Proyecto Final: Cómputo estadístico (primer avance)

J. Antonio García Ramírez

26 de septiembre, 2018

The task

- **Predecir** el año de lanzamiento de una canción a partir de las características del audio, utilizando PLS Y PCR.

Dataset

- Canciones comerciales que varían en su año de lanzamiento entre 1922 y 2011
- Subconjunto del famoso **Million Song Dataset**
- 515,345 observaciones con 90 variables y etiqueta
- Las covariables corresponden a los 12 promedios y las 78 covarianzas del timbre [1] a lo largo de 12 segmentos de la canción extraidas con la api The Echo Nest API

Ejemplo: Un *la* de 440 Hz emitido por una flauta es distinto del *la* que emite una trompeta aunque estén tocando la misma nota

[1] El timbre es la cualidad que caracteriza un sonido. Se trata de una de las cuatro cualidades esenciales del sonido (junto con el tono, la duración y la intensidad)

Ejemplo de observación

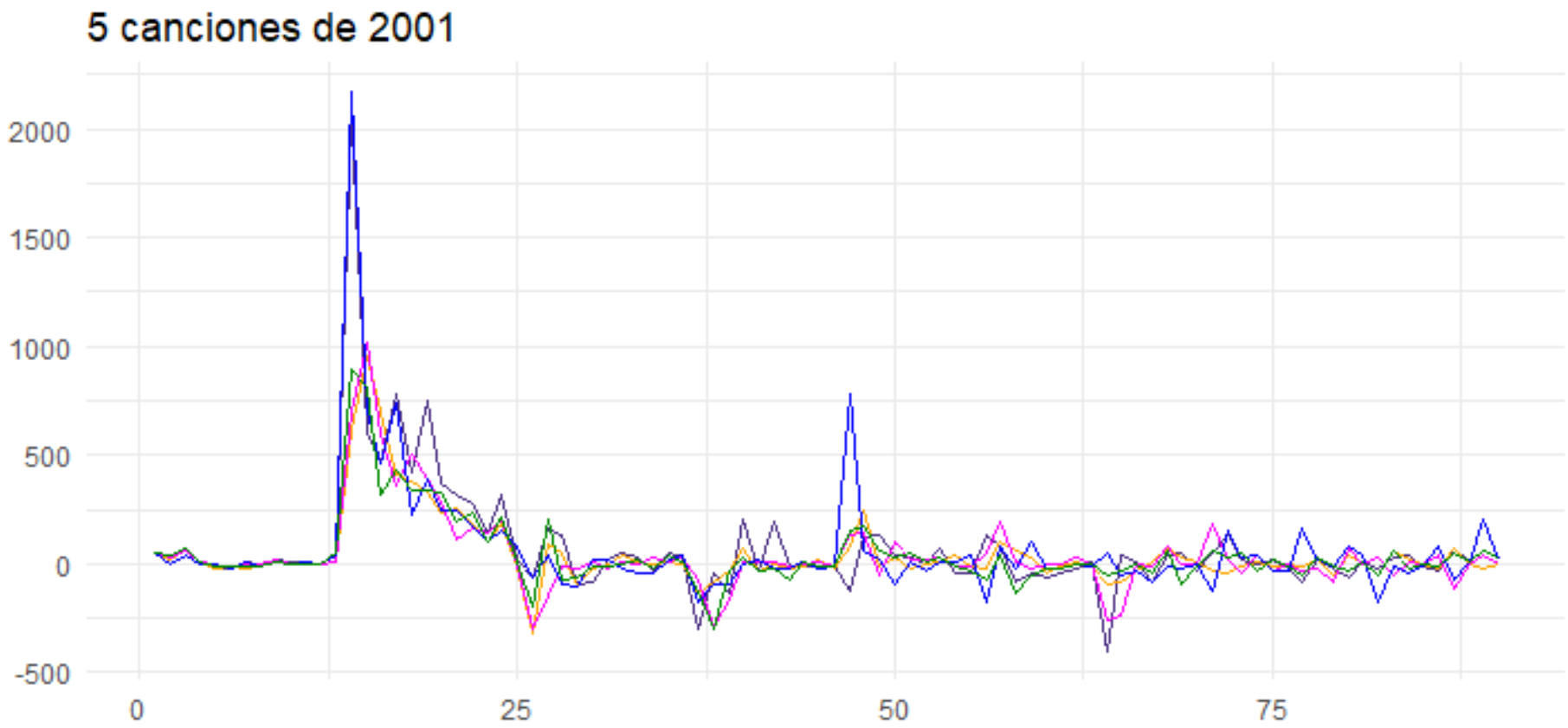

Figure 1:

Distribución de la respuesta

Figure 2:

Situación actual

- Dataset limpio, no requirió de técnicas de imputación
- Resultados sin muestreo, usando todo el dataset

Esquema

1. Determinar si estamos en una situación del tipo *maldición de la dimensionalidad*, cálculo explícito

   - En caso afirmativo hablar sobre la geometría de los conjuntos de datos *chaparros* [2], con base en los resultados propuestos en Hall, P., Marron, J. S. and Neeman, A. (2005) Geometric representation of high dimension low sample size data

2. Resultado base, predicción con:

   - Un modelo apropiado de *OLS* y cv
   - Un modelo estimado con *Ridge* y cv
   - Un modelo estimado con *Lasso* y cv

[2] High Dimension Low Sample Size (HDLSS)

Esquema

3. Resultado central, predicción con:

   - PCR y cv
   - PLS y cv

4. Comparación de resultados (Restricción a 2 fold cv):

   - Con una competencia en Kaggle
   - Utilizando MSE

Posibles anexos:

- Enfoque de PCR usando Parallel analysis: Imperativo una simulación eficaz
- Clusterización de observaciones: ¿Correspondencia entre años?

Cereza

- PLS y ecuaciones estructurales o PLS-Path modeling