Ciencia de datos 1: Usando PCA

Jorge Luis Ramos Zavaleta
Centro de Investigación en Matemáticas. Unidad Monterrey.
Email: jorge.ramos@cimat.mx

***Resumen*—En este reporte se muestran las principales conclusiones que obutve al realizar los ejercicios de la primera tarea concerniente a la materia de Ciencia de datos.**

I. Primer ejercicio

En este ejercicio se nos presenta una base de datos concernientes a dos tipos de vinos: blanco y rojo, y una evaluación que se tiene de ellos en función de su calidad. Y queremos analizar si visualmente podemos encontrar si alguna de las variables fisico-quimicas nos permitiría establecer una posible clasificación entre los 2 tipos de vinos, y después observar si existen dentro de estas variables fisico-quimicas algunas que nos puedan permitir distinguirlos con respecto cada tipo por su calidad. Para iniciar el análisis primero hice un histograma de la variable de calidad.

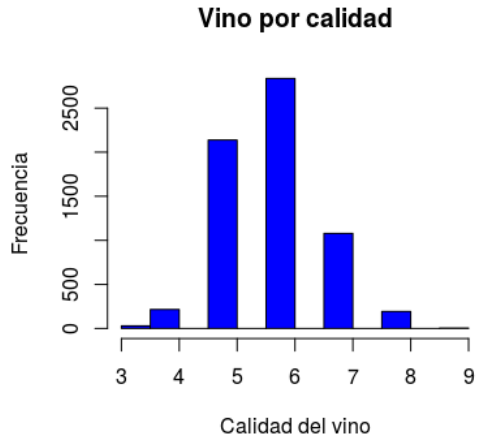

Figura 1. Frecuencias en las calidades de los vinos

En el histograma podemos observar no existen respuestas de tipo 1,2 y 10, y en general se tienen datos de vinos de calidad media, lo cual puede sesgar nuestro análisis ya que al no estar presentes estos datos los resultados que podemos obtener pueden estar sesgados. Después separe los datos clasificandolos entre los 2 tipos de vinos y realice boxplots para observar si las çajas"no se intersectaban lo que podría dar un primer indicio de una variable que permitira hacer la clasificación.

Encontre 3 variables que lograban esto de alguna manera: $density$, $sulphates$ y $free\ sulfur\ dioxide$ debido a que esta relación la encontre en otras variables también decidí contrastar el boxplot contra el boxplot eliminando algunos cuantiles (0 al cuantil 5, y del cuantil 95 al 100) para eliminar la presencia de valores atípicos y observar si la relación se preservaba.

Como se puede observar la relación se preserva, y las cajas casi no se intersectan, lo que podría sugerir que estas 3 variables podrían lograr una posible clasificación por si mismas. Sin embargo no debe pasarsenos por alto que la caja no son todos los datos por lo que las variables podrían no lograr la clasificación por su propia cuenta.

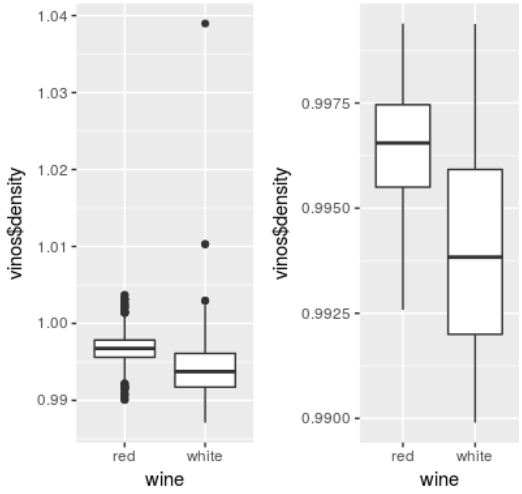

Figura 2. Boxplots por densidad y tipo de vino

Para dar cuenta de este último hecho realice gráficas de violín para ver si las concentraciones de los datos no se intersectaban o que bien la intersección se daba por un lado con una alta concentración de datos y en el otro con una pequeña concentración de ellos.

En la figura 5 encontramos que la densidad en el caso del vino rojo se concentra en una parte donde los datos del vino blanco no estan tan concentrados lo cual puede sugerir que en combinación con alguna otra variable o variables la densidad puede ayudar a lograr la clasificación entre vino rojo y blanco.

En la figura 6 se puede observar que las concentraciones altas de datos de sulfatos se dan generalmente de manera que el otro tipo de vino tenga concentraciones pequeñas. Sin embargo hay una parte donde las altas concentraciones

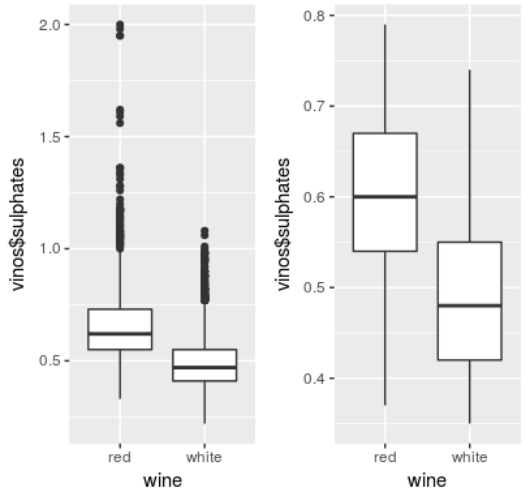

Figura 3. Boxplots por sulfatos y tipo de vino

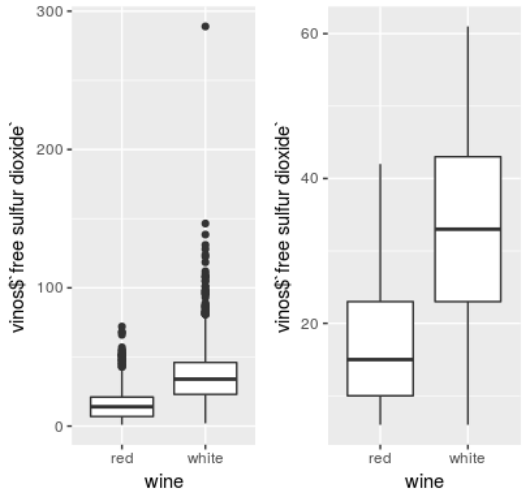

Figura 4. Boxplots por sulfato libre de dioxido y tipo de vino

chocan en ambos tipos lo que sugiere que esta variable puede no ayudar mucho en la clasificación de los tipos de vino.

Y por último en esta parte del análisis en la figura 7 se puede observar que en general las concentraciones de los datos del sulfuro libre de dióxido se encuentran en cantidades muy bajas de éste en el caso del vino rojo mientras que en este caso las concentraciones altas en el caso del vino blanco se logran en donde se tienen menos concentraciones en el caso del vino rojo. Lo cual sugiere que esta variable en conjuntos tal vez con las dos anteriores pueden permitir generar un clasificador decente para conocer el tipo de vino.

Para la segunda parte que es indicar visualmente si es posible encontrar alguna variable fisico-quimica que permita clasificar el vino dado su tipo por nivel de calidad. Para ello agrupe los datos en 3 categorías: $mala$, $regular$ y $excelente$. Para establecer estas categoria construí la tabla I

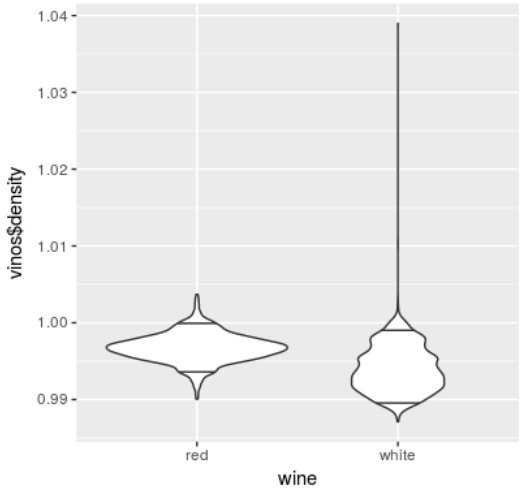

Figura 5. Graficos de violin para densidad y tipo de vino

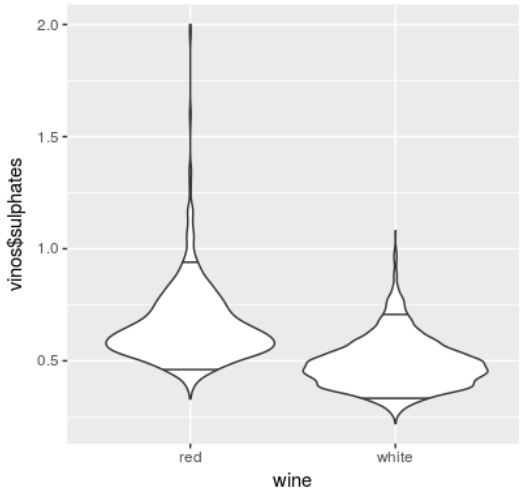

Figura 6. Graficos de violin para sulfatos y tipo de vino

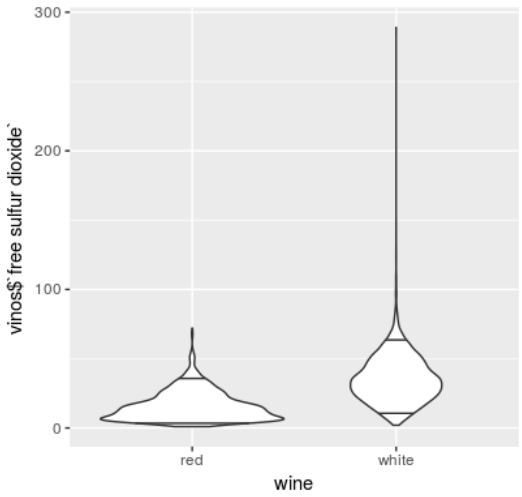

Figura 7. graficos de violin para sulfato libre de dioxido y tipo de vino

Cuadro I
TOTAL DE VINOS POR CALIDAD

wine	3	4	5	6	7	8	9
red	10	53	681	638	199	18	NA
white	20	163	1457	2198	880	175	5

para establecer los límites para cada categoría. De la tabla tenemos que tanto la calidad 1,2 y 10 no se encuentran presentes y la calidad 9 tiene muy pocos datos en un tipo y 0 en el otro, por lo que decidí separar en mala calidad los que tengan calidad menor o igual a 4, en regular los que tengan calidad 5 o 6, y en excelente a los que tuvieran de 7 en adelante.

Para establecer posibles limites sobre el análisis grafique las frecuencias de las 3 categorias para cada tipo de vino. En la figura 8 podemos observar que la mayoría de los datos se encuentran las categorías de regular para ambos tipos y que tenemos casi el triple de datos de vino tipo blanco que del rojo, lo cual puede darnos conclusiones erroneas sobre como afectan las variables a cada categoria de calidad, y esto en particular se puede hacer mayor el error en el caso de las categorias excelente y mala que casi no tienen datos comparadas con la categoria regular.

En esta parte analice los gráficos de violin de todas las variables pero solo considere que de relevancia pueden ser los sulfatos, la densidad, el nivel de alcohol, el acído cítrico y la volatilidad ácida.

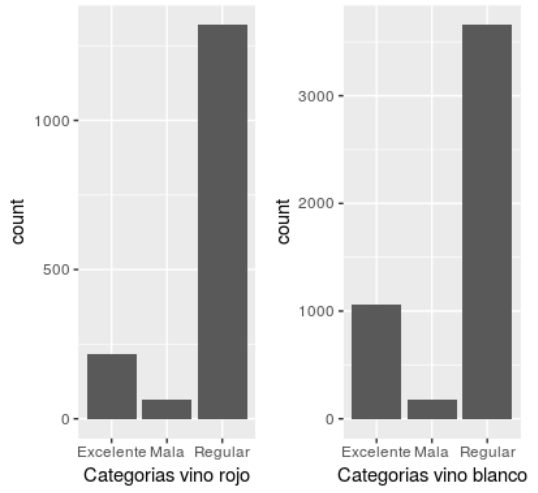

Figura 8. Frecuencias por tipo de vino y categoria de calidad

Comenzando por el nivel de alcohol, en la figura 9 en ambos tipos de vino se puede observar que una mayor concentración de alcohol en el vino tiende a indicar una calidad excelente. Por otra parte en la figura 10 tenemos la volatilidad ácida y nos indica que una alta volatilidad ácida

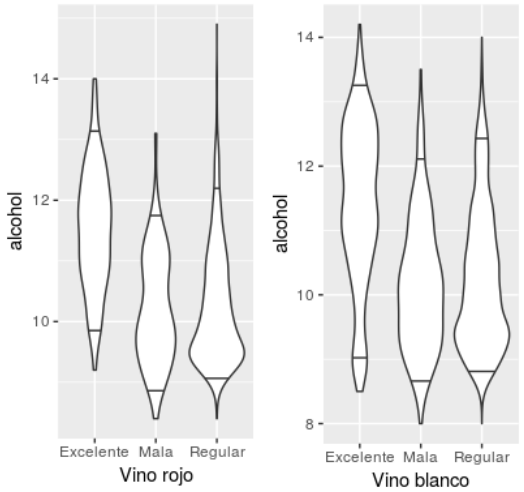

Figura 9. graficos de violin por nivel de alcohol, tipo de vino y categoria de calidad

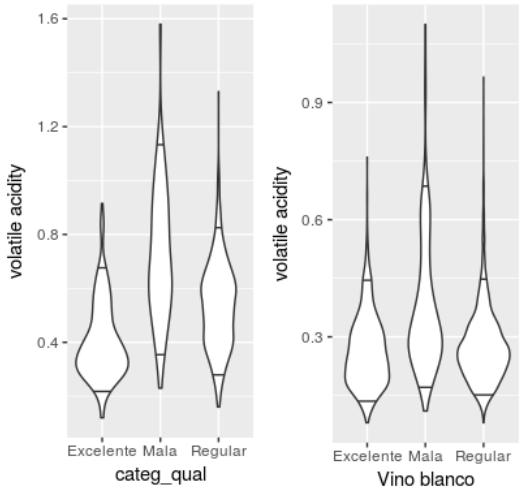

Figura 10. graficos de violin por volatilidad ácida, tipo de vino y categoria de calidad

puede determinar que un vino rojo caiga en la categoria de mala calidad mientras que al diminuir parece indicar una mejora en la calidad.

En el caso de la volatilidad ácida que se encuentra en la figura 11 tenemos que para el caso del vino rojo entre menor sea el valor tiende a tener menor calidad y al aumentar tiende a mejorar la calidad, por su parte en el caso del vino blanco no se ve que exista una tendencia en general con respecto a esta variable.

Para el caso de los sulfatos la figura 12 sugiere en el caso del vino rojo que una menor presencia de sulfatos tiende disminuir la calidad del vino rojo, mientras que en el caso del vino blanco no parece haber una tendencia. Por último en el caso de la densidad que se observa en la figura 13 parece que tenemos una tendencia en el caso del vino blanco

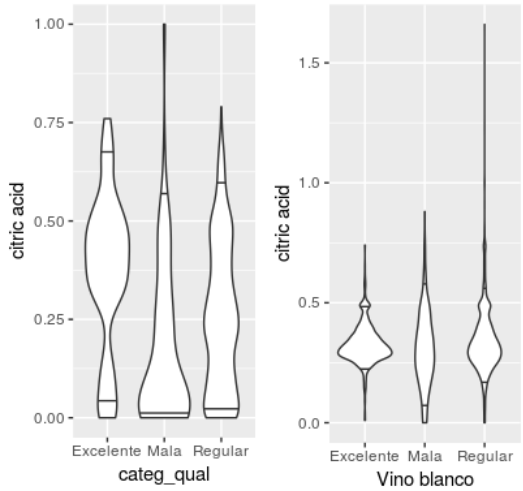

Figura 11. graficos de violin por acidez cítrica, tipo de vino y categoria de calidad

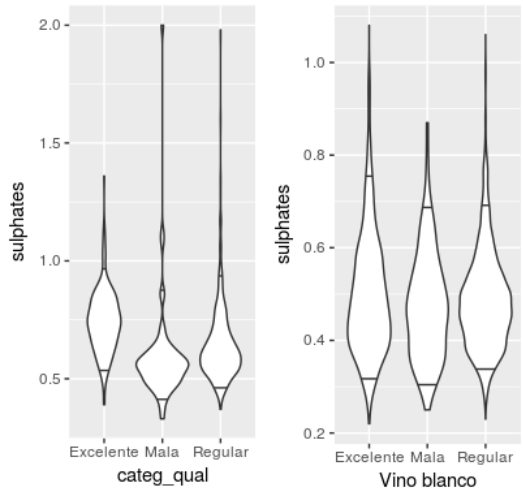

Figura 12. graficos de violin por sulfatos, tipo de vino y categoria de calidad

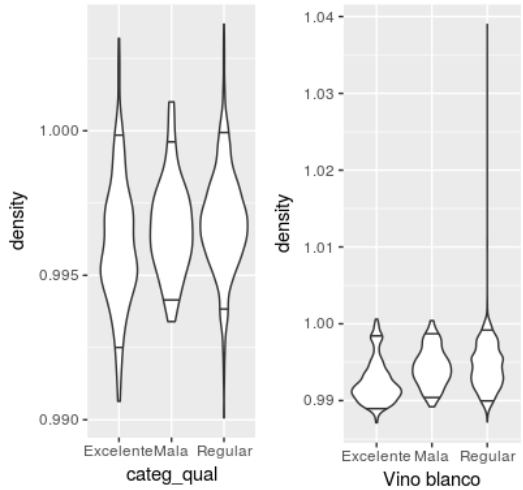

Figura 13. graficos de violin por densidad, tipo de vino y categoria de calidad

donde encontramos una mayor concentración con una menor densidad en la categoria de excelente y en el caso del vino blanco parece no haber una tendencia.

Del presente análisis podemos concluir que no hay una variable que por si misma pueda permitirnos clasificar el vino en sus 2 tipos: rojo y blanco, pero que en conjunto es posible que se pueda lograr dicha separación. En la segunda parte del análisis no logramos encontrar como tal variables que nos permitan por si solas establecer la clasificación en categorias de calidad pero observamos algunas tendencias con respecto a algunas variables, lo que sugiere que tal vez si es posible lograr crear un buen clasificador.

II. Segundo ejercicio

Debemos realizar PCA a la siguiente matriz que supondre que es una matriz de covarianzas debido a la siguiente parte del ejercicio

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

Con $\rho > 0$. Para calcular sus componentes principales solo requerimos obtener sus valores propios y sus vectores propios. Por lo que haciendo el algebra obtenemos que sus valores propios son $\lambda_1 = 1 + \rho$ y $\lambda_2 = 1 - \rho$ por lo que al menos tenemos un valor propio positivo y por tanto al menos una componente principal.

En el caso de los vectores propios tenemos que el vector propio normalizado asociado a λ_1 es $v_1 = \left(\frac{1}{\sqrt{2}} \ \frac{1}{\sqrt{2}}\right)^t$ y el asociado a λ_2 es $v_2 = \left(\frac{1}{\sqrt{2}} \ -\frac{1}{\sqrt{2}}\right)^t$. Por lo que nuestra matriz de proyección será la formada utilizando los vectores propios como columnas de dicha matriz. Además debido a que $\frac{1+\rho}{1+\rho+1-\rho} = \frac{1+\rho}{2}$ entonces tenemos que $\rho \in (0, 1)$.

Para la segunda parte debemos considerar calcular PCA con la matriz considerando la varianza de CX_1 y manteniendo la de x_2. Por lo que nuestra matriz tiene la siguiente forma:

$$\begin{pmatrix} c^2 & c\rho \\ c\rho & 1 \end{pmatrix}$$

Realizando el algebra para obtener sus valores propios obtenemos el siguiente polinomio característico

$$p(\lambda) = \lambda^2 - \lambda(c^2 + 1) + c^2 - c^2\rho^2$$

Que para mantener consistencia con la parte anterior requerimos que $0 < \rho < 1$ por lo que para que nuestro polinomio tenga raíces reales requerimos que $c \in [-1/3, 1/3]$. Suponiendo que λ_1 es positivo calculamos su vector propio para lo cual requerimos resolver el sistema

$$\begin{aligned} (c^2 - \lambda_1)x_1 + c\rho x_2 &= 0 \\ c\rho x_1 + (1 - \lambda_1)x_2 &= 0 \end{aligned} \tag{1}$$

De donde obtenemos

$$\frac{\lambda_1 - c^2 + c\rho}{\lambda_1 - 1 + c\rho} x_1 = x_2$$

por lo que el vector propio normalizado asignado a λ_1 es

$$v_1 = \left(\sqrt{1+\left(\frac{\lambda_1 - c^2 + c\rho}{\lambda_1 - 1 + c\rho}\right)^2}\right)^{-1} \begin{pmatrix} 1 \\ \frac{\lambda_1 - c^2 + c\rho}{\lambda_1 - 1 + c\rho} \end{pmatrix}$$

De aquí podemos ver que el primer componente principal depende de c por lo que dependiendo del escalamiento se pueden obtener direcciones contrarias o un aumento artificial de la variación con respecto al componente principal obtenido anteriormente.

La siguiente parte del ejercicio es aplicar PCA normalizando y sin normalizar a una base de datos sobre causas de muerte en EE.UU. aparentemente en los años 80's. Ambos PCA se realizaron centrando los datos para solo observar la influencia de la normalización en el PCA.

Al realizar ambos PCA se observan cambios gigantescos en la varianza explicada por cada componente en cada versión del PCA. En el caso del PCA no normalizado el primer componente tiene una proporción de la varianza de 0.9363, mientras que en el normalizado la primera componente principal solo explica 0.4859 de la varianza. Además de esto en la primera componente del no normalizado solo una variable tiene un valor muy alto de -0.9363, mientras que en el normalizado se reparte los valores más altos entre 2 variables $card$, $canc$.

Con respecto a dichos coeficientes en valor absoluto de ambos PCA, para el primer componente se obtienen que las variables $canc$ y $card$ tienen valores de 0.5157479 y 0.4996192 respectivamente. Mientras que en el caso no normalizado se tiene que la variable $card$ tiene un valor de 0.936363587, lo que implica un peso enorme y por tanto la convierte en practicamente la única variable explicativa o de valor en el modelo.

En el caso del según componente principal tenemos para el caso no normalizado que en contraste con las otras variables solo $canc$ tiene un valor absoluto de 0.85829848, mientras que en el caso normalizado $pneu$ tiene el valor más alto en valor absoluto, con un valor de 0.69503875, por lo que esto junto con lo anterior si muestra una diferencia significativa entre la versión normalizada y no normalizada.

Debido a que la normalización se hace con el fin de eliminar problemas con diferencias entre escalas entre las variables entonces en general es mejor hacer uso de la normalización para realizar el PCA. Aunque en algunos casos donde la escala en que se mida si tenga relevancia se podría hacer uso de la versión no normalizada.

Se agregan los biplots con respecto a los primeros dos componentes principales y separando con elipses las regiones los datos para ambos casos del PCA. En la figura 14 se observa una flecha muy grande con respecto a $card$ que era lo que

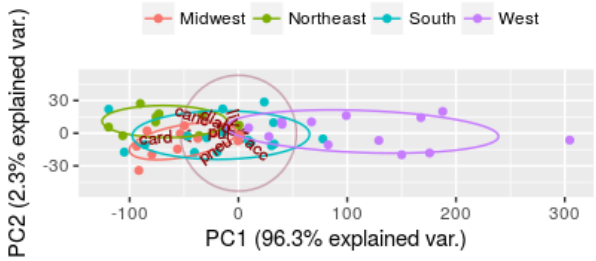

Figura 14. Biplot del PCA no normalizado

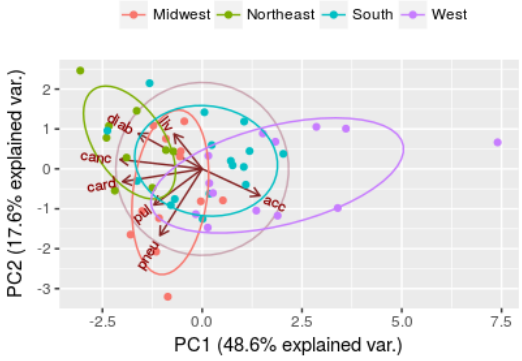

Figura 15. Biplot del PCA normalizado

indicamos al principio con respecto al primer componente de la no normalizada. Y en el caso de la versión normalizada no se observan patrones donde tengamos que una sola variable explique toda la variabilidad en los datos, lo cual parece correcto en este caso.

III. Tercer ejercicio

Para efectos de la realización correcta de un indicador debe establecerse correctamente el constructo sobre el cual se va a realizar dicho índice. En este sentido es preciso establecer cual es la definición de marginación a emplearse para realizar dicho proceso.

Para efectos de este trabajo debe hacerse la distinción entre marginalidad y marginación dado que estos conceptos tienden a ser confundidos. La marginalidad hace referencia

a personas que estan en el sentido de marginación debido a que no forman ni parecen tener forma de pertenecer debido a falta de oportunidades a la sociedad en la que se encuentran residiendo. Por su parte marginación se refiere a un estado en conjunto de una población donde no tienen los recursos para salir de esa posición ante la sociedad.

En este sentido cabe aclarar cuales son considerados dichos recursos para la elaboración del indicador. La CONAPO desde 1990 ha estado realizando este indicador para lo cual considera variables de escolaridad, acceso a servicios públicos en vivienda y carencia de bienes. La idea de elaborar un indicador suficientemente objetivo puede permitir focalizar de mejor manera los esfuerzos del gobierno para contrarrestar este fenomeno.

El índice de marginación por localidad es en resúmen una suma ponderada de 8 indicadores. Para la ponderación se debe considerar la forma de establecer los pesos particulares para cada indicador. En el caso particular de este indicador se hace uso de una técnica conocida como PCA para obtenerlos.

El resultado obtenido es

$$\begin{aligned} IML_{2010} = {} & 0{,}3165331x_1 + 0{,}4225538x_2 + \\ & 0{,}3008837x_3 + 0{,}2865479x_4 + 0{,}3833827x_5 + \\ & 0{,}3849674x_6 + 0{,}2916162x_7 + 0{,}4103562x_8 \end{aligned}$$

Donde

- x_1: Porcentaje de población de 15 años o más analfabeta
- x_2: Porcentaje de población de 15 años o más sin primaria completa
- x_3: Porcentaje de viviendas particulares habitadas sin excusado
- x_4: Porcentaje de viviendas particulares habitadas sin energía eléctrica
- x_5: Porcentaje de viviendas particulares habitadas sin disponibilidad de agua entubada
- x_6: Promedio de ocupantes por cuarto en viviendas particulares habitadas
- x_7: Porcentaje de viviendas particulares habitadas con piso de tierra
- x_8: Porcentaje de viviendas particulares habitadas que no disponen de refrigerador

Una consecuencia directa de conocer los pesos usando PCA es que dichos pesos en términos absolutos nos sugieren los indicadores que mayor variación presentan en los datos, y que pueden darnos una idea en este sentido de cual indicador es al que debe voltearse a ver prioritariamente. En este caso los indicadores de población de 15 años o más sin primari a completa y el porcentaje de viviendas que no tienen refrigerador parecen ser las que mayor impacto tienen en el indicador y son categorias que deben voltearse a ver primero para ayudar a disminuir la marginación.

Una de las actividades económicas predominantes que se dan usualmente en los círculos de marginación es la preparación y comercialización de alimentos. Por lo que un indicador extra que se podría incluir es el porcentaje total de viviendas habitadas que cuenten con gas.

Este indicador puede tener un peso mayor también debido a que de no contarse con gas los alimentos que se consuman pueden no ser saludables o pueden ser utilizadas alternativas al gas que pueden afectar su salud e imposibilitar aun más a la localidad.

IV. Cuarto ejercicio

Para este ejercicio se realiza una regresión PCA para clasificar imagenes digitales que contienen los dígitos del 1 al 9 escritos a mano y digitalizados para su procesamiento.

Las imagenes originales son de 16x16 pixeles por lo que consideramos 1 variable por cada pixel, por lo que tendremos 256 variables a analizar. Debido a que la dimensionalidad es muy grande para intentar analisis exploratorios multivariados, y además debido a que las variables pueden estar correlacionadas (un pixel contiene información que con otro podríamos obtener) entonces lo adecuado sería intentar eliminar la correlación y de paso intentar disminuir la dimensionalidad para poder trabajar más comodamente con técnicas simples pero poderosas como lo es la regresión.

Para eliminar la correlación e intentar disminuir las dimensiones a analizar aplique PCA a las 256 variables para obtener una matriz de $scores$ y la utilizare para realizar una regresión multivariada evaluando un posible menor número de variables sin perder poder de predicción.

Para la realización del PCA contamos con un conjunto de datos de entrenamiento para con ellos obtener los coeficientes de la regresión, y también contamos con un conjunto de datos de prueba que nos servirán para observar que tan bueno es nuestro modelo para clasificar, pero en este caso más que su bondad para clasificar revisaremos el error cuadrado medio, para con ello lograr elegir el número de componentes adecuado para lograr una buena clasificación.

Entrenando el modelo y evaluando los errores cuadrado medios por el número de componente principales podemos observar en la figura 16 que el error cuadrado medio para ambos conjuntos tanto el de entrenamiento como en el de prueba va disminuyendo conforme se aumenta el número de componentes principales, lo cual parece una conclusión lógica.

El número de componentes principales a elegir debe estar en función del mínimo error cuadrado medio pero también debe considerarse una disminución de la dimensión suficiente para hacer más efectivo el tiempo de cómputo perdiendo el

mínimo poder de predicción posible.

En este caso puede verse en la figura 16 que el descenso de la curva de prueba empieza a desacelerar a partir de los 25 componentes principales, por lo que es aconsejable usar entre 25 y 30 componentes principales para lograr el objetivo de mantener una buena predicción con una dimensionalidad tan alta.

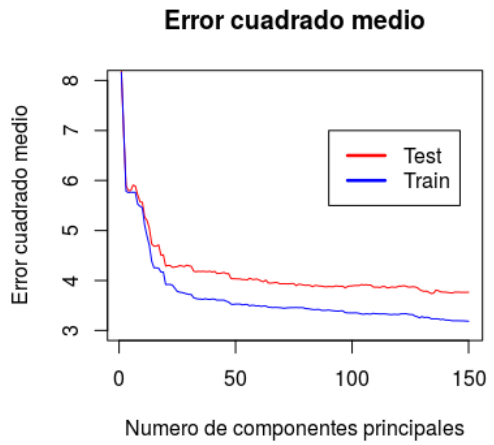

Figura 16. Error cuadrado medio de los conjuntos de entrenamiento y prueba por número de componentes principales