TP3 - Métodos numéricos y Optimización: SVD y reducción de la dimensionalidad

Martina Grünewald y Victoria Lynch

8 de Noviembre 2023

1. Introducción

Este informe de investigación se enfocó en dos aspectos clave dentro del campo de la ciencia de datos y el procesamiento de imágenes. En la primera sección, se exploraron los procesos de compresión de imágenes utilizando la Descomposición de Valores Singulares (SVD). Se analizó cómo la representación de baja dimensión influye en la similitud entre imágenes, empleando métricas de distancia y compresión con diversas dimensiones. Además, se investigó el valor mínimo de dimensiones necesario para comprimir una imagen sin exceder cierto porcentaje de error. La segunda sección se centró en la reducción de la dimensionalidad y los Cuadrados Mínimos, evaluando cómo el análisis en espacios de baja dimensión permite revelar estructuras subyacentes y mejorar la predicción de etiquetas. También se exploró la importancia de las dimensiones originales y su relación con los valores singulares. Estos experimentos proporcionaron información valiosa sobre la selección de dimensiones en problemas de análisis y predicción de datos. A través de estos estudios, se obtuvieron percepciones significativas que contribuyeron al avance de la ciencia de datos y la visualización de información.

2. Métodos numéricos utilizados

2.1. Descomposición SVD

La descomposición de valores singulares, o también llamada SVD por sus siglas en inglés es una técnica que se utiliza comúnmente en análisis de datos, procesamiento de imágenes, compresión de datos y en diversos algoritmos de aprendizaje automático. Esta consiste en descopmoner la matriz inicial en tres matrices más simples y así representar la matriz original de manera más compacta. Sea A la matriz a la cual se le quiere implementar SVD:

$$A = USV^T \tag{1}$$

donde A puede tener cualquier tamaño $m * n$ con $m = n$ (es decir que puede ser tanto una matriz cuadrada como no), S es una matriz diagonal de la forma $m*n$ (con ceros fuera de la diagonal) que contiene los valores singulares de nuestra matriz inicial A en orden descendente, U es una matriz ortogonal de tamaño $m*m$ que contiene los vectores singulares izquierdos (salen a partir de los valores singulares), y V^T es la traspuesta de una matriz ortogonal $n * n$ que contiene los vectores singulares derechos (también obtenible con los valores singulares).

2.2. Análisis de los componentes principales (PCA)

El Análisis de Componentes Principales (PCA) es una técnica estadística utilizada para reducir la dimensionalidad de un conjunto de datos, al mismo tiempo que conserva la mayor cantidad posible de la variabilidad presente en los datos originales.

Supongamos que tenemos un conjunto de datos XX con nn observaciones y pp variables. La idea central de PCA es encontrar un nuevo conjunto de variables no correlacionadas, llamadas componentes principales, que sean combinaciones lineales de las variables originales. Estas componentes principales están ordenadas de tal manera que la primera componente captura la mayor variabilidad presente en los datos, y las siguientes componentes capturan la mayor variabilidad restante, sujetas a la restricción de ser ortogonales entre sí.

El procedimiento para llevar a cabo PCA implica los siguientes pasos:

1. Centrar los datos: Restar la media de cada variable para que los datos estén centrados alrededor del origen.

2. Calcular la matriz de covarianza o la matriz de correlación, dependiendo del contexto.

3. Calcular los autovectores y autovalores de la matriz de covarianza o de la matriz de correlación.

4. Ordenar los autovectores por los autovalores asociados y seleccionar los autovectores correspondientes a las componentes principales deseadas.

La proyección de los datos originales sobre las componentes principales resultantes proporciona una representación reducida de los datos, lo que facilita la interpretación y el análisis de conjuntos de datos complejos.

3. Otras herramientas utilizadas

3.1. Distancia euclidiana

La distancia euclidiana es una métrica común en el espacio euclidiano y se utiliza para medir la recta más corta entre dos puntos en este espacio. La distancia euclidiana entre dos puntos en un espacio euclidiano se calcula con la norma L_2.
La norma 2 de un vector $v = v_1, v_2, ...v_n$ es:

$$||\mathbf{v}||_2 = \sqrt{v_1^2 + v_2^2 + \ldots + v_n^2} \tag{2}$$

La distancia euclidiana entre dos puntos p y q en el espacio n-dimensional es:

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \ldots + (p_n - q_n)^2} \tag{3}$$

En este trabajo se utilizó una función de python `euclidean_distances` para calcular dicha distancia. Esta función se proporciona en la biblioteca scikit-learn, que es ampliamente utilizada para tareas de aprendizaje automático y análisis de datos.

3.2. Norma de Frobenius

La norma de Frobenius se utiliza para calcular la magnitud o norma de una matriz cuadrada o rectangular. Se denota $\|A\|_F$ y su fórmula para calcularla es la siguiente:

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2} \tag{4}$$

donde A es la matriz a la cual se le está aplicando la norma, m es el número de filas de la matriz A y n es la cantidad de columnas. En este trabajo se utilizó la librería `numpy` para realizar la norma.

4. Desarrollo experimental

4.1. Compresión de imágenes

Este primer experimento consistió en trabajar con la compresión de imágenes utilizando un conjunto de 16 imágenes contenidas en un dataset. Cada una de estas imágenes, que estaban principalmente representadas como matrices de $p \times p$, se transformaron en vectores $x_i \in \mathbb{R}^{(p \times p)}$ con los cuales se formó una única matriz.

Parte 1: SVD
A esta matriz, que contenía los datos de las 16 imágenes, se le realizó la descomposición de valores singulares (SVD) para obtener una representación de baja dimensión dependiendo de la cantidad de valores singulares utilizados para esta.

Parte 2: Visualización de imágenes reconstruidas con reducción de dimensionalidad.

Se procedió a visualizar las imágenes reconstruidas utilizando la compresión con cierta cantidad de dimensiones. Se exploraron tanto las primeras k dimensiones como las últimas d dimensiones de la descomposición obtenida. Se realizó un análisis comparativo para identificar las diferencias entre estas representaciones y extraer conclusiones relevantes.

Parte 3: Medición de similaridad

Se empleó la compresión con diversos valores de k y d para medir la similaridad entre pares de imágenes en un espacio de baja dimensión. Se utilizó la métrica de distancia euclidiana para calcular la matriz de similitud entre las imágenes originales y las imágenes reconstruidas. Esto se consiguió mediante la función `euclidean_distances` de la biblioteca scikit-learn de Python. El objetivo era analizar cómo variaba la similaridad entre las imágenes al modificar el valor de d.

Parte 4: Determinación del valor mínimo de dimensiones

Se llevó a cabo un experimento para encontrar el número mínimo de dimensiones para la cual se puede realizar SVD, de tal manera que el error entre la imagen comprimida y la original no excediera el 10 % bajo la norma de Frobenius. Este proceso se realizó, principalmente para una única imagen seleccionada y luego, con la matriz V^T de la descomposición con este número óptimo de dimensiones, es utilizó para las demás imágenes del conjunto.

4.2. Reducción de la dimensionalidad y Cuadrados Mínimos

El experimento consistió en realizar un análisis de reducción de dimensionalidad y medidas de similaridad en un conjunto de datos representado por el archivo "dataset.csv". Inicialmente, el conjunto de datos fue transformado en una matriz X de dimensiones nxp, donde n es el número de muestras y p es la dimensión inicial del conjunto de datos.

Parte 1: Reducción de dimensionalidad

Se llevó a cabo un análisis de los componentes principale de la matriz X para reducir su dimensionalidad a d = 2, 4, 6, 20 y p. Se proyectaron los vectores de datos x al espacio de dimensión reducida $V_d^T x$ Se analizó la facilidad de realizar el análisis para cada valor de d y su relación con los valores singulares de X. Se extrajeron conclusiones pertinentes sobre la relación entre la dimensionalidad y los valores singulares.

Parte 2: Medición de similitud

Se determinó la similitud par-a-par entre las muestras en el espacio de dimensión X y en el espacio de dimensión reducida dd utilizando el análisis de componentes principales (PCA) y proyecciones al azar. Se compararon las medidas de similitud obtenidas y se analizaron las diferencias entre ellas.

Parte 3: Determinación de las dimensiones más representativas

Se identificaron las dimensiones originales más representativas del conjunto de datos p en relación con las dimensiones d obtenidas mediante la descomposición en valores singulares (SVD). Se utilizó un método específico para determinar las dimensiones más importantes y se analizó su relevancia en el conjunto de datos.

Parte 4: Modelado lineal

Se utilizó la descomposición en valores singulares (SVD) para encontrar el vector β que minimiza la norma $\|X\beta - y\|_2$, donde y es un vector de etiquetas o variable dependiente. Se identificó la dimensión d que mejoró la predicción del modelo lineal para $\hat{y}$, minimizando el error de predicción.

El objetivo principal fue comprender la estructura y la distribución de las muestras en un espacio de alta dimensión, así como la relación entre la dimensionalidad reducida y los valores singulares, con el fin de facilitar un análisis más eficiente y una mejor comprensión de los datos presentes en el conjunto "dataset.csv".

5. Resultados y discusiones

Luego de llevar a cabo los diferentes experimentos previamente explicados para ambos modelos, se procedió a realizar un análisis de los resultados obtenidos.

5.1. Ejercicio 1: Compresión de imágenes

En este primer experimento, dividido en cuatro partes, se pudo llegar a diferentes conclusiones.

La primera parte, la que consistía en realizar la descomposición de valores singulares (SVD), se compararon diferentes reconstrucciones de la matriz original (la de las fotos), pero con diferentes valores para k; siendo k la cantidad de primeros valores singulares utilizados para la reconstrucción de la matrz). Los valores de k elegidos para el análisis comparativo fueron $k = 5, 10, 15, 16$. Este último valor es la cantidad total de valores singulares que hay dentro de la matriz S, así que, antes de hacer el experimento se sabía con certeza que la reconstrucción con $k = 16$ iba a ser igual a las imágenes orignales. Estos fueron las reconstrucciones con diferentes valores de k:

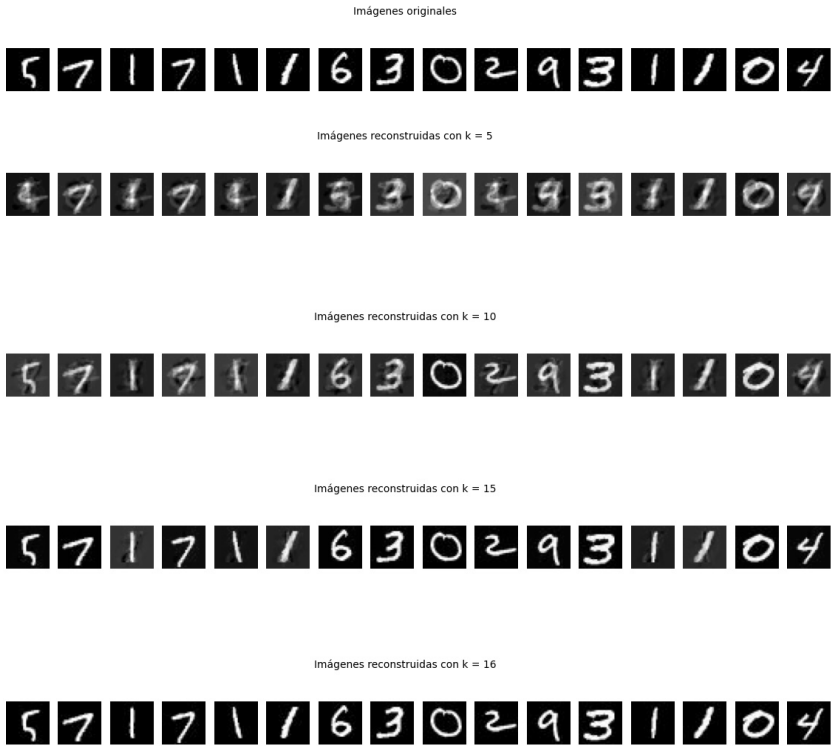

Figura 1: Comparación entre las imágenes originales y las representaciones de baja dimensión que utilizan 5, 10, 15 y 16 valores singulares

Al comparar las reconstrucciones con las imágenes originales fue posible apreciar cómo, mientras más valores singulares/dimensiones tomaba la representación, más parecida a las imágenes originales era. A medida que el valor k se hacía mayor, mejoraba la calidad de la representación; siendo la representación con $k = 5$ una "vagarecreación de las imágenes originales mientras que cuando $k = 15$ el parecido es mucho más evidente. También se pudo demostrar la hipótesis que se había presentado antes de realizar las impresiones de las reconstrucciones: 16 es el número total de valores singulares, por eso su representación es igual a las imágenes originales del dataset.

En la segunda parte de este experimento se comparó qué variaba entre utilizar las k primeras dimensiones o las d últimas. Para esto se volvió a realizar la descomposición SVD, pero esta vez la recontrucción de la matriz se hizo con diferentes valores de d, ya que para la parte anterior ya se habían obtenido las representaciones con los k primeros valores. Estos fueron los resultados obtenidos:

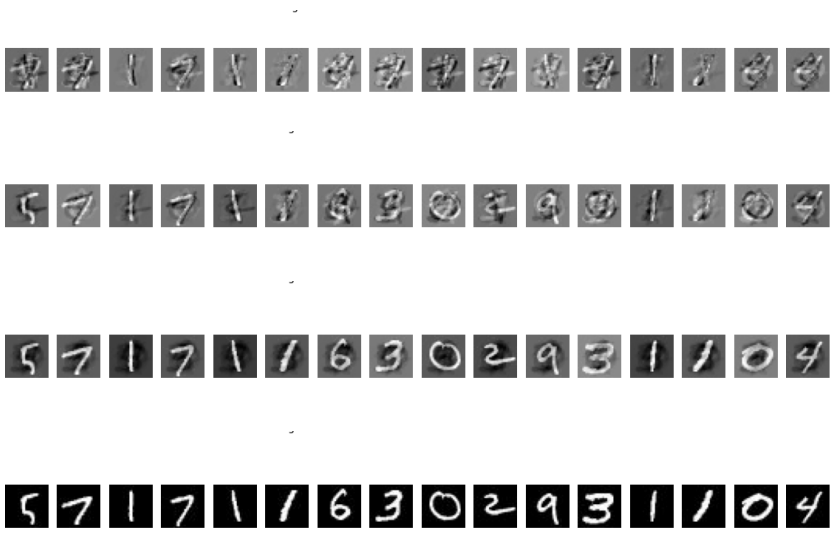

Figura 2: Representaciones de baja dimensión que utilizan las 5, 10, 15 y 16 últimas dimensiones (se muestran en ese órden)

En el caso anterior, en donde se probaban las recontrucciones con las primeras k dimensiones, se veía que mientras más grande k, mejor la representación; pero de igual manera, aún con $k = 5$, las representaciones son considerablemente más similares a las imágenes originales que las que son contruidas con $d = 5$ o $d = 10$. Esto se dió por el siguiente motivo. Al hacer la descomposición SVD, en la matriz de valores singulares (S), quedan los valores singulares ordenados de mayor relevancia a menor relevancia, es por eso que si se toman los primeros 5 valores singulares, va a ser mucho más exacto que si se toman los últimos 5. Al tomar las primeras k dimensiones, si bien mientras más se toman más se parece a la matriz origianl, los valores más importantes para la representación son tomados, porque son los primeros. Esto explica por qué hay una gran diferencia entre la reconstrucción con $d = 15$ y la de $d = 16$ aún si solo se agregó una dimensión, porque el valor 16 de atrás para adelante es el primer valor singular, es decir el más importante. Por eso mientras más grande sea el d (últimos valores singulares tomados) en realidad esos valores se acercan más a los primeros valores, los más importantes. Eso explica que una representación con un d mayor se parezcae más a la orignal.

La tercera parte consistió en comparar mediante a algún método de similaridad cómo la similaridad entre pares de imágenes cambia a medida que se utilizan distintos valores de d. El método escogido para construir esta matriz de similitud fue el de la distancia euclidianda (Norma L_2) entre la matriz de las imágenes orginales y las reconstruídas con diferentes valores de k y d. Estos fueron los resultados obtenidos y algunos análisis.

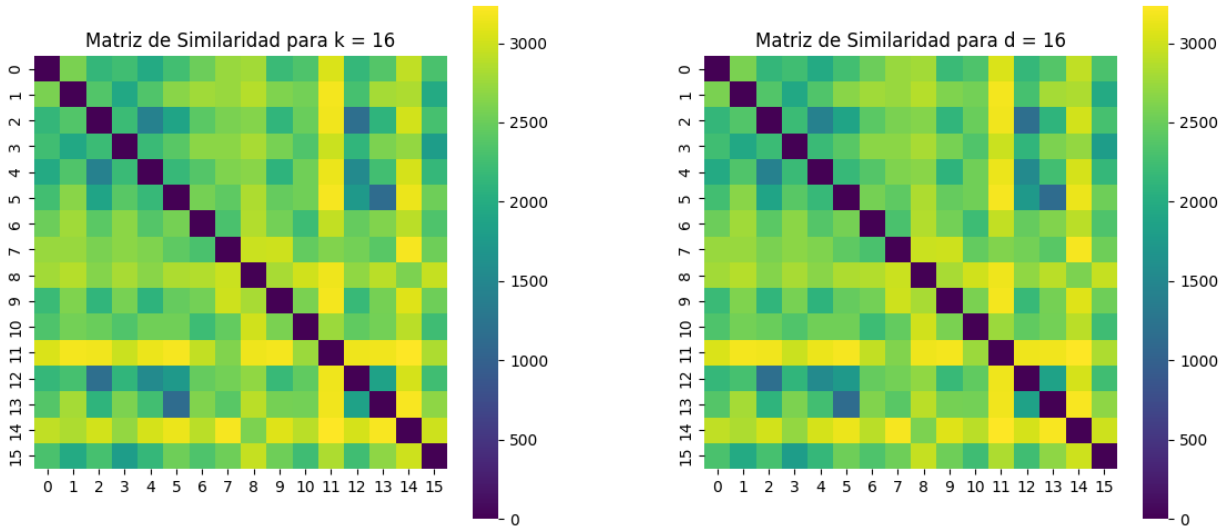

Figura 3: Comparación entre las matrices de similaridad cuando se usan los primeros y últimos 16 valores singulares para la reconstrucción

Esta comparación se realizó para demostrar una vez más que hay 16 dimensiónes en total, entonces es igual si se toman las primeras o últimas 16 dimensiones para recrear las imágenes, porque el resultado va a ser el mismo.

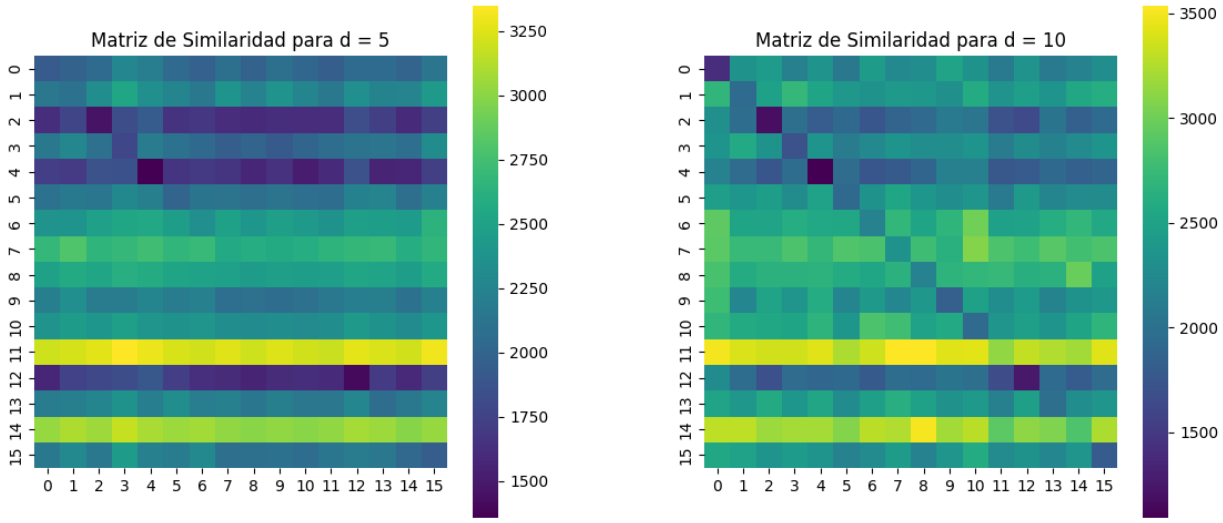

Figura 4: Matriz de similaridad cuando se utilizan los últimos 5 y últimos 10 valores singulares

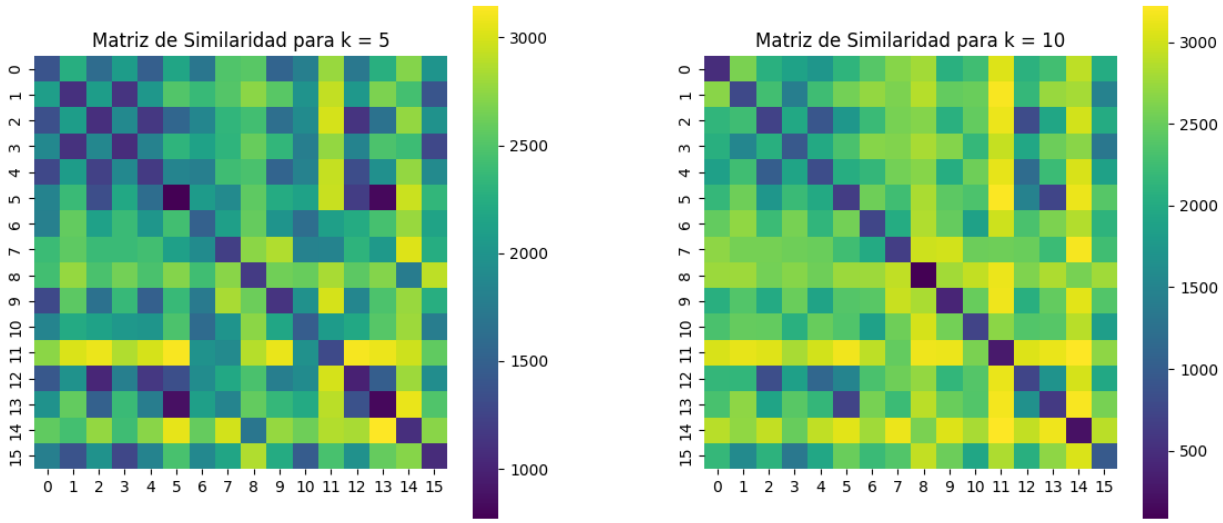

Figura 5: Matriz de similaridad cuando se utilizan los primeros 5 y los primeros 10 valores singulares

Con estas matrices de similaridad se pudo apreciar lo que se obtuvo de la parte anterior del ejercicio: que las reconstrucciones en las cuales se utilizan (la mayor cantidad de) las primeras dimensiones son mucho más similares a la matriz original que las que utilizan los últimos valores singulares, que son los que contienen información menos importante para la recreación. Es por eso que utilizar una poca cantidad de últimas dimensiones es la manera de tener una representación muy diferente a la orginal, pero a medida que el número de d (dimensiones de atrás para adeltante) vaya acercandose a la cantidad de dimensiones totales, en este caso 16, de a poco los valores singulares tendrán más importancia porque se acercarán a los primeros k.

En la cuarta y última parte del experimento se seleccionó una imagen del dataset y se realizó una búsqueda para encontrar un valor para d, la cantidad mínima de dimensiones para que la reconstrucción (con esa cantidad de dimensiones) tenga un error menor al 10 % bajo la norma de Frovenius. Se lo llamó d porque cuenta las dimensiones, no porque sean las últimas d, como se había trabajado en las partes anteriores. Una vez hecha la iteración y encontrado el d para la descomposición SVD, se truncó la matriz V^T para que esta tenga el tamaño correspondiente para la representación de dimensión d; es decir hasta que V^T tuvo de largo d. Una vez obetnida esta reducción de V^T correspondiente a el d óptimo de la imagen seleccionada, se comprimió y descomprimió cada imágen con estos valores de V^T y V, este último obtenido tras hacer $(V^T)^T$. Estos son los resultados para la reconstrucción de las imágenes con el V^T óptimo de solo una de ellas.

En conclusión, la implementación de técnicas de reducción de dimensionalidad y reconstrucción de datos revela la importancia de identificar y preservar las dimensiones que contienen la información más relevante para una representación precisa. Si bien la reducción de dimensionalidad puede mejorar la comprensión de la estructura subyacente de los datos, es fundamental considerar las implicaciones de la pérdida de información al utilizar este enfoque en conjuntos de datos complejos. Por lo tanto, para lograr una representación fiel y una comprensión profunda de los datos, se requiere un equilibrio cuidadoso entre la preservación de la información crucial y la reducción del ruido o la información redundante en el proceso de análisis y modelado.

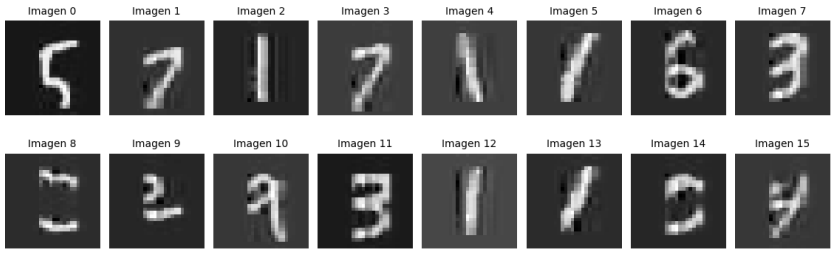

Figura 6: Reconstrucción de las imágenes con los valores óptimos de la imagen 0

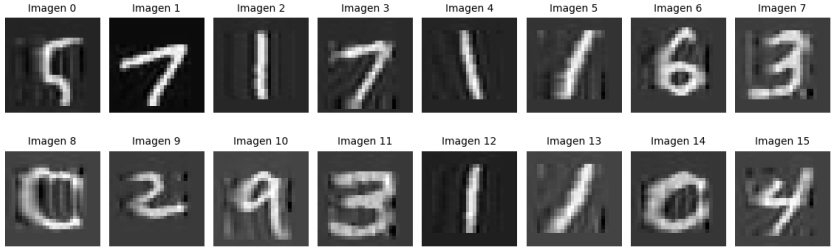

Figura 7: Reconstrucción de las imágenes con los valores óptimos de la imagen 1

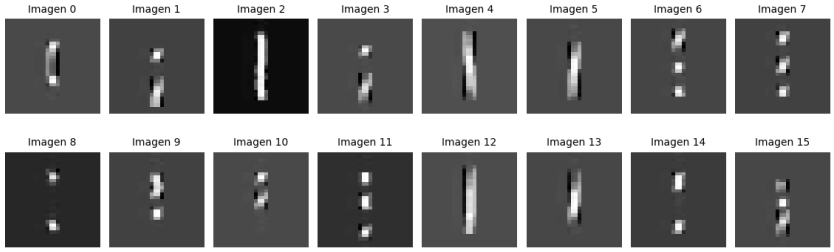

Figura 8: Reconstrucción de las imágenes con los valores óptimos de la imagen 2

Al observar estas tres reconstrucciones fue posible detectar como, la imagen de la cual se tomaron sus valores óptimos es la que mejor calidad posee. Por ejemplo en la Figura 6 (la realizada a partir de los valores óptimos de la primera imagen) es esta la que se ve con mejor definición. Tambien fue observable como son las otras las que se adaptan a las dimensiones óptimas de la otra. Esta adaptación es mejor visualizable en la Figura 8, la representación a partir de las dimensiones óptimas de la imagen 2: al ser una imagen que utiliza pocas dimensiones verticales, sus valores óptimos solo toman estas y descartan las de los costados, haciendo que las demas imágenes al adaptarse queden como recortadas.[en] los costados, viéndose irreconocibles. Esto llevó a la idea de que utilziar las dimensiones óptimas de una sola imagen para reconstruir a todo el conjunto es una manera erronea de reconstruir las imágenes, ya que en el peor de los casos, por ejemplo si una imagen mostrara unicamente un pequeño punto blanco, se podría llegar a perder toda la imagen que se intenta reconstruir.

5.2. Ejercicio 2: Reducción de la dimensionalidad y Cuadrados Mínimos

Luego del desarrollo del experimento se analizaron los resultados.

En primer lugar, se encontró que a medida que se reducía la dimensionalidad de la matriz X, la facilidad de análisis mejoraba significativamente, lo que permitía una mejor comprensión de la estructura y distribución de los datos.

Se identificó una relación directa entre los valores singulares de X y la variabilidad explicada por las dimensiones reducidas, lo que indicaba la importancia de los componentes principales en la representación de los datos. Se determinó que, para los valores de d menores como 2 y 4, como se puede observar en la Figura 9, la dimensionalidad

reducida proporcionaba una representación efectiva de los datos originales sin perder una cantidad significativa de información, lo que facilitaba un análisis más preciso y detallado.

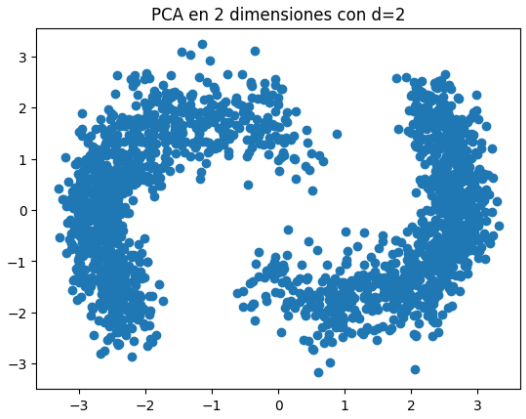

Figura 9: Datos de la matriz X con dimensión 2

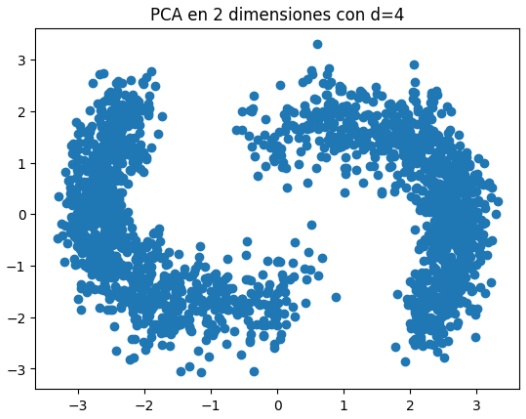

Figura 10: Datos de la matriz X con dimensión 4

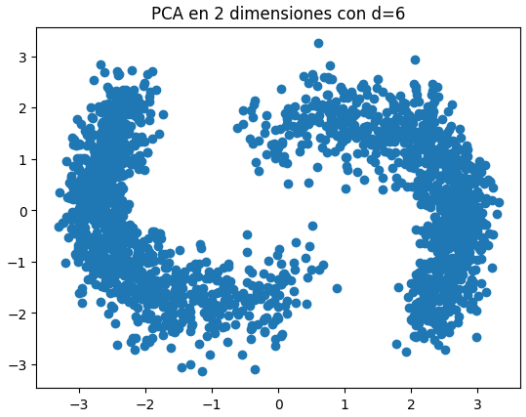

Figura 11: Datos de la matriz X con dimensión 2

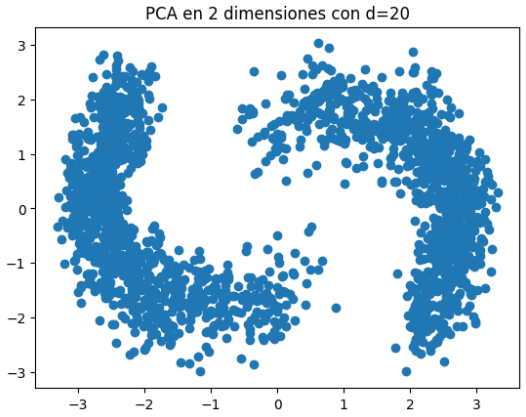

Figura 12: Datos de la matriz X con dimensión 20

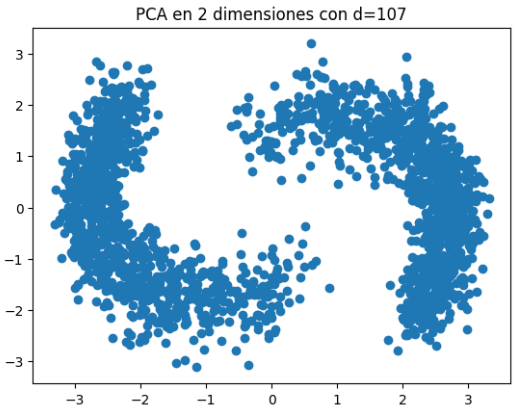

Figura 13: Datos de la matriz X con dimensión 107

Estas conclusiones resaltan la importancia de la reducción de dimensionalidad mediante PCA para comprender la estructura subyacente y la distribución de los datos, lo que puede ser crucial en el análisis de conjuntos de datos complejos y de alta dimensionalidad.

En segundo lugar, se realizó el análisis de los resultados de la medición de similitud par-a-par entre las muestras en el espacio de dimensión X y en el espacio de dimensión reducida d utilizando PCA y proyecciones al azar reveló ciertas observaciones significativas:

Similitud en el espacio de dimensión original (X): Se observó que la similitud par-a-par entre las muestras en el espacio de dimensión original se basaba en la distribución completa de los datos en todas las dimensiones. Esto proporcionó una visión holística de las relaciones entre las muestras en el contexto original de alta dimensionalidad.

Similitud en el espacio de dimensión reducida (d): Se encontró que la similitud par-a-par entre las muestras en el espacio de dimensión reducida se basaba en las relaciones capturadas por los componentes principales seleccionados durante la reducción de dimensionalidad. Se notó que ciertas similitudes se resaltaron más en las dimensiones reducidas, mientras que otras se atenuaron, lo que indicaba una representación simplificada de las relaciones entre las muestras en un espacio de dimensionalidad inferior.

Diferencias entre el PCA y las proyecciones al azar: Se observó que el análisis de componentes principales (PCA) proporcionaba una representación más estructurada y significativa de la similitud entre las muestras en comparación con las proyecciones al azar. Esto indicaba que el PCA capturaba de manera efectiva las relaciones fundamentales presentes en los datos, lo que permitía una interpretación más precisa y fiable de las similitudes entre las muestras en diferentes dimensiones.

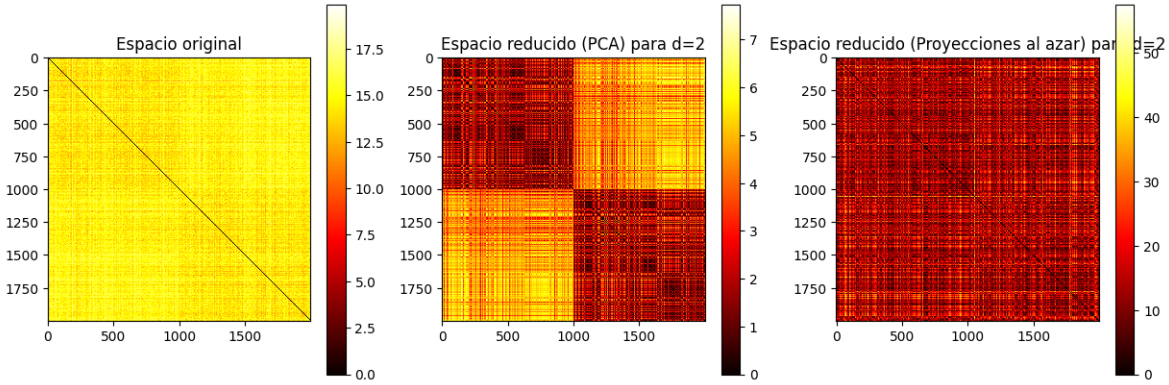

Figura 14: Matrices de similaridad para d = 2

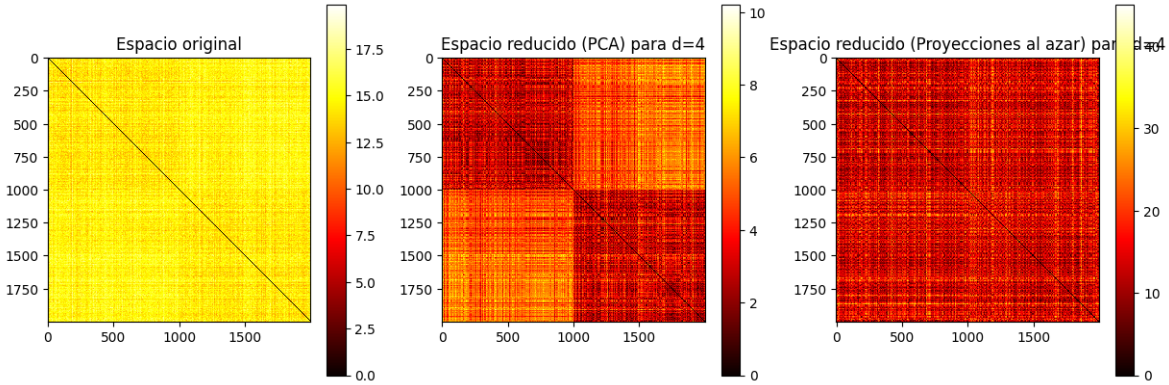

Figura 15: Matrices de similaridad para d = 4

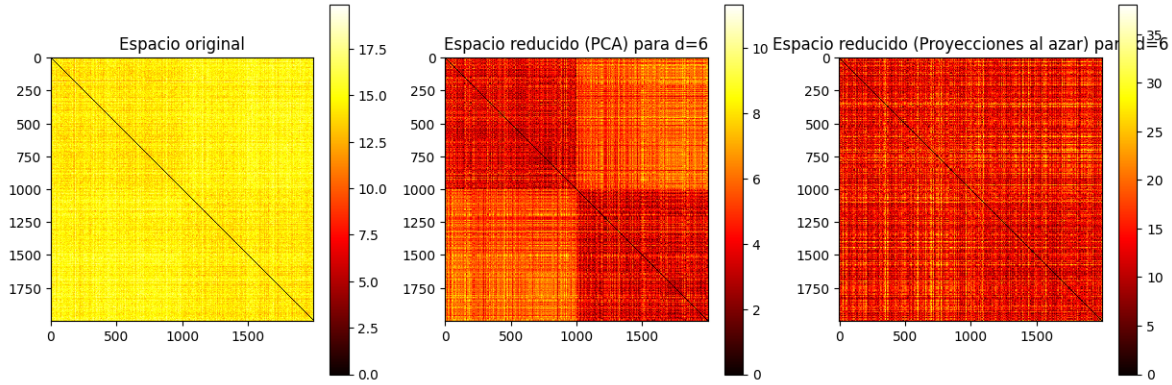

Figura 16: Matrices de similaridad para d = 6

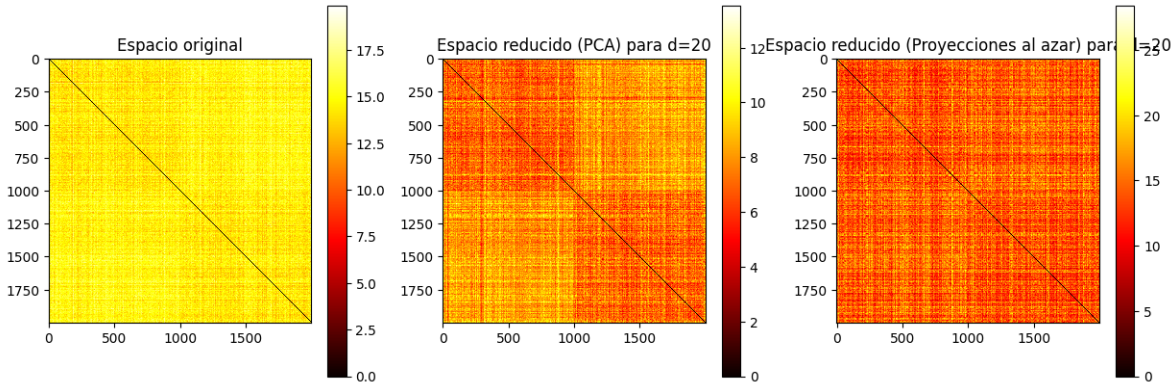

Figura 17: Matrices de similaridad para d = 20

El análisis de similitud proporcionó una comprensión profunda de las relaciones entre las muestras en diferentes espacios de dimensionalidad, lo que ayudó a identificar patrones importantes y estructuras subyacentes en los datos originales y en la representación reducida. Esto fue crucial para comprender la influencia de la reducción de dimensionalidad en la similitud y la relación entre las muestras en un conjunto de datos complejo y de alta dimensionalidad.

En tercer lugar, el análisis de las dimensiones más representativas basado en la descomposición en valores singulares (SVD) reveló las dimensiones originales del conjunto de características que ejercen una influencia significativa en la variabilidad capturada por las nuevas dimensiones generadas. Al considerar diferentes valores de "d", se identificaron las dimensiones más influyentes para cada caso, proporcionando una comprensión más profunda de las características clave dentro del conjunto de datos. Para d=2, las dimensiones más representativas fueron 102 y 101, lo que indica su impacto en la variabilidad capturada por las dos dimensiones principales generadas por el SVD. Al aumentar "d.ª 4 y luego a 6, se observó la contribución de un conjunto más diverso de características, como se evidenció en las dimensiones 102, 101, 57, 10, 32, y 90, entre otras. Finalmente, para d=20, se identificaron 20 dimensiones originales que ejercen una influencia significativa en la variabilidad de los datos capturada por las 20 dimensiones principales generadas.

En el contexto de la descomposición en valores singulares (SVD), los vectores de características se obtienen como parte de la matriz V. Al considerar los elementos de estos vectores, se pueden identificar las dimensiones originales más importantes. Las características con valores absolutos más altos en estos vectores son consideradas como las dimensiones originales más representativas en relación con las dimensiones resultantes obtenidas por SVD. Es importante tener en cuenta que las dimensiones originales más representativas pueden variar entre diferentes ejecuciones debido a la naturaleza estocástica de los métodos de reducción de dimensionalidad. Además, la influencia de las dimensiones individuales puede depender de varios factores, como la distribución de los datos y la escala de las características. Por lo tanto, se recomienda realizar múltiples ejecuciones y considerar una variedad de resultados para obtener una comprensión completa de las dimensiones más influyentes en el conjunto de datos.

En cuarto lugar, se utilizó la descomposición en valores singulares (SVD) para encontrar el vector β que minimiza la norma $\|X\beta - y\|_2$, donde y es un vector de etiquetas o variable dependiente. El análisis realizado demuestra la capacidad del modelo para ajustar los datos mediante la utilización de la descomposición en valores singulares (SVD) y la posterior estimación del vector beta. El vector beta obtenido minimiza la norma de la diferencia entre las predicciones y las etiquetas reales, proporcionando así una representación lineal que se acerca de manera óptima a las etiquetas del conjunto de datos.

El gráfico del error de predicción en función de la dimensión d muestra una caída drástica en el error de predicción en las primeras dimensiones, lo que sugiere que un número limitado de dimensiones es suficiente para capturar la variabilidad presente en los datos y realizar predicciones precisas. La estabilización del error de predicción a partir de la dimensión 3 o 4 indica que la inclusión de dimensiones adicionales no conduce a mejoras significativas en la predicción de las etiquetas y, por lo tanto, no es necesaria.

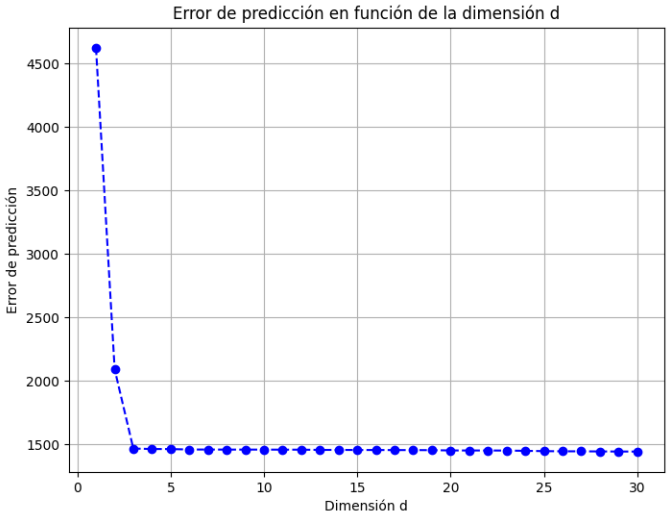

Figura 18: Error de predicción

Estos resultados sugieren que la selección de características es crucial para la precisión de la predicción. Al considerar un subconjunto óptimo de dimensiones que capture la mayor variabilidad de los datos, es posible reducir la complejidad del modelo sin comprometer su capacidad predictiva. Además, el análisis destaca la importancia de la regularización y la selección de un número óptimo de características significativas para evitar el sobreajuste y mejorar la generalización del modelo a datos no vistos.

6. Conclusiones

Luego de realizar estos experimentos se pudo llegar a ciertas conclusiones.

Con los resultados del primer ejercicio fue posible deducir que a medida que se utilicen más dimensiones para hacer la reconstrucción de imágenes, la calidad de la representación se asemejará más a las imágenes originales. También se llegó a la conclusión que las reconstrucciones que toman las primeras dimensiones son mucho más precisas que las que utilizan las últimas, ya que estas dimensiones contienen la información menos relevante para la representación, mientras que las primeras poseen la información más importante. Además, se demostró que utilizar las dimensiones óptimas de una sola imagen para reconstruir todo el resto del conjunto puede llevar a una pérdida significativa de información en algunas imágenes, lo que indica que este enfoque no es adecuado en todos los casos.

El segundo experimento demostró que la reducción de la dimensionalidad mejoró la comprensión de la estructura de los datos. El análisis de similitud reveló diferencias entre las representaciones en dimensiones originales y reducidas. La identificación de dimensiones representativas destacó las características clave del conjunto de datos.

Además, el modelado lineal mostró la importancia de la selección de características óptimas para una predicción precisa y una mejor generalización del modelo. Estos hallazgos enfatizan la utilidad de las técnicas de reducción de dimensionalidad y el análisis eficiente en conjuntos de datos complejos.

En conclusión, la implementación de técnicas de reducción de dimensionalidad y reconstrucción de datos revela la importancia de identificar y preservar las dimensiones que contienen la información más relevante para una representación precisa. Si bien la reducción de dimensionalidad puede mejorar la comprensión de la estructura subyacente de los datos, es fundamental considerar las implicaciones de la pérdida de información al utilizar este enfoque en conjuntos de datos complejos. Por lo tanto, para lograr una representación fiel y una comprensión profunda de los datos, se requiere un equilibrio cuidadoso entre la preservación de la información crucial y la reducción del ruido o la información redundante en el proceso de análisis y modelado.

7. Bibliografía

- Burden, R. L., Faires, J. D., Burden, A. M. (2015). Numerical analysis. Cengage learning.
- Numpy, Biblioteca de Python
- Scikit-learn, Biblioteca de Python
- Seaborn, Biblioteca de Python
- Trefethen, L. N., Bau, D. (2002). Numerical linear algebra (Vol. 181)