Polo del Conocimiento

Pol. Con. (Edición núm. 67) Vol. 7, No 2
Febrero 2022, pp. 1884-1907
ISSN: 2550 - 682X
DOI: 10.23857/pc.v7i2.3687

Detalles de la derivación e implementación de algoritmo básico para la reducción de dimensionalidad con PCA

Details on the algorithm for dimensionality reduction via Principal Components Analysis

Detalhes da derivação e implementação do algoritmo básico para redução de dimensionalidade com PCA

Zenaida Natividad Castillo-Marrero [I]
zenaida.castillo@espoch.edu.ec
https://orcid.org/0000-0002-4424-8652

Gustavo Adolfo Colmenares-Pacheco [II]
gcolmenares@yachaytech.edu.ec
https://orcid.org/0000-0003-4789-0859

Ramón Antonio Abancin-Ospina [III]
ramon.abancin@espoch.edu.ec
https://orcid.org/0000-0002-2417-6671

Víctor Oswaldo Cevallos-Vique [IV]
victor.cevallos@espoch.edu.ec
https://orcid.org/0000-0001-5525-5818

Correspondencia: zenaida.castillo@espoch.edu.ec

Ciencias Técnicas y Aplicadas
Artículo de Investigación

***Recibido:** 04 de enero de 2022 ***Aceptado:** 31 de enero de 2022 * **Publicado:** 21 de febrero de 2022

I. Escuela Superior Politécnica de Chimborazo (ESPOCH), Facultad de Ciencias, Escuela de Matemática, Grupo CIDED, Riobamba, Ecuador.

II. Universidad de Investigación de Tecnología Experimental Yachay, Escuela de Ciencias Matemáticas y Computacionales, Urcuquí, 060150, Ecuador.

III. Escuela Superior Politécnica de Chimborazo (ESPOCH), Facultad de Ciencias, Escuela de Matemática, Grupo CIDED, Riobamba, Ecuador.

IV. Escuela Superior Politécnica de Chimborazo (ESPOCH), Facultad de Administración y Empresas, Escuela de Finanzas, Grupo CIDED, Riobamba, Ecuador.

Resumen

Los problemas que requieren de análisis de datos, a menudo son difíciles de resolver, debido principalmente a la cantidad de variables involucradas en el modelo matemático. Los científicos de datos generalmente trabajan con millones de variables para hacer estimaciones que soportan decisiones importantes. En el procesamiento digital de imágenes, por ejemplo, el número de puntos que representan pixeles en tres dimensiones podría muy grande en imágenes a color. En estos casos, el costo computacional que requiere el manejo de estos datos puede resultar inaceptable, y la reducción de dimensionalidad de estos datos se hace necesaria. Aún cuando se cuente con la tecnología adecuada, la reducción en tiempo de cómputo siempre es deseable. El manejo de datos con alta dimensionalidad, el análisis y la interpretación se dificulta y en el caso de imágenes, su visualización podría verse afectada considerando las limitaciones de memoria. En la mayoría de los casos, estos datos son redundantes, y la información importante puede revelarse con solo parte de los mismos. La reducción de dimensionalidad es el proceso mediante el cual se descarta parte de la data que no aporta información relevante; y uno de los métodos más usados en todos los ámbitos es el análisis de componentes principales, o PCA, el cual se basa en el cálculo de algunos autovalores de la matriz de covarianzas de los datos. En este trabajo revisamos las herramientas matemáticas detrás del análisis del PCA, y detallamos los pasos de un algoritmo para reducir dimensionalidad que luego es implementado en Matlab. Se presenta también un ejemplo de aplicación para ilustrar el proceso.

Palabras claves: PCA; Reducción de dimensionalidad; Autovalores; Análisis de datos.

Abstract

In data analysis, we often deal with millions of variables to make estimations for decisions. Also, in digital image processing the number of data points representing pixels in three dimensions could become very large for color images. In all these cases computational costs for solving associated problems could not be affordable, and a reduction of dimensionality it is necessary. When working with high dimensional data, analysis and interpretation of results is difficult, and visualization of images could be prohibited in terms of memory. Frequently the data contains redundancy, and the important information can be found just with part of the data. In a color image, for example, with four channels, red, blue, green and grey-scale channel, one could combine the first three channels to get the information of the gray-scale channel. Dimensional

reduction is a technique that allow us to work with a reduced part of the data without lost losing important information, and one of the methods used for dimensional reduction is the principal analysis component or PCA, which is based in the computation of a few eigenvalues of the covariance matrix. In this work we review the math tools and the steps to generate a classical algorithm for dimensionality reduction using PCA. We present mathematical elements and concepts involved in a principal component analysis algorithm for the projection of high dimensional data in a lower dimensional subspace. A practical example is presented to show how these mathematical tools work together to produce similar results in lower dimensional spaces.
Keywords: PCA; Dimensionality reduction; Eigenvalues; Data analysis.

Resumo

Problemas que requerem análise de dados são muitas vezes difíceis de resolver, principalmente devido ao número de variáveis envolvidas no modelo matemático. Os cientistas de dados normalmente trabalham com milhões de variáveis para fazer estimativas que apoiam decisões importantes. No processamento digital de imagens, por exemplo, o número de pontos representando pixels em três dimensões pode ser muito grande em imagens coloridas. Nesses casos, o custo computacional necessário para lidar com esses dados pode ser inaceitável, e a redução da dimensionalidade desses dados torna-se necessária. Mesmo quando a tecnologia apropriada está disponível, a redução do tempo de computação é sempre desejável. O manuseio de dados com alta dimensionalidade, análise e interpretação é difícil e, no caso de imagens, sua visualização pode ser afetada por limitações de memória. Na maioria dos casos, esses dados são redundantes e informações importantes podem ser reveladas apenas com parte deles. A redução de dimensionalidade é o processo pelo qual parte dos dados que não fornecem informações relevantes é descartada; e um dos métodos mais utilizados em todas as áreas é a análise de componentes principais, ou PCA, que se baseia no cálculo de alguns autovalores da matriz de covariância dos dados. Neste artigo, revisamos as ferramentas matemáticas por trás da análise de PCA e detalhamos as etapas de um algoritmo para reduzir a dimensionalidade que é implementado no Matlab. Um exemplo de aplicação também é apresentado para ilustrar o processo.
Palavras-chave: PCA; Redução de dimensionalidade; Autovalores; Analise de dados.

Introducción

La reducción de dimensionalidad no es más que la determinación de un número de variables mínimo con el cual se puedan representar los datos, y consiste en eliminar variables que no sean relevantes o que no aporten suficiente información. Actualmente existen técnicas o algoritmos para hacer esta tarea, la cual ha pasado a ser una actividad obligada en disciplinas como la ciencia de datos, una vez que simplifica el manejo de los datos y al mismo tiempo reduce el costo computacional en tiempo de procesamiento y utilización de memoria. Una técnica usada para reducir la dimensionalidad es eficiente mientras logra esta reducción de costos al mismo tiempo que mantiene una buena representación del modelo.

En la actualidad muchas aplicaciones dependen del manejo apropiado de datos, y a menudo los datos suelen presentarse en muchas dimensiones, dependiendo de la cantidad de variables involucradas, que en principio se justifican para la toma de decisiones; por lo tanto se espera que al tener representación en muchas dimensiones la información sea más completa. Sin embargo, el manejo y la interpretación de los datos se hace más complicado, sin contar el costo computacional que a veces es imposible de afrontar, véase por ejemplo (James, 2013)

La reducción de dimensionalidad explota la estructura del problema respetando la correlación entre variables, lo cual permite un manejo más eficiente de los datos sin pérdida de la información relevante.

Este trabajo está orientado a describir en forma simple para lectores que se inician en el tópico, las herramientas matemáticas que se usan en el proceso de la reducción lineal de dimensionalidad a través del análisis de componentes principales o PCA (por sus siglas en inglés). Se mostrará la derivación y posterior implementación en Matlab del algoritmo clásico que define esta técnica para el análisis de datos.

En las próximas secciones estaremos definiendo conceptos básicos asociados al análisis de datos a través de la estadística, con el objetivo de identificar el aporte del análisis de componentes principales al manejo eficiente y significativo de datos con altas dimensiones.

Medidas estadísticas empíricas de los datos

Generalmente, el manejo de los datos, a través de una muestra de ellos, a fin de obtener información de interés, comienza por establecer parámetros estadísticos como la media y la varianza, que nos dan información sobre la dispersión de los datos. En esta sección se presentan

estos conceptos que miden la variabilidad de los datos. Comenzaremos considerando datos en una dimensión (1D) y luego extenderemos los conceptos a dos 2D o más dimensiones.

Sea $\mathrm{D} = \{x_1, x_2, \cdots, x_N\}$, un conjunto de datos u observaciones en una dimensión donde cada $x_i \in \mathbb{R}$, se define la media del conjunto como:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

En este caso, cada $\mathbf{x_i}$ representa una característica o variable que se desea estudiar.

La varianza se define como:

$$V[\mathrm{D}] = s^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

En algunas aplicaciones se divide por $N - 1$ para que sea insesgado.

La varianza, como estadístico de dispersión, mide la variabilidad de los datos de su medida central o media aritmética. Una medida que suele darnos más información sobre los datos es la desviación estándar **s**, o la raíz cuadrada de la varianza, la cual indica el rango de la variabilidad en los datos. La desviación estándar se indica en las mismas unidades de la media, mientras que la varianza se especifica en unidades al cuadrado por lo que no es tan fácil conjeturar sobre la información que brinda.

Se entiende por población a un conjunto bien definido de elementos; por ejemplo, los habitantes de una comunidad o los estudiantes de una Universidad. Si por cada elemento de una población se mide un conjunto de variables estaremos hablando de estadística multivariante, y los estadísticos como la media y la varianza se extienden apropiadamente. Por ejemplo, en el caso de los estudiantes de una Universidad pudiéramos recoger datos como la edad, género, promedio de calificaciones, carrera que cursa, etc., y por cada estudiante obtener esta información con el fin de analizarla para tomar alguna decisión gerencial.

Cada característica o variable puede tomar valores cuantitativos o cualitativos; sin embargo, en la práctica los variables cualitativas como color o género podrán representarse en una escala numérica; de esta manera si tenemos n individuos y medimos k características o variables podemos utilizar una matriz para representar estos datos:

Pol. Con. (Edición núm. 67) Vol. 7, No 2, febrero 2022, pp. 1884-1907, ISSN: 2550 - 682X

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} = \begin{bmatrix} \vec{x}_1^{\,T} \\ \vec{x}_2^{\,T} \\ \vdots \\ \vec{x}_n^{\,T} \end{bmatrix}$$

Donde $\vec{x}_i^{\,T}$ es el vector que representa las características del i-ésimo individuo. En este caso tendremos un vector de medias con **k** componentes, cada una representando la media de los valores de la variable respectiva.

$$\vec{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_k \end{bmatrix}, \text{ donde } \bar{x}_i = \frac{1}{N} \vec{x}_i^{\,T} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

La dispersión o variabilidad de los datos en el caso multivariable puede ser analizada a través de las relaciones lineales entre las variables y representada mediante una matriz llamada matriz de varianzas y covarianzas, o simplemente matriz de covarianza.

$$S = \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1k} \\ s_{21} & s_2^2 & \cdots & s_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ s_{k1} & s_{k2} & \cdots & s_{kk} \end{bmatrix}$$

Las entradas de la diagonal de esta matriz son las varianzas de cada variable, mientras que las entradas fuera de la diagonal son las covarianzas o los indicadores bivariantes. La entrada S_{ij} ($i \neq j$) de esta matriz mide la variación conjunta de las variables x_i, y x_j; si S_{ij} es positiva significa que a valores altos/bajos de la variable xi podemos esperar valores altos/bajos de xi. Si el valor de S_{ij} es negativo significa que la covariación es en sentido inverso. Si la covarianza es cero no se puede establecer una relación entre la variabilidad e las variables.

El cálculo de S también puede realizarse como:

$$S = \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{x})(X_i - \bar{x})^T \in \mathbb{R}^{D \times D}$$

Para efectos prácticos es importante reconocer que la matriz S es simétrica y semidefinida positiva, lo cual garantiza, entre otras cosas, que:

i. Para cualquier vector v de $\mathbb{R}^D$: $v^T S v \geq 0$.

ii. Todos los autovalores de S son positivos o cero.

La covarianza de dos variables x, y, con medias $\bar{x}$, $\bar{y}$ respectivamente, puede calcularse directamente como:

Pol. Con. (Edición núm. 67) Vol. 7, No 2, febrero 2022, pp. 1884-1907, ISSN: 2550 - 682X

$$S = Cov(x,y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N}$$

Cuando los datos están centrados alrededor de la media, la esperanza o valor esperado de la variable aleatoria es E[x]=0. En este caso la matriz de covarianzas es:

$$S = \frac{1}{N}\sum_{i=1}^{N} X_i X_i^T \in \mathbb{R}^{D\times D}$$

Otro estadístico importante que señala en qué proporción una variable ***y*** está explicada por la influencia lineal de otra variable ***x***, es el coeficiente de correlación lineal, denotado por $r(x,y)$.

$$r(x,y) = \frac{Cov(x,y)}{s_x s_y}$$

El coeficiente de correlación lineal toma valores en [-1,1], si $r = 1$ se dice que las variables están directamente relacionadas si una crece/decrece la otra crece/decrece; y si $r = -1$, la relación es contraria, si una crece/decrece la otra decrece/crece. Cuando $r = 0$ decimos que las variables no están correlacionadas.

Una explicación detallada de estos estadísticos con ejemplos incluidos puede ser hallada en (Salazar, 2017, Rencher, 1994).

Transformaciones lineales de los datos

En muchas aplicaciones se desea conocer lo que pasa con estos estadísticos si a los datos se les suma una cantidad o si se multiplican por un escalar; es decir, si son sometidos a transformaciones de desplazamiento o de compresión/estiramiento. Para analizar matemáticamente esta situación, consideremos un conjunto de datos $\mathbf{D} = \{x_1, x_2, \cdots, x_N\}$, donde

$$x_i \in \mathbb{R} \text{ (1-dimensión)}, \ \mathbf{E}[D] = \mu, \text{ y } \mathbf{V}[D] = \sigma^2.$$

Si cada x_i se modifica como $x_i' = \alpha x_i + \mathbf{c}$, **con** para α **y** **c** dados, entonces,

$$\mathbf{E}[\alpha D] = \alpha \mathbf{E}[D] \wedge \mathbf{E}[D + \mathbf{c}] = \mathbf{E}[\mathrm{D}] + \mathbf{c} \Rightarrow \mathbf{E}[\alpha D + \mathbf{c}] = \alpha \mathbf{E}[D] + \mathbf{c}$$

Ilustremos este resultado con un ejemplo, sea $D = \{1,2,3\}$, y tomemos $\alpha = 3$ y $\mathbf{c} = 2$. Entonces

$$\textbf{i)}\ \mathbf{E}[\mathrm{D}] = (1+2+3)/3 = 2 \text{ y } \mathbf{E}[\alpha D] = \mathbf{E}[3D] = (3+6+9)/3 = 6 = 3\mathbf{E}[D].$$

$$\textbf{ii)}\ \mathbf{E}[D + \mathbf{c}] = \mathbf{E}[D + 2] = (3+4+5)/3 = 4 = \mathbf{E}[D] + 2.$$

$$\textbf{iii)}\ \mathbf{E}[\alpha D + \mathbf{c}] = \mathbf{E}[3D + 2] = (5+8+11)/3 = 8 = 3\mathbf{E}[D] + 2.$$

Pol. Con. (Edición núm. 67) Vol. 7, No 2, febrero 2022, pp. 1884-1907, ISSN: 2550 - 682X

Una situación similar se presenta con la varianza.

$$\mathbf{V}[\alpha D] = \alpha^2 \mathbf{V}[D] \wedge \mathbf{V}[D + \mathbf{c}] = \mathbf{V}[D] \Rightarrow \mathbf{V}[\alpha D + \mathbf{c}] = \alpha^2 \mathbf{V}[D]$$

Ejemplificamos estas propiedades con los datos del ejemplo anterior.

$$\mathbf{i)}\ \mathbf{V}[D] = ((-1)^2 + 0^2 + 1^2)/3 = 2/3, \quad \mathbf{V}[\alpha D] = \mathbf{V}[3D] = (9 + 0 + 9)/3 = 3^2 \mathbf{V}[D].$$

$$\mathbf{ii)}\ \mathbf{V}[D + \mathbf{c}] = \mathbf{V}[D + \mathbf{2}] = ((-1)^2 + 0^2 + 1^2)/3 = 2/3 = \mathbf{V}[D].$$

$$\mathbf{iii)}\ \mathbf{V}[\alpha D + \mathbf{c}] = V[3D + \mathbf{2}] = ((-3)^2 + 0^2 + 3^2)/3 = 6 = 3^2 V[D].$$

El detalle de estas propiedades y su deducción puede ser hallado en (De la Puente, 2018, Rencher 1994).

Algunos conceptos de espacios vectoriales

A continuación revisamos brevemente conceptos del álgebra lineal, dentro del tópico de espacios vectoriales, cuyo entendimiento nos permitirá entender cómo proyectar en espaciós de menor dimensión.

Producto escalar: Sean $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ y $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$ dos vectores de $\mathbb{R}^n$, se define su producto escalar o producto punto como:

$$\boldsymbol{x}^T \boldsymbol{y} = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{y}_i.$$

El producto escalar es un producto interior, o producto interno, este último definido como una función que toma dos elementos de un espacio vectorial **V** y devuelve un número real, usualmente se denota como $<:,:>: V \times V \rightarrow \mathbb{R}$, y satisface las siguientes propiedades:

I. Simetría: para todo par de vectores x, y se cumple que $< \mathbf{x}, \mathbf{y} > = < \mathbf{y}, \mathbf{x} >$.

II. Definida positiva: Para $x \neq 0$ se cumple que $< \boldsymbol{x}, \boldsymbol{x} > = \mathbf{0}$, y $< \mathbf{0}, \mathbf{0} > = \mathbf{0}$.

III. Bilineal: Para toto $x, y, z \in V$, y $\lambda \in \mathbb{R}$ se cumple que:

- $< \boldsymbol{\lambda}\mathbf{x} + \mathbf{y}, \mathbf{z} > = \boldsymbol{\lambda} < \mathbf{x}, \mathbf{z} > + < \mathbf{y}, \mathbf{z} >$
- $< \mathbf{x}, \boldsymbol{\lambda}\mathbf{y} + \mathbf{z} > = \boldsymbol{\lambda} < \mathbf{x}, \mathbf{y} > + < \mathbf{x}, \mathbf{z} >$

La norma de un vector $\boldsymbol{x}$ está vinculada al producto interno $\|\boldsymbol{x}\| = \sqrt{< \boldsymbol{x}, \boldsymbol{x} >}$, y podría ser diferente para dos funciones que definan productos internos distintos. Por ejemplo, en $\mathbb{R}^2$ tomando dos vectores genéricos $\boldsymbol{x} = (\boldsymbol{x_1}, \boldsymbol{x_2})$, $\boldsymbol{y} = (\boldsymbol{y_1}, \boldsymbol{y_2})$ pudiéramos definir el siguiente producto interno:

$$< \boldsymbol{x}, \boldsymbol{y} > = \boldsymbol{x_1 y_1} - \frac{1}{2}\boldsymbol{x_1 y_2} - \frac{1}{2}\boldsymbol{x_2 y_1} + \boldsymbol{x_2 y_2}$$

Una inspección de esta función nos hará concluir que en efecto satisface las propiedades I, II y III, mencionadas previamente, que definen un producto interno. En este producto interno, la longitud del vector $\boldsymbol{x} = (1,1)$ será $\|\mathbf{x}\| = \sqrt{<\mathbf{x},\mathbf{x}>} = \mathbf{1}$, mientras que en el producto punto

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^T\mathbf{x}} = \mathbf{2}.$$

Otro ejemplo de funciones que definen un producto interno es la covarianza, que por definición es bilineal, simétrica y definida positiva.

De igual manera podemos predecir que la distancia entre vectores depende del producto interno que se utilice. Si usamos el producto punto, más generalmente conocido, obtenemos lo que se conoce como la distancia euclídea.

En general, dado un producto interno $<:,:>$ la distancia entre dos vectores **x**, **y** se define como:

$$\mathbf{dist}(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\| = \sqrt{<\mathbf{x}-\mathbf{y},\mathbf{x}-\mathbf{y}>}$$

Otro aspecto geométrico a considerar en el análisis de datos es la orientación de los datos, la cual como veremos en la próxima sección puede establecerse mediante direcciones dadas por vectores. Así, la medida del ángulo entre dos vectores puede decirnos qué tan similar son sus direcciones, y esta medida también está relacionada con el producto interno y la norma de los vectores; por ejemplo:

$$\cos(\theta) = \frac{<\mathbf{x},\mathbf{y}>}{\|\mathbf{x}\|\|\mathbf{y}\|}$$

Ortogonalidad: Se dice que dos vectores, distintos de cero, son ortogonales si su producto interno es cero. Por ejemplo, los vectores $\mathbf{x} = (1,2)$ y $\mathbf{y} = (-2,1)$ son ortogonales ya que su producto escalar es cero; haciendo notar que en el producto interno definido en (1) estos vectores no serían ortogonales ya que

$$<\mathbf{x},\mathbf{y}> = x_1y_1 - \frac{1}{2}x_1y_2 - \frac{1}{2}x_2y_1 + x_2y_2 = -2 - 0.5 + 2 + 2 = 1.5.$$

Una base de un espacio vectorial **V** es un conjunto de vectores de **V** que son linealmente independientes y generan a cualquier otro vector de **V** a través de alguna combinación lineal. Por ejemplo, en $\mathbb{R}^2$ los conjuntos $\mathbf{B_1}$ y $\mathbf{B_2}$ cumplen con tales requisitos.

$$\mathbf{B}_1 = \{(1,0),(0,1\}, \quad \mathbf{B}_2 = \{(1,2),(2,1)\}$$

La base **B1** del ejemplo, conocida como la base canónica de $\mathbb{R}^2$, contiene dos vectores ortogonales entre sí y de norma 1, por lo que decimos que es una base ortonormal de $\mathbb{R}^2$. Existen otras bases con estas características que resultan provechosas cuando se hacen proyecciones a

Pol. Con. (Edición núm. 67) Vol. 7, No 2, febrero 2022, pp. 1884-1907, ISSN: 2550 - 682X

espacios de menor dimensión. Todas las bases de un espacio vectorial tienen el mismo número de vectores, y a este número se le conoce como la dimensión del espacio. El concepto se puede extender a $\mathbb{R}^n$ y también a espacios vectoriales de funciones.

Una base ortonormal de vectores en $\mathbb{R}^n$ es un subconjunto de **n** vectores de $\mathbb{R}^n$, los cuales son ortogonales dos a dos, y cada uno tiene norma 1. Siempre es posible ortonormalizar vectores linealmente independientes para producir una base ortogonal, por ejemplo mediante el proceso de Gram Schmidt (Lay, 2012, Datta, 2010).

Los subespacios de un espacio vectorial V son subconjuntos propios de **V** que también son espacios vectoriales y por lo tanto disponen de bases para representar sus elementos.

Proyección en espacios de menor dimensión

Los grandes volúmenes de datos, generados por datos con alta dimensionalidad, son difíciles de analizar o de visualizar, por ejemplo, en el caso de las imágenes a colores, que son representadas por una matriz de pixeles, cada uno de los cuales representa 3 dimensiones, una para cada color (rojo, verde y azul), lo que significa que en alta resolución, la compresión y visualización son tareas que requieren de algoritmos eficientes; afortunadamente, con frecuencia, los datos pueden ser representados con tan solo algunas de las dimensiones. En particular, los procedimientos para comprimir imágenes utilizan técnicas para reducir su tamaño manteniendo la información más relevante o significativa. La reducción de dimensionalidad está fuertemente ligada al concepto de proyección ortogonal ya que se trata de proyectar la data en un subespacio de menor dimensión en el cual se puedan manejar los datos con mayor facilidad y extraer información relevante a menor costo.

Comencemos mostrando como se proyecta un vector $\boldsymbol{x}$ de $\mathbb{R}^n$ en un subespacio de menor dimensión, por ejemplo una recta con vector director $\boldsymbol{v} = (\boldsymbol{v_1}, \boldsymbol{v_2}, \cdots, \boldsymbol{v_n}) \in \mathbb{R}^n$. La recta es un subespacio de dimensión 1 que denotaremos con $\boldsymbol{U}$. La proyección de $\boldsymbol{x}$ sobre $\boldsymbol{U}$ es un vector de $\boldsymbol{U}$ que denotaremos $\boldsymbol{Proy_U}\,(\boldsymbol{x})$, la figura **1** ilustra el proceso.

De acuerdo con la definición, el vector proyección se puede escribir como combinación lineal de los vectores en la base del subespacio $\boldsymbol{U}$, en este caso la base está formada solo por el vector v, por lo tanto $\mathbf{Proy_U}\,(\boldsymbol{x}) = \boldsymbol{\alpha}\mathrm{v}$, para algún escalar $\boldsymbol{\alpha} \in \mathbb{R}$.

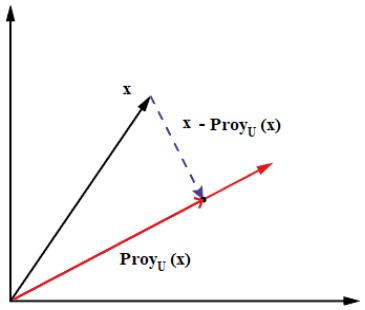

Fig. 1. Proyección en un subespacio de dimensión 1.

En este caso, la proyección es el vector de U más cercano a x, lo que significa que $\|\mathbf{x} - \mathbf{Proy}_U(\boldsymbol{x})\|$ es mínima y en consecuencia el segmento $\boldsymbol{x} - \mathbf{Proy}_U(\boldsymbol{x})$ es ortogonal a $\boldsymbol{v}$, esto se caracteriza como:

$$< v,\ \mathbf{x} - \mathbf{Proy}_U(\boldsymbol{x}) > = < v, \mathbf{x} - \alpha v > = \mathbf{0}\,,\ \text{para algún escalar } \alpha \in \mathbb{R}$$

Tomando en cuenta que el producto interno es bilineal, inferimos que

$$< v, \mathbf{x} > \ - \alpha < v, v > = \mathbf{0} \ \Leftrightarrow \ \alpha = \frac{< v, \mathbf{x} >}{< v, v >}$$

Nótese que conociendo los vectores que generan el subespacio $\boldsymbol{U}$, solo necesitamos el valor de $\boldsymbol{\alpha}$ para obtener la proyección. En el ejemplo, si el vector $\boldsymbol{v}$, que es la base del subespacio de proyección, tiene norma 1, entonces conseguir la proyección de $\boldsymbol{x}$ en $\boldsymbol{U}$ se reduce a hallar el escalar $\boldsymbol{\alpha} = < \boldsymbol{v}, \boldsymbol{x} >$, y posteriormente la proyección $\mathbf{Proy}_U(\boldsymbol{x}) = \alpha v$.

Si usamos el producto punto como producto interno tendremos que

$$\mathbf{Proy}_U(\boldsymbol{x}) = \left(\frac{v^T x}{v^T v}\right) v = \left(\frac{v^T x}{\|v\|^2}\right) v = v\frac{v^T x}{\|v\|^2} = \frac{v v^T}{\|v\|^2} v$$

En esta última expresión notamos que el vector $\mathbf{Proy}_U(\boldsymbol{x}) = \left(\frac{v^T x}{v^T v}\right) v$ es calculado como el escalar $\boldsymbol{\alpha} = \frac{v^T x}{\|v\|^2}$ multiplicado por el vector $\boldsymbol{v}$; sin embargo, esto es equivalente a calcularlo como el producto de la matriz $\mathbf{P} = \frac{v v^T}{\|v\|^2}$ por el vector $\boldsymbol{v}$. A esta matriz $\mathbf{P}$ la llamamos matriz de proyección. Nótese que en este caso, sólo necesitamos al escalar $\boldsymbol{\alpha}$ para representar la proyección, lo cual señala el camino hacia la reducción de la dimensionalidad.

Pol. Con. (Edición núm. 67) Vol. 7, No 2, febrero 2022, pp. 1884-1907, ISSN: 2550 - 682X

Proyección en espacios de gran dimensión

Supongamos ahora que queremos proyectar un vector $x = (x_1, x_2, x_3)$ de un espacio tridimensional en un subespacio U de dos dimensiones generado por los vectores v_1 y v_2. Haciendo la analogía con la proyección en un subespacio de una dimensión, el vector proyección será una combinación lineal de v_1 y v_2.

Adicionalmente, considerando que la proyección ortogonal nos dará la menor distancia, el vector $x - \text{Proy}_U(x)$ será ortogonal al plano expandido por v_1 y v_2 , lo cual significa que es ortogonal a v_1 y a v_2, ver figura **2**.

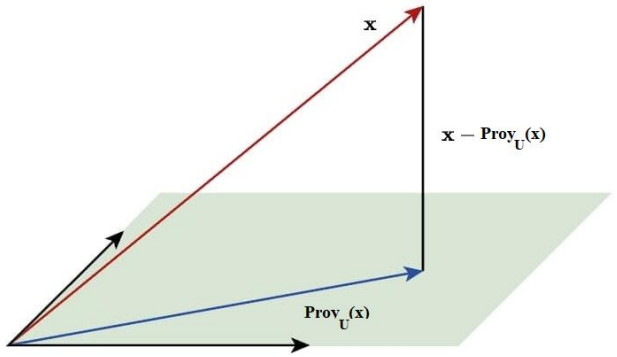

Fig. 2 Proyección ortogonal: $\text{Proy}_U(x) = \alpha_1 v_1 + \alpha_2 v_2$

En términos matriciales, consideremos B la matriz que tiene como columnas a v_1 y v_2, y sea $\alpha = (\alpha_1, \alpha_2)$ el vector de escalares, entonces, las siguientes dos condiciones deben cumplirse:

1) $\text{Proy}_U(x) = B\alpha$

2) $< x - \text{Proy}_U(x), v_1 > = 0$ y $< x - \text{Proy}_U(x), v_2 > = 0$

Sustituyendo (1) en (2) obtenemos

$< x - B\alpha, v_1 > = 0$ y $< x - B\alpha, v_2 > = 0$

$< x, v_1 > - < B\alpha, v_1 > = 0$ y $< x, v_2 > - < B\alpha, v_2 > = 0$

$x^T v_1 - \alpha^T B^T v_1 = 0$ y $x^T v_2 - \alpha^T B^T v_2 = 0$

$x^T [v_1, v_2] = \alpha^T B^T [v_1, v_2] \equiv \alpha^T B^T B = x^T B$

Una vez que los vectores v_1 y v_2 son linealmente independientes, la matriz B^TB es cuadrada de orden 2 con columnas independientes, lo que significa que su inversa $(B^TB)^{-1}$ existe, y en consecuencia los escalares de la combinación pueden obtenerse como:

$$\alpha^T = x^TB(B^TB)^{-1} \equiv \alpha = (B^TB)^{-1}B^Tx$$

Una vez que obtenemos α, la proyección será $\mathbf{Proy}_U(x) = B\alpha = B(B^TB)^{-1}B^Tx$.

A la matriz $P = B(B^TB)^{-1}B^T$ se le conoce como matriz de proyección. Si los vectores columnas de B son ortonormales, entonces B^TB es la matriz identidad y el cálculo de la proyección se reduce a $\mathbf{Proy}_U(x) = BB^Tx$, siendo $P = BB^T$ la matriz de proyección. En este caso, para representar al vector $x = (x_1, x_2, x_3)$ de $\mathbb{R}^3$ en el subespacio U basta con tener $\alpha \in \mathbb{R}^2$, cumpliendo con el propósito de reducir la dimensionalidad.

En el caso general de la proyección de un vector **n**-dimensional x, en un subespacio U de dimensión $k << n$, del cual se conoce una base $B_k = \{v_1, v_2, v_3, \cdots, v_k\}$ de vectores ortonormales, la matriz B tendrá como como columnas los vectores $v_1, v_2, v_3, \cdots, v_k$, y por lo tanto será de $n \times k$, la $\mathbf{Proy}_U(x) = BB^Tx$ será un vector n-dimensional que podrá ser representado por un vector $\alpha = B^Tx$ que es k-dimensional.

Este resultado es la base que sustenta la reducción de dimensionalidad vía análisis de componentes principales.

Análisis de Componentes Principales y reducción de dimensionalidad

El análisis de componentes principales (PCA por sus siglas en inglés) es una de las técnicas más usadas para la comprensión y manejo de datos con altas dimensiones. Se ha utilizado por muchos años en el almacenamiento, compresión y visualización, de imágenes, aprovechando la característica de que generalmente las imágenes pueden ser representadas solo con algunas de las direcciones, y que muchas direcciones están altamente correlacionadas.

La siguiente gráfica ilustra la situación en dos dimensiones, en la cual los datos no están necesariamente en la misma recta; sin embargo su variación en un espacio de una dimensión es poco significativa, por lo que bien pudieran ser modelados por la recta que mejor los represente; este último término debe estar bien establecido, por ejemplo podríamos hablar de la recta de mejor ajuste en algún sentido; un modelo sugerido muchas veces es la recta que minimiza los errores cuadráticos en cada punto, sustentado por el método de los mínimos cuadrados lineales.

Pol. Con. (Edición núm. 67) Vol. 7, No 2, febrero 2022, pp. 1884-1907, ISSN: 2550 - 682X

La figura 3 muestra el caso de variables linealmente correlacionadas, en el cual los datos bien pueden ser modelados por una recta.

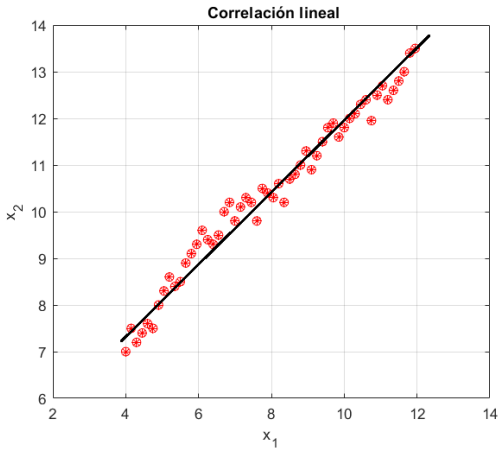

Fig. 3. Correlación lineal entre variables

Supongamos que tenemos un conjunto de datos **X** que contiene **N** vectores $\mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_N}$, cada uno de ellos representando ***d*** características, es decir cada uno es ***d***-dimensional. Hablamos entonces de una muestra de **N** objetos en un espacio **d**-dimensional que quisiéramos representar en un espacio ***k***-dimensional de menor dimensión (***k*** << ***d***) de tal forma que esta representación sea lo más similar posible a la muestra original ***X***. El objetivo del análisis de componentes principales es minimizar el error de reconstrucción promedio de las proyecciones ortogonales con respecto a la data original.

Ahora bien, sea $B_d = \{v_1, v_2, v_3, \cdots, v_d\}$ una base ortonormal de un espacio ***d***-dimensional**,** entonces cada elemento de X_i , del conjunto de datos**,** se puede escribir como una combinación lineal de los vectores de la base B_d**,** esto es, existe un vector de escalares $\alpha = (\alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{id})$ tal que:

$$\mathbf{X_1} = \alpha_{i1} v_1 + \alpha_{i2} v_2 + \cdots + \alpha_{id} v_d$$

Una vez que $v_1, v_2, \cdots v_d$ son ortonormales $v_j^T X_i = \alpha_{ik}$ **,** con $j = 1, \cdots, d$**,** lo cual significa que α_{ij} representa la proyección de X_i sobre el subespacio unidimensional generado por v_j. Si ahora tomamos dos vectores v_j y v_p de B_d, obtendremos que $[v_j, v_p]^T \mathbf{X_1} = [\alpha_{ij}, \alpha_{ip}]$ es una representación de X_i en un espacio de dos dimensiones. Es claro que con este procedimiento

Pol. Con. (Edición núm. 67) Vol. 7, No 2, febrero 2022, pp. 1884-1907, ISSN: 2550 - 682X

podríamos obtener siempre una representación de los datos en un espacio de menor dimensión, y también pudiéramos reconocer que para un cierto k << n la representación es adecuada.

Con todas estas consideraciones, podríamos seleccionar $\boldsymbol{k}$ vectores de $\boldsymbol{B_d}$ ($\boldsymbol{k} << \boldsymbol{d}$) y definir $\boldsymbol{B}$ como la matriz cuyas columnas son los $\boldsymbol{k}$ vectores ortonormales seleccionados, y la proyección del dato $\boldsymbol{X_i}$ sobre el subespacio $\boldsymbol{U}$ generado por estos $\boldsymbol{k}$ vectores se calcula como $\mathbf{Proy}_{\boldsymbol{U}}(\boldsymbol{X_i}) = \boldsymbol{BB^T X_i}$, y puede ser representada por un vector de coordenadas (reconocido como code) $\boldsymbol{\alpha_i} = (\alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{ik}) = \boldsymbol{B^T X_i}$.

Como hemos dicho anteriormente, la proyección $\mathbf{Proy}_{\mathbf{U}}(\mathbf{X_i})$ es un elemento del espacio $\boldsymbol{d}$-dimensional, y por lo tanto se puede expresar como combinación lineal de los vectores de $\boldsymbol{B_d}$, $\mathbf{Proy}_{\mathbf{U}}(\boldsymbol{X_i}) = \boldsymbol{\beta_1 v_1 + \beta_2 v_2 + \cdots + \beta_d v_d}$.

Sin pérdida de generalidad, asúmanos que los **k** vectores que generan el subespacio **U** son los vectores $\boldsymbol{v_1, v_2, \cdots, v_k}$ de $\mathbf{B_d}$, los cuales son las columnas de la matriz B; entonces

$$\mathbf{Proy}_{\mathbf{U}}(\boldsymbol{X_i}) = \beta_1 v_1 + \beta_2 v_2 + \beta_k v_k + \beta_{k+1} v_{k+1} \cdots + \beta_d v_d$$
$$= \sum_{n=1}^{k} \beta_n v_n + \sum_{n=k+1}^{d} \beta_n v_n = \hat{X}_i + + \sum_{n=k+1}^{d} \beta_n v_n.$$

La primera suma es $\hat{X}_i = B\alpha$, la aproximación a los datos en el subespacio $\boldsymbol{U}$, con $\boldsymbol{\alpha} = (\boldsymbol{\beta_1, \beta_2, \ldots, \beta_k})$, y la segunda suma es un vector en el complemento ortogonal de $\boldsymbol{U}$. Lo que se plantea en el análisis de componentes principales es descartar la segunda suma, tal como lo hemos indicado previamente, y definir al subespacio generado por los vectores $\boldsymbol{v_1, v_2, \cdots, v_k}$ como el subespacio principal, aceptando al vector $\tilde{\boldsymbol{X}}_i$ como la representación o aproximación de $\boldsymbol{X_i}$ en el subespacio $\boldsymbol{U}$ de menor dimensión.

Considerando $\boldsymbol{N}$ objetos en el conjunto de datos $\boldsymbol{X_1, X_2, \cdots, X_N}$ el objetivo entonces es hallar el vector de coordenadas $\boldsymbol{\alpha}$ y los vectores $\boldsymbol{v_i}$, con $\boldsymbol{i = 1, \cdots, k}$, tal que el error cuadrático de reconstrucción promedio se minimice. Si denotamos $\mathbf{E_k}$ este error de reconstrucción cuando deseamos representar la data a un espacio **k**-dimensional, el problema consiste en:

$$\boldsymbol{minimizar}\ \mathbf{E_k} = \frac{1}{N}\sum_{n=1}^{N}\|\mathbf{X_n} - \tilde{\mathbf{X}}_n\|^2 = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{X_n} - \tilde{\mathbf{X}}_n)^T(\mathbf{X_n} - \tilde{\mathbf{X}}_n)$$

Con el propósito de facilitar la notación asumimos que la esperanza del conjunto de datos es cero y que disponemos de una base $\mathbf{B_d}$ de vectores ortonormales en el espacio $\boldsymbol{d}$-dimensional donde se encuentran los datos, y que para todo $\boldsymbol{n}$ queremos hallar $\tilde{\boldsymbol{X}}_n = \boldsymbol{B\alpha_n}$ similar a $\mathbf{X_n}$.

Pol. Con. (Edición núm. 67) Vol. 7, No 2, febrero 2022, pp. 1884-1907, ISSN: 2550 - 682X

La medida de similaridad que se plantea es el cuadrado de la distancia euclídea $\|\mathbf{X_n} - \tilde{\mathbf{X}}_n\|^2$, siendo el objetivo minimizar el promedio de los errores cuadráticos o el error de reconstrucción (Pearson, 1901).

Se plantea entonces un problema de optimización, y para hallar el $\tilde{\mathbf{X}}_n$ óptimo, para un ***n*** dado ***n = 1, …, N***, debemos encontrar el conjunto de escalares $\boldsymbol{\alpha_n}$ y el conjunto de vectores $v_1, v_2, \cdots, v_k$ óptimos**,** por lo tanto se trata de un problema de optimización multivariable**.** Una forma simple de lograr este objetivo es hallar primero el conjunto de escalares óptimos dada una base fija de vectores $v_1, v_2, \cdots, v_k$ **,** y luego hallar los vectores óptimos, que conformarán el subespacio principal.

Una vez que hemos escogido la norma euclídea, como la herramienta para medir la similaridad, el conjunto de escalares óptimos, para un $\mathbf{X_n} = (\mathbf{X_{n1}}, \ldots, \mathbf{X_{nd}})$ dado, es aquel que anula las derivadas parciales de $\mathbf{E_k}$ con respecto a cada coordenada en $\boldsymbol{\alpha_n}$**,** las cuales son funciones de $\tilde{\mathbf{X}}_n$, por lo tanto debemos aplicar la regla de la cadena.

$$\frac{\partial \mathbf{E_k}}{\partial \alpha_{in}} = \frac{\partial \mathbf{E_k}}{\partial \tilde{\mathbf{X}}_n} \frac{\partial \tilde{\mathbf{X}}_n}{\partial \alpha_{in}}$$

Considerando que el error de reconstrucción viene dado por $\mathbf{E_k} = \frac{1}{N}\sum_{n=1}^{N}\|\mathbf{X_n} - \tilde{\mathbf{X}}_n\|^2$, tenemos que para un dato particular $\mathbf{X_n}$ el término correspondiente en la sumatoria es:

$$\|X_n - \tilde{X}_n\|^2 = \left\| \begin{pmatrix} X_{n1} \\ X_{n2} \\ \vdots \\ X_{nd} \end{pmatrix} - \begin{pmatrix} \tilde{X}_{n1} \\ \tilde{X}_{n2} \\ \vdots \\ \tilde{X}_{nd} \end{pmatrix} \right\|^2 = \left\| \begin{matrix} (X_{n_1} - \tilde{X}_{n1}) \\ (X_{n2} - \tilde{X}_{n2}) \\ \vdots \\ (X_{nd} - \tilde{X}_{nd}) \end{matrix} \right\|^2$$

$$= \left(X_{n1} - \tilde{X}_{n1}\right)^2 + \left(X_{n2} - \tilde{X}_{n2}\right)^2 + \cdots + \left(X_{nd} - \tilde{X}_{nd}\right)^2$$

Por lo tanto,

$$\frac{\partial E_k}{\partial \tilde{X}_n} = \frac{1}{N}\left(-2\left(X_{n_1} - \tilde{X}_{n_1}\right) - 2\left(X_{n_2} - \tilde{X}_{n_2}\right) - \cdots - 2\left(X_{nd} - \tilde{X}_{nd}\right)\right)$$

$$= \frac{-2}{N}\left(X_n - \tilde{X}_n\right)^T$$

$$\Rightarrow \frac{\partial E_k}{\partial \tilde{X}_n} = \frac{1}{N}\left(-2\left(X_{n_1} - \tilde{X}_{n_1}\right) - 2\left(X_{n_2} - \tilde{X}_{n_2}\right) - \cdots - 2\left(X_{nd} - \tilde{X}_{nd}\right)\right)$$

$$= \frac{-2}{N}\left(X_n - \tilde{X}_n\right)^T$$

Ahora bien, una vez que $\tilde{\mathbf{X}}_n = \alpha_{1n}v_1 + \alpha_{2n}v_2 + \cdots + \alpha_k v_k$ se deduce que:

$$\frac{\partial \tilde{X}_n}{\partial \alpha_{in}} = v_i$$

Por lo tanto, el valor óptimo para el parámetro α_{in} viene dado por:

$$\frac{\partial E_k}{\partial \alpha_{in}} = \frac{\partial E_k}{\partial \tilde{X}_n}\frac{\partial \tilde{X}_n}{\partial \alpha_{in}} = -\frac{2}{N}(X_n - \tilde{X}_n)^T v_i = -\frac{2}{N}(X_n^T v_i - \tilde{X}_n^T v_i) = 0$$

$$\Rightarrow X_n^T v_1 - \alpha_{1n} = 0 \Rightarrow \alpha_{1n} = X_n^T v_1 \qquad (1)$$

De esta manera, para cada X_n $(n = 1, \cdots, N)$ podemos obtener α_{in} $(i = 1, \cdots, k)$, reconociendo que el valor óptimo de α_{in} coincide con aquel que nos da la proyección ortogonal del elemento X_n de la data original en el subespacio 1-dimensional generado por el vector v_i. La generalización de este resultado apoya la propuesta inicial de considerar la proyección ortogonal $\text{Proy}_U(X) = \tilde{X}$ como una representación de la de la data original ***d***-dimensional ***X*** en el subespacio ***k***-dimensional ***U*** (***k*** << ***d***) producido por ***k***-vectores dados. El próximo paso será la escogencia de los vectores $v_1, v_2, \cdots, v_k$ que definen el subespacio principal.

Tal como quedó establecido en el resultado **(1)** cada coordenada del vector de escalares para un determinado X_n se calcula como $\alpha_{1n} = X_n^T v_1$ **,** y la proyección se expresa en términos de la combinación lineal $\tilde{X}_n = \alpha_{1n}v_1 + \alpha_{2n}v_2 + \cdots + \alpha_k v_k$, por lo tanto,

$$\tilde{X}_n = \sum_{i=1}^{k} \alpha_{1n} v_1 = \sum_{i=1}^{k} (X_n^T v_i) v_1 = \left(\sum_{i=1}^{k} v_1 v_1^T\right) X_n = BB^T X_n$$

$$\tilde{X}_n = \sum_{i=1}^{k} \alpha_{1n} v_1 = \sum_{i=1}^{k} (X_n^T v_i) v_1 = \sum_{i=1}^{k} (v_i^T X_n) v_1 \qquad (2)$$

La última igualdad en **(2)** se justifica por la simetría del producto interno.

De igual manera, X_n se expresa como combinación lineal de todos los vectores de la base B_d, lo cual podemos expresar en términos de dos sumatorias, esto es:

$$X_n = \sum_{i=1}^{d} \alpha_{1n} v_1 = \sum_{i=1}^{d} (v_i^T X_n) v_1 = \sum_{i=1}^{k} (v_i^T X_n) v_1 + \sum_{i=k+1}^{d} (v_i^T X_n) v_1$$

$$\Rightarrow X_n = \sum_{i=1}^{d} \alpha_{1n} v_1 = \tilde{X}_n + \sum_{i=k+1}^{d} (v_i^T X_n) v_1$$

$$\Rightarrow X_n - \tilde{X}_n = \sum_{i=k+1}^{d} (v_i^T X_n) v_1 \Rightarrow \|X_n - \tilde{X}_n\|^2 = \left\| \sum_{i=k+1}^{d} (v_i^T X_n) v_1 \right\|^2 \qquad (3)$$

Pol. Con. (Edición núm. 67) Vol. 7, No 2, febrero 2022, pp. 1884-1907, ISSN: 2550 - 682X

Por lo tanto, el error al aproximar X_n por $\tilde{X}_n$, en la medida establecida, es igual a la aproximación que resultaría si usáramos el complemento ortogonal del subespacio principal, y este hecho puede ser usado equivalentemente para hallar la solución óptima, tal como veremos a continuación.

Veamos como reformular la expresión $\mathbf{E_k} = \frac{1}{N}\sum_{n=1}^{N}\|\mathbf{X_n} - \tilde{\mathbf{X}}_n\|^2$ del error cuadrático promedio que queremos optimizar usando el resultado en **(3).**

$$E_k = \frac{1}{N}\sum_{n=1}^{N}\|X_n - \tilde{X}_n\|^2 = \frac{1}{N}\sum_{n=1}^{d}\left\|\sum_{i=k+1}^{d}(v_i^T X_n)\,v_1\right\|^2$$

En esta última expresión podemos identificar escalares $\gamma_1 = (v_i^T X_n),\ \ i = k+1 \ldots . d$ que intervienen en una combinación lineal de los vectores $v_{k+1}, v_{k+2}, \cdots, v_d$ del complemento ortogonal.

$$\sum_{i=k+1}^{d}(v_i^T X_n)\,v_1 = \gamma_{k+1}\,v_{k+1} + \gamma_{k+2}\,v_{k+2} + \ldots + \gamma_d\,v_d$$

Por lo tanto, la norma euclídea al cuadrado no es más que el producto escalar de un vector del complemento ortogonal por si mismo.

$$(\gamma_{k+1}\,v_{k+1} + \gamma_{k+2}\,v_{k+2} + \ldots + \gamma_d\,v_d)^T(\gamma_{k+1}\,v_{k+1} + \gamma_{k+2}\,v_{k+2} + \ldots + \gamma_d\,v_d)$$

En donde $v_i^T v_j = 0$ si $i \neq j$ y $v_i^T v_j = 1$ si $i \neq j,\ \forall\ i,j = k+1,\ldots,d,$ por ser estos vectores ortonormales. De esta manera

$$\left\|\sum_{i=k+1}^{d}(v_i^T X_n)\,v_1\right\|^2 = \sum_{i=k+1}^{d}(v_i^T X_n)^2$$

Luego, sustituyendo este resultado en el objetivo a optimizar, tenemos que:

$$E_k = \frac{1}{N}\sum_{n=1}^{N}\left\|\sum_{i=k+1}^{d}(v_i^T X_n)\,v_1\right\|^2 = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=k+1}^{d}(v_i^T X_n)^2$$

$$\Rightarrow E_k = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=k+1}^{d} v_1^T \mathbf{X_n} v_1^T \mathbf{X_n} = \frac{1}{N}\sum_{n=1}^{N}\sum_{i=k+1}^{d} v_1^T X_n X_n^T v_i^T$$

El último resultado, que obedece a la simetría de del producto interno, y una vez que las sumas son intercambiables, tenemos que:

$$\mathbf{E_k} = \sum_{i=k+1}^{d} v_1^T\left(\frac{1}{N}\sum_{n=1}^{N}\mathbf{X_n}\mathbf{X_n}^T\right)v_i = \sum_{i=k+1}^{d} v_1^T\,\mathbf{S}\,v_i$$

Podemos identificar en esta nueva expresión del error cuadrático de reconstrucción a la matriz $\boldsymbol{S}$ como la matriz de covarianzas de datos centrados.

Con esta reformulación de $\mathbf{E_k}$, el problema se traduce en hallar los vectores v_j, con $\boldsymbol{j}$ en $\{k+1, k+2, \cdots, d\}$ del complemento ortogonal al subespacio principal, sujeto a que estos vectores sean ortonormales. Esto lo podemos escribir como:

minimizar $\mathbf{E_k} = v_{k+1}^T S v_{k+1} + v_{k+2}^T S v_{k+2} + \cdots + v_d^T S v_d$

sujeto a: $v_i^T v_j = \delta_{ij}$ con $\mathbf{i, j} = \{\mathbf{k+1, k+2, \cdots, d}\}$.

Podemos resolver este problema usando el método de los multiplicadores de Lagrange (ver Thomas, 2005). La función Lagrangiana sería:

$$\mathbf{L} = v_{k+1}^T S v_{k+1} + \cdots + v_d^T S v_d + \lambda_{k+1}(1 - v_{k+1}^T v_{k+1}) + \cdots + \lambda_d (1 - v_d^T v_d).$$

La solución se hallaría resolviendo el sistema de Lagrange, para lo cual se igualan a cero las derivadas parciales con respecto a cada multiplicador λ_j, $j = k+1, \cdots, d$, y también las derivadas parciales con respecto a los vectores v_j, $j = k+1, \cdots, d$.

$$\frac{\partial L}{\partial \lambda_j} = 1 - v_j^T v_j = 0 \Leftrightarrow v_j^T v_j = 1$$

$$\frac{\partial L}{\partial v_j} = 2 v_j^T S - 2\lambda v_j^T = 0 \Leftrightarrow S v_j = \lambda v_j$$

Finalmente, podemos notar, que los vectores que optimizan el error cuadrático promedio son autovectores de la matriz de covarianzas. En este caso, el valor del objetivo estaría dado por:

$$\mathbf{E_k} = v_{k+1}^T \lambda_{k+1} v_{k+1} + v_{k+2}^T \lambda_{k+2} v_{k+2} + \cdots + v_d^T \lambda_d v_d$$

$$\Rightarrow \mathbf{E_k} = \lambda_{k+1} v_{k+1}^T v_{k+1} + \lambda_{k+2} v_{k+2}^T v_{k+2} + \cdots + \lambda_d v_d^T v_d$$

$$\Rightarrow \mathbf{E_k} = \lambda_{k+1} + \lambda_{k+2} + \cdots + \lambda_d$$

Así, para saber el valor mínimo de $\mathbf{E_k}$ basta con hallar los ***d-k*** autovalores de menor valor de $\boldsymbol{S}$, que por ser una matriz simétrica y semidefinida positiva tiene autovalores reales no negativos. Sin embargo, los correspondientes autovectores $v_{k+1}, v_{k+2}, \cdots, v_d$ pertenecen al complemento ortogonal del subespacio principal. Por lo tanto, y los vectores que necesitamos en la base $\boldsymbol{B}$ son los autovectores correspondientes a los $\boldsymbol{k}$ autovalores de mayor valor.

Esto puede apreciarse en la figura **4**, que presenta una data en dos dimensiones y señala los dos autovectores de la matriz de covarianzas; se puede observar que el autovector asociado al mayor

Pol. Con. (Edición núm. 67) Vol. 7, No 2, febrero 2022, pp. 1884-1907, ISSN: 2550 - 682X

autovalor (resaltado en rojo) representará la mejor base para la proyección, mientras que el autovector asociado al autovalor de módulo mínimo sugiere la magnitud del error.

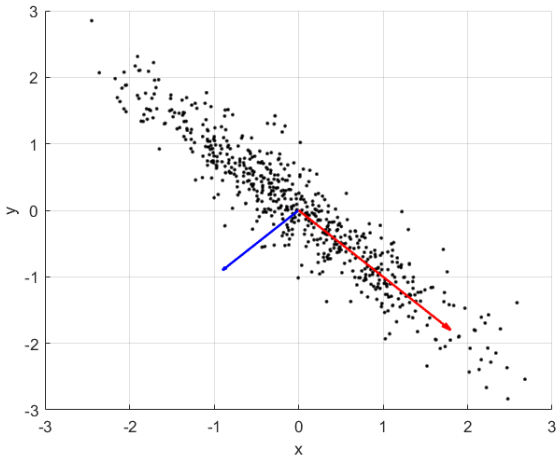

Fig. 4. Representación del subespacio principal

Algoritmo básico (reducción de dimensionalidad vía PCA)

Una vez que entendemos cómo obtener la base del espacio de proyección, podemos retomar el cálculo de los escalares (el code) de la combinación lineal que permite construir la proyección para cada dato $\boldsymbol{X_n}$, con $\boldsymbol{n = 1, \cdots, N}$. Esto nos habilita para definir un algoritmo con los pasos básicos del proceso.

En la solución del problema hemos asumido que la data es centrada, o que la media es cero; solo para disminuir la dificultad en la derivación, pero obtendríamos el mismo resultado con cualquier valor de la media; aunque por cuestiones numérico-computacionales, es preferible restar la media a los datos en este tipo de análisis. Para mayor información sobre estas prácticas y otras técnicas eficientes para implementaciones de PCA se sugiere ver Deisenroth 2020, y James, 2013)

Otra práctica conveniente es normalizar la data; y esto lo hacemos dividiendo en cada dirección por la desviación estándar, lo cual libera la dependencia de unidades y garantiza una varianza unitaria en cada dimensión sin alterar la correlación entre variables.

En resumen, los primeros pasos del algoritmo consisten en centrar y normalizar la data; eso es, por cada punto $\boldsymbol{X_n}$:

$$X_n^{(*)} \leftarrow \frac{X_n - \mu}{\sigma}$$

Luego, debemos hallar la matriz de covarianzas ***S*** de la data $\mathbf{X^*} = \{\mathbf{X_1^*}, \cdots, \mathbf{X_d^*}\}$, y calcular los ***k*** autovalores de mayor valor. Los k autovectores asociados a estos autovalores formarán las columnas de la matriz de proyección, con la cual finalmente obtenemos la data proyectada.

$$\widetilde{X}_n = \text{Proy}_U(X_n^*) = BB^T(X_n^*)$$

Implementación en Matlab

A continuación, se presenta un código básico para la reducción de dimensionalidad; en el cual se crea una data aleatoria, que será sujeta a la reducción de dos dimensiones a una dimensión. La salida es el vector de escalares alfa (en este caso un escalar, ya que el subespacio de proyección es unidimensional).

```
function [alfa]=PCA0(X)
 % Dimentionality reduction via PCA
 % This code takes a random dataset X, with N two-dimensional
 % vectors, and project it onto a one-dimensional subspace, to
 % illustrate the Principal Component Analysis steps.

N=length(X);

 % Centering and normalizing
mu1=mean(X(:,1)); s1=std2(X(:,1));
mu2=mean(X(:,2)); s2=std2(X(:,2));
X(:,1)=(X(:,1)-mu1*ones(N,1))/s1;
X(:,2)=(X(:,2)-mu2*ones(N,1))/s2;

 % Finding the covariance matrix
CovX=cov(X);

 % Identifying the principal component
[V,E]=eig(CovX);
e = diag(E);
[me,ind]=max(e);

 % Defining the projection subspace
u=V(:,ind);

 % Finding the optimal scalars(code)
alfa= X*u;

 % Calculating the projection (Decode)
Proy_u=alfa*u';

 % Visualizing results
figure;
scatter(X(:,1), X(:,2), 'k.');
hold on;
scatter(Proy_u(:,1), Proy_u(:,2), 'g.')
title('PCA: Reducci\'on de dimensionalidad');
xlabel('x');
ylabel('y');
end
```

Fig. 5. Código básico en Matlab para PCA

Pol. Con. (Edición núm. 67) Vol. 7, No 2, febrero 2022, pp. 1884-1907, ISSN: 2550 - 682X

La figura **6** muestra los resultados de la ejecución del código (Fig. 5) para un conjunto **N=500** datos aleatorios.

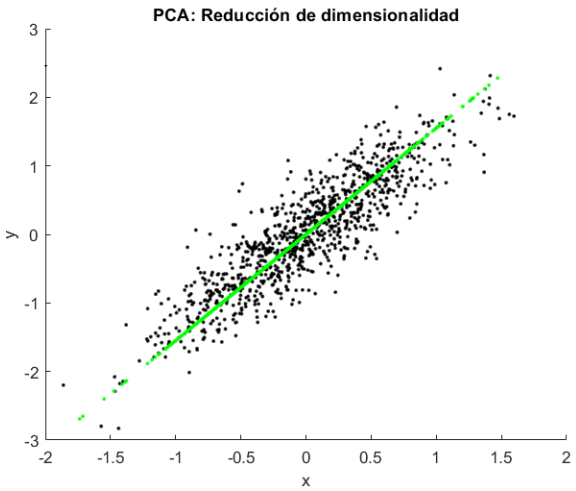

Fig. 6. PCA para reducción de datos de 2 dimensiones a una dimensión

Esta filosofía puede extenderse al caso general de ***N*** datos con ***d*** características que se desee proyectar en un espacio de ***k*** << ***d*** dimensiones; así, seleccionaríamos ***k*** autovectores asociados a los ***k*** autovalores de mayor valor de la matriz de covarianzas.

Nótese que en el paso **20** del código estamos calculando la descomposición espectral de la matriz que representa a los datos; sin embargo, lo conveniente en la práctica es calcular solo **k** autovectores asociados a los autovalores de mayor valor, y esto puede hacerse con métodos de proyección para el cálculo de autovalores; lo que sería recomendable en el caso general de aplicaciones con alta dimensionalidad. Para una explicación detallada de estas técnicas se recomienda ver (Trefethen, 1997). Una vez que la matriz ***S*** es simétrica y deficiente en rango, una propuesta con cálculo estable sería hallar la descomposición en valores singulares truncada de la matriz ***X***. El tópico ha sido tratado ampliamente en bibliografías recientes, ver por ejemplo (Deisenroth, 2020).

Conclusiones

Bajo la suposición de tener un conjunto de **N** datos $\mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_N}$ con media **0**, en un espacio ***d***-dimensional, y disponer de una base ortogonal de vectores para este espacio, se han descrito en

forma detallada los elementos de la matemática que sirven de base para uno de los algoritmos más usado en la reducción de dimensionalidad, como lo es el análisis de componentes principales o PCA. En forma incremental se describieron las bases del algoritmo que soluciona el problema.

El análisis de componentes principales ha sido por muchos años la referencia para hacer reducción de dimensionalidad y existen muchas referencias que lo definen y documentan, generalmente desde perspectivas de alto nivel de especialización en estadística, o bajo el enfoque de nuevos paradigmas como el aprendizaje automático (o Machine Learning) sin embargo, las herramientas básicas para entender el funcionamiento del algoritmo y programarlo en el lenguaje de programación de preferencia, yacen en el álgebra lineal, por lo que puede ser considerado como una aplicación del álgebra lineal.

Otro esquema de derivación del algoritmo consiste en maximizar la varianza en el subespacio de proyección para retener información de interés (Deisenroth, 2020). El enfoque es elegante pero requiere de elementos más sofisticados ligados a la estadística. En este trabajo hemos querido usar herramientas más básicas de la matemática.

Describimos normas relacionadas con diferentes productos internos, aunque usamos el producto escalar clásico en la derivación del algoritmo, por lo cual queda abierta la posibilidad de generar y comparar con otros algoritmos basados en diferentes normas. De igual manera trabajos para el cálculo efectivo y de bajo costo computacional de los autovalores de módulo máximo de la matriz de covarianza se proponen a futuro.

Referencias

1. Datta, B.N (2010). *Numerical Linear Algebra and Applications*. Society for Industrial and Applied Mathematics; 2a. Edición.
2. De la Puente, V., C. (2018). *Estadística descriptiva e inferencial*, Ediciones IDT CB, Madrid, España.
3. Deisenroth, M.P., Faisal, A., Soon, C. (2020). *Mathematics for Machine Learning*. Cambridge University Press. https://mml-book.com.
4. James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An introduction to Statistical learning with applications in R, Springer, New York.
5. Lay, D. (2012). *Álgebra lineal y sus aplicaciones*. Pearson Education, México.

6. Pearson, K. (1901). *On lines and planes of closest fit to systems of points in space*. Philosophical Magazine, 2(6), 559 – 572
7. Rencher, A.C., Bruce Schaalje G.B. (1996). (2008). *Linear models in statistics*. Hoboken, N.J: Wiley-Interscience, 6a. Edición.
8. Thomas, G.B. (2005). *Cálculo. Varias Variables*, Pearson, Addison Wesley, 11a. Edición.
9. Salazar, P., C., Del Castillo, G., S. (2017). *Fundamentos Básicos de Estadística*, http://www.dspace.uce.edu.ec/handle/25000/13720, 1ª. Edición
10. Trefethen, L. N. & Bau III, D. (1997). Numerical linear algebra. Siam, Philadelphia.