METODOS PLS EN ANALISIS MULTIBLOQUES:

DATOS ESPARCIDOS, FALTANTES Y MULTICOLINEALIDAD

VICTOR MANUEL GONZALEZ ROJAS

UNIVERSIDAD DEL VALLE – ESCUELA DE ESTADISTICA

Santiago de Cali, Abril de 2016

INTRODUCCION

A partir de los años 70, aparecieron los métodos denominados Partial Least Squares (PLS).

Procedimientos algorítmicos basados en productos escalares entre vectores, que iteran hasta la convergencia obteniendo las componentes (LV) más relacionadas de varias matrices de datos.

Para el tratamiento de k matrices de datos se creo:

- k=1, Nonlinear estimation by Iterative PLS (NIPALS [ACP])
- k=2, PLS Regression (PLSR: PLS1, PLS2)
- k > 2, PLS Path Modeling (PLS-PM [RGCCA])

PLS-R : Regresión PLS1

REGRESIÓN PLS1

- Se quiere estimar via PLS la regresión múltiple

$$Y = \boldsymbol{\beta_1 X_1} + \cdots + \boldsymbol{\beta_p X_p} + \boldsymbol{\varepsilon}$$

Las variables explicativas $X_1, \ldots, X_p$ pueden presentar:

- **Multicolinealidad**
- **Matrices esparcidas : $p > n$,**
- ***Datos mixtos y faltantes.***

Tanto las X_j como Y que $\in R^n$ se suponen *centradas – reducidas.*

- Se busca en el espacio de las predictoras componentes ortogonales $t = Xw$ correlacionadas con Y, tal que se realice la regresión

$$Y = c_1 t_1 + \cdots + c_H t_H + Y_H$$

para luego mediante desdoblamiento de las $t = f(X)$ que son combinación lineal de las X_j estimar el modelo. Generalmente se tiene que $H < p$.

- La idea es encontrar w tal que se maximice el cuadrado de la covarianza entre el componente $t = Xw$, y la variable respuesta Y bajo $w'w = 1$;

- Sea V la matriz de orden $p\ x\ 1$, el vector de covarianzas de X e Y $(V = X'Y)$, entonces se $maximiza$

$$cov^2(t, Y) = [w'cov(X, Y)]^2 = [w'V]^2 = w'VV'w$$

lagrangiano $\emptyset$ que **maximiza** $\boldsymbol{cov^2(t, Y)}$, sujeta a ortonormalidad de w, es:

- $\emptyset = w'VV'w - \lambda(w'w - 1)$

$$\frac{\partial \emptyset}{\partial w} = 2VV'w - 2\lambda w = 0$$

$$VV'w = \lambda w$$

$\lambda\ y\ w$ autovalor y autovector de VV'. Además

$$\boldsymbol{w} = \frac{\boldsymbol{V}}{\|\boldsymbol{V}\|} = \frac{\boldsymbol{X'Y}}{\|\boldsymbol{X'Y}\|}$$

vector de covarianzas normalizado.

- LA IDEA FUNDAMENTAL ES CONSEGUIR COMPONENTES ORTOGONALES *t1, t2,.. SOBRE LAS CUALES SE REALIZA* LA REGRESIÓN

- $Y \sim c_1 t_1 + c_2 t_2 + \ldots + \varepsilon$

- Las componentes ***t*** *[∈ Rx] que* se obtienen bajo los principios del ACP **guardan relación con *Y*** *; corrigen gracias a su ortogonalidad la multicolinealidad.*

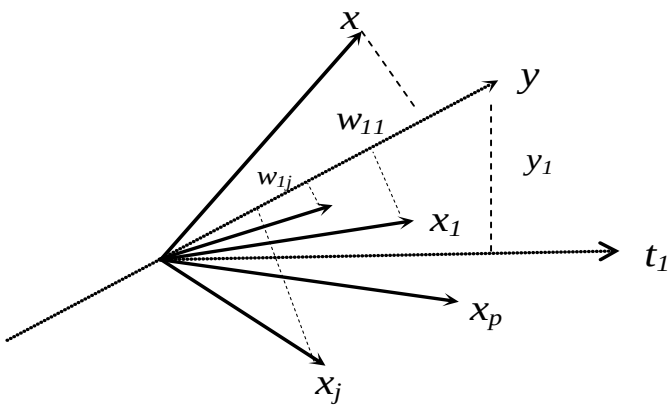
Conseguir en el subespacio R_x engendrado por las $x_1, \ldots, x_p$ de R^n, la"*componente*" t_1 más correlacionada con y, es decir, cuya regresión con y minimice el residuo generado y_1, ver grafica 1.

ACP	$X'N^{1/2}v = \sqrt{\lambda}u$	$Xu = t$
PLS	$\boldsymbol{X'y = w*}$	$\boldsymbol{Xw = t_{pls}}$

Tomo y como el "*eigenvector*" sobre el cual proyecto ortogonalmente, cada x_j, para obtener los coeficientes $w_{1j}^* = cov(x_j, y) = x_j'.y$ derivados de esta regresión. Para garantizar norma 1, se realiza $w_1 = w_1^*/\|w_1^*\|$.

Ahora se calcula la componente PLS t_1

$$\boldsymbol{t_1 = Xw_1 = w_{11}x_1 + \ldots + w_{1p}x_p}$$



Gráfica 1. Obtención de las coordenadas de w_1 por proyección de x_j sobre y.

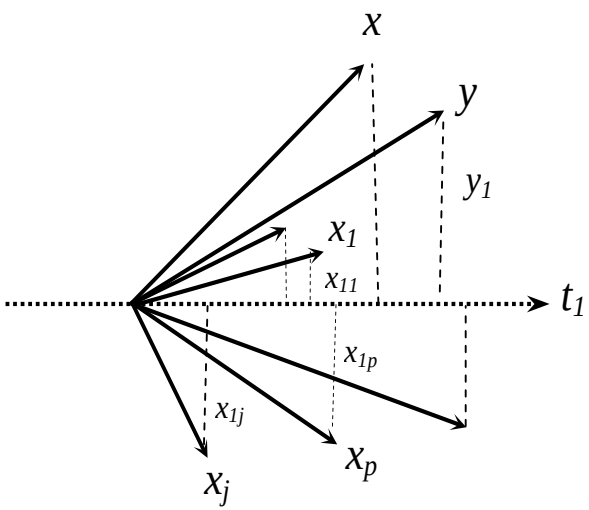
Luego se realiza la *regresión* simple de y sobre t_1 :

$$y = c_1 t_1 + y_1 = c_1 w_{11} x_1 + .. + c_1 w_{1p} x_p + y_1$$

donde $c_1 = y't_1$ es el coeficiente de regresión, y y_1 el vector de residuos [parte de y no explicada por t_1]; estos coeficientes son más fáciles de interpretar por el investigador.

Es posible que aún falte explicar buena parte de los residuos. Si el poder explicativo de la regresión $y \sim t_1$ es muy débil, se construye análogamente, una segunda componente t_2 en el espacio de los residuos de x ortogonales a t_1, y se realiza entonces la regresión

$$y \sim t_1 + t_2.$$



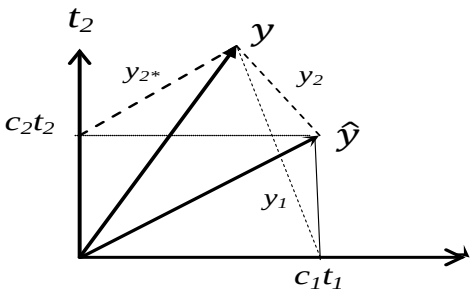
Gráfica 2. Los residuos en x e y son ortogonales a t_1.

En la grafica 2, los residuos x_{1j} (líneas punteadas) forman un subespacio ortogonal a R_x, y representan la parte de las x^s no explicada por t_1; se obtienen de realizar regresiones simples de las x_j con t_1, con lo cual $x_{1j} = x_j - p_{1j}\, t_1$.

La nueva componente t_2 es combinación lineal de estos residuos; $t_2 = w_{21}x_{11}+..+w_{2p}x_{1p}$ y explicará el residuo y_1 también ortogonal a t_1.

Procedemos análogamente como antes, situados ahora en el subespacio de los residuos ortogonal a R_x, tomo como "*vect_p*" el residuo y_1 y proyecto ortogonalmente sobre este los residuos $x1_j$ para obtener sus ponderadores w_{2j} y generar la combinación lineal denominada componente t_2 ortogonal a t_1. Los coeficientes $w_{2j} = cov(x_{1j}, y_1)/\|w_2\|$ conforman el vector "propio" w_2 de norma 1.

Se realiza entonces la regresión $y = c_1t_1 + c_2t_2 + y_2$, la cual es más precisa que la primera; ver grafica 3.



Gráfica 3. Regresión de y sobre las componentes ortogonales t_1, t_2.

Este procedimiento es iterativo, ahora se consiguen los residuos y_2, $x_{21},..,x_{2p}$ de las regresiones de y, $x_1,..,x_p$ sobre el subespacio de dos dimensiones (t_1, t_2). Los residuos x_{2j} se proyectan sobre y_2 y estas coordenadas conformarán el vector w_3 que permitirá además, construir la componente t_3 ortogonal a (t_1, t_2).

Se efectúa entonces la regresión $y = c_1t_1 + c_2t_2 + c_3t_3 + y_3$.

El algoritmo de regresión PLS1

$X_0 = X, \;\; y_0 = y \equiv \psi \;\; comp\ R^n$

para h=1,2,…,a. *(X es de rango a)*

2.1 $w_h = X'_{h-1} y_{h-1} / \|y_{h-1}\|$ [coef regresn x_j sobre $y_{h-1} = w_{hj} = cov(y_{h-1}, x_j)/s_y^2$]

2.2 $w_h / \|w_h\|$: normar w_h a 1 [w_h *vect_p inicial de* R^p]

2.3 $t_h = X_{h-1} w_h / (w'_h w_h)$ [***componente_h de X;*** w_h *vect_p inic* R^p]

2.4 $p_h = X'_{h-1} t_h / (t'_h t_h)$ [*vect-p de* R^p, *coef* x_{hj} *sobre* t_h]

2.5 $X_h = X_{h-1} - t_h p'_h$ [*residuo de X, garantizo ortogonald siguiente* p_h]

2.6 $c_h = y'_{h-1} t_h // (t'_h t_h)$ [*coef de y sobre t*]

2.7 $u_h = y_{h-1} / c_h$ [***compon_h de Y***, *regresión normaliz : 1ª bisectriz*]

2.8 $y_h = y_{h-1} - t_h c_h$ [*residuos en y*]

end h.

Las coordenadas de los vectores w_h, t_h, p_h, e c_h representan las pendientes de las rectas de mínimos cuadrados pasando por el origen y pueden entonces ser calculadas con datos faltantes. En este caso, los cálculos en t y w son de la forma:

$$t_{1i} = \frac{\sum_{\{j:\, x_{ji}\ existe\}} w_{1j}^{''}\, x_{ji}}{\sum_{\{j:\, x_{ji}\ existe\}} (w_{1j}^{''})^2}$$

con

$$w_{1j}^{''} = \frac{w_{1j}^{'}}{\sqrt{\sum_j^p (w_{1j}^{'})^2}} \quad ; \quad w_{1j}^{'} = \frac{\sum_{\{i:\, x_{ji}\ e\ y_i\ existen\}} x_{ji} y_i}{\sum_{\{i:\, x_{ji}\ e\ y_i\ existen\}} y_i^2}$$

Cuando no hay datos faltantes se puede remplazar las etapas 2.1 y 2.2 por:

$$w_h = X_{h-1}^{'} y_{h-1} / \| X_{h-1}^{'} y_{h-1} \| \quad \text{y 2.3 por} \quad t_h = X_{h-1} w_h .$$

Propiedades matemáticas de la Regresión PLS1

De estas regresiones se puede deducir las propiedades de *ortogonalidad*:

$$t_h' X_h = 0 \quad ; \quad t_h' y_h = 0$$

Se describen a continuación un conjunto de propiedades cuando no hay datos faltantes (validas también para PLS2):

a) $t_h' t_l = 0, \;\; l > h$

b) $t_h' X_l = 0, \;\; l \geq h$

c) $w_h' X_l' = 0, \;\; l \geq h$

d) $w_h' w_l = 0, \;\; l > h$

e) $w_h' p_l = 0, \;\; l > h$

f) $w_h' p_h = 1$

g) $X_h = X \prod_j^h (I - w_j p_j'), \;\; h \geq 1$

Estudio de los vectores $\boldsymbol{w_h^*}$

Los componentes t_h se definen a partir de los residuos X_{h-1} , $t_h = X_{h-1}w_h$, pero pueden también ser expresados en función de X (ver prop 1.g)

$$t_h = Xw_h^* \quad ; \quad T_h = XW_h^*$$

con $\boldsymbol{W_h^* = W_h(P_h'W_h)^{-1}}$

Se muestra que la matriz $P_h'W_h$ es triangular superior con todos sus elementos diagonales iguales a 1. Se puede también constatar que $P_h'W_h^* = I$ y que por consiguiente W_h^* es una inversa generalizada de P_h'.

Estudio de la ecuación de regresión PLS

Se puede escribir la fórmula de regresión de Y sobre las componentes $t_1, \ldots, t_h$ en función de las variables X; veamos:

$$\begin{aligned}\hat{Y} &= T_h c_h = t_1 c_1 + \ldots + t_h c_h \\ &= X W_h^* c_h \\ &= X W_h (P_h' W_h)^{-1} c_h = X b_h\end{aligned}$$

donde $\boldsymbol{b_h = W_h (P_h' W_h)^{-1} c_h}$ es el vector de los coeficientes de regresión PLS de Y sobre X utilizando h componentes.

- **El algoritmo de regresión PLS1**

1. $X_0 = X,\ y_0 = y \equiv \psi$ *componente a relacionar*
2. *para h=1,2,…,a.* *(X de rango a)*

2.1 $w_h = X'_{h-1} y_{h-1} / \|y_{h-1}\|$

2.2 $w_h / \|w_h\|$ [normar w_h a 1]

2.3 $t_h = X_{h-1} w_h / (w'_h w_h)$ [Componente PLS]

2.4 $p_h = X'_{h-1} t_h / (t'_h t_h)$

2.5 $X_h = X_{h-1} - t_h p'_h$ [*residuo de X, ortogonalidad*]

2.6 $c_h = y'_{h-1} t_h // (t'_h t_h)$

2.7 $u_h = y_{h-1} / c_h$ [***compon_h de Y***]

2.8 $y_h = y_{h-1} - t_h c_h$ [*residuos en y*]

end h.

Note que no hay iteraciones para la convergencia, debido a que solo se tiene una *y*. Las coordenadas de los vectores w_h , t_h , p_h , e c_h representan las pendientes de las rectas de mínimos cuadrados pasando por el origen y pueden entonces ser calculadas con datos faltantes.

En este caso, los cálculos en *t* y *w* son de la forma:

$$t_{1i} = \frac{\sum_{\{j:\, x_{ji}\ existe\}} w''_{1j}\, x_{ji}}{\sum_{\{j:\, x_{ji}\ existe\}} (w''_{1j})^2}$$

con $$w''_{1j} = \frac{w'_{1j}}{\sqrt{\sum_j^p (w'_{1j})^2}} \quad ; \quad w'_{1j} = \frac{\sum_{\{i:\, x_{ji}\ e\ y_i\ existen\}} x_{ji} y_i}{\sum_{\{i:\, x_{ji}\ e\ y_i\ existen\}} y_i^2}$$

Tratamiento de un ejemplo; *Aplicación PLS1 : datos de Cornell.*

En los datos de Cornell (1990), se registra el índice de octano de motor y en $n=12$ mezclas, para determinar la influencia de $p=7$ componentes, x_1= destilación directa,.., x_7 = esencia natural. Los datos presentan multicolinealidad ya que son proporciones tal que $\sum x_i = 1$. Ver Tabla 1.

	y	x1	x2	x3	x4	x5	x6	x7
1	98.7	0.00	0.23	0.00	0.00	0.00	0.74	0.03
2	97.8	0.00	0.10	0.00	0.00	0.12	0.74	0.04
3	96.6	0.00	0.00	0.00	0.10	0.12	0.74	0.04
4	92.0	0.00	0.49	0.00	0.00	0.12	0.37	0.02
5	86.6	0.00	0.00	0.00	0.62	0.12	0.18	0.08
6	91.2	0.00	0.62	0.00	0.00	0.00	0.37	0.01
7	81.9	0.17	0.27	0.10	0.38	0.00	0.00	0.08
8	83.1	0.17	0.19	0.10	0.38	0.02	0.06	0.08
9	82.4	0.17	0.21	0.10	0.38	0.00	0.06	0.08
10	83.2	0.17	0.15	0.10	0.38	0.02	0.10	0.08
11	81.4	0.21	0.36	0.12	0.25	0.00	0.00	0.06
12	88.1	0.00	0.00	0.00	0.55	0.00	0.37	0.08

Tabla 1. Datos de Cornell.

Bajo R, se tiene:

```
YX <- read.table("Cornell.txt",header=TRUE)

      pls1Cornl <- fPLS1(YX,H)   # H=3 componts

         w <- pls1Cornl[[1]]; T <- pls1Cornl[[2]]
         c <- pls1Cornl[[3]]; P <- pls1Cornl[[4]]
         b <- pls1Cornl[[5]]

  c       c1        c2        c3
[1,] 0.4820365 0.2731127 0.1030689

  b         b1          b2          b3
[1,] -0.21064817 -0.17780112 -0.13909167
[2,] -0.01781672 -0.20482512 -0.20869374
[3,] -0.21081473 -0.17674107 -0.13755531
[4,] -0.17779609 -0.22262082 -0.29316826
[5,]  0.12423073  0.03220166 -0.03843049
[6,]  0.24782555  0.40327511  0.45638984
[7,] -0.18645118 -0.12734749 -0.14338442
```

- Modelo estimado en términos de las t_h :

$$\hat{Y} = 0.4820t_1 + 0.2731t_2 + 0.1031t_3$$

- Modelo estimado en términos de las x_z (estandarizadas):

$$\hat{Y} = -0.1391x_{1z} - 0.2087x_{2z} - 0.1375x_{3z} - 0.2931x_{4z} - 0.0384x_{5z} + 0.4563x_{6z} - 0.1434x_{7z}$$

$$R^2 = 0.99527$$

La previsión.

En el ejemplo de Cornell, suponga que se busca fabricar una mezcla conducente a un índice de octanaje máximo. El modelo obtenido con los datos originales es:

$$\hat{y} = 92.676 - 9.828x_1 - 6.96x_2 - 16.67x_3 - 8.422x_4 - 4.389x_5 + 10.16x_6 - 33.53x_7$$

Para maximizar el índice de octanaje del motor y se puede construir una mezcla tomando los valores más fuertes para las componentes con los coeficientes más elevados, teniendo en cuenta las restricciones. De la regresión múltiple $y = c_1t_1 + .. + c_ht_h + y_h$ con el error $y_h \sim N(0,\sigma)$, se deduce el intervalo de confianza de μ_i y de y_i respectivamente como:

$$\hat{y}_i \pm t_{0.975,n-H-1}\hat{\sigma}\sqrt{\frac{1}{n}+\frac{t_{1i}^2}{t_1't_1}+..+\frac{t_{hi}^2}{t_h't_h}} \quad , \quad \hat{y}_i \pm t_{0.975,n-H-1}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{t_{1i}^2}{t_1't_1}+..+\frac{t_{hi}^2}{t_h't_h}}$$

En el ejemplo de los datos de Cornell, dada la mezcla nº 13,

$$x_1=0,\ x_2=0.14,\ x_3=0,\ x_4=0,\ x_5=0.12,\ x_6=0.74,\ x_7=0$$

Su índice de octanaje es $\hat{y}_{13} = 98.693$; $\hat{\sigma} = 0.7431$ y los intervalos de confianza al 95% son:

IC μ_{13} : [97.744 , 99.693] , IC y_{13} : [96.734, 100.65].

	x_2	x_3	x_4	x_5	x_6	x_7	y
x_1	0.10	0.99	0.37	-0.55	-0.80	0.60	-0.84
x_2		0.10	-0.54	-0.29	-0.19	-0.59	-0.07
x_3			0.37	-0.55	-0.80	0.61	-0.84
x_4				-0.21	-0.64	0.92	-0.71
x_5					0.46	-0.27	0.49
x_6						-0.66	0.98
x_7							-0.74

Matriz de correlaciones.

Se observa que las variables x_5 , x_2 son las menos correlacionadas.

```
fPLS1 <- function(YX)  # bajo R
{
    Xo <- scale(YX[,-1]) ; H <- qr(Xo)$rank
    Yo <- scale(YX[,1])     # == sqrt(n/(n-1))

   pXo <- ncol(Xo); nXo <- nrow(Xo); pYo <- ncol(Yo)

   WH <- matrix(0,pXo,H); TH <- matrix(0,nXo,H)
   CH <- matrix(0,pYo,H); PH <- matrix(0,pXo,H)
   BH <- matrix(0,pXo,H)

   for(hH in 1:H)  # todas las compon t en la regresion.
   {
          whi <- t(Xo)%*%Yo
           nwhi <- as.numeric(sqrt(sum(whi^2)))
       wh <- whi/nwhi      # de norma 1

       th <- Xo%*%wh/as.numeric(t(wh)%*%wh)
          s.th2 <- as.numeric(t(th)%*%th)

       ch <- t(Yo)%*%th/s.th2
            s.ch2 <- as.numeric(t(ch)%*%ch)
```

```
    ph <- t(Xo)%*%th/s.th2       # Los ph no son ortonorm.

    X1 <- Xo -th%*%t(ph); Xo <- X1
    Y1 <- Yo - th%*%t(ch); Yo <- Y1

    WH[,hH]<-wh; TH[,hH]<-th; CH[,hH]<-ch; PH[,hH]<-ph

    BH[,hH] <- WH[,1:hH]%*%(solve(t(PH[,1:hH])%*%WH[,1:hH]))%*%CH[,1:hH]
                # coefs Xjz en Y~c1t1 ; Y~c1t1+c2t2 ; ..

  }  # end hH
     # CH == coef de los TH en la regres lm(Yo~TH-1)

  r.PLS1 <- list(WH,TH,CH,PH,BH)
  return(r.PLS1)

}  # end pls1 datos cuantitativos completos
```

BIBILIOGRAFIA

- Tenenhaus M. La régression PLS théorie et pratique. Editions Technip, Paris 1998.
- Esposito Vinzi V; Chin W; Henseler J; Wang H. Handbook of partial least squares. Concepts, Methods and Applications. Springer, Berlin 2010.
- ..

Gracias por su atencion!!