

TFG análisis de datos...

Carlos Biedma Tapia / Óscar Barquero

30 de marzo de 2015



# Índice general

<b>1. Introducción</b>	<b>7</b>
<b>2. Tipos de aprendizaje</b>	<b>9</b>
<b>3. PCA: Análisis de componentes principales</b>	<b>11</b>
3.1. ¿Por qué PCA? . . . . .	12
3.2. Componentes principales . . . . .	12
<b>4. PCR: Principal Components Regression</b>	<b>17</b>
4.1. Cross Validation . . . . .	18
4.1.1. The Validation Set Approach . . . . .	19
4.1.2. Leave-One-Out Cross-Validation . . . . .	19
4.1.3. k-Fold Cross-Validation . . . . .	20

**5. PLS: Partial Least Squares** **23**

## 5.1. Libro . . . . . 23

# Índice de figuras



# Capítulo 1

## Introducción

El complejo respiratorio porcino (Este nombre habrá que cambiarla por el que realmente es) es un síndrome multifactorial relacionado o provocado por la infección de diferentes patógenos. El objetivo de este estudio se centrará en evaluar la presencia de estos principales patógenos respiratorios entre la población de jabalíes de Extremadura (España) y describir las características patológicas presentes en los pulmones de los animales infectados.

Se realizó una evaluación patológica a partir de 210 pulmones de 210 jabalíes cazados en estado salvaje. La presencia de pulmones infectados con *Mycoplasma hyopneumoniae*, *Haemophilus parasuis*, *Actinobacillus pleuropneumoniae*, *Pasteurella multocida* y *Circovirus porcina tipo 2 (PCV2)* fue evaluado a través del uso de ensayos PCR específicos.

Los jabalíes salvajes infectados con *Mycoplasma hyopneumoniae*, *Haemophilus parasuis* or *Pasteurella multocida* muestran las lesiones más severas.

*Nombre de la enfermedad* es un síndrome respiratorio que produce importantes pérdidas económicas en el área porcina en todo el mundo. Este síndrome afecta mortalmente al rededor del 2 al 10  $X_1, X_2, \dots, X_p$ , de los jabalíes.





# Capítulo 2

## Tipos de aprendizaje

En este trabajo trataremos de encontrar la relación entre una o varias variables de entrada (variables independientes, variables características o predictores) y una o varias de salida (variable dependiente o respuesta). El estudio de esta relación se puede realizar con dos objetivos principales:

- Estudiar cuales son las variables de entrada que más afectan a la salida, cuál es la relación entre la salida y cada una de las variables de entrada, saber si se pueden relacionar con una ecuación lineal o no...
- Predecir la salida en función de unos valores de entrada.

De un modo más general, si tenemos  $p$  predictors en el estudio de un problema  $X_1, X_2, \dots, X_p$ , asumimos que existe cierta relación entre dichas variables de entrada y la salida  $Y$  que puede expresarse de la forma:

$$Y = f(x) + e \tag{2.1}$$

Donde  $f$  es una función desconocida de  $X$  y  $e$  es el término de error irreducible de estimación. Dicho error aparece por cualquier causa relacionada

con el problema estadístico en el que nos encontremos, como por ejemplo el estado de los pacientes en la consulta de un médico, el tiempo atmosférico en el estudio de cultivos agrarios...

Este tipo de problemas se engloban dentro de los denominados problemas de *aprendizaje supervisado*. El objetivo del aprendizaje supervisado es el de crear una función después de haber visto una serie de ejemplos, denominados *datos de entrenamiento*. Trata de ajustar un modelo que relacione la salida con la entrada, con el objetivo de poder predecir respuestas para futuras observaciones, ya sea sobre datos numéricos o de cualquier otra clase, o entender la relación entre ambas.

Por otro lado, el otro tipo principal de problemas estadísticos son los denominados problemas de *aprendizaje no supervisado*. En este tipo de problemas, no estamos interesados en la predicción de un valor de salida, ya que no tenemos una variable de salida  $Y$ . Más bien, nuestro interés radica en descubrir relaciones interesantes sobre los datos de entrada. Como por ejemplo descubrir la mejor forma para visualizar los datos, descubrir grupos o subgrupos entre las medidas...

Los problemas no supervisados requieren una interpretación bajo un punto de vista mucho más subjetivo que los supervisados. Al no partir de un objetivo claro como podría ser la predicción de un valor de salida o ajustar el modelo mediante una función  $f$  que dependa de los datos de entrada, es difícil evaluarlo de acuerdo a unos resultados concretos, ya que no existe una respuesta absoluta o perfecta a su problema.

Los problemas no supervisados tienen una gran importancia en la aplicación de multitud de ciencias como el estudio de pacientes con riesgo de cáncer, estudios de mercado, etc; y se han escrito infinidad de artículos científicos sobre su uso.

## Capítulo 3

# PCA: Análisis de componentes principales

Análisis de componentes principales o PCA es una técnica utilizada en los problemas en los que la matriz original de los datos de entrada  $X$  está compuesta por multitud de variables características o predictors. El objetivo de esta técnica está en reducir la dimensionalidad del conjunto de datos  $X$  y así simplificar en la medida de lo posible la complejidad del problema y al mismo tiempo permitir una visualización de los datos de entrada perdiendo la menor cantidad de información posible.

A las nuevas variables de entrada se las denomina *componentes principales*, y son aquellas que logran albergar la mayor variabilidad posible de los datos de entrada, o lo que es lo mismo, la mayor varianza.

PCA es una técnica estadística no supervisada, ya que únicamente opera sobre unos datos de entrada  $X$  y no tiene ninguna salida  $Y$  asociada.

### 3.1. ¿Por qué PCA?

Cuando se recoge la información de una muestra de datos, lo más frecuente es tomar el mayor número posible de variables (variables características o predictores  $p$ ). Sin embargo, si tomamos demasiadas variables sobre un conjunto de objetos, por ejemplo 20 variables, tendremos que considerar  $\binom{20}{2} = 180$  posibles coeficientes de correlación; si son 40 variables dicho número aumenta hasta 780. Evidentemente, en este caso es difícil visualizar relaciones entre las variables.

Además, otro problema que suele surgir en la toma de medidas estadísticas es la fuerte correlación que suele existir entre muchas de estas variables. Lo normal es que estén relacionadas o que midan lo mismo bajo distintos puntos de vista. Por ejemplo, en estudios médicos, la presión sanguínea a la salida del corazón y a la salida de los pulmones están fuertemente relacionadas.

Por tanto, surge la necesidad de reducir el número de variables de entrada y así disminuir la complejidad del problema.

### 3.2. Componentes principales

Las  $p$  variables originales de la matriz  $X$  correlacionadas entre sí se pueden transformar en otro conjunto de nuevas variables incorreladas entre sí (que no tenga repetición o redundancia en la información). A estas nuevas variables las denominamos *componentes principales*.

Estas nuevas variables serán combinaciones lineales de las  $\mathbf{p}$  variables anteriores, y se construyen según el orden de importancia en cuanto a la variabilidad total que recogen. Es decir, la primera componente principal  $Z_1$  será aquella que mayor variabilidad recoja en sus datos,  $Z_2$  la segunda y así sucesivamente. Es importante resaltar el hecho de que el concepto de mayor variabilidad se relaciona con el de mayor información o varianza. Cuanto mayor sea la variabilidad de los datos (varianza) se considera que existe ma-

por información, lo cual está relacionado con el concepto de entropía.

Matemáticamente, la primera componente principal se expresa de la manera:

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \dots + \phi_{p1}X_p \quad (3.1)$$

Los componentes  $\phi_{11}, \dots, \phi_{p1}$  son los denominados *coeficientes de la primera componente principal*. Juntos, estos coeficientes construyen el vector de la primera componente principal,  $(\phi_{11}, \phi_{21}, \dots, \phi_{p1})^T$ . Este vector define la dirección en el espacio sobre la que los datos varían más. Aquí podría explicar lo de la varianza o no...

Dada una matriz  $n \times p$ , la primera componente principal se puede construir a partir de combinaciones lineales de las muestras de la matriz de la forma:

$$Z_{i1} = \phi_{11}X_{i1} + \phi_{21}X_{i2} + \dots + \phi_{p1}X_{ip} \quad (3.2)$$

Donde  $Z_{11}, \dots, Z_{n1}$  son los valores de la primera componente principal. Por tanto, una proyección de las muestras originales de la matriz sobre el vector de la primera componente principal, los valores proyectados serían  $Z_{11}, \dots, Z_{n1}$ .

Una vez calculada la primera componente principal  $Z_1$ , podemos calcular la segunda componente principal  $Z_2$ . Esta segunda componente principal se construye a partir de combinaciones lineales de las variables características de la matriz original  $X_1, \dots, X_p$  (al igual que la primera componente principal) de modo que la variable obtenida  $Z_2$ , esté incorrelada con  $Z_1$ . Como puede suponerse, los valores de esta segunda componente principal se calculan del mismo modo que los de la primera componente  $Z_1$ :

$$Z_{i2} = \phi_{12}X_{i1} + \phi_{22}X_{i2} + \dots + \phi_{p2}X_{ip} \quad (3.3)$$

Donde  $\phi_2$  compone el vector de la segunda componente principal.

Como se ha mencionado antes,  $Z_1$  y  $Z_2$  están incorreladas entre sí. Esto supone al mismo tiempo que las direcciones de los vectores  $\phi_1$  y  $\phi_2$  son perpendiculares uno respecto al otro.

Otra interpretación distinta de las componentes principales, y quizá algo más sencilla, es la siguiente:

El vector de la primera componente principal es aquel vector que está lo más cerca posible de los datos originales (teniendo en cuenta la distancia euclidiana como medida de cerca”). Aquí poner la imagen 6.15, página 234.

En consecuencia, la primera y segunda componente principal componen el plano que está lo más cerca posible a las muestras originales, las tres primeras componentes principales componen el cubo más ajustado a los datos y así sucesivamente.

Aquí poner la imagen 10.2 de la página 384

Un concepto muy importante al hablar de PCA es el de la varianza explicada. Nos interesa saber cuanta información estamos perdiendo al reducir la dimensionalidad de los datos, o lo que es lo mismo, cuanta información conseguimos captar en cada nueva componente. Recordamos que el concepto de información está íntimamente ligado con el de varianza.

La varianza explicada en un conjunto de datos cualquiera se calcula de la forma:

$$\sum_{j=1}^p Var(X_j) = \sum_{j=1}^p \frac{1}{n} \sum_{i=1}^n X_{ij}^2 \quad (3.4)$$

Y la varianza explicada para una componente principal cualquiera se calcula de la manera:

$$\frac{1}{n} \sum_{i=1}^n Z_{im}^2 = \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^p \phi_{jm} X_{ij} \right)^2 \quad (3.5)$$

Este valor nos es extremadamente útil para decidir cuantas componentes principales son requeridas para un problema estadístico concreto. Por regla general, nos interesa elegir el mínimo número de componentes posible que garanticen una varianza explicada mínima. Para ello, podemos hacer uso de una representación gráfica en la que se muestra el número de componentes frente a la varianza explicada. Sin embargo, no existe una respuesta absoluta al problema de cuantas componentes principales escoger, y dependerá del criterio escogido y la finalidad del estudio (*aprendizaje no supervisado*).

Poner imagenes de la página 388 y explicarla

Como es lógico, al calcular la varianza explicada nos daremos cuenta de que la primera componente principal es la que recoge la mayor parte de la información de los datos, y que a medida que el número de la componente aumenta, la varianza explicada por dicha componente es cada vez menor. Al final, la suma de todas las varianzas explicadas por cada una de las componentes supondrá el 100 % del total.





## Capítulo 4

# PCR: Principal Components Regression

Principal Components Regression se basa en la aplicación de PCA sobre una matriz de muestras originales para después realizar un modelo de regresión lineal sobre las nuevas dimensiones de los datos.

Recordando el capítulo anterior, cada componente principal  $Z_m$  se puede calcular como una combinación lineal de los predictores originales de la forma:

$$Z_m = \sum_{j=1}^p \phi_{jm} X_j \quad (4.1)$$

Con estas componentes principales podemos ajustar el modelo de regresión lineal:

$$y_i = \theta_0 + \sum_{m=1}^M \theta_m z_{im} + e_i, \quad (4.2)$$

Con  $i = 1, \dots, n$ . Automáticamente nos damos cuenta de que gracias al uso de PCA, hemos reducido la complejidad del problema considerablemente, pues hemos pasado de tener que estimar  $p+1$  coeficientes de regresión a  $m+1$  coeficientes nuevos, siendo  $m < p$ . La idea principal de este método es que un número pequeño de componentes principales es suficiente para recoger la mayor parte de la información de los datos originales, así como su relación con la variable de salida  $Y$ .

En la página 237 explican lo del overfitting pero no me acuerdo lo que era, tengo que buscarlo.

Al aplicar PCR sobre una matriz de datos es importante estandarizar los predictores antes de calcular las nuevas componentes principales. De este modo se garantiza que todas las variables están en la misma escala. En caso contrario, se puede obtener un efecto negativo al aplicar el modelo debido al mayor peso que se le conceden a las variables  $p$  con mayor varianza.

## 4.1. Cross Validation

Cross Validation ( $CV$ ) es una técnica de remuestreo ampliamente utilizada en la estadística moderna para validar la calidad de un modelo de regresión lineal. Se basa en ajustar el modelo múltiples veces, cada una de ellas con un conjunto de valores de entrenamiento distinto. Lógicamente, este método conlleva una gran carga computacional, pero nos permite obtener mejores resultados gracias a la información adicional que obtenemos al repetir el proceso en repetidas ocasiones con muestras distintas de la variable original.

En esta sección hablaremos de tres técnicas distintas para la aplicación de  $CV$ :

#### 4.1.1. The Validation Set Approach

Este método es el más sencillo de todos y por tanto el que menos carga computacional conlleva. La idea es simple, consiste en dividir el conjunto de muestras en dos partes iguales, una de ellas para el conjunto de entrenamiento y la otra para el conjunto de validación. Como es de esperar, el conjunto de entrenamiento se emplea para ajustar el modelo de regresión lineal, mientras que el conjunto de validación nos sirve para predecir los valores de salida de nuestro modelo ya ajustado. Con estos valores estimados calcularemos la calidad de nuestro modelo a través del  $MSE$  (en el caso de tener variables cuantitativas).

Uno de los inconvenientes que nos encontramos en este método es el grado de aleatoriedad al que está sometido. Al dividir el conjunto de muestras en dos grupos de manera totalmente aleatoria, nos damos cuenta de que el valor de  $MSE$  obtenido será distinto en función de las muestras que se encuentren en cada uno de estos dos grupos. Es decir, si realizáramos el proceso repetidas veces, cada vez con dos grupos de entrenamiento y validación distintos, obtendremos distintos valores de  $MSE$ , y por tanto no tendremos una medida absoluta de la calidad de nuestro modelo.

#### 4.1.2. Leave-One-Out Cross-Validation

Leave-one-out cross-validation ( $LOOCV$ ) consiste en una mejora del *The Validation Set Approach* Aquí poner referencia al anterior. En este caso, el conjunto de muestras también se divide en dos grupos de muestras distintos (entrenamiento y validación), pero la diferencia radica en que en este caso estos dos grupos no serán del mismo tamaño. Por el contrario, el conjunto de validación estará constituido por una única muestra, mientras que todas las restantes formarán parte del grupo de entrenamiento.

Este proceso se repite  $n$  veces, de tal forma que en cada iteración la muestra de validación será distinta. De este modo, al final del proceso todas las muestras originales habrán formado el conjunto de validación y se habrán hecho  $n$  estimaciones de la variable de salida. Además, en cada una de estas

iteraciones de calculará el  $MSE$  a partir de la muestra de validación  $i$ :

$$MSE_i = (y_i - \hat{y}_i)^2 \quad (4.3)$$

Y el coeficiente de  $CV$  final será igual al promedio de todos los  $MSE$  calculados:

$$CV = \frac{1}{n} \sum_{i=1}^n MSE_i \quad (4.4)$$

La principal ventaja de este modelo es la reducción considerable del error de estimación. Al estar usando un número mucho mayor de muestras para el ajuste del modelo, es de esperar que los resultados sean mucho mejores y por tanto la muestra estimada sea mucho más parecida a la real en comparación con el Validation Set Approach.

Además, este proceso ya no está sujeto a ningún grado de aleatoriedad, pues al aplicar el algoritmo  $n$  veces y usar cada muestra como muestra de validación, el resultado final será siempre el mismo para un conjunto dado de observaciones.

### 4.1.3. k-Fold Cross-Validation

k-Fold Cross-Validation es un proceso intermedio entre Validation Set Approach y LOOCV. En este caso, el conjunto de muestra se divide en  $k$  grupos de igual tamaño. Una vez tenemos estos grupos creados, el proceso es muy similar al del LOOCV. Se itera el proceso  $k$  veces, considerándose en cada una de estas iteraciones un grupo distinto como muestras de validación del modelo, quedando el resto  $(k-1)$  de los grupos como muestras de entrenamiento. De este modo, el coeficiente de  $CV$  final será:

$$CV = \frac{1}{k} \sum_{i=1}^k MSE_i \quad (4.5)$$

Lógicamente, este método no obtendrá tan buenos resultados en el ajuste del modelo como LOOCV, pero por otra parte, será mucho menos costoso computacionalmente, ya que el número de iteraciones que se realizan es mucho menor.



# Capítulo 5

## PLS: Partial Least Squares

### 5.1. Libro

Como hemos estudiado anteriormente, PCR realiza una reducción de los  $p$  predictores originales de una manera no supervisada, es decir, no hace uso de la variable de salida  $Y$ . Esto significa que aunque PCR nos proporciona una reducción de la dimensionalidad del problema para la visualización y estudio de los datos de entrada, no tiene por qué ofrecernos la mejor combinación para la predicción del valor de salida  $Y$ .

PLS se presenta como una variación de PCR en la que las nuevas variables características  $Z_1, Z_2, \dots, Z_m$  se siguen obteniendo a partir de combinaciones lineales de las variables de entrada pero en este caso se obtienen de una manera supervisada, es decir, haciendo uso de la variable de salida  $Y$ .

¿Cómo funciona PLS?

Al igual que en el caso de PCR, PLS parte de la siguiente aproximación:

$$Y = XB + residual \quad (5.1)$$

O lo que es lo mismo:

$$Y = b_0 + b_1x_1 + b_2x_2 + \dots + b_Nx_N \quad (5.2)$$

Estimamos  $Y$  a partir de combinaciones lineales de la matriz de observaciones original  $X$  y el conjunto de coeficientes  $B$ .

Por notación algebraica sabemos que podemos despejar  $B$  de la siguiente manera:

$$\begin{aligned} X^{-1}XB &= X^{-1}Y \\ IB &= X^{-1}Y \\ B &= X^{-1}Y \end{aligned} \quad (5.3)$$

Una vez despejada la matriz de coeficientes  $B$ , aplicando la propiedad de la matriz de...(preguntar a Óscar como se llamaba):

$$\begin{aligned} (X^T X)B &= (X^T X)X^{-1}Y \\ (X^T X)B &= X^T IY \\ (X^T X)B &= X^T Y \\ (X^T X)^{-1}(X^T X)B &= (X^T X)^{-1}X^T Y \\ IB &= (X^T X)^{-1}X^T Y \end{aligned} \quad (5.4)$$

Finalmente obtenemos:

$$B = (X^T X)^{-1}X^T Y \quad (5.5)$$



Esta última ecuación nos proporciona una sencilla solución para obtener la matriz de coeficientes  $B$ . Sin embargo, en ocasiones esta expresión no nos proporciona buenos resultados debido al comportamiento de las observaciones de  $X$  (explicar lo de la alta correlación entre variables...).

Por ello, se plantea una solución alternativa en la que no estimamos  $Y$  a partir de  $X$  directamente, sino que realizamos un paso intermedio. La estimación de  $Y$  se realiza a través de una nueva matriz de coeficientes  $T$  con unas condiciones más favorables que las que tenía  $X$ . Esto es:

$$Y = TC \quad (5.6)$$

O lo que es lo mismo:

$$Y = c_0 + c_1 t_1 + c_2 t_2 + \dots + c_N t_N \quad (5.7)$$

A su vez, cada componente de  $T$  ( $t_1, t_2, \dots, t_k$ ) está compuesto por una combinación lineal de  $X$ :

$$t_k = w_k^0 + w_k^1 X_1 + w_k^2 X_2 + \dots + w_k^N X_N \quad (5.8)$$

Por tanto, al final nos damos cuenta como  $Y$  se estima a partir de una combinación lineal que a su vez está formada por otra combinación lineal en la que todas las componentes de  $X$  están implicadas:

$$Y = c_0 + c_1(Xw_1) + c_2(Xw_2) + \dots + c_N(Xw_N) \quad (5.9)$$

(No entiendo muy bien en esta última fórmula si hace uso de  $Y$  para estimar, la última transparencia del 2015-1 tb la quiero preguntar).



# Bibliografía

- [1] Francisco Javier Ceballos: Enciclopedia de Microsoft Visual Basic. Editorial Ra-Ma. Madrid, 1999.
- [2] Dieter Staas: PHP 5 Espresso!. Franzis Verlag GmbH. Poing, 2004.
- [3] Ralph Pfeiffer. Diplomarbeit: Planung und Erstellung einer im Höhenwinkel nachführbaren Photovoltaikanlage sowie Reali
- [4] Lars Ortlieb, Diplomarbeit: Vernetzung alternativer Energiesysteme unter Verwendung moderner Kommunikationstechnik, Fachhochs
- [5] Mike Schumacher. Datentechnische Erfassung einer nachgeführten Photovoltaikanlage in einem alternativen Energieverbund mittels modern
- [6] TAC Xenta 511 Engineering Manual
- [7] TAC Web-Seite: <http://www.tac-global.com>
- [8] PHP Web-Seite: <http://www.php.net>