

Práctica 6

Análisis de componentes principales

Javier Herrer Torres (NIP: 776609)



Aprendizaje automático
Grado en Ingeniería Informática



**Escuela de
Ingeniería y Arquitectura
Universidad** Zaragoza

Escuela de Ingeniería y Arquitectura
Universidad de Zaragoza
Curso 2020/2021

1. Objetivo

El objetivo es utilizar técnicas de reducción de dimensión basadas en el análisis de componentes principales. En la primera parte, utilizaremos SVD para comprimir una imagen. En la segunda parte evaluaremos el uso de PCA para la reconstrucción de dígitos y su clasificación. Para ello deberéis tener preparado el clasificador que utilizasteis con MNIST.

2. Estudio previo

Descomposición en valores singulares

Sea M una matriz de tamaño $m \times n$. M es factorizable de la siguiente manera:

$$M_{m \times n} = U_{m \times m} \Sigma_{m \times n} V_{n \times n}^T$$

- $U_{m \times m}$ ortogonal, genera el espacio de filas de M .
- Σ diagonal no negativa, contiene los valores singulares λ_i .
- V^T ortogonal, genera el espacio de las columnas de M .

3. Compresión de imágenes

Descomposición en valores singulares

Se ha utilizado SVD para comprimir una foto de mi cara (*script P61.m*). Para simplificar, se ha utilizado una imagen en escala de grises. Se han seguido las instrucciones que aparecen como comentarios en el programa.

Una vez se ha leído la imagen, convertido a blanco y negro y los datos a *double*; se ha aplicado SVD (*Singular Value Descomposition*).

Después, se han graficado las primeras 5 componentes. En la figura 1, cada término captura variaciones a diferente frecuencia.

Posteriormente, en la figura 2 se ha graficado la reconstrucción mediante sumas parciales.

Escoger el valor de k

Para encontrar el valor de k que mantiene al menos el 90 % de la variabilidad se ha empleado:

$$\frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^n \lambda_i} > 0,9$$

, obteniendo $k = 143$.

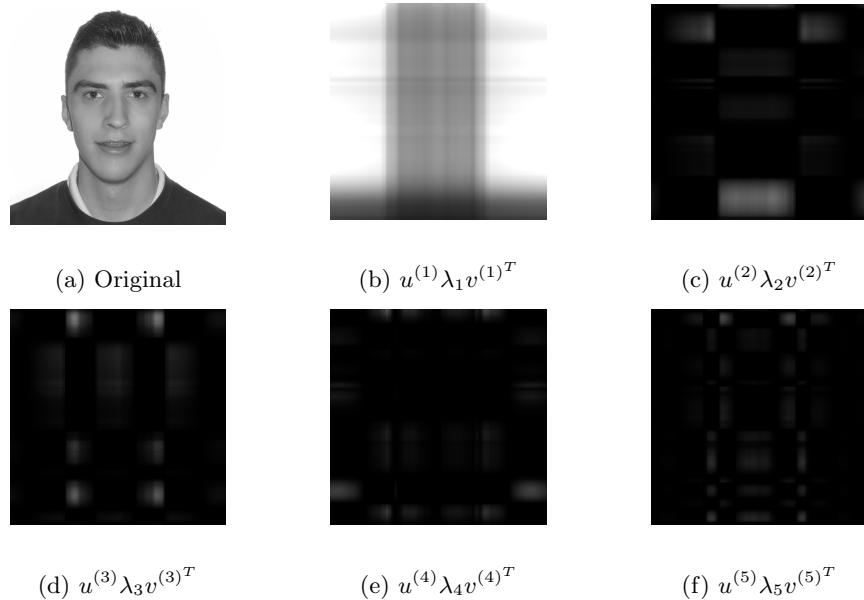


Figura 1: Primeras 5 componentes

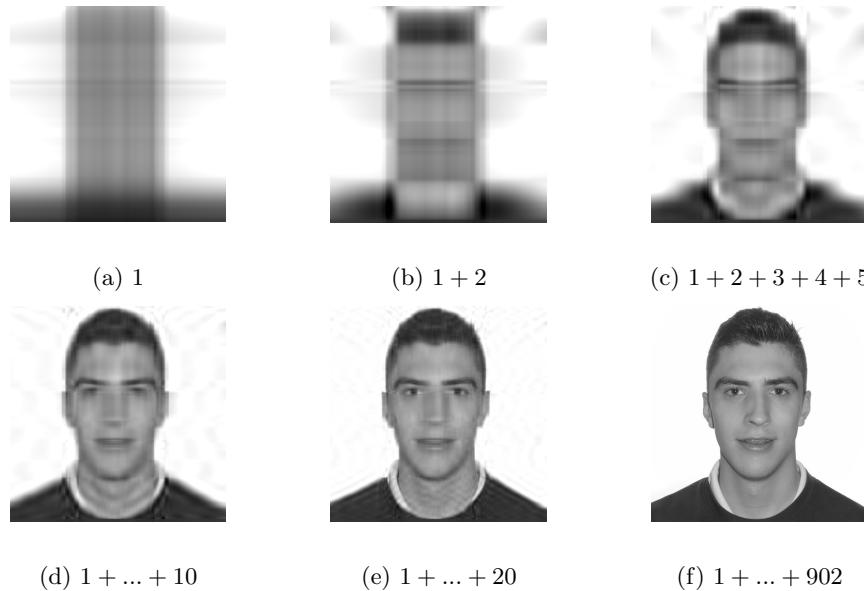


Figura 2: Reconstrucción mediante sumas parciales

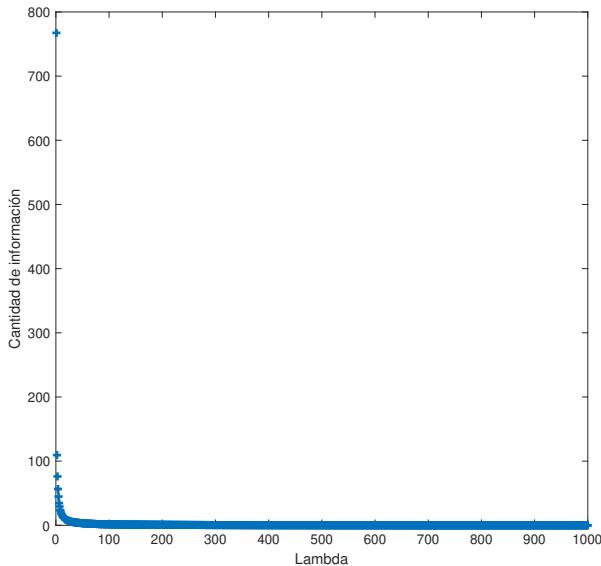


Figura 3: Valores singulares

Ratio de compresión

En la figura 3 se observa la cantidad de información que aporta cada valor propio. Se observa que λ_1 aporta 8 veces más información que la siguiente: λ_2 . Esa diferencia continua reduciéndose a medida que aumentamos la lambda. Mediante la compresión tratamos de escoger un valor de k que haga que la cantidad de información que se pierde sea despreciable.

Por otro lado, en la figura 4 se puede observar que el espacio que ocupa la imagen original es 500 veces más, guardando únicamente la primera componente. El ahorro se va reduciendo a medida que guardo más componentes. Si ampliamos esta gráfica podemos ver el punto en el que el ratio es menor que 1, es decir, la imagen original ocupa menos espacio. Esto se produce a partir de la componente 500 (incluida).

Además, para el valor seleccionado de $k = 143$, se obtiene que *la imagen original es 3,49476 veces mayor*.

4. PCA sobre MNIST

Se ha empleado el mismo conjunto de imágenes que se utilizó en la práctica 5. Se ha utilizado la misma función del apartado 2 para visualizar las componentes (*script P62.m*).

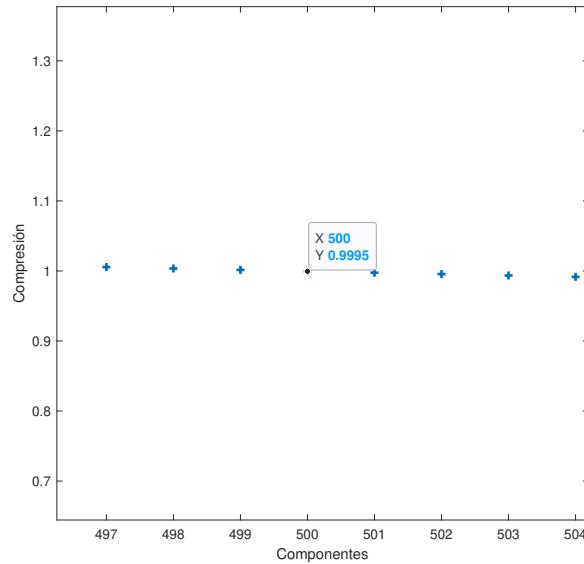
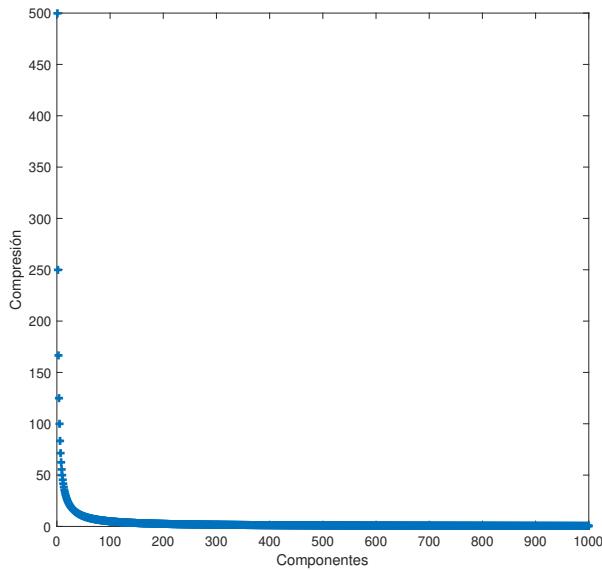


Figura 4: Ratio de compresión

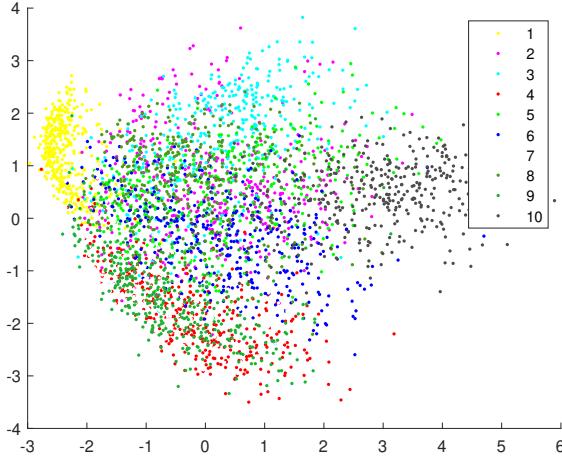


Figura 5: Representación de las dos componentes principales

Visualizar sobre 2 componentes

En primer lugar, se han cargado los datos y han sido permutados aleatoriamente. Posteriormente, se han estandarizado los datos. Es decir, se ha calculado μ y se le ha restado a los m datos de la siguiente manera:

$$x^{(1)} - \mu, x^{(2)} - \mu, \dots, x^{(m)} - \mu$$

Después, se ha aplicado PCA y se han representado en la figura 5 las dos componentes principales de cada imagen con un color diferente para cada dígito.

Se ha calculado la correlación mediante:

$$\rho_{ij} = \frac{\sum_{k=1}^m (x_i^{(k)} - \bar{x}_i)(x_j^{(k)} - \bar{x}_j)}{\sqrt{\sum_{k=1}^m (x_i^{(k)} - \bar{x}_i)^2} \sqrt{\sum_{k=1}^m (x_j^{(k)} - \bar{x}_j)^2}}$$

, obteniendo $\rho_{ij} = 0$. Es decir, que las componentes no están correladas.

Además, de la figura se puede extraer la siguiente conclusión: no se pueden emplear estos datos de menor dimensión para obtener buenos resultados en la clasificación ya que, excepto los dígitos 1 y 0, el resto de clases se entremezclan en la representación.

Clasificación de dígitos

Tal y como se ha mencionado en el análisis del anterior apartado, los dos dígitos que se pueden clasificar usando solo estas dos componentes son el 1 y el 0. Además, también se puede afirmar que los dígitos 2 y 9 no son clasificables mediante estas dos componentes.

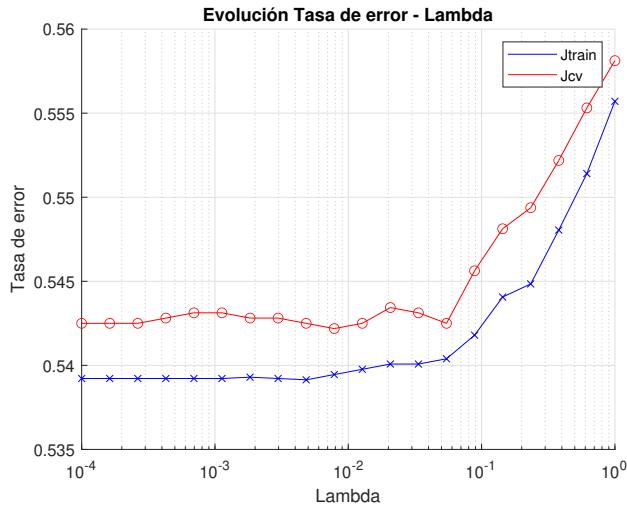


Figura 6: Gráfica de la tasa de errores en función del parámetro landa

Para comprobar estas afirmaciones se ha usado el clasificador de la práctica anterior. En primer lugar se ha separado un 20% de los datos para test. Después, se ha empleado kfold-cross-validation para escoger el mejor parámetro de regularización. En la figura 6 se encuentra dibujada la gráfica de tasa de errores en función del parámetro landa. Se puede observar sub-ajuste a ambos lados del punto ideal que se encuentra en:

$$\text{best_}\lambda = 0,0078$$

Posteriormente, se ha entrenado con ese mejor valor de λ y se han obtenido las siguientes tasas de error:

$$E_{\text{train}}(\theta) = 0,54$$

$$E_{\text{test}}(\theta) = 0,5350$$

, que confirman la conclusión del anterior apartado.

Se ha re-entrenado con todos los datos para el mejor valor de λ , y se han utilizado los datos de test para calcular la matriz de confusión de la figura 7 junto con los valores de precisión y recall para cada dígito, y los globales. Se observa que los dígitos mejor clasificados son el 1 y el 0. Y los dos dígitos con más problemas a la hora de clasificar son el 2 y el 9. Por lo tanto, se confirman las suposiciones realizadas previamente. Es decir, si sólo se consideraran los dígitos 1 y 0 en lugar de todas las clases, los resultados de la clasificación mejorarían.

Dígito Predicho													
Dígito Real	1	2	3	4	5	6	7	8	9	0	Suma	Recall	
	1	81					1				82	0,9878	
	2	4	15		10	11	5	24	1	9	79	0	
	3	1	53		3	9		19			85	0,6235	
	4	3		46		2	30	1	6		88	0,5227	
	5	8	13		7	17	6	23	8	5	87	0,0805	
	6	5		1	5	2	48	3	7	6	5	82	0,5854
	7	5		23		6	34	1	4		73	0,4658	
	8	2	20	3	8	8		34	2		77	0,4416	
	9	1		19		5	39		6	1	71	0,0845	
	0		1		5	7				63	76	0,8289	
Suma	110	0	103	96	35	114	117	109	33	83	800	0,4621	
Precision	0,74	#####	0,51	0,48	0,20	0,42	0,29	0,31	0,18	0,76	#DIV/0!		

Figura 7: Matriz de confusión