Universidad de
SanAndrés

Herramientas del Álgebra Computacional para lidiar con la Alta Dimensionalidad

Lucas Ercolano lercolano@udesa.edu.ar
Tomás Castro tacastro@udesa.edu.ar

Métodos Numéricos y Optimización

Resumen

Este informe presenta un análisis exhaustivo de los métodos de descomposición en valores singulares (SVD) y reducción de dimensionalidad aplicados a dos problemáticas específicas: reducción de dimensionalidad y compresión de imágenes. En la primera parte, se examina un conjunto de datos de alta dimensionalidad, buscando identificar las dimensiones más representativas y evaluar la similitud entre muestras tras aplicar SVD y PCA. La segunda parte se centra en la compresión de imágenes, utilizando SVD para aprender representaciones de baja dimensión y evaluar la calidad de las imágenes reconstruidas.

Palabras clave: Cuadrados Mínimos, SVD, Valores Singulares, Dimensionalidad, Optimización, Similaridad, Varianza.

1. Introducción

La reducción de dimensionalidad es una técnica esencial en el análisis de datos y la optimización, permitiendo simplificar conjuntos de datos complejos sin perder información crucial. Este informe se enfoca en la aplicación de la Descomposición en Valores Singulares (SVD) y el Análisis de Componentes Principales (PCA) para mejorar la interpretación y procesamiento de datos de alta dimensionalidad y compresión de imágenes.

El objetivo principal es demostrar cómo SVD y PCA pueden revelar estructuras subyacentes en datos complejos, facilitar su visualización y mejorar la eficiencia en su almacenamiento y procesamiento. A través de experimentos en dos contextos distintos, análisis de datos numéricos y compresión de imágenes se evaluará la efectividad de estos métodos en la reducción de dimensionalidad.

En la primera parte, se analiza un conjunto de datos numéricos medidos por un sensor, explorando la similitud entre muestras y determinando las dimensiones más representativas tras la reducción. Además, se evalúa la capacidad de predicción de un modelo lineal en el espacio reducido.

En la segunda parte, se aplica SVD para comprimir un conjunto de imágenes, evaluando la calidad de las imágenes reconstruidas y la similitud entre ellas según el grado de compresión. También se determina la mínima dimensionalidad necesaria para mantener un error de compresión aceptable.

Este estudio ilustra la aplicación práctica de SVD y PCA en la mejora del análisis de datos y compresión de imágenes, evidenciando su capacidad para simplificar y optimizar el manejo de datos complejos.

2. Terminología

2.1. Features

Los "features"son las variables o atributos que describen cada muestra en un dataset. Pueden ser numéricos o categóricos y son cruciales para entrenar modelos de machine learning, ya que proporcionan la información necesaria para que el algoritmo aprenda y haga predicciones. La calidad y relevancia de los features afectan directamente el rendimiento y precisión del modelo.

2.2. Kernel

Un kernel es una función utilizada en machine learning, especialmente en métodos de clasificación y regresión, que transforma los datos originales en un espacio de características de mayor dimensión para hacerlos más separables.

2.3. Kernel Gaussiano

El Kernel Gaussiano es una medida entre dos puntos en un espacio n-dimensional. Para dos vectores x_i y x_j, se define como

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}\right) \tag{1}$$

donde $\|x_i - x_j\|$ es la norma euclidiana (distancia) entre los vectores x_i,x_j y σ es un parámetro de escala que ajusta la sensibilidad de la similitud.

Si los vectores son muy similares, la distancia será pequeña, lo que hace que $K(x_i, x_j)$ sea cercano a 1. Si los vectores son muy diferentes (ortogonales), la distancia será grande, haciendo que $K(x_i, x_j)$ sea cercano a 0.

2.4. Matriz de Similaridad

Una matriz de similaridad es una matriz cuadrada que contiene las medidas de similitud entre un conjunto de vectores. Cada entrada K_{ij} de la matriz representa la similitud entre el i-ésimo y el j-ésimo vector. Las medidas de similitud pueden basarse en diferentes métricas, como distancia lineal, el kernel gaussiano o cualquier otra métrica adecuada.

2.5. Varianza

La varianza es una medida de la dispersión de un conjunto de datos. Indica cuánto varían los datos respecto a su media. Matemáticamente, para un conjunto de datos X con media μ, la varianza se define como

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 \tag{2}$$

donde X_i son los datos individuales y n es el número total de datos. Una varianza alta indica que los datos están más dispersos, mientras que una varianza baja indica que los datos están más concentrados alrededor de la media.

2.6. Clusters

Los clusters son grupos de datos similares entre sí y diferentes de otros grupos. Los algoritmos de clustering identifican estas agrupaciones basándose en criterios de similaridad o distancia. Los clusters ayudan a descubrir patrones ocultos y estructuras en los datos, permitiendo la segmentación y la identificación de relaciones subyacentes. Visualizarlos es esencial para interpretar los resultados y entender la distribución en el dataset.

2.7. Denoising

El denoising es el proceso de eliminar el ruido de los datos para mejorar su calidad y la precisión de los modelos. El ruido puede ser cualquier tipo de información aleatoria o irrelevante que contamine los datos y afecta negativamente los resultados del análisis. La eliminación efectiva del ruido permite que los algoritmos se enfoquen en los patrones relevantes de los datos, mejorando así la exactitud de las predicciones y las decisiones basadas en esos datos.

2.8. Manifold

Un manifold es una estructura matemática que generaliza la noción de superficies curvas a espacios de mayor dimensión. En el aprendizaje automático, se utiliza para describir la geometría subyacente de un conjunto de datos de alta dimensionalidad. La hipótesis del manifold sugiere que, aunque los datos puedan residir en un espacio de alta dimensión, en realidad, se distribuyen a lo largo de una superficie de menor dimensión. Identificar y comprender el manifold de los datos permite realizar reducciones de dimensionalidad más efectivas, preservando la estructura esencial y revelando patrones y relaciones que no son evidentes en el espacio original.

3. Métodos y técnicas empleadas

3.1. Descomposición en Valores Singulares (SVD)

La descomposición en valores singulares (SVD) es una técnica de álgebra lineal que descompone una matriz A de dimensiones $m \times n$ en tres matrices específicas: U, S y V^T. Matemáticamente, esto se expresa como

$$A = USV^T \tag{3}$$

Donde A es la matriz original de dimensiones $m \times n$, U una matriz ortogonal de dimensiones $m \times m$, cuyas columnas son los vectores singulares izquierdos de A, S una matriz diagonal de dimensiones $m \times n$, cuyos elementos no negativos son los valores singulares de A. Los valores singulares están ordenados en orden descendente y V^T es la Transpuesta de una matriz ortogonal V de dimensiones $n \times n$, cuyas columnas son los vectores singulares de A.

La SVD descompone la matriz A en una suma de productos exteriores de vectores, donde cada producto está ponderado por un valor singular. Esencialmente, U y V^T proporcionan las direcciones de los espacios de entrada y salida, respectivamente, mientras que S proporciona la escala en estas direcciones. Esta descomposición permite la reducción de dimensionalidad al truncar los valores singulares más pequeños, simplificando la estructura de A sin perder mucha información.

3.2. Pseudo-inversa

La pseudo-inversa de una matriz es una generalización de la inversa de una matriz cuadrada, que se aplica a matrices que no son necesariamente invertibles, es particularmente útil para resolver sistemas de ecuaciones lineales que no tienen una solución única o que son sobredeterminados o subdeterminados (matemáticamente, se denota como A^+). Cuando se usa la Descomposición en Valores Singulares (SVD), se puede calcular de la siguiente manera

$$A^+ = VS^+U^T \tag{4}$$

(S^+, se consigue invirtiendo cada valor singular no nulo en S y luego transponiendo la matriz diagonal resultante).

La pseudo-inversa es una herramienta poderosa en el análisis de datos y aprendizaje automático, ya que permite resolver problemas de mínimos cuadrados y encontrar soluciones aproximadas a sistemas de ecuaciones lineales que no tienen soluciones exactas.

3.3. Análisis de Componentes Principales (PCA)

El Análisis de Componentes Principales (PCA) es una técnica de reducción de dimensionalidad que transforma un conjunto de datos de alta dimensionalidad en un nuevo conjunto de datos con menos dimensiones, mientras se preserva la mayor parte de la varianza original. Esto se logra encontrando las direcciones (componentes principales) que maximizan la varianza de los datos. Para esto se calcula la SVD de la matriz y los vectores singulares (V) proporcionan las direcciones de las componentes principales mientras que las proyecciones de los datos originales en estas nuevas direcciones (componentes principales) son los nuevos datos reducidos.

3.4. Cuadrados Mínimos

El método de *Cuadrados Mínimos* es una técnica matemática utilizada para encontrar la mejor aproximación de un conjunto de datos mediante la minimización de la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por un modelo. En el contexto de la regresión, los cuadrados mínimos se utilizan para ajustar una línea o una curva a un conjunto de datos de manera que la suma de los cuadrados de los residuos (las diferencias entre los valores observados y los valores predichos) sea lo más pequeña posible.

Matemáticamente, si tenemos un conjunto de datos $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, el objetivo es encontrar los coeficientes del modelo $y = f(x)$ que minimicen la función de costo

$$S = \sum_{i=1}^{n} (y_i - f(x_i))^2 \tag{5}$$

3.5. Regresión Lineal

La Regresión Lineal es un método estadístico utilizado para modelar la relación entre una variable dependiente y y una o más variables independientes x. En el caso más sencillo (regresión lineal simple), se considera una variable independiente. La regresión lineal asume que la relación entre las variables es lineal, y el objetivo es encontrar la línea recta que mejor se ajuste a los datos.

El ajuste del modelo de regresión lineal se realiza utilizando el método de cuadrados mínimos, que minimiza la suma de los cuadrados de los residuos, es decir, las diferencias entre los valores observados y los valores predichos por el modelo, de esta manera

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \tag{6}$$

donde y es la variable dependiente, x es la variable independiente, β_0 es la ordenada al origen (intercepto) y β_1 es la pendiente de la línea de regresión

4. Métricas de desempeño

4.1. Norma de Frobenius

La norma de Frobenius de una matriz A de dimensiones $m \times n$ es una medida de su magnitud y se define como

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} \tag{7}$$

donde a_{ij} es el elemento en la posición i, j de la matriz A. La norma de Frobenius es equivalente a la raíz cuadrada de la suma de los cuadrados de todos los elementos de la matriz. Esta norma es útil para medir la diferencia entre una matriz original y su aproximación, especialmente en el contexto de la compresión de imágenes, asegurando que las comparaciones y cálculos se realicen de manera consistente.

4.2. Errores de Compresión

El error de compresión mide la diferencia entre la matriz original y su aproximación tras la compresión. En el contexto de SVD, esto implica comparar la matriz original A con su aproximación $\hat{A}$, obtenida al truncar la SVD para mantener solo los k valores singulares más grandes. El error de compresión puede medirse utilizando la norma de Frobenius

$$\|A - \hat{A}\|_F \tag{8}$$

donde $\hat{A}$ es la matriz aproximada

$$\hat{A} = U_k S_k V_k^T \tag{9}$$

Aquí, U_k, S_k, y V_k^T son las matrices truncadas que corresponden a los k valores singulares más grandes. La norma de Frobenius del error de compresión proporciona una medida cuantitativa de la pérdida de información debido a la compresión.

4.3. Similitud Par-a-Par

La similitud par-a-par entre muestras en los espacios de dimensión original y reducida se compara para diferentes valores de d. Esto se visualiza utilizando matrices de similaridad y ayuda a identificar cómo la reducción de dimensionalidad afecta la estructura del conjunto de datos.

5. Desarrollo Experimental y Resultados

En este análisis, se implementan las librerías de Python `numpy` y `matplotlib` para analizar y representar gráficamente los cálculos de `SVD` y `PCA`. Con el objetivo de comprender la distribución de las muestras en el espacio original de alta dimensionalidad, se calculó primero la matriz de similaridad. Esta matriz permite visualizar cómo las muestras se agrupan en el espacio original, lo cual contribuye a identificar clusters naturales en los datos. Sin embargo, como se observa en la Figura 1, debido a la alta dimensionalidad y al ruido inherente en algunas dimensiones, esta visualización puede ser complicada e imprecisa, dificultando la interpretación clara de los datos.

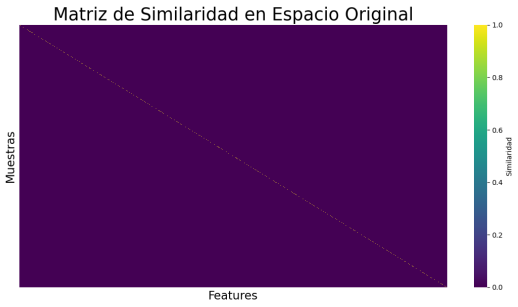


Fig. 1: Matriz de similaridad en espacio original.

El parámetro σ actúa como un factor de escala, influyendo en cómo se ponderan las distancias entre muestras. Matemáticamente, es proporcional al valor singular al cuadrado o al desvío estándar de los datos, afectando directamente al kernel gaussiano utilizado para calcular la similaridad.

Cuando σ es muy pequeño, las diferencias entre las distancias se amplifican, resultando en una matriz de similaridad que resalta solo las muestras muy cercanas entre sí. Esto puede ser útil para identificar clusters muy definidos, ya que solo las muestras extremadamente próximas aparecerán como similares. Sin embargo, puede resultar en una matriz dispersa que no captura la estructura global del manifold.

A medida que σ aumenta, las diferencias se suavizan, y más muestras se consideran similares entre sí. Esto puede llevar a una matriz de similaridad más densa, donde las relaciones entre los datos son más homogéneas. Aunque esta aproximación puede perder detalles finos, es útil para identificar tendencias generales y agrupamientos más amplios en el espacio de características.

En la Figura 2 se presentan las matrices de similaridad con diferentes valores de σ, ilustrando cómo varía la representación de las similitudes entre las muestras y cómo esto afecta la identificación de clusters y la comprensión de la estructura de los datos en dimensiones reducidas.

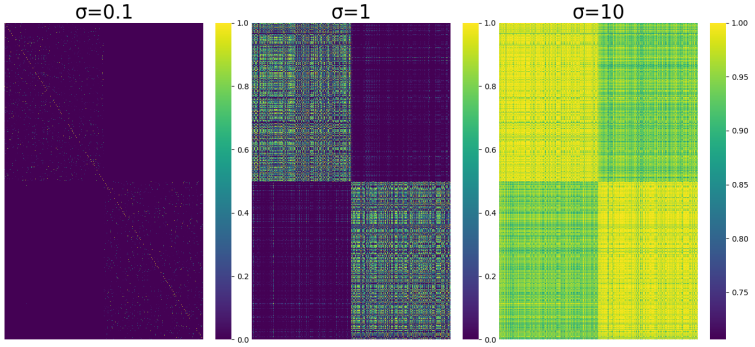


Fig. 2: Matrices de similaridad variando σ ilustrando su impacto en la representación de similitudes.

Dado que el espacio de alta dimensionalidad contiene ruido que puede oscurecer la estructura subyacente de los datos, es crucial realizar una reducción de dimensionalidad para obtener una representación más clara y manejable. La reducción de dimensionalidad permite trabajar en un espacio donde las relaciones entre las muestras sean más evidentes y el ruido se minimice (denoising). Para entender cómo la reducción de dimensionalidad afecta la estructura de los datos, se realizó una descomposición en valores singulares (SVD) y se analizaron los datos en espacios proyectados en 2, 6, 10 y p (siendo $p = 205$) dimensiones (procedimiento detallado en el Apéndice A).

Cuando se proyecta en 2 o 3 dimensiones, se puede visualizar los datos en gráficos bidimensionales o tridimensionales, lo que facilita la interpretación visual de las agrupaciones, tal como se muestra en la Figura 3. Estas representaciones permiten observar cómo las muestras se distribuyen en un espacio de menor dimensionalidad, haciendo más evidente la formación de clusters y permitiendo una interpretación más intuitiva de las relaciones entre las muestras.

Para proyecciones en mayores dimensiones, se utilizaron matrices de similaridad para mantener una representación comprensible de las relaciones entre las muestras. Esto proporciona una forma de visualizar la estructura de los datos incluso cuando no es posible representarla directamente en un gráfico convencional, aunque no se incluirán los gráficos de matrices de similaridad ya que son similares a los anteriormente mostrados. Esta comparación permite evaluar cuánto de la estructura y de las relaciones originales entre las muestras se preservan en los espacios de menor dimensionalidad, ayudando a entender el impacto del ruido y la utilidad de la reducción dimensional en la visualización y el análisis de datos.

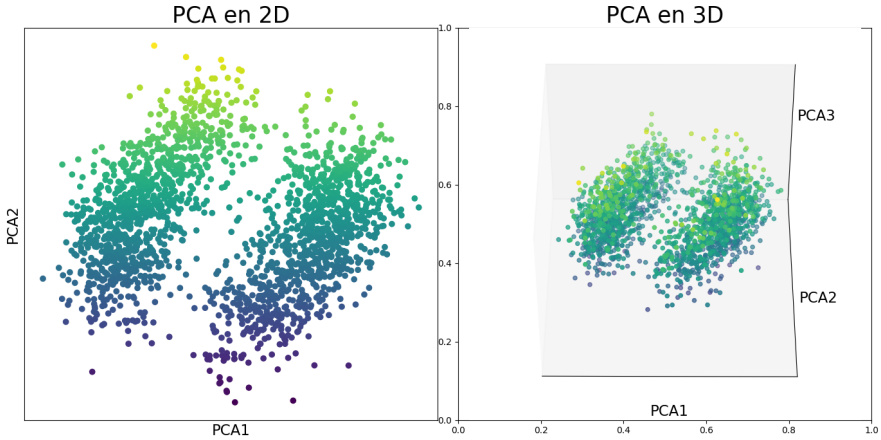


Fig. 3: Proyección de las muestras en dimensiones R^2 y R^3, mostrando la estructura preservada de los datos en espacios de menor dimensionalidad.

La descomposición SVD proporciona una lista de valores singulares que indican la importancia en la representación de los datos originales. Al graficar el porcentaje de representación de cada valor singular, como se observa en la Figura 4, se nota que los primeros valores singulares tienen mucho más peso que el resto. Esto coincide también con la Figura 5, donde al observar los valores singulares, se puede ver claramente que los dos primeros valores tienen una magnitud significativamente mayor en comparación con los demás. Esto indica que las primeras dos componentes principales capturan la mayor parte de la variación presente en los datos originales, aunque no es cierto que el resto de componentes no sean importantes, ya que en tal caso serían exclusivamente ruido.

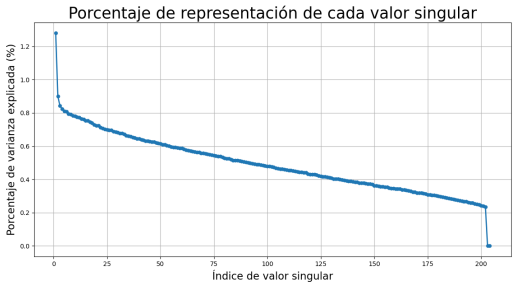


Fig. 4: Porcentaje de Representación.

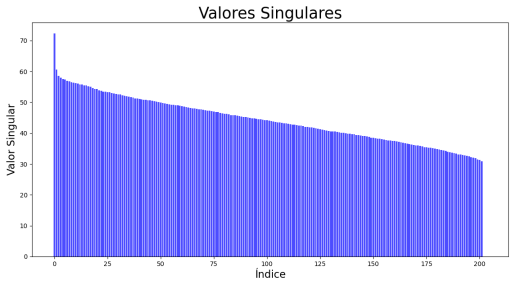


Fig. 5: Magnitud de los Valores Singulares.

Para identificar las dimensiones más significativas, se tomaron los primeros valores singulares y se analizaron sus componentes, ya que los autovectores asociados a los valores singulares más grandes nos muestran las combinaciones de features que más modifican al espacio. En la Figura 6 se ve el peso de cada feature sobre el vector, se nota una gran diferencia que marca la poca importancia de muchos componentes y el gran peso de los primeros. Esto también se puede observar al hacer un mapa de calor del dataset original (Figura 7), donde se identifica mucho ruido, pero a su vez una cierta estructura sobre las últimas columnas.

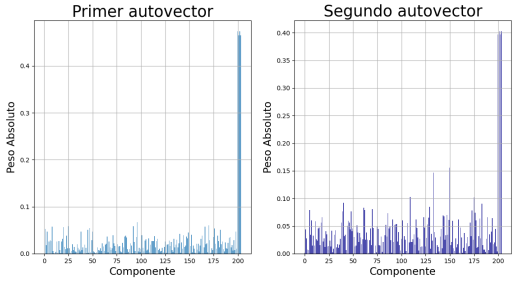


Fig. 6: Peso absoluto de características del vector.

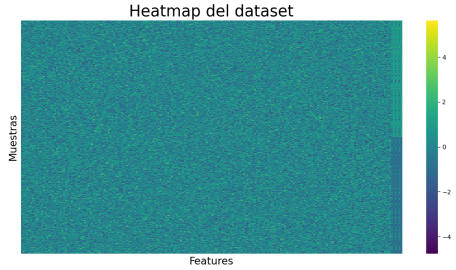


Fig. 7: Mapa de calor del dataset original.

Para simplificar el análisis y al mismo tiempo retener la mayor cantidad de información posible, se optó por reducir la dimensionalidad del espacio original a dos dimensiones. Esta decisión permite una representación más manejable y visualmente interpretable de los datos sin sacrificar una cantidad considerable de variación. Al reducir las dimensiones, se pueden visualizar los clusters de datos y comprender mejor la estructura subyacente del manifold en un espacio de características reducido.

Al determinar la dimensión d que optimiza la predicción de las etiquetas y a partir de los datos X, se normalizó el dataset y se utilizó SVD para descomponer la matriz. Para cada valor de d, se generó una versión reducida de X utilizando los primeros d componentes principales, observando cómo la dimensionalidad afecta la capacidad predictiva del modelo. Los datos fueron divididos en conjuntos de entrenamiento y prueba para evaluar el rendimiento del modelo, enfocándose en no "overfitear" el modelo. Se calculó la pseudo-inversa de X para estimar el vector β que minimiza la norma $\|X\hat{\beta} - y\|^2$, representando los coeficientes del modelo lineal en el espacio reducido.

Después de evaluar el error de predicción para cada dimensión d, se encontró que la mejor dimensión para la predicción es $d = 2$, indicando que un espacio de características reducido a dos dimensiones es el más eficaz para modelar la relación entre X e y en este caso. Esta noción se puede visualizar en la Figura 8.

La elección de dos dimensiones como la óptima puede explicarse por la alta variabilidad capturada por los primeros dos valores singulares, permitiendo que el modelo capture las relaciones más significativas sin el ruido adicional de dimensiones menos informativas. Con esto se procedió a visualizar los datos y el plano de regresión en el espacio reducido, como se muestra en la Figura 9. Esta visualización ayuda a comprender mejor cómo el modelo lineal se ajusta a los datos reducidos y a observar la relación entre las características principales y la variable de salida.

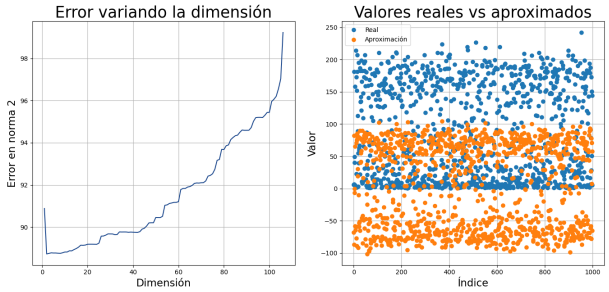


Fig. 8: Comparación entre los valores reales y la aproximación del modelo.

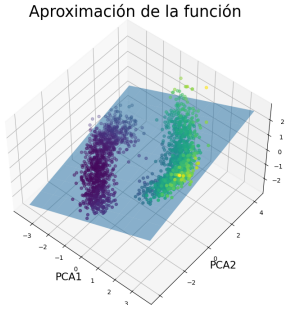


Fig. 9: Visualización del plano de regresión lineal en el espacio reducido de dos dimensiones.

6. Compresión de Imágenes y Análisis de Similaridad

En esta sección se exploran técnicas de compresión de imágenes y análisis de similaridad utilizando la Descomposición en Valores Singulares (SVD). Se contemplan dos datasets distintos: el primero (Figura 10) contiene 19 imágenes de números dibujados de diversas formas y ángulos, mientras que el segundo (Figura 11) incluye 8 imágenes (cuatro del número 2 y cuatro del número 8). Además, se experimentó aplicando técnicas de procesamiento de imágenes utilizando filtros de convolución para mejorar la calidad de los datos y la precisión del análisis. Los detalles de la aplicación y los resultados de estos filtros se presentan en el Apéndice D.

Las imágenes de ambos datasets fueron representadas como matrices $p \times p = 24 \times 24$ y posteriormente vectorizadas y apiladas, formando matrices de datos adecuadas para el análisis.

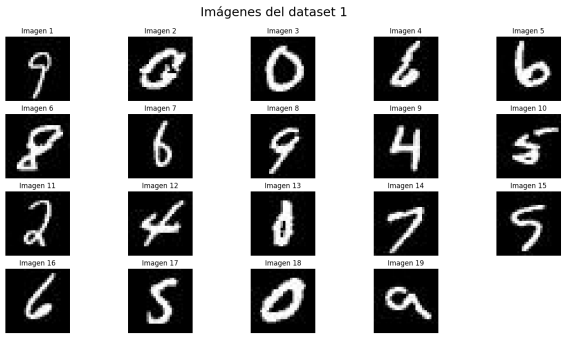


Fig. 10: Conjunto de imágenes del dataset 1: Números 9, 0, 6, 8, 4, 5, 2, 7.

Fig. 11: Conjunto de imágenes del dataset 2: Números 2 y 8.

Dado que $p \times p = 24 \times 24$, enfrentamos un problema de alta dimensionalidad, con un espacio de 576 dimensiones para representar simplemente números que van del 0 al 9. Por esta razón, se implementa la compresión considerando que sería más óptimo contemplar un espacio de menor dimensión que abarque mejor la distribución de los datos. El objetivo es utilizar la SVD para descomponer la matriz de datos en componentes que representan las características más importantes de las imágenes en un espacio de menor dimensión, y analizar cómo esta reducción afecta la calidad de la reconstrucción y la similaridad entre pares de imágenes. Así, la SVD permite reducir el ruido en los datos, logrando una representación más compacta y eficiente.

Para evaluar la calidad de la reconstrucción, se comparan las imágenes originales con las imágenes reconstruidas considerando diferentes dimensiones y utilizando las componentes principales, es decir, los autovectores más significativos. Inicialmente, se utilizan $d = 2$, $d = 10$ y $d = 19$ dimensiones para la reconstrucción. Para simplificar el análisis y mejorar la calidad de los gráficos, se muestran 9 imágenes clusterizadas en lugar de las 19 del dataset, destacando así la mayor variedad.

En primer lugar, se analiza la reconstrucción con $d = 2$. Como se observa en la Figura 12, las imágenes resultantes son muy borrosas y carecen de detalles, lo que indica que solo se retienen las características más básicas y se pierde mucha información relevante.

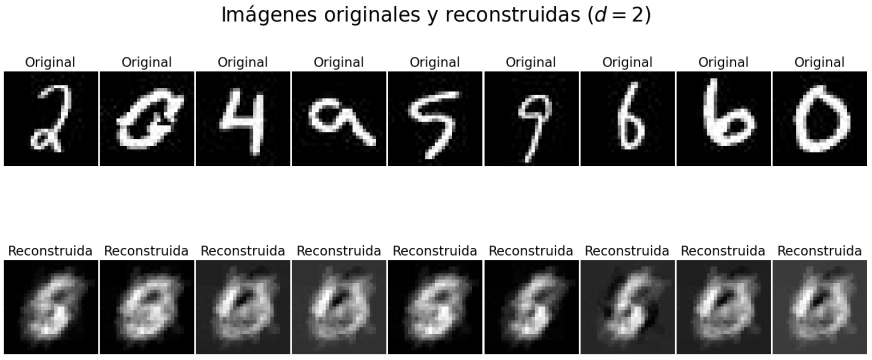


Fig. 12: Imágenes originales y reconstruidas con $d = 2$.

En contraste, las imágenes reconstruidas con $d = 10$ muestran una mejora considerable en la calidad. La Figura 13 ilustra cómo se conservan más detalles, aunque aún se nota una pérdida de calidad en comparación con las originales. Este nivel de compresión logra un equilibrio entre la reducción de dimensionalidad y la preservación de detalles importantes.

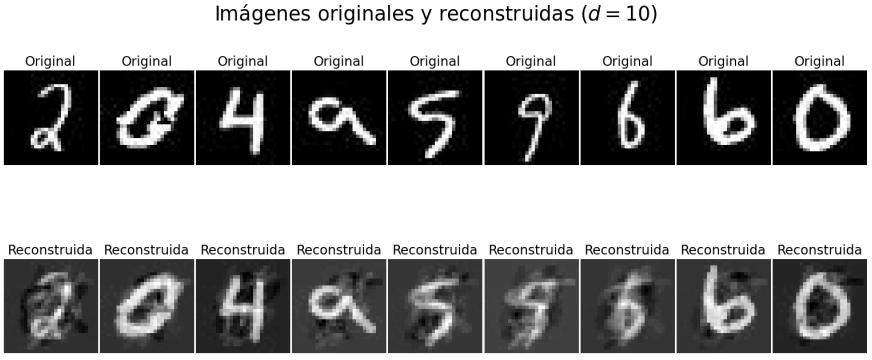


Fig. 13: Imágenes originales y reconstruidas con $d = 10$.

Finalmente, para $d = 19$, las reconstrucciones son bastante fieles a las imágenes originales, como se observa en la Figura 14. Este resultado muestra que una mayor dimensionalidad captura mejor las características importantes de las imágenes, logrando una representación que es casi indistinguible de las originales.

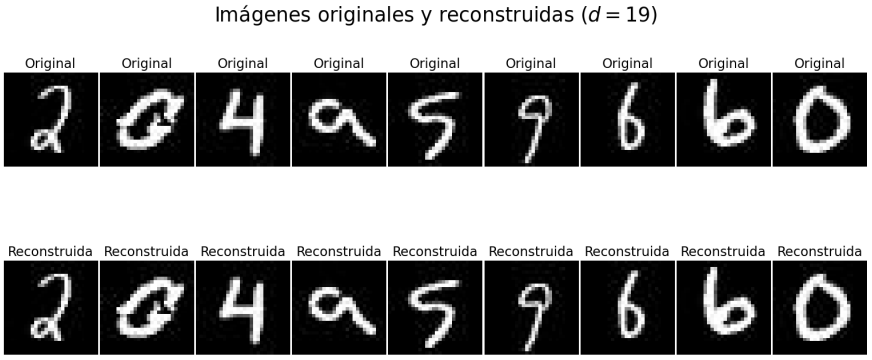


Fig. 14: Imágenes originales y reconstruidas con $d = 19$.

Cabe destacar que, más allá de $d = 19$, no se observan mejoras visuales significativas en la calidad de la reconstrucción (ver Apéndice B). Esto sugiere que después de alcanzar $d = 19$, no se obtienen beneficios adicionales en la reconstrucción de las imágenes debido al tamaño limitado del dataset. Con solo 19 imágenes, las primeras 19 dimensiones ya capturan la mayoría de las variaciones y estructuras presentes en los datos. A medida que se aumenta d, las dimensiones adicionales tienen menos variabilidad para capturar, lo que limita la capacidad de mejorar la representación visual.

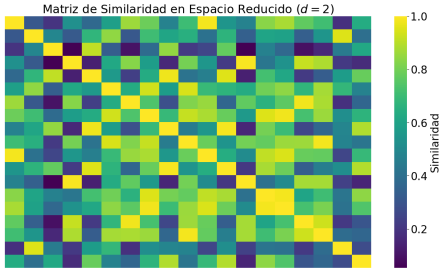
Además, para profundizar el análisis, se compararon las reconstrucciones con $d = 2$ y $d = 10$ utilizando los últimos autovectores (que explican menor varianza). La Figura 15 ilustra este resultado.

Fig. 15: Comparación de imágenes originales y reconstruidas utilizando los últimos autovectores para diferentes valores de d. Se aprecia una reconstrucción significativamente menos precisa.

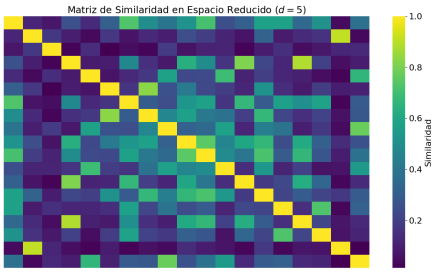
La reconstrucción de imágenes utilizando los últimos autovectores es notablemente menos precisa en comparación con la utilización de los primeros autovectores. Este fenómeno se explica por la naturaleza de la descomposición en valores singulares (SVD). En la SVD, los valores singulares (autovalores) están ordenados de mayor a menor, con los primeros autovalores capturando la mayor parte de la varianza y la estructura principal de los datos.

Los últimos autovectores corresponden a direcciones con menor varianza y contienen información menos relevante. Como resultado, las reconstrucciones basadas en estos autovectores carecen de detalles significativos y presentan una calidad reducida. Esto se observa claramente en la Figura 15, donde las imágenes reconstruidas con los últimos autovectores muestran menos precisión y claridad en comparación con la reconstrucción utilizando los primeros, reflejando una representación menos efectiva de la estructura y características importantes de las imágenes originales.

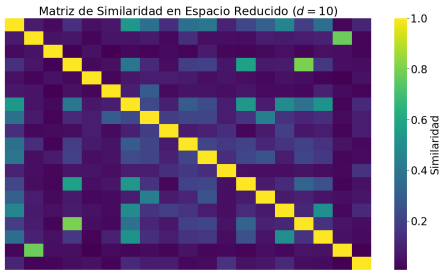
Para medir la similaridad entre pares de imágenes en el espacio reducido, se utilizó una métrica basada en el kernel gaussiano. Las matrices de similaridad para diferentes valores de d muestran cómo cambian las relaciones entre las imágenes con la dimensionalidad (ver Apéndice B).



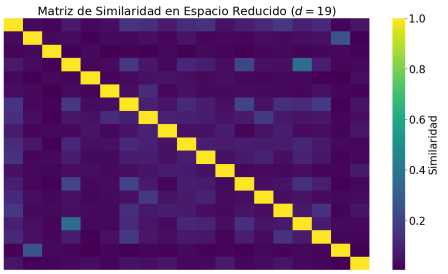
(a) Matriz de Similaridad ($d = 2$)



(b) Matriz de Similaridad ($d = 5$)



(c) Matriz de Similaridad ($d = 10$)



(d) Matriz de Similaridad ($d = 20$)

Fig. 16: Matrices de Similaridad en Espacios Reducidos con Diferentes Valores de d

La Figura 16 muestra las matrices de similaridad para $d = 2$, $d = 5$, $d = 10$, y $d = 19$. Se observa que para $d = 2$ (Figura 16a), los patrones son difusos y hay pocas imágenes claramente agrupadas, indicando que la baja dimensionalidad no captura adecuadamente las relaciones entre imágenes. Este fenómeno se debe a que, con dimensiones bajas, las imágenes reconstruidas no utilizan suficiente información específica de cada imagen, sino que dependen más de la información compartida entre todas las imágenes. Como resultado, hay muchas similitudes entre las imágenes reconstruidas, lo que se refleja en la variedad de colores y el "desorden"visible en la matriz de similaridad.

Para $d = 5$ (Figura 16b), los patrones comienzan a tomar forma más clara, pero todavía hay cierta superposición y falta de definición en los grupos de imágenes. En contraste, cuando $d = 10$, los grupos de similaridad son más evidentes. La reducción en la variedad de colores y un agrupamiento más claro indican que la representación con $d = 10$ (Figura 16c) preserva mejor las características individuales de las imágenes. Finalmente, para $d = 19$ (Figura 16d), la matriz de similaridad es detallada y muestra agrupaciones claras, indicando que esta dimensionalidad preserva bien las relaciones entre

las imágenes. Con una mayor dimensionalidad, las imágenes son reconstruidas con más información específica, lo que resulta en una matriz de similaridad donde predominan los colores uniformes fuera de la diagonal principal.

Este análisis revela que, para dimensiones bajas, las imágenes no son completamente reconstruidas y dependen de la información de otras imágenes, lo que resulta en matrices de similaridad con muchas similitudes entre las imágenes. A medida que se aumenta la dimensionalidad, las reconstrucciones se vuelven más precisas y las relaciones entre las imágenes se preservan mejor, reflejándose en matrices de similaridad con patrones más claros y definidos.

Además del análisis de reconstrucción y similaridad, se exploraron los primeros cuatro autovectores singulares para identificar las direcciones más significativas en el espacio de características. Las Figuras 17 y 18 muestran estos autovectores, que representan combinaciones difuminadas de los números en los datasets.

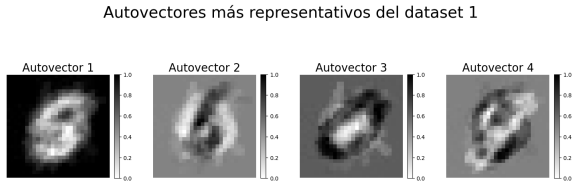


Fig. 17: Primeros cuatro autovectores singulares del dataset 1

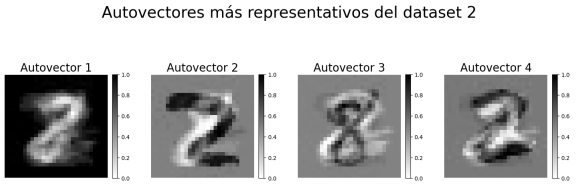


Fig. 18: Primeros cuatro autovectores singulares del dataset 2

Estos autovectores corresponden a las componentes principales que capturan la mayor parte de la variabilidad en los datos, proporcionando una base fundamental para la reconstrucción y el análisis de similaridad. Al visualizar estos autovectores, se observa que combinan patrones de los diferentes números, reflejando las características más relevantes y comunes en las imágenes del dataset.

Finalmente, se realizó una reconstrucción cruzada utilizando las imágenes del dataset 1 y el dataset 2. Específicamente, se intentó reconstruir imágenes del dataset 1 usando las características de las imágenes del dataset 2. La Figura 19 muestra nueve imágenes del dataset 1 y sus correspondientes reconstrucciones basadas en el dataset 2.

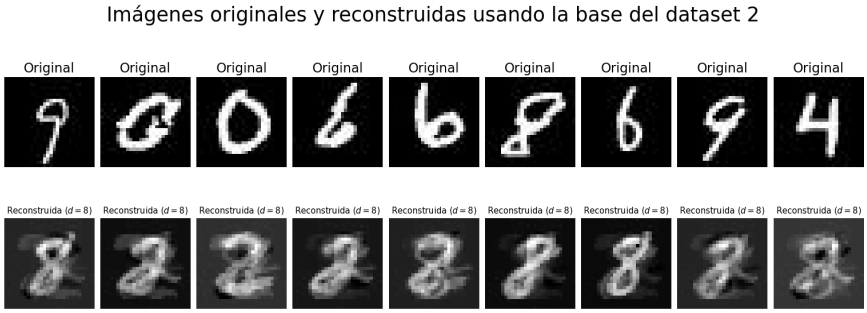


Fig. 19: Reconstrucción cruzada de imágenes del dataset 1 utilizando el dataset 2 mostrando la capacidad del modelo para generalizar entre conjuntos de datos distintos.

La reconstrucción de imágenes utilizando la operación $A^* = AVV^T$ implica la transformación de vectores a una base reducida y su retorno al espacio original con menos direcciones. En este proceso,

la matriz V^T rota los vectores del espacio original a uno reducido, preservando solo las componentes más relevantes. Al aplicar la matriz V, los vectores vuelven al espacio original, aunque con una pérdida de información debido a la reducción inicial de dimensiones.

En este análisis, se parte de un espacio de dimensiones 24×24, donde V^T se aplica utilizando autovectores específicos para retener solo las componentes esenciales que representan los números 2 y 8 en el dataset 2, reduciendo así a 8 componentes. Al aplicar V, se obtiene un espacio de dimensión 8 generado por estos autovectores. Sin embargo, estos vectores, aunque existen en un espacio de 24×24, no pueden generar todas las imágenes originales debido a la reducción de dimensiones.

Así, la operación $A^* = AVV^T$ implica que A actúa como una matriz que reduce el espacio dimensional y luego lo retorna al espacio original con menos componentes. Esta operación proyecta todo el espacio en las direcciones principales definidas por los números 2 y 8, lo que resulta en una reconstrucción de A basada en las imágenes del dataset 2. De esta manera, las imágenes del dataset 1 fueron reconstruidas según su similaridad con la base del dataset 2.

Para determinar el número mínimo de dimensiones d necesarias para que el error de reconstrucción no exceda el 10 % bajo la norma de Frobenius, se analizó cómo el error varía con d. En la Figura 20 se muestra la relación entre el error de reconstrucción y el valor de d.

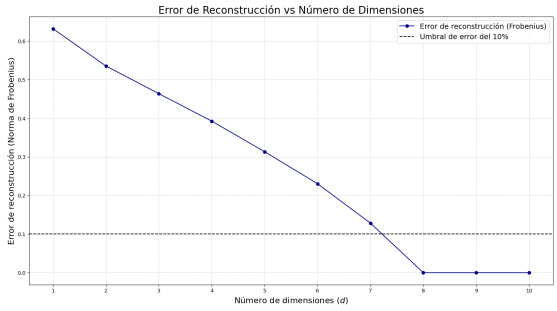


Fig. 20: Error de reconstrucción en función del número de dimensiones.

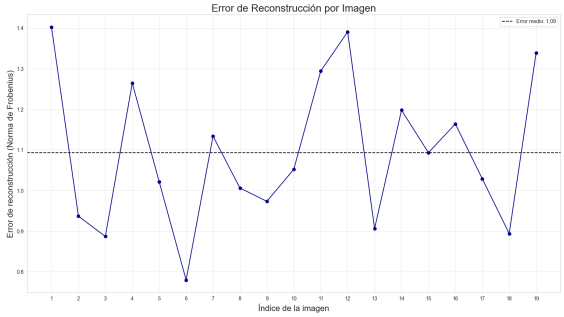


Fig. 21: Error de reconstrucción para cada imagen en el conjunto de datos.

La figura 20 revela que el error de reconstrucción disminuye a medida que aumenta d. Encontrar el valor mínimo de d que mantiene el error de reconstrucción dentro del 10 % es crucial para lograr una compresión eficaz sin perder calidad significativa, y como puede apreciarse en el gráfico, el valor mínimo que cumple con esa condición resultó ser $d = 8$ (procedimiento en Apéndice C).

Al mismo tiempo, la Figura 21 ofrece una visión detallada del error de reconstrucción para cada imagen en el conjunto de datos. Se observa que algunas imágenes contienen características o patrones más complejos, los cuales son más difíciles de reconstruir con un número reducido de dimensiones, lo que conduce a un incremento en el error de reconstrucción. Por otro lado, las imágenes con características más simples o menor variabilidad pueden ser reconstruidas con mayor precisión, incluso con un número menor de dimensiones, lo que se refleja en una disminución del error. El error de reconstrucción que se obtuvo al reconstruir las imágenes del dataset 1 con la base de $d = 8$ dimensiones obtenidas del dataset 2 fue aproximadamente 0.7406.

Este análisis proporciona una comprensión clara de cómo la reducción de dimensionalidad puede afectar la representación y comparación de imágenes, ofreciendo una base sólida para aplicaciones prácticas en compresión y análisis de imágenes.

7. Conclusiones

La investigación realizada proporciona una comprensión profunda de dos aspectos clave en el análisis de datos: la reducción de la dimensionalidad y el ajuste de modelos utilizando el método de cuadrados mínimos.

En primer lugar, la investigación se centra en comprender la distribución y similaridad entre muestras en el espacio original de alta dimensionalidad, destacando la influencia crucial de la elección de la métrica de similaridad, especialmente el kernel gaussiano, en la interpretación y la identificación de clusters en los datos. Además, se explora la reducción de dimensionalidad del conjunto de datos X mediante la descomposición en valores singulares (SVD) y la aplicación de PCA. Este análisis resalta la importancia de elegir la dimensión reducida d de manera adecuada para capturar la estructura subyacente de los datos minimizando la pérdida de información y proporciona insights valiosos sobre cómo las dimensiones originales del conjunto de datos se relacionan con las dimensiones reducidas obtenidas por SVD, permitiendo tomar decisiones informadas sobre la conveniencia de diferentes valores de d para el análisis.

En segundo lugar, se ha explorado la compresión de imágenes utilizando técnicas de SVD. Se ha aprendido una representación de baja dimensión para las imágenes, lo que ha permitido reconstruirlas con diferentes valores de d. La visualización de las imágenes reconstruidas ha proporcionado una comprensión clara de cómo la calidad de la reconstrucción varía con d, lo que ha llevado a conclusiones importantes sobre la cantidad de información retenida en diferentes niveles de compresión.

En resumen, este estudio destaca la importancia de considerar cuidadosamente la dimensionalidad de los datos y la elección de d en el análisis y modelado de datos. Al comprender cómo la reducción de la dimensionalidad y el ajuste de modelos se interrelacionan, se pueden tomar decisiones más informadas en el diseño y aplicación de técnicas analíticas para una variedad de aplicaciones prácticas. Este enfoque proporciona una base sólida para futuras investigaciones y aplicaciones en el campo del análisis de datos y el aprendizaje automático.

Apéndice A: Procedimiento de Cálculos para SVD y PCA

En este apéndice, se detalla el procedimiento de cálculos necesarios para realizar la descomposición de X en sus valores singulares, reducir la dimensionalidad y proyectar las muestras en un nuevo espacio reducido. Se muestran los cálculos para $d = 6$ (mismo procedimiento para cualquier d).

Procedimiento de Cálculo

1. Descomposición en Valores Singulares (SVD)

Dados los datos X de dimensión $m \times p$:

$$X = U\Sigma V^T$$

donde U es una matriz $m \times m$, Σ es una matriz diagonal $m \times p$ con los valores singulares, y V^T es una matriz $p \times p$.

2. Reducción de Dimensionalidad

Seleccionamos los primeros d valores singulares y sus vectores correspondientes:

$$X_d \approx U_d \Sigma_d V_d^T$$

donde U_d es una matriz $m \times d$, Σ_d es una matriz diagonal $d \times d$, y V_d^T es una matriz $d \times p$.

3. Proyección al Espacio Reducido

Proyectamos los datos X en el nuevo espacio reducido Z:

$$Z = XV_d^T$$

Cálculos Específicos para $d = 6$

Primero, realizamos la SVD de X: $X = U\Sigma V^T$.

Luego, tomamos las primeras 6 columnas de U y V, y los primeros 6 valores de Σ:

$$U_6 = U[:, 0:6]$$
$$\Sigma_6 = \Sigma[0:6, 0:6]$$
$$V_6^T = V^T[0:6, :]$$

Finalmente, proyectamos X en el espacio reducido:

$$Z_6 = XV_6^T$$

Este mismo procedimiento se aplica para otros valores de d, como $d = 2$ y $d = 10$, simplemente ajustando el número de columnas y valores tomados de U, Σ y V.

Apéndice B: Impacto del Parámetro σ y de la Dimensionalidad

Durante el análisis, se exploró el impacto conjunto del parámetro σ y la dimensionalidad d en la representación y análisis de datos mediante matrices de similaridad.

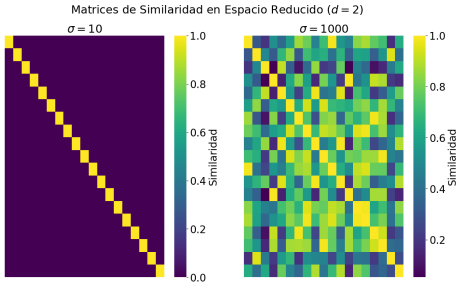


Fig. 22: Comparación de Matrices de Similaridad para $\sigma = 10$ y $\sigma = 1000$.

Fig. 23: Comparación de reconstrucciones para $d = 19$ y $d = 50$. A pesar de aumentar la dimensionalidad, no se observan mejoras.

En la Figura 22, se ilustra esta influencia comparando la matriz de similaridad para $d = 2$ calculada con $\sigma = 10$ y $\sigma = 1000$. Un valor bajo de σ puede no capturar las variaciones relevantes entre las muestras, mientras que un valor más alto permite una mejor diferenciación y formación de clusters más precisos. Este aspecto subraya la importancia de seleccionar cuidadosamente σ para optimizar la interpretación y análisis de los datos.

Además, la Figura 23 muestra la comparación entre reconstrucciones para $d = 19$ y $d = 50$, en donde se observa que después de alcanzar $d = 19$, no se aprecian mejoras visuales significativas en la calidad de las reconstrucciones, lo cual demuestra que extenderse más allá de las dimensiones del dataset no aporta beneficios adicionales en la representación visual de los datos.

Apéndice C: Reducción de Dimensionalidad y Reconstrucción de Imágenes

Para encontrar el número mínimo de dimensiones d a los que se puede reducir la dimensionalidad del dataset `dataset_imagenes2.zip`, primero se define el error de reconstrucción como la diferencia entre cada imagen comprimida y su original bajo la norma de Frobenius:

$$\text{Error de Reconstrucción} = ||A - A'||_F$$

donde $A' = U_d \Sigma_d V_d^T$ es la matriz reconstruida con las d dimensiones principales obtenidas mediante la descomposición en valores singulares (SVD) del dataset `dataset_imagenes2.zip`.

Se realiza la SVD del dataset `dataset_imagenes2.zip` para diferentes valores de d, desde 1 hasta un máximo predefinido. Por cada valor de d, se calcula A' y se determina el error de reconstrucción. Se busca el valor mínimo de d tal que el error de reconstrucción sea menor o igual al 10 %.

Una vez obtenido el valor mínimo de d, se utiliza la misma base de d dimensiones obtenida del dataset 2 para reconstruir las imágenes del dataset `dataset_imagenes1.zip`. Se calcula el error de reconstrucción para este nuevo conjunto de imágenes.

Este proceso permite determinar el número mínimo de dimensiones necesarias para reducir la dimensionalidad del dataset `dataset_imagenes2.zip` y evaluar cómo se generaliza esta representación aprendida al reconstruir imágenes de otro conjunto de datos.

Apéndice D: Aplicación de Filtros de Convolución

Los filtros de convolución son una herramienta fundamental en el procesamiento de imágenes y señales. Se utilizan para extraer características importantes, reducir el ruido y resaltar patrones específicos en los datos. En este apéndice, se aplican filtros de convolución a los datos originales para mejorar la calidad de la representación visual y la precisión del análisis.

Se aplicaron diferentes tipos de filtros de convolución a las imágenes del conjunto de datos, incluyendo filtros de suavizado (como el filtro de promedios y el filtro Gaussiano) y filtros de detección de bordes (como el filtro de Sobel y el filtro de Canny). Los pasos específicos para aplicar cada filtro son los siguientes:

- **Filtro de Promedios**: Se utilizó una máscara de 3×3 para suavizar la imagen, reduciendo el ruido y las fluctuaciones menores.
- **Filtro Gaussiano**: Se aplicó un filtro Gaussiano con un sigma de 1.0 para suavizar la imagen y eliminar el ruido de alta frecuencia.
- **Filtro de Sobel**: Se utilizó para detectar bordes en la imagen, calculando las derivadas en las direcciones x e y.
- **Filtro de Canny**: Se implementó para una detección de bordes más precisa, utilizando umbrales de histéresis para evitar la detección de bordes falsos.

A continuación, se presentan las imágenes resultantes después de aplicar cada tipo de filtro. Estas imágenes muestran cómo cada filtro resalta diferentes características de las imágenes originales.

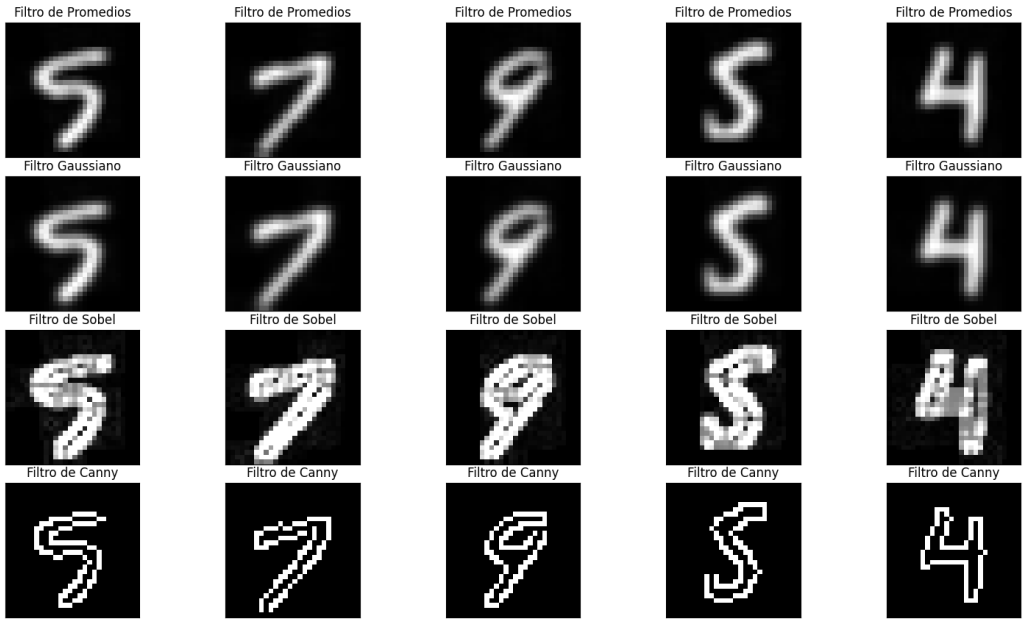


Fig. 24: Resultados de la Aplicación de Filtros de Convolución en la Reconstrucción de las imágenes.

El análisis de las imágenes filtradas muestra que los filtros de convolución son efectivos para resaltar diferentes características de las imágenes. El filtro de promedios y el filtro Gaussiano son útiles para la reducción de ruido, mientras que los filtros de Sobel y Canny son más adecuados para la detección de bordes y la extracción de características geométricas importantes.

Bibliografía

[1] Richard L. Burden and J. Douglas Faires. *Numerical Analysis*. Cengage Learning, 9th edition, 2010.

[2] L. N. Trefethen and D. Bau III. *Numerical Linear Algebra*. SIAM, 1997.

[3] Ricardo G. Durán, Silvia B. Lassalle and Julio D. Rossi. *Elementos de cálculo numérico*. Editorial de la Universidad de Buenos Aires, 2007.

[4] I.T. Jolliffe. *Principal Component Analysis*. Springer, 2nd edition, 2002.

[5] J. Scott Armstrong and Fred Collopy. *Error Measures For Generalizing About Forecasting Methods: Empirical Comparisons*. *International Journal of Forecasting*, 8(1):69–80, 1992.

[6] Geosci. *Root mean square error (RMSE) or mean absolute error (MAE)? –Arguments against avoiding RMSE in the literature. Geosci. Model Dev.*, 7:1247–1250, 2014.

[7] Carina Pfister. *SVD Tutorial*. `http://carina.fcaglp.unlp.edu.ar/mpp/notebooks/SVD.html`

[8] Gilbert Strang *Linear Algebra and Learning from Data*. `https://math.mit.edu/~gs/learningfromdata/`

[9] James, G., Witten, D., Hastie, T., Tibshirani, Python. (2013) *An Introduction to Statistical Learning*. `https://hastie.su.domains/ISLP/ISLP_website.pdf.download.html`