UTILIZACIÓN COMBINADA DE MÉTODOS EXPLORATORIOS Y CONFIRMATORIOS PARA EL ANÁLISIS DE LA ACTIVIDAD ANTIBACTERIANA DE LA CEFALOSPORINA (PARTE II)

Yunier E. Tejeda Rodríguez*, Valia Guerra Ones**, Jesús E. Sánchez García ** y Ramón Carrasco Velar*
* Universidad de las Ciencias Informáticas, MIC
** Instituto de Cibernética, Matemática y Física, CITMA

RESUMEN
En el presente trabajo se aplican los resultados de la estrategia combinada del Análisis de Componentes Principales Robusto y la descomposición matricial CUR que aparecen en la parte I a un problema práctico para el ajuste de un modelo de regresión mediante el uso de los mínimos cuadrados parciales (PLS). Se dan las conclusiones acerca de las ventajas de utilizar este enfoque, así como también información acerca de los modelos obtenidos.

ABSTRACT:
In the paper the results of the combined strategy formed by Robust Principal Component Analysis and CUR matrix decomposition are applied to a practical problem for adjusting a regression model by using Partial Least Squares (PLS). The conclusions concerning their advantages of such an approach as well information about the adjusted models are given.

KEYWORDS: Regression Analysis. Partial Least Squares (PLS). Regression with PLS

MSC:62P10

1. INTRODUCCIÓN

En Tejeda *et al.* (201-) se presentó la etapa exploratoria de los datos de cefalosporina. Sobre la base de los resultados obtenidos se pasa en el presente artículo a la etapa de modelación.
En la primera sección se presenta una breve descripción de los mínimos cuadrados parciales (PLS) y se discuten sus posibilidades para la modelación en este caso. A continuación se presentan los resultados de la modelación y se hace una discusión acerca de su interpretación.

2. PLANTEAMIENTO DEL PROBLEMA

Como se mencionó en Tejeda *et al.* (201-) , un aspecto de interés en el estudio de los fármacos es analizar la posibilidad de establecer un modelo que relacione la actividad biológica de los mismos con sus características estructurales. En la presente investigación se plantea la necesidad de obtener un modelo de relación estructura-actividad de cefalosporinas frente a cepas de *Streptococcus aureus* (*S. aureus*) y *Escherichia coli* (*E. coli*) con respecto a un conjunto de descriptores moleculares.

¿Por qué Regresión Mínimos Cuadrados Parciales (PLS)?

La regresión es una de las técnicas estadísticas que más se utiliza por parte de usuarios que aplican esta disciplina. Un problema con el que se tropieza frecuentemente es la presencia de multicolinealidad, esto es: altas correlaciones entre las variables predictoras. La forma más usual de resolver este problema es el empleo del análisis de componentes principales.

A este problema se ha unido en tiempos recientes el aumento extraordinario del número de variables con lo que es muy usual encontrarse con que el requisito tradicional de que haya más individuos (n) que variables (p) no se cumple y, por tanto, los métodos tradicionales no deben emplearse. Un ejemplo ya clásico de esta problemática lo constituyen los microarrays, tan usados en las investigaciones contemporáneas de genética. Si se unen ambas situaciones, se tiene una situación de difícil solución con las técnicas tradicionales de la estadística. Está claro que una parte del problema se resuelve con la selección de variables y para ello nuevamente puede utilizarse el análisis de componentes principales.

En el año 1966 apareció el trabajo de H. WOLD en el que se presenta por primera vez lo que se conoce actualmente como Partial Least Squares o PLS. A éste le seguirían otros artículos en que se elaboró más la técnica y, con posterioridad, los trabajos fueron continuados por su hijo S. WOLD (ver, por ejemplo: SJÖSTRÖM et al., 1983) acompañado de un grupo de especialistas noruegos entre los que se puede señalar de manera especial a H. MARTENS y T. NAES, autores del conocido libro Multivariate Calibration (1989). La idea básica del PLS (GELADI & KOWALSKI, 1986), (BOULESTEIX & STRIMMER, 2007) es la reducción de la dimensión en regresión múltiple, con la garantía de que las primeras componentes ortogonales mejoran la predicción. Como es sabido, esta es una característica que no poseen las componentes principales (véase MARDIA et al., 1979). BETZIN (2000) resume en una oración lo que podría ser la ventaja de PLS sobre ACP: "la estimación de las ponderaciones según el análisis de componentes principales es "óptima", pero se pierden las relaciones de dependencia. Sobre esto mismo, puede añadirse el comentario de LOHMÖLLER (1984): "(PLS) es una generalización del ACP en el sentido de la introducción de relaciones".

3. REGRESIÓN PLS

Se supone que existen *q* variables dependientes $Y_1, \ldots, Y_q$ de *p* variables independientes $X_1, \ldots, X_p$. Se dispone de *n* observaciones y se desea ajustar un modelo de regresión. Los datos se resumen en forma matricial: $Y_{n\times q}$ y $X_{n\times p}$, respectivamente.

La idea básica es hallar una descomposición en factores latentes *T* tal que:

$$Y = TQ' + F$$
$$X = TP' + E$$

Donde *T* es una matriz de $n \times c$, que contiene las componentes latentes de las *n* observaciones. Por su parte, *P*, de $p \times c$, y *Q*, de $q \times c$, son matrices de coeficientes. *E* y *F*, de dimensiones $n \times p$ y $n \times q$, respectivamente, son matrices de errores aleatorios.

PLS es un método para construir una matriz *T* que sea una transformación lineal de *X:*

$$T = XW$$

Una vez obtenida, esta transformación se utiliza en la regresión en lugar de la matriz original. Finalmente, el modelo se expresa en las variables originales, haciendo la transformación "inversa". Esto es:

$$Q' = (T'T)^{-1}T'Y$$

que no es más que la matriz de coeficientes para el modelo transformado. Al multiplicar *Q´* por *T*, se obtiene la matriz de los coeficientes asociados a las variables originales:

$$B = WQ'$$

Un caso particular: Regresión simple

En este caso, la matriz de variables dependientes se reduce a un vector de dimensión *n.* Aquí, el PLS puede considerarse como una transformación de las variables independientes, teniendo en cuenta su relación con la dependiente. Esta es precisamente la gran diferencia con el ACP en el que la transformación se aplica sólo a la matriz *X* (TEJEDA, 2011).

A continuación se presenta el algoritmo muy simplificado para la obtención de las componentes PLS (VEGA, 2004):

1. Entrada de datos: $X_{n\times p}$, $Y_{n\times 1}$
2. Para i = 1 hasta *p*
3. $w = \mathrm{cov}(Y,X)$, $w = \frac{w}{\|w\|}$
4. $T = Xw$
5. $v = (T'Y)/(T'T)$
6. $b = (T'X)/(T'T)$
7. $X = X - Tb = X - \hat{X}$
8. $Y = Y - Tv = Y - \hat{Y}$
9. Fin i

En los pasos 3 y 4 está uno de los puntos esenciales del método propuesto: la variable latente se conforma a partir de la covarianza entre las variables independientes y, en este caso, la dependiente. En los pasos 5 y 6 es fácil ver que *v* no es más que el coeficiente de regresión simple de *Y* sobre *T,* y *b* es un vector de dimensión *p* cuyas componentes no son más que los coeficientes de regresión simple de cada variable independiente, X_i, sobre *T.* Los pasos 7 y 8 del algoritmo son los de actualización de los valores.

El algoritmo expuesto es iterativo y en cada una de las iteraciones calcula una componente latente.

A partir de lo expuesto anteriormente puede llegarse a lo que constituye el fundamento del método de PLS: La maximización del cuadrado de la covarianza entre la componente latente y la variable dependiente, sujeto a la restricción de $w'w = 1$. Esto lleva a la aplicación de los multiplicadores de Lagrange y la solución no es más que el vector de covarianzas normalizado.

Selección del número de componentes

El criterio para la selección del número de componentes es la minimización de la suma de cuadrados de los residuos. Los criterios más empleados son:

- Estimación de la suma de cuadrados de los residuos mediante validación cruzada
- Estimación de la suma de cuadrados de predicción PRESS (por sus siglas en inglés: **Pre**diction **S**um of **S**quares)

Validación cruzada

La validación cruzada es un método general de estimación muy utilizado en Estadística, cuya idea básica es utilizar solamente parte de la muestra y luego de estimado el modelo en cuestión, hallar la estimación para esas observaciones que no se incluyeron. El método para la selección de componentes en PLS parte de dividir la muestra en *k* partes. Cada una de las partes debe contener aproximadamente n/k observaciones. La literatura especializada recomienda $k = 3, 10$ ó n (en este caso particular se obtiene PRESS). En cada paso se excluye una de las *k* partes y se ajusta el modelo con las restantes observaciones. Una vez estimado los

parámetros correspondientes se hallan los estimados de las dejadas fuera y se halla la suma de cuadrados de los residuos. El procedimiento se repite para cada una de las *k* partes y finalmente se halla el promedio:

$$SCR_j = \sum_{\{i:x_i \in V_j\}} (y_i^{(j)} - \hat{y}_i^{(j)})^2 \ , \quad j = 1,...,k \ ,$$

donde $y_i^{(j)}$ y $\hat{y}_i^{(j)}$ son la i-ésima observación del j-ésimo conjunto y su estimación, respectivamente. A continuación, una vez barrido el índice *j*, se calcula el promedio:

$$SCR_{VC} = \frac{1}{n} \sum_{j=1}^{k} SCR_j$$

Suma de cuadrados de predicción

Este no es más que un caso particular de lo visto anteriormente y se conoce también con el nombre de leave-one-out.

Como su nombre lo indica, aquí se deja fuera una observación cada vez, se estima el modelo y se evalúa esa observación. Así se procede hasta que todas hayan sido dejadas fuera una vez. A continuación se calcula el PRESS promedio:

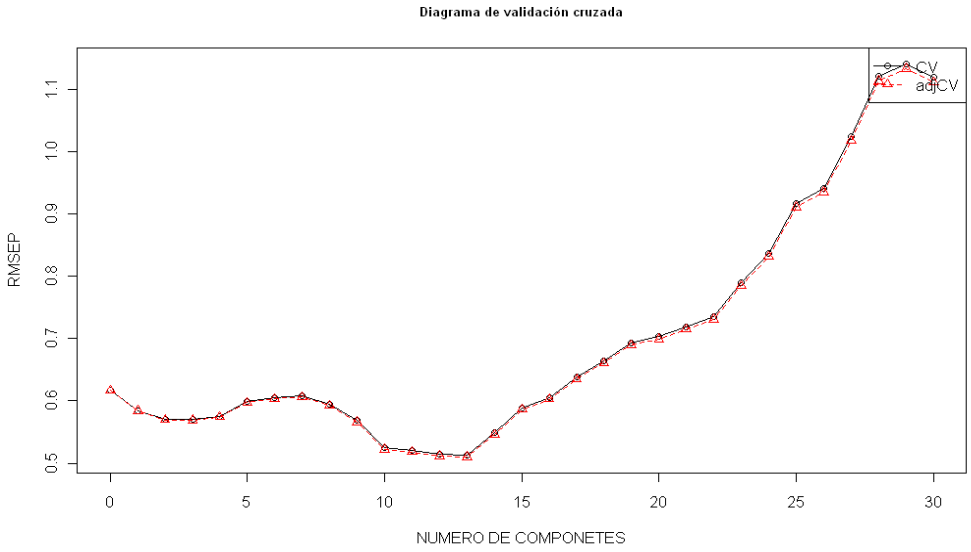
$$PRESS_{med} = \frac{1}{n} \sum_{i=1}^{n} e_{(i)}^2 \text{ , donde } e_{(i)} = y_{(i)} - \hat{y}_{(i)} \text{ , } i = 1,\ldots.n$$

4. RESULTADOS Y DISCUSIÓN

La aplicación del ACP robusto en Tejeda *et al.* (201-), llevó a la eliminación de 32 compuestos, por lo que las dimensiones de la matriz de datos fueron de 72 x 97. Este hecho, unido a la presencia de la multicolinealidad, implicó el uso de la regresión PLS. La obtención de los modelos de regresión se hizo con los tres algoritmos más conocidos de PLS, a saber: NIPALS, Kernel PLS y SIMPLS. Para la estimación de las regresiones se utilizó el código público R-UCA 2.11.
Como variables dependientes se tomaron las actividades biológicas del *S. aureus* y del *E. coli,* respectivamente.

Modelo de regresión PLS usando el algoritmo NIPALS para correlacionar las estructuras de cefalosporinas con su actividad frente al *S. aureus*:

Se utilizó el algoritmo NIPALS que se encuentra en el *paquete pls* del software R R DEVELOPMENT CORE TEAM (2010) para la obtención de las componentes latentes. Para determinar el número de estas, se graficó la raíz cuadrada de los errores cuadrados medios de predicción (RMSEP, por sus siglas en inglés: Root Mean Squared Error Prediction) (VARMUZA & FILZMOSER**,** 2008) contra el número de componentes seleccionando aquella que tuviera menor RMSEP. El criterio recomienda tomar el número de componentes igual al correspondiente a un mínimo en la gráfica. En este caso fueron 13.
De la salida del software R se puede observar que el modelo es adecuado pues el *p-valor* para la prueba F es 2,2e-16, menor que el nivel de significación que es 0,05. El R cuadrado múltiple es 0,858 lo que significa que el modelo explica el 86% de la variabilidad de los datos. Los residuos son relativamente pequeños siendo el primer cuartil -0,14428 y el tercer cuartil 0,12258. Casi todos los coeficientes son significativos pues sus *p-valores* son menores que 0,05, exceptuando sólo los coeficientes 11, 12 y 13.

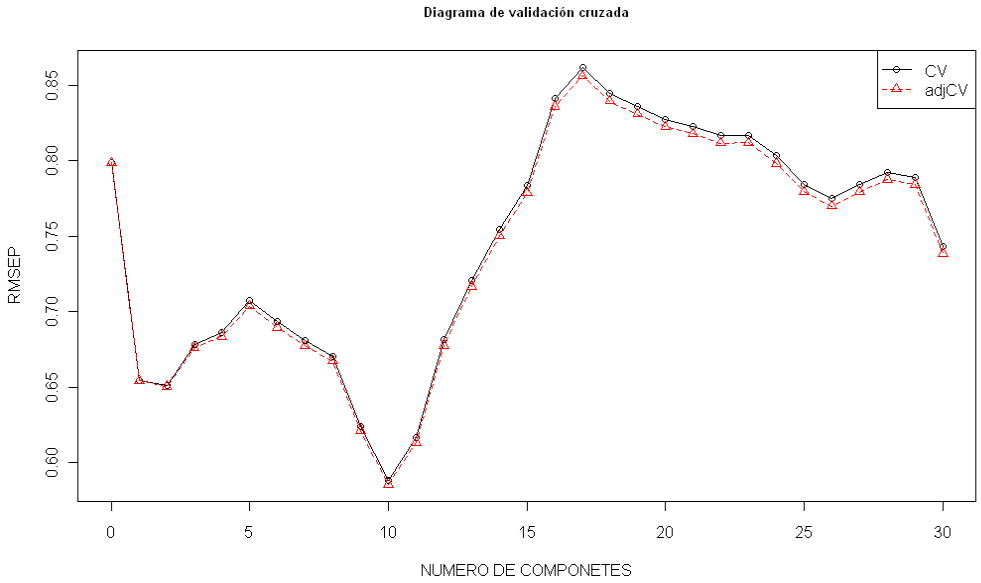



```
Residuals:
     Min       1Q   Median       3Q      Max
-0.67932 -0.14428  0.02154  0.12258  0.57447

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.91225    0.03009 196.488  < 2e-16 ***
X1           0.04101    0.00470   8.725 3.78e-12 ***
X2           0.08720    0.01047   8.331 1.71e-11 ***
X3           0.11194    0.01233   9.077 9.89e-13 ***
X4           0.12858    0.02426   5.301 1.87e-06 ***
X5           0.11734    0.02486   4.719 1.54e-05 ***
X6           0.11621    0.02752   4.223 8.62e-05 ***
X7           0.11958    0.02788   4.289 6.88e-05 ***
X8           0.09299    0.02924   3.180  0.00236 **
X9           0.08732    0.02871   3.041  0.00354 **
X10          0.10746    0.04034   2.664  0.00999 **
X11          0.04333    0.02885   1.502  0.13847
X12          0.09024    0.04795   1.882  0.06485 .
X13          0.07070    0.03772   1.874  0.06595 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2553 on 58 degrees of freedom
Multiple R-squared: 0.8579,     Adjusted R-squared: 0.8261
F-statistic: 26.95 on 13 and 58 DF,  p-value: < 2.2e-16
```

Observación: Los modelos de regresión PLS con los algoritmos Kernel PLS y SIMPLS dieron los mismos resultados.

Modelo de regresión PLS usando el algoritmo NIPALS para correlacionar las estructuras de cefalosporinas con su actividad frente al *E. Coli*:

Se utilizó el algoritmo NIPALS que se encuentra en el *paquete pls* del software R para la obtención de las componentes latentes. Para determinar el número de estas, se graficó el RMSEP contra el número de componentes seleccionando aquella que tuviera menor RMSEP. El criterio recomienda tomar el número de componentes igual al correspondiente a un mínimo en la gráfica. En este caso fueron 10.



De la salida del software R se puede observar que el modelo es adecuado pues el *p-valor* para la prueba F es 2,2e-16, menor que el nivel de significación que es 0,05. El R cuadrado múltiple es 0,898 lo que significa que el modelo explica el 90% de la variabilidad de los datos. Los residuos son relativamente pequeños siendo el primer cuartil -0,17563 y el tercer cuartil 0,17531. Todos los coeficientes son significativos pues sus *p-valores* son menores que 0,05.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.58344 -0.17563  0.03052  0.17531  0.52519

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.135221   0.032220 190.414  < 2e-16 ***
X1          0.069470   0.004645  14.956  < 2e-16 ***
X2          0.157332   0.015476  10.166 9.45e-15 ***
X3          0.099824   0.011232   8.888 1.33e-12 ***
X4          0.180094   0.023673   7.608 2.09e-10 ***
X5          0.136140   0.024569   5.541 6.79e-07 ***
X6          0.115919   0.028410   4.080 0.000133 ***
X7          0.084407   0.032354   2.609 0.011409 *
X8          0.058858   0.029210   2.015 0.048317 *
X9          0.065994   0.032362   2.039 0.045766 *
X10         0.155713   0.048981   3.179 0.002322 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2734 on 61 degrees of freedom
Multiple R-squared: 0.8979,     Adjusted R-squared: 0.8811
F-statistic: 53.64 on 10 and 61 DF,  p-value: < 2.2e-16
```

Observación: Los modelos de regresión PLS con los algoritmos Kernel PLS y SIM-PLS dieron los mismos resultados.

5. CONCLUSIONES

1. El ACP Robusto es una manera para mejorar la información empleada en la modelación
2. Se logró un buen ajuste tanto para la *E. coli* como para *S. aureus* en función de las variables estructurales, mediante el uso de la regresión con PLS

RECEIVED SEPTEMBER, 2011
REVISED MARCH, 2012

REFERENCIAS

[1] BETZIN, J. (2000): PLS-Pfadanalyse und KFA-Modelle – Globale und lokale Entdeckungen, **Psychologische Beiträge**, 42, 469-493

[2] BOULESTEIX, A.-L. and STRIMMER, K. (2007): Partial Least Squares: A Versatile Tool for the Analysis of High-dimensional Genomic Data, **Bioinformatic**s, 8, . 32-44

[3] CARRASCO, R. (2008): **Nuevos descriptores atómicos y moleculares para estudios de estructura-actividad: Aplicaciones.** Tesis en opción al grado de Doctor en Ciencias Químicas, La Habana.

[4] GELADI, P. and KOWALSKI, B. (1986): Partial Least Squares Regression: A tutorial, **Analytica Chmica Acta,** 185, 1-17

[5] LOHMÖLLER, J.-B. (1984): Das Programmsystem LVPLS für Pfadmodelle mit Latenten Variablen, *ZA*-**Information Nr. 15.**

[6] MARDIA, K.V., KENT, J. T. and BIBBY, J.M. (1979): **Multivariate Analysis**, Academic Press, Londres.

[7] MARTENS, H. and NAES, T. (1989): **Multivariate Calibration**, Wiley, Nueva York

[8] SJÖSTRÖM, M., WOLD, S., LINDBERG, W., PERSSON, J.-A. and MARTENS, H. (1983): A Multivariate Calibration Problem in Analytical Chemistry Solved by Partial Least-Squares Models in Latent Variables, **Analytica Chimica Acta**, 150, 61-70

[9] R DEVELOPMENT CORE TEAM (2010): **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

[10] TEJEDA, Y. (2011): **Utilización combinada de métodos exploratorios y confirmatorios para el análisis de la actividad antibacteriana de la cefalosporina**. Tesis en opción al grado de Máter en Ciencias Matemáticas, La Habana.

[11] TEJEDA, Y., GUERRA, V., SÁNCHEZ, J. and CARRASCO, R. (201-): Utilización Combinada de métodos exploratorios para el análisis de la actividad antibacteriana de la cefalosporina (Parte I), Acptado,**Revista de Investigación Operacional.**

[12] VARMUZA, K. and FILZMOSER, P. (2008): **Introduction to multivariate statistical analysis in chemometrics,** CRC Press, Boca Raton

[13] VEGA VILCA, J.C. (2004): **Generalizaciones de mínimos cuadrados parciales con aplicación en clasificación supervisada.** Tesis Presentada en Opción al grado de Dr. en Filosofía, Universidad de Puerto Rico

[14] WOLD, H. (1966): Estimation of principal components and related models by iterative least squares, En: KRISHNAIAH, P.R. (Ed.): **Multivariate Analysis**, Academic Press, Nueva York