Centro de Investigación en Matemáticas (CIMAT). Unidad Monterrey

Tarea 2

Enrique Santibáñez Cortés

26 de febrero de 2021

1. Problema 1

Considera una matriz de datos $X_{n\times d}$. PCA puede formularse también como el problema de encontrar un subespacio (ortonormal) de baja dimensión de forma tal que se minimicen los errores de las proyecciones de los datos en tal subespacio.
Si consideramos una base ortonormal $\{u_j\},\ j = 1, \cdots, d$, ya vimos que una observación x_i puede expresarse como una combinación lineal

$$\mathbf{x_i} = \sum_{\mathbf{j=1}}^{\mathbf{d}} \alpha_{\mathbf{ij}}\mathbf{u_j}.$$

Por la ortogonalidad de $\mathbf{u_j}$, podemos expresar $\alpha_{ij} = \mathbf{x_i'u_j}$. Entonces

$$\mathbf{x_i} = \sum_{\mathbf{j=1}}^{\mathbf{d}} (\mathbf{x_i'u_j})\mathbf{u_j}. \tag{1.1}$$

Ahora, considera una aproximación basada en los primeros $p < d$ vectores de la base de acuerdo al modelo lineal:

$$\mathbf{\hat{x}_i} = \sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{z_{ij}u_j} + \sum_{\mathbf{j=p+1}}^{\mathbf{d}} \mathbf{b_j u_j}. \tag{1.2}$$

Observa que los coeficientes z_{ij} *dependen* de la observación i, mientras que b_j son constantes para todas las observaciones.
Considera la minimización de la siguiente función de costo:

$$L = \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{x_i} - \mathbf{\hat{x}_i}||^{\mathbf{2}}. \tag{1.3}$$

1.1. Encuentra los valores de z_{ij}, b_j que minimizan la función de costo(1.3)

Tenemos que la base $\{u_j\},\ j = 1, \cdots, d$ es ortonormal, es decir, se cumple que

- $u_i \cdot u_j = 0$, donde $i, j = 1, \cdots, d, |i \neq j$.
- $u_i \cdot u_i = 1$, donde $i = 1, \cdots, d$.

Una vez aclarado la definición de base ortonormal, procedemos a encontrar los puntos de inflexión para z_{ij} ocupando el criterio de primera derivada de la función de costo (1.3), pero primeros reescribiremos la función de costo ocupando (1.1) y (1.2) de la siguiente manera:

$$L = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} ||\mathbf{x_i} - \mathbf{\hat{x}_i}||^{\mathbf{2}} = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \left|\left| \sum_{\mathbf{j=1}}^{\mathbf{d}} (\mathbf{x_i'u_j})\mathbf{u_j} - \sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{z_{ij}u_j} - \sum_{\mathbf{j=p+1}}^{\mathbf{d}} \mathbf{b_j u_j}. \right|\right|^{\mathbf{2}}. \tag{1.4}$$

Entonces ocupando la primera derivada de la función de costo(1.4) con respecto a z_{kl} donde $k \in \{1, 2, \cdots, n\},\ \ l \in \{1, 2, \cdots, p\}$ es

$$\frac{\partial \mathbf{L}}{\partial \mathbf{z_{kl}}} = \frac{\mathbf{2}}{\mathbf{n}} \left(\sum_{\mathbf{j=1}}^{\mathbf{d}} (\mathbf{x_k'u_j})\mathbf{u_j} - \sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{z_{kj}u_j} - \sum_{\mathbf{j=p+1}}^{\mathbf{d}} \mathbf{b_j u_j} \right) \cdot (-\mathbf{u_l})$$

Ocupando (1.1) tenemos que $u_j \cdot u_l = 0$ si $j \neq l$ y $u_j \cdot u_l = 0$ si $j = l$, por lo que podemos simplificar la derivada como

$$\begin{aligned}
\frac{\partial \mathbf{L}}{\partial \mathbf{z_{kl}}} &= \frac{\mathbf{2}}{\mathbf{n}} \left(- \left(\sum_{\mathbf{j=1}}^{\mathbf{d}} (\mathbf{x_k'u_j})\mathbf{u_j} \right) \cdot \mathbf{u_l} + \left(\sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{z_{kj}u_j} \right) \cdot \mathbf{u_l} + \left(\sum_{\mathbf{j=p+1}}^{\mathbf{d}} \mathbf{b_j u_j} \right) \cdot \mathbf{u_l} \right) \\
&= \frac{\mathbf{2}}{\mathbf{n}} \left(- \sum_{\mathbf{j=1}}^{\mathbf{d}} (\mathbf{x_k'u_j})\mathbf{u_j} \cdot \mathbf{u_l} + \sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{z_{kj}u_j} \cdot \mathbf{u_l} + \sum_{\mathbf{j=p+1}}^{\mathbf{d}} \mathbf{b_j u_j} \cdot \mathbf{u_l} \right) \\
&= \frac{\mathbf{2}}{\mathbf{n}} \left(- \sum_{\substack{j=1 \\ j \neq l}}^{\mathbf{d}} (\mathbf{x_k'u_j})\mathbf{u_j} \cdot \mathbf{u_l} + (\mathbf{x_k'u_l})\mathbf{u_l} \cdot \mathbf{u_l} + \sum_{\substack{j=1 \\ j \neq l}}^{\mathbf{p}} \mathbf{z_{kj}u_j} \cdot \mathbf{u_l} + \mathbf{z_{kl}u_l} \cdot \mathbf{u_l} + \sum_{\mathbf{j=p+1}}^{\mathbf{d}} \mathbf{b_j u_j} \cdot \mathbf{u_l} \right) \\
&= \frac{\mathbf{2}}{\mathbf{n}} \left(- \sum_{\substack{j=1 \\ j \neq l}}^{\mathbf{d}} (\mathbf{x_k'u_j}) \cdot \mathbf{0} + (\mathbf{x_k'u_l}) + \sum_{\substack{j=1 \\ j \neq l}}^{\mathbf{p}} \mathbf{z_{kj}} \cdot \mathbf{0} + \mathbf{z_{kl}} + \sum_{\mathbf{j=p+1}}^{\mathbf{d}} \mathbf{b_j} \cdot \mathbf{0} \right) \\
&= \frac{\mathbf{2}}{\mathbf{n}} \left(-(\mathbf{x_k'u_l}) + \mathbf{z_{kl}} \right)
\end{aligned}$$

Ocupando el criterio de primera derivada tenemos que el punto de inflexión es

$$\begin{aligned}
\frac{\mathbf{2}}{\mathbf{n}} \left(-(\mathbf{x_k'u_l}) + \mathbf{z_{kl}} \right) &= 0 \\
-(\mathbf{x_k'u_l}) + \mathbf{z_{kl}} &= 0 \\
\mathbf{z_{kl}} &= \mathbf{x_k'u_l}.
\end{aligned}$$

Ahora, la segunda derivada de la función de costo con respecto a z_{kl} es

$$\frac{\partial L}{\partial^2 z_{kl}} = \frac{2}{n}. \tag{1.5}$$

Por lo tanto, como la segunda derivada es no negativa podemos concluir usando el criterio de segunda derivada que el punto de inflexión $\mathbf{z_{kl} = x'_k u_l}$ es un mínimo.

Realizando la misma metodología para encontrar el punto que minimiza la función de costo para b_j, tenemos que la primera derivada de la función de costo (1.4) con respecto a b_h donde $h \in \{p+1, p+2, \cdots, d\}$ es

$$\frac{\partial \mathbf{L}}{\partial \mathbf{b_h}} = \frac{\mathbf{2}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \left(\sum_{\mathbf{j=1}}^{\mathbf{d}} (\mathbf{x_i' u_j}) \mathbf{u_j} - \sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{z_{ij} u_j} - \sum_{\mathbf{j=p+1}}^{\mathbf{d}} \mathbf{b_j u_j} \right) \cdot (-\mathbf{u_h})$$

Ocupando (1.1) tenemos que $u_j \cdot u_h = 0$ si $j \neq h$ y $u_j \cdot u_h = 0$ si $j = h$, por lo que podemos simplificar la derivada como

$$\begin{aligned}
\frac{\partial \mathbf{L}}{\partial \mathbf{b_h}} &= \frac{\mathbf{2}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \left(- \left(\sum_{\mathbf{j=1}}^{\mathbf{d}} (\mathbf{x_i' u_j}) \mathbf{u_j} \right) \cdot \mathbf{u_h} + \left(\sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{z_{ij} u_j} \right) \cdot \mathbf{u_h} + \left(\sum_{\mathbf{j=p+1}}^{\mathbf{d}} \mathbf{b_j u_j} \right) \cdot \mathbf{u_h} \right) \\
&= \frac{\mathbf{2}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \left(- \sum_{\mathbf{j=1}}^{\mathbf{d}} (\mathbf{x_i' u_j}) \mathbf{u_j} \cdot \mathbf{u_h} + \sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{z_{ij} u_j} \cdot \mathbf{u_h} + \sum_{\mathbf{j=p+1}}^{\mathbf{d}} \mathbf{b_j u_j} \cdot \mathbf{u_h} \right) \\
&= \frac{\mathbf{2}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \left(- \sum_{\substack{j=1 \\ j \neq h}}^{\mathbf{d}} (\mathbf{x_i' u_j}) \mathbf{u_j} \cdot \mathbf{u_h} - (\mathbf{x_i' u_h}) \mathbf{u_h} \cdot \mathbf{u_h} + \sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{z_{ij} u_j} \cdot \mathbf{u_h} + \sum_{\substack{j=1 \\ j \neq h}}^{\mathbf{d}} \mathbf{b_j u_j} \cdot \mathbf{u_h} + \mathbf{b_j u_h} \cdot \mathbf{u_h} \right) \\
&= \frac{\mathbf{2}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \left(- \sum_{\substack{j=1 \\ j \neq h}}^{\mathbf{d}} (\mathbf{x_i' u_j}) \cdot \mathbf{0} - (\mathbf{x_i' u_h}) + \sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{z_{ij}} \cdot \mathbf{0} + \sum_{\substack{j=1 \\ j \neq h}}^{\mathbf{d}} \mathbf{b_j} \cdot \mathbf{0} - \mathbf{b_j} \right) \\
&= \frac{\mathbf{2}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \left(-\mathbf{x_i' u_h} + \mathbf{b_h} \right) = -\frac{\mathbf{2}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{x_i' u_h} + \mathbf{2 b_h} = \mathbf{-2 \bar{x}' u_h} + \mathbf{2 b_h}.
\end{aligned}$$

Ocupando el criterio de primera derivada tenemos que el punto de inflexión es

$$\begin{aligned}
\mathbf{-2 \bar{x}' u_h} + \mathbf{2 b_h} &= 0 \\
\mathbf{2 b_h} &= \mathbf{2 \bar{x}' u_h} \\
\mathbf{b_h} &= \mathbf{\bar{x}' u_h}.
\end{aligned}$$

Ahora, la segunda derivada de la función de costo con respecto a b_h es

$$\frac{\partial \mathbf{L}}{\partial \mathbf{^2 b_h}} = \mathbf{2}.$$

Por lo tanto, como la segunda derivada es no negativa podemos concluir usando el criterio de segunda derivada que el punto de inflexión $\mathbf{b_h = \bar{x}' u_h}$ es un mínimo.

Ocupando los dos resultados anteriores, es decir, $z_{ij} = x_i' uj$ y $b_j = \bar{x}' u_j$ minimizan la función de

costo tenemos que

$$\begin{aligned}
\mathbf{x_i} - \hat{\mathbf{x}}_\mathbf{i} &= \sum_{\mathbf{j=1}}^{\mathbf{d}} (\mathbf{x_i'u_j})\mathbf{u_j} - \sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{z_{ij}u_j} - \sum_{\mathbf{j=p+1}}^{\mathbf{d}} \mathbf{b_j u_j} \\
&= \sum_{\mathbf{j=1}}^{\mathbf{d}} (\mathbf{x_i'u_j})\mathbf{u_j} - \sum_{\mathbf{j=1}}^{\mathbf{p}} (\mathbf{x_i'uj})\mathbf{u_j} - \sum_{\mathbf{j=p+1}}^{\mathbf{d}} (\bar{\mathbf{x}}'\mathbf{u_j})\mathbf{u_j} \\
&= \sum_{\mathbf{j=1}}^{\mathbf{p}} (\mathbf{x_i'u_j})\mathbf{u_j} + \sum_{\mathbf{j=p+1}}^{\mathbf{d}} (\mathbf{x_i'u_j})\mathbf{u_j} - \sum_{\mathbf{j=1}}^{\mathbf{p}} (\mathbf{x_i'uj})\mathbf{u_j} - \sum_{\mathbf{j=p+1}}^{\mathbf{d}} (\bar{\mathbf{x}}'\mathbf{u_j})\mathbf{u_j} \\
&= \sum_{\mathbf{j=p+1}}^{\mathbf{d}} (\mathbf{x_i'u_j})\mathbf{u_j} - \sum_{\mathbf{j=p+1}}^{\mathbf{d}} (\bar{\mathbf{x}}'\mathbf{u_j})\mathbf{u_j u_j} = \sum_{\mathbf{j=p+1}}^{\mathbf{d}} (\mathbf{x_i'u_j})\mathbf{u_j} - (\bar{\mathbf{x}}'\mathbf{u_j})\mathbf{u_j} \\
&= \sum_{\mathbf{j=p+1}}^{\mathbf{d}} (\mathbf{x_i'u_j} - \bar{\mathbf{x}}'\mathbf{u_j})\mathbf{u_j} = \sum_{\mathbf{j=p+1}}^{\mathbf{d}} ([\mathbf{x_i'} - \bar{\mathbf{x}}']\mathbf{u_j})\mathbf{u_j} = \sum_{\mathbf{j=p+1}}^{\mathbf{d}} ([\mathbf{x_i} - \bar{\mathbf{x}}]'\mathbf{u_j})\mathbf{u_j}.
\end{aligned}$$

Es decir, la "desviación" está en el espacio ortogonal de los componentes principales.
Por último, ocupando el resultado anterior podemos reescribir la función de costo original (1.3) como

$$\begin{aligned}
L &= \frac{1}{n}\sum_{i=1}^{n} ||\mathbf{x_i} - \hat{\mathbf{x}}_\mathbf{i}||^\mathbf{2} \\
&= \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \left|\left| \sum_{\mathbf{j=p+1}}^{\mathbf{d}} ([\mathbf{x_i} - \bar{\mathbf{x}}]'\mathbf{u_j})\mathbf{u_j} \right|\right|^\mathbf{2} \\
&= \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \sqrt{\sum_{\mathbf{j=p+1}}^{\mathbf{d}} ([\mathbf{x_i} - \bar{\mathbf{x}}]'\mathbf{u_j})\mathbf{u_j} \cdot \sum_{\mathbf{j=p+1}}^{\mathbf{d}} \mathbf{u_j^T}(\mathbf{u_j^T}[\mathbf{x_i} - \bar{\mathbf{x}}])}^{\mathbf{2}} \\
&= \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \left(\sum_{\mathbf{j=p+1}}^{\mathbf{d}} ([\mathbf{x_i} - \bar{\mathbf{x}}]'\mathbf{u_j})\mathbf{u_j} \cdot \sum_{\mathbf{j=p+1}}^{\mathbf{d}} \mathbf{u_j^T}(\mathbf{u_j^T}[\mathbf{x_i} - \bar{\mathbf{x}}]) \right) \\
&= \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \left(\sum_{\mathbf{j=p+1}}^{\mathbf{d}} \big([\mathbf{x_i} - \bar{\mathbf{x}}]'\mathbf{u_j} \big)^\mathbf{2} + \sum_{\mathbf{j=p+1}}^{\mathbf{d}} \sum_{\substack{\mathbf{k=p+1} \\ \mathbf{k \neq j}}}^{\mathbf{p}} ([\mathbf{x_i} - \bar{\mathbf{x}}]'\mathbf{u_j}) \underbrace{\mathbf{u_j u_k^T}}_{\mathbf{=0}} (\mathbf{u_k^T}[\mathbf{x_i} - \bar{\mathbf{x}}]) \right) \\
&= \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \left(\sum_{\mathbf{j=p+1}}^{\mathbf{d}} \big([\mathbf{x_i} - \bar{\mathbf{x}}]'\mathbf{u_j} \big)^\mathbf{2} \right) = \frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} \sum_{\mathbf{j=p+1}}^{\mathbf{d}} \mathbf{u_j^T}[\mathbf{x_i} - \bar{\mathbf{x}}][\mathbf{x_i} - \bar{\mathbf{x}}]'\mathbf{u_j} \\
&= \sum_{\mathbf{j=p+1}}^{\mathbf{d}} \mathbf{u_j^T} \left(\frac{\mathbf{1}}{\mathbf{n}} \sum_{\mathbf{i=1}}^{\mathbf{n}} [\mathbf{x_i} - \bar{\mathbf{x}}][\mathbf{x_i} - \bar{\mathbf{x}}]' \right) \mathbf{u_j} = \sum_{\mathbf{j=p+1}}^{\mathbf{d}} \mathbf{u_j^T S u_j} \quad \blacksquare
\end{aligned}$$

Es decir, la solución se puede obtener resolviendo un problema de valores propios. Esta propiedad ayuda a entender desde otro enfoque la metodología de PCA relacionándolo con la matriz de covarianzas.

2. Índice de Marginación en Nuevo León.

2.1. Introducción.

La marginación se refiere como el conjunto de problemas (desventajas) sociales de una comunidad o localidad y hace referencia a grupos de personas y familias. Parte importante de estudiar la marginación nace de poder encontrar lugares en los que la marginación es más alta para poder ayudar con practicas reguladoras para cada estado, es decir, la marginación es un problema para la economía de los lugares, entonces a menores índices de marginación se esperaría una economía más alta o estable.

2.2. Metodología

Considerando que la marginación es importante, entonces hay que determinar como medirla. El Consejo Nacional de Población (CONAPO) ha presentado una metodología para medir la marginación a nivel localidad. La marginación se puede reflejar en tu nivel educativo, el lugar en que vives y el ingreso que percibes, por lo que para poder medirla hay que considerar estos tres aspectos: **educación, vivienda e ingreso económico**.

Con la ayuda del Censo 2010 que realizó el Instituto Nacional de Estadística, Geografía e Informática es posible medir estos tres aspectos para poder posteriormente tener una única medida de marginación.
Para medir la educación consideramos dos indicadores, el primero se relaciona con la capacidad de las personas de leer y escribir un recado (analfabetas) y el segundo indicador se refiere a las personas que cursaron la primaria. Para la vivienda propuso 5 indicadores, los cuales exploran las condiciones de las viviendas: carencia de excusado, carencia de servicio de energía eléctrica, uso de agua entubada, número de cuartos por personas en la vivienda y uso de piso de tierra. Y para el ingreso económico un indicador, debido a que los datos del censo no presentan a un nivel localidad los ingresos económicos (por motivos de privacidad) se buscó un *proxy*. La disponibilidad de refrigerador se encuentra condicionada por el ingreso del que se dispone en las viviendas, ya sea por trabajo o transferencias monetarias o en especie.

En resumen, los 8 indicadores a considerarla para medir la marginación a nivel localidad son:

1. Porcentaje de población de 15 años o más analfabeta.
2. Porcentaje de población de 15 años o más sin primaria completa.
3. Porcentaje de ocupantes en viviendas particulares habitadas sin drenaje ni excusado.
4. Porcentaje de ocupantes en viviendas habitadas particulares habitadas sin energía eléctrica.
5. Porcentaje de ocupantes en viviendas particulares habitadas sin agua entubada.
6. Porcentaje viviendas particulares habitadas con algún nivel de hacinamiento.
7. Porcentaje de ocupantes en viviendas particulares habitadas con piso de tierra.
8. Porcentaje de viviendas particulares habitadas que no disponen de refrigerador.

Para ver los detalles explícitamente de los cálculos conforme a las variables del Censo 2010 ver la sección de anexos, pero como sus nombres lo dice solo son porcentajes, es decir, el calculo se reduce a considerar la división de las personas/viviendas que si tienen una cierta características entre el número total de personas/viviendas.

Una vez calculado todos los índices anteriores, entonces para crear el índice de marginación tenemos que *reducir* los 8 indicadores económicos a un nuevo indicador, de tal manera que este indicador pueda explicar la mayor información de los otros 8. Una de la manera de hacerlo más sencilla sería crearlo a partir del promedio de estos 8 índices, pero este tiene muchas desventajas ya que no considera la relación de los índices entre si por lo que no sería muy útil.

Una metodología más sotisficada es el Análisis de Componentes Principales (PCA por sus siglas en inglés). **Este método de manera muy general consiste o tiene como objetivo que de un conjunto de datos con múltiples atributos poder simplificar la mayor de información posible en un número de atributos menores que los originales, es decir, se encarga de disminuir la dimensión de nuestros conjuntos de datos y esto provoca un porcentaje en la perdida de la información total.** De manera más técnica, consiste en realizar proyecciones de los datos en una base ortonormal donde si la base es de la misma dimensión que el número de atributos iniciales la perdida de información es nula. Una forma más sencilla de entender es considerar el Ejemplo 1.

Ejemplo: 1. ***(PCA y la fotografía).*** *Imagina que eres fotógrafo profesional. Tu trabajo es tomar una fotografía que pueda capturar la forma específica (altura, ancho y largo) de una mesa (Ver Figura 2.1). ¿Qué pasaría si tomas la fotografía desde la parte de arriba? ¿Cuál sería la mejor manera de hacerlo? Para tomar la mejor la fotografía tienes que considerar varios ángulos para hacerlo, si tomas de arriba no podrás saber la longitud de la mesa, pero si la tomas de un ángulo muy frontal posiblemente no podrás saber el ancho o el largo de la mesa.*

Entonces, pasa de tener 3 dimensiones al tomar la foto haces una proyección a dos dimensiones a este hecho se le llama reducción de dimensión a la hora de hacer PCA. Es decir, PCA busca el mejor ángulo para poder proyectar los datos un dimensiones más pequeñas.

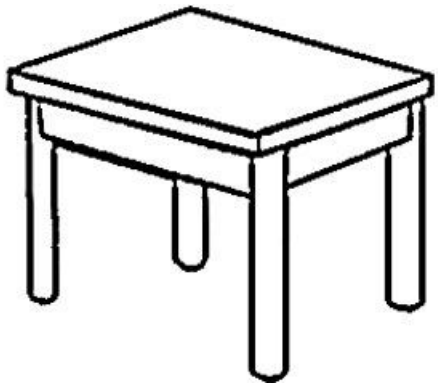

Figura 2.1: Ejemplo de mesa a fotografía.

Por lo tanto, se usará PCA con a los 8 indicadores económicos se procede a calcular el nuevo ´índice de marginación. El cuál sería proyectar nuestros indicadores en el primer componente, por definición este tendrá la mayor proporción de la información (varianza) de los datos. Cabe mencionar el PCA no es robusto a escalas, lo que se recomienda en la mayoría de los casos normalizar antes.

2.3. Replicación del Índice.

En esta sección replicamos la metodología explicada en la sección anterior de CONAPO pero solo al estado de Nuevo León. **Hacer este pequeño cambio cambiara un poco los índices, debido a que se estaría perdiendo información de la relación de los índices fuera del estado de Nuevo León. Cabe mencionar que el archivo que se nos paso del Censo 2010 (censonl.csv) esta desactualizado con la versión del índice calculado de CONAPO,** esta afirmación se basa a que para crear el índice 2 (de la lista anterior) se necesitaban ciertas columnas que no estaban en el archivo ni en el diccionario de datos. Para arreglar esto y poder replicar completamente la metodología de CONAPO se descargo el archivo del Censo 2010 más actualizado. Con este ajuste la diferencia en los cálculos ahora son mínimos o nulos [2].

Nuevo León cuenta con un total de 5262 localidades registradas de las cuales 3224 no tiene ningún registro para calcular los índices. Estas localidades tiene su información confidencial por lo que no podemos realizar un método de imputación ya que estarías sesgando al indicador por que representan más del 50 % de las localidades, pero estas 3224 localidades no representan ni el 1 % de la población total por lo que podemos ignorarlas para el calculo (esto mismo se hace en el documento de CONAPO). Entonces, el calculo de los índices socioeconómicos se realizo a 2037 localidades, de las cuales solo 1 tenía un valor nulo por lo que se procedió a imputar con la media del municipio considerando el peso que representa la localidad.

Antes de aplicar PCA a nuestro conjunto de datos (2037×8) los estandarizamos, es decir, a cada observación le restamos la media y esto lo dividimos entre su desviación estándar. Observando la figura 2.2 podemos observar que **la correlación más alta se dan entre el índice de viviendas sin energía eléctrica y viviendas que no disponen de refrigerador con 0.72**, lo cual tiene sentido debido a que si no se cuenta con energía eléctrica esto implica que no tenga un refrigerado ya que no funcionara.

Figura 2.2: Matriz de correlación de los índices socieconomicos.

La varianza explicada por el primer componente principal es de 37,14 % (ver Cuadro 2.1) que a diferencia del reportado por CONAPO hay una diferencia de 9 puntos abajo, esto se explica por el hecho de que no se considera los demás estados.

Componentes principales	λ_i	Varianza explicada	Varianza acumulada
1	2.97	37.14	37.14
2	1.28	15.93	53.07
3	0.25	11.84	64.91
4	0.44	9.63	74.54
5	0.95	8.83	83.37
6	0.64	8.03	91.4
7	0.71	5.47	96.87
6	0.77	3.13	100

Cuadro 2.1: Valores propios de la matriz de correlaciones y porcentaje de varianza explicada, 2010

Entonces ocupemos los valores del primer componente principal (ver Cuadro 2.2) para calcular el índice de marginación por localidad, para ello solo basta con multiplicar para cada localidad los coeficientes encontrados por los índices socioeconomicos.

Indicador socioeconómico	coeficientes
Porcentaje de población de 15 años o más analfabeta.	0.291
Porcentaje de población de 15 años o más sin primaria completa.	0.321
Porcentaje de ocupantes en viviendas particulares habitadas sin drenaje ni excusado.	0.36
Porcentaje de ocupantes en viviendas habitadas particulares habitadas sin energía eléctrica.	0.425
Porcentaje de ocupantes en viviendas particulares habitadas sin agua entubada.	0.312
Porcentaje viviendas particulares habitadas con algún nivel de hacinamiento.	0.17
Porcentaje de ocupantes en viviendas particulares habitadas con piso de tierra.	0.385
Porcentaje de viviendas particulares habitadas que no disponen de refrigerador.	0.478

Cuadro 2.2: Coeficientes de la primera componente principal por indicador socioeconómicos, 2010

Una vez calculado el valor del índice para las 2037 localidades, se procedió a clasificarlas en uno de los cinco grupos proporcionados por CONAPO (ver el Cuadro 2.3). Usando estos límites por grupo tenemos la distribución de las localidades que se ve en la Figura 2.3.

Grado de marginación	Límites del IM
Muy bajo	[−1,83197, −1,32309]
Bajo	(−1,32309, −1,06870]
Medio	(−1,06870, −0,81425]
Alto	(−0,81425, 0,71231]
Muy alto	(0,71231, 8,34515]

Cuadro 2.3: Clasificación del grado de marginación.

El ejemplo más emblemático es el de Nuevo León, aunque se ubica entre las entidades con muy bajo grado de marginación, por localidad se observa que de sus 2 037 localidades, 124 reportan un

grado de marginación muy alto (6,1 %), 1030 tienen grado alto (50,6 %), 404 están con grado medio (19,8 %), 321 en grado de marginación bajo (15,8 %) y 158 tienen grado de marginación muy bajo (7,8 %). Esta distribución de las localidades por grado de marginación se parecen bastante a los reportados por CONAPO usando su índice, por lo que podemos concluir que nuestro índice es muy cercano al que reporta CONAPO para Nuevo León.

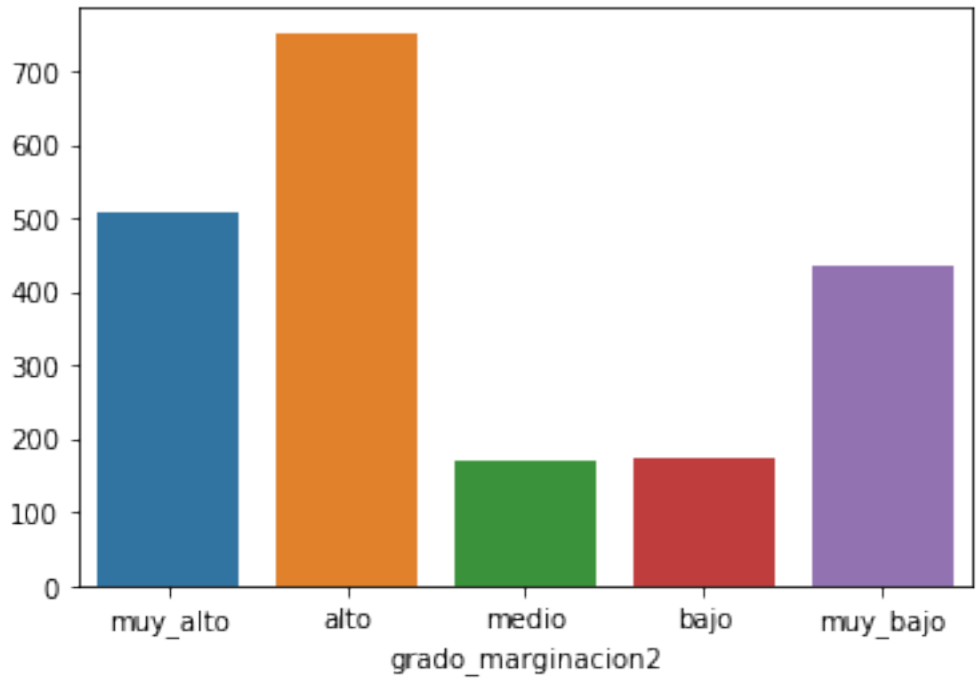

Figura 2.3: Localidades por nivel de marginación

Y por último, **algo importante que resaltar es que la localidad de San Pedro Garza García se encuentra con un índice de marginación muy bajo, están en el top 20 menos marginados lo que hace pensar que el índice es bueno**. Aunque creo que se hubiera esperado que esta localidad estuviera en el top 5 ya que es una de las localidades más importantes.

2.4. Propuesta para un nuevo índice.

Los dos planteamientos planteo para mejorar el índice de CONAPO sería: **ampliar las variables del ámbito individual y integrar aspectos generales.** Con el primer planteamiento me refiero a que las aspectos que considero el CONAPO no es el adecuado o faltan profundizar más. Por ejemplo, el indicador *proxy* sobre el ingreso fue considerando si las viviendas tenían refrigerador, pero observamos que esta variable estaba muy relacionado con si la vivienda contaba electricidad y por lo que pienso que no es un buen estimador del ingreso. Entonces yo sugeriría considerar como el porcentaje de personas que se dedican al sector primario, secundario y terciario, o incluso considerar si la vivienda cuenta con internet o uso del automóvil sería un mejor estimador del ingreso pero tengo entendido que en el Censo 2010 no se tenía esa información.

Por otro lado, el segundo planteamiento me refiero a considerar aspectos generales de las localidades y así complementar el índice propuesto. **Con aspectos más generales me refiero por ejemplo así en la localidad existen bancos, hospitales, escuelas, centros comerciales, ¿cuántos**

existen? Y si no hay en la localidad, entonces cuál es la distancia a la localidad más cerca que los tenga. De cierta manera considerar estos aspectos dan un enfoque general de la población en general, porque contar con los servicios más utilizados (salud, finanzas, etc) en los últimos años esta relacionado con el grado de marginalidad de cada localidad.

Una manera de poder crear este índice sería considerando el Directorio Estadístico Nacional de Unidades Económicas (DENUE) el cual realiza el INEGI. Este directorio cuenta con cada unidad económica del país, de donde podríamos calcular el número el número de hospitales, escuelas, bancos, centros comerciales o la distancia más cerca a estos lugares por localidad. Y con esto podrías realizar un nuevo índice para complementar a los anteriores. **Otro aspecto general que se puede considerar sería la delincuencia**, este índice se esperaría que en los lugares más marginales y con los niveles más marginales tengan un índice que delincuencia menor en cambio los lugares con niveles moderados de marginación tendrían un mayor más grande. Este índice se puede construir con los reportes diarios de delitos estatales. Y por último, **podríamos considerar el porcentaje de calles que se encuentran pavimentadas o rusticas**, esto nos daría un enfoque de esta la construcción de las localidades lo cual estaría muy relacionado debido a que es más difícil pavimentar en lugares con más marginación es más difícil y caro llevar el material, trabajadores y maquinaria que en lugares menos marginales.

Un problema por el cuaĺ no se calcularon el nuevo índice propuesto fue debido a que actualmente no se encuentra las limitaciones de las localidades de México, es decir, no se tienen geodelimitadas las localidades. Esto hace imposible calcular por ejemplo el número de hospitales en cada localidad, ya que no sabríamos saber a que localidad le pertenece.

2.5. Conclusiones.

En general podría decir que el índice que propone CONAPO es adecuado para clasificar el grado de marginación. Pero algo importante a notar es que no es están bueno para aspectos más específicos. Como el presupuesto de cada municipio depende de este índice creo que debería comprender más aspectos para que este mejor. Lo anterior se debe a la clasificación de la localidad de San Pedro Garza García ya que no se encuentra en un nivel más bajo.

Por otro lado, creo que este índice no es tan informativo. Ya que no podríamos saber que factor es que tendría que mejorar para tener una mayor clasificación. Por lo que si estamos interesados en políticas públicas para disminuir el grado de marginación en las localidades sería complicado buscar propuestas para hacerlo.

3. Eigenfaces

En este ejercicio nos enfocamos al conjunto de datos Labelled Faces in the Wild, que consiste en fotografías de rostros recolectados de internet y contenido en `sklearn`. Solo se consideraron aquellas personas que tienen al menos 70 fotografías de su rostro en su tamaño original de la imagen (125×94).

3.1. a) Primeros dos componentes.

El número total de imágenes analizadas fueron 1288, algo importante a resaltar es la proporción de fotografías que se encuentra en la muestra por cada personaje. Observamos que existe una muestra

desbalanceada lo que puede influir en los errores, dándoles más prioridad a George W Bush por ser el que tiene más imágenes a la muestra (Ver Cuadro 3.1).

Nombre	Número de Fotografia
George W Bush	530
Colin Powell	236
Tony Blair	144
Donald Rumsfeld	121
Gerhard Schroeder	109
Ariel Sharon	77
Hugo Chavez	71

Cuadro 3.1: Número de fotografías por personaje en la muestra.

Se realizó una partición de nuestros datos en un conjunto de entrenamiento (%80) y conjunto de prueba. Posteriormente llevamos acabo la metodología de PCA para encontrar los eigenfaces en nuestros conjunto de entrenamiento. Observando los dos primeros componentes principales podemos dar una .explicación de estos", el primer componente principal se enfoca en los rasgos del rostro: los ojos, nariz y boca. Y el segundo componente principal se enfoca en el fondo de la imagen.

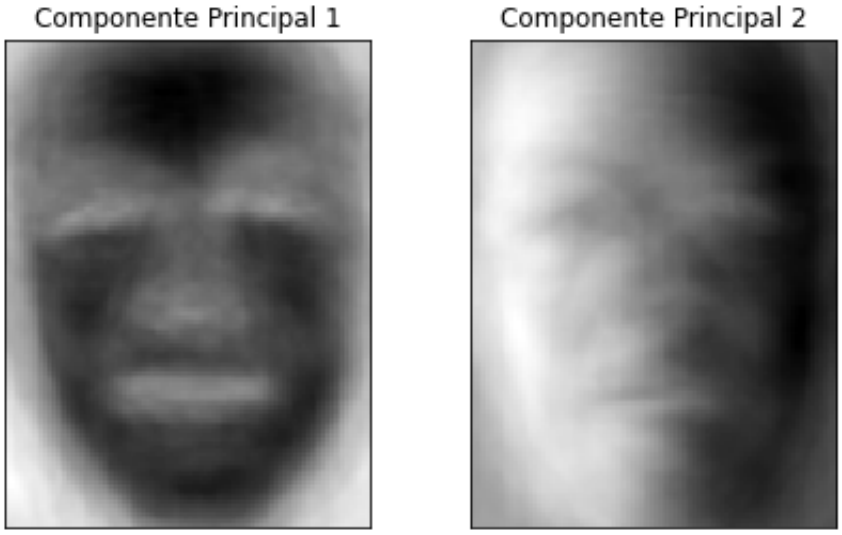

Figura 3.1: Primeros 2 componentes principales.

Por otro lado, la varianza explicada por estos dos componentes es aproximadamente 38 %, es un valor muy bajo por lo que se sugeriría usar un número mayor de componentes principales para tener mejores resultados en cualquier análisis.

3.2. Ubicación de cada individuo.

Ahora consideremos las proyecciones de los datos en los primeros dos componentes principales encontrados.

Al graficarlos en el plano cartesiano no podemos tener una segmentación de los datos, es decir, se esperaría que podamos separar las proyecciones de los datos en 8 grupos diferentes y cada grupo significará una persona (Ver Figura 3.2).

Las dos razones que puedo dar por la cual no se observa lo esperado son: la primera se puede deber a que como los dos componentes principales utilizados no explican un porcentaje alto de la varianza esto hace que no se este recopilando mucha información de los datos para obtener una segmentación por cada personaje. Y la segunda es debido a que no exista una segmentación como nosotros la vemos, es decir, en lugar de separar por persona los componentes principales estén segmentando por características del rostro: ojos grandes, nariz corpulenta, uso de anteojos, etc.

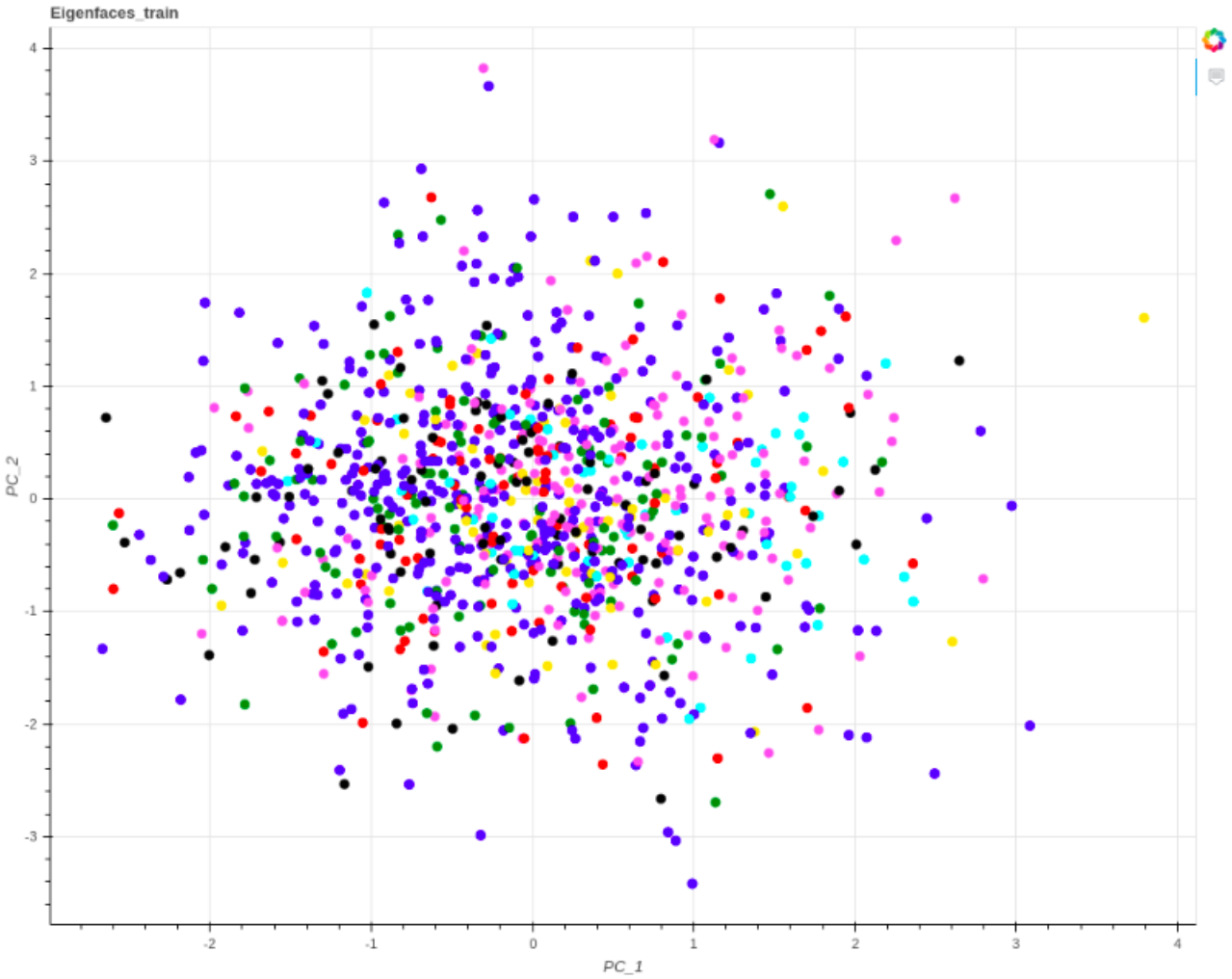

Figura 3.2: Proyecciones de los datos en los 2 primeros componentes principales.

Ahora proyectemos los datos de prueba en la figura 3.2 y veamos algunos ejemplos de qué fotografía es el más cercana. En la figura 3.3 se pueden observar la fotografía 125 y 15 de los datos de entrenamiento, cada una se proyecta en los primeros dos componentes y se encuentra la fotografía más cerca (considerando la distancia euclidiana) de los primeros dos componentes. La primera imagen que muestra a Gerhard Schoreder podemos ver que la imagen más cerca no es de el sino de Colin Powell, en cambio la segunda imagen de George W Bush observamos que la imagen más

cerca si le pertenece a el.

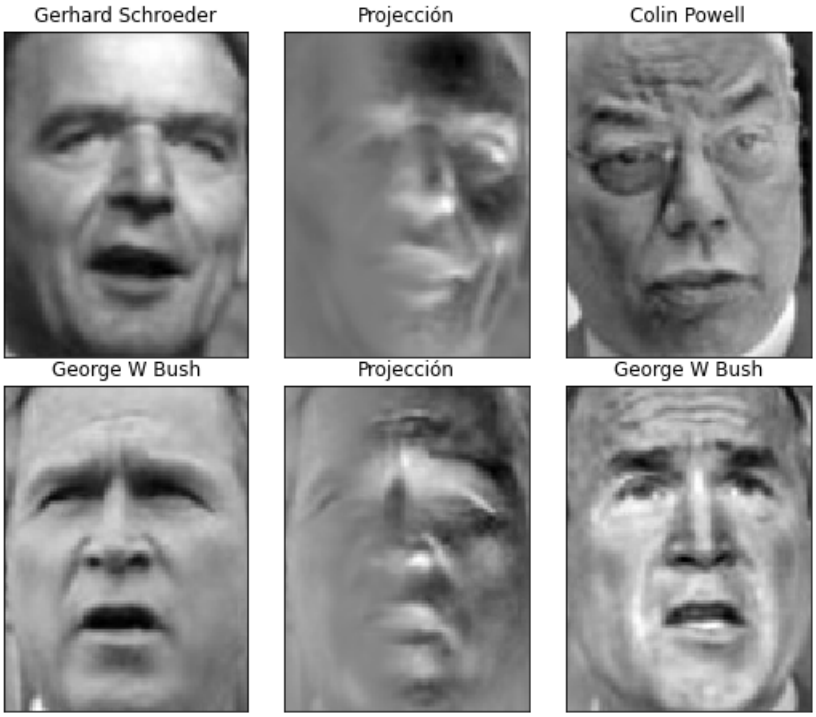

Figura 3.3: Proyección de los datos de prueba en dos componentes principales.

Si buscamos la imagen más cerca a cada una de las proyecciones de las imagenes de prueba tendríamos que 24,8 % de las proyecciones estaría cerca a una imagen con el personaje que le corresponde, lo que implica que en el 75 % de las veces no estaría ubicando bien a los individuos.

3.3. Identificación de la persona en la imagen.

El objetivo de esta sección es poner clasificar las imágenes según el personaje que aparece en estas. Para ello ocupamos la metodología del vecino más cercano, utilizando la distancia euclidiana como medida de disimilaridad. En la sección anterior se puede interpretar como el vecino más cercano considerando solo dos componentes principales, y ya vimos que no generan buenos resultados con pocos componentes.

Tengamos en cuenta que el número de componentes principales a ocupar en las proyecciones y el número de vecinos más cercano son parámetros que afectan en el los errores de clasificación. Estos parámetros son muy influyentes en la clasificación de las imágenes, por lo que es importante saber que combinación de parámetros elegir.

Para ello, un enfoque para saber la cantidad de componentes principales a ocupar es tomando en cuenta la varianza acumulada por los componentes principales. Si consideramos los primeros 50 componentes principales tenemos aproximadamente 80 % de la varianza de los datos explicada, si consideramos los primeros 100 componentes principales tenemos aproximadamente 90 % de la varianza. Entonces si consideramos 50 ya sería "sucifiente"para tener un buen predictor.

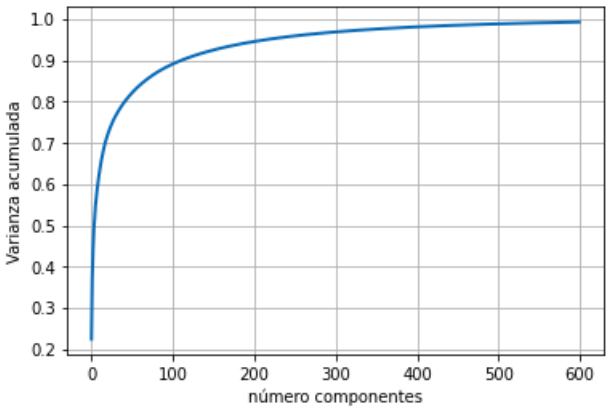

Figura 3.4: Varianza acumulada explicada por los componentes principales.

La segunda manera de abordar la incertidumbre que proporcionan los parámetros es considerar es considerar diversas combinaciones de parámetros y considerar la combinación con el error más pequeño, utilizando validación cruzada. Esta metodologías conocida como Hyperparameter tuning methods se enfocan básicamente en configurar de antemano los hiperparámetro de los algoritmos de aprendizaje automático para obtener un rendimiento predictivo óptimo mediante un procedimiento de ajuste. [3] Entonces, **ocupando lo anterior con la ayuda de la librería sklearn pudimos encontrar que ocupando** $p = 60$ **y** $n\_neighbors = 6$ **es la combinación de parámetros con el menor error de todas las combinaciones que se probaron, el error utilizando validación cruzada fue de 0.703 (sin utilizar validación cruzada fue de 0.846)**.

```
from sklearn.pipeline import Pipeline
from sklearn.model_selection import GridSearchCV

pca = PCA(svd_solver='randomized', whiten=True).fit(X_train)
knn = KNeighborsClassifier(p=2)

pipe = Pipeline(steps=[('pca', pca), ('knn', knn)])

param_grid = {
'pca__n_components': [5, 15, 45, 60, 90, 120, 135, 150, 120],
'knn__n_neighbors': [1, 3, 6, 9, 10, 12, 13]
}

search = GridSearchCV(pipe, param_grid, n_jobs=-1)
search.fit(X_train, y_train)
```

Ahora, considerando la combinación anterior **el error obtenido en el conjunto de prueba fue de 0.732. Por lo que podemos concluir que el ajuste del modelo es adecuado.** En la Figura 3.5 podemos observar la imagen sin modificaciones, normalizada, las proyecciones en los 60 componentes principales y el vecino más cercano con la probabilidad de ser clasificado con ese personaje. Es decir, en la primeras imágenes observamos al ex-presidente George W. Bush el cual el algoritmo del vecino más cercano lo clasificó como George W. Bush con una probabilidad de 1. En las imágenes posteriores observamos al ex-presidente alemán Gerhard Schroeder el cuál el cual el algoritmo del vecino más cercano lo clasificó como Gerhard Schroeder con una probabilidad de 0.5, es decir, de los otros 7 personajes tiene el resto de la probabilidad.

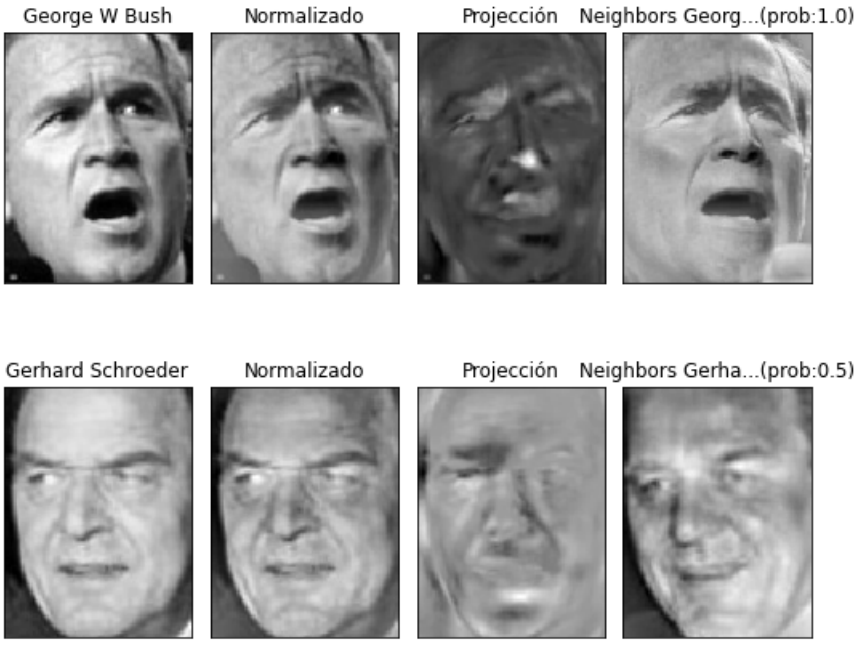

Figura 3.5: Identificación utilizando KNN.

3.4. Prevención de casos.

Una manera de disminuir el error o posibles errores de clasificación es considerar un umbral más estricto en el algoritmo del vecino más cercano. Nosotros podemos obtener la probabilidad de cada personaje cuando se quiere clasificar una imagen, entonces si observamos que la probabilidad máxima de todos los personajes es menor a 50 % entonces podríamos clasificar la nueva imagen como: imposible clasificar.
En la figura 3.6 podemos observar la imagen del ex presidente de México Vicente Fox, el cual si lo proyectamos en los 60 componentes principales y usamos buscamos los vecinos más cercanos, tendríamos que la probabilidad más alta la tiene Colin Powell por lo que con nuestra regla de decisión a esta nueva imagen la clasificaríamos como: imposible de clasificar.

Figura 3.6: Vicente Fox clasificación: imposible de clasificar.

Otro enfoque sería considerar las distribución de la distancias de los 6 vecinos más cercanos. Y si la distancia de la nueva a sus 6 vecinos más cercanos no cae en los 2 y 3 cuantiles podríamos considerar que esa foto no es de alguien de la muestra. Pero este enfoque es equivalente a obtener las probabilidades que proporciona el algoritmo de vecino más cercanos programado en `sklearn`.

4. Problema 4

Dado un conjunto de datos centrados $X_{n\times d}$, vimos que hacer PCA, es realizar la descomposición espectral de la matriz de covarianzas muestral, que puede estimarse como $S = X'X$ (omitimos el coeficiente $n-1$). Ahora, considera la matriz $K_{n\times n} = \mathbf{XX'}$.

4.1. Realizar PCA utilizando la matriz S es equivalente a hacerlo en con la matriz K

Recordemos la definición de valor propio y vector propio de una matriz.

Definición: 1. *Sea A una matriz de $n \times n$. El escalar λ es un valor propio de A si el vector $x(x \neq 0)$ se cumple*

$$A\mathbf{x} = \lambda\mathbf{x}.$$

El vector x es el vector propio de A correspondiente a λ. [1]

Ocupando la Definición 1, sea λ un valor propio de la matriz $\mathbf{S}$ y $\mathbf{u}$ su vector propio correspondiente, entonces

$$\begin{aligned}
\mathbf{Su} &= \lambda\mathbf{u} \\
\mathbf{X'Xu} &= \lambda\mathbf{u} \\
\mathbf{XX'Xu} &= \mathbf{X}\lambda\mathbf{u} \\
\mathbf{XX'}(\mathbf{Xu}) &= \lambda(\mathbf{Xu}) \\
\mathbf{K}(\mathbf{Xu}) &= \lambda(\mathbf{Xu}).
\end{aligned}$$

Por lo anterior, podemos concluir que λ también es valor propio de la matriz K con vector propio correspondiente Xu. También veamos que,

$$\begin{aligned}
\mathbf{Su} &= \lambda\mathbf{u} \\
\mathbf{X'Xu} &= \lambda\mathbf{u} \\
\mathbf{u^T X'Xu} &= \mathbf{u^T}\lambda\mathbf{u} \\
||\mathbf{Xu}||^{\mathbf{2}} &= \lambda\mathbf{u^T u} = \lambda. \Rightarrow ||\mathbf{Xu}|| = \sqrt{\lambda}.
\end{aligned}$$

Entonces podemos concluir que el vector propio de λ normalizado para la matriz K es $\frac{Xu}{||Xu||} = \lambda^{-1/2}Xu$. Y de igual manera, sea λ valor propio de K y $\mathbf{v}$ su vector propio correspondiente, entonces

$$\begin{aligned}
\mathbf{Kv} &= \lambda\mathbf{v} \\
\mathbf{XX'v} &= \lambda\mathbf{v} \\
\mathbf{X'XX'v} &= \mathbf{X'}\lambda\mathbf{v} \\
\mathbf{X'X}(\mathbf{X'v}) &= \lambda(\mathbf{X'v}) \\
\mathbf{S}(\mathbf{X'v}) &= \lambda(\mathbf{X'v}).
\end{aligned}$$

Entonces, podemos concluir que λ también es valor propio de la matriz S con vector propio correspondiente $X'v$. También veamos que,

$$\begin{aligned}
\mathbf{Kv} &= \lambda\mathbf{u} \\
\mathbf{XX'v} &= \lambda\mathbf{v} \\
\mathbf{v^T XX'v} &= \mathbf{v^T}\lambda\mathbf{v} \\
||\mathbf{X'v}||^\mathbf{2} = \lambda\mathbf{v^T v} = \lambda. &\Rightarrow ||\mathbf{X'v}|| = \sqrt{\lambda}.
\end{aligned}$$

Entonces podemos decir que el vector propio de λ normalizado para la matriz S es $\frac{X'v}{||X'v||} = \lambda^{-1/2\prime}X'v$. **Por lo tanto, se ha demostrado que** $(\lambda^{-1/2})Xu, \lambda$ **es un par de vector propio normalizado y valor propio de la matriz K, y a su vez,** $(\lambda^{-1/2})X^Tv, \lambda)$ **es un par de vector propio normalizado y valor propio de la matriz S, donde u y v son vectores propios de S y K respectivamente. Y lo anterior implica que realizar PCA en S es equivalente a realizarlo ocupando la matriz K.**

4.2. Ejemplo de PCA con la matriz **K**.

Para observar que hacer PCA en la matriz **K** es equivalente a hacerlo en la matriz **S**, procedemos a obtener los valores propios de ambas matrices y compararemos los primeros 5 componentes principales en el conjunto de datos de las imágenes LFW de la sección anterior. Realizando esto ocupando la librería `sklearn` obtenemos los eigenfaces que se observan en la figura 4.1. En ella podemos observar los primeros eigenfaces obtenidos con la matriz **S** en la parte de arriba y en la parte de abajo los eigenfaces obtenidos con la matriz **K**. Lo cual podemos observar claramente que son los mismos. Al tomar el tiempo de ejecución vimos un tiempo más pequeño cuando se ocupo la matriz **K**.

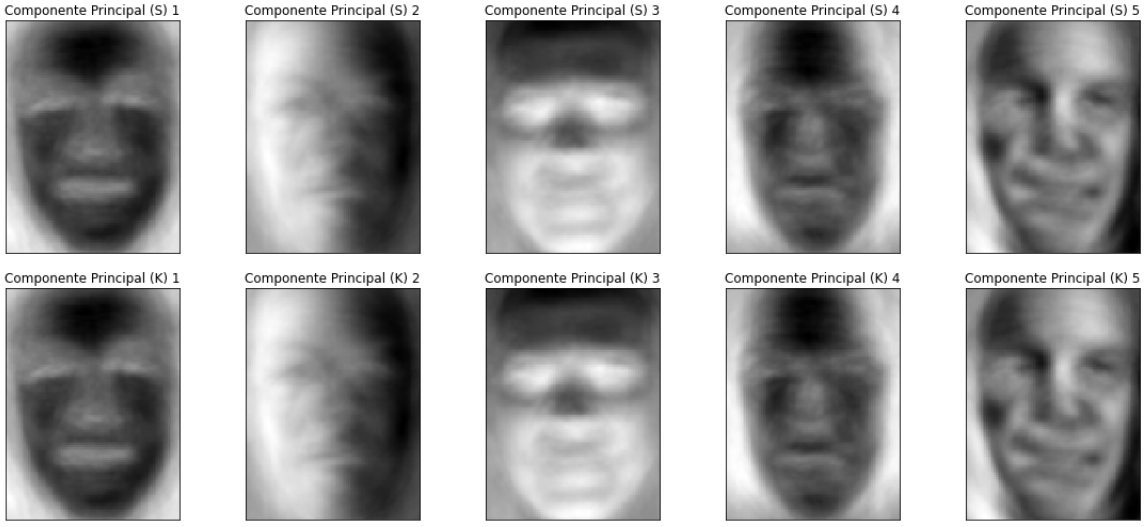

Figura 4.1: Eigenfaces usando la matriz **S** y **K** respectivamente.

En el caso cuando el número de registros n es más grande que el número de variables del conjunto de datos, es mejor recomendable trabajar con la matriz **S**, ya que tendrá dimensiones más pequeñas que la matriz **K**. Ahora, en el caso que el número de registros sea menor que el número de variables (esto es muy común en áreas como la genética) entonces es recomendable utilizar la matrizz **K**

ya que será de un tamaño más pequeño y esto facilitaría los cálculos para encontrar los valores y vectores propios. Esto es lo que paso en este ejemplo, en donde el número de registros era mucho más pequeño que las variables del conjunto de datos, ya que el tiempo de ejecución fue menor cuando se utilizó la matriz **K** para encontrar los eigenfaces.

5. Anexos.

5.1. Códigos.

Todos los códigos utilizados para estos resultados se pueden encontrar en mi página personal de Gitgub: Enriquesec. En el repositorio Ciencia_de_Datos/Tareas/Tareas2/solution.ipynb

5.2. Calculos de los Indicadores socioeconómicos.

Cómo ya se menciono en la sección 2.3 el conjunto de datos que se nos proporciono estaba desactualizado, por lo que se descargo de la página https://www.inegi.org.mx/programas/ccpv/2010/?ps=microdatos. Para calcular los 8 índices socioeconomicos se ocuparon las siguientes columnas y formulas:

- Porcentaje de población de 15 años o más analfabeta.
  - Población de 15 años o más $\left(P_i^{15+}\right)$. Columna: **P_15YMAS**.
  - Población de 15 años o más analfabeta $\left(P_i^{anal}\right)$: Columna: $\mathbf{P15YM_AN}$.

$$\text{Calculo del índice } I_{i,1} = \frac{P_i^{anal}}{P_i^{15+}} * 100.$$

- Porcentaje de población de 15 años o más sin primaria completa.
  - Población de 15 años o más sin escolaridad $\left(P_i^{15+se}\right)$. Columna: **P15YM_SE**.
  - Población de 15 años o más con primaria incompleta $\left(P_i^{15+pin}\right)$. Columna: **P15PRI_IN**.
  - Población de 15 años o más con primaria completa $\left(P_i^{15+pco}\right)$. Columna: **P15PRI_CO**.
  - Población de 15 años o más con secundaria incompleta $\left(P_i^{15+sse}\right)$. Columna: **P15SEC_IN**.
  - Población de 15 años o más con secundaria completa $\left(P_i^{15+sco}\right)$. Columna: **P15SEC_CO**.
  - Población de 18 años o más con educación pos-básica $\left(P_i^{18+pb}\right)$. Columna: **P18YM_PB**.

$$\text{Calculo del índice } I_{i,2} = \frac{P_i^{15+se} + P_i^{15+pin}}{P_i^{15+se} + P_i^{15+pin} + P_i^{15+pco} + P_i^{15+sse} + P_i^{15+sco} + P_i^{15+pb}}$$

- Porcentaje de viviendas particulares habitadas sin excusado.
  - Total de viviendas particulares habitadas$\left(P_i^{tvv}\right)$. Columna: **TVIVPARHAB**.
  - Total viviendas particulares habitadas que disponen de excusado o sanitario$\left(P_i^{tvv_exc}\right)$. Columna: **VPH_EXCSA**.

$$\text{Calculo del índice } I_{i,3} = \frac{P_i^{tvv} - P_i^{tvv\_exc}}{P_i^{tvv}}.$$

- Porcentaje de viviendas particulares habitadas sin energía eléctrica.
  - Viviendas particulares habitadas que no disponen de luz eléctrica$\left(P_i^{vv_sl}\right)$. Columna: **VPH_S_ELEC**.
  - Viviendas particulares habitadas que disponen de luz eléctrica$\left(P_i^{vv_cl}\right)$. Columna **VPH_C_ELEC**.

$$\text{Calculo del índice } I_{i,4} = \frac{P_i^{vv\_sl}}{P_i^{vv\_sl} + P_i^{vv\_cl}}.$$

- Porcentaje de viviendas particulares habitadas sin disponibilidad de agua entubada.
  - Viviendas particulares habitadas que no disponen de agua entubada en el ámbito de la vivienda$(P_i^{vv_sa})$. Columna: VPH_AGUADV.
  - Viviendas particulares habitadas que disponen de agua entubada en el ámbito de la vivienda$\left(P_i^{vv_cl}\right)$. Columna VPH_AGUAFV.

$$\text{Calculo del índice } I_{i,5} = \frac{P_i^{vv\_sa}}{P_i^{vv\_sa} + P_i^{vv\_ca}}.$$

- Promedio de ocupantes por cuarto en viviendas particulares habitados (ya calculado) $I_{i,6}$. Columna: **PRO_OCUP_C**.
- Porcentaje de viviendas particulares habitadas con piso de tierra.
  - Viviendas particulares habitadas con piso de tierra$\left(P_i^{vv_ct}\right)$. Columna: **VPH_PISOTI**.
  - Viviendas particulares habitadas con piso diferente de tierra$\left(P_i^{vv_st}\right)$. Columna: **VPH_PISODT**.

$$\text{Calculo del índice } I_{i,7} = \frac{P_i^{vv\_ct}}{P_i^{vv\_st} + P_i^{vv\_ct}}.$$

- Porcentaje de viviendas particulares habitadas que no disponen de refrigerador.
  - Viviendas particulares habitadas totales$\left(P_i^{tvv}\right)$. Columna: $TVIVPARHAB$.
  - Viviendas particulares habitadas que disponen de refrigerador$(P_i^{vv_cr})$. Columna: **VPH_REFRI**.

$$\text{Calculo del índice } I_{i,8} = \frac{P_i^{vv\_cr}}{P_i^{tvv}}.$$

Con lo anterior sería suficiente de calcular de los índices de las localidades de Nuevo León sin ningún problema.

Referencias

[1] Howard Anton y Chris Rorres. *Elementary Linear Algebra: Applications Version*. Eleventh. Wiley, 2014. ISBN: 9781118434413 1118434412 9781118474228 1118474228.

[2] INEGI. *Censo de Población y Vivienda 2010. Principales resultados por localidad (ITER)*. URL: `https://www.inegi.org.mx/contenidos/programas/ccpv/2010/doc/fd_iter_2010.pdf`.

[3] Philipp Probst, Bernd Bischl y Anne-Laure Boulesteix. *Tunability: Importance of Hyperparameters of Machine Learning Algorithms*. 2018. arXiv: `1802.09596 [stat.ML]`.