc - \begin{bmatrix} 2.2129 \\ -6.3887 \\ 1.9629 \\ 2.4629 \end{bmatrix} \cdot \begin{bmatrix} 0.8210 \\ -0.5710 \end{bmatrix}^T$

$Xc' = \begin{bmatrix} -0.0667 & 0.0135 \\ -0.0051 & 0.1021 \\ -0.8615 & -1.1292 \\ 0.7281 & 1.1563 \end{bmatrix}$

Repetimos el proceso con Xc'
para extraer el 2º factor :

10

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Descomposición de una matriz de rango k

(10)

X

Z

BEN SO₂

z1 z2

48 5.0 1.1411 -0.4583

44 4.8 0.8317 -0.6110

46 5.2 0.9864 -0.3055

42 4.8 0.6770 -0.6110

25 4.9 -0.6383 -0.5346

23 4.7 -0.7930 -0.6874

16 7.5 -1.3346 1.4511

22 7.9 -0.8704 1.7567

Cov = R = $\begin{bmatrix} 1.0000 & -0.6179 \\ -0.6179 & 1.0000 \end{bmatrix}$

$V = \begin{bmatrix} 0.7071 & -0.7071 \\ -0.7071 & -0.7071 \end{bmatrix}$

$\Lambda = \begin{bmatrix} 1.6179 \\ 0.3821 \end{bmatrix}$ ($\lambda_1 = 81\%$ correlación)
($\lambda_2 = 19\%$ correlación)

- La varianza total de **Z**, al haber sido autoescalada por columnas, es 2
- Ya que queremos descomponer buscando las direcciones de mayor varianza, hay que **calcular Cov (Z)**
- Al haber autoescalado, la matriz de covarianzas **coincide con R**
- Con este tratamiento la SVD proporciona las direcciones (**V**) y los autovalores (**Λ**)

11

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Descomposición de una matriz de rango k

(11)

- Las puntuaciones normalizadas se obtienen a partir de las fórmulas de transición y **V**:
 $U = T \cdot \Lambda^{-1} = (X \cdot V) \cdot \Lambda^{-1}$
- Los dos vectores singulares columna normalizados (**U**) son:

U2

U1

2.5
2
1.5
1
0.5
-0.5
-1
-1.5

más cont.

más org.

6
5
4
3
2
1

7
8


4
3
2
1

$U = \begin{bmatrix} 0.6990 & -1.2637 \\ 0.6305 & -0.4084 \\ 0.5646 & -1.2601 \\ 0.5629 & -0.1220 \\ -0.0453 & 2.1706 \\ -0.0462 & 2.7396 \\ 1.2175 & -0.2158 \\ -1.1481 & -1.6402 \end{bmatrix}$

- Al haber usado el 100% de la información, la disposición de los datos coincide con la representación **z_{BENZ}** vs. **z_{SO₂}**, (excepto en la orientación)
- PC1** contrasta el origen de la contaminación: **contaminación orgánica** a la derecha e **inorgánica** a la izquierda
- PC2** indica la cantidad de contaminante

12

4



MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
 PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Reducción de dimensiones (12)


- Una de las aplicaciones más importantes de la rotación propia
- La rotación propia modela los datos obteniendo contribuciones ordenadas, y **el modelo** se puede **depurar** para incluir únicamente lo relevante
- Los **modelos blandos** dividen la matriz de datos en dos partes:

Estructura - la parte relevante o informativa

Ruido - la parte irrelevante o indeseable



- La reconstrucción puede hacerse tan precisa como se desee: es función del número de **factores (k)** o componentes principales escogidos

	<p>MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA</p> <p>MÓDULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica</p>
<div data-bbox="185 1358 832 1383"><h2>Reducción de dimensiones (13)</h2></div> <div data-bbox="185 1417 848 1805"><ul style="list-style-type: none">• Cuantos más componentes principales se usen (mayor sea k)...<ul style="list-style-type: none">• más perfecta será la reconstrucción de X• más deficiente será la separación entre la parte relevante de X y el ruido, E• En el caso de PCA obtenido por SVD, la reconstrucción de la matriz depurada se realiza:<ul style="list-style-type: none">• Con las k primeras columnas de U• Con las k primeras columnas de V (o filas de V^T)• Con una submatriz diagonal de Λ incluyendo los k primeros λ• Tras la rotación, normalmente basta con observar las distribuciones en los planos PC1-PC2 y ocasionalmente en PC1-PC3 y PC2-PC3</div>	

14

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Reducción de dimensiones

(14)

Imaginemos que tenemos **3 sensores (S1-S3)**, y tomamos las medidas correspondientes en **5 muestras (M1-M5)**

Datos originales

	S1	S2	S3
M1	10.12	22.72	19.52
M2	12.56	28.41	20.69
M3	14.45	31.64	23.85
M4	17.56	37.38	28.86
M5	19.45	41.29	30.17

Pretratamiento:
Centrado por columnas

Datos centrados

$$X = \begin{pmatrix} -4.708 & -9.571 & 5.100 \\ -2.268 & -3.877 & -3.923 \\ -0.378 & -0.649 & -0.770 \\ 2.732 & 5.095 & 4.243 \\ 4.622 & 9.002 & 5.549 \end{pmatrix} = \underset{5}{\overset{3}{X}}$$

15

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Reducción de dimensiones

(15)



nPC	Eigen	Var(%)	AcVar(%)
1	89.51	99.16	99.156
2	0.76	0.84	99.999
3	0.0012	0.0013	100.00

16

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Reducción de dimensiones

(16)

La rotación propia descompondrá a la matriz **X** en dos matrices, **T** y **P**

- Las puntuaciones **T** describen las muestras (= "objetos") en el sistema de componentes principales
- Las cargas **V** describen las variables (= "características") en el sistema de componentes principales

Cada columna en **T** y **V** corresponderá a un componente principal diferente

Matriz de las puntuaciones

$$\begin{matrix} 5 & \mathbf{T} & 3 \\ \begin{pmatrix} -11.7888 & 0.8988 & 0.0210 \\ -5.8417 & -1.1977 & -0.0270 \\ -1.0338 & -0.2967 & 0.0151 \\ 7.1254 & 0.8110 & -0.0474 \\ 11.5388 & -0.2154 & 0.0384 \end{pmatrix} \\ \begin{matrix} \uparrow & \uparrow & \uparrow \\ \text{PC}_1 & \text{PC}_2 & \text{PC}_3 \end{matrix} \end{matrix}$$

Matriz de las cargas

$$\begin{matrix} 3 & \mathbf{V} & 3 \\ \begin{pmatrix} 0.3964 & -0.0604 & 0.9161 \\ 0.7717 & -0.5187 & -0.3681 \\ 0.4974 & 0.8528 & -0.1590 \end{pmatrix} \\ \begin{matrix} \uparrow & \uparrow & \uparrow \\ \text{PC}_1 & \text{PC}_2 & \text{PC}_3 \end{matrix} \end{matrix}$$

17

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Reducción de dimensiones

(17)

$$\mathbf{X} = \mathbf{X}_1^* + \mathbf{X}_2^* + \mathbf{X}_3^* = \mathbf{T}_1 \cdot \mathbf{V}_1' + \mathbf{T}_2 \cdot \mathbf{V}_2' + \mathbf{T}_3 \cdot \mathbf{V}_3'$$

Contribución del PC1	Contribución del PC2	Contribución del PC3
$\begin{pmatrix} -4.673 & -9.097 & -5.864 \\ -2.316 & -4.508 & -2.906 \\ -0.410 & -0.798 & -0.514 \\ 2.824 & 5.498 & 3.544 \\ 4.574 & 8.904 & 5.739 \end{pmatrix}$	$\begin{pmatrix} -0.054 & -0.466 & 0.767 \\ 0.072 & 0.621 & -1.021 \\ 0.018 & 0.154 & -0.253 \\ -0.049 & -0.421 & 0.692 \\ 0.013 & 0.112 & -0.184 \end{pmatrix}$	$\begin{pmatrix} 0.019 & -0.008 & -0.003 \\ -0.025 & 0.010 & 0.004 \\ 0.014 & -0.006 & -0.002 \\ -0.043 & 0.017 & 0.008 \\ 0.035 & -0.014 & -0.006 \end{pmatrix}$
\mathbf{X}_1^*	\mathbf{X}_2^*	\mathbf{X}_3^*

El primer componente principal es responsable de la mayor parte de la variación observada en los datos (99.16 %)

El segundo no aporta información relevante y puede ser despreciado (0.84 %)

El tercero es únicamente ruido (<0.01%)

Podemos comparar cada contribución con la matriz original :

$$\mathbf{X} = \begin{pmatrix} -4.708 & -9.571 & 5.100 \\ -2.268 & -3.877 & -3.923 \\ -0.378 & -0.649 & -0.770 \\ 2.732 & 5.095 & 4.243 \\ 4.622 & 9.002 & 5.549 \end{pmatrix}$$

18

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Reducción de dimensiones (18)

Así, podemos **reconstruir** la matriz **X** con **sólo información relevante** ...

Si reconstruimos **X** con sólo el primer componente principal, **X**= X_1^*

$$\mathbf{X} = \begin{pmatrix} -4.673 & -9.097 & -5.864 \\ -2.316 & -4.508 & -2.906 \\ -0.410 & -0.798 & -0.514 \\ 2.824 & 5.498 & 3.544 \\ 4.574 & 8.904 & 5.739 \end{pmatrix}$$

Estructura X_1^*

... y somos capaces de explicar el **99.16 %** de la variación observada.

El error será entonces ...

$$\begin{pmatrix} -0.035 & -0.474 & 0.763 \\ 0.048 & 0.631 & -1.017 \\ 0.032 & 0.148 & -0.255 \\ -0.092 & -0.403 & 0.699 \\ 0.048 & 0.098 & -0.190 \end{pmatrix}$$

Error (**E**) $X_2^* + X_3^*$

Si reconstruimos **X** con los dos primeros componentes principales, **X**= $X_1^* + X_2^*$

$$\mathbf{X} = \begin{pmatrix} -4.727 & -9.564 & -5.097 \\ -2.243 & -3.887 & -3.927 \\ -0.392 & -0.644 & -0.767 \\ 2.776 & 5.078 & 4.236 \\ 4.587 & 9.016 & 5.556 \end{pmatrix}$$

Estructura $X_1^* + X_2^*$

... y somos capaces de explicar el **99.99 %** de la variación observada.

El error será entonces ...

$$\begin{pmatrix} 0.019 & -0.008 & -0.003 \\ -0.025 & 0.010 & 0.004 \\ 0.014 & -0.006 & -0.002 \\ -0.043 & 0.017 & 0.008 \\ 0.035 & -0.014 & -0.006 \end{pmatrix}$$

Error (**E**) X_3^*

Si reconstruimos **X** con todos los componentes principales, **X**= $X_1^* + X_2^* + X_3^*$ y **E**=0

19

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Reducción de dimensiones (19)

PC	Eigen	Var(%)	AcVar(%)
1	3.10	77.48	77.48
2	0.87	21.73	99.21
3	0.03	0.75	99.96
4	0.00	0.04	100.00

Con todas las variables (BEN, TOL, PS y SO₂)

Con 2 variables (BEN, SO₂)

PC2

PC1

El **signo** de los PCs es indeterminado (sólo es significativa su dirección)

Aspecto similar debido a la fuerte correlación de PS con SO2 (r=0.990) y BEN con TOL (r=0.973)

20

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Selección del número óptimo de PCs (20)

Necesario establecer **criterios** para determinar el **valor idóneo de k** para modelar la estructura:

Número de fuentes significativas de **varianza** deducidas a partir de la naturaleza del problema

Porcentaje de **varianza acumulada explicada** si se conoce la cantidad típica en muestras similares (e.g., 60%)

Gráfico de **autovalores u otros similares** punto en el que la caída se atenúa o se alcanza la estabilización (demasiado subjetivo)

Gráfico de **cocientes de autovalores sucesivos** ($\lambda_{i+1} / \lambda_i$) para exaltar las diferencias en intensidad de las caídas, de modo que las zonas de reducción fuerte se convierten en picos

Los procedimientos más seguros se basan en determinar la **calidad predictiva real** (validación)

21

7

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Selección del número óptimo de PCs (21)

Gráfico de cocientes de autovalores sucesivos

Café de grano (k=2)

Harinas (k=3)

Café

Harinas

Orden en el denominador (i)

22

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Calibración: Objetivo general

Sustituir

Medida costosa

Y

Variable Predicha (respuesta)

estableciendo una relación numérica

Función de calibración

...por

Medida accesible

X

Variable de Predicción (predictor)

con el propósito de utilizar X para predecir Y

...debido a razones económicas técnicas éticas peligrosidad

No interesa el modelo en sí, sino su rendimiento en las predicciones de Y

La función de calibración puede ser muy compleja o incluso no existir formalmente (redes neuronales artificiales)

23

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Calibración multivariante

$$(y_1, y_2, ..., y_q) = F (x_1, x_2, ..., x_n)$$

En apariencia ...

Se usa más de una variable predictora (x) para predecir una o varias respuestas (y)

pero en realidad ...

Se modela más de una fuente de varianza (=causa de variación en las respuestas), aunque la variable de respuesta sea una sola

24

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Por ejemplo...

Determinación simultánea de **varios constituyentes** de una disolución desconocida, a partir de los **espectros UV-visible completos** de varias mezclas

✓

• Usamos *más de un predictor* (=λ)

✓

• Hay *más de una fuente significativa de varianza* (=compuesto)

Es un problema multivariante

25

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Calibración

Univariante

Multivariante

• Sencilla (gráficas, calculadoras)

• Modelos simples y muy manejables

• En presencia de interferencias...

• Compleja (ordenadores)

• Modelos abstractos, incluso “ausencia” de modelos

• En presencia de interferencias...

Predicciones incorrectas

Predicciones correctas

Otra razón más ...

Hoy en día disponemos de **instrumentos** que proporcionan **muchas medidas** de una muestra única a un **coste cada vez más bajo**...

...¿no resultará indeseable **renunciar** a una **información** rica y extensa, potencialmente muy valiosa, reduciéndola a una medida única?

Calibración multivariante

26

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Pero no todo son ventajas

• Se requieren instrumentos de cálculo potentes (**ordenadores** y **software**)

• Modelos más complejos de interpretar y crear (**inviable cálculo manual**)

• Se abre la puerta a dos problemas: la entrada de información:

• **innecesaria** (= irrelevante para el fenómeno bajo estudio)

• **redundante** (= que aporta información muy similar)

• El modo en que se afronta la información **innecesaria** y **redundante** en los datos da origen a diversas técnicas multivariantes (**MLR**, **PCR** y **PLS**)

• En esta sesión nos centraremos **MLR**

27

9

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

MLR, PCR y PLS - Dos enfoques son posibles (1)

Calibración de **retroceso** (CLS)

absorbancias = F (concentraciones)

- Llamada también:
calibración clásica
calibración directa
(cuidado: denominación conduce a equívocos)
- Frecuente en problemas **univariantes**
- Aparentemente un **enfoque ilógico**
- Comprende **tres etapas**:
(1) Ajuste del modelo (calibración), (2) Inversión, (3) Predicción

28

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

MLR, PCR y PLS - Dos enfoques son posibles (2)

Calibración de **avance** (ILS)

concentraciones = F (absorbancias)

- Llamada a veces
calibración inversa
(cuidado: denominación conduce a equívocos)
- El enfoque **más usado en problemas multivariantes**
- Aparentemente el **enfoque más lógico**
- Comprende **dos etapas**:
(1) Ajuste del modelo (calibración), y (2) Predicción

29

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Caso de estudio

9 espectros UV-visible de disoluciones conteniendo mezclas ternarias de los compuestos **a** , **b** y **c**.

Las composiciones de los patrones siguen el siguiente esquema o **diseño experimental**

30

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MÓDULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
 PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Conjunto de calibración o de entrenamiento

9 disoluciones

	Absorbancias y longitudes de onda (nm)								Concentración		
	380	430	480	530	580	630	680	730	a	b	c
1	0.0010	0.0082	0.0363	0.0835	0.0552	0.0068	0.0004	-0.0002	50	50	50
2	0.0013	0.0167	0.0692	0.1334	0.0822	0.0115	0.0014	0.0000	100	50	50
3	0.0009	0.0082	0.0376	0.1120	0.0759	0.0075	0.0005	0.0001	50	100	50
4	0.0020	0.0163	0.0713	0.1617	0.1027	0.0124	0.0012	-0.0002	100	100	50
5	0.0006	0.0074	0.0360	0.0899	0.0605	0.0071	0.0005	0.0002	50	50	100
6	0.0016	0.0161	0.0691	0.1396	0.0875	0.0125	0.0009	-0.0002	100	50	100
7	0.0006	0.0076	0.0379	0.1176	0.0816	0.0082	0.0008	0.0005	50	100	100
8	0.0017	0.0165	0.0717	0.1674	0.1091	0.0134	0.0002	0.0002	100	100	100
9	0.0012	0.0126	0.0538	0.1263	0.0822	0.0099	0.0003	-0.0001	75	75	75

...datos que
 corresponde a los
 siguientes 9
 espectros
 simplificados,
 muestreados
 cada 50 nm:

Bloque X

8

A

9 × 8

9

Bloque Y


3

C

9 × 3

9

31



MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MÓDULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Conjunto de validación o de test

30 disoluciones de concentraciones aleatorias situadas en el intervalo cubierto por el conjunto de entrenamiento

Bloque X

A

30 x 8

30


Bloque Y

C

30 x 3

30

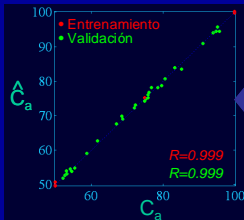
- Es habitual que, si se usa como conjunto de validación un set externo, éste incluya muchas menos muestras que el de calibración (1/3)
- En este caso el problema es sintético y queremos visualizar mejor los fallos de cada enfoque



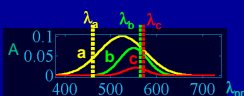
MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

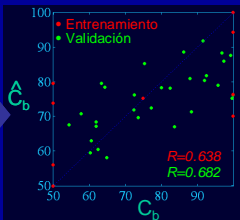
MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
 PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Resolución calibración univariante de avance con las λ óptimas

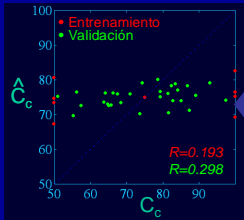


Compuesto a
 $\lambda_a = 458 \text{ nm}$





Compuesto b
 $\lambda_b = 564 \text{ nm}$



Compuesto c
 $\lambda_c = 566 \text{ nm}$

33

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Resolución mediante calibración univariante, con las λ óptimas

- Con calibración univariante:
 - Podremos predecir correctamente **a**, pero obtendremos
 - Predicciones de **b** y **c** completamente erróneas
- Carecemos de **diagnósticos intrínsecos** que nos revelen la presencia de interferencias
- Habrà que recurrir a nuevas formas de calibración

34

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Calibración con enfoques “duros”

- Regresión Lineal Múltiple de Avance
- Regresión Lineal Múltiple de Retroceso

- Un **modelo duro** considera que *todas las fuentes de varianza* en el conjunto de entrenamiento son *conocidas* y están *correlacionadas con la respuesta*, pero *no entre sí* (i.e., son independientes)
- Los **modelos duros** dan lugar a *calibrados erróneos* cuando:
 - Existe *correlación entre los descriptores* (o entre las respuestas en ciertos casos)
 - El conjunto de entrenamiento no es *suficientemente informativo*

35

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

1. Regresión Lineal Múltiple (MLR) de avance

- Si se midiera a una λ única, la concentración del analito i sería:
$$c = b_0 + b_1 \cdot A_i$$
- Si se miden n longitudes de onda tendríamos:
$$c = b_0 + b_1 \cdot A_1 + b_2 \cdot A_2 + \dots + b_n \cdot A_n$$
- Con m muestras, debiera cumplirse:

$$\left. \begin{aligned} c_1 &= b_0 + b_1 \cdot A_{11} + b_2 \cdot A_{12} + b_3 \cdot A_{13} + b_4 \cdot A_{14} + \dots + b_n \cdot A_{1n} \\ c_2 &= b_0 + b_1 \cdot A_{21} + b_2 \cdot A_{22} + b_3 \cdot A_{23} + b_4 \cdot A_{24} + \dots + b_n \cdot A_{2n} \\ c_3 &= b_0 + b_1 \cdot A_{31} + b_2 \cdot A_{32} + b_3 \cdot A_{33} + b_4 \cdot A_{34} + \dots + b_n \cdot A_{3n} \\ &\vdots \\ c_m &= b_0 + b_1 \cdot A_{m1} + b_2 \cdot A_{m2} + b_3 \cdot A_{m3} + b_4 \cdot A_{m4} + \dots + b_n \cdot A_{mn} \end{aligned} \right\}$$

Un sistema de **m ecuaciones** con **$n+1$ incógnitas** que tendrá solución siempre que m sea *igual* (solución exacta) o *mayor* (solución de mínimos cuadrados) que **$n+1$**

36

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

1. Regresión Lineal Múltiple (MLR) de avance

Esta expresión puede ser reescrita matricialmente como:

$$\begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} 1 & A_{11} & A_{12} & \cdots & A_{1n} \\ 1 & A_{21} & A_{22} & \cdots & A_{2n} \\ 1 & A_{31} & A_{32} & \cdots & A_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} \cdot \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

↑
Variable predicha

Offset Tabla de absorbancias
Variables predictoras

↑
Parámetros del modelo

... o aún más esquemáticamente:

9

C_i

1

=

9

A

8+1

·

8+1

B

1

... ya que hay 9 patrones y la tabla contiene 8 longitudes de onda

37

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Olvidemos por un momento que los datos son matriciales y que nos interesa un ajuste de mínimos cuadrados

Consideremos el compuesto **i**, cuya concentración debe ser predicha con el modelo es $C_i = A \cdot B_i$

El valor de B_i correspondiente a la solución es “ $B_i = C_i / A$ ”

Esta operación **no se puede hacer con matrices**, pero existe una equivalente :

$$B_i = (A^T \cdot A)^{-1} \cdot A^T \cdot C_i$$

Esta operación sería la **equivalencia matricial** de dividir “ C_i / A ”

Además, **es la solución de mínimos cuadrados** al problema

38

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Es fácil deducir la solución de mínimos cuadrados para la **calibración univariante de avance** a partir de $B = (A^T \cdot A)^{-1} \cdot A^T \cdot C$. Basta con sustituir:

$$\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ A_1 & A_2 & A_3 & \cdots & A_m \end{pmatrix} \cdot \begin{pmatrix} 1 & A_1 \\ 1 & A_2 \\ 1 & A_3 \\ \vdots & \vdots \\ 1 & A_m \end{pmatrix}^{-1} \cdot \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ A_1 & A_2 & A_3 & \cdots & A_m \end{pmatrix} \cdot \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_m \end{pmatrix}$$

Es decir :

$$\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} m & \Sigma A \\ \Sigma A & \Sigma A^2 \end{pmatrix}^{-1} \cdot \begin{pmatrix} \Sigma c \\ \Sigma A c \end{pmatrix}$$

de donde:

$b_0 = \frac{\Sigma A^2 \cdot \Sigma c - \Sigma A \cdot \Sigma c}{m \cdot \Sigma A^2 - (\Sigma A)^2}$

$b_1 = \frac{m \cdot \Sigma A c - \Sigma A \cdot \Sigma c}{m \cdot \Sigma A^2 - (\Sigma A)^2}$

Análogamente, en **calibración univariante de retroceso**:

$$\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ c_1 & c_2 & c_3 & \cdots & c_m \end{pmatrix} \cdot \begin{pmatrix} 1 & c_1 \\ 1 & c_2 \\ 1 & c_3 \\ \vdots & \vdots \\ 1 & c_m \end{pmatrix}^{-1} \cdot \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ c_1 & c_2 & c_3 & \cdots & c_m \end{pmatrix} \cdot \begin{pmatrix} A_1 \\ A_2 \\ A_3 \\ \vdots \\ A_m \end{pmatrix}$$

Es decir :

$$\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{pmatrix} m & \Sigma c \\ \Sigma c & \Sigma c^2 \end{pmatrix}^{-1} \cdot \begin{pmatrix} \Sigma A \\ \Sigma c A \end{pmatrix}$$

de donde:

$b_0 = \frac{\Sigma c^2 \cdot \Sigma A - \Sigma c \cdot \Sigma A}{m \cdot \Sigma c^2 - (\Sigma c)^2}$

$b_1 = \frac{m \cdot \Sigma c A - \Sigma c \cdot \Sigma A}{m \cdot \Sigma c^2 - (\Sigma c)^2}$

39

13

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

El ajuste de la concentración de **a** ha sido satisfactorio, pero **no ha mejorado** los resultados de la calibración univariante. Y lo que es peor ...

Los compuestos **b** y **c** tienen sus bandas de absorción *demasiado próximas*.

Si repetimos el tratamiento anterior sustituyendo **C_a** por **C_b** y **C_c** obtenemos:

Compuesto **b**

Perfectas predicciones del **conjunto de entrenamiento**
Fallos catastróficos en el **conjunto de validación**

Compuesto **c**

Inaceptable

43

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

¿Qué ha ido mal?

Si analizamos la expresión que proporciona la solución de mínimos cuadrados de MLR de avance, vemos que tiene un punto débil:

$B = (A^T \cdot A)^{-1} \cdot A^T \cdot C$

$A^T \cdot A$ es una matriz simétrica de dimensiones:

$$\begin{matrix} n+1 \\ A^T \end{matrix} \cdot \begin{matrix} m \\ A \end{matrix} = \begin{matrix} n+1 \\ P \end{matrix}$$

En el ejemplo: (9x9)

Ese producto **P** se denomina **matriz de momentos**⁽¹⁾, y describe la **asociación entre columnas** o **filas** de la matriz **A**. La matriz **P** **debe poder invertirse**

La **inversión de P** **no será posible** (división por cero) **cuando** alguna fila o columna sea combinación lineal de otras (el **rango** de **P** sea menor que **n+1**):

- Si existen **menos patrones independientes que longitudes de onda**
- Si existe **colinealidad entre espectros** (espectros muy parecidos)

(1) Si la matriz **A** ha sido centrada **P** es proporcional a la matriz de varianzas- covarianzas. Si **A** ha sido autoescalada, **P** \propto matriz de correlaciones

44

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

¿Qué significa colinealidad?

Para ilustrarlo, supongamos que registramos únicamente 2 longitudes de onda. El modelo será **C = A · B = B₀ + A(λ₁) · B₁ + A(λ₂) · B₂**, cuya representación geométrica es un plano:

Si hemos podido **ajustar bien ese plano**, podremos deducir cuál es la concentración de la muestra problema con facilidad

Pero si los datos son **colineales**, el plano **queda indeterminado**. No se podrá saber cuál es el modelo correcto: muchos serán compatibles

45

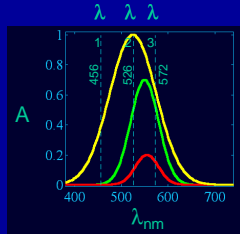
Estrategias para predecir mejor b y c

- **Reducir la información** (estrategia clásica): usar menos longitudes de onda, seleccionando las “mejores” para que \mathbf{P} sea una matriz de rango pleno
- **Comprimir la información**: usar “*variables latentes*” (PCR y PLS)
- Es posible **realizar una “pseudo-inversión**” de \mathbf{P} , usando sólo información relevante (*autovalores* significativos). Veremos después esta posibilidad, que permite usar los espectros completos.
- Otra estrategia es modificar la diagonal principal de \mathbf{P} (**regresión sesgada**)

Seleccionemos las mejores λ ...

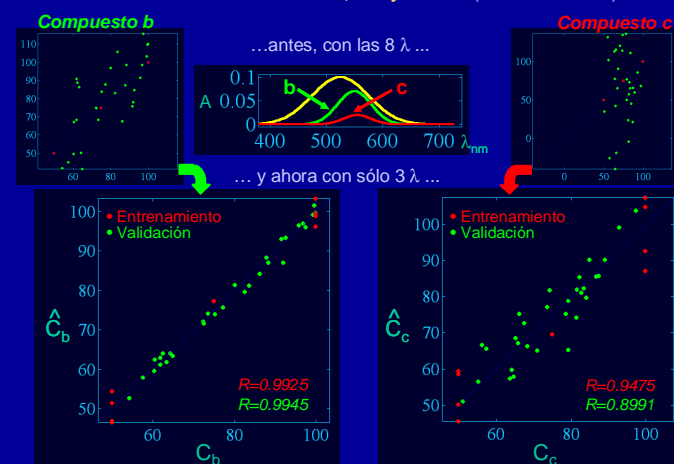
(= reducir la información)

- Para encontrar las mejores λ existen muchos métodos
- Algunos se encuentran implementados en paquetes estadísticos como SPSS o *Statistica*
- Aquí he usado un algoritmo genético para elegir las 3 longitudes de onda que conjuntamente ofrecen las mejores predicciones posibles de la concentración de **a**, **b** y **c**, a la vez



46

Utilizamos sólo las absorbancias a 456, 526 y 572 nm (A= matriz 8×4)



Mejora, pero estamos **desperdiciando información**. Lo ideal es usar **todo** el espectro

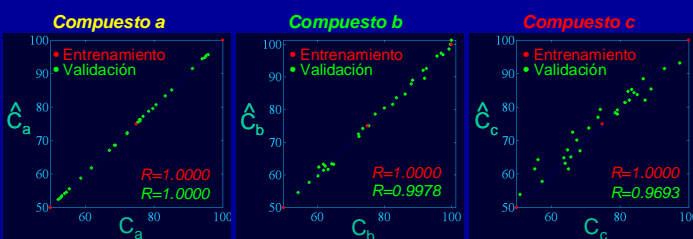
47

Otro modo de afrontar la colinealidad: “Pseudo-inversión” de A :

- Una matriz con dependencias lineales puede invertirse numéricamente a partir de una técnica denominada **descomposición en valores singulares**, emparentada con el método de regresión en componentes principales que después veremos
- También es posible despejar **B**, en lugar de $B = (A^T \cdot A)^{-1} \cdot A^T \cdot C_1$, del siguiente modo:

$$\mathbf{B}_i = \mathbf{A}^T \cdot (\mathbf{A} \cdot \mathbf{A}^T)^{-1} \cdot \mathbf{C}_i$$

- De cualquiera de estas formas, se puede usar espectros completos para predecir las concentraciones, y éstos son los resultados:



48

- En la segunda etapa predecimos las concentraciones :

Dividiendo el espectro de cualquier muestra **m** entre la matriz de espectros puros **B**,
obtendremos su composición **C_m** : Si **A_m=C_m · B**, entonces “ **C_m = A_m / B** ”

...cuya traducción matricial es: $C_m = A_m \cdot B^T \cdot (B \cdot B^T)^{-1}$

Aplicamos este tratamiento al ejemplo *con todo el espectro*. Predecimos **a**, **b** y **c** :



Con el enfoque multivariante de retroceso hemos conseguido mejoras muy notables

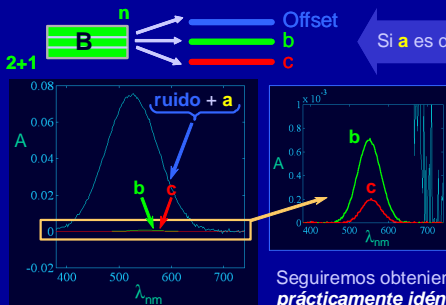
52

¿Qué sucedería si no supiéramos que había 3 sustancias presentes?

Imaginemos que *no supiésemos que existe a* (= el componente mayoritario)

$$\begin{matrix} & n \\ \boxed{A} & = & \boxed{C} \cdot \boxed{B} \\ m & & m \end{matrix}$$

Si **a** es desconocido



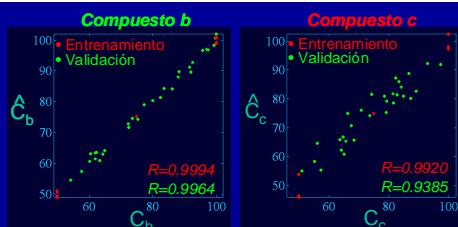
El término del blanco recoge en este caso dos contribuciones:

- El ruido
- El interferente (**a**)

Seguiremos obteniendo espectros de **b** y **c** *prácticamente idénticos* a los anteriores

- **El blanco nos revela que algo va mal:** hay una fuente de varianza desconocida
- **Obtendremos predicciones** de concentración de **b y c** en las muestras **correctas**
- **Naturalmente no podremos cuantificar a**, aunque detectaremos su presencia

53

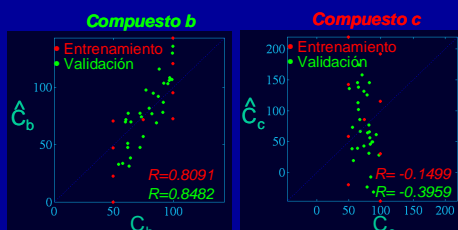


El hecho de ignorar la existencia de **a** apenas ha afectado al rendimiento

En **algunos libros** se ignora la necesidad de modelar la línea base + interferentes ...

...lo que se traduce en unos **resultados catastróficos** ...

De ahí los comentarios que a veces se oyen sobre el planteamiento de retroceso



54

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Sumario: Calibración por MLR de retroceso

Ventajas...

- Podemos utilizar **todo el espectro** : El uso de información redundante incrementa la seguridad de los análisis (*“promedio de réplicas”*)
- En consecuencia se obtienen **estimaciones de concentración más seguras** (más precisas y más exactas)
- Es **imprescindible** incluir los **términos** necesarios para poder **modelizar** posibles **sustancias desconocidas**

... e inconvenientes

- La **planificación de los patrones** debe ser **muy cuidadosa** :
 - Deben **representar** suficientemente el **dominio** de concentraciones
 - No** debe haber **dependencia lineal** en el diseño experimental de los estándares (=no puedan construirse unos por “mezcla” de otros)
- Se necesita una cantidad de **patrones igual o mayor que la cantidad de sustancias** presentes (inversión de **$C^T \cdot C$**)

55

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

MLR: Recapitulando...

- La mayor parte de los problemas radican en:
 - La presencia de **colinealidad**:
 - en los **espectros**: espectros demasiado similares
 - en los **patrones**: patrones mal diseñados
 - La presencia de **información redundante**: uso de longitudes de onda que aportan información casi idéntica
 - La **insuficiencia de información**: menos patrones que los necesarios para determinar los compuestos presentes
- El **enfoque de avance** es más atractivo y flexible, pero puede conducir a predicciones peores
- El **enfoque de retroceso** predice mejor al poder usar el espectro completo, pero es imprescindible un diseño de los patrones más cuidadoso

¿Sería posible usar un enfoque de avance con espectros completos, disponiendo sólo de una información parcial sobre la composición de las muestras y patrones? ...

Resultados de la validación

Univariante

$R_a = 0.9987$
 $R_b = 0.6818$
 $R_c = 0.2977$

MLR avance

(1) Las 8 λ
 $R_a = 0.9972$
 $R_b = 0.7346$
 $R_c = 0.0183$
(2) 3 λ óptimas
 $R_a = 0.9993$
 $R_b = 0.9943$
 $R_c = 0.8993$

MLR retroceso

(1) Con a, b y c
 $R_a = 0.9999$
 $R_b = 0.9978$
 $R_c = 0.9653$
(2) Sólo con b y c
 $R_b = 0.9964$
 $R_c = 0.9385$

56

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Validación de un modelo de regresión

(2)

- Pero además, la validación se usa también para seleccionar el **número óptimo de PCs** a incluir para que el modelo sea el mejor posible.
- En casos sencillos la naturaleza del problema permitirá **conocer a priori las causas significativas de variación** y por tanto el número de PCs, pero el modo más seguro de decidir cuántos componentes usar es validar
- En modelos blandos, la validación es una etapa **imprescindible e ineludible**, ya que se debe evitar a toda costa el riesgo de:
 - Sobreajuste (overfitting)**: Se ajusta ruido por incluir demasiados PCs
 - Infraajuste (underfitting)**: Los PCs del modelo son insuficientes para modelar correctamente todas las fuentes de varianza en los datos

57

19

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Validación de un modelo de regresión(3)

Test F de Malinowski

Si los autovalores corresponden a varianzas, se puede aplicar el test de Malinowski, que es un test de Fischer-Snedecor modificado:

$$F(r^*) = \frac{S_{ESTRUC}^2(r^*)}{S_{RESID}^2(r^*)} = \frac{\sum_{k=r^*+1}^r s_k^2}{s_{r^*}^2}$$

(α = 0.05)

r*	r-r*	λ²	V _r (r*)	F (r*)	F _{crit}
1	12	0.8606	0.0383	22.48	4.75
2	11	0.1534	0.0278	5.52	4.84
3	10	0.1074	0.0199	5.41	4.96
4	9	0.0464	0.0169	2.75	5.12
5	8	0.0403	0.0140	2.88	5.32
6	7	0.0329	0.0113	2.92	5.59
7	6	0.0231	0.0093	2.48	5.99
8	5	0.016	0.0080	2.01	6.61
9	4	0.0131	0.0067	1.96	7.71
10	3	0.0116	0.0050	2.30	10.13
11	2	0.0102	0.0025	4.16	18.51
12	1	0.004	0.0009	4.44	161.45
13	0	0.0009	(-)	(-)	(-)

Café

58

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Validación de un modelo de regresión(4)

Medidas más usuales de los errores de validación y calibración / predicción

- La validación implica definir medidas de error que describan la calidad de los modelos
- Dichas medidas de discrepancia entre predicciones y valores experimentales pueden hacerse...

(1) Como varianzas (o residuos al cuadrado)

$$PRESS = \sum_{i=1}^k (\hat{C}_i - C_i)^2$$

PRESS

Suma de los cuadrados de los residuos en la predicción

(2) Como desviaciones estándar

$$RMSEP = \sqrt{\frac{\sum_{i=1}^k (\hat{C}_i - C_i)^2}{k}}$$

RMSEP o SE y variantes (RMSECV, RMSEP, RMSEC)

Raiz cuadrada del residuo de predicción medio al cuadrado, es decir: RMSEP=(PRESS / n)^{1/2}

59

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Validación de un modelo de regresión(5)

Validación externa

- La validación externa bien ejecutada es la forma de validación más rigurosa (debe escogerse siempre que sea posible)
- Sin embargo, requiere duplicar el esfuerzo experimental, lo que la convierte en muchos casos en inviable

Validación por corrección de influencia

Se incrementa la importancia de los residuales correspondientes a puntos alejados del centro modelo:

$$PRESS \text{ corregido} = \sum_{i=1}^k \frac{1}{1 - f_i} (\hat{C}_i - C_i)^2$$

Problema: es demasiado optimista (los errores reales son mayores). Sin embargo, es útil en las etapas iniciales de construcción del modelo.

60

20

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Validación de un modelo de regresión (6)

Validación cruzada o crossvalidación

Valida las muestras con el mismo conjunto de entrenamiento con el que se ha construido el modelo (no requiere datos externos)

- Unas pocas muestras son separadas formándose el modelo con las demás
- Las muestras separadas son entonces predichas:
 - Si la selección es **sistemática**: El proceso se repite hasta que todas las muestras han sido predichas
 - Si la selección es **aleatoria**: Se repite la selección aleatoria una cierta cantidad de veces
- Se calculan los errores de predicción asociados al total de muestras excluidas
- Se repite el proceso para modelos progresivamente más completos (más PCs)

Ventajas e inconvenientes

- Minimiza los errores asociados a una división desigual de los datos y proporciona estimaciones del error predictivo real realistas
- Es un método laborioso que enlentece el proceso de construcción de los modelos

61

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Validación de un modelo de regresión (7)

Validación cruzada (objeto excluido)

(1) Construimos el modelo con un determinado número de componentes principales, utilizando **todas las muestras excepto una**

(2) Calculamos el error de predicción de la muestra que hemos dejado fuera del modelo

(3) Repetimos el proceso tantas veces como muestras hay, prediciéndolas todas

(4) Sumamos los errores y calculamos **RMSE** o **PRESS**

(5) Repetimos el proceso para cantidades crecientes de componentes principales, y representamos como en el caso anterior

Así, si examinamos un modelo de **k** PCs:

X

Y

Construir un modelo parcial
 $Y = F(X, k, j)$

Aplicar el modelo a
 $y_j = F(x_j, k, j)$

y_j experimental

\hat{y}_j calculado

residuo de j con k PCs

62

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Validación de un modelo de regresión (8)

(1) Validación completa (leave-one-out, o método del objeto excluido)

(2) Validación segmentada sistemática (venetian blind, o persiana veneciana)

(3) Validación segmentada aleatoria (random, o validación cruzada aleatoria)

(4) Intercambio de sets (test-set switch, o método del intercambio de conjuntos)

63

21

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Selección del número de componentes principales (9)

Modelo infraajustado | Modelo sobreajustado (interferentes, ruido)

RMSE, PRESS (etc)

MODELO OPTIMO
Número correcto de PCs

número de PCs

64

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Validación de un modelo de regresión (10)

Ejemplo 1: Determinación de agua en 55 muestras de trigo a partir de medidas espectrales en el infrarrojo cercano (método de regresión: PLS1)

- Validación correspondiente tras eliminar los outliers presentes (5), por el método del objeto excluido
- La **varianza de calibración** se hace siempre más pequeña al incluir más PCs
- La **varianza de validación** en cambio alcanza un mínimo nítido para 3 PCs

Varianza residual de **calibración**

Varianza residual de **validación**

65

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Validación de un modelo de regresión (11)

Ejemplo 2: Validación del ejemplo de las 9 muestras conteniendo mezclas de las sustancias a, b y c (método de regresión: PCR)

- La sustancia **a** podría ser descrita con únicamente 2 PCs
- Las sustancias **b** y **c** requieren 3 PCs

Validación externa

Validación cruzada

66

22

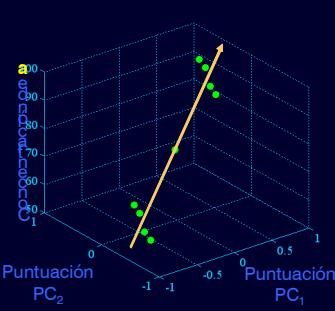
MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

3. Regresión de componentes principales (PCR)

Un modelo de PCR es muy parecido a uno de MLR de avance:

- En MLR relacionamos los espectros (**A**) con las concentraciones a predecir (**C**), de modo que el modelo es : **C** = **A** · **B**
- En PCR relacionamos la estructura latente de **A** con **C** : **C** = **T** · **B**



Volvamos al ejemplo de los 9 espectros

En PCR relacionamos en lugar de los espectros, **A**, su matriz de puntuaciones, **T**

De este modo, si en MLR el conjunto de parámetros se obtenía mediante:

$$\mathbf{B} = (\mathbf{A}^T \cdot \mathbf{A})^{-1} \cdot \mathbf{A}^T \cdot \mathbf{C}$$

... en PCR, la solución será ...

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

Como **T** es una matriz ortogonal, la inversión de **T^T · T** no será problemática

El único problema será decidir **cuántos componentes principales** debemos usar

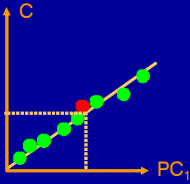
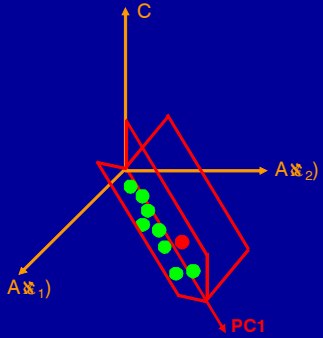
67

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Geoméricamente, la colinealidad ha dejado de ser un problema en PCR:

Para ilustrarlo, veamos de nuevo el ejemplo del ajuste a un plano



Al usar **1 PC**, el problema del ajuste a un plano se ha convertido en el **ajuste de una recta**

Hemos **reducido la dimensionalidad** del problema

Anteriormente nos era **imposible encontrar el mejor plano** de ajuste

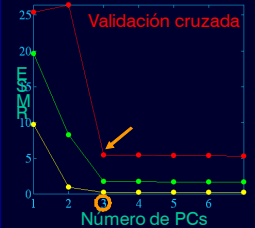

El que los datos sean colineales va a significar **más seguridad** en los resultados

68

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

En el ejemplo que usamos, tanto la validación externa como la cruzada nos indican que debemos usar **3 componentes principales**, tal como era de esperar



Nótese que el compuesto **a** puede ser descrito con **sólo 2 PCs**

Compuesto **c**
Compuesto **b**
Compuesto **a**

... de modo que el **modelo PCR** (**C** = **T** · **B**) en este caso **relacionará**:

- Las **3 primeras columnas** de la matriz **T** (= puntuaciones de **A**)
- La matriz de concentraciones, **C**

Lo correcto es construir **un modelo independiente para cada compuesto**

$$\mathbf{T} = \begin{bmatrix} -0.267 & -0.030 & 0.002 & 0.001 & 0.001 & 0.001 & 0.000 & 0.002 \\ 0.063 & -0.069 & 0.002 & 0.000 & 0.002 & -0.002 & 0.000 & -0.001 \\ -0.103 & 0.045 & 0.004 & 0.002 & -0.001 & 0.001 & 0.000 & -0.002 \\ 0.228 & 0.005 & 0.004 & -0.002 & -0.001 & 0.000 & -0.001 & 0.001 \\ -0.228 & -0.005 & -0.004 & 0.000 & -0.002 & -0.002 & -0.001 & 0.000 \\ 0.102 & -0.045 & -0.004 & 0.000 & 0.000 & 0.003 & 0.000 & -0.001 \\ -0.064 & 0.069 & -0.002 & -0.002 & 0.002 & 0.000 & 0.000 & 0.000 \\ 0.268 & 0.029 & -0.002 & 0.003 & 0.000 & -0.001 & 0.000 & 0.001 \\ 0.000 & 0.000 & 0.000 & 0.000 & -0.001 & 0.000 & 0.003 & 0.000 \end{bmatrix}$$

69


23

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Estos son los resultados de PCR mediante la regresión $B = (T^T \cdot T)^{-1} \cdot T^T \cdot C$:

Compuesto a



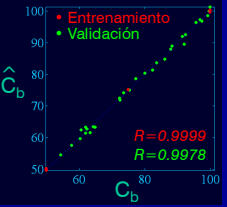
Entrenamiento

Validación

$R = 1.0000$

$R = 1.0000$

Compuesto b



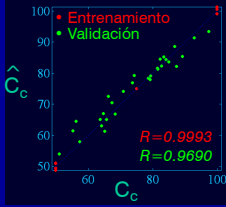
Entrenamiento

Validación

$R = 0.9999$

$R = 0.9978$

Compuesto c



Entrenamiento

Validación

$R = 0.9993$

$R = 0.9690$

Podemos compararlos con los obtenidos hasta el momento:

Univariante	MLR avance	MLR retroceso	...y PCR
$R_a = 0.9987$	(1) Las 8 + $R_a = 0.9972$	(1) Con a, b y c $R_a = 0.9999$	$R_a = 1.0000$
$R_b = 0.6818$	$R_b = 0.7346$	$R_b = 0.9978$	$R_b = 0.9978$
$R_c = 0.2977$	$R_c = 0.0183$	$R_c = 0.9653$	$R_c = 0.9690$
	(2) 3 + óptimas $R_a = 0.9993$	(2) Sólo con b y c $R_b = 0.9964$	<i>PCR ha sido el método que mejores resultados ha conseguido</i>
	$R_b = 0.9943$	$R_c = 0.9385$	
	$R_c = 0.8993$		

70

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Sumario: Calibración por PCR

Ventajas...

- Hemos convertido a la **colinealidad** de los datos en una **ventaja**
- Podemos separar las contribuciones** del ruido y otras fuentes de variación que no describen la propiedad de interés, de la parte relevante de **A**

... e inconvenientes

- Existen **dos etapas independientes**: la rotación propia de **A** (bloque **X**), y la construcción de la relación entre las puntuaciones y **C** (bloque **Y**)
- La rotación de **A** se hace **sin tener en cuenta a C**, de modo que ...
- ... **no hay garantía** de que la descomposición que hemos hecho **sea realmente la mejor** (= conduzca a lo que realmente queremos)
 - Pueden permanecer **contribuciones irrelevantes** en los primeros PCs
 - Se pueden **perder contribuciones** muy significativas en PCs elevados, que serían excluidos por su pequeña magnitud (e.g., *un banda característica muy útil para describir Y será ignorada si es muy pequeña*)

Aunque a primera vista lo parecía, PCR no es la solución ideal. Hay un método mejor...

71

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

3. Regresión de Mínimos Cuadrados Parciales (PLS) (1)

- Con PCR hemos convertido a la **colinealidad** de los datos en una **ventaja**
- Podemos separar las contribuciones** del ruido y otras fuentes de variación que no describen la propiedad de interés, de la parte relevante de **A**

¿Es PCR el “método perfecto” para calibración multivariante?

- Existen **dos etapas independientes**: la rotación propia de **A** (bloque **X**), y la construcción de la relación entre las puntuaciones y **C** (bloque **Y**)
- La rotación de **A** se hace **sin tener en cuenta a C**, de modo que ...
- ... **no hay garantía** de que la descomposición que hemos hecho **sea realmente la mejor** (= conduzca a lo que realmente queremos)
 - Pueden permanecer **contribuciones irrelevantes** en los primeros PCs
 - Se pueden **perder contribuciones** muy significativas en PCs elevados, que serían excluidos por su pequeña magnitud (e.g., *un banda característica muy útil para describir Y será ignorada si es muy pequeña*)

CONCLUSION: Aunque a primera vista lo parecía, PCR no es la solución ideal

72

24

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Minimos cuadrados parciales(5)

- El efecto de **maximizar la correlación X-Y se puede imitar también en PCR** ordenando las direcciones extraídas en PCR no en cuando a varianza decreciente de X sino por las correlaciones que ofrece cada **t** incluido en **T** con **Y**
- PLS usa otra solución:

Buscar las **direcciones de compromiso** que, maximizando la varianza explicada en los bloques X e Y, mejoran todo lo posible la correlación entre la variable latente considerada y la respuesta
- En PLS existen **dos tipos de cargas para el bloque X**:
 - Los pesos de las cargas, **W (loading weights)**, que describen la asociación de las variables del bloque **X** con **Y**
 - Las cargas ordinarias, **P (loadings)**, que describen las direcciones de mayor variación en **X**, exactamente igual que en PCA (aunque están sesgadas)
- Naturalmente, las **puntuaciones y cargas** de PLS no coinciden con las de PCR

76

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Minimos cuadrados parciales(6)

Dos tipos de PLS son posibles...

- PLS2** (**Y es una matriz**) se realiza la **rotación propia interconectada** tanto de **X** como de **Y**. Se obtienen cargas y puntuaciones de ambas matrices: **T / U** y **P / Q**
- PLS1** (**Y es un vector**) **Y no es descompuesta**, pero sí utilizada para descomponer **X**. Se obtienen cargas y puntuaciones de **X** : Las puntuaciones de **Y** son **U = Y**

- Se puede aplicar **PLS 1** a **matrices o vectores Y** ; **PLS 2** sólo a **matrices Y**
- Con **PLS 1** construimos **modelos individuales** para cada propiedad a describir
- Con **PLS 2** construimos un **modelo global** que describe **todas** las variables **Y**

Así pues, tenemos varias posibilidades

Atendiendo a la naturaleza de Y

Si Y es un vector.....

Si Y es una matriz se puede tratar

Individualmente cada variable Y.....

Colectivamente todas las variables Y.....

PLS 1

PLS 2

➡ En el ejemplo de los nueve mezclas, **Y** es una matriz : **PLS 1** y **PLS 2** son válidos

77

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Minimos cuadrados parciales(7)

- PLS 1** es el método de regresión más preciso
- PLS 2** tiene **ventajas exploratorias** muy interesantes, ya que permite detectar asociaciones entre las variables **X**, **Y** y relaciones entre **X** e **Y**
 - Diagramas dobles (biplots) de **T** y **P**
 - Diagramas dobles (biplots) de **Q** y **W**
- Cuando se sospecha la existencia de **correlación entre las variables Y**, **PLS2** proporciona modelos globales casi tan buenos como **PLS1**
- Cuando hay **una sola variable Y** o queremos simplemente **máxima precisión**, el modelo más adecuado es **PLS1**
- Tanto **PLS1**, como **PLS2**, como **PCR**, proporcionan **gran cantidad de información exploratoria** (son tan ricos o más que PCA)

78

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Estos son los resultados de PCR mediante la regresión $B = (T^T \cdot T)^{-1} \cdot T^T \cdot C$:

Compuesto a

Compuesto b

Compuesto c

Podemos compararlos con los obtenidos hasta el momento:

Univariante	MLR avance	MLR retroceso	...y PCR
$R_a = 0.9987$	(1) Las 8 λ	(1) Con a, b y c	$R_a = 1.0000$
$R_b = 0.6818$	$R_a = 0.9972$	$R_a = 0.9999$	$R_b = 0.9978$
$R_c = 0.2977$	$R_b = 0.7346$	$R_b = 0.9978$	$R_c = 0.9690$
	$R_c = 0.0183$	$R_c = 0.9653$	
	(2) 3 λ óptimas	(2) Sólo con b y c	
	$R_a = 0.9993$	$R_b = 0.9964$	
	$R_b = 0.9943$	$R_c = 0.9385$	
	$R_c = 0.8993$		

79

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

PLS 1 ofrece los mejores resultados de entre todos los métodos ensayados

Compuesto a

Compuesto b

Compuesto c

PLS 2 ofrece unos resultados casi idénticos a PLS 1 y equivalentes a PCR

Compuesto a

Compuesto b

Compuesto c

80

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Sumario: Calibración por PCR y PLS

PCR: Ventajas

- Hemos convertido a la **colinealidad** de los datos en una **ventaja**
- Podemos separar las contribuciones** del ruido y otras fuentes de variación que no describen la propiedad de interés, de la parte relevante de **A**

PCR: inconvenientes

- Existen **dos etapas independientes**: la rotación propia de **A** (bloque **X**), y la construcción de la relación entre las puntuaciones y **C** (bloque **Y**)
- La rotación de **A** se hace **sin tener en cuenta a C**, de modo que ...
- ... **no hay garantía** de que la descomposición que hemos hecho **sea realmente la mejor** (= conduzca a lo que realmente queremos)
 - Pueden permanecer **contribuciones irrelevantes** en los primeros PCs
 - Se pueden **perder contribuciones** muy significativas en PCs elevados, que serían excluidos por su pequeña magnitud (*e.g.*, *un banda característica muy útil para describir Y* será *ignorada si es muy pequeña*)

CONCLUSION: Aunque a primera vista lo parecía, **PCR no es la solución ideal**

81

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

Sumario: Calibración por PCR y PLS

PLS: Ventajas

- El proceso de obtención de **B** tiene lugar en **una sola etapa global**
- Conduce a resultados equivalentes a los obtenidos con PCR, utilizando un **menor número de componentes principales**
- Es más interpretable que PCR e **igualmente robusto**
- Enfoca su atención sólo en la parte de la **estructura de X** que es **relevante** para predecir **Y**
- PLS1** da las predicciones más perfectas, mientras que **PLS2** es más útil para screening, o cuando interesa conseguir una relación global

PLS: inconvenientes

- Desde un punto de vista estadístico, PLS resulta **oscuro** (no es posible determinar rigurosamente la significatividad del modelo)
- Mayor **lentitud** de cálculo, especialmente en la validación cruzada por el método del objeto excluido
- PLS, al igual que PCR o MLR, es un **modelo lineal**

82

MASTER EN TÉCNICAS EXPERIMENTALES EN QUÍMICA POR LA UNIVERSITAT DE VALÈNCIA

MODULO II. Asignatura: LABORATORIO DE Calibración y tratamiento de datos
PROFESOR/ES: José Ramón Torres. DEPARTAMENTO: Química Analítica

A modo de resumen final

- La **Regresión Univariante** es incapaz de revelar la presencia de interferencias
- La **Regresión Multivariante** puede revelar la presencia de interferencias
 - En el **enfoque de avance** cada término del modelo **B** es un **peso** o **influencia**
 - En el **enfoque de retroceso** cada término del modelo **B** es una **sensibilidad**
- MLR** usa la **totalidad de información** disponible para establecer la mejor correlación entre X e Y
 - MLR de avance** permite trabajar con **datos desestructurados** (más flexible)
 - MLR de retroceso** requiere un **diseño muy cuidadoso de los patrones**, pero puede conducir a muy buenas predicciones
- PCR** usa **sólo la información relevante en X** pero ignora en el proceso de obtención si esta relevancia se relaciona o no con Y
- PLS** usa la **la información relevante en X en cuanto a la predicción de Y**
 - PLS1** relaciona **individualmente** X con cada una de las respuestas
 - PLS2** relaciona **colectivamente** X con todas las respuestas

83
