

# 4

## QUIMIOMETRÍA EN ESPECTROSCOPIA NIR

---

### 4.1. INTRODUCCIÓN

Un método de análisis instrumental proporciona un conjunto de datos, más o menos complejo, al que se denomina señal analítica. Dicha señal es una magnitud física que incorpora información química y/o física de la muestra analizada. Para obtener dicha información, se debe correlacionar la señal obtenida con la propiedad analítica que se pretende medir. Este proceso no siempre es directo, ya que la señal obtenida suele ser contribución de más de una especie o propiedad. En métodos cromatográficos, hay una separación previa de los componentes de la muestra, por lo que cada señal analítica obtenida corresponde a analitos discretos, no así en métodos espectroscópicos, donde la señal proporcionada por el instrumento no es exclusiva de un solo componente de la muestra.

Los avances instrumentales, la automatización y la incorporación de los ordenadores en el control y adquisición de señales de instrumentos, permite obtener gran cantidad de información analítica en tiempos muy cortos. Poder extraer la información útil de la que no lo es, y ser capaz de interpretar los datos para que puedan ser utilizados y relacionados con el parámetro a determinar, se convierte en una tarea compleja dado el gran volumen de información. Esta problemática ha propiciado el desarrollo de métodos quimiométricos, basados en cálculos matemáticos, estadísticos y de lógica formal, que permiten diseñar o seleccionar procedimientos de medida óptimos y obtener la máxima información relevante de los datos analíticos<sup>1</sup>. La quimiometría se ha convertido así en una parte muy importante de la química analítica y su uso ha ido en aumento aplicándose al tratamiento de todo tipo de datos espectroscópicos, cromatográficos, polarográficos, etc.

La quimiometría posee gran cantidad de aplicaciones y de muy diversa naturaleza. En concreto, en los trabajos desarrollados en esta memoria se ha aplicado la quimiometría a la

resolución de diversas problemáticas que aparecen en el desarrollo de métodos de análisis mediante espectroscopia NIR, de las cuales, las más comunes son:

- Filtrado de señales: los espectros NIR pueden incorporar información no relevante. La aplicación de determinados métodos quimiométricos permiten mejorar la calidad de la señal instrumental, eliminando o reduciendo el ruido.
- Métodos de reconocimiento de pautas: debido a las características del espectro NIR, no puede utilizarse directamente para identificación visual de compuestos o grupos funcionales. Los métodos de reconocimiento de pautas permiten establecer relaciones entre muestras para agruparlas según la similitud de sus espectros o crear agrupaciones de muestras capaces de clasificar nuevas muestras como pertenecientes a un grupo determinado de una forma objetiva.
- Métodos de calibración multivariable: de la misma forma, los espectros NIR requieren de métodos de calibración multivariable que hagan posible relacionar la señal analítica con la propiedad a determinar. Estos métodos son muy diversos y utilizados actualmente.

## **4.2. ETAPAS DEL PROCESO DE MODELADO**

El proceso de modelado de un conjunto de datos, ya sea para clasificación o bien para determinación cuantitativa de un parámetro determinado, tiene como finalidad obtener un modelo capaz de predecir propiedades de nuevas muestras. Para la obtención de un modelo robusto se deben seguir las siguientes etapas: selección de un conjunto de calibración, determinación de la propiedad o identidad por métodos de referencia adecuados, obtención de la señal analítica, cálculo del modelo, validación del modelo, aplicación del modelo al análisis de muestras desconocidas, análisis de rutina y monitorización y por último transferencia de modelos. A continuación se explica más detalladamente cada una de las etapas implicadas en el modelado de datos.

### **4.2.1. Selección del conjunto de calibración**

Se debe seleccionar un conjunto limitado de muestras que debe ser representativo de toda la variabilidad química y física que pueda darse durante el análisis de rutina, con la finalidad de obtener buena capacidad predictiva. El conjunto de calibración (o training)

deberá incorporar variabilidad de distinta naturaleza según la finalidad del modelo. Así, para una calibración multivariable, las muestras deberán ser representativas de todo el intervalo del parámetro a determinar (p.ej: cubrir todo el intervalo de concentraciones en una determinación de un principio activo), mientras que en modelos de clasificación las muestras deben incorporar toda la posible variabilidad que pueda darse para cada tipo de muestras (proveedor, tamaño de partícula, grados de pureza...).

#### **4.2.2. Métodos de referencia**

Se determinan las concentraciones o propiedades de las muestras o bien las identidades de las mismas mediante métodos de referencia adecuados. Estos métodos de referencia deben proporcionar valores precisos y exactos, ya que de ello dependerá la exactitud del modelo multivariable obtenido.

#### **4.2.3. Obtención de la señal analítica**

Para obtener la señal analítica se debe someter a las muestras al procedimiento instrumental que se haya elegido. Si se trata de técnicas espectroscópicas se deben registrar los espectros de las muestras. En el caso de métodos cromatográficos, se introducen las muestras en el sistema cromatográfico para la obtención de los datos. Una vez el sistema instrumental proporciona los datos resultantes, éstos son almacenados y quedan disponibles para su posterior tratamiento.

#### **4.2.4. Cálculo del modelo**

La construcción del modelo implica una serie de tratamientos previos de la señal para a continuación hallar la relación más simple entre la señal analítica y la propiedad a determinar, ya sea estableciendo la relación con la concentración de un analito o con parámetros físicos de la muestra, o bien estableciendo las características que definen una clase y fijando las fronteras que la separan de otra u otras clases.

Los pretratamientos de la señal reducen las contribuciones del ruido. El procesado previo de la señal analítica permite corregir el efecto de los componentes ajenos a la información buscada, mejorando así la exactitud y precisión de los resultados. Ejemplos típicos de tratamientos de señales son la utilización de filtros de transformada de Fourier, correcciones de línea base... En espectroscopia, sobretodo en la zona del infrarrojo cercano, estos pretratamientos son necesarios para corregir además otros efectos indeseados en la

señal. Debido a su importancia, los pretratamientos espectrales habituales en NIR se discutirán detalladamente más adelante.

Una vez corregidos los efectos del ruido, el modelo puede ser construido teniendo en cuenta las bases teóricas que explican la relación entre la magnitud física de la señal analítica con la propiedad a medir (Lambert-Beer en absorción, ecuación de Nernst en potenciometría...) o bien por relaciones totalmente empíricas. En calibración multivariable para el cálculo del modelo se utilizan gran variedad de algoritmos matemáticos, así como técnicas estadísticas para evaluar la calidad del mismo y optimizarlo.

#### **4.2.5. Validación del modelo**

Se trata de aplicar el modelo calculado a un número limitado de muestras de las cuales se conoce la propiedad que el modelo predice. Estas muestras son externas al conjunto de calibración. Los resultados obtenidos mediante el modelo se comparan estadísticamente con los valores de referencia, comprobando que ambos resultados no son distintos, es decir que el modelo predice correctamente y por lo tanto es válido para la determinación de la propiedad.

#### **4.2.6. Aplicación del modelo**

Una vez validado el modelo, éste se aplica a la predicción de muestras desconocidas. Los resultados obtenidos para estas muestras son generalmente aceptables, por lo que esta fase es tan sólo una segunda comprobación de la capacidad predictiva del modelo.

#### **4.2.7. Análisis de rutina y Monitorización**

Una vez establecida la capacidad predictiva del modelo, se debe realizar un seguimiento del mismo para comprobar que continúa proporcionando resultados exactos y precisos. Del mismo modo, se debe realizar un seguimiento del instrumento para detectar cualquier posible deterioro del mismo.

#### **4.2.8. Transferencia del modelo**

Una última etapa del proceso de modelado puede conllevar la transferencia de ese modelo a otro instrumento en el que se deban predecir muestras en control de rutina. Para que un modelo prediga de forma correcta muestras desconocidas en el nuevo instrumento, en ocasiones será necesario realizar una serie de correcciones del modelo, comprobando estadísticamente que una vez transferido proporciona buenos resultados. Este último aspecto

del proceso de modelado está siendo objeto de discusión en numerosas publicaciones, existiendo distintas opciones para transferir modelos de un instrumento a otros.

### **4.3. PRETRATAMIENTOS ESPECTRALES**

Los componentes o efectos no deseados presentes en las señales obtenidas en espectroscopia se denominan comúnmente ruido. Este ruido puede tener diferentes causas u orígenes, por lo que puede afectar de forma distinta al espectro. Puede haber ruido debido a los componentes de la instrumentación utilizada para el registro del espectro (ruido instrumental), debido a variaciones de temperatura, humedad u otras condiciones ambientales durante el registro (ruido ambiental) o bien variaciones en la señal debidas a la propia naturaleza de la muestra. Para corregir o reducir este ruido de los datos espectrales es necesario un pretratamiento espectral.

A continuación se comentan los tratamientos espectrales más habituales en espectroscopia, tanto en la región NIR como en el resto del espectro, aplicados con el fin de reducir o eliminar la contribución del ruido.

#### **4.3.1. Promediado de espectros**

El ruido de alta frecuencia es de naturaleza aleatoria, por lo que un promediado de varias señales de una misma muestra reduce la contribución del ruido aumentando la relación señal/ruido. Es un procedimiento muy usual en espectroscopia, siendo el espectro final el resultado de promediar un número determinado de espectros individuales.

#### **4.3.2. Suavizado espectral**

El suavizado espectral se aplica en aquellos casos en los que el promediado de espectros no es suficiente ya que la relación señal/ ruido es pequeña. En estos casos existen algunos algoritmos matemáticos que, aplicados al espectro reducen el ruido suavizando la señal. Los métodos de suavizado más habituales son los basados en filtros de Savitzky-Golay<sup>2</sup> y en transformadas de Fourier<sup>3</sup>.

#### **4.3.3. Corrección de línea base**

Es un tipo de pretratamiento que intenta corregir determinadas tendencias en la línea base que aportan ruido a la señal. Existen varios tipos de corrección de línea base según el efecto que se desea corregir. Un tipo de corrección es el que modela la línea base como una

función simple de longitud de onda y sustrae esta función a todos los datos espectrales. Este tipo de corrección es muy frecuente en MIR pero menos utilizada en NIR.

Cuando se trabaja en NIR con Transformada de Fourier, es frecuente utilizar una corrección de la línea base llamada “Tilting”<sup>14</sup> que consiste en corregir el espectro, eliminando la curvatura de manera que se cumplan los requisitos de periodicidad para la transformación.

Un último tipo de corrección de línea base es el llamado Ajuste de Línea Base (o De-Trending) que ajusta el espectro original a una función cuadrática, sustrayéndola posteriormente del espectro original, eliminando la característica curvatura ascendente a medida que aumenta la longitud de onda. De esta forma, el conjunto de espectros corregidos tiene como media el valor 0 y una varianza diferente de cero. Este pretratamiento se aplica a cada espectro individual, por lo tanto no depende de ningún espectro de referencia.

#### **4.3.4. Derivación**

Es uno de los pretratamientos más utilizados en espectroscopia para minimizar desviaciones de línea base causados por efectos de dispersión. En espectroscopia NIR se utiliza para disminuir los problemas más característicos: solapamiento de bandas y variaciones de línea base. La utilización de la primera derivada elimina los términos constantes a todas las longitudes de onda, es decir, desplazamientos de línea base, mientras que la segunda derivada corrige además las desviaciones causadas por los términos que varían linealmente con la longitud de onda. Generalmente no se utilizan derivadas de orden superior, puesto que la primera y segunda derivada suelen ser suficientes. La aplicación de las derivadas permite un aumento de la resolución de bandas, pero por contra aumenta el ruido. Para efectuar el cálculo de derivación los dos métodos más utilizados son el método de ventana móvil (*Moving Averaged Segment Convolution*) propuesto por Norris y Williams<sup>4</sup> y el propuesto por Savitzky y Golay<sup>2</sup>.

#### **4.3.5. Corrección del efecto multiplicativo de la dispersión**

Denominado habitualmente MSC (*Multiplicative Scatter Correction*), este método, propuesto por Geladi y colaboradores<sup>5</sup>, tiene la finalidad de corregir desplazamientos espectrales debidos al diferente tamaño de partícula de muestras. Es por lo tanto un método muy utilizado en NIR, ya que este efecto se presenta en medidas de reflectancia. La corrección se lleva a cabo aplicando a cada espectro original los parámetros de la recta obtenida por mínimos cuadrados representando los valores de absorbancia de un intervalo de

longitudes de onda en el que el analito no absorbe frente a los valores de absorbancia de un espectro de referencia (normalmente el espectro medio del conjunto a tratar). Su principal inconveniente es que si el conjunto de espectros originales es modificado, el espectro de referencia cambia, por lo que los parámetros de la recta obtenida deben recalcularse de nuevo.

#### **4.3.6. Variable normal estándar**

La variable normal estándar o SNV es otro pretratamiento utilizado en la corrección de los efectos de tamaño de partícula en espectroscopia NIR<sup>6</sup>. A diferencia de MSC, SNV se aplica individualmente a cada espectro, sin depender de ningún espectro de referencia. De esta forma se obtiene una escala común a todos los espectros. El procedimiento de corrección consiste en centrar cada espectro, restando el valor medio de absorbancia del espectro a cada uno de los valores de absorbancia a cada longitud de onda, dividiendo el resultado por la desviación estándar del espectro. Se ha demostrado que existe una relación lineal entre MSC y SNV<sup>7</sup>, por lo que, aplicados correctamente, proporcionan resultados equivalentes.

Hasta aquí se han revisado los pretratamientos espectrales más comunes en NIR. En ocasiones resulta indicado utilizar más de un método de pretratamiento espectral para mejorar la calidad de los resultados. Así, es frecuente encontrar combinaciones de SNV-DT o DT-SNV, y combinaciones de SNV-derivadas que permiten eliminar en mayor medida el ruido espectral. Sin embargo, esto puede conllevar también una eliminación de información útil, por lo que en cada caso se deberá evaluar qué tratamiento es mejor o bien si es necesario aplicar más de uno para la corrección de los datos.

#### **4.4. REDUCCIÓN DE VARIABLES-PCA**

Debido al gran volumen de información que proporcionan los métodos instrumentales, al que se ha hecho referencia al inicio del capítulo, uno de los campos de estudio de la quimiometría ha sido el desarrollo de métodos de cálculo capaces de reducir este gran volumen de información para que ésta quede contenida, sin pérdidas relevantes, en un número reducido de variables. Uno de los métodos más utilizados para reducción de variables, es la descomposición de los datos en componentes principales (PCA, *Principal Component Analysis*)<sup>8, 9</sup>. La mayoría de las técnicas quimiométricas utilizadas en esta

memoria se basan en un análisis en componentes principales previo al modelado de los datos, por lo que a continuación se explicarán los fundamentos de este tratamiento.

#### 4.4.1. Tratamiento previo de los datos

Los procedimientos de reducción de variables no suelen ser aplicados a los datos originales, sino que estos son previamente tratados para eliminar posibles efectos que puedan afectar a la descomposición. Los dos tratamientos habituales son el centrado y el autoescalado, cuyos efectos han sido discutidos en la bibliografía<sup>10, 11</sup>.

Considerando una matriz  $\mathbf{X}$  de datos donde cada fila corresponde al espectro de una muestra y cada columna corresponde a una longitud de onda (variable), el centrado y el autoescalado transforman esta matriz de la siguiente forma:

- Centrado: se calcula el valor medio de cada variable ( $\bar{x}_k$ ) del conjunto de calibración (de cada columna de la matriz), y se resta este valor a cada punto ( $x_{ik}$ ) de la columna.

$$x_{ik}^{\text{centrado}} = x_{ik} - \bar{x}_k \quad (4.1)$$

El valor medio corresponde al centro del modelo, y los valores de todas las variables están ahora referidos a dicho centro. Este tratamiento permite seguir manteniendo las unidades originales.

- Autoescalado: consiste en, después de centrar cada columna, dividir el resultado por la desviación estándar de la misma,  $s_k$ , de esta forma la varianza de cada variable vale la unidad.

$$x_{ik}^{\text{autoescalado}} = \frac{x_{ik} - \bar{x}_k}{s_k} \quad (4.2)$$

Geoméricamente es equivalente a cambiar la longitud de los ejes de coordenadas; todos los ejes tienen la misma longitud y cada variable tiene la misma influencia en el cálculo. Si los datos son espectros de muestras puede ser más interesante no escalarlo, ya que al escalar se daría igual importancia a las variables con baja absorbancia, o con principalmente solo ruido, que a los máximos de absorción. Por otra parte, el escalar puede favorecer la determinación de compuestos minoritarios en presencia de absorbentes

mayoritarios. Para cada problema en concreto hay que adoptar la solución más conveniente.

#### 4.4.2. Análisis en componentes principales

El espectro de una muestra registrado a  $k$  longitudes de onda puede describirse como un vector con  $k$  coeficientes. Se puede construir un espacio de  $k$  dimensiones de forma que cada una de ellas sea la señal a cada una de las longitudes de onda, y se puede representar la muestra como un punto en este espacio. Si se tienen  $m$  muestras, cada una de ellas se puede representar como un punto en el espacio de  $k$  dimensiones. Si estas muestras no tienen nada en común aparecerán dispersadas en el espacio. Si por el contrario tienen algo en común o están relacionadas los  $m$  puntos aparecerán agrupados.

El objetivo del PCA es hallar las direcciones que explican la máxima variabilidad de las muestras y utilizarlas como nuevos ejes de coordenadas, denominados componentes principales (PCs). De esta forma, se reduce la dimensionalidad de un espacio de  $k$  dimensiones a un espacio de  $a$  dimensiones ( $a < k$ ), manteniendo intacta la información relevante del sistema. Geométricamente, un PCA es un cambio de ejes, representando las muestras en un nuevo sistema de coordenadas con un número inferior de ejes al utilizado inicialmente. El primer componente principal es combinación lineal de las  $k$  variables que explica la máxima variabilidad de las muestras. El segundo PC se escoge de forma que sea ortogonal al primero y que explique la máxima variabilidad de las muestras una vez restada la explicada por el primer PC. Para definir matemáticamente estos nuevos ejes se utilizan los *loadings* que son los cosenos de los ángulos que forman los nuevos ejes con los originales. Los *scores* son las coordenadas de las muestras en estos nuevos ejes.

Numéricamente, la matriz de datos  $\mathbf{X}$  (datos espectrales) se descompone en el producto de dos matrices: una matriz de scores  $\mathbf{T}$  y una matriz de loadings  $\mathbf{P}$ , quedando un residual representado por la matriz  $\mathbf{E}$ :

$$\mathbf{X} = \mathbf{TP}^t + \mathbf{E} \quad (4.3)$$

El PCA tiene la propiedad de que la matriz de loadings  $\mathbf{P}$  puede encontrarse mediante un ajuste por mínimos cuadrados de  $\mathbf{X}$  en la matriz de scores  $\mathbf{T}$ , del mismo modo que  $\mathbf{T}$  se puede encontrar por un ajuste de mínimos cuadrados de  $\mathbf{X}$  en  $\mathbf{P}$ . Esto ha llevado al desarrollo de algoritmos de cálculo para la obtención de  $\mathbf{T}$  y  $\mathbf{P}$ . El más conocido es el algoritmo

NIPALS <sup>12</sup> (*Nonlinear Iterative Partial Least Squares*), que permite obtener sólo los primeros PCs sin necesidad de calcular todos los vectores propios de una matriz.

Cada componente principal contiene información de diferente relevancia. Los primeros PCs describen la fuente de variación más importante de los datos. La representación completa de la matriz  $\mathbf{X}$  implica  $\mathbf{k}$  vectores de loadings y scores. Sin embargo, puesto que la finalidad de la descomposición es la reducción de la dimensionalidad del sistema, se representa la matriz original con un número menor de vectores  $\mathbf{a}$  de la forma:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^t + \mathbf{t}_2\mathbf{p}_2^t + \dots + \mathbf{t}_a\mathbf{p}_a^t + \mathbf{E} \quad (4.4)$$

El conjunto de datos  $\mathbf{X}$ , descrito inicialmente por variables correlacionadas, queda de esta forma definido por un nuevo conjunto de variables (PCs) no correlacionadas entre si en un nuevo sistema de ejes ortogonales.

#### **4.5. TÉCNICAS QUIMIOMÉTRICAS EN ANÁLISIS CUALITATIVO**

Las técnicas quimiométricas han sido ampliamente aplicadas a la resolución de distintos problemas de clasificación de sustancias mediante una serie de características o magnitudes físicas de las muestras medidas o determinadas previamente. Las técnicas quimiométricas utilizadas en problemas de análisis cualitativo se conocen de forma general con el nombre de Métodos de Reconocimiento de Pautas (*PRM, Pattern Recognition Methods*).

##### **4.5.1. Métodos de Reconocimiento de Pautas**

Los métodos de reconocimiento de pautas son un conjunto de herramientas quimiométricas que permiten establecer agrupaciones de muestras en función de características comunes o relaciones que existan entre ellas o bien definir métodos de clasificación para muestras desconocidas. Existen gran variedad de métodos de reconocimiento de pautas y continuamente aparecen nuevas variantes de los ya existentes. La mayoría de métodos de reconocimiento de pautas se basan en la medida de la similitud, parámetro que indica en qué medida un objeto es igual a otro. La manera más común de expresar la similitud es a través de medidas de correlación o distancias:

-Medidas de correlación: se basan en el cálculo del coeficiente de correlación entre dos muestras:

$$\tilde{n}_{jk} = \frac{\sum_{i=1}^p (\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)(\mathbf{x}_{ik} - \bar{\mathbf{x}}_k)}{\sqrt{\sum_{i=1}^p (\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)^2 \sum_{i=1}^p (\mathbf{x}_{ik} - \bar{\mathbf{x}}_k)^2}} \quad (4.5)$$

Oscila de  $-1$  a  $1$ . El valor de  $1$  indica coincidencia total entre los dos espectros<sup>13</sup>.

-Medidas de distancia: estas medidas se basan en el cálculo de una distancia que representa cuán diferente es una muestra de otra o bien de un punto en el espacio que represente el modelo de una clase. El cálculo de la distancia  $D$  entre una muestra  $x_i$  y el centroide de una clase  $\mu$  se determina:

$$D^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' \Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) \quad (4.6)$$

siendo  $\Sigma$  la matriz de dispersión de los datos alrededor del centroide. Dependiendo del valor de  $\Sigma^{-1}$  adopte esta matriz de dispersión tendremos diferentes tipos de distancias<sup>14</sup>:

- Cuando  $\Sigma^{-1} = I$  la distancia se conoce con el nombre de distancia Euclídea. En este tipo de distancia todas las clases tienen forma esférica (hiperesferas). Sólo puede aplicarse cuando las variables no están correlacionadas (p.ej.: habiendo efectuado un análisis en componentes principales y calculado las distancias en el espacio de los componentes principales).
- Cuando  $\Sigma^{-1}$  es la inversa de la matriz de varianza-covarianza, la distancia es conocida como distancia de Mahalanobis. Con estas distancias, la forma de una clase es elipsoidal. Dependiendo de si se aplica la misma matriz de dispersión para todas las clases o una distinta para cada clase, estos elipsoides tendrán formas, tamaños y orientaciones iguales o diferentes. En concreto, podemos diferenciar entre tres maneras de calcular la distancia de Mahalanobis:
  1.  $\Sigma^{-1}$  es la matriz de varianza-covarianza aplicada a todo el data set. Esto hace que todas las clases tengan el mismo elipsoide (tamaño, forma y orientación).
  2.  $\Sigma^{-1}$  es la matriz de varianza-covarianza aplicada a todo el data set, pero

normalizando las distancias por el tamaño de cada clase. Los elipsoides de las clases poseen la misma forma y orientación, pero su tamaño varía para cada clase.

3.  $\sigma^{-1}$  se calcula independientemente para cada clase. Cada clase posee su propio elipsoide, con forma, orientación y tamaño distintos.

Los métodos de reconocimiento de pautas pueden clasificarse según se conozca a priori o no, la pertenencia de los objetos a clases determinadas siendo denominados respectivamente métodos supervisados y métodos no supervisados.

#### 4.5.1.1. Métodos no supervisados

Se basan en descubrir agrupaciones de pautas en el espacio de N-dimensiones sin saber a priori a qué clase pertenece cada muestra. Algunos de los más comunes son:

- Análisis de clusters: bajo esta denominación quedan englobados toda una serie de métodos diseñados para entender la estructura de una gran matriz de datos, reconociendo similitudes entre objetos (o variables), y así llegar a distinguir algunas clases, que serán conjuntos de objetos similares<sup>15</sup>.
- Minimal Spanning Tree (MST): se basa en conectar puntos (objetos) de forma que la longitud total es la mínima de todas las combinaciones posibles. El algoritmo busca de todas las distancias entre objetos la máxima y divide los objetos conectados en dos clusters. Sigue haciendo lo mismo en cada uno de los nuevos clusters, hasta cumplir una condición impuesta previamente<sup>16</sup>.
- Redes neuronales no supervisadas (Kohonen): este tipo de red halla la neurona que se parece más a un objeto presentado a ella y modifica sus pesos para que se parezca más al ejemplo presentado. Después de un número determinado de entradas de los datos a la red, diversas zonas de la red de Kohonen responden a diferentes tipos de las clases presentes en el conjunto de datos<sup>17</sup>.

#### 4.5.1.2. Métodos supervisados

En estos métodos, la clasificación se basa en un aprendizaje previo del sistema, con conjuntos de calibración (o *training*) de objetos que definirán cada clase. Estos objetos son de conocida pertenencia a una de las clases. La calidad de los resultados de clasificación vendrá influenciada por la calidad de los conjuntos de entrenamiento. Estos métodos pueden

ser divididos en dos subgrupos: métodos discriminantes y métodos de modelado<sup>18</sup>.

### Métodos discriminantes

Dividen el espacio en tantas regiones como clases haya en el conjunto de *training*, creando unos límites compartidos por los espacios. Siempre clasifican una muestra desconocida como perteneciente a una de las clases. Los más comunes son:

- Análisis discriminante (DA): están basados en el concepto de una función discriminante que divide el espacio en regiones características para cada una de las clases, creando fronteras entre cada una de ellas<sup>19</sup>. Los dos métodos más conocidos son Análisis Discriminante Lineal (LDA) y Análisis Discriminante Cuadrático (QDA).
- KNN (k-nearest neighbour): este método clasifica un objeto test como perteneciente a la clase en la cual la mayoría de objetos pertenecientes a ella son más cercanos al objeto test. Generalmente se utiliza la distancia euclídea como medida de distancia<sup>20, 21</sup>.
- Potential Function Methods (PFM): estos métodos, de los cuales el más conocido en química analítica es el denominado ALLOC<sup>22, 23</sup>, pueden considerarse como generalizaciones del método KNN. En estos métodos, cada objeto de un training set es considerado como un punto en el espacio rodeado por un campo de potencial. La clasificación de un objeto test dentro de una de las clases es determinada mediante el potencial acumulado de la clase en la posición del objeto desconocido. El potencial acumulado se obtiene sumando los potenciales individuales de los objetos de la clase en la posición del objeto desconocido. El objeto test se clasifica dentro de la clase que da el mayor potencial acumulado.

### Métodos de modelado

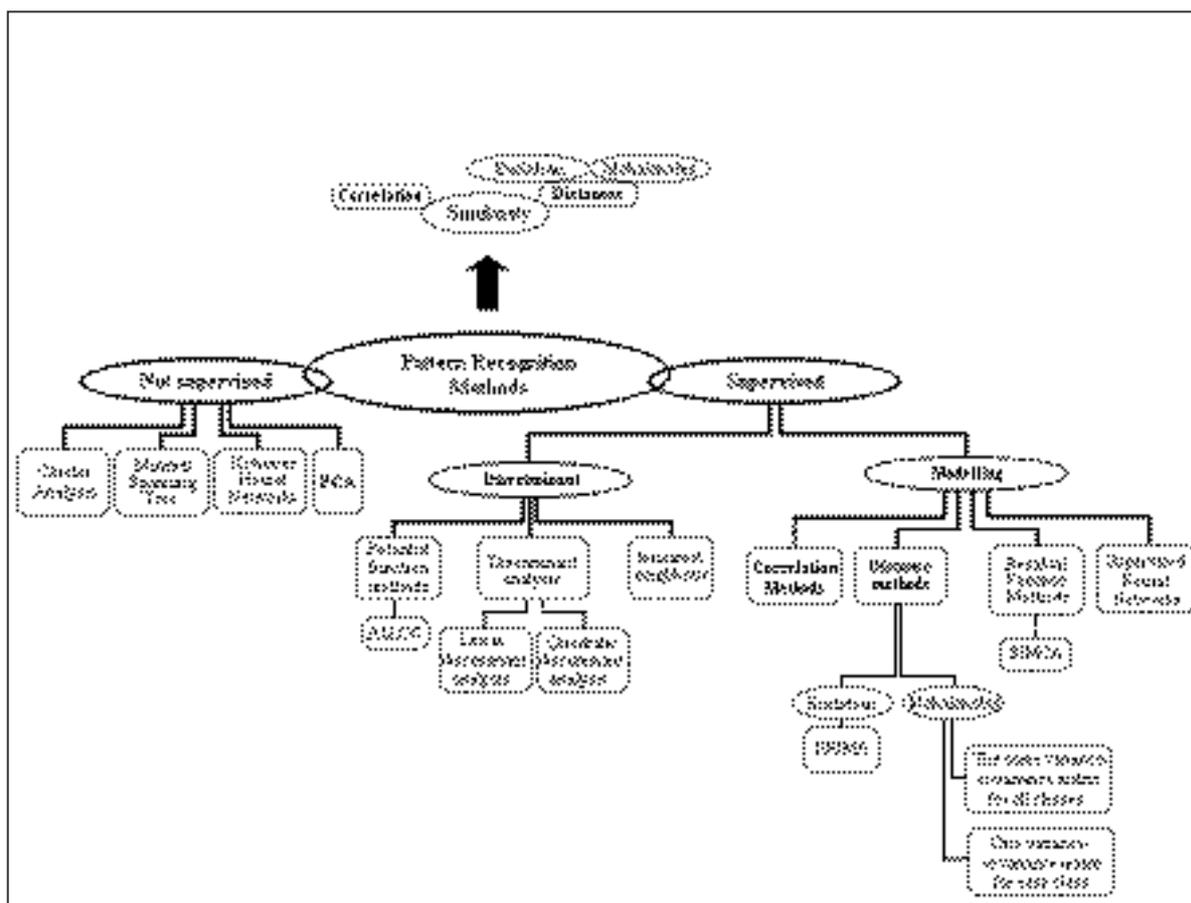
Los métodos de modelado se basan en la creación de unos volúmenes en el espacio, cada uno de ellos con unos límites distintos para cada una de las clases. Mediante estos métodos, una muestra puede clasificarse como perteneciente a alguna de las clases o a ninguna de ellas. Algunos de los más utilizados son:

- PRIMA (Pattern Recognition by Independent Multicategory Analysis): en

este método cada clase es modelada mediante su centroide, teniendo el modelo una forma esférica alrededor de éste. La semejanza entre un objeto y el modelo de cada clase se mide como la distancia euclídea calculada con las variables autoescaladas<sup>24</sup>. La distancia crítica ( $d_{crit}$ ) define el radio de las esferas de cada clase.

- UNEQ: el modelo se construye mediante el centroide de la clase, teniendo forma de hiper-elipsoide, siendo el centroide la media de la población y definiendo su tamaño mediante el valor de la  $d_{crit}$ . Teniendo en cuenta que el modelo de cada clase se construye mediante el cálculo de las distancias de Mahalanobis de forma individual, siendo la matriz de dispersión distinta para cada clase, UNEQ es especialmente adecuado para clases que presenten diferente dispersión (UNEQ=unequal dispersed classes)<sup>25</sup>.
- Métodos de varianza residual: estos métodos se basan en PCA de cada uno de los training sets, creando un modelo para cada uno de ellos<sup>26</sup>. Un espectro test se reconstruye según todos los modelos, y los residuales obtenidos son utilizados para calcular la probabilidad de que la muestra desconocida pertenezca o no a alguna de las clases. SIMCA es probablemente el más conocido de todos los métodos de varianza residual<sup>27</sup>.
- Redes neuronales artificiales supervisadas: estos métodos muestran un gran potencial en el campo de la clasificación de sustancias, debido a su gran capacidad de modelado, hecho que las hace especialmente adecuadas para la resolución de una amplia variedad de problemas. Existen diferentes tipos de redes neuronales artificiales según sea su proceso de entrenamiento. De entre todas, la más utilizada en el campo químico es la Multi-Layer Perceptron<sup>28</sup>.

Los métodos comentados en este punto son los más habituales para clasificación de objetos, aunque existen otros que son producto de combinaciones o variantes de los comentados. En la figura 4.1 se muestra esquemáticamente la clasificación de los métodos de reconocimiento de pautas propuesta.



**Figura 4.1.** Clasificación de los métodos de reconocimiento de pautas.

#### 4.5.2. Bibliotecas de identificación-clasificación de espectros

Una de las aplicaciones de los métodos de reconocimiento de pautas es la creación de bibliotecas de espectros que permitan clasificar, una vez creadas, muestras desconocidas como pertenecientes a alguna de las clases de la biblioteca o bien a ninguna de ellas. Son de gran utilidad en espectroscopia IR y NIR, ya que permiten llevar a cabo la identificación de un producto desconocido mediante su espectro de una forma objetiva, rápida y fiable, sin tener que depender de observaciones visuales, sujetas a interpretaciones según los conocimientos y experiencia del analista.

No todos los métodos de reconocimiento de pautas existentes son igualmente apropiados para la construcción de una biblioteca de identificación de espectros. Los más adecuados son los métodos supervisados de modelado ya que permiten crear clases a partir de muestras de identidad conocida, creando un modelo que represente cada clase. La identificación de un producto desconocido se realiza por comparación de su espectro con

cada uno de los modelos, siendo identificado como la clase con la que posea una mayor similitud, siempre que el resultado se encuentre entre los límites establecidos como aceptables en la construcción de la biblioteca. En el apartado de Metodología y Discusión Global de los Resultados, se describe uno de los trabajos presentados en esta memoria, consistente en la creación de bibliotecas de espectros NIR para la identificación de materias primas de interés farmacéutico, las cuales demuestran ser una herramienta muy potente para su aplicación en análisis cualitativo.

#### **4.6. TÉCNICAS QUIMIOMÉTRICAS EN ANÁLISIS CUANTITATIVO**

El proceso de calibración permite establecer la relación entre la respuesta instrumental y la propiedad del analito a determinar, utilizando a tal efecto un conjunto de muestras representativas. Como se ha comentado al principio del capítulo, los métodos cromatográficos separan los analitos de una muestra antes de medir la respuesta. Por ello, para cada analito se obtiene una sola variable respuesta, lo que simplifica el proceso de calibración. En cambio, las técnicas espectroscópicas proporcionan un gran número de variables respuesta para cada muestra, variables que en general no pueden ser asignadas a un solo analito. Esto ha propiciado el desarrollo de métodos de calibración capaces de relacionar múltiples variables con la propiedad a determinar. Estos métodos son conocidos como Métodos de Calibración Multivariable.

En general, los métodos de calibración pueden clasificarse siguiendo varios criterios<sup>29</sup>. De esta forma podemos tener métodos de calibración:

- Univariados-multivariados: en función del número de variables de la respuesta.
- Lineales-no lineales: según la función que relaciona las variables.
- Directos-indirectos: si los parámetros de la calibración se calculan a partir de la señal de los componentes puros o a partir de mezclas.
- Clásica-inversa: en función de cuál sea la variable independiente utilizada (concentración o señal analítica respectivamente).
- Espectro completo-Selección de variables: según el número de variables utilizadas.
- Rígidos-flexibles: en los primeros se debe disponer de información de todas las especies que contribuyen a la señal, mientras que en los segundos únicamente es necesario tener información de los analitos que se desea cuantificar.

Existe una gran variedad de estos métodos, por lo que la elección del más adecuado depende del tipo de muestra a analizar, el conocimiento del sistema y de los datos disponibles.

#### 4.6.1. Regresión lineal múltiple

El método de regresión lineal múltiple (MLR, *Multiple Linear Regression*) calcula, para el análisis de muestras con más de un componente, una relación lineal entre la señal y la concentración de cada componente mediante un ajuste por mínimos cuadrados. Se deben distinguir dos variantes de MLR llamadas clásica e inversa, según la relación establecida entre señal y concentración.

##### 4.6.1.1. Regresión lineal múltiple clásica

Asume el cumplimiento de la ley de Lambert-Beer para cada uno de los componentes de una mezcla en todo el intervalo de trabajo y la aditividad de las absorbancias en las muestras. La señal debida a una muestra es función de la concentración de cada componente por un coeficiente de proporcionalidad, de la forma expresada en la siguiente ecuación:

$$a_j = k_{j1}c_1 + k_{j2}c_2 + k_{j3}c_3 + \dots + k_{jn}c_n + e_j \quad (4.7)$$

siendo esta ecuación válida para una muestra que contenga n componentes, donde  $e_j$  es el error aleatorio de la medida,  $c_i$  es la concentración del componente i y  $k_{ji}$  es el coeficiente de proporcionalidad de cada muestra, igual al producto de su absorptividad molar por el camino óptico ( $\epsilon \cdot b$ ). Si se registran medidas a k longitudes de onda ( $k \geq n$ ), se obtendrá un sistema de ecuaciones, que al ser resuelto dará las constantes de proporcionalidad de la ecuación.

Éste es un método de espectro completo, por lo que puede proporcionar mejoras frente a otros métodos en los que se trabaja con un número limitado de variables<sup>30</sup>. Sin embargo, para que este método pueda aplicarse a la determinación simultánea de varios componentes, deben cumplirse las siguientes condiciones:

- Se deben conocer todos los componentes que contribuyen a la señal analítica.
- No se deben producir interacciones químicas ni físicas entre los analitos, ni tampoco entre analito y matriz.
- El ruido debe ser aleatorio en todo el intervalo de longitudes de onda y con la menor magnitud posible.

#### 4.6.1.2. Regresión lineal múltiple inversa

La regresión lineal múltiple inversa (ILS, Inverse Least Squares) asume una relación entre la absorbancia y la concentración que es inversa a la ley de Lambert-Beer. La concentración es función de la absorbancia a cada longitud de onda y se calcula multiplicando los valores de absorbancia por unos coeficientes de proporcionalidad:

$$y_i = b_0 + b_1x_1 + b_2x_2 + \dots + b_{K-1}x_{K-1} + e_i \quad (4.8)$$

Este método tiene la ventaja de que no es necesario conocer la concentración o la contribución de todas las especies absorbentes de la mezcla para la cuantificación de los analitos de interés. Sin embargo, los componentes no incluidos en la cuantificación deben estar presentes en todas las muestras y son modelados implícitamente.

La principal desventaja de este método es que el número de muestras debe ser mayor que el número de variables. La selección de las variables es una cuestión compleja, ya que la utilización de un número demasiado elevado de variables puede provocar problemas de colinealidad (diferentes variables independientes aportan la misma información, ya que están altamente correlacionadas), disminuyendo así la precisión de los resultados.

### 4.6.2. Métodos basados en reducción de variables

Estos métodos se basan en que la información contenida en las variables de la señal puede estar contenida en un número menor de variables sin que haya pérdida de información relevante (ver punto 4.4). El proceso de calibración se realiza, no sobre los datos originales, sino sobre estas nuevas variables, simplificando el modelo y la interpretación de los resultados.

Este tipo de métodos de calibración son de espectro completo, no presentan problemas de colinealidad ni las consecuencias derivadas de ella. Por estas razones, la tendencia actual es la utilización de métodos de calibración basados en una reducción de variables previa al cálculo del modelo.

Generalmente, de la misma forma que ocurría en PCA, los procedimientos de reducción de variables no son realizados sobre los datos originales sino que se centran o autoescalan previamente.

#### 4.6.2.1. Regresión en componentes principales

La Regresión en Componentes Principales (PCR, *Principal Component Regression*) aprovecha las propiedades de la descomposición en componentes principales (PCA), realizando una regresión múltiple inversa (ILS) de la propiedad a determinar sobre los scores obtenidos en el PCA en lugar de realizarla sobre los datos originales. No existe pérdida de información útil, ya que los scores contienen la misma información que los datos originales pero habiendo eliminado el ruido.

Si tenemos una muestra con un conjunto de  $P$  especies absorbentes, tendremos  $P$  variables  $y_1, y_2, y_3 \dots y_p$  representando la concentración de cada componente que pueden ser escritas en forma de vector  $\mathbf{y}$ . El espectro de la misma, registrado a  $K$  longitudes de onda constituye un conjunto de  $K$  variables independientes  $x_1, x_2, x_3 \dots x_k$  que pueden ser escritas en forma de vector  $\mathbf{x}$ . Si se construye un conjunto de calibración con  $M$  objetos, se pueden agrupar los vectores que describen cada uno de ellos en dos matrices: la matriz  $\mathbf{Y}$ , que contiene las concentraciones de cada componente en cada muestra, de dimensiones  $(M \times P)$  y la matriz  $\mathbf{X}$ , que contiene los espectros de cada muestra, de dimensiones  $(M \times K)$ . De este modo, dentro de las matrices la información espectral o las propiedades a determinar de cada muestra están escritas en una fila, mientras que cada columna contiene la información de una variable concreta para todas las muestras presentes.

El primer paso del PCR consiste en realizar una descomposición de la matriz  $\mathbf{X}$  en sus componentes principales, de la forma indicada en el punto 4.4.2.

$$\mathbf{X} = \mathbf{TP}^t + \mathbf{E} = \sum_{a=1}^A \mathbf{t}_a \mathbf{p}_a^t + \mathbf{E} \quad (4.9)$$

Una vez elegido el número de  $A$  componentes principales que se considere como óptimo para describir la matriz  $\mathbf{X}$  (ver procedimientos de elección de componentes principales en el punto 4.6.2.4), ésta se puede representar por su matriz de scores  $\mathbf{T}$ .

$$\mathbf{T} = \mathbf{XP} \quad (4.10)$$

Hasta aquí lo que se ha realizado es un PCA, obteniendo a partir de la matriz de datos  $\mathbf{X}$  la matriz de scores  $\mathbf{T}$  y la de loadings  $\mathbf{P}$ . A continuación, la matriz de datos  $\mathbf{Y}$  se puede calcular según la expresión:

$$\mathbf{Y} = \mathbf{TB} + \mathbf{E} \quad (4.11)$$

siendo **B** la matriz de regresores que se halla por mínimos cuadrados conociendo los valores de **Y** del conjunto de calibración:

$$\mathbf{B} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{Y} \quad (4.12)$$

El símbolo  $\hat{\phantom{x}}$  indica valores calculados. Una vez calculados los regresores, y por lo tanto establecido el modelo de calibración, se pueden realizar los cálculos para predecir un conjunto de nuevas muestras. En primer lugar, la matriz de datos espectroscópicos del conjunto de muestras de predicción  $\mathbf{X}^*$ , se centra o autoescala utilizando los valores calculados a partir de la matriz de datos **X** empleada en la calibración. A partir de la matriz de loadings calculada en la calibración, para el número A de componentes principales óptimo, se calculan los scores de las muestras de predicción,  $\mathbf{T}^*$ .

$$\mathbf{T}^* = \mathbf{X}^*\mathbf{P} \quad (4.13)$$

y se utiliza la matriz de regresores calculada también en la calibración, junto con los scores de estas muestras, para el cálculo de la propiedad a determinar en las muestras desconocidas.

$$\mathbf{Y} = \mathbf{T}^*\mathbf{B} \quad (4.14)$$

Uno de los principales problemas con PCR es que los componentes principales que mejor representan la matriz de los datos espectroscópicos, **X**, pueden no ser los más apropiados para la predicción de las propiedades de los analitos que queremos determinar. Por este motivo se ha desarrollado otra técnica de calibración que intenta concentrar el máximo poder predictivo en los primeros componentes principales. Este nuevo método es la regresión parcial por mínimos cuadrados.

#### 4.6.2.2. Regresión parcial por mínimos cuadrados

El método de Regresión Parcial por Mínimos Cuadrados (PLS, Partial Least-Squares Regression) fue desarrollado por H. Wold en 1975<sup>31</sup>. La principal diferencia con respecto a PCR radica en el hecho de que en PLS se intenta contener la mayor información para la predicción de las muestras en los primeros componentes principales. Para ello, durante la etapa de calibración, el algoritmo PLS utiliza tanto la información contenida en la matriz de datos espectroscópicos (matriz **X**) como la información contenida en la matriz de la propiedad a determinar (matriz **Y**), obteniéndose unas variables auxiliares llamadas variables latentes, factores o componentes PLS.

Una vez centradas o autoescaladas las matrices  $\mathbf{X}$  e  $\mathbf{Y}$ , cada una de las matrices se descompone en una suma de  $A$  factores, calculando simultáneamente:

$$\mathbf{X} = \mathbf{TP}^t + \mathbf{E} = \sum_{a=1}^A \mathbf{t}_a \mathbf{p}_a^t + \mathbf{E} \quad (4.15)$$

$$\mathbf{Y} = \mathbf{UQ}^t + \mathbf{F} = \sum_{a=1}^A \mathbf{u}_a \mathbf{q}_a^t + \mathbf{F} \quad (4.16)$$

La ecuación (4.15) es la descomposición de la matriz de datos espectroscópicos  $\mathbf{X}$ , siendo  $\mathbf{T}$  es la matriz de scores,  $\mathbf{P}$  la de loadings y  $\mathbf{E}$  la matriz de residuales. En la ecuación (4.16), que es la descomposición de la matriz de la propiedad a determinar  $\mathbf{Y}$ ,  $\mathbf{U}$  es la matriz de scores,  $\mathbf{Q}$  la matriz de loadings y  $\mathbf{F}$  la matriz de residuales. Si tenemos  $M$  muestras,  $A$  factores,  $K$  variables y  $P$  analitos, la dimensionalidad de las matrices es la siguiente:  $\mathbf{T}$  y  $\mathbf{U}$  ( $M \times A$ ),  $\mathbf{P}^t$  ( $A \times K$ ) y  $\mathbf{Q}^t$  ( $A \times P$ ). En este caso los loadings no coinciden con las direcciones de máxima variabilidad de las muestras como en el caso de PCA, ya que están corregidos para obtener la máxima capacidad predictiva para la matriz  $\mathbf{Y}$ .

La descomposición de ambas matrices no es independiente, sino que se realiza de forma simultánea, estableciéndose una relación interna entre los scores de los bloques  $\mathbf{X}$  e  $\mathbf{Y}$  de forma que para cada componente  $a$  se cumpla:

$$\hat{u}_a = b_a t_a \quad (4.17)$$

donde el símbolo  $\hat{}$  indica que es una cantidad calculada y  $b_a$  es el coeficiente de regresión para cada uno de los factores PLS. Si se desea calcular una única concentración de las presentes en la matriz  $\mathbf{Y}$ , el algoritmo recibe el nombre de PLS1. El algoritmo PLS1<sup>32</sup> ha sido el empleado en casi todas las calibraciones multivariantes de los trabajos que conforman esta memoria, por lo que en las tabla 4.1 y 4.2 se muestra una descripción detallada del proceso de calibración y predicción respectivamente mediante PLS1.

Tabla 4.1. Descripción del algoritmo PLS.

Calibración	
	Una vez se tienen los datos centrados ( $\mathbf{X}_0$ ) para cada factor $a=1,2,\dots,A$ se realizan los pasos del 1 al 5:
Paso 1	Se calculan los llamados “pesos” ( <i>weights</i> ) $\mathbf{w}_a$ por mínimos cuadrados, utilizando la variabilidad restante en $\mathbf{y}$ expresada como vector $\mathbf{y}_{a-1}$ : $\mathbf{X}_{a-1} = \mathbf{y}_{a-1} \mathbf{w}_a^T + \mathbf{E}$ escalando a continuación el vector $\mathbf{w}_a$ : $\mathbf{w}_a^T = c \mathbf{X}_{a-1}^T \mathbf{y}_{a-1}$ siendo $c$ un factor de escalado que hace que la longitud del vector final $\mathbf{w}_a$ sea igual a 1.
Paso 2	Se estiman los scores $\mathbf{t}_a$ utilizando los pesos obtenidos en el modelo local: $\mathbf{X}_{a-1} = \mathbf{t}_a \mathbf{w}_a^T + \mathbf{E}$ siendo los scores calculados por mínimos cuadrados: $\mathbf{t}_a = \mathbf{X}_{a-1} \mathbf{w}_a$
Paso 3	Con los scores calculados, se estiman los loadings $\mathbf{p}_a$ de los datos espectroscópicos. Con el modelo local: $\mathbf{X}_{a-1} = \mathbf{t}_a \mathbf{p}_a^T + \mathbf{E}$ de nuevo por mínimos cuadrados se calcula: $\mathbf{p}_a = \mathbf{X}_{a-1}^T \mathbf{t}_a / \mathbf{t}_a^T \mathbf{t}_a$
Paso 4	Lo siguiente es calcular el loading de concentración $\mathbf{q}_a$ utilizando a tal efecto el modelo: $\mathbf{y}_{a-1} = \mathbf{t}_a \mathbf{q}_a^T + \mathbf{f}$ lo que da como solución: $\mathbf{q}_a = \mathbf{y}_{a-1}^T \mathbf{t}_a / \mathbf{t}_a^T \mathbf{t}_a$
Paso 5	Se crean una matriz $\mathbf{X}_{a-1}$ nueva y un residual $\mathbf{y}_{a-1}$ sustrayendo el efecto de este factor. Ahora $a$ habrá aumentado en una unidad ( $a=a+1$ ) $\mathbf{X}_{a-1}^{(nueva)} = \mathbf{X}_{a-1} - \mathbf{t}_a \mathbf{p}_a^T$ $\mathbf{y}_{a-1}^{(nueva)} = \mathbf{y}_{a-1} - \mathbf{t}_a \mathbf{q}_a^T$
Paso 6	Se determina el número $A$ de factores PLS y se calculan los predictores $\mathbf{b}_0$ y $\mathbf{b}$ para ser utilizados en la predicción de nuevas muestras: $\mathbf{b} = \mathbf{W}(\mathbf{P}^T \mathbf{W})^{-1} \mathbf{q}$ $\mathbf{b}_0 = \bar{y} - \bar{x}^T \mathbf{b}$
Predicción	
La concentración $y$ de una muestra desconocida cuyo espectro es $\mathbf{x}$ se calcula utilizando los regresores obtenidos en el paso 6 de la forma: $y = b_0 + \mathbf{x}^T \mathbf{b}$	

#### 4.6.2.3. Evaluación de la capacidad predictiva del modelo

El objetivo de la calibración es obtener unos parámetros que permitan calcular la propiedad a determinar en futuras muestras de forma que los valores obtenidos difieran lo menos posible de los valores reales.

En modelos de clasificación la capacidad predictiva se evalúa por los aciertos en la clasificación de muestras que no hayan formado parte de la calibración.

En modelos de cuantificación, se realizan estudios cuantitativos de los resultados obtenidos. Se debe utilizar un parámetro que permita evaluar el error medio de toda la población, no sólo de una muestra. Para ello, se puede utilizar el sumatorio del cuadrado de los residuales ( $(\hat{y}_{ij} - y_{ij})^2$ ), denominado habitualmente PRESS (Predicted Residual Error Sum of Squares) o su valor medio obtenido dividiendo el PRESS por el número de muestras de predicción ( $n_p$ ), conocido como MSEP (Mean Square Error of Prediction) ( $(\hat{y}_{ij} - y_{ij})^2 / n_p$ ). También se puede utilizar la raíz cuadrada del MSEP, denominada RMSEP (*Root Mean Square Error of Prediction*).

Para la construcción de un modelo se suelen utilizar dos conjuntos de muestras: de calibración y de validación (o *test set*). Si se procede de esta forma, se calcula el MSEP o el RMSEP para cada componente principal. De igual forma, también se pueden calcular estos parámetros para el conjunto de calibración (MSEC, RMSEC). En calibración, en algunos casos no se divide por el número de muestras del conjunto de calibración sino por los grados de libertad reales ( $n_c - 1 - a$ ) siendo  $a$  el número de componentes principales para el cual se está calculando el parámetro.

Si el número de muestras disponibles es relativamente pequeño, la metodología a seguir es la validación cruzada (cross validation)<sup>10</sup>, la cual utiliza, para comprobar el ajuste del modelo, muestras del conjunto de calibración. Mediante este método, el conjunto de calibración se divide en varios bloques o segmentos. El modelo se construye tantas veces como número de segmentos se ha elegido, utilizando un segmento como bloque de datos para comprobar resultados y el resto para construir el modelo, de forma que se deja un segmento fuera cada vez. Esto se realiza para cada factor, calculando el MSE para cada segmento y acumulándolo de forma que se consiga una buena estimación del poder predictivo de las muestras de calibración. Cuando el número de segmentos es igual que el número de muestras de calibración, se le denomina *leave one out* (LOO). Este procedimiento deja cada vez una muestra fuera, utilizando el resto para el cálculo del modelo, repitiendo el proceso un número de veces igual al número de muestras. El MSECv (Mean Square Error of Prediction by Cross Validation) se expresa:

$$MSECV = \frac{\sum_{i=1}^{n_c} (\hat{y}_{ij} - y_{ij})^2}{n_c} \quad (4.18)$$

#### 4.6.2.4. Elección del número de componentes principales

La elección del número de componentes principales o factores que configura el modelo óptimo es el punto clave en la utilización de cualquier técnica de calibración que realice reducción de variables. Se han propuesto formas para reducir el número de PCs, que se basan, en general, en el análisis del error de predicción al utilizar distinto número de éstos. Puesto que el método más usual de construcción del modelo es el de validación cruzada, una forma muy popular de seleccionar el número de factores óptimo es la sugerida por Wold<sup>33</sup>, consistente en representar el valor de MSECV frente al número de PCs y elegir como óptimo el mínimo de la curva. Se parte de que el error disminuye al aumentar el número de PCs que se utilizan en el modelo, hasta que llega un momento en que los nuevos PCs únicamente explican ruido y el MSECV aumenta debido al sobreajuste del modelo. Aunque a priori parece una idea razonable, el hecho de utilizar únicamente un número limitado de muestras hace que el método esté sujeto a error y que genere un sobreajuste de los datos<sup>34</sup>.

Cuando no hay un mínimo, se utiliza el criterio de Haaland y Thomas<sup>35</sup> para la elección del número de componentes. Este método consiste en escoger el número de componentes cuyo valor de MSECV no es significativamente diferente del valor mínimo del modelo. El mínimo valor de MSECV vendrá dado por un número de componentes que denominamos  $a^*$ . Cada valor de MSECV obtenido con un número de PCs menor que  $a^*$  se compara con el valor mínimo mediante una prueba de Fischer. Para cada componente  $a=1,2,\dots,a^*$  se calcula

$$F(a) = \frac{MSECV(a)}{MSECV(a^*)} \quad (4.19)$$

Como número de componentes óptimo se escoge el menor tal que  $F(a) < F_{\alpha; gdl; gdl}$ , donde  $F_{\alpha; gdl; gdl}$  es el valor tabulado para una prueba F unilateral con un porcentaje de nivel de significación de  $(1-\alpha)$ . Basándose en un criterio meramente empírico, el valor de  $\alpha$  recomendado por Haaland y Thomas es de 0.25. Los grados de libertad del numerador y denominador son los mismos ( $n_c$  en el caso de PLS1).

#### 4.7. CORRECCIÓN ORTOGONAL DE LA SEÑAL

La corrección ortogonal de la señal (OSC) es una novedosa técnica de tratamiento de datos espectroscópicos desarrollada por Wold et al.<sup>36</sup> cuya idea principal es corregir la matriz de datos espectroscópicos  $\mathbf{X}$ , eliminando de ésta la parte que no aporta información sobre la matriz de concentraciones  $\mathbf{Y}$ . Éste tratamiento se aplica conjuntamente a todos los espectros presentes en la matriz de calibración. Una vez aplicado este tratamiento a la matriz de calibración, esta corrección puede ser aplicada a un conjunto de predicción externo para evaluar la capacidad predictiva del modelo construido con los datos tratados.

Debido a la naturaleza de esta técnica, que resulta una mezcla entre pretratamientos espectrales y calibración multivariable (se usa para corregir espectros, pero utilizando a tal efecto un método de calibración multivariable), y a que es el tema abordado en uno de los trabajos de esta tesis, se ha preferido tratarla en un apartado aparte del resto y tras haber tratado el resto de temas.

El algoritmo utilizado en este tipo de corrección es similar al algoritmo NIPALS utilizado comúnmente en PCA y PLS y está descrito con detalle en el trabajo de Wold et al.<sup>36</sup>. En cada paso de dicho algoritmo, se modifica el vector de pesos ( $w$ ), imponiéndole la condición que  $t=X \cdot w$  sea ortogonal a la matriz de concentraciones  $\mathbf{Y}$ , siendo  $t$  el score correspondiente. Para el cálculo OSC, en lugar de imponer en el algoritmo la condición de que los pesos ( $w$ ) sean calculados para maximizar la covarianza entre  $\mathbf{X}$  e  $\mathbf{Y}$ , tal y como se hace en PLS, se intenta todo lo contrario: minimizar dicha covarianza, haciendo a  $t$  lo más cercano a la ortogonalidad con  $\mathbf{Y}$ . Se busca modelar toda aquella información de la matriz de datos espectroscópicos que no contenga información de la matriz de concentraciones. El resultado de este cálculo es una matriz de scores y una matriz de loadings, que modelan toda aquella información no relacionada con la concentración. Una vez modelada, dicha información es eliminada de los datos espectroscópicos, restando a la matriz de datos originales ( $\mathbf{X}$ ) la matriz de scores ortogonales por la matriz de loadings:

$$\mathbf{X}_{\text{OSC}} = \mathbf{X} - \sum_{i=1}^n \mathbf{t}_i \mathbf{p}_i' \quad (4.20)$$

siendo  $i$  el número de veces que se efectúa el tratamiento (pasadas de filtro o factores OSC).

Este tipo de tratamiento matemático ha sido ya aplicado con éxito al pretratamiento de datos espectroscópicos NIR en transferencia de calibraciones<sup>37</sup>.

## BIBLIOGRAFÍA

---

- <sup>1</sup> D. L. Massart, B. G. M. Vandeginste, S. N. Deming, Y. Minchotte and L. Kaufmann, *Chemometrics: A textbook*, Elsevier, Amsterdam, 1988.
- <sup>2</sup> A. Savitzky and M. J. E. Golay, *Anal. Chem.*, 1964, **36**, 1627.
- <sup>3</sup> G. Horlick, *Anal. Chem.*, 1972, **44**, 943.
- <sup>4</sup> K. H. Norris and P. C. Williams, *Cereal Chem.*, 1984, **61**, 158.
- <sup>5</sup> P. Geladi, D. McDougall and H. Martens, *Appl. Spectrosc.*, 1985, **39**, 491.
- <sup>6</sup> R. J. Barnes, M. S. Dhanoa and S. J. Lister, *Appl. Spectrosc.*, 1989, **43**, 772.
- <sup>7</sup> M. S. Dhanoa, S. J. Lister, R. Sanderson and R. J. Barnes, *J. Near Infrared Spectrosc.*, 1994, **2**, 43.
- <sup>8</sup> J. E. Jackson, *User's Guide to Principal Components*, Wiley, New York, 1991.
- <sup>9</sup> S. Wold, K. Esbensen and P. Geladi, *Chemom. Intell. Lab. Syst.*, 1987, **2**, 37.
- <sup>10</sup> P. Geladi and B. R. Kowalski, *Anal. Chim. Acta*, 1985, **185**, 1.
- <sup>11</sup> M. Blanco, R. Boqué, R. Cela, J. Coello, S. MasPOCH, M. C. Ortiz, J. Riba, X. Rius, A. Ruiz, L. A. Sarabia and X. Tomás, *Avances en Quimiometría Práctica*, Universidad de Santiago de Compostela, 1994.
- <sup>12</sup> H. Wold, *Multivariate Analysis*, Ed. P. R. Krishnaiah, Academic Press, New York, 1966.
- <sup>13</sup> M. Blanco, J. Coello, H. Iturriaga, S. MasPOCH and C. de la Pezuela, *Anal. Chim. Acta* 298, 183 (1994).
- <sup>14</sup> D. A. Burns and E. Ciurczac, *editors in Handbook of Near Infrared Analysis*, Marcel Dekker, Inc, New York (1992). Capítulo 13: *Qualitative discriminant analysis* by H. Mark.
- <sup>15</sup> N. Bratchell, *Chemom. Intell. Lab. Syst.*, 1989, **6**, 105.
- <sup>16</sup> O. Strouf, *Chemical Pattern Recognition*, Research Stuides Press LTD, Letchworth, Hertfordshire, England, 1986.
- <sup>17</sup> P. Cáceres-Alonso and A. García-Tejedor, *J. Near Infrared Spectrosc.*, 1995, **3**, 97.
- <sup>18</sup> M. P. Derde and D. L. Massart, *Anal. Chim. Acta*, 1986, **191**, 1.
- <sup>19</sup> P. A. Lachenbruch, *Discriminant Analysis*, Hafner Press, New York, 1975.
- <sup>20</sup> D. Coomans and D. L. Massart, *Anal. Chim. Acta*, 1982, **136**, 15.
- <sup>21</sup> D. Coomans and D. L. Massart, *Anal. Chim. Acta*, 1982, **138**, 153.
- <sup>22</sup> D. Coomans and I. Broeckaert, *Potential Pattern Recognition*, Wiley, New York, 1986.
- <sup>23</sup> D. Coomans, M. P. Derde, D. L. Massart and I. Broeckaert, *Anal. Chim. Acta*, 1981, **133**, 241.
- <sup>24</sup> I. Jurickskay and G. E. Veress, *Anal. Chim. Acta*, 1981, **171**, 61.
- <sup>25</sup> M. P. Derde and D. L. Massart, *Anal Chim. Acta*, 1986, **184**, 33.

- 
- <sup>26</sup> *VISION User Manual*, Foss NIRSystems, Silverspring, Maryland, USA, 1998.
- <sup>27</sup> P. J. Gemperline and L. D. Webber, *Anal. Chem.*, 1989, **61**, 138.
- <sup>28</sup> E. Bertran, M. Blanco, J. Coello, H. Iturriaga, S. Maspoch and I. Montoliu, *J. Near Infrared Spectrosc.*, 2000, **8**, 45.
- <sup>29</sup> H. Martens and T. Naes, *Multivariate Calibration*, John Wiley & Sons, Chichester, England, 1989.
- <sup>30</sup> D. M. Haaland and R. G. Easterling, *Appl. Spectrosc.*, 1980, **34**, 539.
- <sup>31</sup> H. Wold, *Soft Modelling by Latent Variables; the Non-Linear Iterative Partial Least Squares Approach*, en *Perspectives in Probability and Statistics*, Paper in Honour of M. S. Barlett, ed. J. Gani, Academic Press, London, 1975.
- <sup>32</sup> *The Unscrambler 6.0 User's Guide*, CAMO AS, Trondheim, Norway, 1996.
- <sup>33</sup> S. Wold, *Technometrics*, 1987, **20**, 397.
- <sup>34</sup> D. W. Osten, *J. Chemometrics*, 1988, **2**, 39.
- <sup>35</sup> D. M. Haaland and E. V. Thomas, *Anal. Chem.*, 1988, **60**, 1193.
- <sup>36</sup> S. Wold, H. Antti, F. Lindgren and J. Öhman, *Chemom. Intell. Lab. Syst.*, 1998, **44**, 175.
- <sup>37</sup> J. Sjöblom, O. Svensson, M. Josefson, H. Kullberg and S. Wold, *Chemom. Intell. Lab. Syst.*, 1998, **44**, 229.

## METODOLOGÍA Y DISCUSIÓN

### GLOBAL DE LOS RESULTADOS

---

#### 1. INTRODUCCIÓN

El control de calidad en cualquier tipo de industria exige la determinación de numerosos parámetros químicos y también físicos mediante los cuales se quiere asegurar que el producto que se pone a la venta cumplirá todos los requisitos que exige el consumidor. Este control es aún más estricto en la industria farmacéutica, al ser el producto final destinado al consumo humano, aumentando consiguientemente la cantidad y variedad de parámetros a establecer así como la estrechez del intervalo de variación de los mismos. Para poder asegurar estos niveles de calidad, la industria farmacéutica se ve obligada a utilizar una gran variedad de técnicas físicas y químicas y de metodologías distintas que desde el punto de vista de la productividad pueden presentar limitaciones. Entre las limitaciones más destacables de estas técnicas podemos destacar:

- Elevados tiempos de análisis
- Importante consumo de reactivos, principalmente disolventes
- Pretratamientos laboriosos y largos de muestra, y como consecuencia resultados que pueden depender de la habilidad y experiencia del analista.

La tendencia actual de las industrias farmacéuticas es la utilización de métodos instrumentales, más robustos y fiables que los métodos tradicionales, y que permiten superar algunas de las limitaciones que éstos presentan. No obstante, los métodos instrumentales más utilizados, como HPLC, implican todavía un elevado consumo de disolventes y generación de residuos. Actualmente existen métodos instrumentales de análisis que permiten reducir e incluso resolver algunos de los problemas citados anteriormente.

El objetivo principal de esta tesis es el desarrollo de nuevas metodologías de análisis mediante técnicas instrumentales de reciente implantación, cuyas características las hacen

especialmente adecuadas para llevar a cabo el control de calidad, superando algunas limitaciones de los métodos actuales.

En esta memoria se propone la utilización de dos técnicas instrumentales de análisis: la electroforesis capilar y la espectroscopia en el infrarrojo cercano, ambas técnicas de reciente implantación y cuyo uso en la industria farmacéutica no está muy extendido. Ambas técnicas pueden satisfacer muchas de las exigencias del control de calidad en la industria farmacéutica.

Con estas dos técnicas se ha pretendido, no sólo desarrollar metodologías de análisis para control de calidad farmacéutico, puesto que esto ha sido ya realizado en numerosas ocasiones por este mismo grupo de investigación, sino también abordar problemas conceptuales que aparecen durante la etapa de implantación de estos métodos, y que debido a la reciente introducción de estas técnicas instrumentales no han sido completamente establecidos.

## **2. METODOLOGÍA**

La distinta naturaleza de las técnicas utilizadas (la primera de ellas una técnica de separación, mientras que la segunda una técnica espectrofotométrica) conduce a metodologías y estrategias de trabajo distintas en cada caso. Por ello, en cada uno de los puntos que conciernen a metodología, tanto instrumental, experimental, como de tratamiento de datos, se distingue entre electroforesis capilar y NIR.

### **2.1. METODOLOGÍA INSTRUMENTAL**

Los métodos de análisis presentados en los trabajos adjuntos en los anexos han sido desarrollados mediante las técnicas instrumentales de electroforesis capilar y espectroscopia NIR.

La electroforesis capilar (CE) es una técnica de separación que presenta una serie de ventajas con respecto a las técnicas cromatográficas, que la hacen especialmente adecuada como técnica analítica. Es una técnica que posee una alta capacidad de resolución y permite obtener elevadas eficacias en el proceso de separación. Las reducidas dimensiones de los capilares comportan un bajo consumo de reactivos y de muestra, así como una baja generación de residuos. Posee una gran versatilidad debido a los distintos modos de aplicación y al gran número de parámetros variables para mejorar la separación (aditivos, disolventes orgánicos, tensoactivos, selectores quirales...). Por último, se debe destacar el alto grado de automatización que permite, lo que facilita el trabajo de análisis, mejorando así

la calidad de los resultados. Este conjunto de características permite pensar que puede llegar a sustituir a HPLC en un futuro próximo. Sin embargo, también presenta una serie de inconvenientes como que no mejora la sensibilidad y presenta una mayor irreproducibilidad en el volumen de inyección y en los tiempos de migración.

La espectroscopia NIR es una técnica que permite efectuar medidas directamente sobre la muestra a analizar, ya sea sólida o líquida, sin realizar un tratamiento previo de la misma. Esto sin duda reduce los tiempos de análisis y elimina el consumo de reactivos y por consiguiente, de residuos generados, además de permitir mejorar la precisión de los resultados, ya que no hay procesos de preparación entre la toma de muestra y el resultado final. La naturaleza del espectro NIR, que contiene información tanto física como química, permite aplicarla a una gran variedad de análisis, como determinación de características físicas y químicas, posibilidad de realizar análisis cualitativo y cuantitativo de principios activos y excipientes farmacéuticos... Aunque esta técnica no ha sido aceptada todavía en las farmacopeas, ya hay grupos trabajando para este objetivo. Uno de los mayores inconvenientes de la técnica es la necesidad de utilización de técnicas de calibración multivariable, que alargan el proceso de calibración y requieren personal entrenado. Sin embargo, su aplicación puede ser realizada por personal no cualificado.

## **2.2. METODOLOGÍA EXPERIMENTAL**

Para cada una de las dos técnicas utilizadas, la metodología experimental seguida es distinta puesto que cada una de ellas requiere de una serie de condiciones y estrategias diferentes. A continuación se expone la metodología seguida para cada técnica.

### **2.2.1. Metodología experimental CE**

La metodología experimental que se debe seguir para el desarrollo de un método de análisis mediante electroforesis capilar debe incluir los siguientes puntos:

- Elección del pH de trabajo, en base al estudio de las características de los componentes de la muestra problema y su posible comportamiento.
- Elección del tampón, con capacidad tamponante suficiente al pH elegido, naturaleza del mismo, etc.
- Posible adición de aditivos al tampón, para mejorar la separación. Estos aditivos pueden ser selectores quirales (separaciones quirales), tensoactivos (separaciones de compuestos neutros) o aditivos orgánicos (modificadores del flujo electroosmótico).

- Fijar parámetros instrumentales de trabajo, como potencial y temperatura. Deben establecerse de forma empírica lo que exige varias experiencias para determinar las mejores condiciones.
- Acondicionamiento del capilar. Es importante seguir siempre una misma metódica de acondicionamiento ya que de ello dependerá el estado de la superficie del capilar y por consiguiente la magnitud del flujo electroosmótico (EOF).
- Establecimiento de la ecuación de calibración, inyectando muestras de concentraciones conocidas de cada uno de los componentes a determinar, utilización de patrones internos.
- Reacondicionamiento del capilar. Mantener una superficie reproducible influirá directamente en la reproducibilidad de los resultados. Se vuelve a preparar el capilar para las siguientes inyecciones.

### **2.2.2. Metodología experimental NIR**

La espectroscopia NIR permite desarrollar muy diversas metodologías de análisis, ya que posee gran cantidad de módulos distintos y accesorios que facilitan el trabajo (módulos de líquidos, módulos para sólidos, sondas remotas de fibra óptica, reflectores, cubetas de flujo...). Debido a esto, la metodología experimental seguida para todos los trabajos desarrollados por NIR ha sido diferente debido a la diferente finalidad de cada uno de los métodos desarrollados. Sin embargo, existen unas etapas comunes que todo usuario de la técnica NIR debe seguir para el desarrollo de una metódica de análisis:

- Elección del módulo de medida NIR. Para ello es necesario conocer la naturaleza de la muestra a analizar: sólida, líquida, gel, granulada, comprimidos intactos... En función de las características de la muestra se escoge el módulo o se diseña un montaje experimental que permita obtener espectros reproducibles.
- Registro de espectros. Una vez comprobado el correcto funcionamiento del instrumento se registran los espectros NIR de las muestras. Cada muestra suele registrarse por triplicado para ser promediados los tres espectros posteriormente.
- Método de referencia. Es necesario analizar las muestras que se utilizan para la construcción del modelo mediante un método de referencia adecuado que permita obtener la concentración del componente o magnitud de la propiedad a determinar, o bien que permita establecer la identidad de la muestra en el caso de bibliotecas de identificación.

## **2.3. METODOLOGÍA DEL TRATAMIENTO DE DATOS**

La principal diferencia entre los métodos desarrollados mediante EC y los desarrollados por NIR es que los primeros se realizan mediante calibración univariable mientras que los otros requieren de métodos multivariados.

### **2.3.1. Tratamiento de datos de CE**

Para desarrollar un método cuantitativo por CE generalmente se utiliza calibración univariable, siendo la magnitud a relacionar con la concentración la llamada área de pico corregida. Se calculan las curvas de calibración para cada uno de los analitos, utilizando los parámetros de las rectas para la predicción de nuevas muestras, que se realiza por interpolación. En ocasiones puede resultar útil el uso de un patrón interno que permita corregir posibles variaciones instrumentales del sistema electroforético.

### **2.3.2. Tratamiento de datos NIR**

Las características del espectro NIR hacen que, frecuentemente, se le aplique un pretratamiento que elimine o reduzca los efectos indeseados en la señal. Existe una notable variedad de pretratamientos, y su selección es empírica aunque se pueden utilizar ciertas pautas en función de la corrección a aplicar (ver Introducción-Quimiometría...).

Los métodos de calibración que se utilizan en NIR son de naturaleza multivariable y la selección del método más adecuado depende de la finalidad de la metodología de análisis y de las características del producto a analizar. Una vez elegido el método de calibración se realiza el cálculo del modelo. A continuación, se estudian los resultados del modelo de calibración para detectar y eliminar posibles puntos anómalos (outliers) que reduzcan la calidad de los resultados. El siguiente paso consiste en la validación del modelo, aplicando éste a un conjunto de muestras, distintas a las utilizadas para la construcción del modelo, de las cuales se conoce la propiedad a determinar mediante métodos de referencia. Los resultados, tanto de calibración como de validación, se evalúan mediante el uso de herramientas estadísticas que permiten comparar los resultados NIR con los de referencia. Tras la validación del modelo, éste puede comenzar a ser aplicado para análisis de rutina, realizando un seguimiento continuo para detectar posibles desviaciones del mismo.

## **3. DISCUSIÓN DE RESULTADOS**

Este apartado recoge aquellos aspectos más remarcables de cada una de las metodologías de análisis presentadas, enmarcándolas en la problemática concreta para la cual han sido desarrolladas. Esta parte del trabajo no pretende ser redundante con respecto a

lo expuesto en los artículos, que se adjuntan en los anexos, sino que intenta dar una visión más amplia de aquellos puntos que suponen un avance con respecto a otros métodos de análisis vigentes. Como es normal, todo nuevo método tiene sus limitaciones, por lo que también son comentadas a la vez que se proponen posibles estrategias para poder superarlas.

### **3.1. MÉTODOS DE ANÁLISIS POR ELECTROFORESIS CAPILAR**

El control de calidad farmacéutico no debe limitarse a asegurar la calidad de un preparado farmacéutico en el momento de su fabricación sino que también debe asegurar su estado de conservación a lo largo del tiempo. El principio activo o algunos de los excipientes y/o preservantes pueden verse sometidos a procesos de transformación, que pueden provocar la pérdida de potencia del fármaco. Es necesario disponer de métodos de análisis que permitan controlar el estado de conservación en el que se encuentra el preparado farmacéutico, asegurando así que el preparado se mantiene estable y es apto para el consumo durante el tiempo especificado en su fecha de caducidad.

Evacuol es el nombre comercial de un preparado farmacéutico con propiedades laxantes que contiene picosulfato sódico (SPS) como principio activo, metilparabeno (MPB) como preservante y sorbitol como edulcorante. Está descrito en la bibliografía que los parabenos sufren una reacción de transesterificación en presencia de polioles, por lo que el metilparabeno se transforma en sorbitolparabeno (SPB) con el tiempo. A la vez, tanto metilparabeno como sorbitolparabenos pueden hidrolizarse para dar ácido p-hidroxibenzoico (PHBA). Los productos resultantes de estos procesos no poseen actividad preservante conocida, por lo que deben ser cuantificados para evaluar así el estado de conservación del preparado. Para llevar a cabo la determinación de los productos de degradación se requiere un método que permita separar todos los productos presentes en el preparado y cuantificarlos cada uno por separado, y que pueda también ser aplicado en el control de rutina para la determinación del contenido en principio activo. La separación de los productos de degradación es compleja debido a la gran semejanza de los mismos y la electroforesis capilar se ha propuesto por su alta capacidad de resolución como alternativa a los métodos cromatográficos habituales.

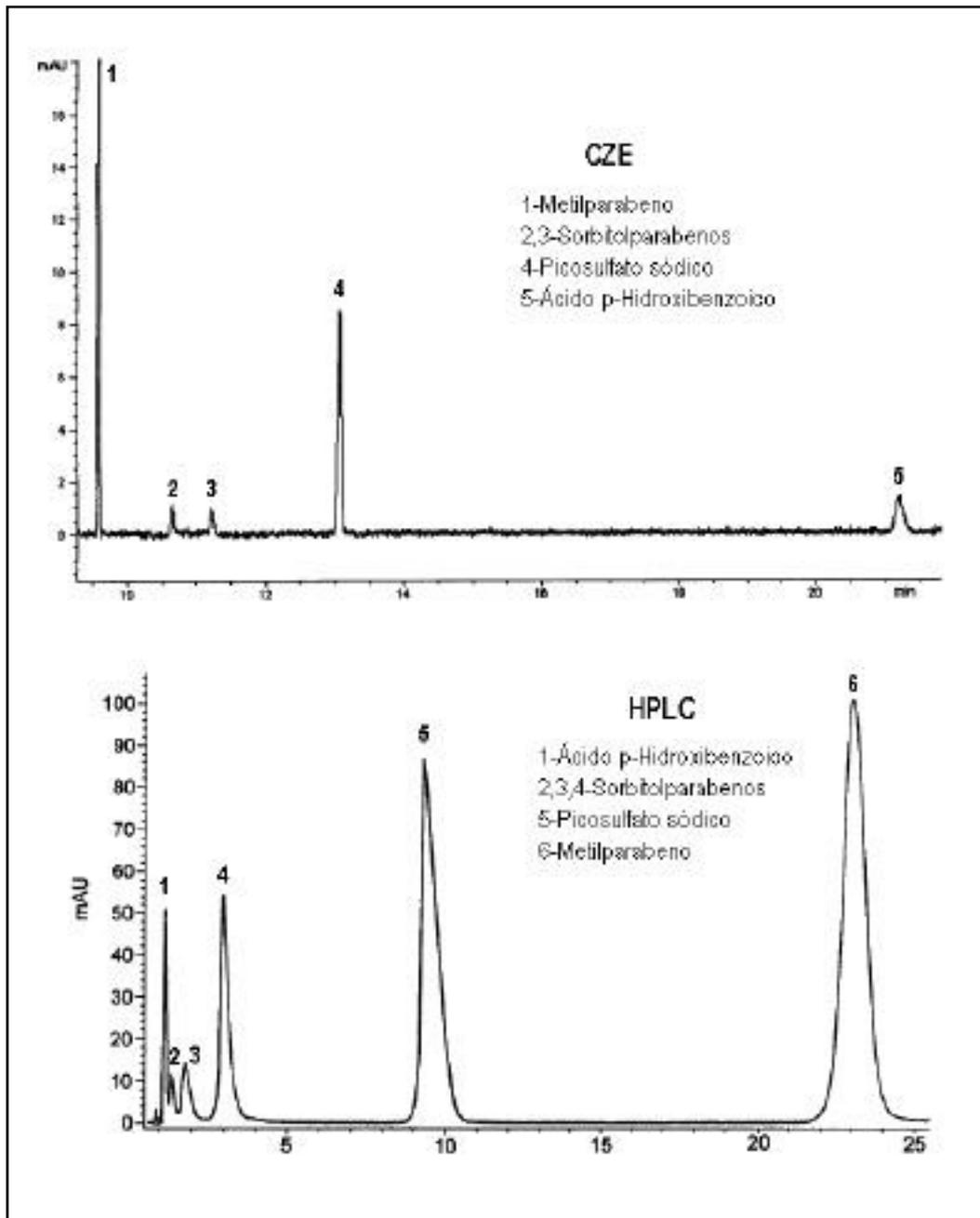
#### **3.1.1. Análisis de Evacuol por CZE**

En el trabajo presentado en el anexo I se ha desarrollado un método para la determinación del principio activo y la evaluación del estado de conservación a lo largo del tiempo del preparado farmacéutico Evacuol. Con este fin se ha utilizado la electroforesis

capilar, en el modo de aplicación de electroforesis capilar de zona (CZE) dado el carácter iónico del tipo de sustancias que componen el preparado farmacéutico. Se han hallado unas condiciones experimentales que permiten obtener una completa separación de los productos que conforman el fármaco. En la figura 1 se presentan de forma comparativa las separaciones obtenidas mediante CZE y HPLC. El mayor poder de resolución de la electroforesis capilar, junto con el distinto mecanismo de separación, favorece la separación completa de todos los componentes. La sensibilidad obtenida para todos los productos es suficiente para determinarlos con exactitud y precisión, incluso para los productos minoritarios.

El orden de aparición de los componentes en CZE concuerda con el esperado atendiendo a la relación carga radio electroforético de carga/radio. En polaridad directa (detector en el cátodo(-)), los cationes migran atraídos por el cátodo, apareciendo ordenados por relación carga/radio decreciente. En nuestro caso los productos a separar son aniones: MPB y SPB adquieren al pH de trabajo una carga negativa, mientras que SPS y PHBA adquieren dos cargas negativas. Los aniones migran en sentido contrario al detector, atraídos por el ánodo pero son detectados porque el flujo electroosmótico es mayor que la movilidad de los aniones y éstos son arrastrados hacia el cátodo (-). Por ello, aparecen ordenados inversamente a los cationes: los aniones de mayor tamaño y menos cargados aparecen los primeros (relación carga/radio creciente).

Según este mecanismo, el SPB debería alcanzar el detector antes que el MPB, puesto que tiene un mayor tamaño, pero en este caso llega después. Este hecho lo hemos atribuido a una posible interacción entre SPB y el tampón borato, que ha sido descrita en la bibliografía y que incluso ha sido utilizada por algunos autores para la separación de carbohidratos mediante electroforesis capilar. La formación del complejo borato-SPB aumenta la fracción de carga negativa del SPB, lo que explica que éste aparezca después del MPB. El hecho de que SPB aparezca formando dos picos, mientras que en HPLC aparece en formando tres picos es debido al distinto número y proporción de isómeros posicionales de SPB que se separan mediante una y otra técnica. Este hecho es abordado en el siguiente trabajo de electroforesis (anexo II), en el cual se realiza un estudio de los distintos posibles isómeros posicionales de SPB.



**Figura 1.** Comparativa entre la separación de los componentes del preparado farmacéutico mediante CZE y HPLC.

Condiciones CZE: tampón borato 20 mM pH=10, 22kV, 25 °C.

Condiciones HPLC: Na<sub>2</sub>SO<sub>4</sub> 0.05 M/ MeOH (70:30), 1 ml/min.

El último componente (PHBA) aparece en CZE a un tiempo de 22 minutos, que es un tiempo ligeramente inferior al que tardan en eluirse todos los componentes en HPLC. Se ha intentado optimizar el voltaje aplicado para disminuir este tiempo, pero voltajes superiores al utilizado (22 kV) no reducían significativamente el tiempo de migración y proporcionaban un empeoramiento de la relación señal/ruido. Para intentar disminuir el tiempo de análisis en CZE se puede variar la temperatura u otros parámetros (tampón, pH...). En este caso no se ha realizado ninguna de estas variaciones por considerar que la resolución era correcta y que los resultados eran aceptables para la finalidad de nuestro estudio. La posible implantación de este método en control de calidad, aconsejaría efectuar un diseño experimental para hallar las mejores condiciones de análisis, que permitieran reducir el tiempo de análisis, sin disminuir la calidad de los resultados.

Uno de los aspectos más destacables de la separación CZE en comparación con la separación cromatográfica es la diferencia que a simple vista se puede observar en las anchuras de los picos. En CE los picos son mucho más estrechos, lo que permite obtener mejor resolución y una eficacia muy superior. Como ejemplo, destacar que metilparabeno tiene una anchura de pico de aproximadamente 0.1 minuto en CZE mientras que en HPLC es de aproximadamente 2.5 minutos.

Al construir las curvas de calibración se obtienen buenas relaciones lineales para todos los componentes de la muestra, tanto para CZE como para HPLC, cuyos valores se han utilizado como método de referencia para la posterior validación del método electroforético y los límites de detección obtenidos son similares a los obtenidos por HPLC.

Se ha realizado la validación del método electroforético, siguiendo para ello las indicaciones de las ICH Guidelines para la validación de métodos de análisis. En concreto se han evaluado los parámetros de selectividad, linealidad, precisión y exactitud, obteniéndose resultados satisfactorios para todos ellos, y en concordancia con los obtenidos para el método cromatográfico. En la tabla 1 se recogen de forma resumida el procedimiento seguido y los resultados obtenidos. La validación del método corrobora su idoneidad como método para el control del preparado farmacéutico, útil tanto para la determinación del principio activo como para el control de la estabilidad del preparado.



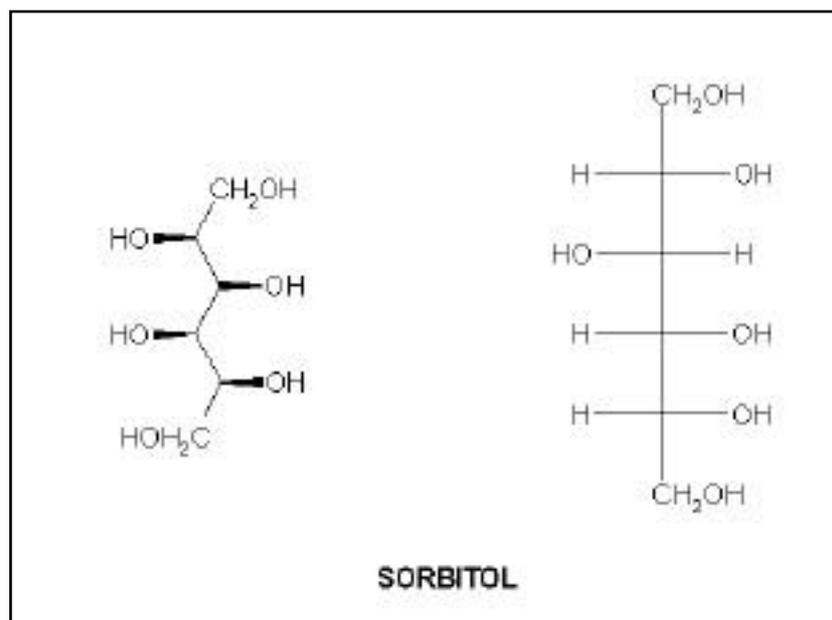
**Tabla 2.** Resultados de control de estabilidad de dos lotes del preparado Evacuol.

Lote Evacuol	Antigüedad	Analito	Valor CZE (mg/ml)	Valor HPLC (mg/ml)	% MPB Reaccionado
A	1 año	SPS	8.120	8.230	10 %
		MPB	2.046	2.066	
		PHBA	0.070	0.070	
		SPB	0.223	0.198	
B	2 años	SPS	8.030	7.960	16 %
		MPB	1.891	1.881	
		PHBA	0.076	0.078	
		SPB	0.421	0.398	

Con esto se demuestra que el método electroforético desarrollado permite llevar a cabo tanto la determinación cuantitativa del principio activo del preparado como el control de la estabilidad del mismo, de la misma forma que cualquier otro método cromatográfico. Sin embargo, la diferencia entre ambos métodos es muy clara con respecto al gasto de disolventes y generación de residuos: un cromatógrafo trabajando unas 8 horas diarias (1 ml/min) consume aproximadamente medio litro de disolventes mientras que el consumo en CZE era de aproximadamente 5 ml de tampón. Además, por el alto contenido de sorbitol que se adsorbe en las columnas HPLC, éstas requieren un mayor acondicionamiento para asegurar su correcto funcionamiento y evitar que se degraden, mientras que el sistema electroforético no se ve influido por este compuesto.

### 3.1.2. Separación de los Sorbitolparabenos

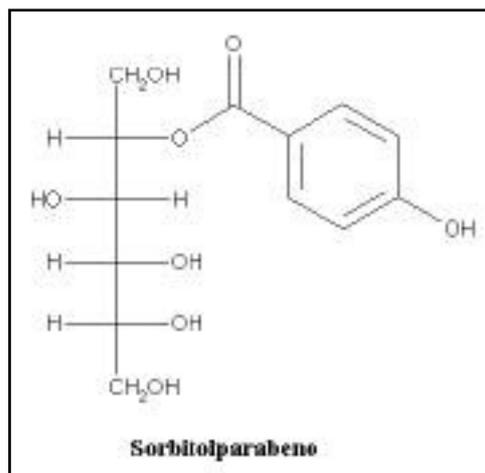
La reacción del metilparabeno en el fármaco Evacuol es la transesterificación del metilparabeno por sorbitol para formar un nuevo éster: p-hidroxibenzoato de sorbitol. El sorbitol es un C-6 poliol, cuya estructura puede observarse en la figura 2.



**Figura 2.** Estructura del C-6 polioli Sorbitol.

La disposición geométrica de los hidroxilos en cada uno de los carbonos quirales (carbonos 2, 3, 4 y 5) hace que no exista un plano de simetría en la molécula. Cuando se forma el transéster Sorbitolparabeno (SPB), el sorbitol puede sustituir cualquiera de los seis hidroxilos formando hasta seis isómeros posicionales. Puesto que no existe simetría en la molécula, los isómeros formados no son equivalentes dos a dos (isómeros 1-6, 2-5 y 3-4). La existencia de estos seis posibles isómeros, junto con los datos contradictorios que habíamos obtenido mediante HPLC y CZE, por los cuales obteníamos distinto número de picos y proporciones que no se correspondían, hizo que nos planteáramos estudiar la separación de los isómeros posicionales, de los cuales en la figura 3 se muestra uno de los posibles.

En el trabajo anterior (anexo I) habíamos visto que el sorbitolparabeno (SPB) se separaba en dos picos de relación 50:50 en CZE, y en tres picos de proporciones 5:20:75 en HPLC. Ninguna de las combinaciones posibles de los picos de HPLC se corresponde con los obtenidos en CZE. La mejor explicación es que existe un diferente mecanismo de separación en CZE, lo que podría explicarse por la interacción entre borato y sorbitolparabenos, que también es la responsable del retraso en la migración del SPB.



**Figura 3.** Uno de los 6 posibles isómeros posicionales de sorbitolparabeno.

Para potenciar la capacidad de separación utilizamos la cromatografía electrocinética micelar (MEKC), por ser un modo de aplicación de CE que asocia las características de la electroforesis capilar con las ventajas de la cromatografía líquida, puesto que un aditivo del tampón actúa como fase pseudoestacionaria hidrófoba, estableciéndose una interacción similar a la que se establece entre analitos y fase estacionaria de HPLC en fase reversa. En el trabajo presentado en el anexo II se ha realizado un estudio comparativo de la separación de los isómeros posicionales de sorbitolparabeno mediante HPLC, CZE y MEKC, intentando hallar una explicación a las distintas áreas relativas de los picos hallados por cada técnica.

En MEKC se utiliza como aditivo un tensoactivo capaz de formar micelas que migran por el capilar. Los analitos a separar migran por el capilar y van interaccionando con las micelas de forma distinta. Cuando las micelas son aniónicas, migran hacia el ánodo (+) por lo que aquellos analitos que interaccionan con la fase micelar son retrasados con respecto a aquellos que no lo hacen. Este mecanismo es similar al que se establece en HPLC de fase reversa, en el que las columnas están rellenas de una fase estacionaria apolar. Por ello en ocasiones se denomina al aditivo MEKC fase pseudoestacionaria.

El aditivo elegido ha sido el dodecilsulfato sódico (SDS). Para favorecer la interacción entre los sorbitolparabenos y las micelas se ha trabajado a un pH en el que los parabenos están en su forma neutra (pH=7). El ensayo de distintas proporciones de SDS ha revelado que, a medida que aumenta la proporción de tensoactivo, van apareciendo nuevos picos de SPB, atribuibles a isómeros posicionales distintos. La concentración de 100 mM, la máxima concentración posible por limitaciones operacionales, permite resolver en 5 picos los SPBs. En la figura 4 se muestran comparativamente los resultados obtenidos mediante

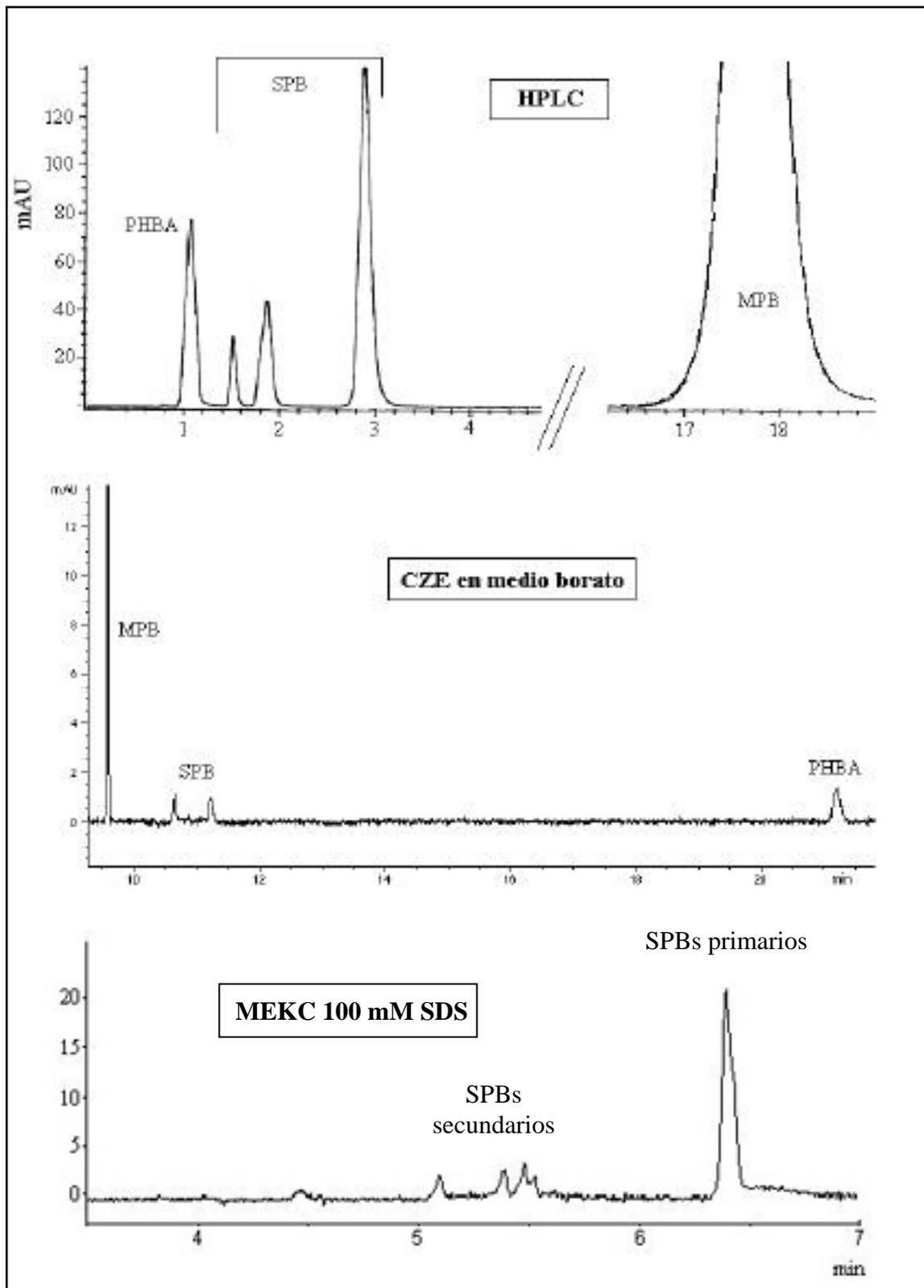
HPLC, CZE y MEKC. Se puede observar la similitud entre las separaciones HPLC y MEKC, obteniéndose mayor resolución de los SPBs por esta última.

El pico que aparece como mayoritario de SPB en MEKC supone aproximadamente el 75 % del total de SPBs, igual que el pico mayoritario que aparecía en HPLC. En otros trabajos que tratan la formación de estos transésteres se había identificado este pico mayoritario como la mezcla de isómeros primarios de SPB. Parece lógico pensar que por efectos estéricos se forman en mayor proporción los isómeros primarios, con los carbonos 1 y 6 del sorbitol. Las proporciones halladas en MEKC y la evolución que se ha observado al aumentar la concentración de tensoactivo hacen pensar que hemos separado los isómeros secundarios (los 4 posibles). No obstante, la máxima concentración de SDS no consigue la separación de los isómeros primarios y estos aparecen formando sólo un pico, aunque la forma asimétrica que adquiere al aumentar la concentración de SDS, parece indicar que se estuviera a punto de dividir en dos picos de los dos isómeros.

La proporción de las áreas de los picos obtenidos por CZE, HPLC y MEKC es distinta y difícil de relacionar; mediante cromatografía líquida preparativa, reproduciendo las condiciones descritas para la separación HPLC, se ha recogido la fracción correspondiente al pico mayoritario (mezcla de isómeros primarios) y se ha inyectado en CZE en medio borato con las condiciones descritas para el trabajo anterior.

Se separan dos picos a unos tiempos de migración similares a los obtenidos para SPB, pero ahora los picos no tienen áreas equivalentes, sino que están en la proporción de 64:36. Puesto que se había inyectado solamente la fracción correspondiente a la mezcla de isómeros primarios de SPB, la aparición de estos dos picos parece indicar que los isómeros primarios se separan en CZE al interactuar con el medio borato.

Utilizando MEKC y la combinación de dos técnicas cromatográficas hemos establecido la existencia de los seis isómeros posicionales de SPB. La técnica MEKC permite resolver con éxito todos los isómeros secundarios aunque no consigue resolver los dos primarios, demostrando ser una potente alternativa para la separación de moléculas neutras que, mediante la interacción con una fase pseudoestacionaria, permite obtener altas resoluciones.



**Figura 4.** Separación de isómeros posicionales de Sorbitolparabeno mediante HPLC (MeOH/H<sub>2</sub>O 30:70, 1 ml/min), CZE (Borato 20 mM, pH=10, 22 kV) y MEKC (Fosfato 25 mM, pH=7, 25 kV, 100 mM SDS).

### **3.2. MÉTODOS DE ANÁLISIS POR ESPECTROSCOPIA EN EL INFRARROJO CERCANO (NIR)**

En esta memoria se ha aplicado la espectroscopia NIR en el desarrollo de diversas metodologías de análisis, dirigidas al control de calidad de diversas fases de producción de preparados farmacéuticos, desde el control de las materias primas hasta el análisis del producto acabado. El orden seguido en todos los trabajos se corresponde a las distintas etapas de la construcción de un modelo de calibración NIR: preparación y selección de las muestras, obtención de los valores de referencia, registro de los espectros, pretratamiento de datos, construcción del modelo y validación del mismo.

#### **3.2.1. Bibliotecas NIR para la identificación de materias primas**

La identificación cada producto, ya sea principio activo o excipiente, es uno de los ensayos que debe realizarse durante el control de calidad farmacéutico. Los métodos oficiales de identificación son en ocasiones laboriosos, costosos y poco selectivos, por lo que resulta frecuente la aplicación de más de un método para identificar y/o caracterizar completamente un producto. El número de métodos propuestos por las farmacopeas para la identificación es elevado por lo que un laboratorio que deba realizar la identificación de muchos productos se ve obligado a formar a sus analistas en numerosas metodías de análisis y a que éstos tomen decisiones que en ocasiones vienen influidas por la presión que impone el proceso de producción.

En el trabajo presentado en el anexo III: **Bibliotecas NIR**, se recogen una serie de recomendaciones básicas para cualquier usuario de la técnica NIR que pretenda desarrollar una biblioteca para la identificación de productos farmacéuticos, que pueden extrapolarse a cualquier ámbito industrial. Estas recomendaciones incluyen consideraciones tanto de la fase de diseño como de la fase de construcción y validación de la biblioteca. Como ejemplo se exponen los resultados obtenidos en la construcción de una biblioteca NIR para la identificación de materias primas.

Una biblioteca de identificación debería contener la totalidad de los productos que maneja la empresa, lo que permitirá identificar todos los productos con un mismo método. Sin embargo, no resulta sencillo conseguir la total discriminación de todos ellos en una sola biblioteca por lo que en este trabajo se ha presentado una estrategia de construcción de bibliotecas en cascada, en la que una biblioteca general identifica el grueso de los productos, mientras que las sub-bibliotecas de clasificación distinguen los productos con espectros parecidos que no pueden diferenciarse en la biblioteca general. La potencialidad de esta

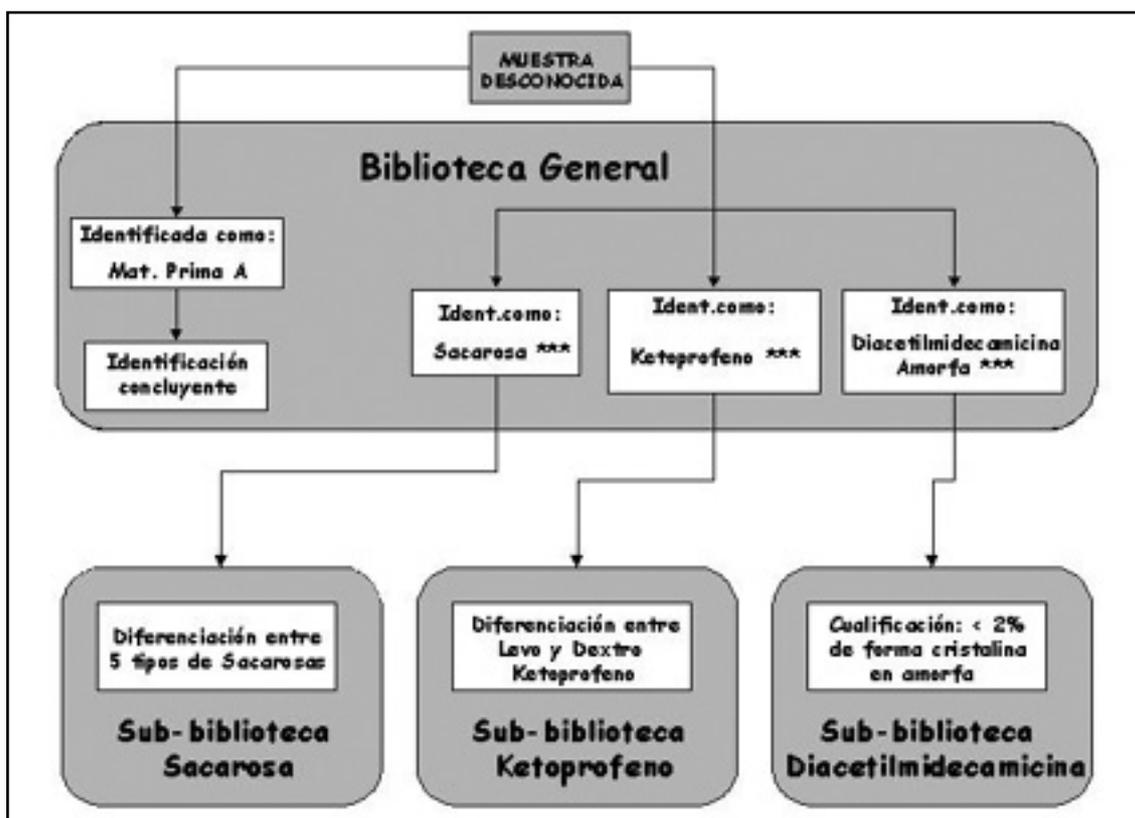
metodología se ha demostrado construyendo una biblioteca general que contiene varias situaciones que requieren de la construcción de sub-bibliotecas en cascada; para ilustrar su funcionamiento hemos seleccionado tres ejemplos.

La biblioteca general se ha construido utilizando el coeficiente de correlación como criterio diferenciador entre espectros. Este método resulta rápido, fiable y permite una fácil construcción y posterior ampliación de la biblioteca. Además, este criterio no requiere de un elevado número de espectros para definir cada producto y no es muy sensible a pequeñas variaciones instrumentales. En cambio, las sub-bibliotecas se construyen mediante la utilización de métodos quimiométricos de mayor poder discriminante como distancia de Mahalanobis, Varianza residual..., que permiten obtener un mayor grado de diferenciación.

Se ha construido una biblioteca para la identificación de 125 materias primas farmacéuticas. Esta biblioteca consigue identificar casi la totalidad de estas materias tan sólo registrando su espectro NIR y realizando el cálculo de identificación, lo cual no lleva más de 1 minuto por cada muestra. Esto supone un avance significativo en métodos rápidos de análisis cualitativo.

Para ilustrar el procedimiento de identificación en cascada se han elegido tres grupos de compuestos que han requerido de un segundo paso para obtener una completa identificación. Estos productos no son totalmente caracterizados en la biblioteca general por lo que es necesario construir una sub-biblioteca para cada uno de ellos. Esto suele ocurrir en casos en los que hay productos de una misma familia con distintas presentaciones, estructuras cristalinas, tamaños de partícula, etc. En la figura 5 se describe esquemáticamente el funcionamiento de la biblioteca y de estas tres sub-bibliotecas construidas.

La sub-biblioteca de sacarosa consigue diferenciar los cinco tamaños de partícula distintos de Sacarosa de forma concluyente, sin errores ni confusiones. Esta diferenciación resulta sencilla utilizando la distancia de Mahalanobis. En la tabla 3 se muestran los resultados que proporciona esta sub-biblioteca.



**Figura 5.** Esquema de funcionamiento de la biblioteca general y las tres sub-bibliotecas de cualificación construidas.

La sub-biblioteca de diacetilmidecamicina resuelve un problema de polimorfismo de esta sustancia. La diacetilmidecamicina se encuentra en dos formas en la biblioteca: en la forma amorfa, que es la de interés, y en la forma cristalina. Estas dos formas se diferencian claramente en la biblioteca general. Sin embargo, la forma amorfa puede estar contaminada con la forma cristalina en pequeña proporción. La sub-biblioteca construida permite diferenciar las muestras que tienen un contenido superior al 2% de la forma cristalina. En la tabla 3 se recogen resultados que muestran la diferenciación de muestras contaminadas con forma cristalina.

Por último, la sub-biblioteca de ketoprofeno se ha construido para resolver las confusiones aparecidas en la biblioteca general al identificar las formas dextro y levo ketoprofeno. La diferenciación de ketoprofeno racémico de los enantiómeros puros es posible en la biblioteca general, debido a que existen diferencias espectrales atribuibles a diferencias en las estructuras cristalinas de la forma racémica y dextro-levo. El fenómeno de polimorfismo entre formas racémicas y formas enantioméricamente puras está descrito, lo que nos hace pensar que es la causa de las diferencias espectrales. Sin embargo, también

hemos detectado pequeñas diferencias espectrales entre las formas cristalinas de *levo* y *dextro* ketoprofeno. Los enantiómeros no pueden ser diferenciados por espectroscopia NIR, pero las diferencias observadas no son atribuibles a ruido instrumental ni a distintos grados de pureza. Por ello pensamos que podían ser debidas a distintas formas de empaquetamiento y que podíamos aprovechar este fenómeno para diferenciar las formas *dextro* y *levo*. La sub-biblioteca que hemos construido permite diferenciarlos correctamente y se ha demostrado con diferentes muestras de ambas formas correspondientes a nuevos lotes de producción (Tabla 3). También se ha demostrado que es posible detectar la presencia de un enantiómero en el otro (impureza) por encima de un valor límite, recristalizando muestras de un enantiómero que contenían pequeñas cantidades del otro enantiómero. El hecho que se hayan podido diferenciar dos enantiómeros no quiere decir que cualquier par de enantiómeros pueda diferenciarse mediante espectroscopia NIR. En realidad con esta sub-biblioteca no se está efectuando una diferenciación enantiomérica sino más bien una discriminación polimórfica de las dos distintas formas cristalinas de los enantiómeros.

La biblioteca y sub-bibliotecas construidas son ejemplos claros del gran potencial para análisis cualitativo de la espectroscopia NIR. Además la metodología de bibliotecas en cascada permite llegar a resolver diferentes problemas que se presenten en la identificación. Mediante este procedimiento se pueden llegar a tener grandes bibliotecas de identificación, capaces de identificar una gran cantidad de productos distintos (más de 500) y resolviendo las confusiones que se den en la biblioteca general mediante sub-bibliotecas de clasificación adecuadas. Resulta por lo tanto un método idóneo para identificación farmacéutica con el consiguiente ahorro económico y de tiempo que resulta en comparación con los métodos tradicionales de identificación.

**Tabla 3.** Tabla de identificaciones cruzadas de las sub-bibliotecas construidas.

Sub-biblioteca	Muestra	Identificada como:	Resultado	Discriminante
<b>Sacarosa</b> Distancia Mahalanobis de Identificación positiva: Threshold < 0.89	Cristal	Cristal	Positivo	0.693
		Otras	Negativo	1.000
	Polvo A	Polvo A	Positivo	0.100
		Otras	Negativo	0.999-1.000
	Polvo B	Polvo B	Positivo	0.530
		Otras	Negativo	1.000
	Sémola	Sémola	Positivo	0.127
		Otras	Negativo	0.999-1.000
	Granulada	Granulada	Positivo	0.550
		Otras	Negativo	1.000
<b>Ketoprofeno</b> Varianza Residual Identificación positiva: Threshold < 0.87	Dexketoprofeno (5 muestras)	Dexketoprofeno	Positivo	0.216-0.510
		Levoketoprofeno	Negativo	0.999-1.000
	Levoketoprofeno (5 muestras)	Levoketoprofeno	Positivo	0.326-0.611
		Dexketoprofeno	Negativo	1.000
<b>Diacetilmidecamicina</b> Varianza Residual Identificación positiva: Threshold < 0.88	+ 0.1 % Cristalina	< 2 % cristalina	Positivo	0.608
	+ 0.3 % Cristalina	< 2 % cristalina	Positivo	0.724
	+ 0.6 % Cristalina	< 2 % cristalina	Positivo	0.686
	+1.5 % Cristalina	< 2 % cristalina	Positivo	0.789
	+2 % Cristalina	< 2 % cristalina	Positivo	0.796
	+3 % Cristalina	< 2 % cristalina	Negativo	0.932
	+4 % Cristalina	< 2 % cristalina	Negativo	0.975
	+6 % Cristalina	< 2 % cristalina	Negativo	1.000
	+8 % Cristalina	< 2 % cristalina	Negativo	1.000
	+10 % Cristalina	< 2 % cristalina	Negativo	1.000

### 3.2.2. Estrategias de calibración NIR

Los espectros NIR incorporan información de la muestra, tanto química como física y la contribución de las características físicas de la muestra puede resultar de utilidad en la determinación de alguna propiedad. Sin embargo, esta contribución introduce nuevas fuentes de variabilidad que conducen a modelos de calibración más complejos y menos robustos. Para solventar este problema, se suelen aplicar pretratamientos espectrales a los espectros NIR que reducen las contribuciones indeseables del espectro, consiguiendo así modelos de

calibración más simples. Existen una serie de pretratamientos que podrían denominarse como estándares, porque suelen ser los más utilizados: derivadas, SNV, MSC, además de aparecer nuevos pretratamientos a menudo.

Para construir un modelo de calibración con buena capacidad predictiva es necesario disponer de un conjunto representativo de muestras que incorpore ambos tipos de información: física y química de la muestra. Las muestras de producción poseen una pequeña variabilidad de concentración, por lo que se deben seguir otras metodologías para obtener un conjunto representativo de muestras con un intervalo de la propiedad a determinar adecuado. El procedimiento más sencillo es preparar muestras por pesada y mezcla de los componentes hasta conseguir extender el intervalo de concentraciones hasta el deseado (muestras sintéticas). Mediante este procedimiento las muestras pueden presentar diferencias espectrales importantes con respecto a las de producción debido al diferente modo de preparación, diferencias que se intentan resolver mediante la aplicación de pretratamientos espectrales. Otro de los procedimientos más utilizados es dopar muestras de producción con excipientes o principio activo para conseguir muestras, por debajo o por encima respectivamente, del contenido nominal en principio activo (muestras dopadas).

A continuación se exponen los resultados obtenidos en dos trabajos que abordan la construcción de modelos de calibración para la determinación de nimesulide, principio activo de un granulado con propiedades antiinflamatorias, desde posiciones distintas. En el trabajo presentado en el anexo IV se comparan dos procedimientos distintos en la construcción y validación de un método de análisis para la determinación de un principio activo en un granulado farmacéutico. Se muestran los resultados obtenidos mediante ambos procedimientos tanto en la calibración como en la validación y además se exponen las diferencias observadas al predecir muestras de un tipo con el modelo construido con las otras muestras, debido a las diferencias espectrales observadas entre ambos tipos de muestras. A continuación, en el trabajo presentado en el anexo V, se intentan reducir esas diferencias espectrales entre muestras mediante la aplicación de un nuevo pretratamiento espectral con la finalidad de obtener modelos de calibración más sencillos y robustos. Los resultados de aplicar este nuevo pretratamiento se comparan con los obtenidos mediante otros pretratamientos espectrales.

Para la construcción de los modelos de calibración (Anexo IV) se disponen de 3 tipos de muestras: muestras sintéticas, muestras dopadas y muestras de producción, estas últimas suministradas por la empresa fabricante del preparado. Las muestras dopadas y sintéticas se dividen para formar dos conjuntos, de calibración y validación para cada tipo de muestra. A

cada conjunto, se le añaden algunas muestras de producción, para introducir la variabilidad del proceso de fabricación. Se ha añadido el mismo conjunto de muestras de producción a los dos conjuntos de calibración, y otro conjunto a los de predicción. Los modelos con muestras de laboratorio y con muestras dopadas que proporcionan los mejores resultados son similares, tanto en pretratamiento como en número de factores PLS. Ambos modelos, el de dopados y el de sintéticos poseen similares capacidades predictivas, siendo ligeramente mejor el modelo construido con muestras dopadas (Tabla 4).

**Tabla 4.** Comparación de los modelos de calibración construidos con muestras sintéticas y con muestras dopadas.

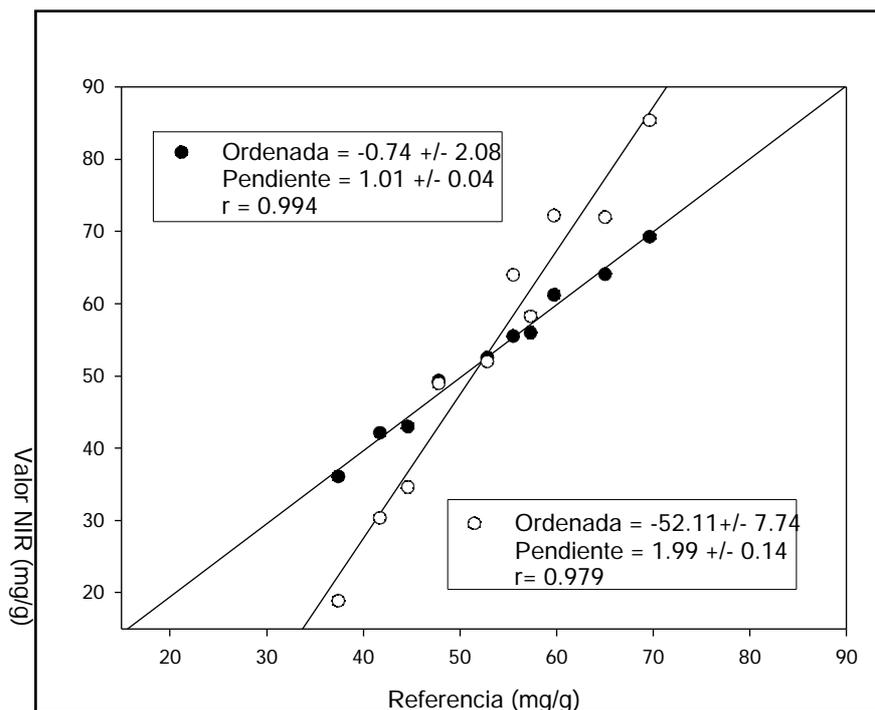
	<b>Modelo Muestras sintéticas</b>	<b>Modelo Muestras Dopadas</b>
<b>Intervalo Espectral</b>	1100-2200 nm	1100-2200 nm
<b>Pretratamiento</b>	1a derivada	1a derivada
<b>Factores PLS</b>	4	4
<b>RMSEP (C)</b>	1.4	0.9
<b>RMSEP (P)</b>	1.7	1.0

La validación de ambos modelos se ha realizado para determinar si alguno de ellos era claramente mejor que el otro. Sin embargo la validación muestra buenos resultados para ambos en todos los parámetros evaluados. Sólo en los parámetros relacionados con la precisión se obtiene una menor dispersión de los resultados para el modelo de dopados (Tabla 5). En vista de los resultados se puede concluir que ambos modelos son adecuados para su uso en control de rutina para la determinación del principio activo en el preparado farmacéutico. La consecuencia directa de estos resultados es que no hay argumentos claros para decantarse a favor de uno o de otro. Para este caso concreto ambas estrategias de calibración serían correctas y proporcionarían resultados aceptables.

**Tabla 5.** Resumen de la validación de los modelos de calibración construidos con muestras sintéticas y con muestras dopadas.

<b>Parámetros</b>	<b>Procedimiento</b>	<b>Modelo Sintéticos</b>	<b>Modelo Dopados</b>
<b>Linealidad</b>	Valor NIR= a + b·ValorRef (a = 0 ; b = 1)	b = 1.02 ± 0.06 a= -2.18 ± 2.83 corr. = 0.999	b = 1.01 ± 0.09 a= -0.75 ± 4.81 corr. = 0.994
<b>Exactitud</b>	Test t de diferencias entre valores NIR y valores referencia de 15 lotes de producción	Media dif.= -0.77 Desv. Est = 3.54 t exp = 1.6 t crítica = 2.14	Media dif.= -0.37 Desv Est.= 1.63 t exp = 1.13 t crítica = 2.14
<b>Repetitividad</b>	Una muestra analizada 6 veces por el mismo operador	$\bar{x}$ = 50.57 Desv. Est= 1.41 % C. V. = 2.82	$\bar{x}$ = 50.54 Desv Est = 0.37 % C. V. = 0.72
<b>Precisión Intermedia</b>	Muestra analizada 3 días por 2 operadores distintos	$\bar{x}$ = 52.65 Desv Est = 2.16 % C. V. = 4.10	$\bar{x}$ = 51.50 Desv Est.= 0.68 % C. V. = 1.33
<b>Robustez</b>	Comparación entre valores NIR y de referencia en 30 muestras a lo largo de un año y medio. Test t de diferencias.	Media dif.= 0.2 Desv. Est.= 2.8 t exp = 0.64 t crítica = 2.03	Media dif.= 0.15 Desv Est.= 1.39 t exp = 0.91 t crítica = 2.03

Durante la etapa de validación del modelo construido con muestras sintéticas, al plantear cómo evaluar la linealidad del modelo, pensamos en preparar muestras que extendieran el rango de concentraciones, cuya composición fuera lo más similar posible a las de producción. Las muestras que mejor cumplen esta premisa son las muestras dopadas. No obstante descubrimos que, debido a las diferencias espectrales entre muestras sintéticas y de laboratorio, al predecir muestras de un tipo con el modelo construido con las muestras del otro tipo se obtenían resultados incorrectos: la recta obtenida al representar valores NIR vs valores de referencia se desvía de la que debiera ser cuando la predicción es cruzada (muestras dopadas con modelo sintéticos y viceversa). En la figura 6 se muestra cómo el modelo de dopados cuantifica correctamente las muestras dopadas, no así el modelo de sintéticos.



**Figura 6.** Muestras dopadas predichas con el modelo dopados (•) y con el modelo sintéticos (○) .

Para reducir estas diferencias se han aplicado distintos pretratamientos espectrales, modelos conjuntos con ambos tipos de muestras e incluso construyendo modelos MLR sobre los scores de distintas combinaciones de componentes principales, intentando eliminar aquellos que no contengan información de concentración. Sin embargo, ninguno de los procedimientos ha demostrado ser eficaz para eliminar estas diferencias. Así pues, si estas diferencias existen y a priori no pueden ser eliminadas mediante ningún procedimiento convencional, es importante detectar la existencia de éstas y evaluar su magnitud, siempre que no interfieran en la capacidad predictiva de los modelos construidos para muestras de producción. Si hay diferencias entre muestras dopadas o sintéticas, se debe comprobar que ambos procedimientos son igualmente efectivos, elegir uno de ellos y utilizar siempre el mismo tipo de muestras para realizar calibración y validación. Esta elección debe estar basada en los resultados obtenidos que en cada caso serán distintos. En base a lo obtenido en este trabajo y en otros realizados en nuestro grupo de investigación, las muestras sintéticas no consiguen incorporar la variabilidad de la producción, no así las muestras dopadas. Dependiendo de la magnitud

de las diferencias espectrales, los modelos obtenidos mediante uno y otro tipo de muestras tendrán capacidades predictivas más o menos parecidas.

En esta línea, el trabajo presentado en el anexo V aborda la problemática de resolver diferencias espectrales entre muestras según su origen, utilizando para ello un nuevo pretratamiento espectral: la corrección ortogonal de la señal (OSC de *Orthogonal Signal Correction*) en el desarrollo del método de calibración.

El OSC consiste en modelar aquella información del espectro no correlacionada con la propiedad a determinar para después eliminarla y poder construir el modelo de calibración. Se puede considerar este procedimiento como dos calibraciones consecutivas, en la primera de las cuales se modela la información ortogonal con la propiedad a determinar, se elimina, y a continuación se modela la propiedad. El OSC funciona de la misma forma que lo hace el tratamiento PLS: descomponiendo la información en factores o variables latentes (similares a los componentes principales del PCA). Estos componentes se ortogonalizan con la matriz de concentraciones, de manera que contienen todo lo que no se corresponde a información de concentración. Una vez finalizada la descomposición, esta información es eliminada de los espectros originales y los espectros así corregidos se utilizan para la calibración PLS.

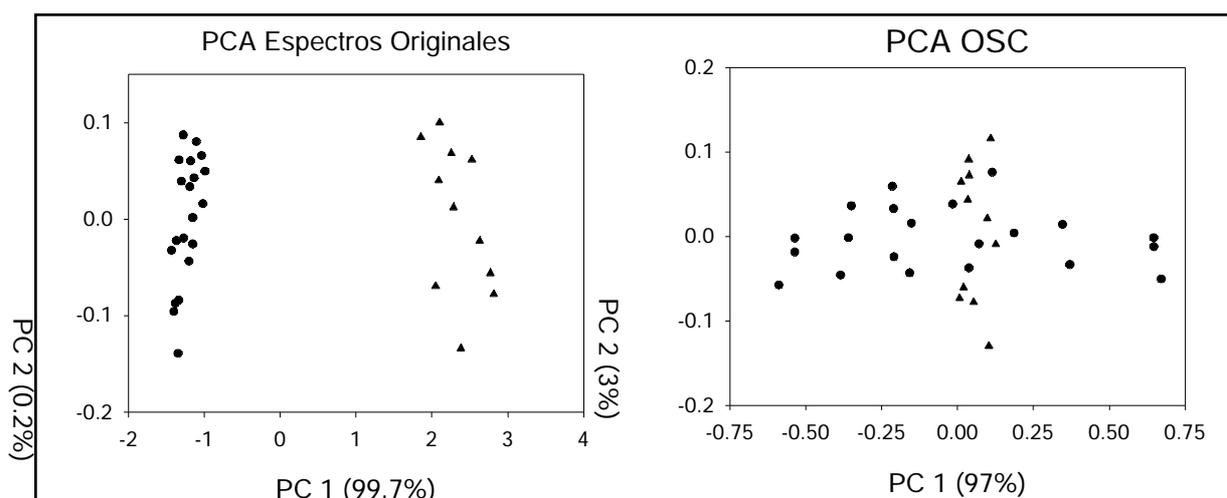
Se han utilizado dos tipos de muestras: muestras sintéticas y muestras de producción para ilustrar las diferencias existentes entre ellas y cómo OSC es capaz de corregirlas, mejorando la calidad de los modelos de calibración. Para efectuar una corrección efectiva, es necesario que el conjunto de “entrenamiento” del cálculo OSC incorpore muestras de ambos tipos, para detectar la fuente de variabilidad debida al origen de las muestras. Este es un proceso laborioso, en el que se ensayan distintos conjuntos, variando la cantidad de muestras de cada origen. Una vez halladas las condiciones que han proporcionado los mejores resultados, se procedió a la corrección de los espectros y a la posterior construcción del modelo de calibración. Al comparar los resultados obtenidos con los diferentes pretratamientos se puede comprobar que utilizando el método de corrección OSC se consiguen unos resultados similares a los obtenidos utilizando SNV y MSC, que proporcionan los menores errores de calibración-predicción. Sin embargo, el tratamiento OSC elimina tal cantidad de información no relacionada con la concentración que permite obtener un modelo PLS con sólo 1 factor, mientras que los otros pretratamientos requieren 5 factores PLS (Tabla 6). El OSC elimina toda aquella información que no es atribuible a concentración, incluyendo las

diferencias espectrales entre muestras de producción y muestras sintéticas, diferencias que permanecen al aplicar los otros pretratamientos.

**Tabla 6.** Resultados de Calibración-Predicción para los mismos conjuntos de muestras aplicando distintos pretratamientos espectrales.

Modelo	Factores PLS	% RSEC	% RSEP
Absorbancia	7	2.7	5.6
1ª Derivada	3	4.1	5.3
SNV	5	2.4	2.8
MSC	5	2.4	3.4
<b>OSC</b>	<b>1</b>	<b>3.4</b>	<b>2.9</b>

Incluso un PCA de los espectros corregidos mediante OSC muestra que no existe segregación entre los dos tipos de muestras, no así al aplicar SNV, MSC y derivadas, donde se forman dos sub- grupos correspondientes a muestras de producción y sintéticas (Figura 7).



**Figura 7.** Representación del 1er score frente al 2º de los análisis en componentes principales de los espectros originales y corregidos mediante OSC. Muestras sintéticas: ● y muestras producción: ▲.

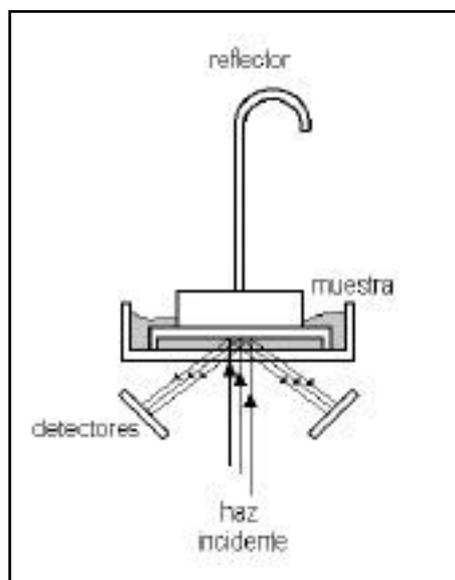
Los espectros corregidos mediante OSC permiten incluso realizar una calibración univariante a la longitud de onda de máxima correlación con la concentración (2172nm) obteniéndose buena capacidad predictiva para muestras externas.

Este pretratamiento puede ser aplicado a diversos problemas que requieran de una corrección espectral y en cada caso se deben optimizar un buen número de parámetros, puesto que la corrección no es directa y hay que efectuar una etapa de optimización de los parámetros necesarios para el cálculo OSC. OSC se encuentra todavía en el periodo inicial de aplicación y todavía queda un camino por recorrer para demostrar todo su potencial y facilitar su aplicación práctica.

### **3.2.3. Determinación NIR del principio activo de un hidrogel**

Una de las ventajas de la espectroscopia NIR es que permite la medida directa sobre el producto a analizar. La mayoría de productos farmacéuticos son sólidos (granulados, polvos, comprimidos) por lo que su medida directa es viable. También existen módulos especialmente diseñados para la medida directa de líquidos. Sin embargo, cuando el preparado es un gel, su medida directa no resulta tan sencilla. Registrar un espectro de reflectancia no es viable ya que no permite fijar el camino óptico, mientras que el registro por transmitancia requiere introducir la muestra en cubetas, lo que resulta difícil, poco práctico y laborioso a la hora de limpiarlas. En el trabajo presentado en el anexo VI, se propone la utilización del modo de transflectancia, en el desarrollo y validación de un método NIR para la determinación del contenido en principio activo dexketoprofeno en un gel. Para ello se utiliza un accesorio reflector (figura 8) el cual permite realizar medidas con los módulos habituales de reflectancia.

Uno de los mayores inconvenientes en el desarrollo del método viene dado porque el gel contiene un 60 % de agua y un 35 % de etanol, compuestos altamente absorbentes en NIR, además del bajo contenido (1.25 %) en principio activo Dexketoprofeno (DKP) 1.25 %, valor cercano al considerado como límite de cuantificación en NIR (habitualmente situado entorno al 1%).



**Figura 8.** Accesorio reflector para medidas de transfectancia NIR

Para la construcción del modelo de calibración se han preparado una serie de muestras en el laboratorio a partir de muestras de producción del gel, las cuales han sido sobre y sub-dosificadas para ampliar el intervalo de concentraciones. Los mejores resultados se han obtenido eliminando aquellas zonas del espectro donde el placebo (agua-etanol) presentaba una fuerte absorción y restringiendo el intervalo a aquellas zonas donde el dexketoprofeno presenta mayor señal. En la tabla 7 se muestran las características del modelo de calibración construido.

**Tabla 7.** Parámetros de construcción del modelo de calibración PLS para la determinación del DKP en el gel.

<b>Intervalo espectral</b>	1600-1800 nm
<b>Pretratamiento espectral</b>	2 <sup>a</sup> derivada
<b>Algoritmo de cálculo</b>	PLS1
<b>Nº de factores PLS</b>	3
<b>RMSEP Calibración</b>	0.08 mg /g
<b>RMSEP Predicción</b>	0.25 mg /g

De la misma forma que para los demás métodos NIR validados se han seguido las indicaciones de las normas ICH para validación de métodos analíticos de análisis, adaptando a la técnica aquellos puntos que difieren de los métodos tradicionales de análisis. Se han evaluado los parámetros de selectividad, linealidad, repetitividad, precisión intermedia, exactitud y robustez obteniéndose buenos resultados para todos ellos, tal y como se muestra en la tabla 8, demostrando así la idoneidad del método NIR como método de análisis alternativo a los habituales para la determinación del principio activo DKP en el gel.

**Tabla 8.** Validación del modelo PLS para la determinación de DKP en el gel.

<b>Parámetros</b>	<b>Procedimiento</b>	<b>Resultados</b>
Selectividad	Identificación en una biblioteca que contiene espectros del gel, del placebo y de los componentes mayoritarios.	Biblioteca: Varianza residual 1100-2500 nm 2ª derivada Threshold: 0.84  Diferenciación completa de la muestra de los demás componentes de la biblioteca
Linealidad	Valor NIR= a + b Valor Ref.	Pendiente (b)= 0.97 ± 0.05 Ordenada (a) = 0.11 ± 0.60 r = 0.988
Exactitud	Test t de diferencias entre valores NIR y valores de referencia de 9 muestras de producción	Diferencia media = 0.12 Desv.Est. Dif.= 0.19 t experimental = 1.89 t teórica = 2.31
Repetitividad	6 determinaciones de la misma muestra realizadas por el mismo operador	Promedio = 12.21 Desv. Est. = 0.08 % C.V. = 0.7 %
Precisión Intermedia	1 muestra analizada por 2 operadores en 3 días distintos	Promedio = 12.32 Desv. Est. = 0.14 % C.V. = 1.1 %
Robustez	Seguimiento de muestras de producción durante un año. Comparación entre valor NIR vs valor de referencia	Error relativo = 1.6 %