Taller de Análisis de datos - Problema 1

Jésica Charaf e Ignacio Spiousas

6 de noviembre de 2023

Problema 1-4

Los siguientes datos corresponden a un trabajo para determinar la composición de un conjunto de vasijas de vidrio de un yacimiento arqueológico. Como el análisis espectrométrico es más barato que el análisis químico, se procuró calibrar el primero para que reemplace al segundo. Con este objetivo se tomó una muestra de $n = 180$ vasijas, a las que se realizó una espectrometria de rayos X sobre 1920 frecuencias, y también un análisis de laboratorio para determinar el contenido de 13 compuestos químicos, a saber:

$$Na_2O,\ MgO,\ Al_2O_3,\ SiO_2,\ P_2O_5,\ SO_3,\ Cl,\ K_2O,\ CaO,\ MnO,\ Fe_2O_3,\ BaO\ y\ PbO$$

Cada fila del archivo `Vessel_X` es el espectro de una vasija, limitado a las frecuencias 100 a 400, pues las demás tienen valores casi nulos. O sea, para cada $i = 1, ..., n, x(i, j(j = 1, .., 301))$ es la energía correspondiente a la frecuencia j (en realidad la frecuencia es j+99, pero podemos dejar eso de lado).

Cada fila del archivo `Vessel_Y` tiene los contenidos de los 13 compuestos en esa vasija. Vamos a comparar distintos métodos para predecir el compuesto 4 (P_2O_5).

Para familiarizarse con los datos, grafique en función de la frecuencia las medias y varianzas de X, y también algunos espectros (o sea, $x(i, j)$ en función de j para algunos i). Aplique los métodos que le parecen adecuados para este problema, y encuentre el que muestra menor error de predicción.

Para el estimador que mejor funciona:

- Grafique los coeficientes (pendientes) en función de la frecuencia.
- Haga el clásico gráfico de residuos vs. ajustados.
- Si ve algo llamativo (outliers, residuos con estructura) tome las medidas correctivas que le parezcan adecuadas.

Resolución

Análisis exploratorio

Lo primero que vamos a hacer es a graficar el contenido de `Vessel_X.txt`, es decir, la energía por banda de frecuencia de cada una de las 150 vasijas. Esto puede verse en líneas continuas de colores en la Figura 1 junto con el promedio en línea sólida negra. En la figura pareciera indicarse que las diferencias entre vasijas ocurren a determinadas frecuencias (en las que la amplitud es distinta de cero y hay más diferencia entre las mediciones individuales) y, por lo tanto, resulta esperable que la información contenida en esas bandas de frecuencias sea la que más aporte a la determinación del contenido de P_2O_5 (aunque bajo condiciones particulares podría no ser el caso).

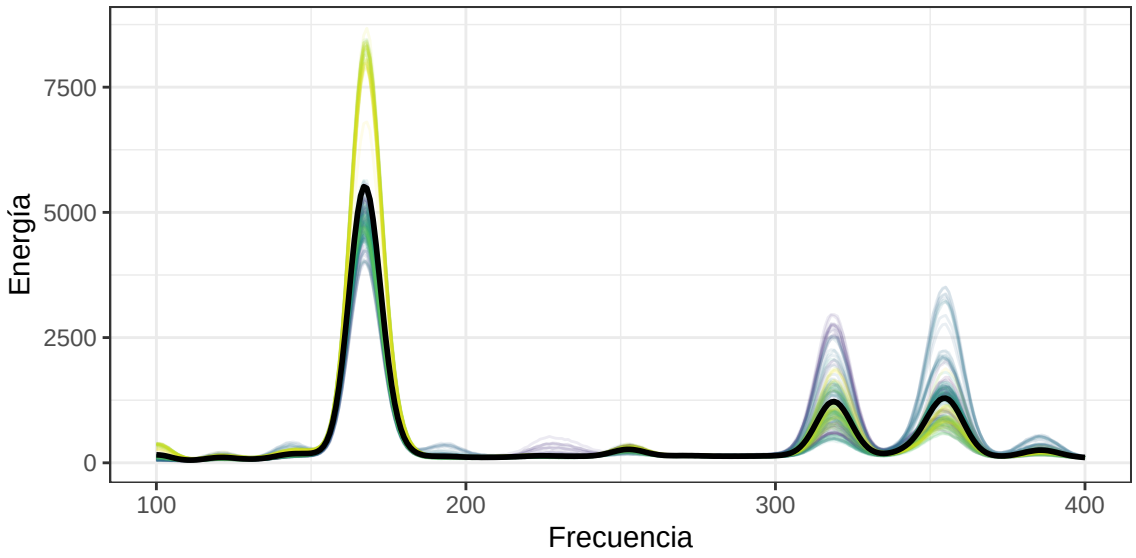

Figure 1: Energía en función de las frecuencias. Cada línea de color representa una vasija mientras que la línea negra representa al promedio de las 150 vasijas.

Para explorar esta ídea un poco más allá podemos ver en la Figura 2 el error estándar de la media en función de la banda de frecuencia. En esta figura vemos cuantificada la intuición que generamos en la Figura 1 de que, efectivamente la variabilidad en las mediciones se concentra en unas pocas bandas de frecuencia.

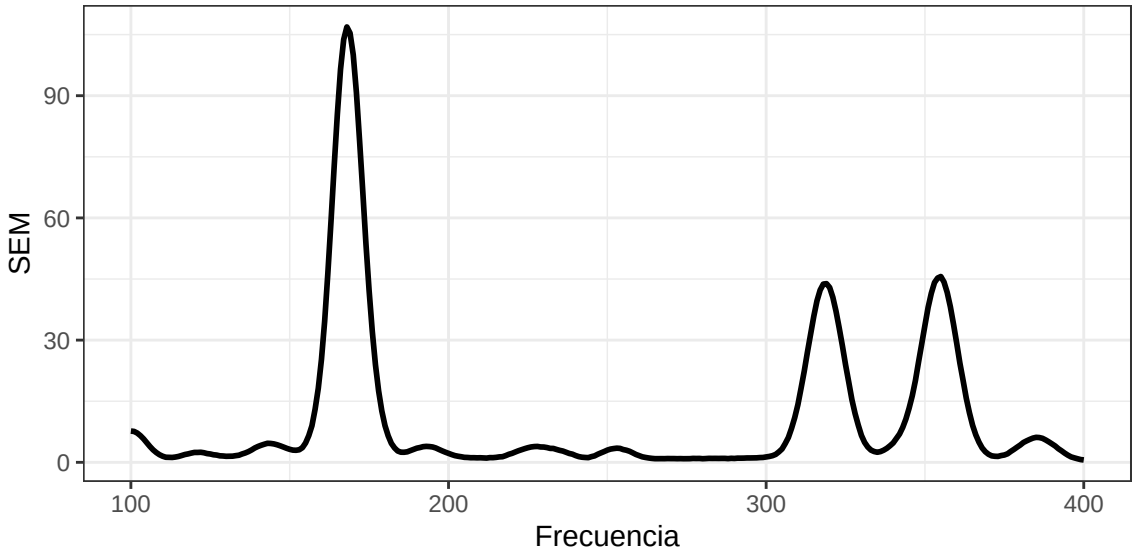

Figure 2: Error estándar en función de las frecuencias.

Luego, lo que queremos ver es como se distribuye la cantidad de P_20_5 en las muestras, para ver si esto tiene algún patrón. En la Figura 3 podemos ver el histograma y la densidad estimada para esta magnitud. En la misma se ve que no pareciera haber valores atípicos y que la distribución es unimodal y con una cola pesada a la izquierda (hacia valores más bajos). Esta asimetría podría llegar a influir en el supuesto no normalidad de los residuos del modelos a ajustar, más adelante lo evaluaremos.

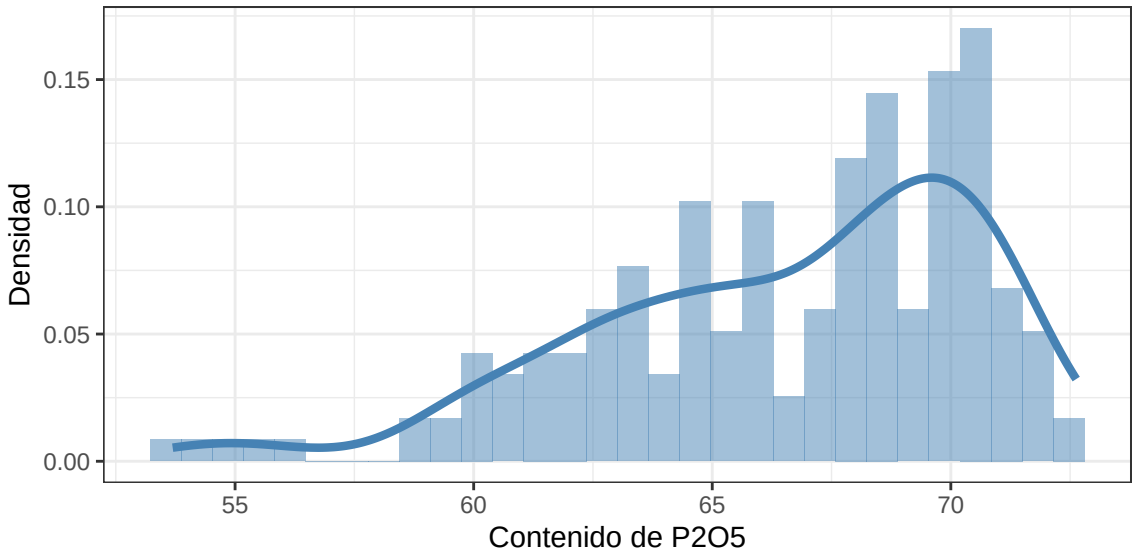

Figure 3: Histograma y estimación de la densidad de la cantidad de P205 en las muestras.

Finalmente, y a modo exploratorio, vamos a calcular el coeficiente de correlación entre la energía de cada banda de frecuencia y la cantidad de P_20_5. De esta forma queremos seguir indagando sobre qué bandas de frecuencia deberían ser más importantes en el modelo de predicción. En la Figura 4 puede verse el valor absoluto del coeficiente de correlación de Pearson en función de la banda de frecuencia. Retomaremos los resultados de esta figura luego de ajustar un modelo.

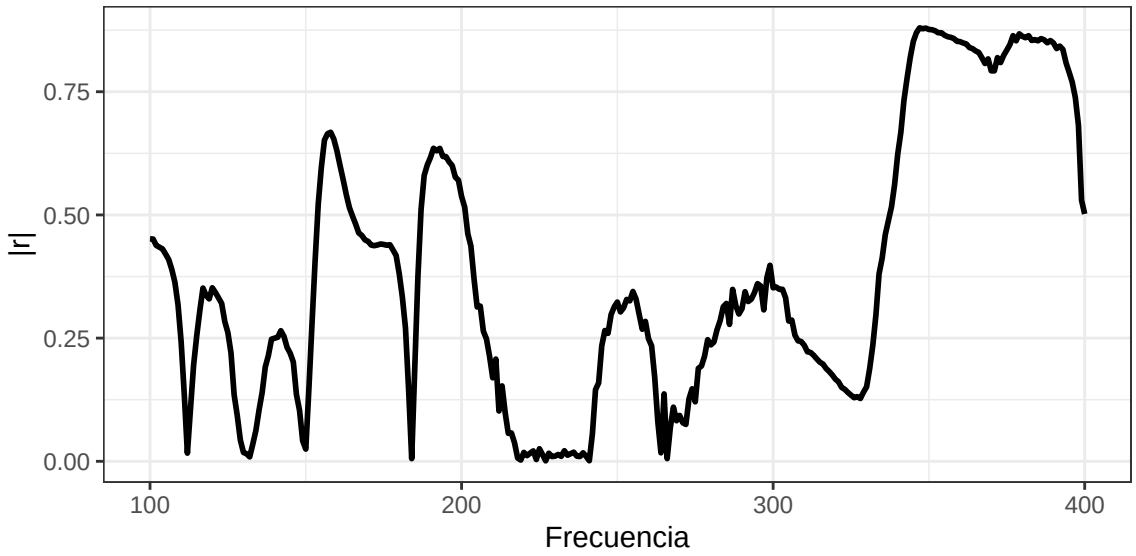

Figure 4: Energía en función de las frecuencias. Cada línea de color representa una vasija mientras que la línea negra representa al promedio de las 150 vasijas.

Creación de un modelo predictivo

Como lo que queremos hacer es entrenar un modelo predictivo, la métrica que vamos a utilizar para evaluarlo es el error cuadrático medio (MSE[1]) calculado a partir de una muestra de *testeo*. Para esto vamos a dividir los datos en dos partes, dejando dos tercios de los datos (120 vasijas) en el set de entrenamiento o *training* y un tercio de los datos (60 vasijas) en el set de validación o *testing*. Para evitar que haya una representación desigual de cantidad de P_20_5 en las muestras de entrenamiento y validación, vamos a hacer esta división estratíficada[2] utilizando la función `initial_split` del paquete `tidymodels`.

Vamos a considerar dos familias de modelos para resolver este problema. Primero exploraremos los modelos lineales con regularización (Ridge, Lasso o Elastic Net) utilizando el paquete `glmnet`. Luego exploraremos modelos basados en regresión de componentes principales (PCR[3]) utilizando el paquete `pls`.

Modelos lineales con regularización

El primer modelo que vamos a ajsutar es un modelo lineal con regularización de Lasso ($\alpha = 1$ en `glmnet`). Para esto vamos a utilizar el set de datos de entrenamiento, y para encontrar el mejor λ utilizaremos como criterio el MSE y una validación cruzada con 10 *folds*.

Para la regresión con regularización Lasso, el lambda que minimiza el MSE es 0.034911, con un valor de MSE de 0.6228984.

En la Figura XXX podemos ver que la mayoría de las componentes no nulas del modelo corresponden con variables con alta correlación con Y. Esto es meramente exploratorio ya que para la selección de variable también resulta relevante que tan correlacionadas están las variables entre sí.

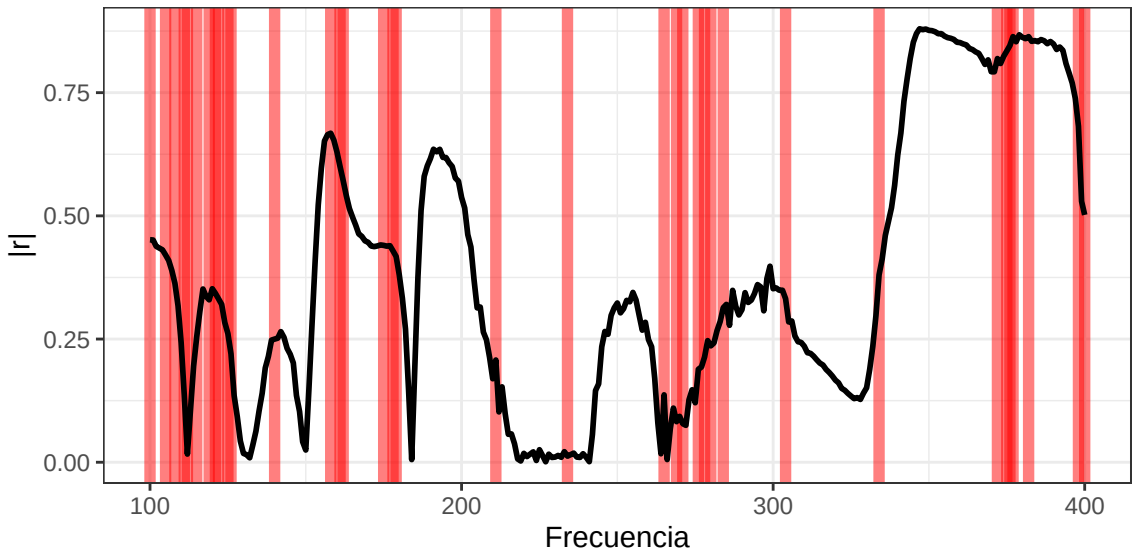

Figure 5: MSE para un modelos glmnet en función de alpha (de Ridge a Lasso) para el mejor lambda obtenido a partir de cross-validation con 10 folds. El panel b muestra un zoom para valores de alpha de 0.5 a 1

Buscando los mejores parámetros λ y α Un posible paso siguiente es el de, además de buscar el mejor λ, también hacer una búsqueda en una grilla del parámetro α. Este parámetro es el responsable de convertir el modelo de Ridge ($\alpha = 0$) a Lasso ($\alpha = 1$) pasando por Elastic Net ($\alpha = 0.5$). Haremos la búsqueda para una grilla de valores intermedios entre 0 y 1 con paso de 0.01.

[1] Del inglés *Mean Squared Error*.
[2] Como se trata de una variable numérica, la función estratifica la separación a partir de los cuartiles.
[3] Del inglés *Principal Components Regression*.

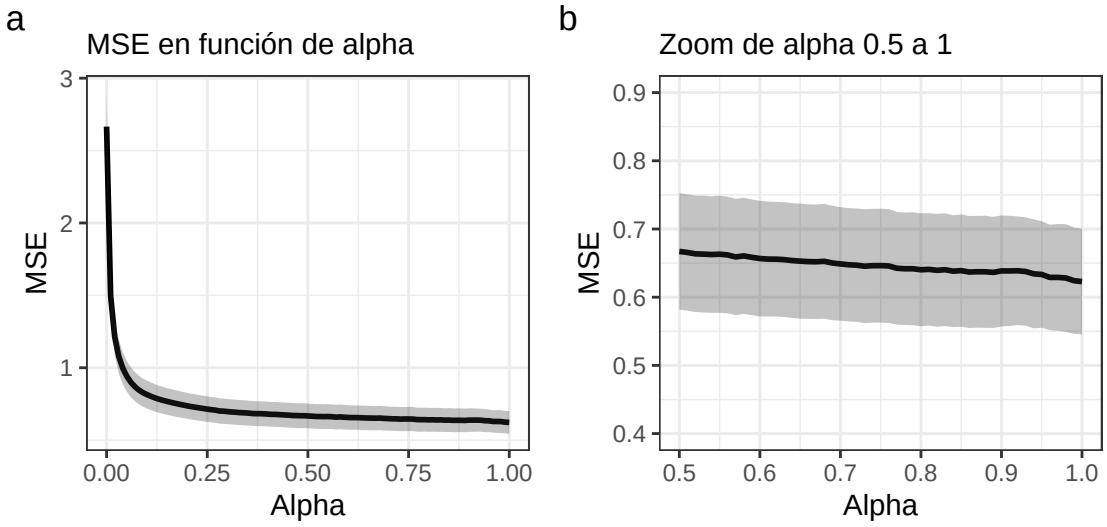

Figure 6: MSE para un modelos glmnet en función de alpha (de Ridge a Lasso) para el mejor lambda obtenido a partir de cross-validation con 10 folds. El panel b muestra un zoom para valores de alpha de 0.5 a 1

Puede verse que el MSE disminuye monotonamente con el α, llegando a un mínimo de 0.623 para $\alpha = 1$. Es decir, la regresión regularización Lasso es la más conveniente. Para este valor de α el mínimo error se obtuvo para un λ de 0.03, con un número de componentes no nulas igual a 36.

Regresión de componentes principales

Otro enfoque que exploramos para abordar el problema es el de reducción de la dimensión. Consideramos el método PCR que consiste en ajustar un modelo de regresión lineal utilizando como variables predictoras un sobconjunto de las componentes obtenidas a partir del análisis de componentes principales (PCA[4]). De esta manera, se reduce la cantidad de variables predictoras del modelo.

Para implementar este método utilizamos la función `pcr` de la librería `pls`. Esta función calcula las componentes principales y ajusta el modelo de regresión lineal con la cantidad de componentes que se desee. Para elegir la cantidad de componentes usamos *Cross-Validation* sobre la muestra de entrenamiento, separando en 10 folds y buscando el número que minimice el error de predicción. En la figura XX se muestra un gráfico con los resultados del MSE obtenido por CV en función de la cantidad de componentes.

[4]Del inglés *Principal Components Analysis*.

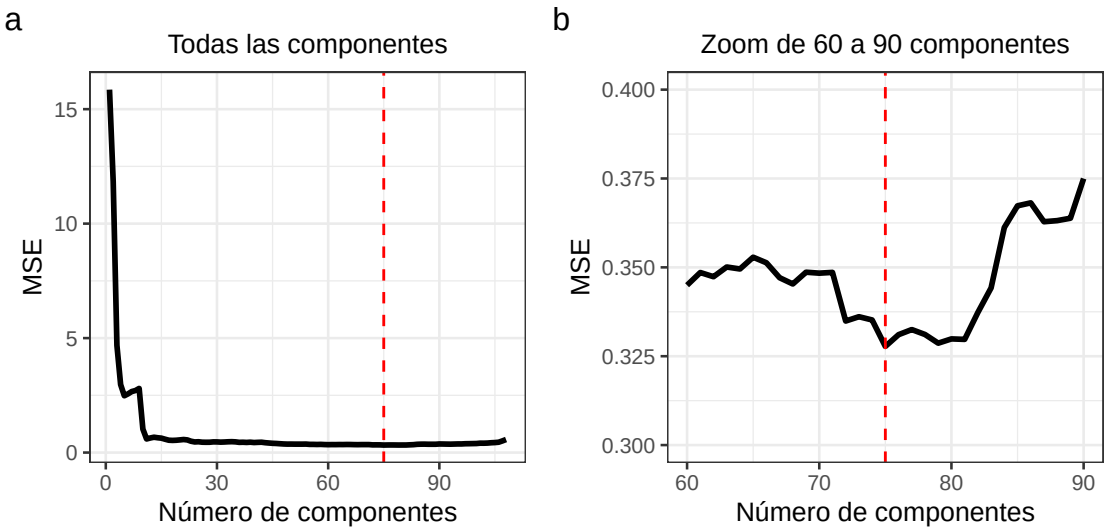

Figure 7: Energía en función de las frecuencias. Cada línea de color representa una vasija mientras que la línea negra representa al promedio de las 150 vasijas.

El óptimo se encuentra en 75 componentes y el valor obtenido para el error de predicción por CV sobre la muestra de entrenamiento es 0.328.

Comparación de la performance de los modelos predictivos

Métricas con el set de testeo

Conclusiones