Reducción de Dimensionalidad

Rafael Villca Poggian
Álgebra Lineal II

1. Introducción

En el análisis de datos para el ajuste de modelos de Machine Learning, es útil tener una idea de la estructura de la información como puntos en $\mathbb{R}^n$, estos se representan como una matriz $X \in \mathcal{M}_{m\times n}(\mathbb{R})$, dónde "punto" es un vector fila $x \in \mathbb{R}^n$. Cuando $n > 3$ no se puede graficar la el dataset, es debido a esto que se desarrollan métodos para proyectar los puntos en dimensiones altas a baja dimensionalidad, tal que represente lo mayor posible de su geometría.

Existen distintas maneras de descomponer una matriz en operaciones simples que tengan interpretación geométrica, para aprovechar estas propiedades y manipular la estructura de los datos.

2. Diagonalización

También llamada eigendescomposición, es una factorización matricial expresada en función de los eigenvectores y eigenvalores de una matriz.

Geométricamente, una matriz $A \in \mathcal{M}_{m\times n}(\mathbb{R})$ es una función que toma puntos de R^n y los mapea en R^m, manteniendo las líneas paralelas, por lo que se puede expresar la transformación de la base canónica como $A \cdot I$.

Para la eigendescomposición se requiere que A sea cuadrada de $n \times n$, y diagonalizable, de modo que se factoriza como

$$A = V \cdot D \cdot V^{-1} \tag{1}$$

dónde V tiene como columnas v_i los eigenvectores de A, y D es una matriz diagonal con entradas $D_{ii} = \lambda_i$ los eigenvalores de A.

Por el teorema espectral, si A es simétrica, sus eigenvectores son ortogonales y todos sus eigenvalores son reales, de modo que si se normalizan las columnas de V, esta se convierte en una matriz ortogonal, y la fórmula expresión se convierte en:

$$A = V \cdot D \cdot V^{T} \tag{2}$$

Si se considera la expresión equivalente $A = VDV^{-1}I$, la interpretación geométrica es que se expresa I en la base V de los eigenvectores, para luego escalar los ejes por D y restaurar la transformación.

Así, la exponenciación matricial se puede ver como un escalado sucesivo de estos ejes en V, y en el caso que $\exists_i : \lambda_i = 0$ la información en la dirección v_i colapsa, y es imposible de restaurar la correspondencia, por lo que una matriz con algún eigenvalor 0 no es invertible.

3. Descomposición de Valor Singular

Se considera la generalización de la eigendescomposición para matrices no cuadradas, y toda matriz tiene descomposición SVD.

3.1. Valores singulares

Sea $A \in \mathcal{M}_{m\times n}(\mathbb{R})$ una matriz no cuadrada, la matriz A^TA es simétrica positiva semi-definida, de modo que $\forall_i$, tiene eigenvalores $\lambda_i \geq 0$ y n eigenvectores v_i ortogonales entre si.

Desarrollando:

$$\begin{aligned} A^TAv_i &= \lambda_i v_i \\ \langle A^TAv_i, \lambda_i v_i\rangle &= \langle \lambda_i v_i, \lambda_i v_i\rangle \\ \lambda_i\langle A^TAv_i, v_i\rangle &= \lambda_i^2\langle v_i, v_i\rangle \\ \langle Av_i, Av_i\rangle &= \lambda_i\langle v_i, v_i\rangle \\ ||Av_i|^2 &= \lambda_i||v_i||^2 \\ \lambda_i &= \frac{||Av_i||^2}{||v_i||^2} \\ \sqrt{\lambda_i} &= \frac{||Av_i||}{||v_i||} \end{aligned} \tag{3}$$

Si se normaliza cada vector v_i y se lo renombra a x_i tal que $||x_i|| = 1$, la fórmula 3 se convierte en:

$$\sqrt{\lambda_i} = ||Ax_i|| \tag{4}$$

Además sean μ_i los eigenvalores de A en caso que fuera cuadrada, con $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_k > 0$, se tiene que:

$$\begin{aligned} \langle Ax_i, Ax_i\rangle &= \langle \mu_i x_i, \mu_i x_i\rangle \\ &= \mu_i^2\langle x_i, x_i\rangle \\ ||Ax_i||^2 &= \mu_i^2 \\ \lambda_i &= \mu_i^2 \end{aligned}$$

Por lo que

$$\sqrt{\lambda_i} = \mu_i \tag{5}$$

de este modo se relacionan los eigenvalores de A con la raíz cuadrada de los eigenvalores de A^TA. A los $\sqrt{\lambda_i}$ se les llama valores singulares de A, y son la generalización de sus autovalores en caso de que no sea cuadrada y no existan estos.

3.2. Vectores singulares

Si trabajamos a partir de ahora con los autovectores normalizados de A^TA, y se desarrollan como sigue:

$$\begin{aligned} A^TAx_i &= \lambda_i x_i \\ AA^T(Ax_i) &= \lambda_i Ax_i \\ AA^Tu_i &= \lambda_i u_i \end{aligned}$$

se tiene que $u_i = Ax_i$ son los eigenvectores de AA^T, además observamos que $A^TA_{n\times n}$ y $AA^T_{m\times m}$ comparten los primeros $min\{n, m\}$ eigenvalores distintos de 0, y por 5, tenemos que:

$$\begin{aligned} u_i &= Ax_i \\ ||u_i|| &= ||Ax_i|| \\ ||u_i|| &= \mu_i \\ ||u_i|| &= \sqrt{\lambda_i} \end{aligned}$$

con esto podemos escribir la fórmula de los eigenvectores normalizados de AA^T como

$$y_i = \frac{Ax_i}{\sqrt{\lambda_i}} \tag{6}$$

a los x_i y y_i se les llama vectores singulares.

3.3. La descomposición

La SVD se puede obtener a partir de la ecuación 6 reescrita como $y_i\sqrt{\lambda_i} = Ax_i$, llevándola a una forma matricial apilando las columnas, se tiene que

$$\begin{bmatrix} | & | & & | \\ y_1\sqrt{\lambda_1} & y_2\sqrt{\lambda_2} & \dots & y_m\sqrt{\lambda_m} \\ | & | & & | \end{bmatrix} = A \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \dots & x_n \\ | & | & & | \end{bmatrix}$$

$$\begin{bmatrix} | & | & & | \\ y_1 & y_2 & \dots & y_m \\ | & | & & | \end{bmatrix} \begin{bmatrix} \begin{bmatrix} \sqrt{\lambda_1} & 0 & & 0 \\ 0 & \sqrt{\lambda_2} & & 0 \\ & & & \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & & \sqrt{\lambda_m} \end{bmatrix} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} = A \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \dots & x_n \\ | & | & & | \end{bmatrix}$$

$$\begin{aligned} U\Sigma &= AV \\ A &= U\Sigma V^{-1} \end{aligned}$$

Como A^TA y AA^T son simétricas, sus eigenvectores son ortogonales entre si respectivamente, es por esto que si se construyen matrices con sus eigenvectores normalizados como columnas, las matrices resultantes serán ortogonales, es debido a esto que la fórmula de la factorización SVD se escribe como

$$A = U\Sigma V^T \tag{7}$$

Como U y V son matrices ortogonales, estas preservan las distancias, por lo que son isometrías que representan rotación o reflexión respecto a un eje.

3.4. Geometría del SVD

Si A es una matriz tal que

$$\begin{aligned} A\colon \mathbb{R}^n &\to \mathbb{R}^m \\ x &\mapsto y. \end{aligned}$$

entonces se tiene que las transformaciones realizadas por los factores de su descomposición son

$$\begin{array}{ccc} \mathbb{R}^n & \xrightarrow{A} & \mathbb{R}^m \\ {\scriptstyle V^T}\downarrow & & \uparrow{\scriptstyle U} \\ \mathbb{R}^n & \xrightarrow[\Sigma]{} & \mathbb{R}^m \end{array}$$

El procedimiento geométrico es similar al de la eigendescomposición, una interpretación útil es considerar la esfera unitaria en $\mathbb{R}^n$, aplicarle una rotación con V^T tal que ahora los vectores singulares v_i son los nuevos ejes canónicos, estirar sus ejes mediante Σ, lo cual nos da los semiejes de la elipse proyectada en $\mathbb{R}^m$, para finalmente restaurar la rotación original con V.

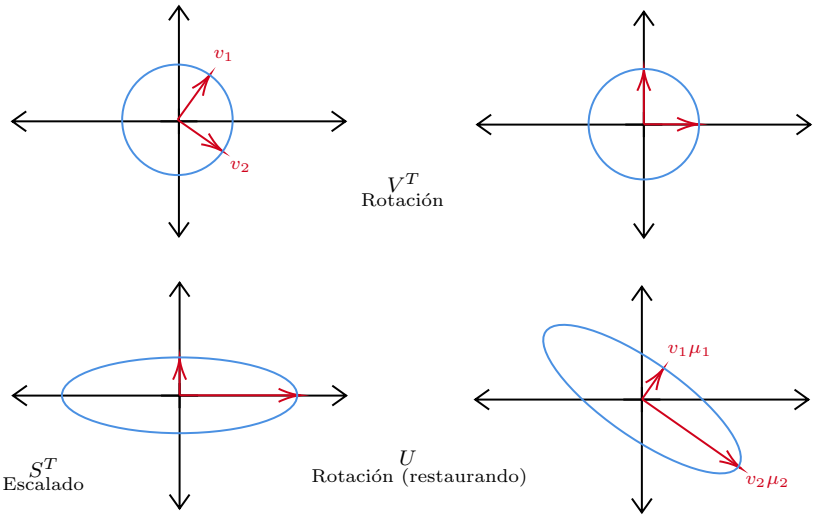

Figura 1: Interpretación gráfica de las etapas del SVD

3.5. SVD compacto

Cuando se descompone la matriz $A_{m\times n}$, si $m > n$, se tienen $|m-n|$ filas de ceros al final de Σ, de manera similar, si $n > m$, se tienen $|m-n|$ columnas.

Esta idea se extiende a cuando $\rho(A) = k \leq min\{m, n\}$ con $\rho(A)$ el rango de A, como algunos valores singulares serán 0, se eliminan filas de V^T sin pérdida de información.

Un ejemplo considerando $\mu_m = 0$ debido a que $n = m+1$ y $\rho(A) = m$

$$A = \begin{bmatrix} | & & | & | \\ y_1 & \dots & y_{m-1} & y_m \\ | & & | & | \end{bmatrix} \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \dots & x_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} \begin{bmatrix} \mu_1 & & 0 & 0 \\ & \ddots & 0 & 0 \\ 0 & & \mu_{m-1} & 0 \\ 0 & & 0 & \mu_m \end{bmatrix} & \mathbf{0} \end{bmatrix} \begin{bmatrix} — & x_1^T & — \\ & \vdots & \\ — & x_{n-1}^T & — \\ — & x_n^T & — \end{bmatrix}$$

el resultado de multiplicar

$$\begin{bmatrix} \begin{bmatrix} \mu_1 & & 0 & 0 \\ & \ddots & 0 & 0 \\ 0 & & \mu_{m-1} & 0 \\ 0 & & 0 & \mu_m \end{bmatrix} & \mathbf{0} \end{bmatrix} \begin{bmatrix} — & x_1^T & — \\ & \vdots & \\ — & x_{n-1}^T & — \\ — & x_n^T & — \end{bmatrix} \text{ y } \begin{bmatrix} \begin{bmatrix} \mu_1 & & 0 & 0 \\ & \ddots & 0 & 0 \\ 0 & & \mu_{m-1} & 0 \\ 0 & & 0 & \mu_m \end{bmatrix} & \mathbf{0} \end{bmatrix} \begin{bmatrix} — & x_1^T & — \\ & \vdots & \\ — & x_{n-1}^T & — \end{bmatrix}$$

es el mismo, ya que la fila n se vuelve cero debido al vector al final de Σ.

3.6. SVD Truncado

El método anterior permite almacenar la información con menos peso, basándose en esa idea, si se está dispuesto a sacrificar la precisión de la reconstrucción, se pueden hacen cero los μ_i más pequeños, de esta manera, se eliminan las filas o columnas de V^T asociadas a esos valores singulares como en el ejemplo anterior.

Si nos quedamos con los primeros k valores singulares, estamos reduciendo el rango de A artificialmente, de modo que al reconstruir la matriz, se obtiene A_k tal que $\rho(A) = k$, la cual es una aproximación de bajo rango de la matriz A original, pero a cambio de esta pérdida de precisión, se reducen las dimensiones de todas las matrices factores a $U_{m\times k}$, $\Sigma_{k\times k}$ y $V^T_{k\times n}$.

3.7. Reducción de dimensionalidad con SVD

Las columnas de V forman una base para las filas de A, por la interpretación geométrica del SVD, cada vector base de V representa uno de los ejes de un elipsoide que encierra a los datos, dónde μ_i indica cuánto se estira la esfera unitaria en cada dirección.

Cuando los valores singulares son muy pequeños, geométricamente en las direcciones asociadas la esfera se "aplasta" más que en las demás, de modo que si se vuelven ceros, colapsa esas dimensiones y se proyecta sobre un subespacio de menor dimensionalidad.

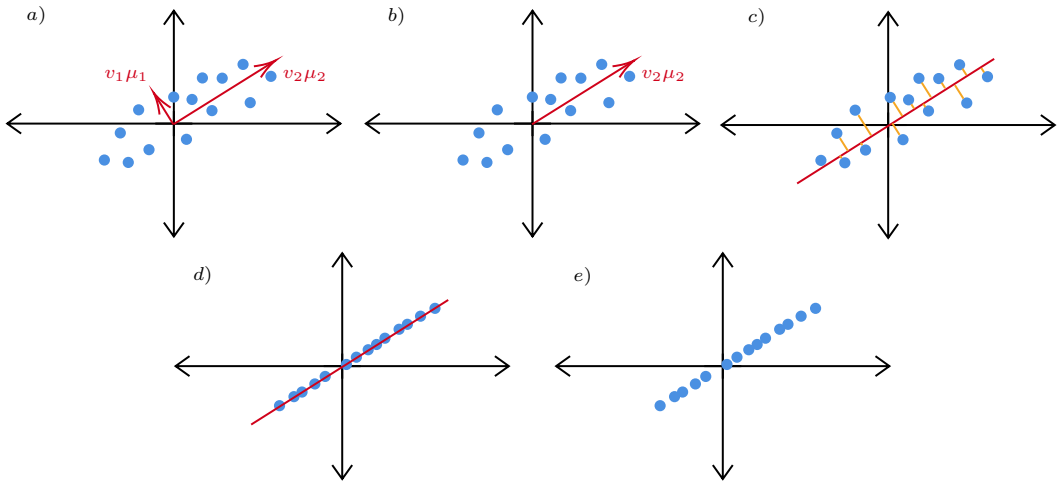

Figura 2: a) vectores singulares y puntos de la matriz A, b) se anula el menor valor singular $\mu_1 = 0$, c) se observa la correspondencia ortogonal de los puntos sobre $span v_1$, d) se proyectan los puntos sobe el subespacio 1D y se reconstruyen, e) puntos de la matriz A_k

Así la proyección ortogonal de los vectores se define como:

$$AP = \tilde{A} \tag{8}$$

con $P = \begin{bmatrix} | & & | \\ x_1 & \dots & x_k \\ | & & | \end{bmatrix}$ las primeras k columnas de V

De manera similar al ser sus columnas ortonormales, la reconstrucción está dada por

$$\tilde{A}P^T = A_k \tag{9}$$

Con $A_k \approx A$ la aproximación de bajo rango.

Nótese que no se utiliza la matriz diagonal, ya que interesa rotación y el colapso de las dimensiones pequeñas, pero mantener los ejes grandes sin tocar, para luego restaurar la rotación con la información sacrificada por espacio.

4. Análisis de Componentes Principales

Es una técnica usada para la reducción de dimensionalidad, compresión y visualización de datos, también llamada la transformación de Karhunen-Loève. De manera similar a la proyección mediante SVD, el método proyecta puntos $x \in \mathbb{R}^n$ no dispersos en un espacio lineal de menor dimensión llamado *subespacio principal* de manera que se minimice la suma del cuadrado de las distancias entre el punto x y un hiperplano en $\mathbb{R}^m$, también llamado el costo medio de proyección o reconstrucción.

Teorema 1. *Sea A una matriz simétrica con valores propios $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ y los vectores propios correspondientes $v_1, v_2, \ldots, v_n$, entonces el máximo de la forma cuadrática*

$$x^T Ax \quad \textit{sujeto a } ||x|| = 1$$

se alcanza en el vector propio v_1 asociado a λ_1 tal que x sea ortogonal a los demás vectores propios.

Generalizando

Teorema 2. *Sea A una matriz simétrica con valores propios $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ y los vectores propios correspondientes $v_1, v_2, \ldots, v_n$, entonces el máximo de*

$$tr(V^T AV) \quad \textit{sujeto a } V^T V = I$$

se alcanza en la matriz cuyas columnas son los vectores propios v_i ordenados de manera que v_1 es la primera columna, v_2 la segunda y así sucesivamente hasta v_n la última columna de V.

Demostración. Las formas cuadráticas $v_i^T Av_i$ componen la diagonal de $V^T AV$, entonces por el teorema 1, el máximo de cada expresión, está dado por el eigenvector v_i asociado al eigenvalor λ_i más grande de A.

Así el máximo de $v_1^T Av_1$ está dado por λ_1 tal que $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_k$, pero el máximo de $v_2^T Av_2$ no puede ser λ_1 ya que sino $\langle v_1, v_2 \rangle \neq 0$ no son ortogonales y no se cumple la restricción $V^T V = I_k$, así, como el máximo dado por v_1 es una cota superior, el siguiente vector que nos da el valor más grande menor que la cota, es v_2 asociado a λ_2. Continuando con este razonamiento inductivo, tenemos que el k-ésimo máximo menor a los $k-1$ anteriores está dado por el eigenvector v_k asociado al eigenvalor λ_k de A.

Construyendo V como $V = \begin{bmatrix} | & & | \\ v_1 & \ldots & v_k \\ | & & | \end{bmatrix}$, al multiplicar las matrices $V^T AV$ los elementos de sus diagonales son $v_i^T Av_i$ tales que $v_1^T Av_1 \geq v_2^T Av_2 \geq \cdots \geq v_n^T Av_n > x^T Ax$ con x cualquier vector de $\mathbb{R}^n$ que cumpla $x \neq v_1, \ldots, x \neq v_n$ □

La diferencia entre PCA y SVD está en que en PCA se trabaja sobre los datos centrados restando la media, de este modo las operaciones se realizan sobre la matriz de convarianzas, permitiendo una mejor interpretación de los estadísticos de la información.

Por definición la covarianza está dada por $Cov(X,Y) = \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu_x)(y_i - \mu_y)$, de modo que su generalización en forma matricial considerando una matriz $X_{m \times n}$ con cada punto en $\mathbb{R}^n$ una fila de X

$$K_{XX} = \frac{1}{m} \left(X - E(X)\right)^T \left(X - E(X)\right) \tag{10}$$

Así su definición se resume en que queremos encontrar un conjunto ortogonal de k vectores base $w_i \in \mathbb{R}^n$, y su correspondiente proyección $z \in \mathbb{R}^k$, tal que se minimice el error medio de reconstrucción

$$\mathcal{J} = \frac{1}{m}\sum_{i=1}^{m} ||x_i - \hat{x}_i||^2 \tag{11}$$

con $\hat{x}_i = Wz_i$. Además se restringe W tal que $W^T \cdot W = I_{k\times k}$.

Reescribiendo $\mathcal{J}$ en términos de la norma de Frobenius dada por $||X||_F = \sqrt{tr(X^T X)}$, se tiene

Teorema 3. *Sea una función objetivo a minimizar, siendo X los datos con la media sustraida por columnas*

$$\mathcal{J} = ||X - \hat{X}||_F^2 \tag{12}$$

la solución optima se obtiene en $\tilde{W}_{n\times k}$ con w_i los eigenvectores asociados a los k eigenvalores más grandes de la matriz de covarianzas K como sus columnas, y además la codificación óptima en menos dimensiones está dada por $z_i = W^T x_i$, la cual es una proyección ortogonal de los datos en el espacio de columnas generado por los eigenvectores.

Demostración. La técnica está condicionada por la elección del la función de restauración del vector codificado al más al cercano original, para esto , se elige una matriz $W_{n\times k}$ tal que la función de decodificación es

$$\begin{aligned} g\colon \mathbb{R}^k &\to \mathbb{R}^z \\ z &\mapsto Wz. \end{aligned}$$

Ahora debemos definir una función que devuelva el código óptimo z^* dado un punto x. Formulando el problema de optimización como:

$$z^* = \underset{z}{\operatorname{argmin}} ||x - g(z)||_2^2 \tag{13}$$

reescribiendo en forma vectorial tenemos

$$\begin{aligned} z^* &= \underset{z}{\operatorname{argmin}}\ (x - g(z))^T (x - g(z)) \\ &= \underset{z}{\operatorname{argmin}}\ x^T x - x^T g(z) - g(z)^T x + g(z)^T g(z) \\ &= \underset{z}{\operatorname{argmin}}\ x^T x - 2x^T g(z) + g(z)^T g(z) \\ &= \underset{z}{\operatorname{argmin}}\ x^T x - 2x^T Wz + (Wz)^T Wz \\ &= \underset{z}{\operatorname{argmin}}\ x^T x - 2x^T Wz + z^T W^T Wz \\ &= \underset{z}{\operatorname{argmin}}\ x^T x - 2x^T Wz + z^T I_k z \\ &= \underset{z}{\operatorname{argmin}}\ x^T x - 2x^T Wz + z^T z \\ &\text{Minimzando igualando el gradiente a 0} \\ &\Rightarrow \nabla_z (x^T x - 2x^T Wz + z^T z) = 0 \\ &\Rightarrow -2x^T W + 2z = 0 \\ &\Rightarrow z = x^T W = W^T x \end{aligned}$$

Así la codificación óptima está dada por $z = f(x) = W^T x$. Considerando que los vectores x^T son las

filas de X, entonces

$$\begin{aligned} z &= W^T x \\ Wz &= WW^T x \\ (Wz)^T &= (WW^T x)^T \\ z^T W^T &= x^T WW^T \end{aligned}$$

Apilando los vectores x^T y verticalmente como filas de X, se tiene

$$\tilde{X} = XWW^T \quad \text{y} \quad Z = XW \tag{14}$$

Ahora se debe encontrar $\tilde{W}$ tal que minimice la distancia entre los puntos originales y su reconstrucción mediante la formulación:

$$\begin{aligned} W^* &= \underset{W}{\text{argmin}} \ \mathcal{J} \\ &\text{Sujeto a} : W^T W = I_k \end{aligned} \tag{15}$$

Por 14 en 12, se desarrolla como:

$$\begin{aligned} W^* &= \underset{W}{\text{argmin}} \ ||X - XWW^T||_F^2 \\ &= \underset{W}{\text{argmin}} \ tr\left(\left(X - XWW^T\right)^T \left(X - XWW^T\right)\right) \\ &= \underset{W}{\text{argmin}} \ tr(X^T X - X^T XWW^T - WW^T X^T X + WW^T X^T XWW^T) \\ &= \underset{W}{\text{argmin}} \ tr(X^T X) - tr(X^T XWW^T) - tr(WW^T X^T X) + tr(WW^T X^T XWW^T) \end{aligned}$$

El término sin W es constante para la minimización y se descartan sin pérdida de generalidad

$$= \underset{W}{\text{argmin}} \ - tr(X^T XWW^T) - tr(WW^T X^T X) + tr(WW^T X^T XWW^T)$$

La traza es invariante a las permutaciones cíclicas de sus argumentos

$$\begin{aligned} &= \underset{W}{\text{argmin}} \ - 2tr(X^T XWW^T) + tr(WW^T X^T XWW^T) \\ &= \underset{W}{\text{argmin}} \ - 2tr(X^T XWW^T) + tr(X^T XWW^T WW^T) \end{aligned}$$

Por la restricción $W^T W = I_k$

$$\begin{aligned} &= \underset{W}{\text{argmin}} \ - 2tr(X^T XWW^T) + tr(X^T XWI_k W^T) \\ &= \underset{W}{\text{argmin}} \ - 2tr(X^T XWW^T) + tr(X^T XWW^T) \\ &= \underset{W}{\text{argmin}} \ - tr(X^T XWW^T) \end{aligned}$$

Minimizar $-f(x)$ es equivalente a maximizar $f(x)$

$$= \underset{W}{\text{argmax}} \ tr(X^T XWW^T)$$

Así nos queda resolver el problema de maximización

$$\begin{aligned} W^* &= \underset{W}{\text{argmax}} \ tr(W^T X^T XW) \\ &\text{Sujeto a} : W^T W = I_k \end{aligned} \tag{16}$$

Por el teorema 2, el máximo de $tr(W^T X^T XW)$ está dado por $W = \begin{bmatrix} | & & | \\ w_1 & \dots & w_k \\ | & & | \end{bmatrix}$

Así el menor error de reconstrucción $\mathcal{J}$ al proyectar los puntos en X de $\mathbb{R}^n$ a $\mathbb{R}^k$ está dado por la matriz W cuyas columnas son los eigenvectores w_i asociados a los eigenvalores λ_i más grandes de X^TX, la matriz de covarianzas de X sin el término de escalado $\frac{1}{m}$, el cual se puede descartar sin pérdida de generalidad. □

4.1. PCA y SVD

Para encontrar la matriz de proyección de PCA, calculamos los eigenvectores y eigenvalores de la matriz de covarianzas $K = \frac{1}{m}X^TX$, lo cual es equivalente a encontrar los valores singulares y los vectores singulares derechos de X la matriz de datos estandarizados.

Es por esta razón que si no se estandarizan los datos restando la media en cada columna, PCA y SVD dan resultados ligeramente diferentes, además como la información está centrada, al calcular los valores singulares, está encontrado las direcciones de máxima varianza dada por los vectores singulares derechos y en PCA llamados **Componentes Principales**, por eso al quedarnos con los valores singulares más grandes, estamos manteniendo la mayor cantidad posible de varianza original de los datos.

Para obtener la varianza explicada por cada vector singular, se ponen en un vector columna $\mathbf{s}$ los valores singulares y se calcula la matriz de covarianza empirica de estos

$$C_{\mathbf{s}} = \frac{\mathbf{s}\mathbf{s}^T}{m-1} \tag{17}$$

en la diagonal de $C_{\mathbf{s}}$ se encuentran las variazas explicadas por cada componente principal, si se divide entre la varianza total y se suman los primeros k términos, se tiene el porcentaje total explicado.

Referencias

[1] Charu C. Aggarwal. *Linear Algebra and Optimization for Machine Learning - A Textbook*. Springer, 2020.

[2] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[3] Kevin P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press, Cambridge, Mass. [u.a.], 2013.

[4] Gilbert Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, Wellesley, MA, fourth edition, 2009.