4

QUIMIOMETRIA

4.1 INTRODUCCIÓN 53

4.2 ETAPAS DEL PROCESO DE DESARROLLO DE MODELOS 54

4.3 HERRAMIENTAS QUIMIOMÉTRICAS MULTIVARIABLES 56

4.3.1 Regresión Lineal Multiple (MLR) 56

4.3.2 Análisis en Componentes Principales (PCA) 57

4.3.3 Regresión en Componentes Principales (PCR) 63

4.3.4 Regresión Parcial en Mínimos Cuadrados (PLS) 64

4.3.5 Métodos de Resolución 68

4.3.6 Métodos Empíricos de Modelado 74

4.3.7 Uso conjunto de Modelos Empírico y de MCR-ALS 79

4.3.8 Análisis Factorial Paralelo (PARAFAC) 81

4.3.9 Uso conjunto de modelos PARAFAC y MLR 82

4.1 INTRODUCCIÓN

La definición del término quimiometría es un tema de discusión abierto y no existe un consenso unánime sobre el mismo, a pesar de que existen dos publicaciones científicas internacionales y numerosas sociedades tanto nacionales como internacionales que usan esta palabra en sus títulos.

Svante Wold empleó la palabra "chemometrics" en 1972 para describir la disciplina de extraer información química relevante partiendo de un sistema químico experimental [1]. En 1995 trató de redefinir su significado:

"*La quimiometría responde a cómo extraer información química relevante a partir de un conjunto de medidas químicas, cómo representar, visualizar e interpretar dicha información y, por último, cómo hacer útil dicha información*" [2].

Aunque, desde mi punto de vista, la definición más completa y precisa se encuentra en el libro de Massart et al. [3], 1997.

"*La quimiometría es la parte de la química que se sirve de las matemáticas, estadística y lógica formal para: diseñar o seleccionar procedimientos experimentales óptimos; proporcionar información química relevante a partir del análisis de señales analíticas y, finalmente, adquirir conocimiento de los sistemas químicos*".

Esta última definición es muy parecida a la formulada por Svante Wold y Bruce Kowalski cuando fundaron la primera "Sociedad de Quimiometría" en 1974.

La quimiometría abarca una gran cantidad de herramientas de análisis, útiles en áreas tales como: el tratamiento de señales [4], la calibración

---

[1] Wold S. "Spline Funcions, a new tool in data-analyis", Kemisk Tidskrift, **1972**, 84(3), 34-37.

[2] Wold S. "Chemometrics; What do we mean with it, and what do we want from it?", Intell. Chem. Lab. Syst., **1995**, 30(1), 109-115.

[3] Massart, D.L., Vandeginste B.G.M., Buydens L.C.M., De Jong, S., Lewi, P.J., Smeyers-Verbeke J., "Handbook of Chemometrics and Qualimetrics: Part A", **1997**, 1st ed. Elsevier.

[4] Chaminade P., Baillet A., Ferrier D., "Data treatment in near infrared spectroscopy", Analusis, **1998**, 26(4), M33-M38.

multivariable [5-7], la resolución y modelado de datos [8-9], el reconocimiento de pautas [10-12] y la monitorización y el control de procesos.

4.2 ETAPAS DEL PROCESO DE DESARROLLO DE MODELOS MULTIVARIABLES

Un modelo multivariable es aquel en el cual se relaciona varias variables (por ejemplo, un espectro NIR) con propiedades de uno o varios analitos de una muestra (por ejemplo, la concentración).

El principal objetivo de los métodos multivariables es establecer modelos que sean capaces de predecir propiedades de nuevas muestras. Para que estas predicciones sean fiables se han de establecer modelos robustos, y para ello se deben seguir las etapas que se describen a continuación:

- **Selección del conjunto de calibración.** Seleccionar un conjunto limitado de muestras que debe ser representativo de toda la variabilidad química y física que pueda encontrarse dentro de la población no es una etapa trivial, especialmente cuando la variabilidad poblacional es múltiple y diversa, por ejemplo en productos naturales. Existen numerosos métodos de selección de muestras que buscan la

---

[5] Martens H., "Multivariate calibration - Direct and indirect regression methodology – Discussion and comments", Scandinavian Journal of Statistics, **1999**, 26, 193-196.

[6] Smilde A K, Tauler R, Saurina J, Bro R. "Calibration methods for complex second-order data". Anal. Chim. Acta, **1999**; 398, 237-251.

[7] Gemperline P. J. "Developments in Nonlinear Multivariate Calibration". Chemometrics Intell.Lab. Syst., **1992**, 15, 115-126.

[8] Jaumot J., Gargallo R., de Juan A., et al. "A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB" Chem. Intell. Lab. Syst., **2005**, 76 (1), 101-110.

[9] Gemperline P.J., Cash E. "Advantages of Soft versus Hard Constraints in Self-Modeling Curve Resolution Problems. Alternating Least Squares with Penalty Functions", Anal. Chem., **2003**; 75(16), 4236 – 4243.

[10] Forina M., Armanino C., Raggio V. "Clustering with dendrograms on interpretation variables", Anal. Chim. Acta, **2002**, 454(1), 4.

[11] Albano C., Dunn W. J., Edlund U., Johansson E., Nordén B., Sjöström M., Wold S., "Four levels of pattern recognition", Anal. Chim. Acta, **1978**, 103, 429-443.

[12] Lindon J.C., Colmes E., Nicholson J.K., "Pattern recoginition methods and applications in biomedical magnetic resonante". Progr. in Nucl Magn. Reson. Spectrosc. **2001**, 39, 1-40.

parsimonia [13] en conjuntos de calibración y ésta se consigue no aumentando el conjunto de calibración con nuevas muestras, sino a través de una selección adecuada de las mismas [14].

- **Métodos de referencia.** A través de ellos se determinan las concentraciones o propiedades de las muestras mediante los métodos analíticos pertinentes, que deben proporcionar valores precisos y exactos, ya que de ellos dependerá la exactitud del modelo multivariable obtenido.
- **Obtención de la señal analítica.** En esta memoria la información que va a ser utilizada es el espectro NIR de las muestras, el cual contiene implícitamente la información química deseada.
- **Pretratamiento de los datos.** En esta etapa se minimizan las contribuciones no deseadas, presentes en la señal analítica, que disminuyen la reproducibilidad y pueden provocar no linealidades u otros efectos que darían lugar a estimaciones menos sólidas.
- **Construcción del modelo.** A través de la herramienta quimiométrica elegida, pero buscando siempre seleccionar un modelo que establezca la relación más simple posible entre la propiedad a determinar y la señal analítica (parsimonia).
- **Validación del modelo.** Aplicación del modelo establecido a un número de muestras de las que se conoce la propiedad a determinar y que no han sido utilizadas en la etapa de construcción. De esta manera se verifica que el modelo construido constituye una correcta descripción del conjunto de datos experimentales.
- **Predicción de nuevas muestras.** Una vez construido (calibrado) y validado el modelo, éste puede ser aplicado en la predicción de nuevas muestras.

---

[13] Seasholtz M.B., Kowalski B. "The parsimony principle applied to multivariate calibration", Anal. Chim. Acta, **1993**, 277, 165-177.

[14] Cruz S.C., Rothenberg G., Westerhuis J.A., "Tackling calibration problems of spectroscopic analysis in high-throughput experimentation", Anal. Chem., **2005**, 77 (7), 2227-2234.

4.3 HERRAMIENTAS QUIMIOMÉTRICAS MULTIVARIABLES

Proporcionar una descripción completa de métodos multivariables de análisis es una tarea ardua, ya que la quimiometría está en constante desarrollo buscando el pragmatismo a través de la extracción de información analítica útil. Sin embargo, existen buenos libros de textos donde se recopilan de manera exhaustiva los algoritmos quimiométricos más extendidos, utilizados y contrastados [3], [15 - 20].

En este capítulo se presenta brevemente la base teórica de las principales herramientas utilizadas en esta memoria.

4.3.1 REGRESIÓN LINEAL MULTIPLE (MLR)

La regresión lineal múltiple es expresada en notación matricial como:

$$\mathbf{Y=XB+E} \qquad \textbf{[4.1]}$$

En la figura 4.1 se representa gráficamente la expresión matricial **[4.1]**, donde **Y** es la matriz de datos de referencia de la propiedad analítica que se quiere modelar, **X** es la matriz de datos espectrales y **B** es la matriz de parámetros de regresión estimados. Las dimensiones de las matrices hacen referencia a **m** el número de muestras, **p** el número de propiedades analíticas estudiadas o variables dependientes y **n** el número de canales, longitudes de onda o variables independientes. Desde un punto de vista formal, MLR es la generalización del problema de mínimos cuadrados univariantes o regresión lineal simple (**n**=1; **p**=1) a un estado multivariable en el cual **p**≥1, **n**>1.

---

[15] Sharaf M.A., Illman D.L., Kowalski B.R., "Chemometrics", **1986**, Ed. John Wiley & Sons, (New York).

[16] Martens H. and Næs T., "Multivariate calibration", **1989**, Ed. John Wiley & Sons, (England).

[17] Kramer R., "Chemometrics Technics for Quantitative Analysis", **1998**, Ed. Marcel Dekker (New York).

[18] Otto M., "Chemometris. Statistics and Computer Application in Analytical Chemistry", **1999**, Ed. Wiley-VCH. (New-York).

[19] Tauler R. "Anàlisi de Mescles Mitjançant Resolució Multivariant de Corbes", **1997**, Ed. Institut d'Estudis Catalans.

[20] Smilde A., Bro R., Geladi P. "Multi-way Analysis with Applications in the Chemical Sciences", **2004**, Ed. John Wiley & Sons, (England).

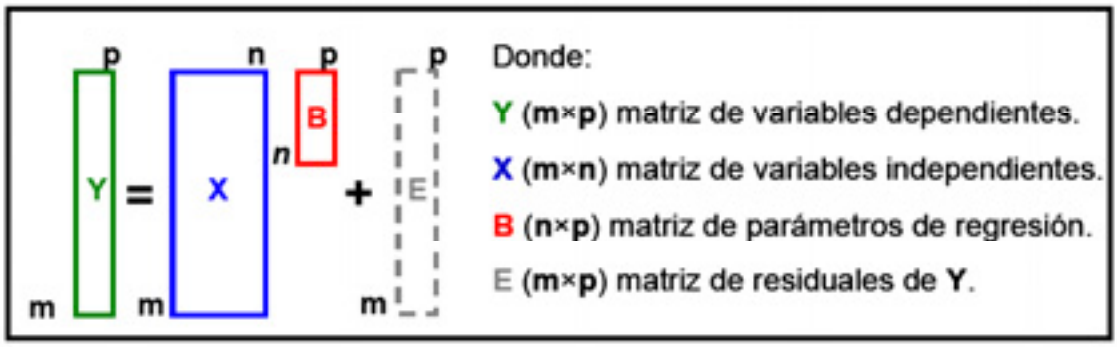


Figura 4.1. Representación gráfica de la expresión matricial de la regresión lineal múltiple.

La solución de mínimos cuadrados a la ecuación **[4.1]** viene dada por la expresión **[4.2]**:

$$\mathbf{B}=(\mathbf{X}^{T}\mathbf{X})^{-1}\mathbf{X}^{T}\mathbf{Y} \qquad [4.2]$$

MLR proporciona los mejores parámetros lineales insesgados (BLUE, *Best Linear Umbiased Estimator*). Sin embargo, este método tiene dos importantes restricciones:

- El número de muestras debe ser superior al número de canales o longitudes de onda empleado.
- La información espectral no debe estar correlacionada ya que la matriz $\mathbf{X}^{T}\mathbf{X}$ sería singular y su inversa inestable. Matemáticamente, se dice que el problema está mal condicionado (*ill-conditioned*).

Al trabajar con espectroscopia NIR nos encontramos con estás dos restricciones, usualmente el número de muestras es inferior al número de longitudes de onda y, además, el espectro NIR está altamente correlacionado debido a sus características bandas de absorción anchas.

4.3.2 ANÁLISIS EN COMPONENTES PRINCIPALES (PCA)

Esta herramienta ha constituido el pilar central a partir del cual se han desarrollado multitud de métodos quimiométricos.

Conceptualmente, PCA recoge la idea de condensar una gran cantidad de datos de partida en unos pocos parámetros representativos (denominados componentes principales, factores latentes o variables latentes) que capturan la

máxima variabilidad existente entre objetos y variables. Varias revisiones sobre PCA han sido publicadas en diferentes campos [21.-26].

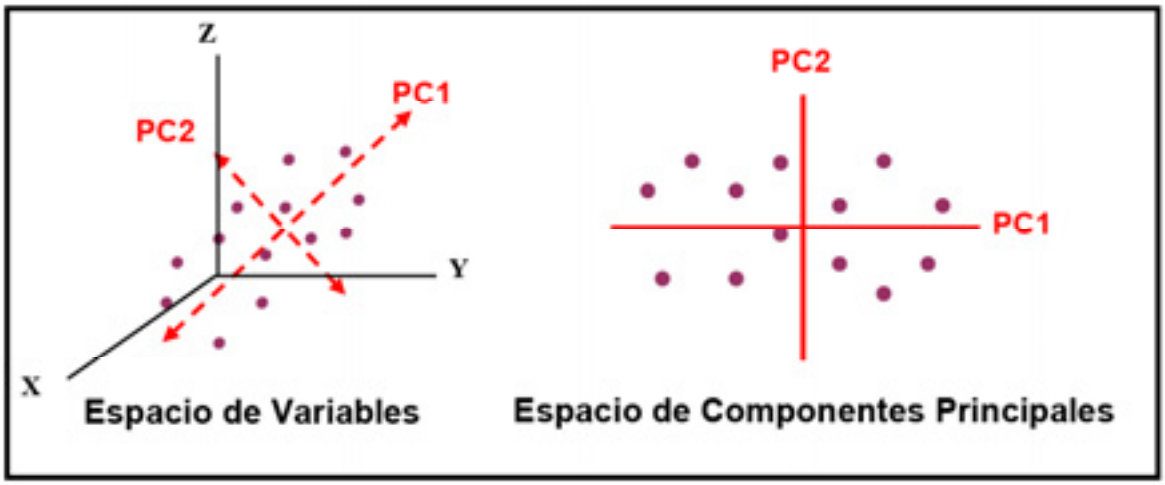


Figura 4.2. Representación gráfica de la reducción de la dimensionalidad llevada a cabo por el Análisis en Componentes Principales (PCA) desde un espacio de tres variables a uno de dos componentes principales.

A modo de ejemplo se puede considerar un conjunto de espectros organizados matricialmente, **X**(**m**×**n**), como un conjunto de **m** puntos en un espacio **n**-dimensional (longitudes de onda), ver figura 4.2. El objetivo del PCA es encontrar las direcciones, ortogonales entre sí, en las cuales existe la máxima variabilidad espectral. De esta manera, un PCA puede ser interpretado como un cambio de un sistema de coordenadas **n**-dimensional a otro **r**-dimensional **X'**(**m**×**r**) en el cual los nuevos ejes, denominados componentes principales, son

[21] Kruskal, J.B. "Factor analysis and principal components", **1978**, International Encyclopedia of Statistics, Ed. The Free Press (New York).

[22] Jackson, J. E. "Principal components and factor análisis: part I- principal components", J. of Quality Tech., **1980**, 12, 201-213.

[23] Wold S., Esbensen K.H., Geladi P. "Principal components analysis", Chemom. Intell. Lab. Syst., **1987**, 2, 37-52.

[24] Johnson G.W., Ehrlich R., "State of the art report on Multivariate chemometric methods in environmental forensics". Environm. Forens., **2002**, 3, 59-79.

[25] Stanimirova I., Walczak B., Massart D.L., Simeonov V., "A comparison between two robust PCA algorithms". Chemom. Intell. Lab. Syst., **2004**, 71, 83– 95.

[26] Tzeng D., Berns R. S., "A Review of Principal Component Analysis and Its Applications to Color Technology", Color Research Applic., **2005**, 30, 2.

perpendiculares entre sí, han sido creados por combinación lineal de las **n** variables originales y recogen la máxima variabilidad espectral.

Matemáticamente, la matriz de datos **X** se descompone en el producto de dos matrices, **T** (matriz de *scores*) y **P** (matriz de *loadings*), más una matriz **E** de residuales de **X**, **[4.3]**:

$$\mathbf{X}=\mathbf{T}\mathbf{P}^{T}+\mathbf{E} \qquad [4.3]$$

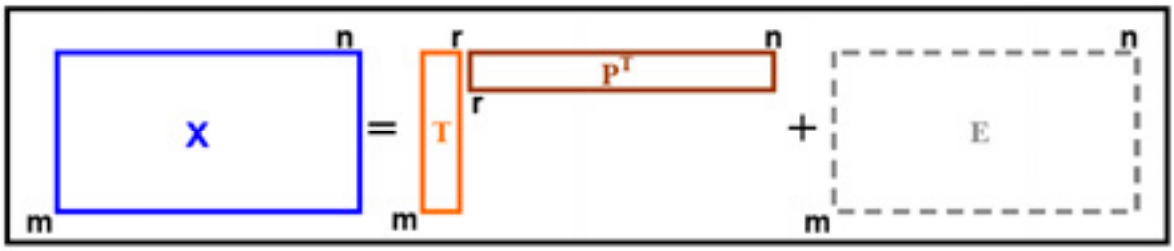


Figura 4.3. Representación gráfica de la expresión matricial de Análisis en Componentes Principales (PCA).

De forma análoga, cada elemento de la matriz **X**, $\mathbf{x}_{ij}$, puede ser expresado como:

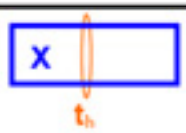
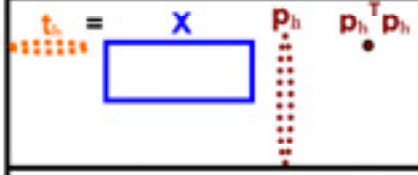
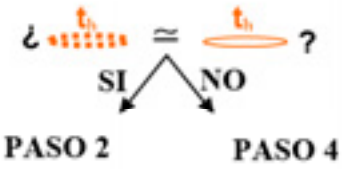
$$\mathbf{x}_{ij}=\sum_{R=1}^{r}\mathbf{t}_{iR}\mathbf{p}_{jR}^{T}+\mathbf{e}_{ij} \qquad [4.3a]$$

La matriz **T** recoge las coordenadas de las muestras originales en el nuevo sistema de ejes ortogonales. La matriz $\mathbf{P}^{T}$ recoge los cosenos directores de los nuevos ejes respecto del sistema de ejes original. La dimensión **r** del nuevo subespacio viene determinada por el rango de la matriz **X**.

Existen diferentes algoritmos de cálculo para obtener las matrices **T** y **P**. El más conocido es el algoritmo NIPALS (Nonlinear Iterative Partial Least Squares) [27]. Este método permite calcular solamente el número de componentes deseado, con lo cual, se ahorra un considerable tiempo de cálculo. Los pasos de este algoritmo iterativo están recogidos en la tabla 4.1, en élla, para facilitar la compresión, se muestra de una manera esquemática la dimensionalidad de los diferentes elementos: los rectángulos representan matrices, los vectores se representan con elipses excéntricas y los escalares con puntos.

[27] Wold H., "Multivariate Analysis", **1966**, Ed. Krishnaiah, P.R. Academic Press, (New York).

Tabla 4.1. Esquematización del proceso iterativo PCA-NIPALS.

Paso 1	
	Selección de una columna de la matriz **X** como vector de partida $\mathbf{t}_h$
Paso 2	
	Proyectar **X** sobre $\mathbf{t}_h^T$ para encontrar el correspondiente *loading* $\mathbf{p}_h^T$ $\mathbf{p}_h^T = \frac{\mathbf{t}_h^T \mathbf{X}}{\mathbf{t}_h^T \mathbf{t}_h}$
Paso 3	
	Normalizar $\mathbf{p}_{h\,norm}^T$ $\mathbf{p}_{h\,norm}^T = \frac{\mathbf{p}_h^T}{\lVert \mathbf{p}_h^T \rVert}$; $\mathbf{p}_h^T = \mathbf{p}_{h\,norm}^T$
Paso 4	
	Proyectar **X** sobre $\mathbf{p}_h$ para encontrar el correspondiente *score* $\mathbf{t}_h$ $\mathbf{t}_h = \frac{\mathbf{X}\mathbf{p}_h}{\mathbf{p}_h^T \mathbf{p}_h}$
Paso 5	
	Comparar $\mathbf{t}_h$ utilizado en el paso 2 con el obtenido en el paso 4. Si son significativas diferentes repetir el proceso desde el paso 2. Si son iguales $\mathbf{E}_h = \mathbf{X} - \mathbf{t}_h \mathbf{p}_h^T$ y comenzar de nuevo en 1, seleccionando un vector columna de $\mathbf{E}_h$

Si se tiene en cuenta que en las ecuaciones de los pasos 2 y 4, $\mathbf{t}_h^T\mathbf{t}_h$ y $\mathbf{p}_h^T\mathbf{p}_h$ son escalares (producto vectorial de dos vectores) y se sustituye dichos valores por un valor constante genérico $\mathbf{K}_i$, se obtienen las expresiones:

$$\mathbf{K}_1\mathbf{p}_h^T = \mathbf{t}_h^T\mathbf{X} \qquad [4.4]$$

$$\mathbf{K}_2\mathbf{t}_h = \mathbf{X}\mathbf{p}_h \qquad [4.5]$$

Si se despeja $\mathbf{t}_h$ en la expresión **[4.5]**, se sustituye en **[4.4]** y se opera la transpuesta:

$$K_3 p_h^T = (X p_h)^T X; \qquad K_3 p_h^T = p_h^T X^T X;$$

Se obtiene la expresión:

$$K_3 p_h = X^T X p_h \qquad [4.6]$$

Igualmente, si se sustituye el valor de t_h de la expresión **[4.4]** en **[4.5]** se obtiene:

$$K_4 t_h = X X^T t_h \qquad [4.7]$$

Las expresiones **[4.6]** y **[4.7]** son las ecuaciones de valor y vector propio para las matrices $\mathbf{X^TX}$ y $\mathbf{XX^T}$. Con lo cual, bajo condiciones de convergencia las soluciones proporcionadas por NIPALS son iguales a las proporcionadas a través del cálculo de la fórmula del vector propio [28].

Actualmente, con los avances en la velocidad de los equipos informáticos, el cálculo de un número limitado de componentes principales ya no es tan necesario. De esta manera han ganado importancia otros métodos de resolución como la descomposición en valores singulares (SVD, *Single Value Decomposition*). Este método proporciona la descomposición de la matriz de datos en tres matrices **U**, **S**, y **V**

$$\mathbf{X=USV^T} \qquad [4.8]$$

Siendo **U**, matriz de *left eigenvectors,* que contiene la misma información que la matriz de *scores* **T** pero normalizados a longitud uno. **S** es la matriz diagonal que contiene la raíz cuadrada de los valores propios de la matriz $\mathbf{X^TX}$. $\mathbf{V^T}$, matriz de (vectores propios) *right eigenvectors*, que es exactamente igual a la matriz de *loadings* $\mathbf{P^T}$.

Desde un punto de vista algebraico y considerando la factorización de la matriz **X**(**m**×**n**) a través de la descomposición SVD, podemos considerar **X**(**m**×**n**) como una aplicación en base canónica entre dos espacios vectoriales de dimensiones **n** y **m**. La matriz **S**(**m**×**n**) representa la misma aplicación que la matriz **X**(**m**×**n**) pero entre las base ß y θ. La matriz **V**(**n**×**n**) es la matriz de cambio de base entre las bases ß y canónica en el espacio **n**-dimesional y, paralelamente, la matriz **U(m×m)** es la matriz de cambio de base entre las bases θ y canónica en el espacio **m**-dimesional. Ver figura 4.4.

---

[28] Strang G., "Linear Algebra and its Applications", **1988**, Ed. Harcourt Brace Jovanovich Publishers.

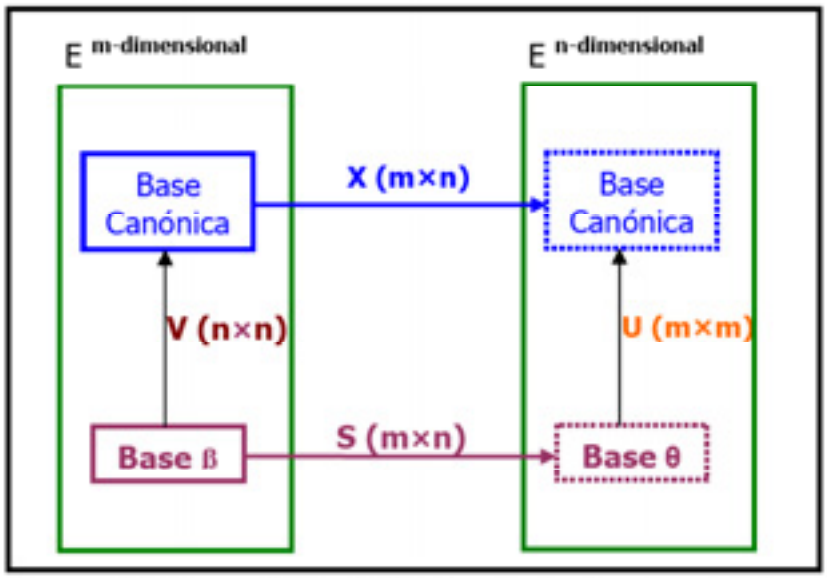


Figura 4.4. Interpretación algebraica de la descomposición en valores singulares (SVD).

Independientemente del método empleado para calcular los componentes principales, uno de los aspectos más delicados y fundamentales en PCA es la determinación del rango o número óptimo de componentes que describen la matriz **X**. En la mayoría de *softwares* comerciales, ésta es la única decisión que el usuario debe tomar. Para decidir el número de componentes existen varios métodos heurísticos y estadísticos como el porcentaje de varianza explicada o *scree test* y la validación cruzada o *cross validation*.

El **porcentaje de varianza explicada** se puede utilizar si se dispone de suficiente experiencia en la manipulación de conjuntos de datos similares. La fracción de varianza acumulada (s_e^2) se calcula a partir de la relación de la suma de **d** valores propios respecto a la suma de todos los valores propios, **p**, calculados a través de la expresión:

$$s_e^2 = \frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \qquad [4.9]$$

El **gráfico de varianza explicada *o scree test*** es la representación gráfica del número de componentes principales frente a la varianza explicada (o residual).

A partir de este gráfico se escoge el número de componentes más bajo que presenta una variación significativa de la varianza. En los casos en los que la disminución de la varianza explicada es prácticamente asintótica este criterio puede acabar siendo poco objetivo.

La **validación cruzada o *cross validation*** es un método iterativo en el cual se extrae un espectro (o espectros) del conjunto de calibración y se calcula las matrices **T** y **P** con el resto del conjunto. Los espectros extraídos se predicen y se calcula el error residual utilizando, en cada iteración, un componente principal adicional. Al igual que el gráfico de la varianza explicada, el número de componentes seleccionados es el mínimo que presenta una variación significativa en la varianza.

4.3.3 REGRESIÓN EN COMPONENTES PRINCIPALES (PCR)

Las propiedades de los métodos PCA y MLR se aúnan en la regresión en componentes principales (PCR). Por un lado, la descomposición de la matriz espectral **X** en componentes principales permite reducir el número de variables y eliminar los efectos de colinealidad entre variables. Por otra parte, la regresión lineal múltiple entre la matriz de *scores* obtenidos y los datos analíticos de referencia proporciona una solución de mínimos cuadrados.

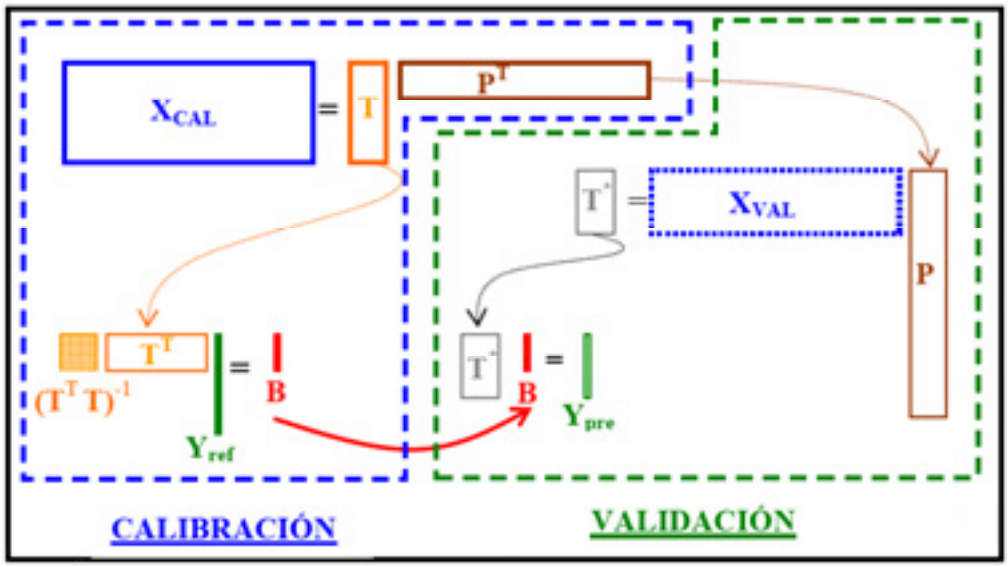


Figura 4.5. Representación conjunta de las etapas de calibración y validación en el proceso de creación de modelos PCR.

En la figura 4.5 se muestran como se combinan estos dos métodos en las etapas de calibración y predicción. En un primer paso se descompone la matriz de datos espectroscópicos de calibración, $\mathbf{X}_{cal}$, en las matrices de *scores*, **T**, y *loadings*, **P**, a través de un PCA. Con la matriz de *scores* obtenidos y los datos de referencia asociados a la matriz de datos espectroscópicos, $\mathbf{Y}_{ref}$, se calcula por mínimos cuadrados la matriz **B** de regresores, concluyendo, de esta manera, la etapa de calibración. La etapa de validación o de predicción de nuevas muestras comienza con la obtención de los *scores* $\mathbf{T}^*$ de los espectros pertenecientes al conjunto de validación, $\mathbf{X}_{val}$, y los *loadings* obtenidos en la etapa de calibración. Por último, los valores analíticos predichos $\mathbf{Y}_{pre}$ se calculan a partir de las matrices $\mathbf{T}^*$ y **B**.

Paradójicamente, la ventaja que supone reducir el número de variables a través de un análisis en componentes principales se convierte en el punto débil de este método, ya que los componentes principales calculados mediante PCA son aquellos que mejor describen la varianza espectral, pero son calculados sin tener en cuenta los valores analíticos de referencia.

4.3.4 REGRESIÓN PARCIAL POR MÍNIMOS CUADRADOS (PLS)

La regresión parcial por mínimos cuadrados (PLS) es probablemente el método de regresión por reducción multivariante más ampliamente utilizado [16]. Este método no sólo permite trabajar con datos que presentan una alta colinealidad sino que además proporciona una matriz de regresores robustos ya que son calculados teniendo en cuenta la descomposición espectral de la matriz **X** e **y** conjuntamente, o lo que es lo mismo, maximizando la covarianza entre **X** e **y**.

PLS fue desarrollado entre 1975 y 1982 por Herman Wold y colaboradores [29] y desde entonces, este modelo ha sido ampliamente estudiado, aplicado, contrastado y divulgado [30 - 34].

---

[29] Wold H. "Soft Modeling. The Basic Design and Some Extensions. in Systems Under Indirect Observation", **1982**, Eds. Jöreskog K. G., Wold H. (Amsterdam).

[30] Geladi P. "Notes on the history and nature of partial least squares (PLS) modelling", J. Chemom., **1988**, 2, 231-246.

El algoritmo PLS-NIPALS puede ser utilizado para extraer de forma secuencial los factores PLS. La estructura básica del algoritmo es la presentada en la tabla 4.2 y las etapas son similares a las descritas en el algoritmo NIPALS-PCA, con las excepciones de que en lugar de utilizar un vector columna de **X** para iniciar el algoritmo, éste se inicia con el vector de la variable respuesta **y**. De esta manera, en el calculo de los *scores* $\mathbf{t}_h$ (Paso 4) se han utilizado tanto la matriz **X** como el vector **y**. Además en el algoritmo PLS-NIPALS se introduce una nueva etapa (Paso 9) en la cual se calcula, para cada factor, la relación interna existente entre el *score* calculado para la **X** y para la **y**.

La parte más importante de cualquier método de regresión es su uso y habilidad para predecir muestras nuevas. La matriz de regresores **B** que se utiliza en la expresión general **Y=XB** tiene la expresión general [35]:

$$\mathbf{B} = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1} \qquad [4.10]$$

Para obtener la mejor estimación de la matriz **B**, el modelo PLS debe ser calibrado con muestras que abarquen toda la variación existente en **y** y que, al mismo tiempo, sean representativas de las futuras muestras a determinar. El número óptimo de muestras a utilizar en el conjunto de calibración vendrá determinado por la variabilidad y complejidad que presente la variable respuesta.

[31] Höskuldsson A. "PLS regression methods". J. Chemom. **1988**, 2, 211-228.

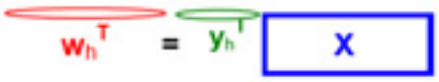
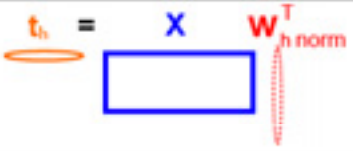
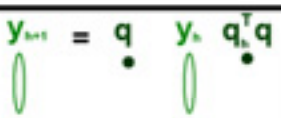
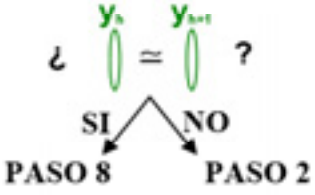
[32] De Jong S. "SIMPLS: an alternative approach to partial least squares regression". Chemom. Intell. Lab. Syst., **1993**, 18, 251-263.

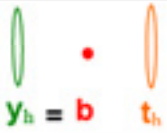
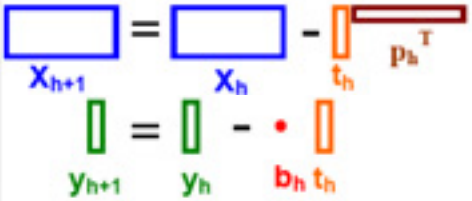
[33] Phatak A., De Jong S., "The geometry of partial least squares". J. Chemom., **1997**, 11, 311-338.

[34] Burnham A. J., MacGregor J. F., Viveros R., "A statistical framework for multivariate latent variable regression methods based on maximum likelihood", J. Chemom., **1999**, 13, 49-65.

[35] Helland I S., "On the structure of Partial Least-Squares Regression" Communications in Statistics-Simulation and Computation, **1988**, 17, 581-607.

Tabla 4.2. Representación esquemática del algoritmo NIPALS-PLS

Paso 1	
	Seleccionar la variable respuesta **y**. $\mathbf{y}=\mathbf{y}_h$.
Paso 2	
	Proyectar **X** sobre $\mathbf{y}_h^T$ para calcular el correspondiente *loading* $\mathbf{w}_h^T$ $\mathbf{w}_h^T=\frac{\mathbf{y}_h^T\mathbf{X}}{\mathbf{y}_h^T\mathbf{y}_h}$
Paso 3	
	Normalizar $\mathbf{w}_{h\,norm}^T$ $\mathbf{w}_{h\,norm}^T=\frac{\mathbf{w}_h^T}{\|\mathbf{w}_h^T\|}$; $\mathbf{w}_h^T=\mathbf{w}_{h\,norm}^T$
Paso 4	
	Proyectar **X** sobre $\mathbf{w}_h$ para encontrar el correspondiente *score* $\mathbf{t}_h$ $\mathbf{t}_h=\mathbf{X}\mathbf{w}_h$
Paso 5	
	Calcular el *loading* $\mathbf{q}^T$ de **y** $\mathbf{q}^T=\frac{\mathbf{t}_h^T\mathbf{y}}{\mathbf{t}_h^T\mathbf{t}_h}$
Paso 6	
	Estimar el valor de $\mathbf{y}_{h+1}$. $\mathbf{y}_{h+1}=\frac{\mathbf{y}_h\mathbf{q}}{\mathbf{q}^T\mathbf{q}}$
Paso 7	
	Comparar $\mathbf{y}_h$ con el nuevo $\mathbf{y}_{h+1}$ calculado en el Paso 6. Si la diferencia es menor que el criterio de convergencia establecido, ir al Paso 8 si no volver al Paso 2 usando $\mathbf{y}_{h+1}$
Paso 8	
	Calcular $\mathbf{p}_h^T$ a partir del *score* $\mathbf{t}_h$ con el cual se ha conseguido la convergencia $\mathbf{p}_h^T=\frac{\mathbf{t}_h^T\mathbf{X}}{\mathbf{t}_h^T\mathbf{t}_h}$

Paso 9	
	Establecer el valor de la pendiente **b** del modelo lineal subyacente entre **y** y $\mathbf{t}_h$ $\mathbf{b}=\frac{\mathbf{t}^T\mathbf{y}}{\mathbf{t}_h^T\mathbf{t}_h}$
Paso 10	
	Calcular los residuales para $\mathbf{X}_{h+1}$ e $\mathbf{y}_{h+1}$: $\mathbf{X}_{h+1}=\mathbf{X}_h-\mathbf{t}_h\mathbf{p}_h^T$ $\mathbf{y}_{h+1}=\mathbf{y}_h-b_h\mathbf{t}_h$ Iniciar el proceso en el Paso 1 utilizando $\mathbf{X}_{h+1}$ e $\mathbf{y}_{h+1}$ en lugar de $\mathbf{X}_h$ e $\mathbf{y}_h$

4.3.4.1 Evaluación de resultados.

Tanto si se emplea PCR como PLS es necesario algún sistema que, a parte de saber si una calibración proporciona una capacidad predictiva apropiada, permita evaluar la conveniencia de utilizar más o menos componentes principales en una determinada calibración.

La determinación del número de componentes se llevó a cabo utilizando como criterios de decisión el análisis de la ganancia en varianza explicada al añadir un nuevo componente al modelo y a través de la comparación de la representación gráfica del RMSE, error cuadrático medio (*Root Mean Square Error*), **[4.11]** frente al número de componentes tanto para calibración RMSEC, como para validación RMSEP. El RMSEC disminuye paulatinamente al aumentar el número de componentes, en cambio, el RMSEP presenta un mínimo o bien una disminución relativa significativamente menor a partir del número óptimo de componentes, siendo **m** el número de muestras, $\mathbf{y}_{REF}$ los valores de referencia e $\hat{\mathbf{y}}_{PLS}$ los valores predichos por el modelo PLS. El RMSE puede ser considerado como el error medio obtenido en el proceso de modelado y está expresado en las mismas unidades que los datos de referencia.

$$RMSE=\sqrt{\frac{\sum_{i=1}^{m}(\hat{y}_{PLS}-y_{REF})^2}{m}} \qquad [4.11]$$

Otros estadísticos básicos, además del RMSE, utilizados para determinar la bondad del modelo y su habilidad predictiva fueron: los obtenidos a través del análisis de residuales (estadístico-t), el coeficiente de determinación entre los valores de referencia y los predichos por el modelo (R^2) y el RSE, error estándar relativo (*Relative Standard Error*) **[4.12]**, calculado tanto para calibración RSEC, como para validación externa RSEP.

$$\mathbf{RSE(\%)=\sqrt{\frac{\sum_{i=1}^{m}(\hat{y}_{PLS}-y_{REF})^2}{\sum_{i=1}^{m}(\hat{y}_{REF})^2}}x100} \qquad \mathbf{[4.12]}$$

El RSE es un estadístico menos optimista, o más realista, que el RMSE para determinar la bondad de un modelo ya que en su determinación se ha tenido en cuenta el ámbito de aplicación del modelo, al dividir el bias por los valores de referencia.

4.3.5 MÉTODOS DE RESOLUCIÓN

Los métodos de resolución descomponen matemáticamente una señal instrumental compleja en las contribuciones debidas a los componentes que forman el sistema, y se pueden expresar en forma matricial de la siguiente manera:

$$\mathbf{A = CS^T + E} \qquad \mathbf{[4.13]}$$

Donde **A (mxn)** es la matriz de datos espectrales adquiridos a diferentes valores de una cierta variable (tiempo, pH, concentración,….) durante un proceso o reacción química. **C (mxr)** es la matriz relacionada con los perfiles de concentración y $\mathbf{S^T}$ **(rxn)** es la matriz de las respuestas unitarias de las especies químicas activas espectroscópicamente. **E (mxn)** es la matriz de los residuales no explicados por las especies químicas definidas en **C** y $\mathbf{S^T}$, que se asumen independientes y de varianza constante. Los parámetros **m** y **n** hacen referencia al número de muestras y al número de canales o longitudes de onda, **r** es el número de especies químicas activas espectroscópicamente.

La factorización de la matriz respuesta **A** está siempre sometida a ambigüedades, no importa que método de resolución se utilice [36]. Esto implica que la resolución de la matriz **A** puede conseguirse, sin perder calidad de ajuste, con infinitos pares de matrices **C** y $\mathbf{S}^T$ distintos.

La expresión **[4.14]** refleja la ambigüedad rotacional dónde para una determinada descomposición, siempre existe un infinito número de matrices **T** invertibles que hacen que la solución no sea única. A esta ambigüedad también se la denomina "de forma" porque en función de cómo sea **T** los perfiles **C** y **C***, al igual que **S** y $\mathbf{S}^{T*}$ pueden ser totalmente diferentes.

$$\mathbf{A} = \mathbf{C}\,\mathbf{S}^T = \mathbf{C}\,\mathbf{T}\,\mathbf{T}^{-1}\,\mathbf{S}^T = (\mathbf{C}\mathbf{T})\,(\mathbf{T}^{-1}\mathbf{S}^T) = \mathbf{C}^*\,\mathbf{S}^{T*} \qquad [4.14]$$

La expresión **[4.15]** refleja las ambigüedades de rotación o de magnitud. Dado cualquier esquema de factorización, siempre se puede extraer un escalar genérico, α, de la matriz **C** y, su inverso 1/ α, de la matriz $\mathbf{S}^T$. De esta manera, $\mathbf{C}^*$ es α veces **C** y $\mathbf{S}^{T*}$ es 1/α veces $\mathbf{S}^T$, aunque $\mathbf{C}^*$ y $\mathbf{S}^{T*}$ tienen la misma forma que **C** y $\mathbf{S}^T$, respectivamente.

$$\mathbf{A} = \mathbf{C}\,\mathbf{S}^T = (\alpha\,\mathbf{C}^*)\,(1/\alpha\,\mathbf{S}^{T*}) = (\alpha \times 1/\alpha)\,(\mathbf{C}^*\,\mathbf{S}^{T*}) = \mathbf{C}^*\,\mathbf{S}^{T*} \qquad [4.15]$$

4.3.5.1 Método de Resolución Multivariante de Curvas mediante Mínimos Cuadrados Alternados (MCR-ALS)

La ecuación **[4.13]** se resuelve de manera iterativa utilizando el algoritmo de mínimos cuadrados alternados, ALS (*Alternating Least Squares*). En la figura 4.6 aparece esquematizado las distintas etapas que componen el citado algoritmo.

[36] Lawton W.H., Sylvestre E.A., " Eliminitaion of linear paramenters in nonlinear regresision" Technometrics, **1971**, 13(3), 461-467.

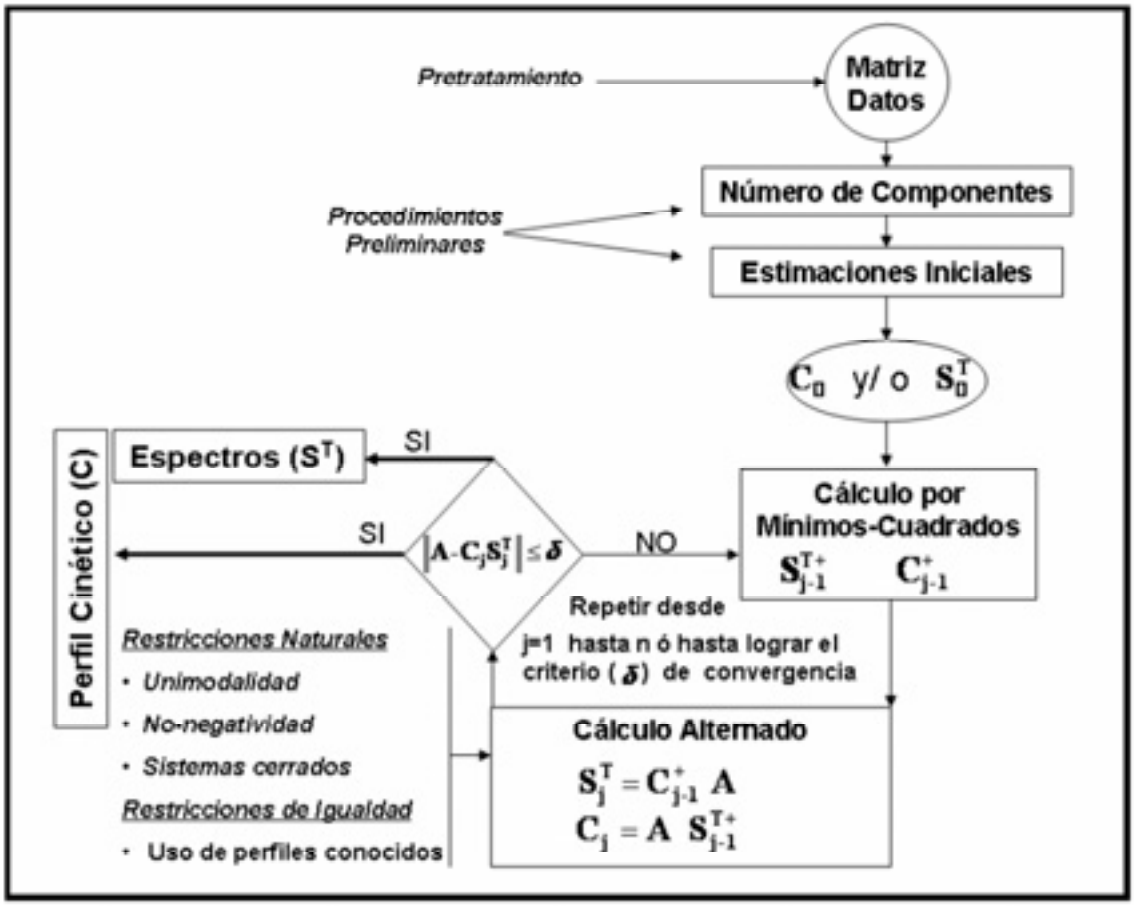


Figura 4.6. Esquema del algoritmo de mínimos cuadrados alternos (ALS)

4.3.5.1.1 Establecimiento del número de componentes.

Una de las etapas claves es establecer el número de especies espectroscópicamente activas que pueden ser monitorizadas. En la ausencia de ruido y otras causas de variabilidad, el rango de la matriz experimental debería ser el mismo que el número de especies químicas. Sin embargo, en sistemas en evolución en los que el número de analitos es mayor que el número de reacciones más uno, (como en el caso de las fermentaciones alcohólicas) existe un problema de deficiencia de rango, esto se traduce en que no se pueden obtener, de una manera directa, los perfiles de todos los analitos involucrados en el proceso [37-39]. Existen diferentes herramientas de ayuda para la determinación

[37] Amrheim M., Srinivasan B., Bonvin D., Schumacher M.M.,"On the rank deficiency and rank augmentation of the spectral measurement matrix", Chemom. Intell. Lab. Syst., **1996**, 33(1), 17-33.

del rango de un sistema fermentativo, como por ejemplo, la descomposición en valores singulares o la representación gráfica de los vectores asociados a los valores propios más significativos.

4.3.5.1.2 Estimaciones iniciales

Para inicializar el algoritmo ALS es necesario partir de unas estimaciones iniciales, bien de los perfiles de concentración, $\mathbf{C}_0$, bien de los perfiles espectrales, $\mathbf{S}_0$. Dichas estimaciones las podemos obtener a través de diferentes métodos de resolución preliminares. Gran parte de dichos métodos fueron desarrollados a principios de los 80 para diferenciar y resolver analitos en sistemas formados por instrumentos cromatográficos acoplados a detectores espectroscópicos multi-canal utilizados como técnica de detección. Los más utilizados son:

- **EFA** (*Evolving Factor Analysis*) [40-42]. Con este método es posible establecer y diferenciar las regiones en las que cambia el número de analitos estudiados, siendo el principio subyacente el Análisis en Componentes Principales, que se realiza sobre la matriz experimental partiendo del primer espectro y aumentándola en la dirección del proceso. Este análisis es realizado de principio a fin del proceso (*forward*) y en la dirección opuesta (*backward*), siendo el resultado final la evolución de los valores propios, una vez consolidadas las dos direcciones de análisis.

[38] Saurina J., Hernández-Cassou S., Tauler R., Izulerdo-Ridorsa A., " Multivariate resolution of rank deficient spectrophotometric data from first-order kinetic decomposition reactions", J. Chemom., **1998**, 12, 183-203.

[39] Garrido M., Lázaro I., Larrechi M.S., Rius F.X.. "Multivariate resolution of rank-deficien NIRS data from the reaction of curing epoxy resins using the rank augmentation strategy and MCR-ALS". Anal. Chim. Acta, **2004**, 58, 47-53.

[40] Gampp H., Maeder M., Meyer C.J., Zuberbühler A.D., "Calculation of Equilibrium Constants from Multiwavelength Spectroscopic Data. Model-free Analysis of Spectrophotometric and ESR Titrations", Talanta, **1985**, 32, 1133-1139.

[41] Maeder M., "Evolving Factor Analysis for the resolution of overlapping chromatographic peaks" Anal. Chem., **1987**, 59, 527-530.

[42] Gemperline P.J., Hammilton C., "Evolving Factor Analysis applied to flow injection analysis data". J. Chemom., **1989**, 3, 455-461.

- **SIMPLISMA** (*SIMPLe-to-use Interactive Self-modeling Mixture Analysis*) [43-44]. Este algoritmo está basado en la selección de lo que se denominan variables puras, definiéndose variable como aquella que su intensidad es debida solamente a uno de los analitos estudiados.
- **ITTFA** (*Iterative Target Transformation Factor Analysis*) [45-46]. Este método utiliza los vectores propios, obtenidos a través de una descomposición en valores singulares (SVD) y busca las variables puras, al igual que SIMPLISMA, pero en el espacio de componentes principales.

4.3.5.1.3 Algoritmo ALS

Una vez que se dispone de las estimaciones iniciales, el modelo general indicado en la expresión **[4.13]**, se resuelve a través de un proceso iterativo de optimización para minimizar la matriz de residuales **E**. Los dos pasos del proceso iterativo son:

$$\mathbf{S}^T = (\mathbf{C})^+ \mathbf{A}^* \qquad [4.16]$$

$$\mathbf{C} = \mathbf{A}^* (\mathbf{S}^T)^+ \qquad [4.17]$$

Donde $(\mathbf{S}^T)^+$ y $(\mathbf{C})^+$ son las respectivas pseudoivnversas de $\mathbf{S}^T$ y $\mathbf{C}$. La matriz $\mathbf{A}^*$ es similar a la matriz de datos experimental **A**, pero reconstruida a partir del número de variables latentes proporcionados por un PCA, su uso proporciona una mayor estabilidad en los cálculos ya que tiene parte del ruido filtrado por la selección de los componentes principales. La optimización iterativa de las matrices $\mathbf{S}^T$ y $\mathbf{C}$ a partir de las ecuaciones **[4.16]** y **[4.17]** siguen el orden propuesto cuando se parte de una estimación inicial de la matriz de concentraciones, si se dispone de estimaciones iniciales de espectros el orden cambiaría. El algoritmo se repite de forma iterativa hasta que se logra la convergencia o hasta que se alcanza un

[43] W. Windig, "Mixture analysis of spectral data by multivariate methods", Chemom. Intell. Lab. Syst., **1988**, 4, 201-213.

[44] Cuesta F., Massart D.L. "Application of SIMPLISMA for the assessment of peak purity in liquid chromatography with diode array detection". Anal. Chimi. Acta, **1994**, 298, 331-339.

[45] Vandeginste B.G.M., Esserst R., Bosman T., Reijnen J., Kateman G., "Three-component curve resolution in liquid chormatograpy with multiwavelength diode array detection", Anal. Chem., **1985**, 57, 971-985.

[46] Gemperline P.J. "Target Transformation Factor Analysis with Linear Inequality Constraints Applied to Spectroscopic-Chromatographic Data", Anal. Chem., **1986**, 58, 2656-2663.

número preseleccionado de iteraciones, definiéndose la convergencia como el cambio relativo de la desviación estándar de los residuales entre dos ciclos consecutivos.

4.3.5.1.4 Restricciones

Uno de las principales características de los métodos de resolución es que no necesitan ningún tipo de información de referencia, ni química ni matemática, para ser aplicados. Sin embargo, toda la información que se tenga del sistema puede ser utilizada en forma de restricciones que se aplican sobre los perfiles espectrales y/o de concentración, con el objetivo de reducir el dominio de las soluciones posibles.

Restricción de no-negatividad.

Esta es la restricción más general en la optimización por mínimos cuadrados. Las concentraciones de las especies químicas han de ser siempre valores positivos o cero.

La aplicación de esta restricción se hace utilizando el algoritmo de mínimos cuadrados no negativos [47-48], forzando los valores negativos a ser cero en cada iteración.

Restricción de unimodalidad.

Esta restricción se puede aplicar en aquellos perfiles en los que se presupone que sólo poseen un máximo. Esta situación es bastante corriente para perfiles de concentración y no tan corriente para los perfiles espectrales. Existen varios algoritmos para la aplicación de unimodalidad [49] aunque la forma más sencilla e intuitiva de aplicar esta restricción es encontrar el máximo del perfil, los valores en el entorno de este máximo han de disminuir de forma monótona. En ciertos casos se permite una desviación del comportamiento monótono, como por ejemplo cuando los perfiles de concentración muestran cierto ruido asociado.

---

[47] Winding W., "Self-modeling mixture analysis of spectral data with continuous concentration profiles", **1992**, 16(1), 1-16.

[48] Maeder M., Zuberbühler A.D., "Nonlinear Least-Squares Fitting of Multivariate Absorption Data", Anal. Chem., **1990**, 62, 2220-2224.

[49] Bro R., Sidiripoulos N.D., "Least Squares algorithms under unimodality and non-negativity constraints" J. Chemom., **1998**, 12, 223-247.

Restricción de sistemas cerrados.

Se pueden emplear restricciones adicionales de normalización, las cuales tienen un efecto importante sobre la ambigüedad de intensidad o escala. Una de las formas más comunes de normalización es la de los sistemas cerrados donde la cantidad total de las especies es constante.

Restricción de igualdad

En la expresión más general del problema de mínimos cuadrados, **y=Xb**, este tipo de restricciones de aproximación se implementan a través de las funciones penalizadoras (*penalty function*) [9], [50]. Si se posee algún tipo de información, tanto en el dominio de concentraciones como espectral, puede ser utilizada como restricciones de igualdad, actuando las funciones penalizadoras como "elementos ponderadores", que hacen que las estimaciones obtenidas en cada uno de los ciclos del algoritmo ALS se adapten más (restricción severa, *hard constraint*) o menos (restricción suave, *soft constraint*) a la información suministrada como restricción de igualdad.

4.3.6 METODOS EMPIRICOS DE MODELADO

Las fermentaciones alcohólicas han sido y, son aún en día, estudiadas a través de modelos empíricos solamente válidos en unas condiciones de trabajo específicas y definidas [51]. Dentro de la gran variedad de modelos empíricos existentes, cobran especial importancia por su grado de difusión, aceptación y utilización los modelos basados y/o inspirados en las leyes mecanicistas de la cinética enzimática. Estos modelos vienen definidos por el siguiente sistema de ecuaciones diferenciales para el crecimiento **[4.18]**, la formación de producto **[4.19]** y el consumo de substrato **[4.20]**.

[50] Bro R. "Multi-way Analysis in the Food industry. Models, Algorithms and Applications", **1998**, PhD thesis.

[51] Marin R.M. "Alcoholic Fermentation Modelling: Current State and Perspectivas" Am. J. Enol. Vitic., **1999**, 50(2), 166-178.

$$\frac{dX}{dt} = \mu X \qquad \textbf{[4.18]}$$

$$\frac{dP}{dt} = \nu X \qquad \textbf{[4.19]}$$

$$\frac{dS}{dt} = -\left(\frac{1}{Y_{x/s}}\frac{dX}{dt}\right) - \left(\frac{1}{Y_{p/s}}\frac{dP}{dt}\right) \qquad \textbf{[4.20]}$$

Las tasas de variación de los principales analitos a lo largo del tiempo son función de los parámetros μ (tasa específica de crecimiento, horas^{-1}) y/o ν (tasa específica de formación de producto, también en horas^{-1}). Los diferentes modelos se diferencian en la forma en la que estos parámetros quedan definidos. En la tabla 4.3 aparecen recogidos las expresiones de dichos parámetros en los modelos más ampliamente utilizados. Los modelos 1 y 2 representan una cinética libre de inhibición, los modelos 3 y 4 incluyen un término de inhibición por sustrato y los modelos del 5 al 7 representan una cinética por inhibición por producto.

Tabla 4.3. Expresión de los parámetros μ, tasa especifica de crecimiento y ν, tasa especifica de formación de producto, para diferentes modelos empiricos.

Modelo	$\mu =$	$\nu =$	Referencia
1	$\mu_{max}\left(\frac{S}{K_{SX}+S}\right)$	$\nu_{max}\left(\frac{S}{K_{SP}+S}\right)$	Monod [52]
2	$\mu_{max}\left(1-\exp\left(-\frac{S}{K_{SX}}\right)\right)$	$\nu_{max}\left(1-\exp\left(-\frac{S}{K_{SP}}\right)\right)$	Teissier [53]
3	$\mu_{max}\left(\frac{S}{K_{SX}+S}\right)\exp\left(-\frac{S}{K_{IX}}\right)$	$\nu_{max}\left(\frac{S}{K_{SP}+S}\right)\exp\left(-\frac{S}{K_{IP}}\right)$	Edwards [54]
4	$\mu_{max}\left(\frac{S}{K_{SX}+S}\right)\left(1-\frac{S}{S_{S,max}}\right)^{n}$	$\nu_{max}\left(\frac{S}{K_{SP}+S}\right)\left(1-\frac{S}{S_{P,max}}\right)^{n}$	Luong [55]
5	$\mu_{max}\left(\frac{S}{K_{SX}+S}\right)(1-K_{PX}P)$	$\nu_{max}\left(\frac{S}{K_{SP}+S}\right)(1-K_{PP}P)$	Hinshelwood [56]
6	$\mu_{max}\left(1-\frac{P}{P_{X\,max}}\right)$	$\nu_{max}\left(1-\frac{P}{P_{P\,max}}\right)$	Ghose Tyagi [57]
7	$\mu_{max}\left(\frac{S}{K_{SX}+S}\right)\exp\left(-K_{PX}P\right)$	$\nu_{max}\left(\frac{S}{K_{SP}+S}\right)\exp\left(-K_{PP}P\right)$	Aiba [58]

[52] Monod J., Ann. Rev. Microbiol., "The Growth of Bacterial Cultures", **1949**, 3, 371-394.

[53] Teissier G., "Croissance des populations bactériennes et quantité d'aliment disponible" Rev. Sci., **1942**, 80, 209-230.

[54] Edwards V.H., "The influence of high substrate concentrations on microbial kinetics", Biotech. Bioeng., **1970**, 12, 679-712.

[55] Luong J.H.T., "Generalization of monod kinetics for analysis of growth data with substrate inhibition" Biotech. Bioeng., **1987**, 29, 242-248.

[56] Hinshelwood C.N. "The Chemical Kinetics of the Bacterial Cells", **1946**, Ed. Oxford Universtiy Press, (London).

[57] Ghose T.K., Tyagi R.D., "Rapid ethanol fermentation of cellulose hydrolysate. II. Product and substrate inhibition and optimization of fermentor design", Biotech. Bioeng., **1979**, 21, 1401-1420.

[58] Aiba S., Shoda M., Nagatani M. "Kinetics of Product Inhibition in Alcoholic Fermentation", Biotech. Bioeng., **1968**, 10, 845-864.

Donde:

X Concentración celular (g (materia seca)/L).

P Concentración de etanol (g/L).

S Concentración de sustrato (g/L).

μ_{max} Tasa específica de crecimiento (h^{-1}).

ν_{max} Tasa máxima especifica de fermentación (h^{-1}).

K_{SX} Constante de crecimiento en la ecuación del consumo de substrato (g/L).

K_{SP} Constante de producto en la ecuación del consumo de substrato (g/L).

K_{PX} Constante de inhibición del crecimiento por el etanol (g/L).

K_{PP} Constante de inhibición de la fermentación por el etanol (g/L).

K_{IX} Constante de inhibición del crecimiento por el substrato (g/L).

K_{IP} Constante de inhibición de la producción de etanol por el substrato (g/L).

$S_{S\max}$ Constante de consumo máximo de substrato en el término de crecimiento biológico (g/L).

$S_{P\max}$ Constante de consumo máximo de substrato en el término de producción de etanol (g/L).

$P_{X\max}$ Constante de producción máxima de producto en el término de crecimiento biológico (g/L).

$P_{P\max}$ Constante de formación máxima de producto en el término de producción de etanol (g/L).

$Y_{x/s}$ Ratio de células producidas por substrato consumido para el crecimiento.

$Y_{p/s}$ Ratio de etanol producido por substrato consumido para la fermentación.

La metodología para encontrar los valores que toman los diferentes parámetros ha evolucionado de forma paralela al desarrollo de equipos informáticos potentes y rápidos. Tradicionalmente, y en los casos más simples, se procedía a linealizaciones [59] o aproximaciones polinómicas [60], aunque estos

[59] Atilio C., Perego P., Lodi A., Parisi F., del Bofghi M. "A kinetic study of Saccharomyces strains: Performance at High Sugar Concentrations", Biotech. and Bioeng., **1985**, 27, 1108-1114.

métodos están sometidos a problemas numéricos debido a que utilizan inversas de datos experimentales o cocientes de variables determinadas experimentalmente, lo que afecta a la precisión en la determinación de los parámetros. Para evitar estos problemas, el modelado puede hacerse por integración simultánea del sistema de ecuaciones [61]. Los resultados obtenidos tras la integración son comparados con los datos experimentales y, a través de una rutina de optimización, el valor de los parámetros va cambiando hasta que el error entre los datos experimentales y los calculados es minimizado.

[60] Bovee J.P., Strehaiano P., Goma G., Sevely Y., "Alcoholic Fermentation: Modelling based on sole substrate and product measurement". Biotech. and Bioeng., **1984**, 26, 328-334.

[61] Gülnur B., Doruker P., Kirdar B., Ilsen Z., Ülgen K. "Mathematical description of ethanol fermentation by Immobilised Saccharomyces cerevisiae", Process Biochem., **1998**, 33(7), 763-771.

4.3.7 USO CONJUNTO DE MODELOS EMPÍRICOS DE MODELADO Y DE MCR-ALS

El uso combinado de modelos empíricos junto a métodos de resolución de curva es un planteamiento novedoso que aprovecha de forma sinérgica las ventajas de ambos métodos. Dicho planteamiento ha sido aplicado con éxito en la resolución de problemas cinéticos [62-63], en problemas de cuantificación con analitos interferentes [64] y en la monitorización de procesos [65-66], por citar algunos ejemplos. La principal ventaja del uso conjunto de ambos métodos es la minimización de la ambigüedad de rotación inherente a los métodos de resolución, lo cual permite la monitorización del sistema en estudio y además, como información adicional, se obtiene una estimación de los diferentes parámetros del modelo empírico.

El modo en el que modelos empíricos y de resolución han sido combinados en este trabajo aparece esquematizado en la figura 4.7.

---

[62] De Juan A., Maeder M., Martinez M., Tauler R., "Combining hard- and soft- modelling to solve kinetic problems", Chem. Intell. Lab. Syst., **2000**, 54, 123-141.

[63] Haario H., Taavitsainen V.M. "Combining soft and hard modelling in chemical kinetic models". Chem. Intell. Lab. Syst., **1998**, 44, 77-98.

[64] Diework J., De Juan A., Maeder M., Tauler R., Lendl B. "Application of a Combination of Hard and Soft Modeling for Equilibrium Systems to the Quantitative Análisis of ph-Modulated Mixture Simples". Anal. Chem. **2003**, 75, 641-647.

[65] Van Sprang E. N. M., Ramaker H., Westerhuis J., Smilde A.K., Wienke D., "Statistical Batch process monitoring using grey models", AIChE Journal, **2005**, 51(3), 931-945.

[66] Gemperline P., Putxy G., Maeder M., Walker D., Tarczynsky F., Bosserman M. "Calibration-Free Estimates of Batch Process Yields and Detection of process upsets using in Situ Spectroscopic Measurements and Nonisothermal Kinetic Models: 4-(Dimethylamino)pyridine-catalyzed esterirfication of Butanol", Anal. Chem., **2004**, 76, 2575-2582.

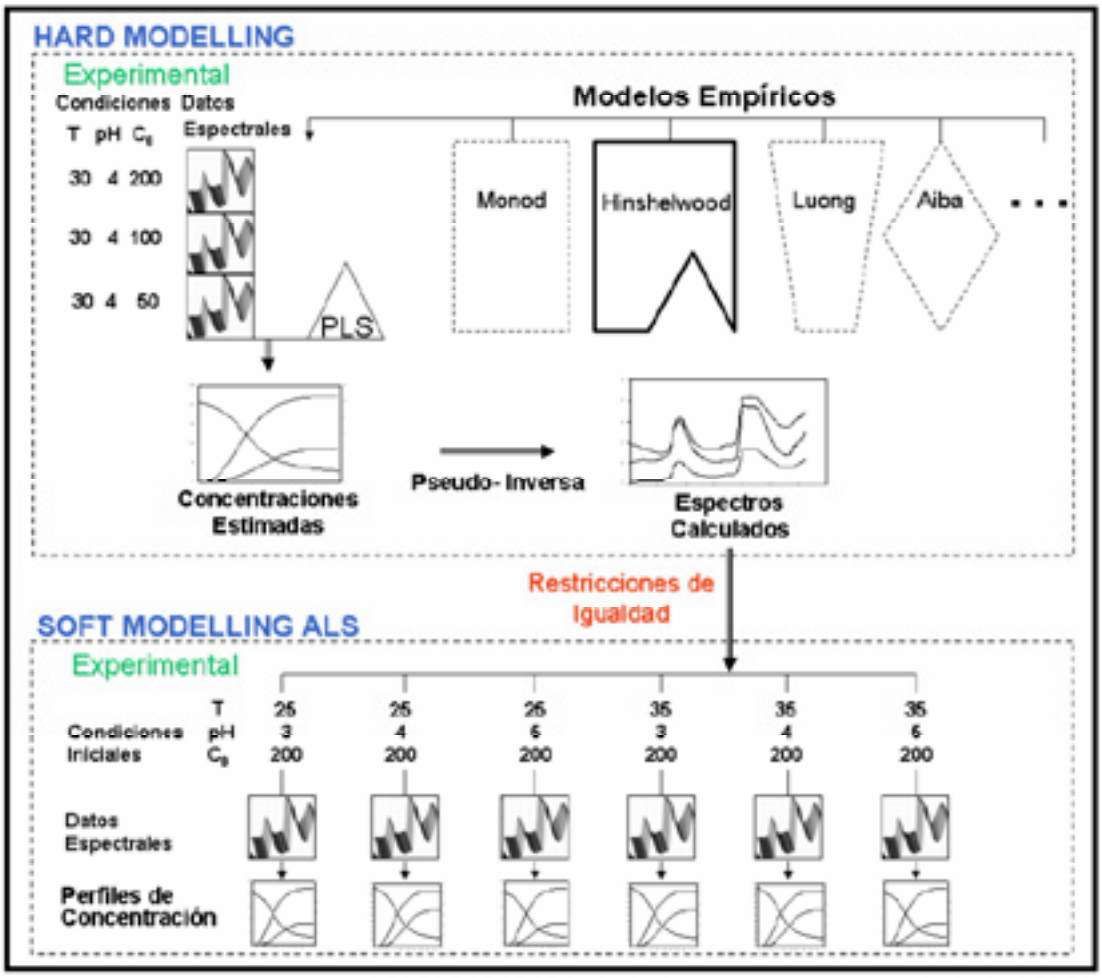


Figura 4.7. Esquema del proceso de combinación de los modelos empíricos y del modelo de resolución ALS por medio de las restricciones de igualdad.

En la primera fase, *hard modelling*, los modelos recogidos en la tabla 4.3 fueron utilizados para evaluar su capacidad descriptiva del proceso fermentativo. El modelo que presentó una mayor coherencia y consistencia de los parámetros, para fermentaciones realizadas a diferentes concentraciones de glucosa, resultó ser el propuesto por Hinshelwood. A partir de los perfiles de concentración estimados por el modelo se calcularon los espectros asociados mediante la pseudoinversa. De esta manera, los espectros calculados sirvieron como nexo de unión entre las fases de *hard- y soft-modelling*, utilizándose como restricciones de igualdad dentro del algoritmo p-ALS (*Alternating Least Squares with Penalty functions*) en la segunda fase, fase de *soft-modelling*.

4.3.8 ANÁLISIS FACTORIAL PARALELO (PARAFAC)

PARAFAC (*PARAllel FACctor analysis*) puede ser considerado como una extensión del análisis en componentes principales a situaciones de multilinealidad. Fue desarrollado en 1970 independientemente por dos grupos de investigación y fue denominado como PARAFAC por Harshman [67] y CANDECOMP (*CANonical DECOMPosition*) por Carrol & Chang [68]

PARAFAC se aplica a conjuntos de datos organizados en estructuras con dimensión superior a dos (matrices), siendo las más comunes, en el campo de la química analítica, las estructuras de datos de dimensión igual a tres denominadas: "cubos", "matrices tridimensionales", "tensores de segundo orden", o, en inglés, "*3-way array*". Ejemplos típicos de estructuras tridimensionales las obtenemos con las técnicas fluorimétricas donde, para cada muestra, obtenemos una matriz de datos, resultado de agrupar los espectros de emisión y excitación. De esta forma, la estructura tridimensional surge al combinar varias muestras.

En este trabajo, las estructuras tridimensionales han sido creadas a partir de espectros NIR de diferentes muestras, siendo la temperatura la dimensión adicional que ha permitido construir estructuras tridimensionales.

Para un cubo de datos de dimensiones **(I x J x K)**, con elementos $\mathbf{x}_{ijk}$, podemos generalizar el modelo bilineal recogido en la ecuación **[4.1]** al modelo PARAFAC:

$$\mathbf{x}_{ijk} = \sum_{R=1}^{r} \mathbf{a}_{iR}\mathbf{b}_{jR}\mathbf{c}_{kR} + \mathbf{e}_{ijk} \qquad \textbf{[4.21]}$$

Donde **r** es el número de componentes usado en el modelo PARAFAC. Cada componente r esta formado por un *score* o *loading* en el primer modo $\mathbf{a}_{iR}$ y dos *loadings vectors* $\mathbf{b}_{jR}$, $\mathbf{c}_{kR}$ para los modos segundo y tercero respectivamente; $\mathbf{e}_{ijk}$ es el término residual que contiene toda la variación no explicada por el modelo. Una descripción gráfica de la descomposición PARAFAC para un modelo con tres factores se encuentra en la figura 4.8.

---

[67] Harshman R.A., "Foundations of the PARAFAC procedure: model and conditions for an explanatory' multi-mode factor analysis" UCLA Working Papers in phonetics, **1970**, 16, 1.

[68] Carroll J.D., Chang J., "Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition", Psychometrika, **1970**, 35, 283.

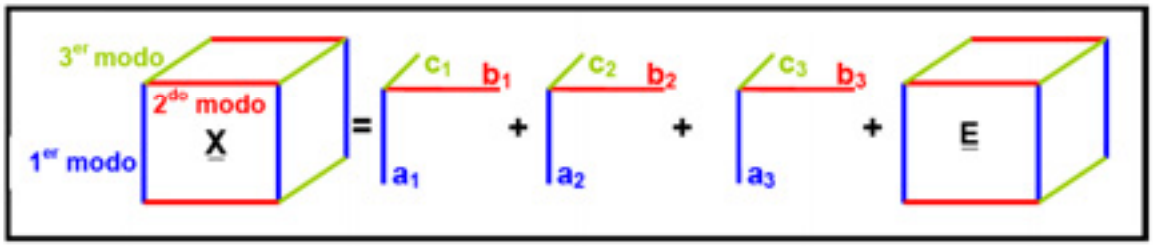


Figura 4.8. Representación gráfica de la descomposición PARAFAC para un sistema con tres componentes.

La principal diferencia entre PCA y PARAFAC es que en éste último la factorización no se hace imponiendo restricciones de ortogonalidad y los loadings obtenidos pueden ser asociados a las fuentes que provocan la variación en cada modo. Esta importante propiedad se denomina unicidad *(uniqueness)* y en sistemas sencillos y trilineales pueden ser utilizados para predecir concentraciones de muestras desconocidas [69]. Desde un punto de vista matemático, de esta importante propiedad se deriva que la descomposición en triadas al ser única, no está sometida a ningún tipo de restricciones y cualquier rotación ejercida sobre el modelo PARAFAC va acompañado de una pérdida de ajuste [70].

4.3.9 USO CONJUNTO DE MODELOS PARAFAC Y MLR

Los sistemas sencillos y trilineales raramente se encuentran en el mundo real salvo en condiciones muy controladas y vigiladas. En estos casos, PARAFAC no puede proporcionar una descomposición que sea capaz de explicar las fuentes de variación causantes de los cambios en el sistema. Pero, sin embargo, sí puede utilizarse como una técnica de reducción de variables, en un estilo similar al PCA, pero con una ventaja importante sobre el análisis en componentes principales. Mientras que PCA establece las nuevas variables latentes o componentes principales a través de la combinación lineal de las variables originales buscando maximizar la varianza espectral, PARAFAC establece los componentes principales a través de un algoritmo ALS en la dirección de las fuentes principales de

[69] Bro R., "PARAFAC. Tutorials and applications", Chemom. Intell. Lab.Syst., **1997**, 38, 149–170.

[70] Leurgans S., Ross R.T., Abel R.B., "A decomposition for three-way arrays, SIAM", J. Matrix Anal. Appl., **1993**, 14, 1064-1076.

variación. Así, en sistemas no trilineales, los componentes PARAFAC no pueden ser relacionados directamente con las fuentes que provocan variaciones en el sistema, pero sí son albaceas de importante información química y/o física que puede ser utilizada para describir el sistema.

En la figura 4.9 se puede ver el procedimiento utilizado en este trabajo para crear modelos combinando PARAFAC y MLR. Por un lado hemos utilizado PARAFAC como técnica de reducción de variables y la información recogida en los *scores* procedentes de la descomposición PARAFAC, junto con los respectivos valores de referencia (en nuestro caso, concentración de analito y temperatura), han sido utilizados para construir un modelo MLR que puede ser utilizado para la predicción de nuevas muestras que no han intervenido en el proceso de calibración.

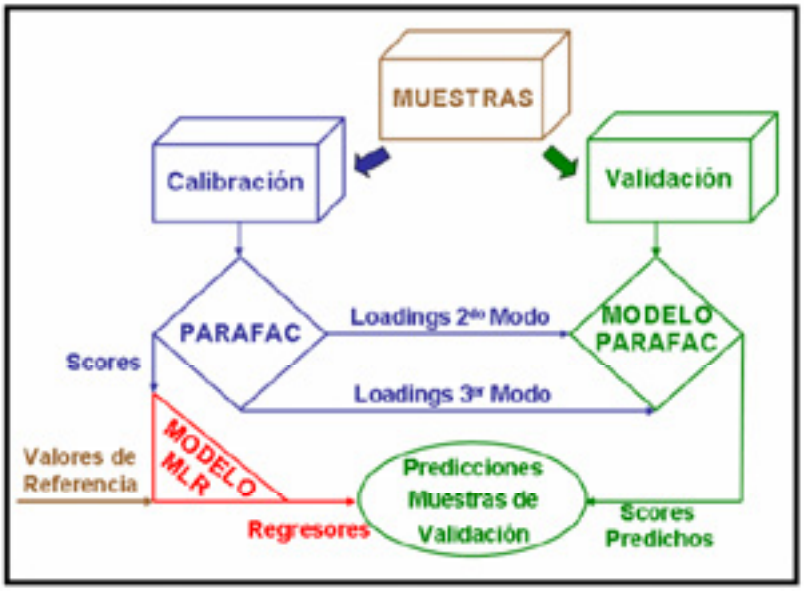


Figura 4.9. Combinación de los métodos PARAFAC y MLR propuesta en este trabajo.

5

METODOLOGÍA Y RESULTADOS

5.1 INTRODUCCIÓN 87

5.2 METODOLOGÍA EXPERIMENTAL 88

5.2.1 Establecimiento de un proceso fermentativo 88

5.2.2 Sistemas para estudiar el efecto de la temperatura en la habilidad predictiva de los modelos 93

5.2.3 Métodos analíticos de referencia 97

5.3 RESULTADOS 108

5.3.1 Análisis de Espectros 108

5.3.2 Modelos PLS utilizados en el seguimiento de fermentaciones alcohólicas. 115

5.3.3 Modelos MCR-ALS 119

5.3.4 Uso conjunto de MCR-ALS y modelos empíricos 127

5.3.5 Modelos PARAFAC-MLR 137

5.1 INTRODUCCIÓN

Como ha sido expuesto en capítulos anteriores, la producción y utilización de bioetanol se encuentran en una situación de expansión. Sin embargo el éxito de su consolidación y aceptación está condicionado a la optimización del proceso de una forma eficaz técnica y económicamente.

Las vías para aumentar la eficiencia del proceso de fermentación son muy diversas y abarcan desde la modificación genética, para buscar cepas de levadura altamente productoras [1], la optimización de las condiciones de fermentación [2], o incluso, la combinación del proceso fermentativo con el de extracción [3].

Otro importante aspecto a tener en cuenta, a la hora de mejorar un proceso productivo, es el desarrollo de una estrategia de análisis y monitorización eficiente que permita obtener información en tiempo real y mantener el proceso bajo condiciones óptimas. Sin embargo, en el caso de los bioprocesos, esta tarea no es fácil debido a la naturaleza compleja del metabolismo microbiano así como a las no-linealidades de su cinética [4]. En este contexto es donde se enmarca este trabajo, cuyo objetivo general es la aplicación in-line de metodologías analíticas basadas en la combinación de medidas espectroscópicas de infrarrojo cercano con métodos quimiométricos de análisis multivariante.

En este capítulo se recoge la metodología instrumental y analítica utilizada así como los resultados de los paulatinos y sucesivos estudios realizados para alcanzar el objetivo general planteado.

---

[1] Helle S., Murria A., Lam J., Cameron D., Duff S. "Xylose fermentation by genetically modified *Saccharomyces cerevisiae* 259ST in spent sulfite liquor". Bioresource Tech., **2004**, 92, 163-171.

[2] Costa A., Atala D., Maugeri F., Maciel F. "Factorial design and simulation for the optimization and determination of control structures for an extractive alcoholic fermentation". Process Biochem.,**2001**, 37, 125-137.

[3] Silva F.L.H., Rodrigues M.I., Maugeri F. "Dynamic modelling, simulation and optimization of an extractive continuous alcoholic fermentación". J. Chem. Tech. Biotech., **1999**, 74, 176-182.

[4] Costa A.C., Meleiro L.A.C., Maciel Filho R. "Non-linear predictive control of an extractive alcoholic fermentation process". Process biochemistry, **2002**, 38, 743-750.

5.2 METODOLOGÍA EXPERIMENTAL

La distinta naturaleza de los sistemas estudiados, de las técnicas utilizadas y de los métodos quimiométricos aplicados conduce a metodologías y estrategias de trabajo diferentes en cada caso. Por ello, en cada epígrafe de esta sección se va a diferenciar entre sistemas fermentativos y sistemas químicos.

5.2.1 ESTABLECIMIENTO DE UN PROCESO FERMENTATIVO

Todas las fermentaciones fueron realizadas con *Saccharomyces cerevisiae* ATCC 1322 (*American Type Culture Collection*). La levadura fue conservada en placas petri con un medio de YPD agar (10 g/L extracto de levadura, 20 g/L peptona, 20 g/L glucosa) a 4ºC.

El medio de cultivo en el que se llevaron a cabo las fermentaciones fue un medio completo (medio de Wickerman) compuesto por 5 g/L de extracto de levadura, 5 g/L de peptona y 3 g/L de extracto de malta, suplementado con una concentración variable de glucosa comprendida entre 200 g/L y 50 g/L. Durante la fermentación el medio se mantuvo en agitación de 400 rpm para mantener la homogeneidad y evitar la floculación de la levadura.

5.2.1.1 Seguimiento en discontinuo

Las fermentaciones, de las cuales se extrajeron muestras que sirvieron para construir y validar los diferentes métodos analíticos y modelos de calibración multivariable, fueron inoculadas con precultivos (*overnight*) de células mantenidas en Erlenmeyer de 250 ml a 25ºC en un medio de Wickerman con 200 g/L de glucosa durante 48 horas. Estas fermentaciones fueron realizadas en un bioreactor de tres litros, equipado con un sistema de agitación y una resistencia eléctrica acoplado a una sonda Pt-100 para el control de la temperatura, aunque ésta no siempre fue controlada. En la figura 5.1 se muestra el bioreactor utilizado.

Figura 5.1. Bioreactor utilizado para llevar a cabo las fermentaciones, de donde se extrajeron las muestras utilizadas para construir los modelos multivariables.

Los espectros NIR, tanto de las muestras procedentes de las diferentes fermentaciones como de las muestras sintéticas realizadas por pesada mediante mezcla, fueron registrados con el equipo NIR que aparece en la figura 5.2. El módulo acoplado al instrumento es un RCA *(Rapid Content Analyzer)* que permite el registro tanto de muestras sólidas como líquidas. La disposición de la batería de detectores que contiene el RCA aparece recogida en la esquina inferior izquierda de la figura 5.2.

El rango espectral fue el comprendido entre 1100 y 2500 nm con un paso óptico de 2 nm, siendo cada espectro el promedio de 32 barridos (*scans*). Igualmente, aparece recogido el dispositivo utilizado para contener la muestra durante el registro espectral que está compuesto por: una cubeta de vidrio óptico de fondo plano y un reflector de oro que, además, sirve para establecer y fijar el camino óptico (1 mm).

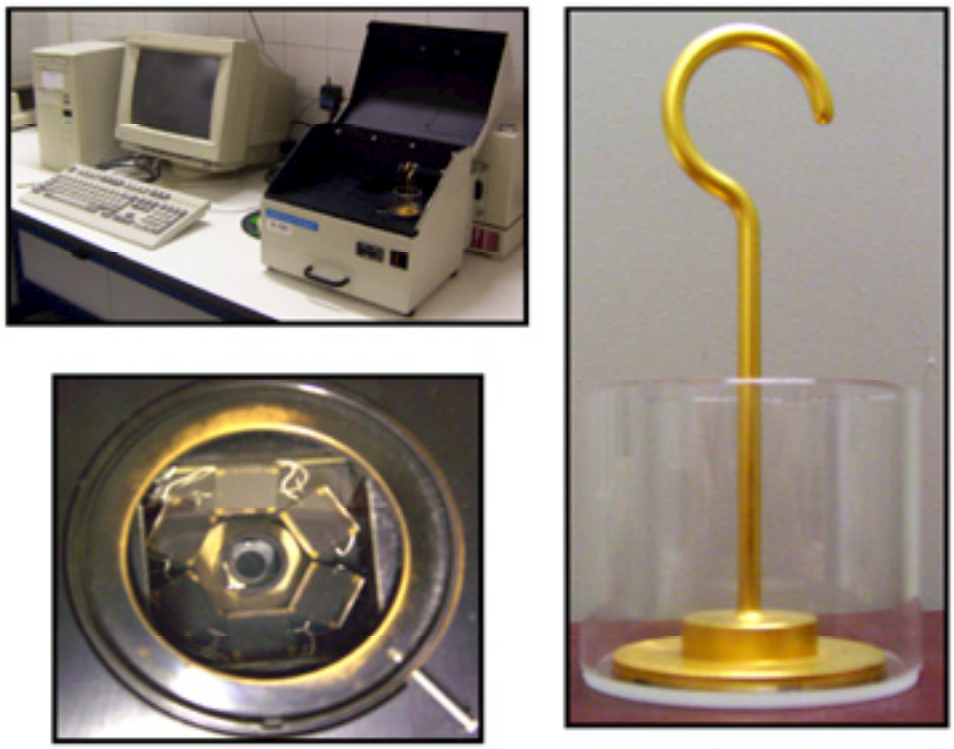


Figura 5.2. Fotografía del módulo RCA utilizado para el registro de espectros (Sup. izq.). Disposición espacial de los elementos que conforman el detector (Inf. izq.). Contenedor de muestra formado por una cubeta de cuarzo y un reflector de oro que además sirve para fijar el camino óptico (dcha.).

5.2.1.2 Seguimiento en continuo

Las fermentaciones monitorizadas "in-line", mediante una sonda de fibra óptica de inmersión, fueron inoculadas directamente transportando la levadura, con la ayuda de un asa de siembra, desde la placa de petri al bioreactor. De esta manera la influencia del inóculo inicial sobre el desarrollo de la fermentación fue minimizada.

En este caso, el bioreactor utilizado tenía un volumen útil de un litro y estaba equipado con una doble camisa para termostatizar por circulación de agua y una tapa de silicona en la cual se encontraban insertos un dispositivo para la toma de muestras y el cabezal de la sonda, ver figura 5. 3.

Figura 5.3. Montaje experimental donde se recoge el bioreactor, con la sonda de fibra óptica insertada y el baño termostático utilizado para mantener constante la temperatura, (izquierda). Bioreactor en funcionamiento tras 72 horas de fermentación (derecha).

El motivo de la utilización de un reactor pequeño y manejable no fue trivial, ya que uno de los aspectos más importantes, para el registro correcto de los espectros, es que el camino óptico de la sonda permanezca constante y, en una fermentación alcohólica, este hecho no resulta fácil debido, por ejemplo, a la producción de CO_2 y a la progresiva acumulación de biomasa en el medio.

La tapa de silicona flexible del bioreactor permitió mover y fijar la sonda en una posición de unos 40º, de tal manera que, el cabezal de la misma, quedaba próximo a la zona de régimen turbulento del vortex originado por la agitación magnética. Esta posición permitió la adecuada monitorización espectral de las fermentaciones, evitando la acumulación de biomasa en el camino óptico y favoreciendo la eliminación de las burbujas generadas, figura 5.4.

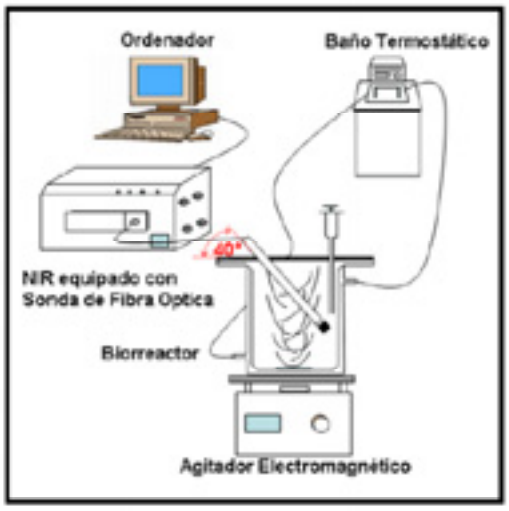


Figura 5.4. Posicionamiento de la sonda de fibra óptica NIR, en relación al vortex originado por la agitación magnética, que permitió el registro correcto de los espectros. El ángulo de unos 40º, formado por la sonda y el borde superior del fermentador, fue el que aportaba una solución de compromiso, proporcionando una mayor inclinación sin tocar las paredes de la camisa de termostatización y quedando en la zona postrera del régimen turbulento.

Otro problema, que presentan ciertas fibras de sonda óptica, es que no disponen de un dispositivo para seleccionar y establecer el camino óptico de una forma reproducible y la utilización de galgómetros no proporciona resultados suficientemente reproducibles. Para establecer y fijar el camino óptico se ha utilizado una arandela tórica de teflón de 0.5 mm que, al ser insertada entre el vástago y el cabezal de la sonda, permite las operaciones de extracción, mantenimiento y limpieza de la sonda y, posteriormente, fijar nuevamente el camino óptico de forma reproducible. En la figura 5.5, se puede observar el instrumento NIR utilizado acoplado a la sonda de fibra óptica utilizada, (izquierda), el cabezal de la sonda con la arandela de teflón utilizada para establecer el camino óptico se recoge en la figura 5.5 derecha.

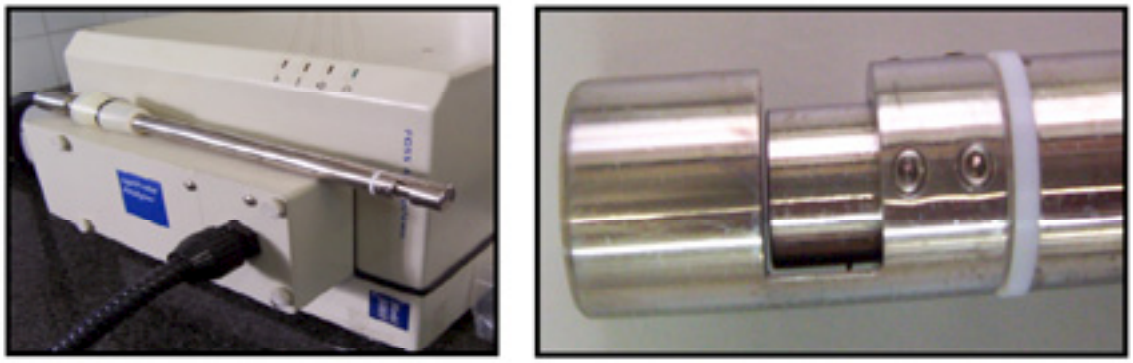


Figura 5.5. Equipo NIR acoplado a la sonda de fibra óptica utilizada (izquierda). Detalle del cabezal de la sonda y de la arandela tórica de teflón, de color blanco (derecha).

5.2.2 SISTEMAS PARA ESTUDIAR EL EFECTO DE LA TEMPERATURA EN LA HABILIDAD PREDICTIVA DE LOS MODELOS

La temperatura es un factor que causa distorsión y variación en los espectros NIR. Estas variaciones espectrales afectan negativamente a la capacidad predictiva de los modelos creados si la temperatura de los conjuntos de calibración y validación (predicción externa) es diferente. El algoritmo propuesto en este trabajo para minimizar este engorroso y molesto problema se ha aplicado a dos conjuntos de datos; uno estaba compuesto por espectros NIR registrados "in-line" con una sonda de fibra óptica de inmersión, entre las longitudes de onda de 1100-2500 nm y el otro lo componían espectros NIR registrados "at-line" entre 400-1100 nm. En los apartados siguientes se describen ambos conjuntos de datos.

5.2.2.1 Sistema en continuo

El registro in-line de espectros del sistema químico se llevó a cabo en un reactor equipado con una unidad de control de temperatura (± 0.2ºC), agitación y dosificación, tal y como aparece esquematizado en la figura 5.6.

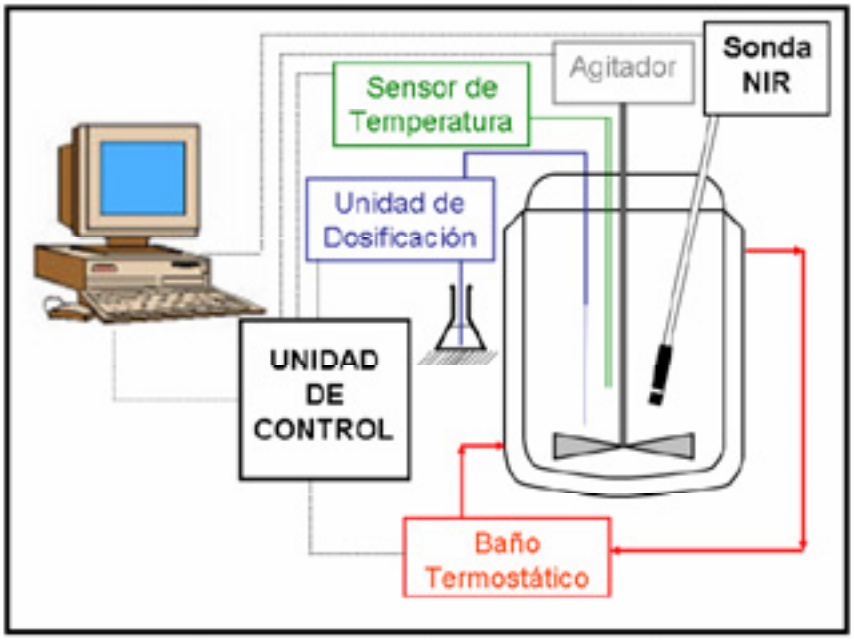


Figura 5.6. Esquema del sistema experimental utilizado para el control analítico de la temperatura.

El conjunto de espectros NIR estuvo compuesto por seis lotes (*batches* o *runs*), a los que se le aplicó un perfil común de variación de temperatura. Tres de los lotes eran de especies puras: agua, etanol y glicerina. Los otros tres lotes estaban formados por mezclas binarias: de glicerina diluida con etanol, de glicerina diluida con agua y de etanol diluido con agua, cuya composición relativa iba evolucionado con el tiempo. En la figura 5.7 aparece recogido el perfil de temperatura, aplicado a todos los lotes, y el patrón de dilución aplicado a los lotes mezcla. Los valores de referencia para temperatura y concentración de analitos que se utilizaron para construir los modelos multivariantes fueron los reales suministrados por la unidad de control al reproducir los perfiles teóricos de la figura 5.7.

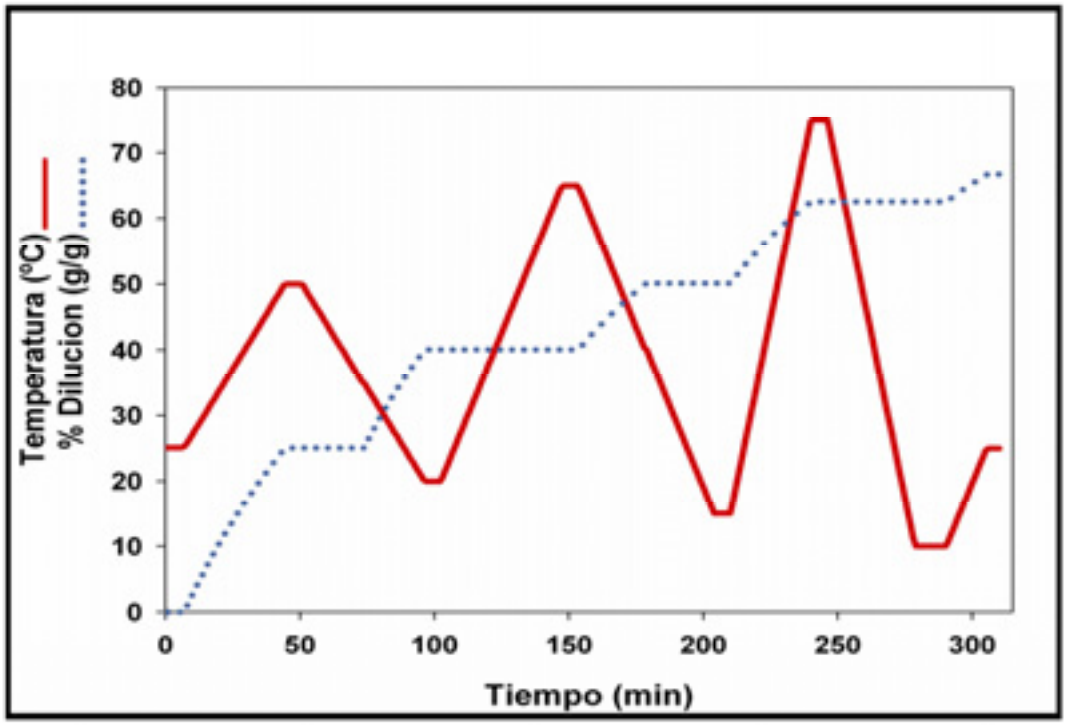


Figura 5.7. Perfil de temperatura, aplicado a todos los lotes (rojo) y patrón de dilución utilizado en los lotes formados por mezcla (azul).

La organización y ensamblado de los diferentes lotes para constituir una estructura numérica tridimensional viene recogida en la figura 5.8. Los dos primeros lotes estaban constituidos por un analito diferente cada uno de ellos.

El tercer lote era una mezcla binaria, de los analitos anteriores, que seguía el perfil de dilución temporal recogido en la figura 5.7.

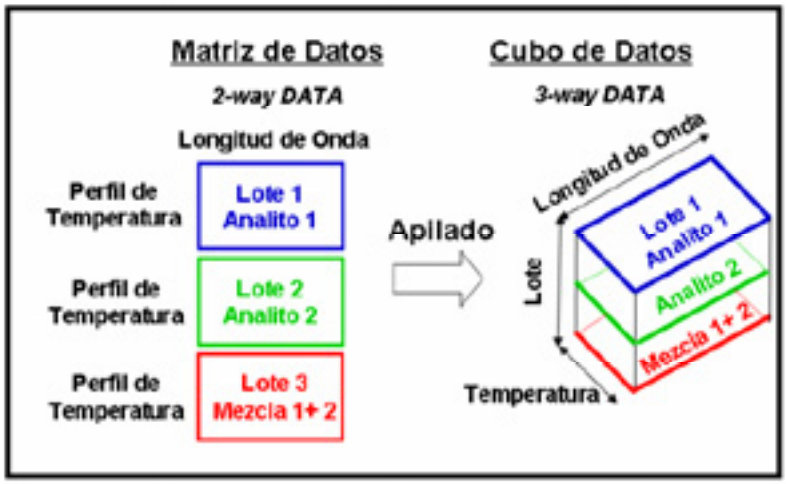


Figura 5.8. Esquema de la formación de estructuras de datos tridimensionales a partir de los diferentes lotes.

5.2.2.2 Sistema en discontinuo

El conjunto de datos registrados por Wülfer et al. [5], se ha convertido en un "banco de pruebas" que es utilizado por diferentes grupos de investigación [6-9] para probar y contrastar la habilidad predictiva de sus algoritmos.

En la tabla 5.1 se recoge la concentración, expresada en fracción molar (%), de las mezclas de etanol, isopropanol y agua, de las muestras utilizadas. En la figura 5.9 se representa la distribución de las mezclas en un gráfico de proporciones. Las muestras en azul son las que fueron utilizadas para construir el modelo (conjunto de calibración), las muestras marcadas en rojo fueron utilizadas

---

[5] Wülfer F., Kok T., Smilde A.K. "Influence of Temperature on Vibrational Spectra and Consequences for the Predictive Ability of Multivariate Models". Anal. Chem., **1998**, 70, 1761-1767.

[6] Swierenga H., Wülfert F., De Noord O.E., De Weijer A.P., Smilde A.K., Buydens L.M.C. "Development of robust calibration models in near-infrared spectrometric applications". Anal. Chim. Acta, **2000**, 411, 121-135.

[7] Eilers P., Marx B. "Multivariate calibation with temperature interaction using two-dimensional penalized signal regression". Chemom. Intell. Lab. Syst., **2003**, 66, 159-174

[8] Thissen U., Üstün B., Melssen W.J., Buydens L.M.C. "Multivariate Calibration with LS Support Vector Machines". Anal. Chem. **2004**, 76, 3099-3105

[9] Marx B., Eilers P., "Multivariate calibration stability: a comparison of methds", J. Chemometrics, **2002**, 16, 129-140

para validar externamente el modelo (conjunto de validación). El espectro NIR de cada una de las muestras fue registrado a las temperaturas de 30, 40, 50, 60 y 70ºC

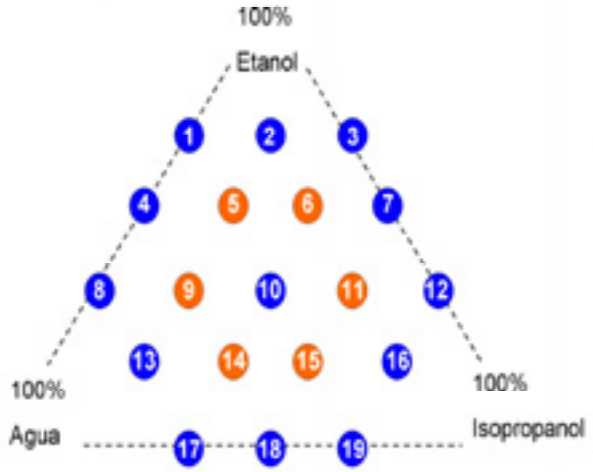


Figura 5.9. Diseño de la composición de las mezclas de etanol, isopropanol y agua. En color azul están representadas las muestras pertenecientes al conjunto de calibración y en rojo las de validación.

Tabla 5.1. Fracción molar (%) de los tres analitos que constituyen las diferentes mezclas.

	Etanol	Agua	iso-propanol
1	66.4	33.6	0.0
2	67.2	16.3	16.5
3	66.6	0.0	33.4
4	50.0	50.0	0.0
5	50.0	33.3	16.7
6	49.9	16.7	33.3
7	50.0	0.0	50.0
8	33.3	66.7	0.0
9	33.2	50.0	16.7
10	33.3	33.4	33.3
11	32.2	16.6	51.2
12	33.5	0.0	66.5
13	16.6	66.7	16.7
14	16.7	50.0	33.3
15	16.6	33.3	50.1
16	16.2	16.3	67.5
17	0.0	66.7	33.3
18	0.0	50.0	50.0
19	0.0	33.4	66.6

La organización y ensamblado de las muestras para conformar una estructura numérica tridimensional o cubo de datos (*3-way data structure*) susceptible de ser utilizada para aplicar PARAFAC, o cualquier otra técnica *3-way*, viene recogida en la figura 5.10. Como se puede ver, cada una de las dimensiones del cubo esta relacionada con una propiedad de la muestra, o bien con su espectro (dirección longitud de onda) o con su composición química (dirección mezclas) o bien con su temperatura. El caso representado en la figura 5.10 es el más general donde el cubo de datos se ha construido con las cinco temperaturas, necesitándose, como mínimo, dos temperaturas para formar la estructura tridimensional.

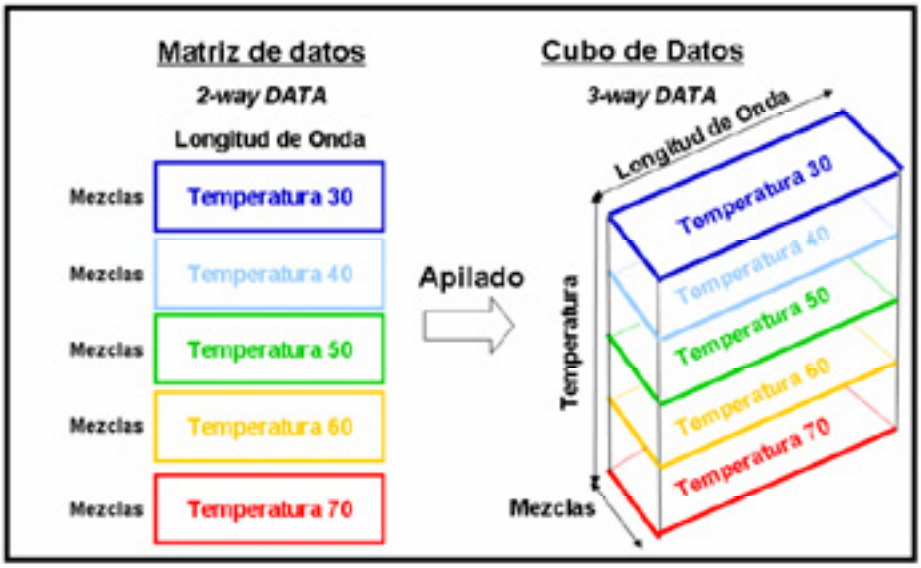


Figura 5.10. Formación de un cubo de datos a partir de los espectros de las muestras registrados a diferentes temperaturas.

5.2.3 MÉTODOS ANALÍTICOS DE REFERENCIA

La exactitud y precisión asociadas a los valores analíticos de referencia de las muestras utilizadas en la calibración de los modelos multivariantes son propiedades esenciales y deseables si se busca crear modelos robustos y fiables. Es por esto, que el establecimiento de métodos analíticos exactos y reproducibles supone una fase básica y de suma importancia en la creación de modelos inversos. En la figura 5.11 aparece esquematizado las técnicas instrumentales utilizadas en el contexto del proceso de creación de modelos PLS.

5.2.3.1 Determinación de acidez

La acidez fue determinada por titración de 5 ml de muestra de fermentación con NaOH 1M utilizando fenoftaleína como indicador. La acidez total fue expresada como contenido en ácido acético.

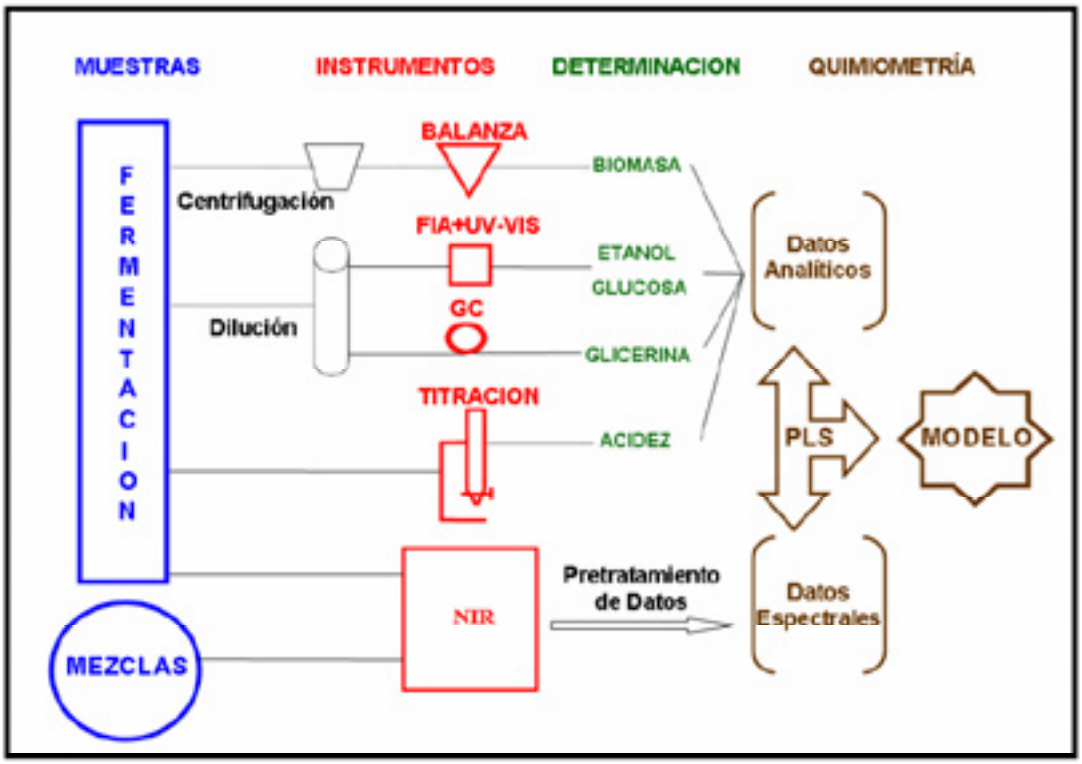


Figura 5.11. Procedimiento utilizado para la obtención de modelos PLS. GC: Cromatografía de Gases. FIA+UV-VIS: Análisis por Inyección en Flujo con detector Ultravioleta-Visible.

5.2.3.2 Determinación de glicerina

El contenido en glicerina fue determinado por cromatografía de gases utilizando un cromatógrafo con detector de ionización por llama. La columna capilar utilizada tenía unas dimensiones de 15 m × 0.25 mm (diámetro interno) y un espesor de la fase estacionaria de 0.25 µm. El método de separación usaba como patrón interno palmitato de metilo y una rampa de temperatura de 7°C/ min desde los 120 a los 270°C.

5.2.3.3 Determinación de biomasa

La determinación más común de biomasa suele realizarse mediante métodos físicos, ya sea por filtración [10] o mediante extracto seco [11]. Aunque ambos métodos presentan precisiones similares [12], éste último fue el elegido, en nuestro caso, por ser más rápido y barato. El protocolo de trabajo seguido fue el siguiente:

1. Se limpian y se secan los tubos de cristal en estufa.
2. Se pesa y anota el peso de cada tubo.
3. Se extrae muestra del fermentador. Se coloca 5 ml de dicha muestra en cada tubo (por triplicado).
4. Se centrifugan las muestras durante 15 minutos a 10.000*g*.
5. Se retira el sobrenadante de los tubos mediante pipeta y se conserva el precipitado.
6. Se resuspende el precipitado en 5 ml de agua desionizada, se agita vigorosamente y se repite la centrifugación.
7. Se colocan los tubos en estufa a 105ºC durante 24 horas.
8. Se retiran los tubos del horno y se colocan en un desecador para que se enfríen.
9. Se pesan los tubos y se repiten los pasos 7 y 8 hasta que la masa permanezca constante.
10. Al peso obtenido se le resta el peso del tubo vacío y, de esta manera, se calcula el peso seco de cada muestra por triplicado.

---

[10] Nolasco C., Matsunaka T., Kobayashi G., Sonomoto K., Ishizaki A. "Synchronized fresh cell bioreactor system for continous L(+)-lactic acid production using *Lactococcus Lactis* in hydrolysed sago starch", J. Biosci. Bioeng., **2002**, 93(3), 281-287.

[11] Zhihong G., Cavinato A.G., Callis J.B. "Noninvasive Spectroscopy for Monitoring Cell Density in a Fermentation Process", Anal. Chem., **1994**, 66, 1354-1362.

[12] Stone K., Roche F., Thornhill N.F. "Dry weight measurement of microbial biomass and measurement variability analysis", **1992**, Biotech. Tech. 6(3), 207-212.

5.2.3.4 Determinación de glucosa y etanol

Al empezar este trabajo y estudiar las técnicas potenciales que podíamos utilizar para la determinación de glucosa y etanol, una de las características deseables que buscábamos era poder extraer la muestra del sistema fermentativo y analizarla directamente sin realizar ningún tipo de operación previa o pretratamiento, por esta razón, se optó por utilizar un sistema de análisis por inyección en flujo FIA (*Flow Injection Analysis*), con un detector Ultravioleta-Visible (UV-VIS).

5.2.3.4.1 Interacción entre analitos

Al diseñar el sistema FIA, se intentó poder determinar simultáneamente ambos analitos y que las muestras pudieran ser inyectadas directamente, así llegamos a una configuración de un sistema FIA complejo, en el cual se perdían las ventajas inherentes a la inyección en flujo. Además, dicho sistema presentaba interferencias mutuas de la glucosa en la determinación de etanol y del etanol en la determinación de glucosa. Este tipo de interferencias también han sido descritas por otros autores [13].

Para evitar estas interacciones entre analitos se replanteó el diseño del sistema FIA y se estableció uno más sencillo, fiable y reproducible, pero en detrimento de tener que realizar un tratamiento previo de dilución a las muestras, antes de introducirlas en el sistema, en una proporción comprendida entre 1/100 y 1/200. En la figura 5.12 aparece esquematizado el sistema FIA empleado.

Se realizaron una serie de estudios previos para determinar si el sistema FIA diseñado presentaba especificidad para la determinación de glucosa y etanol. A modo de ejemplo en la figura 5.13, se muestra el fiagrama obtenido para determinar si la presencia de etanol afectaba significativamente la exactitud en la determinación de glucosa, en él se recoge la señal proporcionada por el análisis de una serie de muestras sintéticas con concentraciones de glucosa (2.02, 1.5, 0.95, 0.58 y 0.25 g/L) a dos niveles de concentración de etanol (0.007 y 0.12 %).

---

[13] Rhee J.I., Shügerl K., "The influence of metabolites on enzyme based flow injection analysis", Anal. Chim. Acta, **1997**, 355, 55-62.

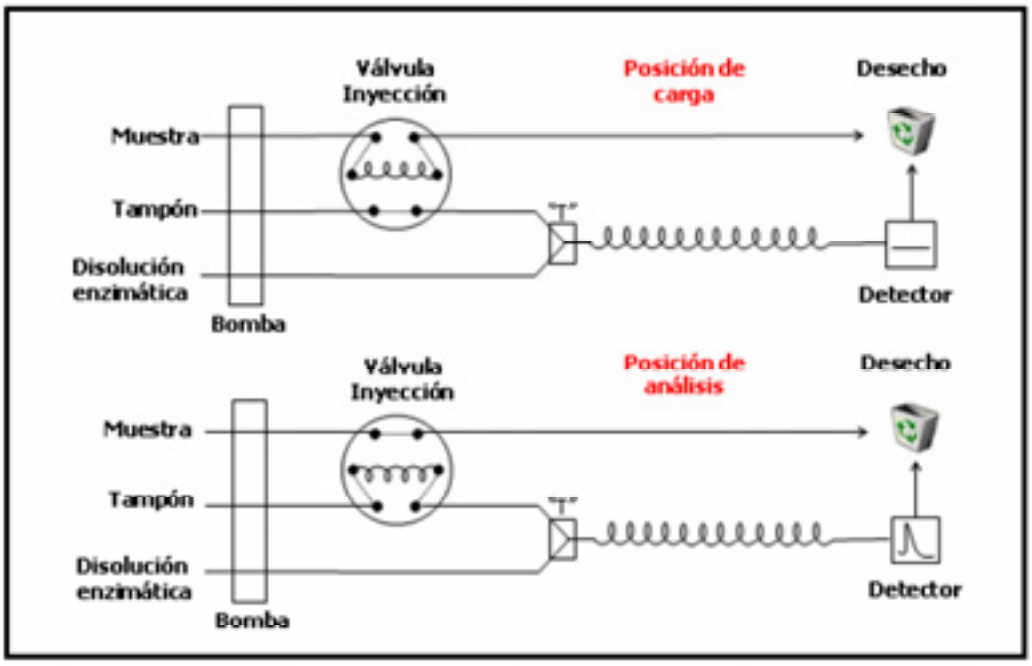


Figura 5.12. Disposición de los elementos que integran el sistema FIA utilizado.

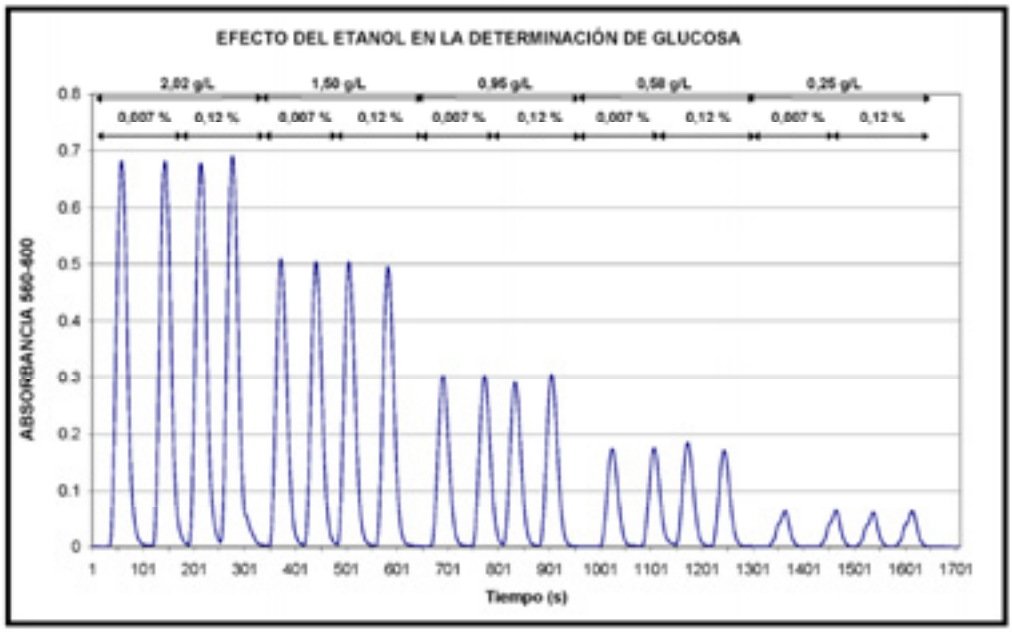


Figura 5.13. Estudio de la influencia de la concentración de etanol en la determinación de glucosa.

Las posibles diferencias físicas (índice de refracción, viscosidad, turbidez, etc.) entre la disolución de la muestra y la disolución portadora donde ésta es insertada, fueron solventadas corrigiendo la absorbancia a la longitud de onda

analítica (340 nm para la determinación de etanol y 506 nm para la determinación de glucosa) por una absorbancia a una longitud de onda de referencia (600 nm en ambos casos) [14-15].

Tabla 5.2. ANOVA para determinar la influencia del etanol en la concentración de glucosa.

Factor	Grados Libertad	Varianza	F	Valor crítico F
Etanol	1	1.2005E-06	0.03	4.96
Glucosa	4	0.24795996	5894.90	3.48
Etanol*Glucosa	4	2.1652E-05	0.51	3.48
Residual	10	4.2063E-05		
Total	19			

Tomando como variable respuesta la absorbancia máxima en cada pico del fiagrama, se realizó un análisis de la varianza (ANOVA) para determinar la significación de los factores: concentración de glucosa, concentración de etanol e interacción glucosa-etanol. Como se puede deducir de la tabla 5.2, el único factor significativo en la determinación de glucosa, a un nivel de significación de un 95%, es la concentración de glucosa ya que el valor crítico de F para la concentración de glucosa (3.48) es menor que el valor F tabulado para glucosa (5894.9) a un nivel de significación del 95%; con lo que se puede concluir que el sistema FIA propuesto presenta especificidad para la determinación de glucosa.

Similares conclusiones fueron obtenidas cuando se estudio la especificidad del sistema para la determinación de etanol.

5.2.3.4.2 Calibración para la determinación de glucosa

La determinación de glucosa se realizó enzimáticamente, basándose en el acoplamiento de las siguientes reacciones a pH 7.0 [16-17].

---

[14] Rothman L.D., Crouch S.R., Ingle J.D., "Theoretical and experimental investigation of factor affecting precision in molecular absorption spectrophotometry", Anal. Chem., **1975**, 47(8) 1226-1233.

[15] Zagatto E., Arruda M., Jacinto A., Mattos I. "Compensation of the Schlieren effect in flow-injection analysis by using dual-wavelength spectrophotometry", Anal. Chim. Acta., **2000**, 234, 153-160.

[16] Medina M.J., Bartrolí J., Alonso J., Blanco M., Fuentes J., "Direct determination of glucose in blood serum using Trinder's reaction", Anal. Lett., **1986**, 17(B5), 385-396.

[17] Valero F., La Fuente J., Poch M., Sola C., "On-line fermentation monitoring using flow injection analysis", Biotechnology and Bioengineering, **1990**, 36, 647-651.

La primera reacción es catalizada por la enzima glucosa oxidasa (GOD) y en ella, en presencia de oxigeno, la glucosa es transformada en ácido glucónico y peróxido de hidrógeno. Éste último, junto con la 4-aminofenazona y el fenol, son los sustratos de la segunda reacción, catalizada por la peroxidasa (POD), produciéndose agua y monoimino-p-benzoquinona-4fenazona, la cual presenta un máximo de absorción a una longitud de onda de 506 nm.

$$\text{Glucosa} + O_2 + H_2O \xrightarrow{GOD} \text{Ácido Glucónico} + H_2O_2$$

$$2\,H_2O_2 + \text{4-aminofenazona} + \text{fenol} \xrightarrow{POD} \text{monoimino-p- benzoquinona- 4 fenazona} + 4\,H_2O$$

Las concentraciones de los analitos presentes en las disoluciones utilizadas fueron las siguientes:

- Disolución tampón:
  - 0.1 M Na_2HPO_4 ajustado a pH 7.0
- Disolución enzimática, preparada a partir de la disolución tampón anterior:
  - 12.5 Unidades GOD/ml
  - 5 Unidades POD/ml
  - 9.3 mM fenol.
  - 1.5mM 4-aminofenazona.

La disolución se prepara inmediatamente antes de realizar el análisis y el exceso puede ser reutilizado durante un tiempo máximo de 3 días, si el manejo es el adecuado y permanece refrigerada a 4ºC.

Las muestras tomadas del bioreactor se diluían en una proporción 1/100 en agua bidestilada antes de ser inyectadas en el sistema FIA. Aunque esta dilución es indeseable ya que aumenta el tiempo de análisis es necesaria por dos motivos:

- Para tener un método de análisis con una función de respuesta lineal entre la concentración medida y la señal obtenida.
- Por la interferencia que supone la presencia de glucosa en la determinación de etanol y viceversa.

Para la cuantificación de la concentración de glucosa presente en las muestras extraídas del sistema fermentativo se realizó una recta de calibración previa mediante cinco muestras sintéticas (estándares o muestras patrón), de concentración de glucosa conocida. En la figura 5.14 se recoge, a modo de ejemplo un fiagrama donde aparecen recogidos las muestras patrón que forman el conjunto de calibración para la cuantificación de glucosa.

La relación entre la concentración de glucosa y la absorbancia máxima para cada estándar pertinente al diagrama recogido en la figura 5.14, es altamente significativo presentando un coeficiente de determinación de 0.9996. La relación entre ellos es lineal, presenta una pendiente de 0.3508 y una ordenada en el origen de –0.0011 con un límite de confianza, para un nivel de significación del 95%, de ± 0.0064; intervalo que incluye el cero.

Es necesario la realización de una nueva recta de calibración cada vez que se prepara disolución enzimática ya que la variabilidad de la actividad enzimática en cada disolución preparada es diferente.

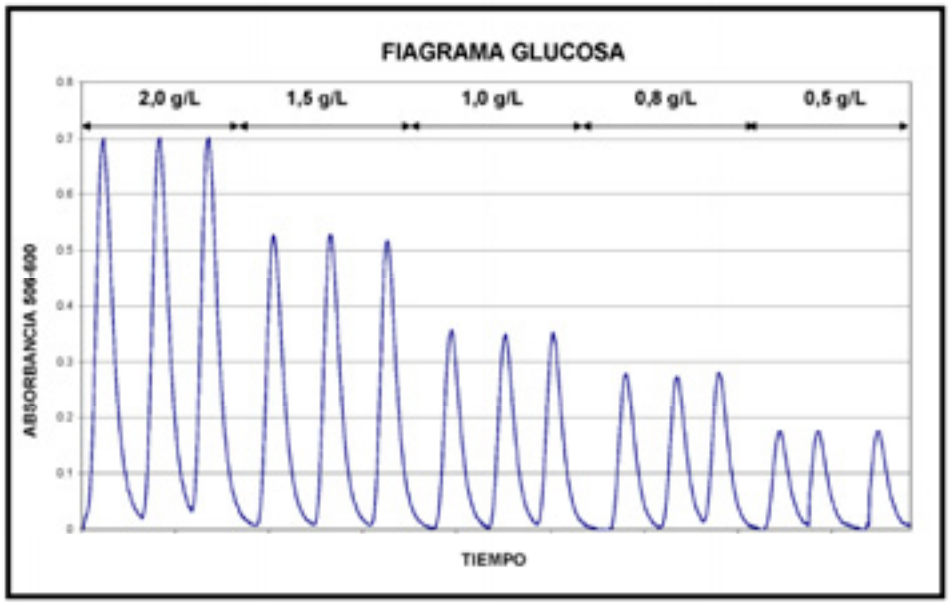


Figura 5.14. Fiagrama donde se recoge cinco muestras patrón de glucosa por triplicado.

5.2.3.4.3 Calibración para la determinación de Etanol

La determinación espectrofotométrica del etanol está basada en la reacción enzimática catalizada por la Alcohol Deshidrogenasa (ADH) a pH 8.0 [18-19].

$$CH_3CH_2OH + NAD^+ \xrightarrow{ADH} CH_3CHO + NADH + H^+$$

En dicha reacción el etanol se oxida en presencia de NAD^+ a acetaldehído y NADH el cual presenta un máximo de absorción a 340nm.

Las concentraciones de los analitos presentes en las disoluciones utilizadas fueron las siguientes:

- Disolución tampón, ajustada a pH 8.0
  - 75mM $Na_4P_2O_7$.
  - 0.15M NaCl.
  - 21 mM Glicina.
  - 75mM Semicarbazida.
- Disolución enzimática, preparada a partir de la disolución tampón anterior:
  - 130 Unidades ADH/ml
  - 1 g NAD^+/L

La disolución enzimática se preparó siempre inmediatamente antes de ser utilizada en el análisis.

Las muestras tomadas del bioreactor eran diluidas en una proporción comprendida entre 1/100 y 1/200 en agua bidestilada antes de ser inyectadas en el sistema FIA, por las razones expuestas anteriormente.

Para la determinación de etanol en las muestras extraídas del sistema fermentativo, se procede de una manera análoga a la explicada para la determinación de glucosa, es decir, previamente a la determinación hay que preparar una curva de calibración a partir de unas muestras sintéticas de

---

[18] Salgado A.M., Folly R. Valdman B., Cos O., Valero F., "Colorimetric method for the determination of ethanol by flow injection analysis", Biotechnol. Lett., **2000**, 22(4), 327-330.

[19] Rangel A., Tóth I.V., "Enzymatic determination of ethanol and glycerol by flow injection parallel multi-site detection", Anal. Chim. Acta, **2000**, 416(2), 205-210.

concentración en etanol conocidas, pero en este caso se utilizan siete estándares, que se introducen en el sistema FIA por triplicado. En la figura 5.15 se muestra, a modo de ejemplo, uno de los fiagrama realizados.

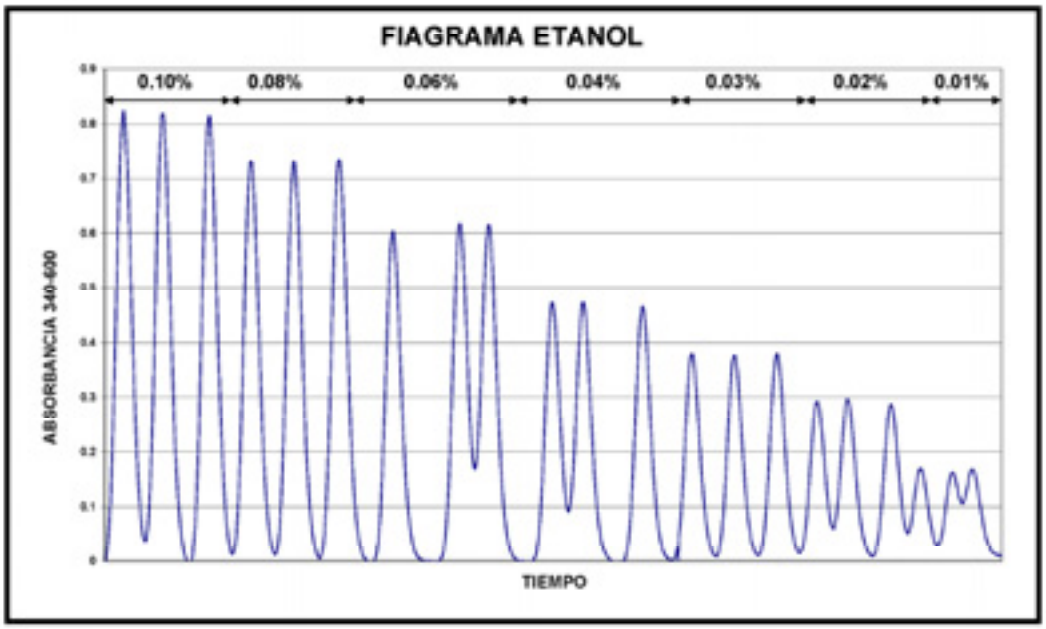


Figura 5.15. Fiagrama donde se recogen siete patrones de etanol por triplicado.

En la Figura 5.16 se recoge la relación entre la concentración de etanol y la absorbancia máxima para cada estándar, correspondiente al fiagrama de la figura 5.15. Como se puede intuir del análisis de esta gráfica y corroborar mirando el análisis de residuales frente a la concentración, figura 5.16 derecha, la relación entre estas dos variables no es lineal de primer orden; esto es debido a que la enzima alcohol deshidrogenasa presenta inhibición por producto, [20-21] y su comportamiento no sigue la típica cinética hiperbólica de Michaelis-Menten, sino que presenta una cinética parabólica [22].

[20] Lázaro F., Luque de Castro M.D., Valcárcel M. "Individual and simultaneous enzymatic determination of ethanol and acetaldehyde in wines by flow injection analysis", Anal. Chim. Acta, **1986**, 185, 57-64.

[21] Gacesa P., Hubble J., "Tecnología de las enzimas", **1990**, Ed. Acribia.

[22] Penasse L., "Les enzymes: Cinétique et Mecanisme d'Action", **1974**, Ed. Masson et al.

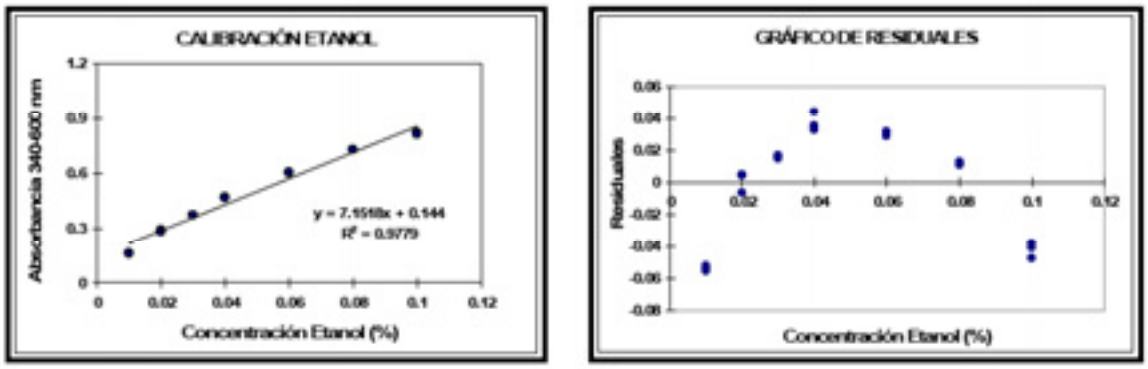


Figura 5.16. Recta de calibración para etanol y gráfico de residuales generados.

De esta forma si a los datos de la figura 5.16 le aplicamos una ecuación lineal de segundo orden, es decir parabólica, la relación obtenida presenta un coeficiente de determinación de 0.9987, figura 5.17 izquierda. La distribución de residuales para esta relación de segundo orden queda reflejada en la figura 5.17 derecha, dichos residuales no siguen ningún patrón, distribuyéndose al azar. Por este motivo y para evitar los errores cometidos en el proceso de dilución, se decidió utilizar un ajuste de segundo grado en la ecuación de calibración para la determinación de etanol.

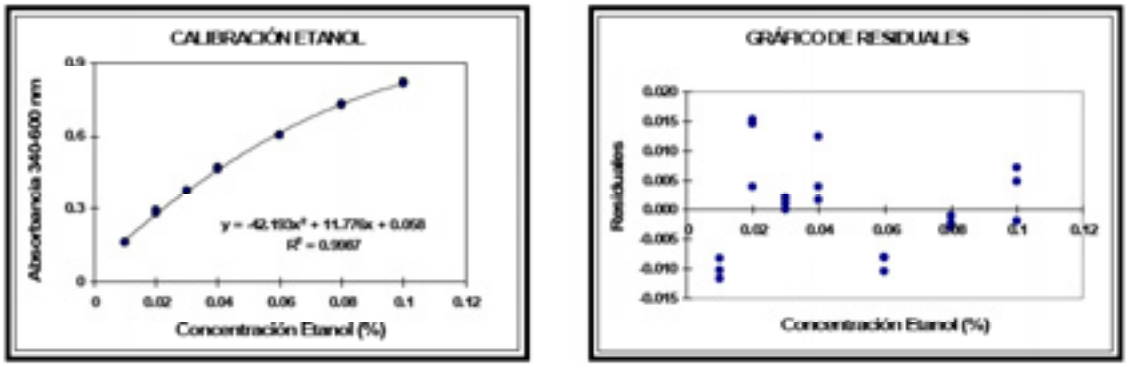


Figura 5.17. Calibración parabólica para etanol y gráfico de residuales generados.

5.3 RESULTADOS

*5.3.1 ANALISIS DE ESPECTROS**

5.3.1.1 Sistema fermentativo

La visualización, análisis y estudio de los espectros NIR registrados es una fase importante y útil del proceso de construcción de modelos multivariables en sistemas en evolución, ya que permite caracterizar y asociar determinadas regiones espectrales, a lo largo del tiempo, con las variaciones en concentración de los analitos estudiados.

En la figura 5.18 se recogen los espectros, de un conjunto de disoluciones patrón, ordenados según su contenido creciente en glucosa; se puede observar un solapamiento y abigarramiento de colores y no existe ningún intervalo espectral donde se aprecie una tendencia o evolución evidente en los espectros. La presencia de bandas de absorción anchas entorno a 1430 nm y 1940 nm corresponden respectivamente al primer sobretono y a la bandas de combinación del enlace O-H.

Al aplicar el tratamiento espectral de primera derivada, que permite eliminar los cambios aditivos en la línea base, a la ordenación espectral anterior, podemos observar una evolución espectral en la región comprendida entre 2030 nm y 2130 nm, figura 5.19, por consiguiente, esta zona espectral la podemos adscribir al contenido en glucosa y, a la hora de realizar los diferentes modelos de calibración, ha de ser considerada con especial atención. Esta región coincide con la zona donde se manifiestan las combinaciones de tono del enlace O-H de la glucosa.

[*] Las figuras que se muestran en esta sección presentan una leyenda colorimétrica que sigue el orden de colores del espectro electromagnético, esto es: rojo, naranja, amarillo, verde, azul, añil y violeta. De esta forma, el espectro que presenta un valor para la propiedad analítica analizada más bajo aparecerá en rojo y el que presente un valor más alto aparecerá en violeta.

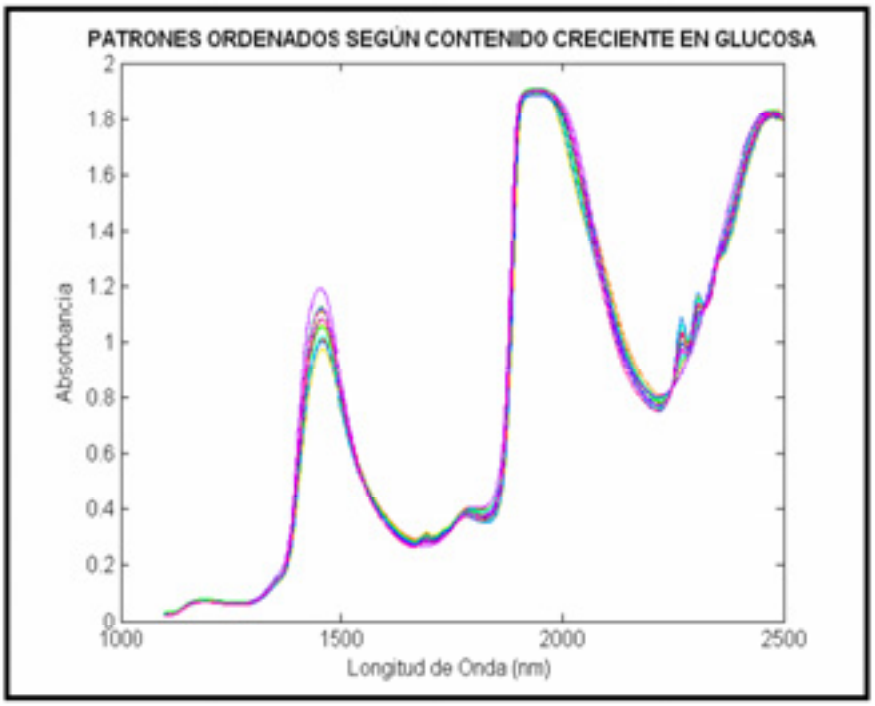


Figura 5.18. Espectros NIR en Absorbancia ordenados según contenido creciente de glucosa.

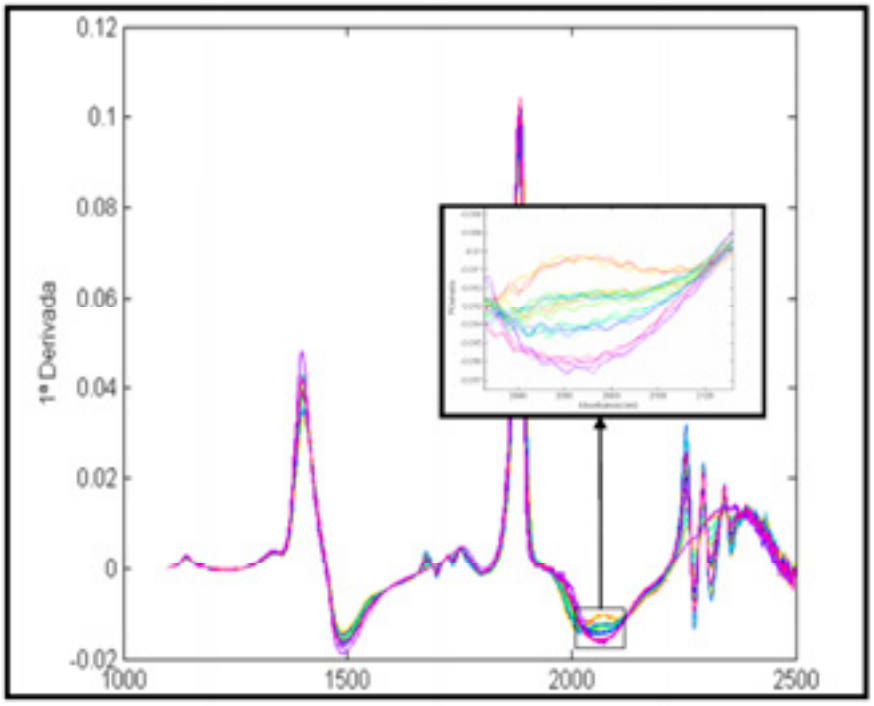


Figura 5.19. Espectros NIR en primera derivada ordenados según contenido creciente de glucosa. Detalle de la ventana espectral entre 2030 y 2130 nm.

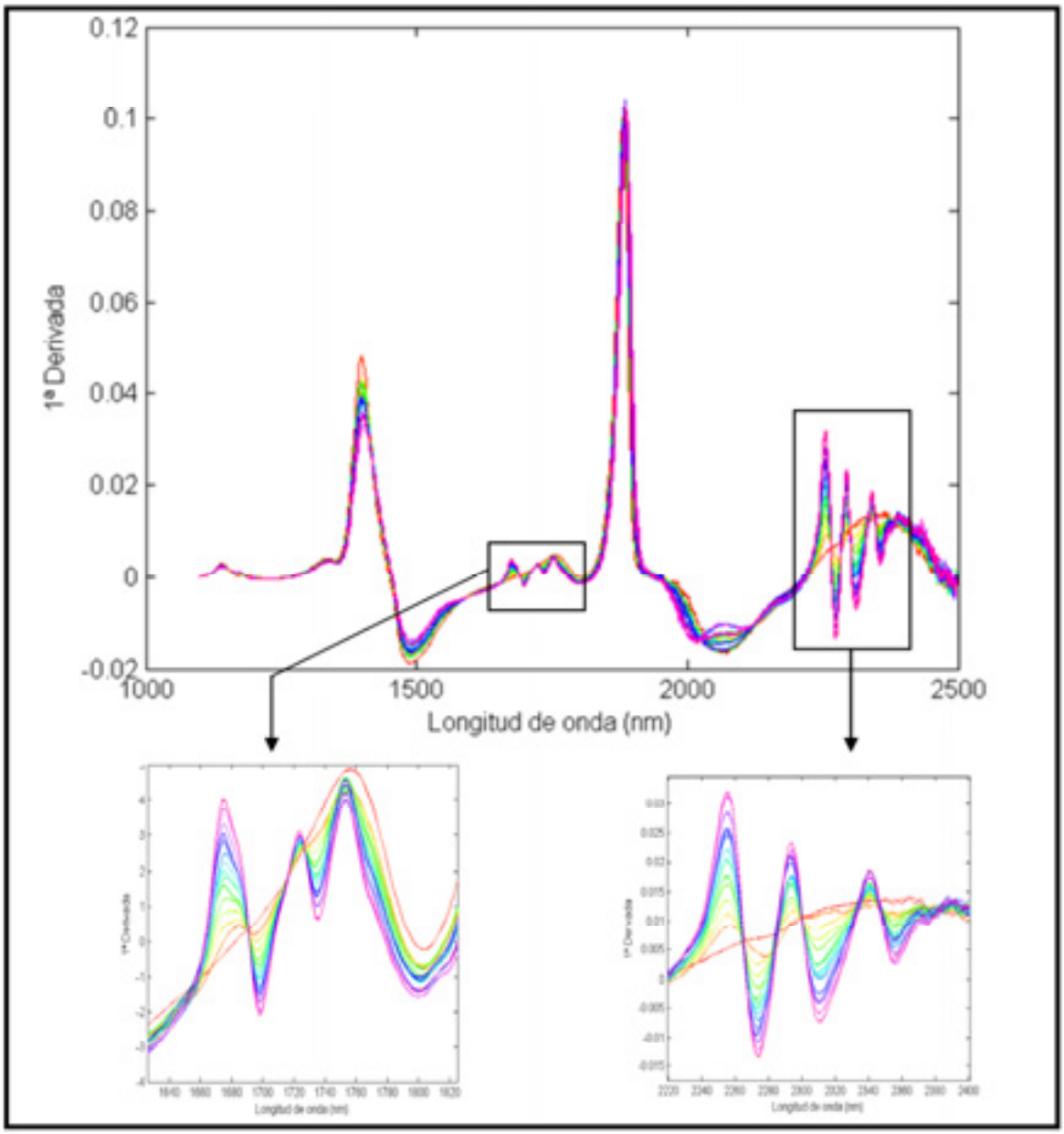


Figura 5.20. Espectros NIR en primera derivada ordenados según contenido creciente de etanol. Ventanas espectrales comprendidas entre 1750 y 1900 nm y entre 2150 y 2350 nm.

Si aplicamos el tratamiento espectral de primera derivada a los espectros ordenados en función de su contenido creciente en etanol, figura 5.20, se observa una zona espectral alrededor de 1660-1800 nm y entre 2220-2360 nm donde aparece un claro ordenamiento de los espectros, desde la concentración más baja en etanol (color rojo), a la más alta (color violeta). Estas dos regiones presentan una serie de tres máximos y tres mínimos concatenados de aspecto parecido

aunque de intensidad diferente, y se han asociado a la combinación de tonos del enlace C-H del etanol.

La región del espectro comprendida entre 2330-2500 nm, aún siendo rica en información, ya que en ella se manifiestan las bandas de combinación de diferentes estructuras, presenta una relación señal/ ruido elevada, por esta razón, esta zona espectral siempre fue considerada con cuidado a la hora de construir los modelos.

Para el resto de analitos analizados se hicieron gráficos y estudios sistemáticos de regiones, de similar forma a la descrita, aunque los resultados no fueron tan gráficos ni evidentes como en los casos descritos para glucosa y etanol.

En la figura 5.21 se puede observar los espectros en absorbancia de un proceso fermentativo realizado a una temperatura constante de 30ºC. Se puede observar como durante el transcurso de la fermentación hay un desplazamiento progresivo de la línea base debido al aumento de la dispersión en el medio de cultivo, causada principalmente por la acumulación de biomasa en el medio.

Al representar el proceso fermentativo anterior, pero en el modo espectral de primera derivada, se observa como los desplazamientos constantes de línea base han sido corregidos por el tratamiento de derivada, para esto no hay nada más que comparar la dispersión presente en la región en torno a 1100-1400 nm entre las figura 5.21 y 5.22.

Si atendemos a la figura 5.22, se observa como se hacen visibles las regiones que anteriormente habíamos adscrito a la glucosa y al etanol, (en el detalle aparece enmarcado en negro), además aparecen otras regiones que no habíamos podido diferenciar, (en el detalle aparece enmarcado en rojo), en la zonas comprendidas entre 1150 y 1250 nm y entre 1350 y 1450 nm que corresponden, respectivamente, al segundo sobretono y primer sobretono de los enlaces CH, CH_2 y CH_3 presentes en los diferentes compuestos orgánicos que van apareciendo a lo largo de la fermentación.

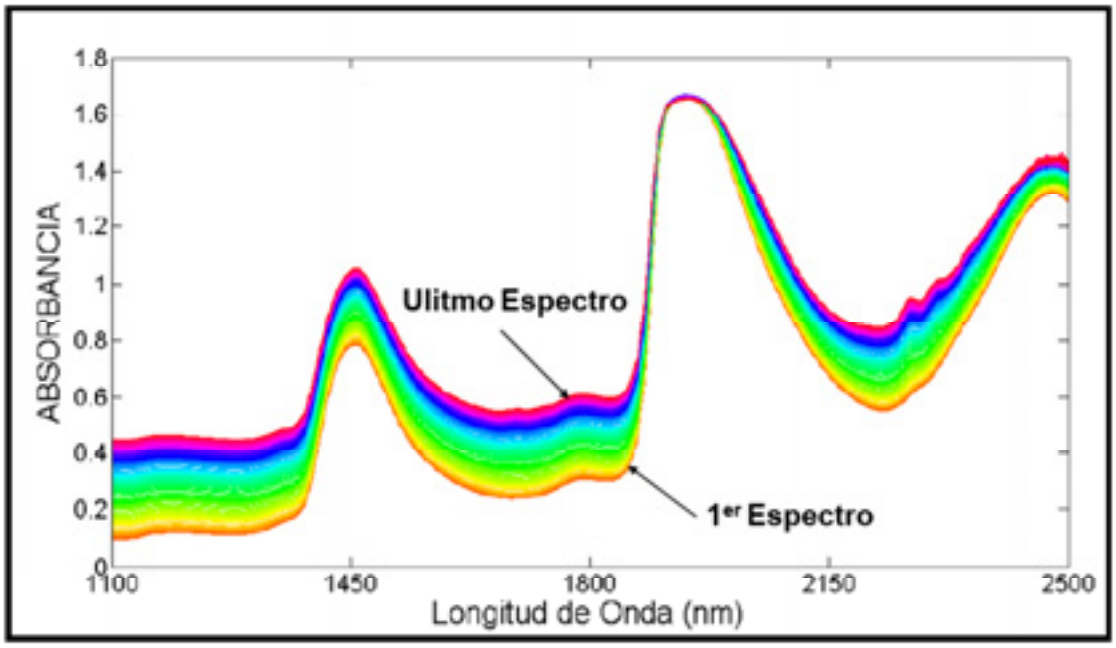


Figura 5.21. Evolución espectral en absorbancia de una fermentación llevada a cabo a 30ºC.

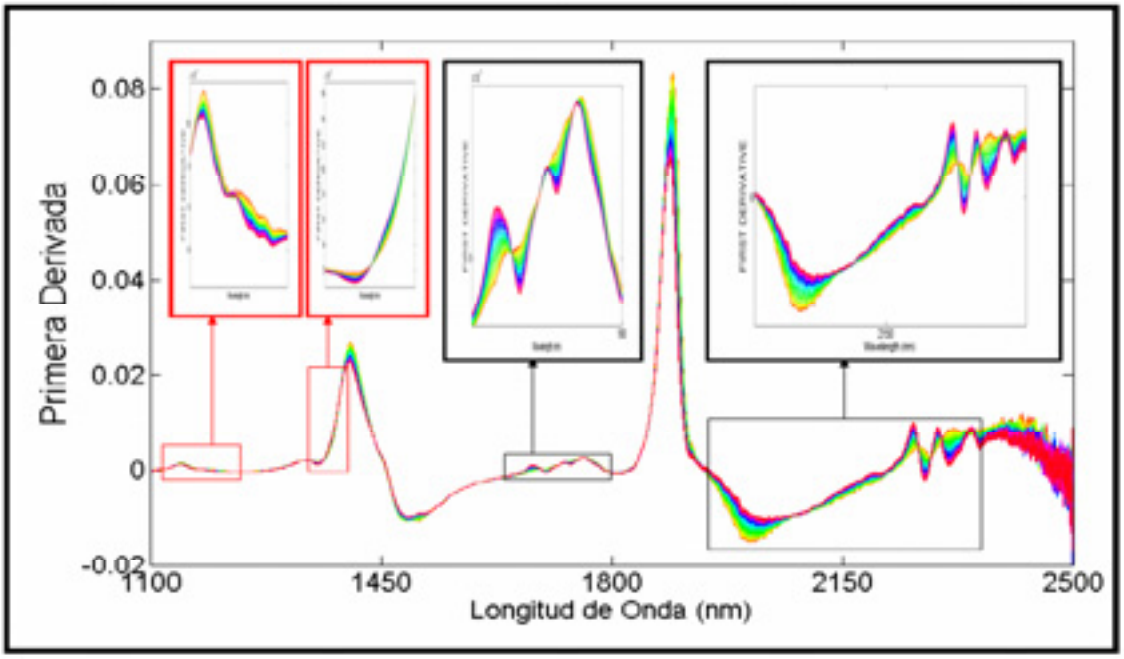


Figura 5.22. Evolución espectral en 1ra derivada de una fermentación alcohólica realizada a 30ºC. Detalle de diferentes regiones dónde se observa una ordenación temporal espectral.

5.3.1.2 Efecto de la temperatura en los espectros NIR de líquidos polares

A pesar de que el efecto de la temperatura en sistemas acuosos sobre los espectros NIR es hartamente conocido, estando bien descrito y documentado [24], no existe un modelo consensuado que describa la evolución del espectro NIR con la temperatura, incluso es un tema de discusión actualmente abierto y en boga si el agua puede ser considerado como un sistema continuo, en el cual los enlaces de hidrogeno se debilitan al incrementar la temperatura o, por el contrario, su comportamiento puede ser descrito como un sistema discreto de dos o más componentes [25].

Los espectros NIR de muestras líquidas sufren cambios muy significativos al variar la temperatura. Al aumentar ésta, las bandas de absorción se estrechan y se desplazan de una forma no lineal [26]. En la figura 5.23, 5.24, 5.25 se pueden observar los efectos descritos en los espectros de los tres lotes puros empleados, (agua, etanol y glicerina), a las temperaturas de 15, 45 y 75ºC.

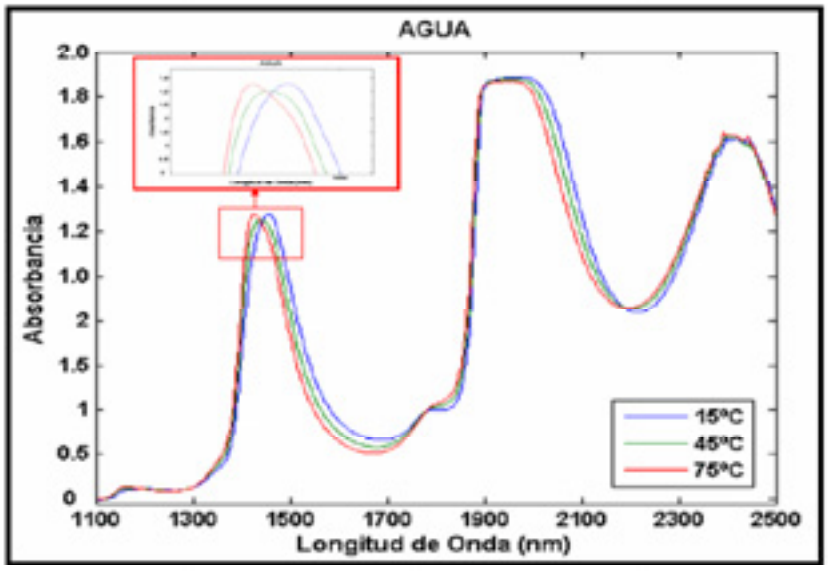


Figura 5.23. Evolución espectral del agua con la temperatura. Detalle en zona de 1400-1550nm.

[24] D. Eisemberg, W. Kauzmann, "The structure and properties of water", **1969**, Ed. Oxford University Press.

[25] Starzak M., Mathlouthi M., "Cluster composition of liquid water derived from laser-Raman spectra and molecular simulation data", Food Chemistry, **2003**, 82(1), 3–22.

[26] Osborne B., Fearn T., "Near Infrared Spectroscopy in Food Analysis", Ed. John Wiley & Sons,1993.

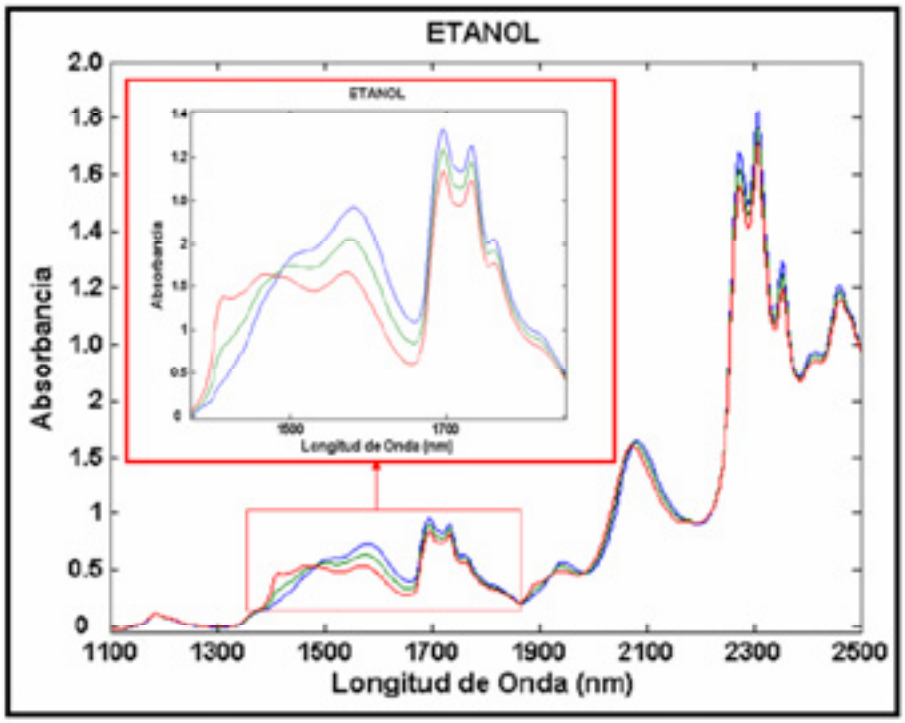


Figura 5.24. Evolución espectral del etanol con la temperatura. Detalle en la zona de 1350-1850nm.

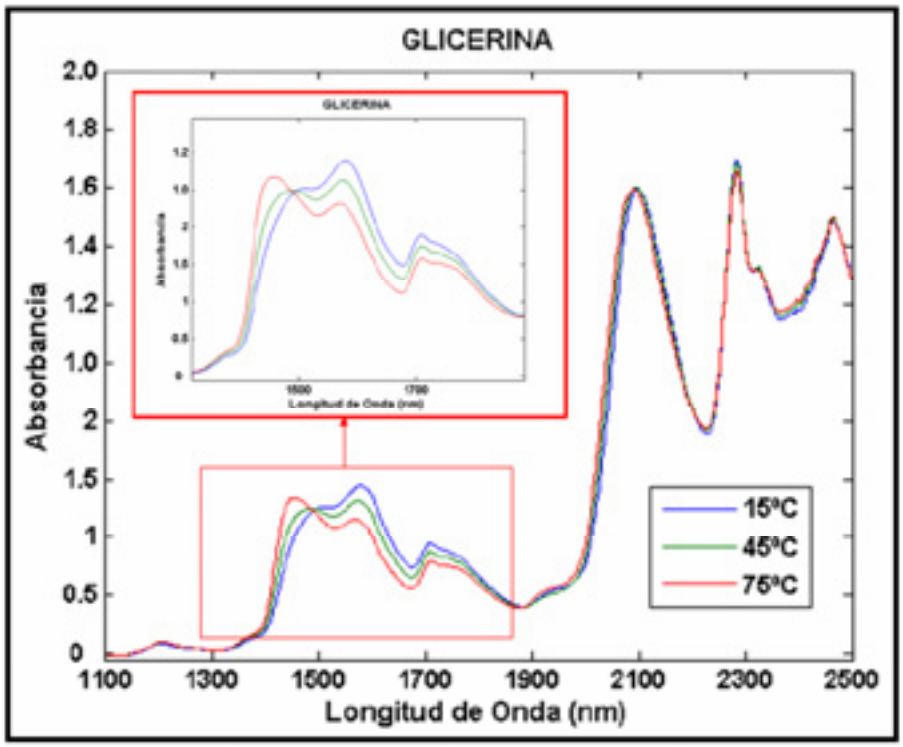


Figura 5.25. Evolución espectral de la glicerina con la temperatura. Detalle en la zona de 1300-1850 nm.

5.3.2 MODELOS PLS UTILIZADOS EN EL SEGUIMIENTO DE FERMENTACIONES ALCOHÓLICAS.

Para romper la correlación entre los diferentes analitos involucrados en el proceso de fermentación alcohólica, el conjunto de calibración estaba compuesto por:

- Muestras extraídas de diferentes procesos de fermentación (realizadas a temperaturas comprendidas entre 25 y 35ºC)
- muestras sintéticas realizadas mediante mezcla por pesada de glucosa, etanol, biomasa, glicerina y ácido acético en proporciones relativas diferentes a las presentes durante la evolución de la fermentación.

En la tabla 5.3 aparece recogido el número de muestras que fueron utilizadas, su origen y el intervalo de concentraciones empleado en la construcción de los cinco modelos creados. El número de muestras utilizado para los diferentes modelos no fue siempre el mismo y fueron seleccionadas en base a su variabilidad espectral utilizando el análisis en componentes principales.

Tabla 5.3. Características de las muestras incluidas en los conjuntos de calibración y validación externa utilizadas en la creación de modelos PLS.

CONJUNTO	TIPO DE MUESTRA	NUMERO DE MUESTRAS PARA GLUCOSA	ETANOL	ACIDO ACÉTICO	GLICERINA	BIOMASA
CALIBRACIÓN	Laboratorio	12	12	8	31	9
	Fermentación	14	14	20	6	12
	Total	26	26	28	37	21
	Rango Concentración	0-225 g/L	0-18 %	0 -50 g/L	0 - 10 g/L	0 - 14 g/L
VALIDACIÓN	Laboratorio	5	5	5	8	0
	Fermentación	18	12	12	9	11
	Total	23	17	17	17	11

Durante el proceso de creación y optimización de modelos, los datos espectrales NIR fueron sometidos a diferentes pretratamientos como la estandarización o SNV (*Standard Normal Variate*), y los tratamientos de primera y segunda derivada utilizando el algoritmo de Savitzky-Golay [27], con un tamaño de ventana de 11 puntos. De igual forma las regiones descritas en la sección anterior fueron consideradas y combinadas de diferentes maneras, al igual que el

[27] Savitzky A., y Golay M.J.E., "Smoothing and differentiation of data by simplified least squares procedures", Anal. Chem., **1964**, 36, 1627-1639.

algoritmo de selección de variables *Jackkniffing* propuesto por Martens y Martens [28] e implementado en Unscramber v9.2 [29].

En la tabla 5.4 se recogen diferentes estadísticos utilizados para la evaluación de la capacidad predictiva de los modelos creados para los cinco analitos estudiados. En todos los casos, los mejores modelos se han obtenido en el modo espectral de primera derivada, excepto para la biomasa, que se consiguió en el modo espectral de segunda derivada. Sorprende que el mejor modelo para biomasa sea en segunda derivada ya que ésta reduce los desplazamientos que el crecimiento de las levaduras produce en el espectro de absorción. Sin embargo, este resultado esta en la línea de los presentados por otros autores que han utilizado el tratamiento espectral de derivadas en los modelos para la determinación de biomasa [30-32].

Tabla 5.4. Estadísticos descriptivos de los conjuntos de calibración y validación.

Analito	Tratamineto Espectral	Rango	Numero Variables	Factores PLS	CALIBRACIÓN		VALIDACIÓN	
					R^2	RSEC	R^2	RSEP
Etanol	1ª Derivada	1650-1820 + 2240-2400	165	2	0.9997	1.60	0.9976	5.04
Glucosa	1ª Derivada	Jackknife	606	3	0.9959	4.07	0.9984	4.81
Glicerina	1ª Derivada	Jackknife	168	9	0.9972	4.87	0.9955	6.20
Acético	1ª Derivada	Jackknife	312	8	0.9937	7.85	0.9842	7.98
Biomasa	2ª Derivada	1100-1500 + 1760-2010 + 2350-2500	400	4	0.9905	4.41	0.9755	6.94

R^2: coeficiente de determinación. RSE: Error estándar relativo tanto para calibración (RSEC), como para validación (RSEP).

[28] Martens H, Martens M., "Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by Partial least squares regression (PLSR)", Food Qual. and Prefer., **2000**, 11, 5-16.

[29] http://www.camo.com, **2005**.

[30] Giavasis I., Robertson I., McNeil B., Harvey L.M., "Simultaneous and rapid monitoring of biomass and biopolymer production by Sphingomonas paucimobilis using Fourier transform-near infrared spectroscopy", Biotechnol. Lett., **2003**, 25, 975–979.

[31] Sivakesava S., Irudayaraj J., Ali D., "Simultaneous determination of multiple components in lactic acid fermentation using FT-MIR, NIR, and Ft-Raman spectroscopic techniques" Process Biochem., **2001**, 37(4), 371–378.

[32] Vaidyanathan S., Stewart W.,., Harvey L.M., McNeil B., "Influence of morphology on the near-infrared spectra of mycelial biomass and its implications in bioprocess monitoring", Biotec. Bioeng., **2003**, 82(6), 715-724.

En la figura 5.26 se recogen los gráficos de evolución conjunta del RMSEC y RMSEP, en función del número de componentes, en los modelos seleccionados para glucosa, etanol, biomasa y glicerina. Como se puede apreciar, los modelos con un menor número de componentes fueron los obtenido para etanol, glucosa y biomasa, hecho que es lógico ya que son los analitos que se encuentran en mayor concentración y que, como se ha mostrado al principio de este capítulo, presentan una señal analítica NIR característica y diferenciadora.

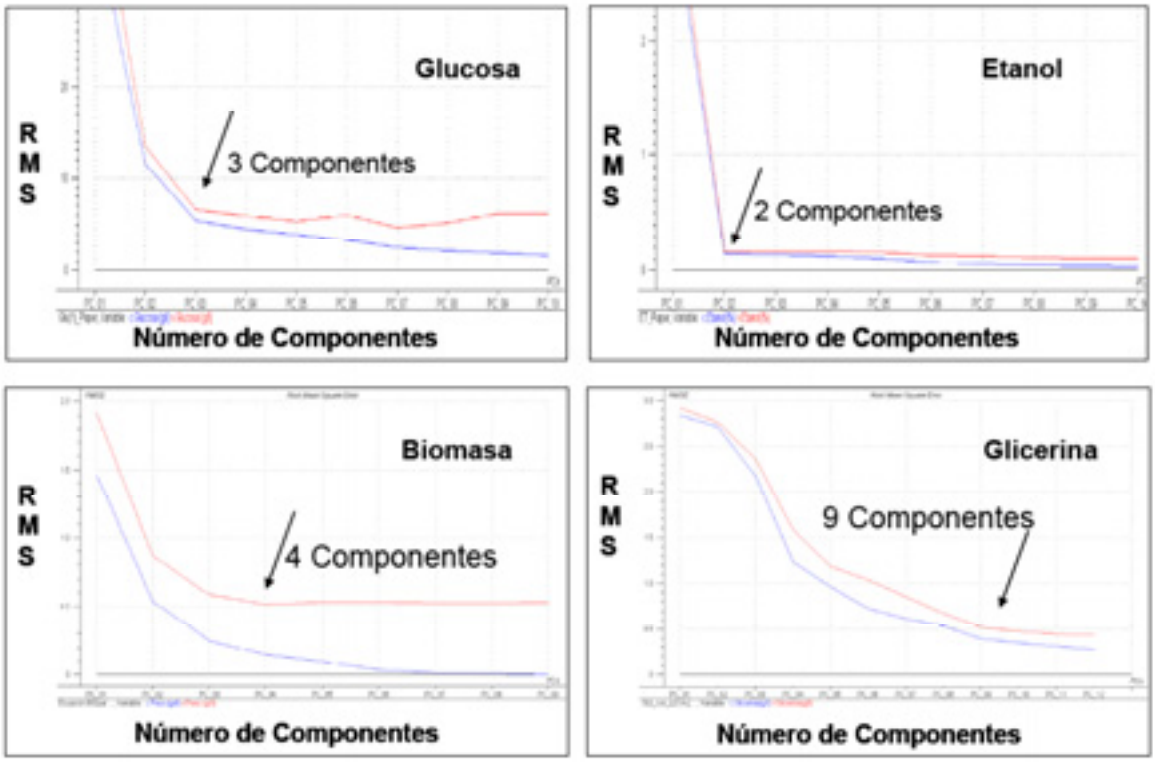


Figura 5.26. Gráfico de la evolución del RMSE frente al número de componentes para calibración (azul) y para validación (rojo).

El test de selección de variables *Jack-knife* ha simplificado los modelos, en término de número de variables, y ha proporcionado los mejores estadísticos, excepto para los modelos de etanol y de biomasa, donde los mejores resultados se han logrado a través de la selección manual de longitudes de onda. La selección de variables a través de *Jack-knife* es rápida, viene implementada en el software utilizado y comprobar si ayuda a simplificar los modelos es una tarea trivial y sencilla que puede aumentar la parsimonia y robustez de los modelos.

La figura 5.27 muestra las predicciones PLS, a lo largo del tiempo, obtenidas al aplicar los modelos seleccionados a una fermentación que fue monitorizada durante 74 horas y utilizada para validar el modelo. El gráfico muestra la disminución continua de la concentración de glucosa y el aumento paulatino y acumulación del etanol en el medio. Las concentraciones de glicerina, acido y biomasa crecen inicialmente hasta que alcanzan un valor casi constante después de un cierto tiempo que es diferente según el analito considerado. Los diferentes símbolos representan los valores PLS predichos para los distintos analitos, la línea continúa representa el ajuste de los valores predichos a un polinomio de tercer grado.

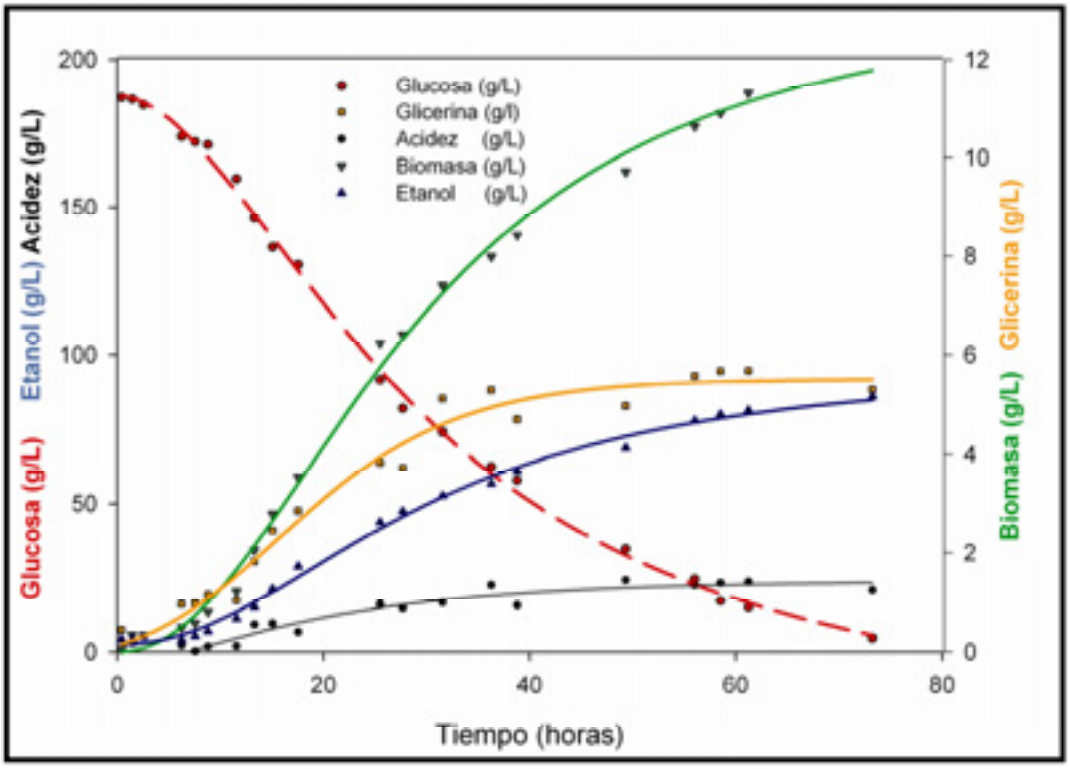


Figura 5.27. Evolución de los principales analitos presentes en una fermentación alcohólica

5.3.3 MODELOS MCR-ALS

5.3.3.1 Introducción

Una vez desarrollados y validados los modelos PLS para los diferentes analitos, se planteó el estudiar la viabilidad de aplicar métodos de resolución de curvas para el seguimiento de las fermentaciones y utilizar los modelos PLS, previamente construidos, para comprobar la bondad del ajuste.

Los métodos por resolución de curvas no requieren información analítica de referencia, aunque cualquier información que se tenga del sistema en estudio puede ser utilizada para mejorar los modelos creados [33]. Este tipo de herramientas han sido aplicadas con éxito en diferentes problemas analíticos [34], sin embargo, a pesar de su potencial, su utilización en la monitorización y control de procesos aún es muy escasa [35 - 36] y, según me consta por las búsquedas bibliográficas realizadas, no se ha divulgado ninguna aplicación de esta técnica en el seguimiento de bioprocesos donde el responsable de la transformación química sea un microorganismo.

El objetivo de este trabajo fue estudiar la capacidad del algoritmo ALS para ser aplicado al seguimiento de fermentaciones alcohólicas, para esto se utilizaron tres procesos fermentativos que partían de una concentración inicial de glucosa de 200 g/L y una temperatura constante de proceso de 25ºC:

- Uno de ellos se utilizó para construir un modelo MCR-ALS. El proceso partió de un pH inicial de 4
- Los otros dos se utilizaron para validar el modelo. El pH inicial fue de 4 para uno y 5 para el otro.

---

[33] De Juan A., Tauler, R., "Chemometrics applied to unravel multicomponent processes and mixtures. Revisiting latest trends in multivariate resolution", Anal. Chim. Acta, **2003**, 500, 195-210.

[34] Jiang J., Liang, Y., Ozaki Y. "Principles and methodologies in self-modeling curve resolution". Chemom. Intell. Lab. Syst., **2004**, 71(1), 1-12.

[35] Tauler R.; Kowalski B.; Fleming S., " Multivariate curve resolution applied to spectral data from multiple runs of an industrial process", Anal. Chem., **1993**, 65(15), 2040-2047.

[36] Garrido, M., Lázaro, I., Larrechi M. S., Rius F. X., "Multivariate resolution of rank-deficient near-infrared spectroscopy data from the reaction of curing epoxy resins using the rank augmentation strategy and multivariate curve resolution alternating least squares approach", Anal. Chim. Acta, **2004**, 515(1), 65-73.

5.3.3.2 Establecimiento del número de componentes

Una etapa fundamental y clave en la aplicación de cualquier algoritmo de resolución es el establecimiento y determinación del número de componentes que pueden ser monitorizados. Para tal fin, se estudiaron tanto diferentes tratamientos como zonas espectrales. Aunque en bibliografía aparecen citados diferentes estadísticos experimentales propuestos para determinar el número de componentes [37], en nuestro caso, la representación gráfica de los vectores propios, asociados a los principales valores propios, obtenidos a través de una descomposición en valores singulares, fue una herramienta útil y dilucidadora que ayudó a la selección del número de componentes.

En la figura 5.28 se muestra, a modo de ejemplo, el gráfico de los cinco principales vectores propios o *eigenvectores* en función del tiempo. Éstos han sido obtenidos utilizando los espectros registrados cada 30 minutos en el transcurso de un proceso fermentativo de 50 horas de duración. Los vectores propios asociados a los componentes 1, 2 y 3 muestran una tendencia y se asemejan a la evolución de concentración de los analitos glucosa, etanol y biomasa respectivamente, sin embargo, los vectores asociados a los componentes 4 y 5 no siguen ninguna tendencia definida haciendo imposible toda interpretación y asociación química, por lo que podemos concluir que el rango químico del sistema en estudio es tres.

[37] E. R. Malinowski, "Statistical F-tests for abstract factor analysis and tartget testing" J. Chemometrics, **1988**, 3, 49-60.

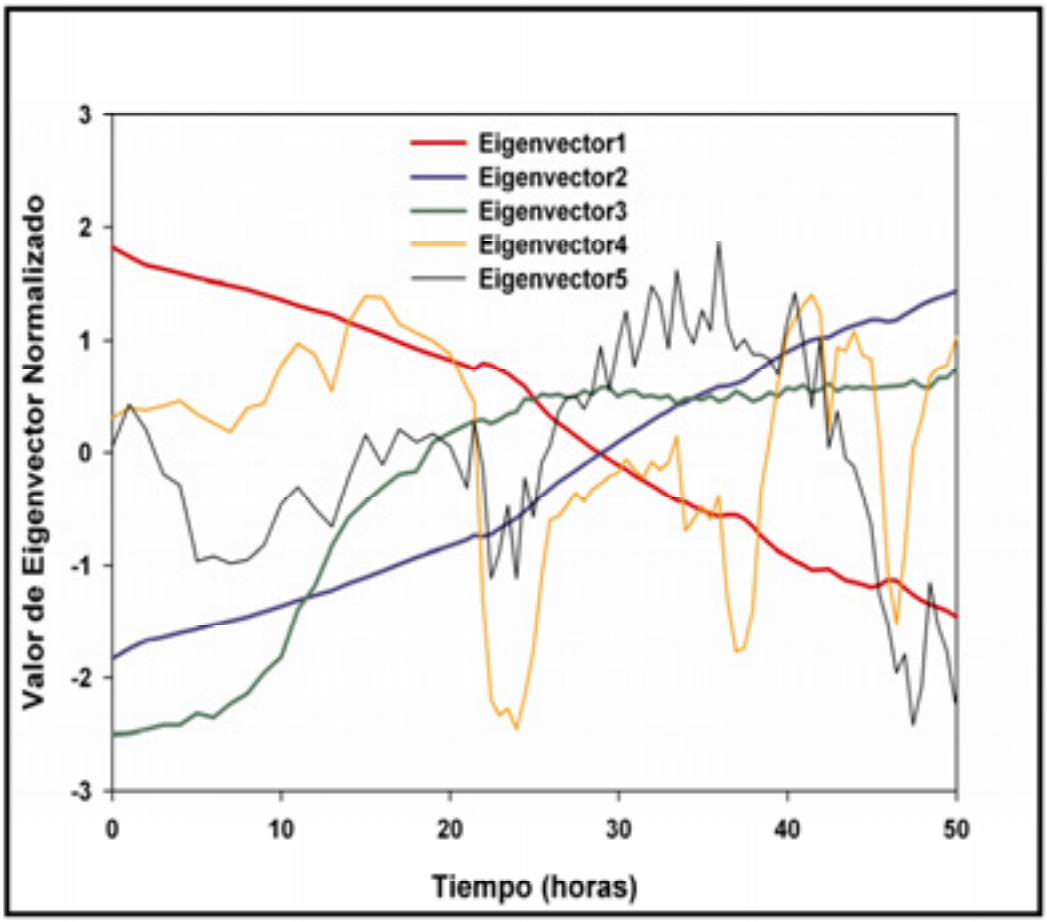


Figura 5.28. Vectores propios normalizados asociados a los cinco componentes con mayor varianza.

5.3.3.3 Restricciones aplicadas

Las restricciones aplicadas al sistema en el dominio de las concentraciones fueron no negatividad y unimodalidad y, de no negatividad, en el dominio espectral. Si estas restricciones no son impuestas, el algoritmo puede no converger hacia una solución que se corresponda con la evolución química de la fermentación. Además, también se ensayó la opción de introducir diferentes combinaciones de restricciones de igualdad. Las opciones consideradas fueron:

- No suministrar información espectral.
- Asignar al componente glucosa el primer espectro registrado, que corresponde al medio de cultivo justo antes de inocular la levadura.

- Asignar al componente glucosa el perfil de concentraciones obtenidos a través del algoritmo EFA
- La unión de las dos opciones anteriores.

Los resultados fueron expresados en términos de Falta de Ajuste, **LOF** (*Lack of Fit*) **[5.1]**, donde a_{ij} es el valor de la absorbancia registrada para la muestra *i* a la longitud de onda *j* y $\hat{a}_{ij}$ es el valor calculado por el modelo. Cuanto mayor sea el **LOF** peor será la capacidad descriptiva del modelo. La bondad de los espectros recogidos fue expresada a través del coeficiente de disimilaridad, **sin_z [5.2]**, donde s_i es el valor del espectro de referencia, bien de glucosa (sin_z_{glu}) o de etanol (sin_z_{et}) y $\hat{s}_i$ es el valor estimado por el modelo. Cuanto más bajo sea el **sin_z** mas semejantes serán la información espectral recogida por el modelo y los espectros de refencia.

$$LOF=\sqrt{\frac{\sum_i\sum_j (a_{ij}-\hat{a}_{ij})^2}{\sum_i\sum_j a_{ij}^2}}\times 100 \qquad [5.1]$$

$$sin\_z=\sqrt{1-\cos^2\frac{s_i^T\hat{s}_i}{\|s_i\|\|\hat{s}_i\|}} \qquad [5.2]$$

En la tabla 5.5, se recogen los estadísticos explicados en las cuatro situaciones estudiadas. Como se puede observar, los estadísticos fueron significativamente peores cuando se introdujeron restricción de igualdad en el dominio de las concentraciones (cifras sin formato de letra, en la tabla 5.5). Esto fue debido a que el algoritmo EFA no proporciona estimaciones buenas en los procesos en los que hay analitos que se acumulan en el medio, a pesar de ser un método muy preciso para determinar el momento en el que la contribución de los diferentes analitos es significativa o deja de serlo. Para las otros dos opciones (aparecen en negrita y cursiva en la tabla 5.5), las diferencias entre los estadísticos no fueron significativas, entre el caso en el que no se introdujo ningún tipo de restricción espectral y el caso en el que se introdujo como restricción el espectro del etanol. Por lo tanto, se optó por no introducir ningún tipo de

restricción de igualdad, para evitar cargar al algoritmo con condicionantes que no mejoran su capacidad descriptiva.

Tabla 5.5. Estadísticos obtenidos al aplicar diferentes combinaciones de restricciones de igualdad al algoritmo ALS.

CONCENTRACIONES	ESPECTROS					
	NINGUNA			PRIMER ESPECTRO		
	LOF	sinz_glu	sinz_et	LOF	sinz_glu	sinz_et
NINGUNA	0.0645	0.0014	0.0197	0.0646	0.0009	0.0201
GLUCOSA	***0.1688***	***0.1467***	***0.1800***	***0.1896***	***0.1654***	***0.2120***

5.3.3.4 Predicciones MCR-ALS

En la figura 5.29 se recoge los perfiles cinéticos para glucosa, etanol y biomasa para la fermentación realizada a partir de una concentración de glucosa inicial igual a 200 g/L y un pH inicial de 4. Las concentraciones se muestran a escala real del proceso, ya que a los perfiles de concentración MCR-ALS se le han aplicado los respectivos coeficientes de regresión lineal (ordenada en el origen y pendiente) entre los valores MCR-ALS y los valores obtenidos a través del modelo de referencia PLS.

Los valores del coeficientes de determinación R^2 entre los valores MCR-ALS y los valores PLS para glucosa, etanol y biomasa fueron respectivamente de 99.73%, 99.55% y 99.15%. Dichas relaciones son aceptables y nos indica que, a pesar de que el modelo MCR-ALS está sometido a las ambigüedades inherentes de rotación y de intensidad, su influencia no impide obtener buenos estadísticos.

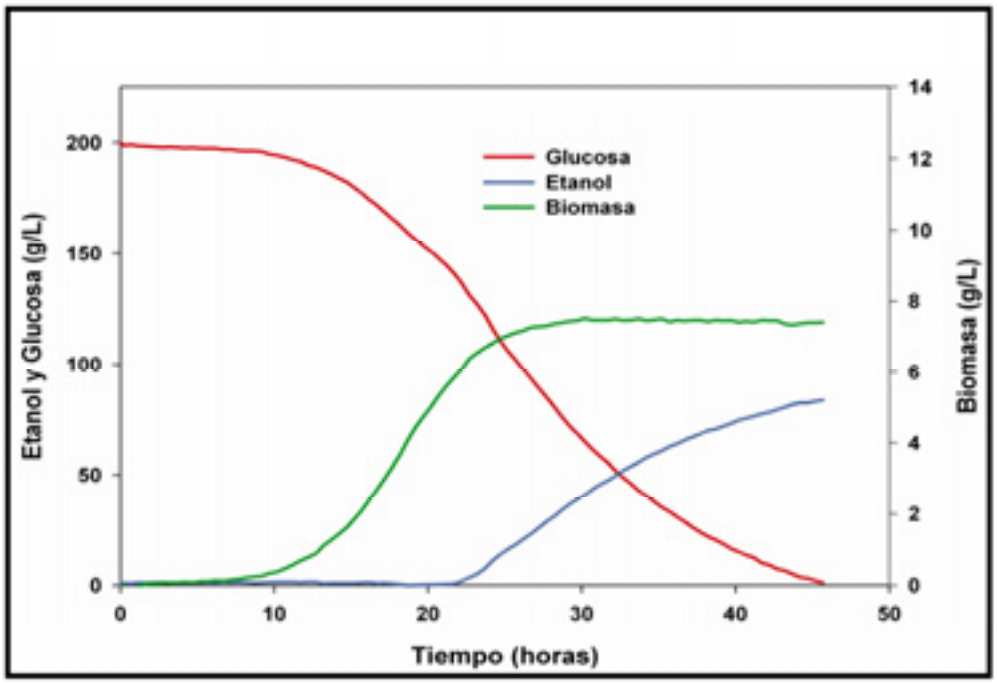


Figura 5.29. Evolución de la glucosa, etanol y biomasa proporcionados por el modelo MCR-ALS.

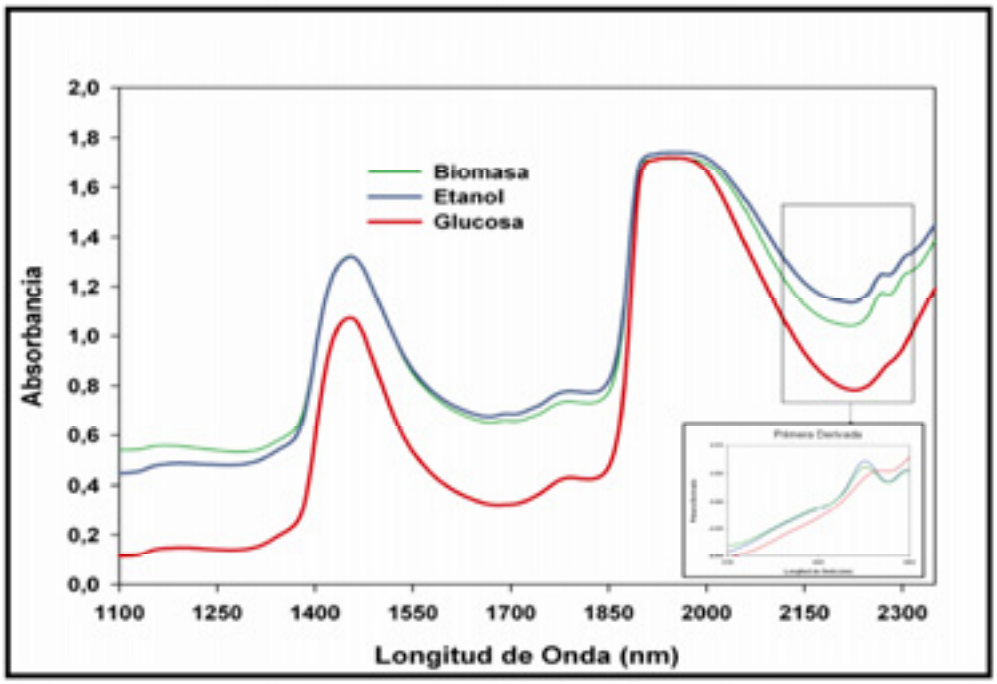


Figura 5.30. Perfiles espectrales obtenidos por el modelo MCR-ALS asociados a la glucosa, etanol y biomasa. En la esquina inferior derecha se muestra detalle en primera derivada.

En la figura 5.30 se recogen los espectros asociados a los perfiles de concentración de los analitos representados en la figura 5.29. En el margen inferior derecho se recoge el detalle de la primera derivada en la zona comprendida entre 2000 y 2300 nm. Tal y como se observa, el espectro asociado a la glucosa es muy parecido al espectro inicial recogido al principio de la fermentación, momento en el que la concentración de glucosa es máxima (comparar con figura 5.21).

El espectro asociado al etanol se diferencia claramente del espectro de glucosa en la zona comprendida en torno a 2100-2300 nm. En esta zona esta comprendida la región que anteriormente asociamos al contenido en etanol.

El espectro asociado a la biomasa, que es la responsable principal de la dispersión espectral producida en el medio de cultivo, presenta un valor de absorbancia mayor que el de glucosa y etanol, en torno a la zona de 1100-1400 nm, que es la zona donde se observa fácilmente el aumento de la dispersión a lo largo de la fermentación, el resto del espectro muestra una gran similitud a los espectros de glucosa y etanol. No hay que olvidar que la fermentación transcurre en medio acuoso y la elevada absortividad de las bandas del agua define y moldea los espectros de los tres analitos.

Tabla 5.6. Estadísticos obtenidos para la relación entre los valores de concentración proporcionados por el modelo MCR-ALS y los obtenidos al aplicar directamente el algoritmo ALS a dos fermentaciones realizadas a pH 3 y pH 5

ANALITO	pH	PENDIENTE	INTERCEPTO	R^2
Glucosa	pH3	0.998 ± 0.001	-0.095 ± 0.001	0.999
	pH5	0.998 ± 0.001	0.200 ± 0.001	0.999
Etanol	pH3	1.000 ± 0.005	-0.095 ± 0.006	0.998
	pH5	0.949 ± 0.046	0.114 ± 0.033	0.994
Biomasa	pH3	0.959 ± 0.075	0.130 ± 0.048	0.959
	pH5	0.949 ± 0.054	0.114 ± 0.031	0.962

El símbolo± indica el intervalo de confianza a una significación del 95%

El modelo MCR-ALS obtenido se aplicó a dos fermentaciones realizadas en condiciones similares, pero con un pH inicial diferente 3 y 5. La aplicación se realizó utilizando la pseudo-inversa de la matriz de datos espectrales del modelo.

En la figura 5.31 se muestran los perfiles de concentración de las dos fermentaciones predichos por el modelo MCR-ALS. Como se puede observar las restricciones de no negatividad y unimodalidad impuestas al modelo inicial no se cumplen estrictamente en las predicciones realizadas.

En la tabla 5.6 se muestran, para los tres analitos estudiados, la ordenada en el origen, la pendiente y el coeficiente de determinación, entre los valores de concentración predichos por el modelo MCR-ALS y los valores obtenidos al aplicar MCR-ALS directamente a las dos fermentaciones utilizadas en la validación. En este caso, las restricciones empleadas fueron las mismas que las utilizadas a la hora de construir el modelo. A pesar de las incertidumbres que presenta el modelo debido a la existencia de ambigüedades, las correlaciones son altamente significativas para los analitos glucosa y etanol y significativas para la biomasa.

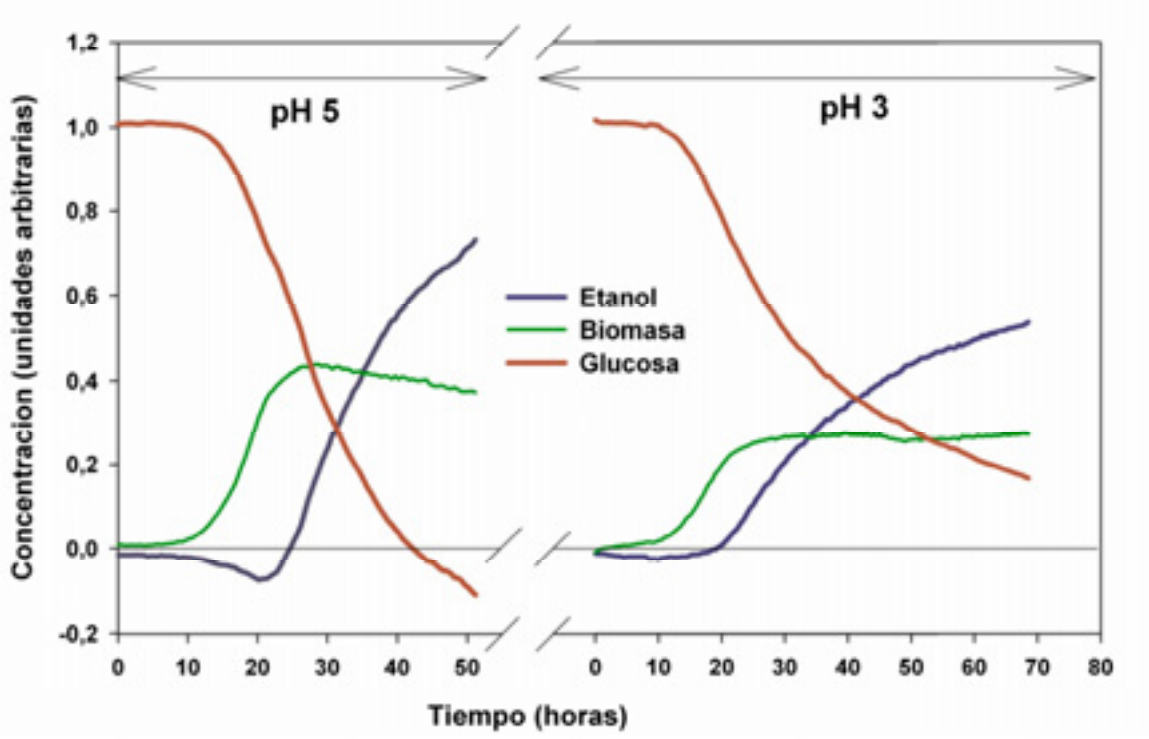


Figura 5.31. Perfiles de concentración obtenidos tras aplicar el modelo MCR-ALS a dos fermentaciones realizadas a pH 3 y pH 5.

5.3.4 USO CONJUNTO DE MCR-ALS Y MODELOS EMPÍRICOS

5.3.4.1 Introducción

Las ambigüedades inherentes a los métodos de resolución, que provoca que el espacio de soluciones factibles sea indeterminado, es uno de los principales inconvenientes de este tipo de métodos. Recientemente, se han desarrollado algoritmos que proporcionan una estimación de los límites máximo y mínimo para cada perfil [38 - 39]. Sin embargo, la única manera de restringir y acotar el rango de soluciones posibles es a través de la introducción de información externa relacionada con el sistema en estudio, surgiendo así los denominados modelos grises (*grey models o semi-soft modelling*), siendo ésta una nomenclatura descriptiva que indica que son métodos basados en métodos de resolución, con baja demanda de información de referencia (*withe modelling o soft modelling*), pero que han sido complementados con información externa.

Tras la obtención de los resultados presentados en la sección anterior, donde los resultados para la biomasa no fueron suficientemente satisfactorios, se decidió conjugar los métodos de resolución MCR-ALS junto con la información procedente de modelos fermentativos empíricos, utilizados tradicionalmente en el diseño y seguimiento de la evolución de diferentes analitos en bioreactores.

El objetivo de este trabajo fue estudiar si la utilización del algoritmo ALS, complementado con la información suministrada en forma de restricciones de igualdad, procedente de modelos empíricos de modelado, puede mejorar los modelos de resolución anteriormente construidos. El método utilizado en este caso fue el de mínimos cuadrados alternados con funciones de penalización (*penalty Alternating Least Squares*), p-ALS.

---

[38] Gemperline P.J., "Computation of the range of feasible solutions in Self-Modeling Curve Resolulton Algorithms", Anal. Chem., **1999**, 71, 5398-5404.

[39] Garrido M., Larrechi M.S., Rius F.X., Tauler R., "Calculation of band boundaries of feasible solutions obtained by MCR-ALS of multiple runs of a reaction monitored by NIRS", Chem. Intell. Lab. Syst., **2005**, 76, 111-120.

Para este estudio se realizaron nueve procesos fermentativos:

- Tres fueron realizados a pH 4 y tres niveles de concentración inicial de glucosa (200, 100, 50 g/L). Estos procesos fueron utilizados para seleccionar el modelo empírico que mejor describía las condiciones de nuestro sistema y para establecer las restricciones de igualdad.
- Los otros seis procesos partieron de una concentración de glucosa inicial de 200 g/L, pero fueron realizadas a tres niveles de pH inicial, pHs 3, 4 y 5. Cada nivel de pH fue realizado a dos temperaturas diferentes (25 y 35ºC).

5.3.4.2 Selección del modelo empírico

Para seleccionar el modelo empírico, de los recogidos en la tabla 4.3, con mayor capacidad descriptiva para las condiciones de trabajo de nuestro sistema fermentativo, se utilizaron dos criterios cualitativos:

- El primero es la consistencia de los valores obtenidos, para los diferentes parámetros, cuando la concentración inicial de sustrato en el medio es diferente.
- El segundo es el ajuste obtenido en la transición, entre la fase exponencial y la fase estacionaria, donde hay una gran variación en la tasa de crecimiento en un corto intervalo de tiempo.

En la figura 5.32 se puede observar el ajuste para glucosa, etanol y biomasa de los valores de referencia, proporcionados por los modelos PLS, al modelo de inhibición por producto propuesto por Hinshelwood. Como se puede observar, el ajuste es altamente aceptable, obteniéndose buenas correlaciones para todos los analitos y, lo que es aún más importante, los valores de los parámetros obtenidos en el ajuste del modelo a las diferentes fermentaciones son de la misma magnitud, además, los valores de los parámetros no son incongruentes ya sea por presentar una escala desmesurada o bien por ser negativos. Sin embargo, esto no ocurrió con el resto de los modelos empíricos ensayados.

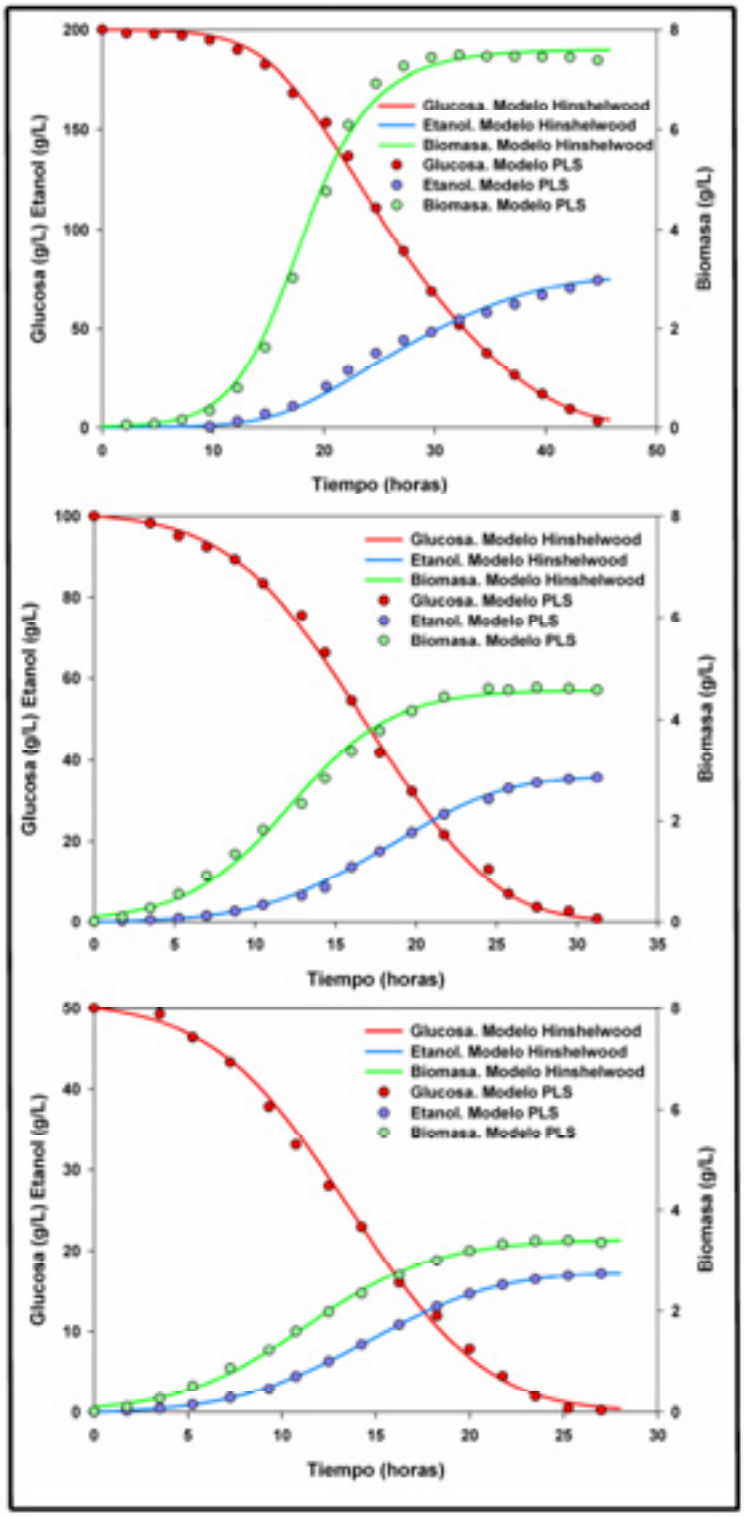


Figura 5.32. Ajuste del modelo empírico de Hinshelwood (línea continua) a los valores de referencia PLS obtenidos para tres procesos fermentativos con concentración de glucosa inicial de 200 g/L (superior), 100 (g/L) (intermedia) y 50 g/L (inferior).

Tabla 5.7. Valores de los parámetros del modelo de Hinshelwood para tres fermentaciones realizadas a pH inicial 4 y Temperatura 30°C y concentración inicial de glucosa de 200 g/L, 100 g/L y 50 g/L.

PARAMETROS DEL MODELO	CONCENTRACIÓN INICIAL DE GLUCOSA (g/L)		
	200	100	50
μ	0.153	0.132	0.137
K_{sx}	36.121	17.748	10.792
K_{px}	0.070	0.098	0.128
ν	0.310	0.297	0.354
K_{sp}	22.851	26.211	16.514
K_{pp}	0.010	0.010	0.011
Y_{xs}	0.544	0.215	0.244
Y_{pp}	0.411	0.453	0.476

En la tabla 5.7 aparecen recogidos los valores de los parámetros obtenidos para el modelo de Hinshelwood cuando fue ajustado a los valores de referencia obtenidos a través de los modelos PLS para glucosa, etanol y biomasa. Los resultados son consistentes, entre las diferentes fermentaciones, si tenemos en cuenta que se tratan de modelos empíricos aplicados a fermentaciones que parten de concentraciones de glucosa diferentes y se encuentran abalados por resultados similares obtenidos por otros autores en trabajos análogos [40].

5.3.4.3 Obtención de las restricciones espectrales de igualdad

Los perfiles espectrales correspondientes a la fermentación realizada a 200 g/L fueron calculados a partir de la matriz **A** de datos experimentales registrados y de la pseudo-inversa de los correspondientes perfiles de concentración, $(C)^+$, a través de la expresión **[5.3]**. Estos perfiles espectrales fueron el elemento clave de unión, entre el modelado empírico y el método de resolución,

[40] Godia F., Casas, C., Sola C.J., "Batch alcoholic fermentation modelling by simultaneous integration of growth and fermentation equations", J. Chem. Tech. Biotech., **1988**, 41(2),155-165.

ya que fueron obtenidos a través de los resultados obtenidos aplicando un modelo empírico y, posteriormente, utilizados como restricciones de igualdad dentro del algoritmo p-ALS.

$$S^T = (C)^+ A \qquad [5.3]$$

La utilización de los perfiles espectrales, en lugar de los perfiles de concentración, como restricciones de igualdad no es caprichosa y responde al hecho de que, al tener las fermentaciones duraciones diferentes, los perfiles de concentración se ven afectados por la dimensionalidad temporal, sin embargo, los perfiles espectrales son funciones independientes del tiempo y puede ser utilizados como restricciones, independientemente de cual sea la duración de la fermentación.

En la figura 5.33 se recogen los perfiles calculados, tal y como se puede apreciar, se asemejan a los recogidos en la figura 5.30, procedentes de la aplicación directa del modelo MCR-ALS. Nuevamente se observa como el perfil adscrito a la glucosa es el que presenta una menor intensidad de la señal, el perfil del etanol presenta una serie de máximos y mínimos en la zona alrededor de 2000 nm. El perfil adscrito a la biomasa es el que presenta una mayor intensidad de la señal en la zona donde más patente se hace la dispersión, 1100-1400 nm.

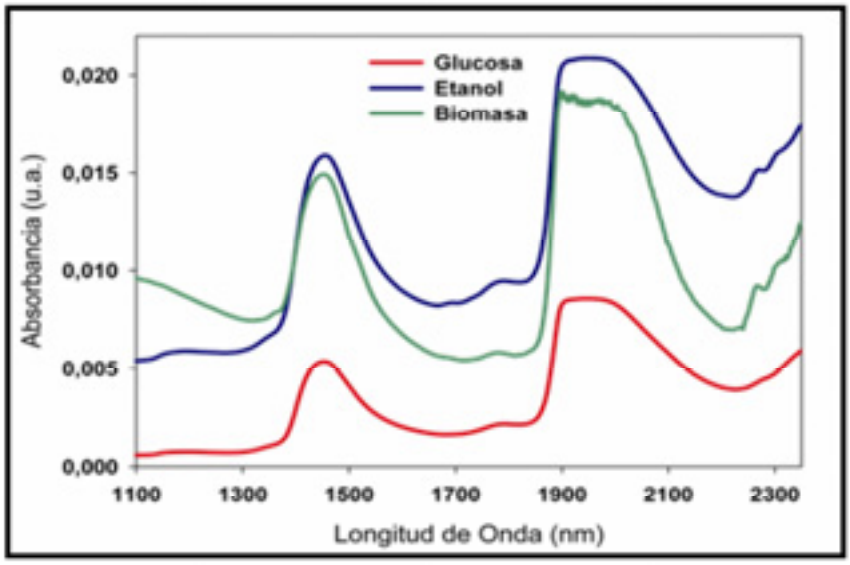


Figura 5.33. Perfiles espectrales para proporcionadas por el modelo de Hinshelwood.

5.3.4.4 Construcción del Modelo p-ALS

A la hora de imponer las restricciones de igualdad, uno de los interrogantes que se planteó fue si era necesario aplicar las restricciones a los tres analitos o, por el contrario, con restringir una o dos especies sería suficiente Con los tres perfiles espectrales obtenidos, existían siete modos diferentes de combinar la información como restricciones de igualdad:

- Una posible combinación es introducir los tres espectros simultáneamente.
- Tres combinaciones resultan de combinar los espectros por parejas.
- Tres combinaciones se obtienen al introducir cada uno de los espectros de forma independiente.

Los resultados obtenidos, al aplicar las restricciones de manera selectiva a uno o dos analitos, proporcionaron perfiles espectrales en una escala relativa del orden de 10^2 - 10^3 veces mayor para las especies restringidas que para las no restringidas. Es decir, las restricciones de igualdad se comportaban como guías conductoras que ponderaban el perfil espectral de las especies restringidas frente al de las especies no sometidas a restricción de igualdad. Para evitar estas desigualdades de escala, las restricciones de igualdad se aplicaron simultáneamente a los tres analitos.

5.3.4.5 Validación del Modelo p-ALS

En la figura 5.34 se muestran los perfiles de concentración obtenidos al aplicar el modelo p-ALS a los seis procesos fermentativos, llevados a cabo en condiciones diferentes de pH y temperatura. En la tabla 5.8 se muestran los resultados obtenidos en términos de coeficiente de determinación entre los valores proporcionados por el modelo p-ALS y los suministrados por el método de referencia PLS para los tres analitos considerados. Como se puede observar la bondad de los resultados obtenidos para los tres analitos, inclusive para biomasa, muestra la robustez ganada con la incorporación de restricciones de igualdad en el

algoritmo. La varianza explicada fue en todos los casos mayor al 99.9% y el LOF menor a 0.1%.

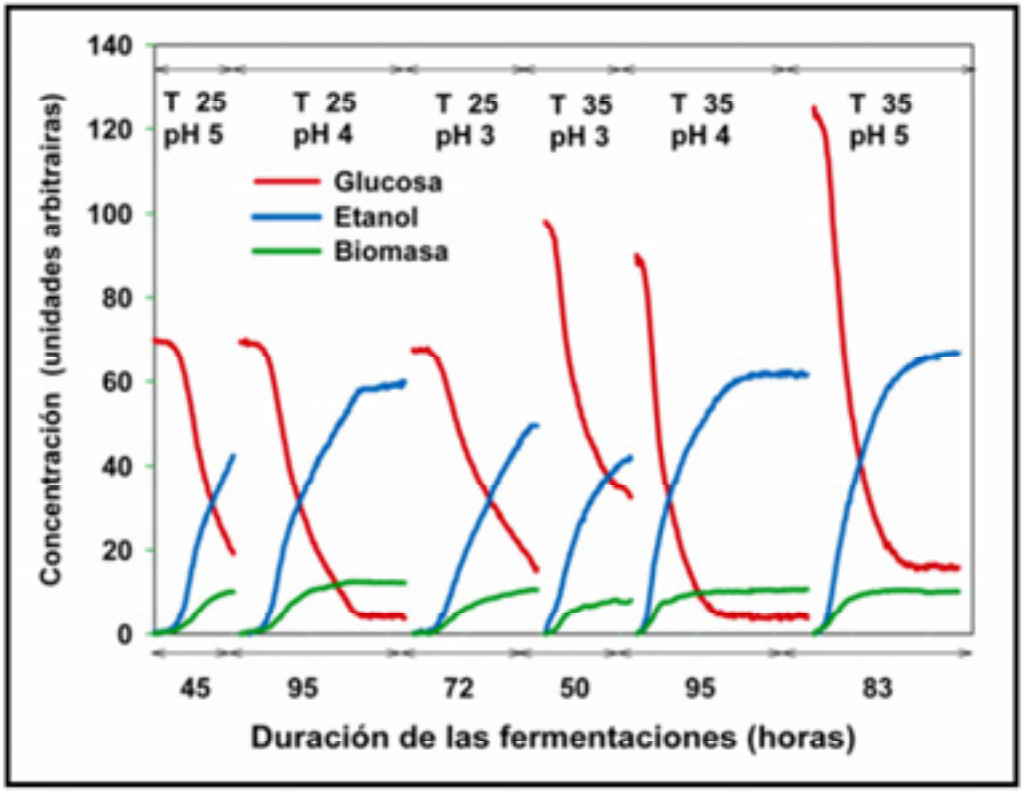


Figura 5.34. Perfiles de concentración obtenidos aplicando el modelo p-ALS con restricciones de igualdad a seis procesos fermentativos.

Tabla 5.8. Coeficiente de determinación entre los valores de referencia PLS y los obtenidos al aplicar el modelo p-ALS, para glucosa, etanol y biomasa.

CONDICIONES		COEFICIENTE DE DETERMINACIÓN		
Temperatura	pH	Glucosa	Etanol	Biomasa
25	5	0.9999	0.9983	0.9937
25	4	0.9945	0.9943	0.9808
25	3	0.9960	0.9949	0.9853
35	3	0.9970	0.9992	0.9764
35	4	0.9930	0.9984	0.9705
35	5	0.9994	0.9997	0.9882

En las figuras 5.35, 5.36, 5.37 se muestran, respectivamente, los perfiles de concentración para glucosa, etanol y biomasa proporcionados por el modelo. En el caso de la glucosa, figura 5.35, los perfiles fueron escalados entre 1 y 0 dividiendo

cada perfil por el valor inicial, puesto que la concentración inicial de todas las fermentaciones era la misma, 200 g/L.

Como se puede observar de un análisis conjunto de las tres figuras, existe un marcado efecto de la temperatura. En las tres figuras se observa una evolución diferente entre las fermentaciones realizadas a 25ºC (líneas con símbolos) y las transcurridas a 35ºC (líneas sin símbolos). De esta manera y, atendiendo al consumo de glucosa, cuando la temperatura es de 35ºC, se observa como ésta empieza a ser consumida desde los primeros estadios de la fermentación, sin embargo a 25ºC, hay un tiempo de latencia de casi 10 horas durante el cual el consumo de glucosa apenas evoluciona.

Si se observa la figura 5.37, donde se representa la evolución de la biomasa, se observa que en los primeros estadios las pendientes que presentan las fermentaciones realizadas a 35ºC son mucho mayores que las que presentan las fermentaciones realizadas a 25ºC, donde las pendientes en las primeras horas son prácticamente horizontales. Este mismo comportamiento se puede observar, de manera aún más marcada y acentuada, en la figura 5.36, donde se representa la evolución temporal del etanol, en las fermentaciones realizadas a 25ºC existe un tiempo refractario de más de 10 horas hasta que empieza a producirse etanol, mientras que a 35ºC existe un producción desde los primeros estadios de la fermentación.

De igual forma, las fermentaciones también muestran una evolución diferente en función del pH inicial del medio, aunque este efecto no es tan marcado como el de la temperatura. Así, observando la figura 5.35, se puede apreciar como la evolución del consumo de glucosa en las fermentaciones realizadas a temperatura 25ºC y pH 4 y 5 es muy parecida, de hecho aparecen solapadas, sin embargo, la fermentación realizada a 25ºC y pH 3 se diferencia claramente de las anteriores, presentando una pendiente y una evolución mucho más suave. Este mismo hecho, se puede observar a temperatura 35ºC, las fermentaciones realizadas a pH 4 y 5 evolucionan de forma paralela y la fermentación realizada a pH 3 se diferencia claramente de ellas y presenta una evolución menos drástica. Dicha fermentación fue interrumpida a las 50 horas de su inicio pero, si se hubiera

dejado más tiempo, posiblemente la fermentación se hubiera estancado. Hay que tener en cuenta que un pH de 3 es un pH extremo y cercano a la letalidad para el desarrollo de *Saccharomyces*. El efecto descrito también puede ser claramente observado si estudiamos las figuras de la producción del etanol y de biomasa. De nuevo podemos observar como, para una misma temperatura, las fermentaciones realizadas a pH 3 presentan una tasa de producción de etanol menor que las fermentaciones realizadas a pH 4 y 5.

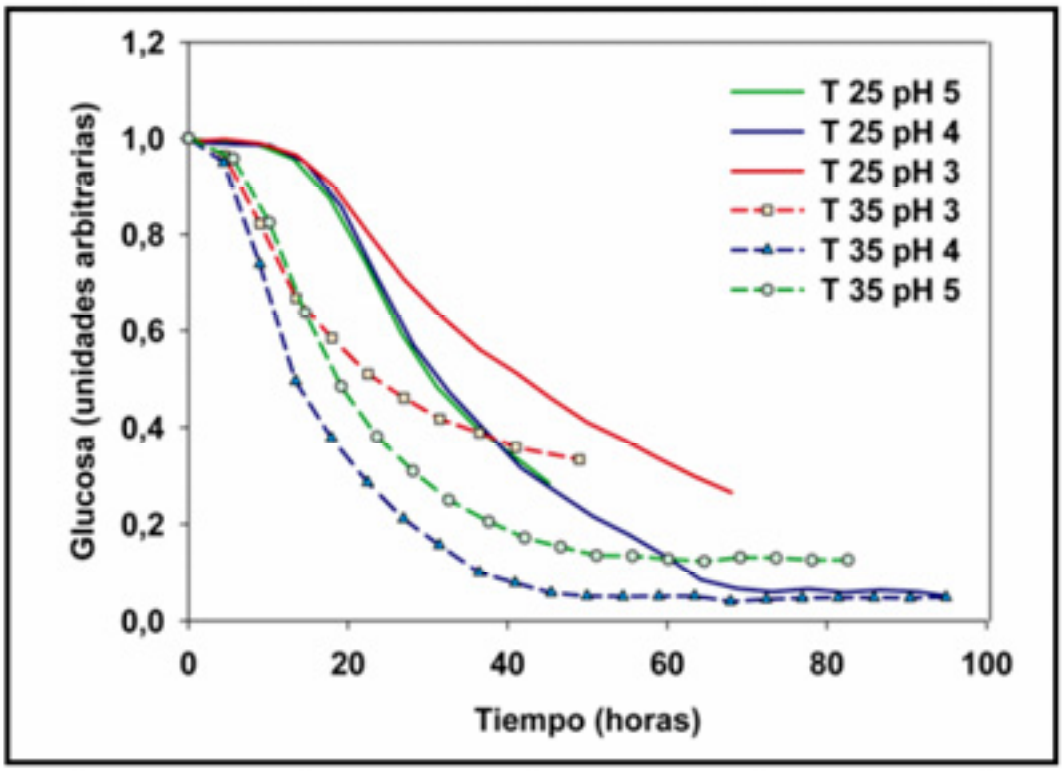


Figura 5.35. Perfiles de concentración de glucosa obtenidos al aplicar el modelo p-ALS a seis fermentaciones llevadas a cabo a diferentes condiciones de temperatura y pH.

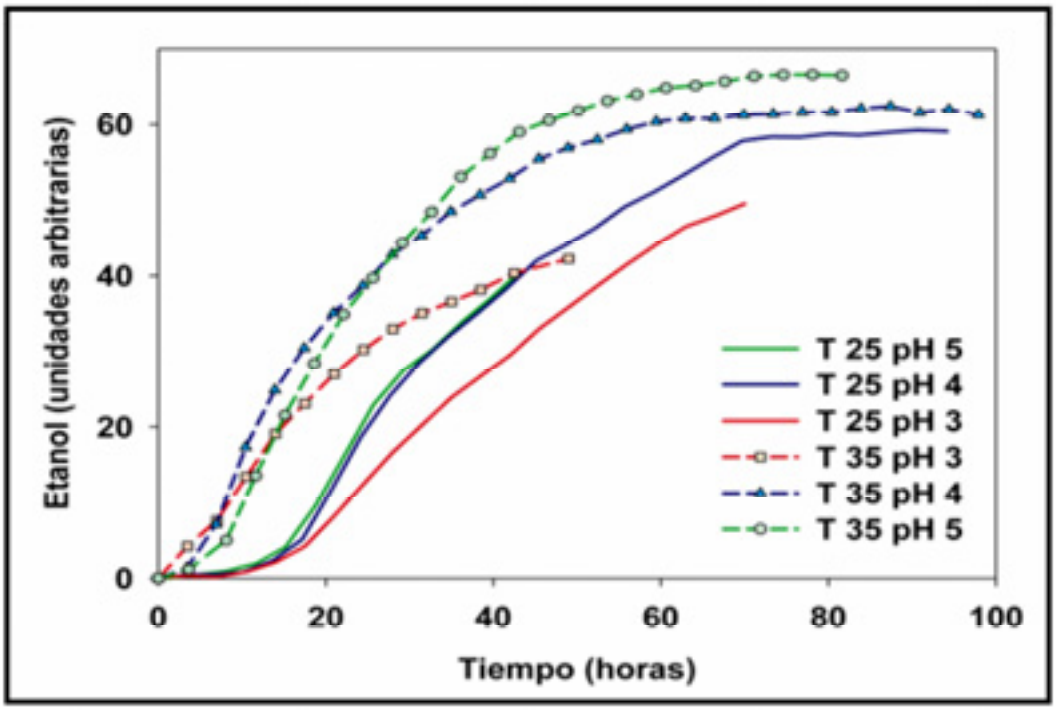


Figura 5.36. Perfiles de concentración de etanol obtenidos al aplicar el modelo p-ALS a seis fermentaciones llevadas a cabo a diferentes condiciones de temperatura y pH.

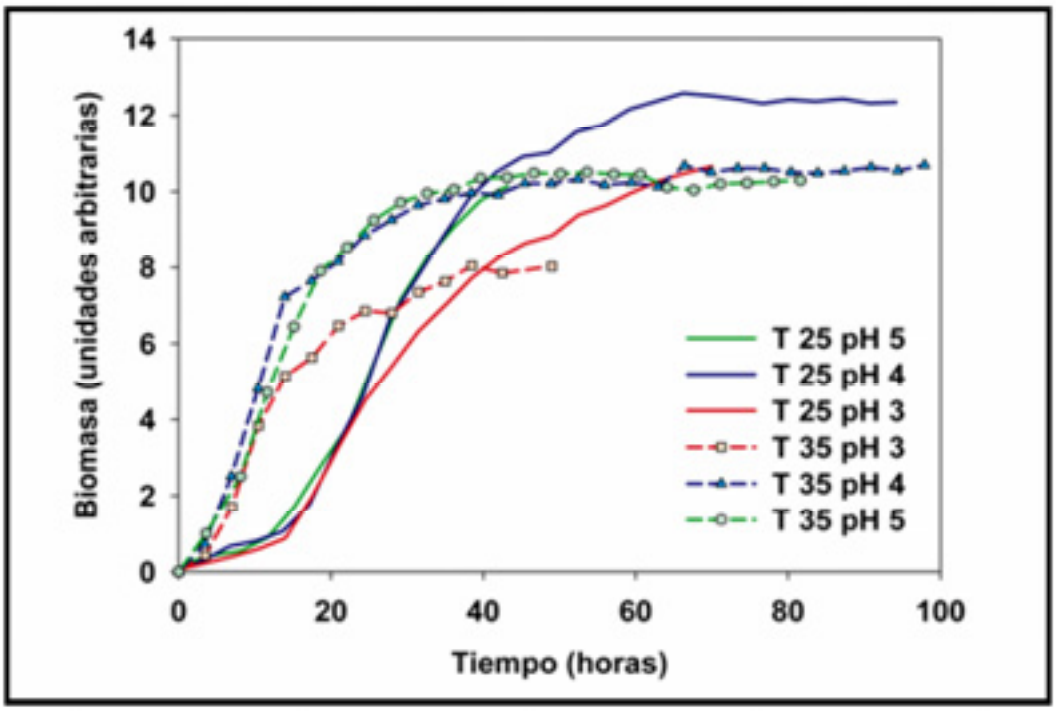


Figura 5.37. Perfiles de concentración de biomasa obtenidos al aplicar el modelo p-ALS a seis fermentaciones llevadas a cabo a diferentes condiciones de temperatura y pH.

5.3.5 MODELOS PARAFAC-MLR

5.3.5.1 Introducción

En los resultados presentados hasta este momento, se ha mostrado como el factor temperatura influye en la cinética de la fermentación de manera altamente significativa y cómo los modelos creados para los analitos estudiados son capaces de monitorizar las variaciones que sufre el proceso fermentativo. Sin embargo, son conocidas las alteraciones que sufren los espectros NIR inducidas por los cambios de temperatura [25], [41]. La consecuencia inmediata de estas distorsiones es la falta de robustez que experimentan los modelos frente a variaciones no modeladas de temperatura, con lo cual, la capacidad predictiva de los modelos se ve alterada de forma negativa. Sirvan las siguientes citas como ejemplos, de aplicaciones recientes, donde se manifiesta el problema descrito [42 - 46].

La versatilidad y capacidad predictiva de los modelos presentados hasta este momento para trabajar a diferentes condiciones de temperatura se debe a que:

- La temperatura ha sido implícitamente modelada al construir los modelos de calibración.
- El rango de temperatura donde se realiza la fermentación alcohólica en condiciones normales es estrecho, 25-35ºC y el efecto distorsionador no llega afectar de una manera problemática.

---

[41] DeBraekeleer K., Cuesta Sánchez F., Halley P.A., Sharp D.C.A., Pettman A.J., Massart D.L., "Influence and correction of temperatura perturbations on NIR spectra during the monitoring of a polymorph conversion process prior to self-modelling mixture analysis", J. Pharma, Biomedical Anal., **1998**, 17, 141-152.

[42] Shujun a., Rajiv N., Carpenter J. F.; Manning, M. C. "Noninvasive determination of protein conformation in the solid state using near infrared (NIR) spectroscopy", J. of Pharma, Science, **2005**, 94(9), 2030-2038.

[43] Zachariassen C. B.; Larsen J., van den Berg F., Balling Engelsen S., "Use of NIR spectroscopy and chemometrics for on-line process monitoring of ammonia in Low Methoxylated Amilated pectin production", Chemom. Intell. Lab. Syst., **2005**, 76(2), 149-161.

[44] Lima F. S. G., Araujo M., Borges L., "Determination of lubricant base oil properties by near infrared spectroscopy using different sample and variable selection methods", J. NIRS, **2004**, 12(3), 159-166.

[45] Chauchard F., Roger, J. M., Bellon-Maurel V., "Correction of the temperature effect on near infrared calibration. Application to soluble solid content prediction", J. NIRS, **2004**, 12(3), 199-205.

[46] Watari M., Ozaki Y., "Prediction of ethylene content in melt-state random and block polypropylene by near-infrared spectroscopy and chemometrics: Influence of a change in sample temperature and its compensation method", Appl. Spectrosc., **2005**, 59(5), 600-610.

- Las condiciones de trabajo fueron isotermas durante todo el proceso.

Sin embargo, el insidioso problema de la temperatura captó nuestro interés ya que en la mayoría de los procesos reales la temperatura no se controla ni se mantiene constante, es más, va cambiando a lo largo del proceso.

Recientemente se ha publicado una revisión donde se recogen las diferentes herramientas quimiométricas que han sido empleadas con éxito para abordar el problema distorsionador de la temperatura [47]. Sin embargo, en él no aparece recogido ningún estudio ni aproximación de los métodos *3-way* para hacer frente a este problema. El objetivo de este trabajo fue estudiar la potencialidad de la combinación PARAFAC y MLR para crear modelos libres del efecto de la temperatura. Una vez desarrollado el algoritmo combinado, se aplicó a dos conjuntos de espectros NIR, tal y como se describió en el apartado de metodología.

5.3.5.2 Conjuntos de datos *in-line*

5.3.5.2.1 Descripción de los resultados PARAFAC

Los resultados proporcionados por PARAFAC tras ser aplicado a la estructura formada por los *batches* agua, etanol y la mezcla de ambos, pueden verse en la figura 5.38.

En el primer modo, en la esquina superior izquierda, se observa como el perfil del primer componente (en rojo) muestra un gran parecido con el perfil de temperatura aplicado, existiendo un coeficiente de correlación entre ambos de 0.9966. Así, el primer componente en el segundo modo, representando en la esquina superior derecha, presenta unos mínimos característicos en la zona donde se manifiesta el enlace OH del agua, en torno 1450 y 1940nm. Se ha colocado el espectro del agua a la misma escala de abscisas (esquina inferior derecha) para facilitar su comparación. Este componente muestra una evidente

[47] Hageman, J.A., Westerhuis J.A., Smilde A.K., "Temperature robust multivariate calibration: an overview of methods for dealing with temperature influences on near infrared spectra", J. Near Infrared Spectrosc., **2005**, 13, 53-62.

relación con el *loading vector* que Libnau et al [48] asignaron a las variaciones provocadas por la temperatura, cuando estudiaron su efecto sobre el espectro puro del agua. Por último, en el tercer modo, el primer componente muestra el valor más alto en los *batches* donde existe agua y su valor es próximo a cero en el *batch* de etanol.

En relación con los componentes segundo y tercero se puede comprobar, atendiendo al primer modo, que presentan una tendencia parecida al perfil de dilución aplicado. Estos componentes, pero en el segundo modo, recogen no sólo las bandas características del agua, sino también otras estructuras presentes en el etanol como son: el primer sobretono del grupo CH_2 y CH_3 alrededor de 1700nm, la banda de combinación de C-OH alrededor de 2000nm y diferentes bandas de combinación del grupo CH_2 alrededor de 2280 nm. El espectro del etanol se ha colocado en la esquina inferior derecha, para facilitar su comparación.

Similares comportamientos se aprecian cuando se analizan los resultados obtenidos a través de PARAFAC de las otras dos estructuras numéricas. En la figura 5.39, aparecen representados los perfiles espectrales, correspondientes a los componentes asociados principalmente con la variación de temperatura, en las tres estructuras de datos estudiadas (glicerina-agua, etanol-agua y glicerina-etanol). Se puede observar como, en las mezclas en las que el agua está involucrada (estructuras etanol-agua y glicerina-agua), los perfiles presentan una forma muy similar, apareciendo prácticamente solapados y, como ya se ha comentado dos parágrafos más atrás, presenta dos mínimos característicos en la zona donde se manifiesta la combinación de tono y el primer sobretodo del enlace OH presente en el agua. Entre estos dos vectores existe un coeficiente de correlación de 0.9975, excluyendo de este cálculo el tramo final de longitudes de onda, comprendido entre 2200-2350 nm adscrito al etanol. Sin embargo, en el *batch* glicerina-etanol, el *loading* adscrito a la variación de temperatura, presenta una forma similar a los *loadings* descritos anteriormente, pero los dos mínimos que

[48] Libnau F.O., Kvalheim O.M.., Christy A.A., Toft J., "Spectra of water in the near- and mid-infrared region", Vib. Spectrosc. **1994**, 7, 243-254.

caracterizan el perfil están desplazados unos 50 nm en la región alrededor de 1400 nm y unos 200 nm en la región en torno a 1880 nm. Estos valores marcan la diferencia entre la zona de absorción donde se manifiestan el primer sobretono y la banda de combinación del grupo OH del agua y del grupo hidroxilo unido a un radical, R-OH.

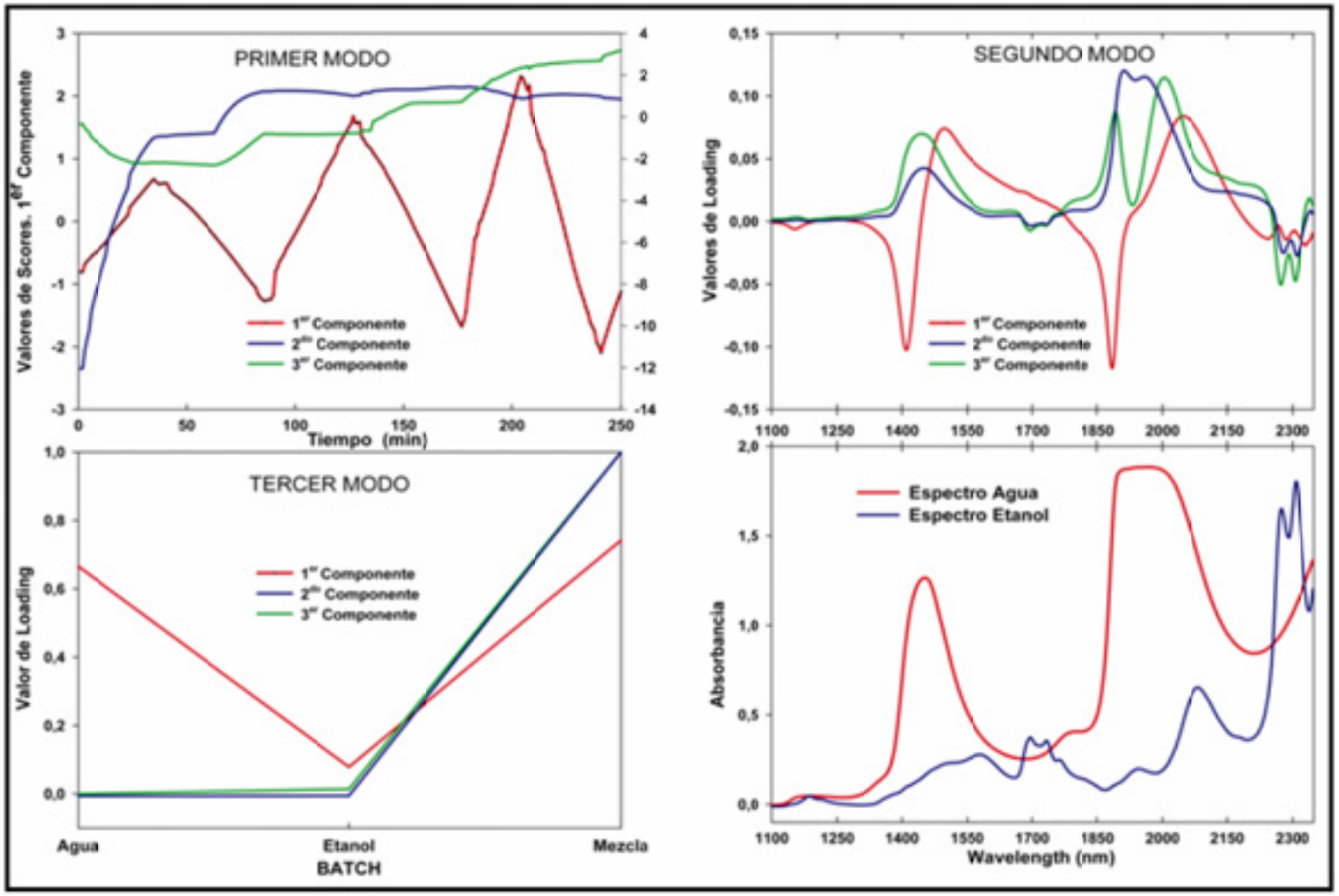


Figura 5.38. Resultados proporcionados por PARAFAC en los tres modos al ser aplicados a un cubo de datos formado a partir de agua y etanol. Los espectros de agua y etanol se han colocado en la esquina inferior derecha, en la misma escala, para facilitar la comparación.

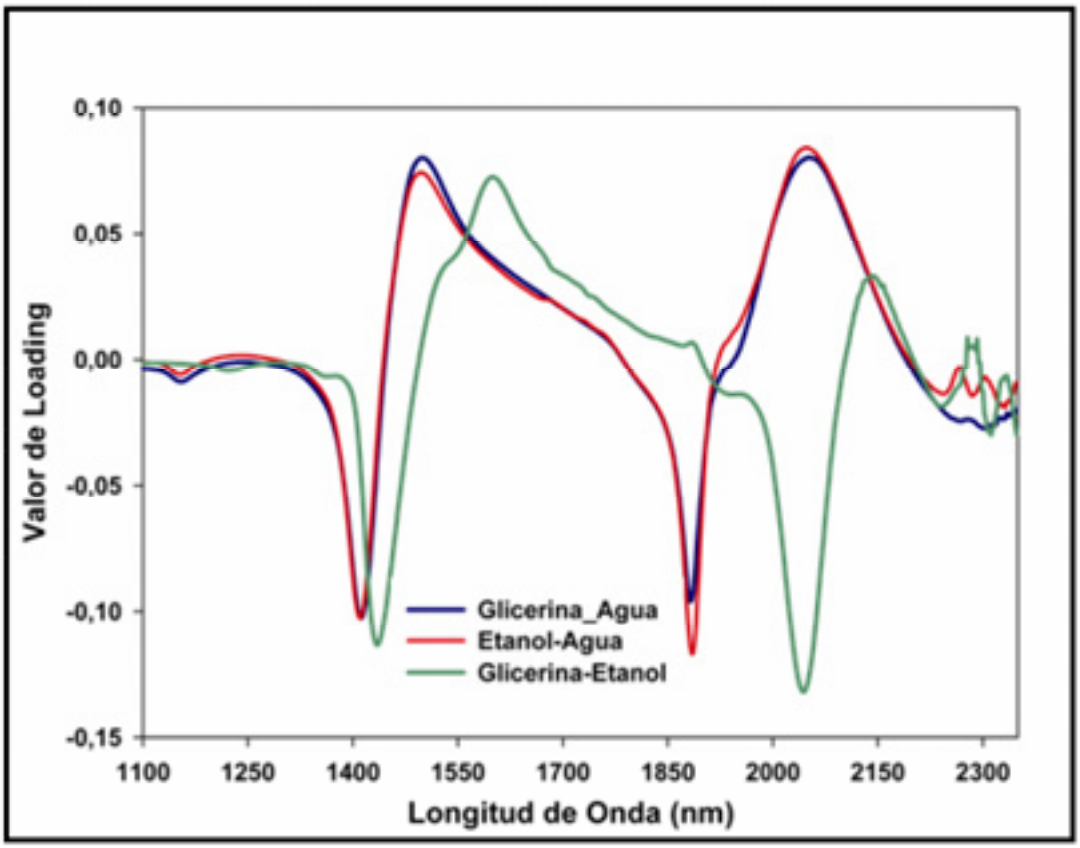


Figura 5.39. ***Loadings*****, proporcionados por PARAFAC, que pueden ser asignadas a la variación de temperatura para las tres estructuras de datos estudiadas.**

5.3.5.2.2 Modelos PARAFAC-MLR

Los *scores* proporcionados por el modelo PARAFAC fueron utilizados para construir los modelos MLR para temperatura y concentración inicial de analito. De cada cinco espectros registrados consecutivamente, uno fue engrosado en el conjunto de calibración y los otros cuatro formaron parte del conjunto de validación.

Se ha de resaltar que todos los modelos PARAFAC-MLR fueron construidos con tres factores y, en este punto, se podría entrar en la discusión de que quizás dos factores bastarían para modelar el sistema, uno para la temperatura y otro para el analito, sin embargo, éste no es un sistema lineal y la utilización de un componente adicional sirve para modelar las no linealidades introducidas por la temperatura.

En la tabla 5.9 aparecen recogidos los estadísticos que avalan la habilidad predictiva de los modelos creados. Donde Nc/Np es el número de muestras utilizadas en el conjunto de calibración frente al número de muestras empleadas en el conjunto de validación externa, RMSEP es la raíz del error cuadrado medio de calibración y R^2 el coeficiente de determinación entre los valores predichos y de referencia. La relación Nc/Np es menor en las estructura glicerina/etanol y glicerina/agua ya que se eliminaron los espectros registrados en las primeras fases del proceso de dilución debido a la deficiente homogeneidad del medio, provocada por la diferencia de densidades entre los dos analitos. Como se puede ver, el coeficiente de determinación para todos los modelos creados es altamente significativo y el intervalo de confianza para la pendiente y la ordenada en el origen, para un nivel de significación del 95%, incluye el 1 y el 0 respectivamente.

Tabla 5.9. Estadísticos obtenidos tras validar externamente el modelo PARAFAC-MLR.

ESTRUCTURA	PROPIEDAD	Nc/Np	RMSEP	R^2	PENDIENTE	INTERCEPTO
Etanol/	Temperatura	51/207	0.0079	0.9984	0.9962 ± 0.0077	0.1468 ± 0.3221
Agua	[Etanol]	51/207	0.0039	0.9997	1.0004 ± 0.0032	0.0032 ± 0.1634
Glicerina/	Temperatura	51/207	0.0095	0.9982	1.0083 ± 0.0085	-0.2895 ± 0.3566
Agua	[Glicerina]	43/165	0.0064	0.9993	1.0026 ± 0.0057	0.2846 ± 0.3188
Glicerina/	Temperatura	47/186	0.0066	0.9991	0.9994 ± 0.0062	0.0475 ± 0.2589
Etanol	[Glicerina]	40/160	0.0070	0.9984	0.9935 ± 0.0088	0.3476 ± 0.5138

5.3.5.3 Conjunto de datos *at-line*

Tal y como se ha comentando, este conjunto de datos se ha convertido en un referente donde diferentes grupos prueban la habilidad de sus algoritmos para hacer frente al efecto distorsionador de la temperatura.

Los conjuntos de calibración y validación utilizados fueron los mismos que los que aparecen citados en la literatura [5] con el propósito de poder establecer y evaluar la utilidad del algoritmo propuesto. El conjunto de calibración constaba de de 14 muestras; 13 de ellas circunscritas en la periferia del diseño experimental más el punto central. El conjunto de validación externa constaba de 6 muestras inscritas entre las muestras de calibración.Ver figura 5.9, (página 96).

5.3.5.3.1 Modelos PARAFAC-MLR

Para seleccionar el número óptimo de componentes y estudiar la influencia de los diferentes pretratamientos espectrales en la capacidad predictiva, se construyeron modelos globales para cada uno de los analitos, es decir, las cinco temperaturas fueron incluidas en la construcción del cubo de datos. Los mejores resultados en términos de RMSEP se consiguieron con cuatro componentes y primera derivada (ventana de 7 puntos ajustada a un polinomio de segundo orden). Los valores RMSEP obtenidos fueron 0.0032 para agua, 0.0067 para etanol y 0.0071 para iso-propanol.

De acuerdo a estos resultados, se construyeron dos modelos locales en primera derivada y con cuatro componentes:

- Uno de ellos fue construido con las muestras a las temperaturas extremas (30 y 70ºC) y se utilizó para validar las muestras registradas a 30ºC y 70ºC y las registradas a 50ºC y 60ºC. Este modelo nos permite estudiar la habilidad de interpolación del algoritmo (en rojo, tabla 5.10).
- El otro modelo fue construido con las muestras registradas a 50ºC y 60ºC y se utilizó para predecir tanto las muestras del conjunto de calibración registradas a 50ºC y 60ºC como las registradas a una temperatura extrema (30 y 70ºC). Este modelo nos permite estudiar la habilidad de extrapolación del algoritmo (en azul, tabla 5.10).

Tabla 5.10. Valores RMSEP para las validaciones de los modelos PARAFAC-MLR y PLS.

Temperatura Validación (ºC)	Especie Química	Temperatura de Calibración (ºC) 30 & 70		50 & 60	
		PARAFAC_MLR	PLS	PARAFAC_MLR	PLS
30 & 70	Agua	0.0044	0.0047 (8)	0.0024	0.0176 (6)
	Etanol	0.0082	0.0149 (9)	0.0147	0.0311 (11)
	Isopropanol	0.0086	0.0129 (9)	0.0143	0.0194 (9)
50 & 60	Agua	0.0035	0.0072 (5)	0.0045	0.0059 (6)
	Etanol	0.0106	0.0257 (6)	0.0099	0.0142 (9)
	Isopropanol	0.0091	0.0177 (5)	0.0096	0.0162 (9)
MODELO GLOBAL		Agua 0.0032	Ethanol 0.0067	Isopropanol 0.0071	

El número de componentes para los modelos PARAFAC-MLR fue siempre de cuatro y para los modelos PLS el indicado entre paréntesis.

En la tabla 5.10 se muestran los resultados obtenidos para los modelos PARAFAC-MLR y para los respectivos modelos PLS, (la cifra en paréntesis indica el número de componentes usados).

Como se puede observar, los mejores resultados fueron para el agua, hecho que era de esperar debido a:

- Mayor absortividad (que implica una mayor sensibilidad) del enlace OH del agua en relación a otras estructuras.
- Los *scores* utilizados en el modelo PARAFAC están directamente relacionados con los perfiles espectrales y éstos recogen mayoritariamente la información de las estructuras que más cambian al variar la temperatura, es decir, del enlace OH del agua.

La bondad de los modelos locales para los tres analitos, cuando fueron validados con muestras registradas a la misma temperatura que las utilizadas en calibración, no fue significativamente diferente de los resultados obtenidos con el modelo global; al igual que tampoco mostraron diferencia con los modelos globales, los modelos locales para el agua, tanto en condiciones de extrapolación (cifras en azul) como de interpolación (cifras en rojo). Sin embargo, los modelos locales para etanol e isopropanol si proporcionaron valores de RMSEP mayores que los obtenidos con los respectivos modelos globales. Pero, a pesar del mayor valor de RMSEP obtenido para estos dos analitos, los resultados son aceptables y no son significativamente diferentes a los publicados recientemente por otro grupo de investigación. En el trabajo divulgaban, en base a sus resultados obtenidos [49], haber eliminado el efecto distorsionador de la temperatura en la habilidad predictiva de los modelos construidos.

Comparando los valores de RMSEP proporcionados por los métodos PARAFAC-MLR y PLS, se observa como, en términos generales, los valores obtenidos son mayores en el caso de PLS. Además, los modelos PLS son complejos en número de componentes, nunca menos de cinco. Cuando para los modelos PARAFAC-MLR, el número de componentes fue cuatro en todos los casos.

[49] Chen Z., Morris J., Martin E., "Correction of Temperature-Induced Spectral Variations by Loading Space Stanardization", Anal. Chem., **2005**, 77, 1376-1384.