TFG análisis de datos...

Carlos Biedma Tapia / Óscar Barquero

1 de abril de 2015

Índice general

1. Introducción **7**

2. Materiales y métodos **9**

3. Resultados **13**

4. Discussion **17**

5. Tipos de aprendizaje **19**

6. PCA: Análisis de componentes principales **21**

6.1. ¿Por qué PCA? 22

6.2. Componentes principales 22

7. PCR: Principal Components Regression **27**

7.1. Cross Validation 28

7.1.1. The Validation Set Approach 29

7.1.2. Leave-One-Out Cross-Validation 29

7.1.3. k-Fold Cross-Validation 30

8. PLS: Partial Least Squares **33**

8.1. Libro 33

Índice de figuras

Capítulo 1

Introducción

El complejo respiratorio porcino (Este nombre habrá que cambiarla por el que realmente es) es un sindrome 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 sindrome respiratorio que produce importantes pérdidas económicas en el área porcina en todo el mundo. Este síndrome afecta mortalmente alrededor del 2 al 10 % de los jabalíes.

En jabalíes domésticos, los animales afectados por este síndrome (poner el nombre) normalmente muestran tos, fiebre, disnea, dismminución en la ingesta de alimento y retraso del crecimiento. Las lesiones pulmonares de los animales infectados se encuentran principalmente en las partes craneoventrales del pulmón,donde se puede observar la consolidación, la decoloración y la atelectasia, aunque estas características pueden variar dependiendo de los patógenos involucrados.

La gravedad de la presentación clínica de esta enfermedad en jabalíes depende de las interacciones entre los agentes patógenos respiratorios involucrados, los factores ambientales, incluyendo los sistemas de gestión y los factores individuales como la edad o el estado inmunológico.

Diferentes agentes patógenos infecciosos pueden estar implicados en el desarrollo de la enfermedad. Los patógenos respiratorios se dividen comúnmente en patógenos primarios, que son capaces de inducir lesiones graves en el tracto respiratorio con su propio efecto; y patógenos oportunistas o secundarios, que por lo general inducen lesiones en el tracto respiratorio en combinación con otros agentes patógenos o factores.

Los principales patógenos respiratorios primarios involucrados en CRP son virus como el síndrome porcino reproductivo y respiratorio (PRRSV), virus de la gripe porcina (SIV), el virus de la enfermedad de Aujeszky (ADV) y circovirus 2 tipo porcino (PCV2); o bacterias como la Mycoplasma hyopneumoniae o Actinobacillus pleuropneumoniae. Por otro lado, Pasteurella multocida o Haemophilus parasuis se encuentran entre los patógenos respiratorios oportunistas más comunes.

El objetivo de este estudio será evaluar la presencia de los principales patógenos respiratorios en cerdos jabalíes del medio oeste de España y describir las características patológicas de los pulmones de los animales naturalmente infectados con patógenos Prdc.

Capítulo 2

Materiales y métodos

Área de estudio y animales

Este estudio se llevó a cabo en un total de 210 jabalíes cazados de 20 fincas de caza en el medio oeste de España. El área de estudio tiene características particulares en cuanto a la ecología y el clima. En pocas palabras, la precipitación media anual alcanza los 623 mm y se concentra en los meses de noviembre a abril. La temperatura promedio anual media es de 17.7ºC, siendo enero el más frío y julio el mes más caluroso del año. La vegetación es típica de bosque mediterráneo, caracterizado por la abundancia de Quercus ilex y Quercus suber. En todas las fincas de caza incluidas en este trabajo, los jabalíes salvajes comporten habitat con el ciervo rojo (Cervus elaphus) y en algunos casos con el gamo (Dama dama) o el corzo (Capreolus capreolus).

Los jabalí incluidos en este trabajo fueron cazados entre octubre de 2011 y febrero de 2013. El sexo y la edad de estos animales se determinaron sobre la base de la observación de sus órganos sexuales y el patrón de erupción dentición respectivamente (no sé que es esto). Los animales fueron divididos en cuatro grupos diferentes de acuerdo a sus edades: lechones (menos de seis meses), jóvenes (seis meses-un año), añales (un año y dos años) y adultos (más de dos años).

La necropsia de todos los animales se realizó en el campo con un detallada inspección macroscópica de los pulmones. Se tomó un pedazo del lóbulo craneal derecho de pulmón y se sumergió un 10 % en formalina tamponada inmediatamente después del examen. A continuación, una pieza adyacente adicional del mismo lóbulo se recogió en bolsas de almacenamiento, manteniéndose frías para el transporte y, en menos de seis horas, congeladas a -20ºC para su posterior análisis en el laboratorio. Finalmente, las muestras de sangre fueron recolectadas desde el corazón o desde la cavidad torácica. Estas muestras se centrifugaron en el laboratorio a 3.000 rpm durante 10 minutos para separar el suero antes de ser almacenadas a -20ºC hasta su utilización.

Estudio histopatológico

Todos los animales estudiados fueron cazados en estado salvaje. Por esta razón, los animales mostraron hemorragias pulmonares frecuentes que podrían conducir a una interpretación errónea en un sistema de puntuación lesional bruto de estos pulmones. Para evitar este hecho, el estudio patológico se basa sólo en el estudio de los parámetros histopatológicos específicos descritos anteriormente, fácilmente distinguibles de los artefactos causados por el tipo de muerte (disparo).

Muestras de tejido pulmonar previamente sumergidos en formalina tamponada fueron procesados siguiendo los procedimientos estándar y se tiñeron de forma rutinaria con hematoxilina y eosina para el estudio histopatológico. Las secciones de pulmón fueron sometidas a un examen a ciegas (sin conocer los resultados de los estudios microbiológicos).

Cinco (a lo mejor son las de 5) parámetros histopatológicos se puntuaron en una escala de 0 a 6 (0: normal; 1: leve multifocal; 2: difusa leve; 3: multifocal moderado; 4: difusa moderada; 5: grave multifocal; 6: difusa grave) siendo la severidad una puntuación subjetiva, pero siempre determinado por los mismos patólogos. De hecho, las lesiones se evaluaron de forma independiente por dos patólogos veterinarios. En los casos de discrepancia (menos del 1 % de los valores eran diferentes), los veterinarios acordaron el valor final. Estos parámetros fueron: presencia de infiltración septal alveolar con células inflamatorias, la cantidad de exudados en los alvéolos y las vías respiratorias, hiperplasia linfoide peribronquial, la cantidad de células inflamatorias en la

lámina propia de los bronquios y los bronquiolos y la presencia de necrosis en las células epiteliales de los bronquios y los bronquiolos. La presencia de lesiones de tuberculosis como granulomatosas y nematodos pulmonares también se evaluó en las secciones de pulmones estudiados, ya que son frecuentes en los animales de la zona de estudio.

Patógenos respiratorios detectados por PCR

Se utilizó un conjunto de técnicas de PCR específicas para detectar la presencia de algunos de los patógenos respiratorios más comunes en los pulmones de jabalí como se ha descrito en estudios similares (Reiner et al 2010;. Reiner et al 2009;. Sibila et al. 2010). El ADN del tejido pulmonar almacenado a -20ºC fue extraído por medio de un kit QIAamp DNA Mini comercial (Qiagen Ltd, Crawley, West Sussex, RH10 9NQ, Reino Unido) siguiendo las recomendaciones del fabricante. Después, PCR específicos para M. hyopneumoniae se llevaron a cabo utilizando ADN extraído anteriormente como plantilla.

Ensayos de inmunohistoquímica (IHC) e hibridación in situ (ISH)

La presencia de antígenos como PRRSV y SIV (poner nombres completos) y su distribución en todo el área pulmonar se evaluó mediante técnicas de inmunohistoquímicas (IHC). La técnica de peroxidasa-avidina biotina se utilizó con los correspondientes anticuerpos específicos para detectar PRRSV y SIV en secciones de pulmón. Por razones de logística y financiación los ensayos IHC para detectar SIV se llevaron a cabo en 127 jabalíes, mientras que la presencia del virus del PRRS se evaluó en 70 animales.

Además, para evaluar la distribución de M. hyopneumoniae y PCV2 en tejidos pulmonares infectados, se llevaron a cabo diferentes técnicas de IHC e ISH en 26 jabalíes. Estos animales fueron elegidos entre los que habían sido evaluados para el estudio de la presencia de SIV y antígenos de PRRSV. Antígenos de PCV2 fueron detectados gracias a IHC y técnicas ISH, llevados a cabo como se describió previamente. Por otro lado, se evaluó la distribución de M. hyopneumoniae en el parénquima pulmonar usando un ensayo de ISH fluorescente.

Análisis estadístico

Los valores medios obtenidos para cada uno de los parámetros histopatológicos estudiados fueron comparados entre los animales infectados y no infectados para cada patógeno. Se empleo un método no paramétrico (Mann-Whitney U-test) para comparar las medias anotados entre los grupos. Para evitar la posible influencia de los nematodos pulmonares en los resultados obtenidos, los animales que mostraron la presencia de parásitos en el examen histopatológico fueron excluidos para el análisis estadístico realizado para correlacionar la presencia de patógenos CRPD y lesiones histopatológicas.

Además, las puntuaciones medias obtenidas para cada parámetro histológico fueron comparados entre los grupos de edad, utilizando la prueba de Kruskal-Wallis. Todos los cálculos se realizaron con el showtware estadístico $SPSS15(SPSSInc., Chicago, Illinois, 60606, EE.UU.)$.

Capítulo 3

Resultados

Resultados de la evaluación de patógenos

Al estudiar los patógenos respiratorios estudiados en diferentes grupos de edad se observó se que un porcentaje moderado de los animales estaban infectados por M. hyopneumoniae (24,80 %), PCV2 (19,5 %) y PRRSV (14,3 %), mientras que el resto de microorganismos estudiados fueron encontrados solamente en un pequeño porcentaje de animales (¡10 %). Las prevalencias fueron muy diferentes entre los grupos de edad en algunos casos. A modo de ejemplo, las prevalencias obtenidas por H. parasuis y P. multocida fueron significativamente mayores en los lechones (p ¡0,001).

Hispatología

Se encontraron lesiones pulmonares histológicos en un 83,34 % de los animales estudiados. Cada parámetro histopatológico estudiado se encontró en un porcentaje diferente en cada animal y mostraba diferente gravedad. Por ejemplo, mientras que se observó la presencia de hiperplasia linfoide peribronquial en un 64,29 % de los animales con una puntuación media de 1,75, la presencia de necrosis en las células epiteliales de los bronquios y los bronquiolos se detectó en sólo un 2,38 % de los animales.

Las lesiones histopatológicas observadas fueron influenciadas por la edad de los animales. Los lechones mostraron una puntuación significativamente mayor en la cantidad de exudados en las vías respiratorias y los alvéolos (p¡0,001), en comparación con los juveniles y los primales, que presentaron una mayor puntuación de la hiperplasia linfoide peribronquial (p = 0,01). Se detectaron lesiones granulomatosas en 42 animales, mientras que se observaron nemátodos pulmonares en 17 animales.

Correlación entre patógenos respiratorios y lesiones histológicas

Algunos de los parámetros evaluados fueron más graves en los animales infectados con M. hyopneumoniae, P. multocida o H. parasuis. Además, se detectó una mayor infiltración significativa de células inflamatorias en los septos alveolares en los animales infectados por H. parasuis (U de Mann-Whitney = 441,5, p = 0,017). Además, las secciones de pulmón que muestran nematodos tenían una cantidad significativamente mayor de los alvéolos y exudados de las vías respiratorias (U de Mann-Whitney = 884,5, p = 0,018), hiperplasia linfoide (U de Mann-Whitney = 712,5, p = 0,016) y marcadas inflamaciones de los bronquios y los bronquiolos (U de Mann-Whitney = 327, p ¡0,001).

La inmunohistoquímica e hibridación in situ

Se detectaron antígenos específicos de PCV2, SIV, PRRSV y M. hyopneumoniae en secciones de pulmón evaluados mediante técnicas de IHC e ISH. Se detectaron antígenos de PCV2 en 14 animales usando IHC en los macrófagos localizados en los tejidos linfoidales asociados a bronquios (BALT), aunque algunos macrófagos alveolares también mostraron la presencia de antígenos de PCV2. Células infectadas con PCV2 también se detectaron usando un ensayo de ISH en ocho de los 14 animales que resultaron positivos a los métodos IHC que muestran una distribución similar. Las células immunopositivas contra antígenos SIV detectadas en animales infectados fueron escasas y sólo unas pocas células epiteliales de los bronquios y los bronquiolos y los macrófagos alveolares mostraron una tinción positiva. Los animales infectados por PRRSV mostraron pocos macrófagos immunopositivos ubicados en BALT y septos alveolares, formando ocasionalmente grupos de células. Se detectó ADN de M. hyopneumoniae en cuatro animales usando el ensayo de ISH, principalmente como bacterias marcadas unidas al borde apical de las

células epiteliales bronquiales y bronquiolares.

Cinco de los 26 animales que fueron probados para todos los patógenos usando técnicas de detección in situ, mostraron células positivas para más de uno de estos patógenos. En general, los animales infectados con virus (sólo en la infección simple o mixta) o infectadas solo con M. hyopneumoniae, mostraron leve infiltración de células inflamatorias en los septos alveolares y linfoide leve hiperplasia. Sin embargo, los dos animales que presentaron infecciones mixtas con virus y M. hyopneumoniae mostraron una bronconeumonía grave.

Capítulo 4

Discussion

Los resultados obtenidos en este trabajo muestran que el jabalí puede ser infectado por los principales patógenos respiratorios involucrados en PRDC. La presencia de lesiones pulmonares histopatológicos estaba ampliamente extendida, encontrando lesiones en un 83,34 % del jabalí estudiado. Sin embargo, las puntuaciones medias obtenidas para cada parámetro histológico mostró que la gravedad de las lesiones fue a menudo leve, aunque también se registraron casos individuales con lesiones pulmonares graves.

Las lesiones neumónicas inducidas por los patógenos respiratorios evaluados fueron variables. Los jabalíes infectados con patógenos como M. hyopneumoniae, H. parasuis o P. multocida mostraron lesiones pulmonares más graves que los animales no infectados. Para otra parte, las infecciones individuales con A. pleuropneumoniae, PCV2 o ADV no estaban relacionadas con lesiones pulmonares. Los animales infectados con M. hyopneumoniae mostraron una cantidad significativamente más alta de las vías respiratorias y los exudados inflamatorios alveolares. Estos resultados están de acuerdo con un reciente trabajo en el que las infecciones por M. hyopneumoniae se han correlacionado con la presencia de lesiones neumónicas y confirman que la presencia de este patógeno también puede afectar a un variedad de parámetros histopatológicos.

Asimismo, las lesiones encontradas en animales infectados con H. parasuis

fueron similares a lo descrito recientemente en un jabalí joven salvaje infectado con este patógeno, que muestra la bronconeumonía grave asociada con la neumonía intersticial. La prevalencia la infección de H. parasuis encontrado en este estudio fue baja (4,7 %), pero es interesante que la mayoría de los animales infectados eran lechones (39,1 % de los lechones resultó infectado). Estos resultados sugieren que la infección de H. parasuis puede conducir a resultados patológicos principalmente en animales jóvenes que, además, son más susceptibles a los patógenos respiratorios.

La detección de células inmunopositivas contra SIV y PRRSV en el parénquima pulmonar no pareció influir en la aparición de lesiones pulmonares de los jabalíes. En la mayoría de los animales infectados por estos virus, se detectó reacción positiva sólo en algunas células, principalmente macrófagos, pero también células epiteliales de los bronquios y los bronquiolos en el caso de SIV. Este hecho podría significar que los animales infectados recibieron dosis bajas de infección o estaban en proceso de eliminar el virus, lo que podría estar relacionado con la falta de lesiones pulmonares detectadas en los animales infectados. Se observaron antígenos de PCV2 y M. hyopneumoniae en las células diana típicos en todo el parénquima pulmonar. En cuanto a la detección de PCV2 in situ, el ensayo de IHC (14 animales positivos) resultó más sensible que el ensayo ISH (ocho animales positivos), como se ha informado previamente.

Se detectaron co-infecciones con algunos de los patógenos respiratorios diagnosticados in situ en varias muestras de pulmón. Las co-infecciones respiratorias pueden llevar a resultados patológicos más graves. De hecho, se ha sugerido que las interacciones entre bacterias tales como M. hyopneumoniae, y los virus respiratorios pueden mejorar el desarrollo de lesiones pulmonares en cerdos domésticos (Brockmeier et al 2002;. Harms et al 2002;. Kim y Chae 2004; Thacker 2001). Los resultados obtenidos en este trabajo parecen estar de acuerdo con esta hipótesis. Los jabalíes infectados con uno o más virus sólo mostraron una neumonía intersticial leve y la hiperplasia linfoide, mientras que dos animales que sufren infecciones concurrentes de virus y M. hyopneumoniae muestra lesiones más graves.

Además, otros patógenos que no participan con frecuencia en el desarrollo PRDC también fueron considerados en el estudio. Las lesiones asociadas a los nematodos pulmonares se registraron en el jabalí estudiado. Los animales infectados por los nematodos pulmonares mostraron grave hiperplasia

linfoide y una bronquitis severa y fibrosis, como se ha descrito previamente en cerdos domésticos. Como la tuberculosis no se incluyó en los objetivos de este trabajo, la presencia de Mycobacterium spp. no fue evaluado de forma sistemática en los jabalíes que presentan lesiones granulomatosas. Sin embargo, varios estudios han mostrado recientemente una alta concordancia entre las lesiones microscópicas similar a la tuberculosis.

Como una enfermedad multifactorial, signos clínicos presentados en los cerdos domésticos afectados por agentes patógenos PRDC pueden aumentar por diversos factores, como la edad o las prácticas de gestión. En este trabajo, las lesiones histológicas más graves se registraron en los pulmones que pertenecen a los lechones (exudados broncoalveolar), juveniles, lo que sugiere que las lesiones neumónicas pueden ser más graves en los animales jóvenes.

Además, la gravedad de los resultados clínicos producidos por patógenos respiratorios porcinos puede ser mayor en granjas con alta densidad de animales y las prácticas de manejo más intensivo. En la actualidad, las prácticas de gestión, tales como la esgrima o el suministro de alimentos adicionales son comunes en la cría de jabalí, que conduce a densidades artificiales más altas de los animales. El número de granjas de jabalíes en la que se aplican estas medidas, está aumentando en todo el mundo para satisfacer la gran demanda de productos de jabalí en la industria cárnica y en el negocio de los juegos. El impacto de patógenos PRDC podría ser mayor en estas granjas "manejo intensivo"que conducen a los resultados clínicos más graves y producen graves pérdidas económicas.

En conclusión, los resultados obtenidos en este trabajo muestran que las infecciones con patógenos respiratorios porcinos como M. hyopneumoniae, P. multocida o H. parasuis, pueden producir lesiones pulmonares en los jabalíes. La presencia de lesiones pulmonares más graves en los pocos animales coinfectados con virus y M. hyopneumoniae sugiere que co-infecciones con patógenos respiratorios pueden aumentar la gravedad de la patología respiratoria. Sin embargo, más investigación, incluyendo un mayor número de animales y teniendo en cuenta otros factores que pueden influir en el desarrollo PRDC, como por ejemplo, la densidad de animales, podrían ser necesarias para confirmarlo.

Capítulo 5

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 $X_1, X_2, ..., X_p$, asumimos que existe cierta relación entre dichas variables de entrada y la salida Y que puede expresarse de la forma:

$$Y = f(x) + e \tag{5.1}$$

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 obervaciones, 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 mediente 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 6

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.

6.1. ¿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.

6.2. 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 **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-

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

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

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + ... + \phi_{p1}X_p \tag{6.1}$$

Los componentes $\phi_{11}, ..., \phi_{p1}$ son los denominados *coeficientes de la primera componente principal*. Juntos, estos coeficientes construyen el vector de la primera componente princial, $(\phi_{11}, \phi_{21}, ..., \phi_{p1})^T$. 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 x p, la primera componente principal se puede construir a partir de combinaciones lineales de las muestras de la matriz de la forma:

$$Z_{i1} = \phi_{11}X_{i1} + \phi_{21}X_{i2} + ... + \phi_{p1}X_{ip} \tag{6.2}$$

Donde $Z_{11}, ..., Z_{n1}$ 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 $Z_{11}, ..., Z_{n1}$.

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 $X_1, ..., X_p$ (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:

$$Z_{i2} = \phi_{12}X_{i1} + \phi_{22}X_{i2} + ... + \phi_{p2}X_{ip} \tag{6.3}$$

Donde ϕ_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 ϕ_1 y ϕ_2 son perpendiculares uno respecto al otro.

Otra intrepretación distinta de las componentes principales, y aquizá 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 cuanta la distancia euclidiana como medida de çerca"). 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:

$$\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} X_{ij}^2 \tag{6.4}$$

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

$$\frac{1}{n}\sum_{i=1}^{n} Z_{im}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\phi_{jm}X_{ij}\right)^2 \tag{6.5}$$

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 cuanta 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 7

PCR: Principal Components Regression

Principal Components Regression se basa en la aplicación de PCA sobre una matriz de muestras originales para despues 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:

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j \tag{7.1}$$

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

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + e_i, \tag{7.2}$$

Con $i = 1, ..., n$. 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 vaviables p con mayor varianza.

7.1. 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:

7.1.1. 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 cuantitaticas).

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.

7.1.2. 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

iteraciónes de calculará el *MSE* a partir de la muestra de validación i:

$$MSE_i = (y_i - \hat{y}_i)^2 \tag{7.3}$$

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

$$CV = \frac{1}{n}\sum_{i=1}^{n} MSE_i \tag{7.4}$$

La principal ventaja de este modelo es la reducción considerabla 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.

7.1.3. 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, quedándo el resto *(k-1)* de los grupos como muestras de entrenamiento. De este modo, el coeficiente de CV final será:

$$CV = \frac{1}{k} \sum_{i=1}^{k} MSE_i \tag{7.5}$$

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 8

PLS: Partial Least Squares

8.1. Libro

Como hemos estudiado anteriormente, PCR realiza una reducción de los p predictors 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 $Z_1, Z_2, ..., Z_m$ 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 las variable de salida Y.

¿Cómo funciona PLS?

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

$$Y = XB + residual \tag{8.1}$$

O lo que es lo mismo:

$$Y = b_0 + b_1x_1 + b_2x_2 + ... + b_Nx_N \tag{8.2}$$

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:

$$\begin{aligned} X^{-1}XB &= X^{-1}Y \\ IB &= X^{-1}Y \\ B &= X^{-1}Y \end{aligned} \tag{8.3}$$

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

$$\begin{aligned} (X^TX)B &= (X^TX)X^{-1}Y \\ (X^TX)B &= X^TIY \\ (X^TX)B &= X^TY \\ (X^TX)^{-1}(X^TX)B &= (X^TX)^{-1}X^TY \\ IB &= (X^TX)^{-1}X^TY \end{aligned} \tag{8.4}$$

Finalmente obtenemos:

$$B = (X^TX)^{-1}X^TY \tag{8.5}$$

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:

$$Y = TC \tag{8.6}$$

O lo que es lo mismo:

$$Y = c_0 + c_1 t_1 + c_2 t_2 + ... + c_N t_N \tag{8.7}$$

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

$$\begin{aligned} t_k &= X w_k \\ t_k &= w_k^0 + w_k^1 X_1 + w_k^2 X_2 + ... + w_k^N X_N \end{aligned} \tag{8.8}$$

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:

$$Y = c_0 + c_1(X w_1) + c_2(X w_2) + ... + c_N(X w_N) \tag{8.9}$$

(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`