# El problema de la reducción dimensional. Análisis de Componentes Principales (PCA)

Sergio A. Pernice[1]

*Universidad del CEMA*
*Av. Córdoba 374, Buenos Aires, 1054, Argentina*

20 de julio de 2022

**Abstract**

En este trabajo de investigación se presenta la técnica de Principal Component Analysis (PCA), y su aplicación práctica en aprendizaje automático (machine learning). La intención es abordar la problemática de la reducción dimensional o compresión de datos. A partir de un análisis intuitivo, se espera acercar a los economistas y otros profesionales de las ciencias sociales estas ideas que, generalmente, resultan ajenas a sus discusiones.

*Keywords*: Principal component analysis, Análisis de componentes principales, aprendizaje no supervisado.

# 1 Introducción: intuición detrás de la reducción dimensional

La combinación de la digitalización de las sociedades, que genera todo tipo de datos fácilmente accesibles, y el crecimiento exponencial en la capacidad de procesar dichos datos, dio lugar a la revolución en el análisis de los mismos que todos conocemos con nombres como "Machine Learning", "Inteligencia Artificial", etc. Las ciencias sociales no son ajenas a dicha revolución, y una creciente proporción de economistas, y científicos sociales están adoptando muchas de las técnicas apropiadas para el eficiente analysis de estos datos.

Los datos a los que hoy se tiene acceso no solo son mucho más numerosos, además suelen ser mucho más ricos en contenido y complejidad. Esto se traduce en que los objetos matemáticos para analizar dichos datos suelen ser vectores, matrices, etc. de muy alta dimensionalidad. La estadística tradicional, la que los economistas y otros científicos sociales aprenden en los cursos formativos en econometría, fue inventada, en general, para el análisis de datos de baja dimensionalidad, por la sencilla razón de que las bases de datos accesibles eran de baja dimensionalidad.

[1] sp@ucema.edu.ar
Los puntos de vista del autor no representan necesariamente la posición de la Universidad del CEMA.

Por ejemplo, típicamente cuatro o cinco, y en general menos de 10 variables, se eligen para explicar alguna otra variable en un modelo de regresión clásica en economía.

Cuando, la dimensionalidad de los datos es mucho mayor a 10, como es crecientemente el caso en estudios empíricos en economía, finanzas, y en general en las ciencias sociales, las técnicas apropiadas para encontrar patrones y estructura en ellos son otras. Más aún, el posicionamiento del investigador frente a los datos es otra, ya que frente a semejante cantidad de variables se descuenta que la intuición, al menos en el sentido tradicional, no será de mucha ayuda.

Las técnicas útiles para análisis de datos de alta dimensionalidad constituyen lo que se conoce con el nombre genérico de "machine learning", y son sorprendentemente poderosas para encontrar patrones en datos altamente complejos. Es entonces natural que un creciente número de investigadores usen, y promuevan, estas técnicas también en las ciencias sociales. Por ejemplo, Susan Athey, ganadora de la John Bates Clark Medal en 2007, junto con Guido Imbens, publicaron en 2019 un influyente artículo titulado "Machine Learning Methods That Economists Should Know About" [Athey-Imbens (2019)].

Mas allá de esfuerzos como el mencionado, sigue siendo el caso que estas técnicas no forman parte de la formación matemática que típicamente reciben economistas y científicos sociales en general, a pesar de que muchos de los trabajos empíricos mas influyentes en la actualidad hacen uso de ellos.

En particular, metodologías basadas en Análisis de Componentes Principales (Principal Component Analysis, o PCA) han tenido un gran impacto en trabajos empiricos en economía, finanzas, y otras ciencias sociales. Por ejemplo, en forecasting macroeconómico con un número grande de predictores [Stock-Watson (2002), Mol-Giannone-Reichlin (2007)], en métodos para entender no-estacionalidad en la data [Bai-Ng (2004), Bai-Ng (2007)], para medir los efectos de la política monetaria [Bernanke-Boivin-Eliasz (2005)], para desarrollar políticas y planificación de la salud efectivas [Vyas-Kumaranayake (2006)], en analysis de posibles cambios de factores pre y post crisis [Stock-Watson (2012)], la estabilidad de los factores ante pequeños inestabilidades estructurales [Bates-PLAGBORG-MoLLER-Stock-Watson (2013)], y la inestabilidad ante grandes inestabilidades estructurales [Cheng-Liao-Schorfheide (2016)]. En general, el desarrollo de modelos para el análisis de panel data de alta dimensionalidad en economía y finanzas, basados directa o indirectamente en PCA es un area muy activa de investigación actual [Chen-Dolado-Gonzalo (2021)].

Este paper intenta aportar en esa dirección, presentando de manera accesible el método de Análisis de Componentes Principales (Principal Component Analysis), un muy poderoso método de análisis de datos que, como lo sugieren los papers citados, típicamente se usa para capturar la estructura mas importante de baja dimensionalidad embebida en datos de alta dimensionalidad, lo que en muchos casos lleva a "entender" datos complejos en términos de un número reducido de factores. Es una técnica tradicional de "Machine Learning no supervisada".

Para ayudar a la intuición detrás de este método, consideremos una nube de puntos en el plano, que pueden pensarse como provenientes de datos empíricos correspondientes a dos variables economicas. Supongamos que las coordenadas $x$ siguen una distribución normal con media cero

y varianza $\sigma_x^2 = 0.8^2 = 0.64$, y coordenadas $y$ una distribución normal con media cero y varianza $\sigma_y^2 = 0.2^2 = 0.04$, ver figura 1.

Dichos puntos "viven" en dos dimensiones, pero si hubiera que elegir una recta, o subespacio de dimension 1, que "mejor" los representa, dicha recta es claramente el eje $x$.

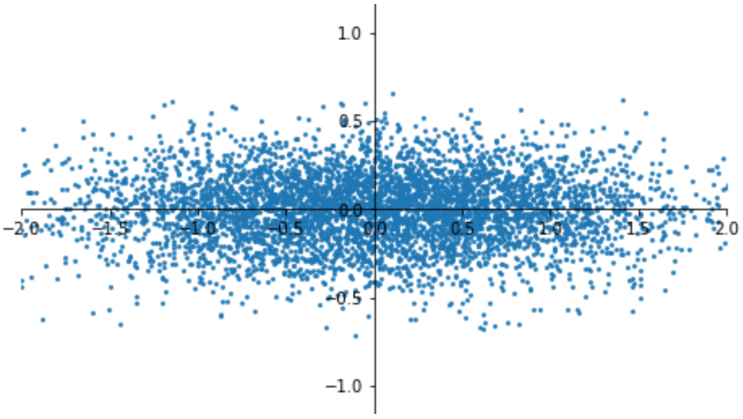



Figure 1: Las coordenadas se distribuyen $x \sim N(0, 0.64)$ y las coordenadas $y \sim N(0, 0.04)$.

¿Qué hay detrás de nuestra intuición de que el eje $x$ es el subespacio de dimension 1 que mejor representa a esos puntos? Hay al menos dos razones. Una es estadística: las coordenadas $x$ siguen una distribución normal $N(0, 0.64)$, mientras que las coordenadas $y$ siguen una distribución $N(0, 0.04)$. El eje $x$ es entonces la recta en la que los puntos muestran mayor dispersión.

La otra es geométrica: dado que los puntos tienen mucha mas dispersión en $x$ que en $y$, y que no hay correlación entre ambas coordenadas, la distancia promedio[2] entre los puntos y el eje $x$ es menos que entre los puntos y cualquier otra recta.

Consideremos ahora un ejercicio similar con los puntos de la figura 2, en los que hay claramente correlación entre la coordenada $x$ y la $y$. Intuitivamente es claro que la "mejor recta" que represente a esos puntos, no puede ser muy diferente de la que aparece en la figura 3. Y las razones son las mismas que antes. La recta roja en la figura 3 es la recta en donde los puntos tienen la mayor varianza, y además, entre todas las posibles rectas, es la que minimiza las distancias a los puntos (haremos mas precisa esta noción de "minimiza las distancias" en la proxima sección).

Se puede pensar en la mejor recta, o la recta mas representativa, de puntos que "vive" en mas dimensiones. Consideremos por ejemplo la nube de puntos en 3-D en la figura 4 En la figura 5 vemos la recta mas representativa de esos puntos, bajo los dos criterios (varianza y distancia) mencionados anteriormente. Nada nos limita a considerar la mejor recta. También podríamos considerar el "mejor plano" (o subespacio afín de dimension 2) para los puntos de la figura 4.

En general, se puede indagar acerca del mejor "$k$-plano", o subespacio afín (no necesariamente pasa por el origen) de $k$ dimensiones, de un conjunto de puntos en $d$ dimensiones ($k < d$).

[2] Apelamos en esta introducción a la noción intuitiva de "distancia promedio" pero luego la reemplazaremos por la suma de los cuadrados de las distancias.

Figure 2: Nube de puntos en el plano.

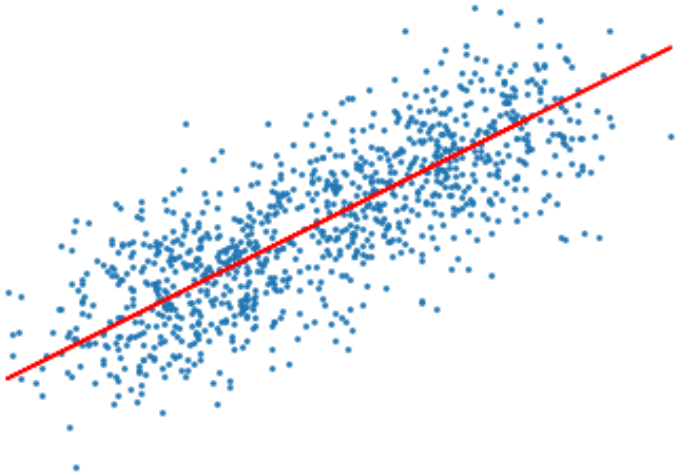


Figure 3: Nube de puntos en el plano con “mejor recta”.

Estructura del paper: en la sección 2 presentamos la técnica PCA y derivamos el correspondientes formulas, enfatizando la intuición geométrica detrás del método, alertando al lector en la sub-sección 2.3 las aplicaciones practica del método usualmente requieren un pre procesamiento de los datos. En la sección 3 mostramos la interpretación estadística de PCA, y probamos la equivalencia entre el aspecto geometrico y el estadístico de PCA. En la sección 4 nos enfocamos en la diferencia entre PCA y regresiones, enfatizando que mientras que las regresiones se pueden pensar como un método de aprendizaje “supervisado”, PCA es un método de aprendizaje “no-supervisado”. Finalmente, en la sección 5 presentamos las conclusiones.

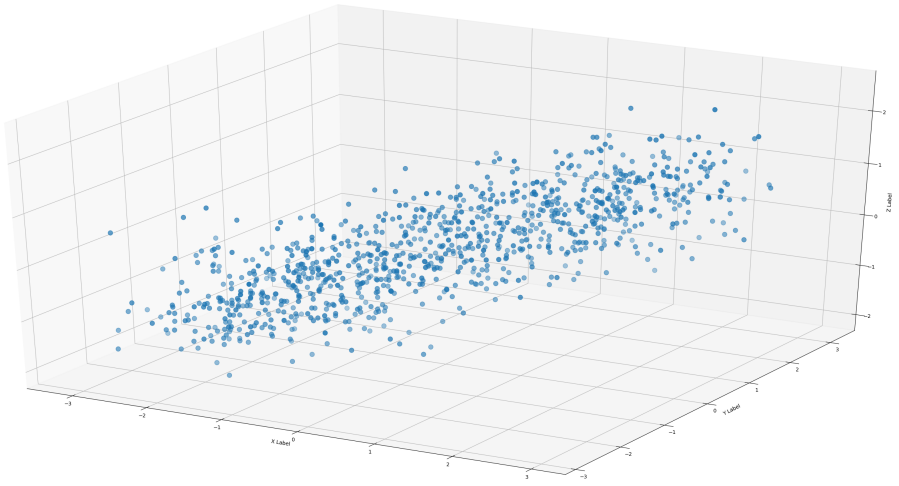


Figure 4: Nube de puntos en el espacio.

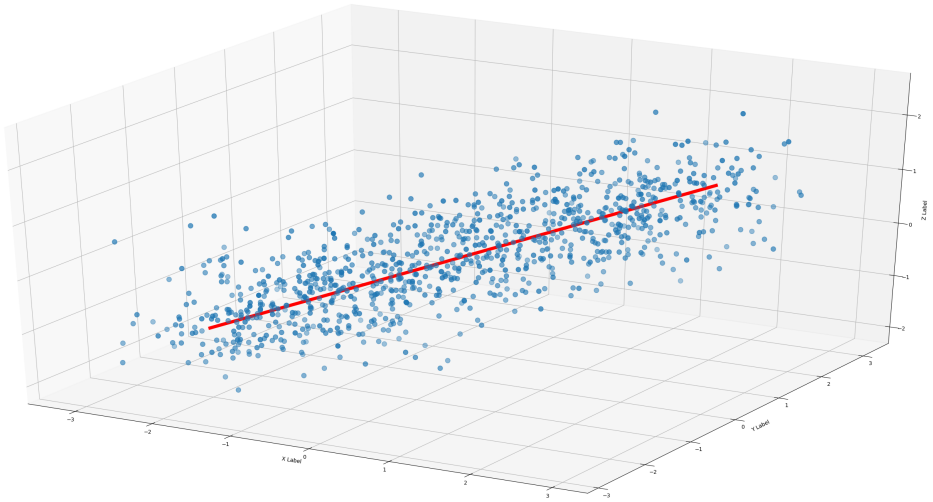


Figure 5: Nube de puntos en el espacio con “mejor recta”.

# 2 La reducción dimensional y el problema geométrico

Supongamos que un trabajo empírico, independientemente de la disciplina, genera $n$ datos, y supongamos que cada uno de esos datos corresponde al valor de $d$ variables. Vamos a pensar en esos datos como $n$ puntos en $\mathbb{R}^d$.

Podemos asociar a dichos puntos los vectores $\mathbf{x}_i \in \mathbb{R}^{d\times 1}$, $i = 0, \ldots, n-1$. El objetivo de PCA es

encontrar aproximaciones de dichos vectores de la forma

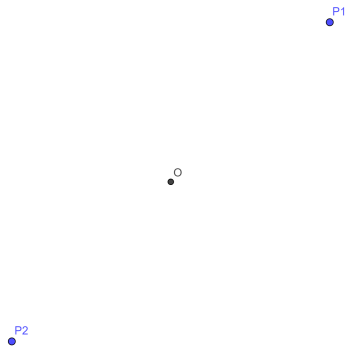
$$\mathbf{x}_i \approx \bar{\mathbf{x}} + \sum_{j=1}^{k} a_{ij}\mathbf{v}_j, \quad i = 1, \ldots, n \tag{2.1}$$

con $k$ vectores $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^n$, donde idealmente $k \ll d$.

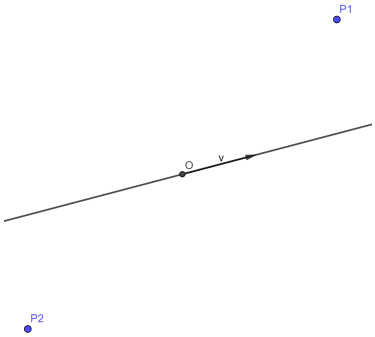
Más específicamente, queremos encontrar el "$k$- plano" (hiperplano de $k$ dimensiones que no necesariamente pase por el origen, también conocido como subespacio afín de $k$ dimensiones) en $\mathbb{R}^d$, que minimice la distancia al cuadrado con los vectores $\mathbf{x}_i \in \mathbb{R}^{d\times 1}$, $i = 0, \ldots, n-1$.

Antes de resolver el problema general, consideremos el caso en el que $n = d = 2$ para ganar intuición, ver figura 6. Como en un plano dos puntos determinan una y solo una recta, el problema es trivial, pero lo analizamos de modo que generalice a $n$ puntos en $d$ dimensiones, para todo $n$ y $d$.

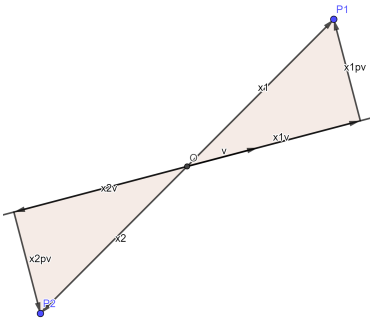

(a) Dos puntos en el plano.




(b) Nos paramos en el medio.




(c) Todo vector $\mathbf{v}$ de norma 1 define una recta.




(d) La distancia del origen a los puntos no depende de $\mathbf{v}$.

Figure 6

En la figura 6a tenemos los dos puntos $P_1$ y $P_2$, notemos que no representamos los ejes de coordenadas, enfatizando que dichos puntos pueden estar en cualquier lugar del plano. En la

figura 6b, representamos, además de los dos puntos, el punto medio entre ellos, que llamamos $O$ por origen:

$$O_x = \frac{x_1 + x_2}{2}, \quad O_y = \frac{y_1 + y_2}{2}$$

Fijado el punto medio, un vector unitario $\mathbf{v}$ con origen en dicho punto determina una única recta, como vemos en 6c. Variando el vector unitario $\mathbf{v}$ y manteniendo fijo el origen en el punto medio $O$, consideramos todas las rectas del plano que pasan por dicho punto.

Como el punto medio $O$ permanece fijo, la distancia entre dicho punto y los puntos $P_1$ y $P_2$ se mantiene constante cuando varía $\mathbf{v}$. Pero tanto las distancias entre los puntos $P_1$ y $P_2$ y la recta determinada por $\mathbf{v}$ ($x_{1pv}$ y $x_{2pv}$ en a figura 6d, los subíndices $pv$ representan "perpendicular a $\mathbf{v}$"), como las proyecciones de dichos puntos sobre la recta ($x_{1v}$ y $x_{2v}$ en la figura 6d, los subíndices $v$ representan "paralelos a $\mathbf{v}$") sí varían.

El teorema de Pitágoras implica que

$$x_1^2 = x_{1pv}^2 + x_{1v}^2, \quad x_2^2 = x_{2pv}^2 + x_{2v}^2 \tag{2.2}$$

lo cual implica que

$$x_1^2 + x_2^2 = \left(x_{1pv}^2 + x_{2pv}^2\right) + \left(x_{1v}^2 + x_{2v}^2\right) \tag{2.3}$$

El lado izquierdo de (2.3) es la suma de los cuadrados de las distancias entre el punto medio $O$ y los puntos $P_1$ y $P_2$. Como ninguno de ellos cambia cuando cambiamos $\mathbf{v}$, el lado izquierdo es constante frente a la variaciones en $\mathbf{v}$.

El primer término del lado derecho de (2.3), $x_{1pv}^2 + x_{2pv}^2$, es la suma de los cuadrados de las distancias entre los puntos $P_1$ y $P_2$ y la recta determinada por $\mathbf{v}$. $x_{1pv}^2 + x_{2pv}^2$ se minimiza (se hace cero) en la "mejor recta", que en este caso es simplemente la (única) recta determinada por $P_1$ y $P_2$, cumpliendo con uno de los criterios mencionados en la introducción.

El segundo término del lado derecho de (2.3), $x_{1v}^2 + x_{2v}^2$, es la suma de los cuadrados de las proyecciones de los puntos $P_1$ y $P_2$ sobre la recta determinada por $\mathbf{v}$. Pensados $P_1$ y $P_2$ como puntos aleatorios generados por alguna distribución, $O$ es la estimación de la media de dichos puntos, y la suma de los cuadrados de las proyecciones de los puntos es proporcional a la varianza en la dirección determinada por $\mathbf{v}$.

Como por un lado el lado izquierdo de (2.3) permanece constante frente a cambios en $\mathbf{v}$, y por el otro, como vimos, $x_{1pv}^2 + x_{2pv}^2$ se minimiza en la "mejor recta", la igualdad (2.3) implica que la suma de los cuadrados de las proyecciones (proporcional a la varianza) se *maximiza* en la "mejor recta", cumpliendo con el otro de los criterios mencionados en la introducción.

Si hay más de dos puntos, vamos a *definir* a la mejor recta como aquella que minimiza la suma de los cuadrados de la distancia entre los puntos y la recta (aunque en general, con mas de dos puntos, esta minimización ya no va dar cero). Tal como ocurrió en el ejemplo de dos puntos, esto va a coincidir con la maximización de la suma de los cuadrados de las proyecciones de los puntos sobre la recta, que a su vez va a ser proporcional a la varianza de dichos puntos en la dirección determinada por $\mathbf{v}$.

En lo que resta de esta sección, generalizamos el procedimiento anterior al problema de encontrar el mejor "$k-$plano" dado un número arbitrario $n$ de puntos-data de dimensión $d$ arbitraria.

Para resolver el problema, tal como hicimos en la figura 6b para el problema de dos puntos, nos va a convenir trasladarnos al "centro" de los $n$ puntos, ver figura 7. Esto se logra haciendo el

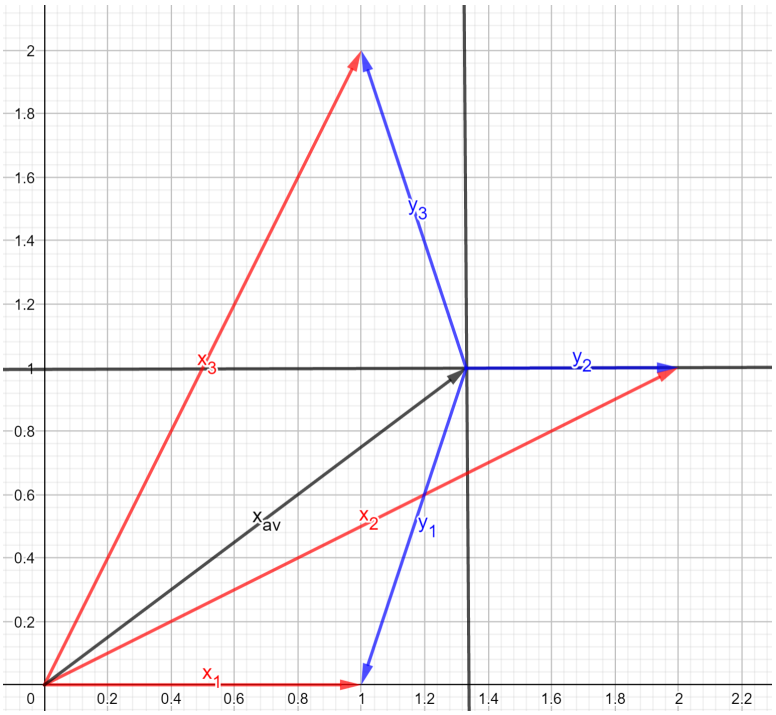



Figure 7: Nos trasladamos a un sistema de coordenadas en el centro de los datos.

cambio de variables:

$$\mathbf{y}_i = \mathbf{x}_i - \bar{\mathbf{x}}, \qquad i = 0, \ldots, n-1 \tag{2.4}$$

donde

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=0}^{n-1} \mathbf{x}_i, \quad \Rightarrow \quad \bar{\mathbf{y}} = \bar{\mathbf{x}} - \bar{\mathbf{x}} = \mathbf{0} \tag{2.5}$$

en este nuevo sistema de coordenadas no vamos a necesitar ordenada al origen, por lo que tratamos de encontrar aproximaciones de cada uno de los $n$ vectores $\mathbf{y}_0, \ldots, \mathbf{y}_{n-1} \in \mathbb{R}^d$ como combinaciones lineales de $k$ vectores $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^d$:

$$\mathbf{y}_i \approx \sum_{j=0}^{k-1} a_{ij} \mathbf{v}_j, \quad i = 0, \ldots, n-1 \tag{2.6}$$

Notar que si bien los vectores $\mathbf{y}_i$ "viven" en $\mathbb{R}^d$, en (2.6) los estamos aproximando con vectores que viven en el $k$-plano determinado por todas las posibles combinaciones lineales de los $k$ vectores $\mathbf{v}_j$, donde idealmente $k \ll d$.

## 2.1 $k = 1$, la mejor recta

Empecemos con $k = 1$. De acuerdo con la sección anterior, planteamos la siguiente función objetivo:

$$\underset{\mathbf{v}:\|\mathbf{v}\|=1}{\operatorname{argmin}} \frac{1}{n} \sum_{i=0}^{n-1} \left( (\text{distancia entre } \mathbf{y}_i \text{ y la línea generada por } \mathbf{v})^2 \right) \tag{2.7}$$

Dada una recta determinada por el vector $\mathbf{v}$ (de norma = 1), el vector $\mathbf{y}_i$ correspondiente a todo punto $i$ se puede escribir como

$$\mathbf{y}_i = d_{\mathbf{v},i}\mathbf{v} + d_{\perp\mathbf{v},i}\mathbf{v}_{\perp,i} \tag{2.8}$$

donde $\mathbf{v}_{\perp,i}$ tiene norma 1 y es ortogonal a $\mathbf{v}$, de ahí la parte $\perp$ del subíndice. La parte $i$ del subíndice es porque en más de dos dimensiones la dirección de $\mathbf{v}_{\perp,i}$, además de depender de $\mathbf{v}$, va a depender también de la posición $\mathbf{x}_i$ del punto $i$.

Como vimos en la figura 6d, el vector $\mathbf{y}_i$ es la hipotenusa del triángulo rectángulo cuyos catetos son $d_{\mathbf{v},i}\mathbf{v}$ y $d_{\perp\mathbf{v},i}\mathbf{v}_{\perp,i}$ [3]. El teorema de Pitágoras indica que para cada punto, $\|\mathbf{y}_i\|^2 = d^2_{\mathbf{v},i} + d^2_{\perp\mathbf{v},i}$. Por otro lado, para cada punto $i$, $\|\mathbf{y}_i\|^2$ está fijo. Por lo tanto

$$\sum_{i=0}^{n-1} \|\mathbf{y}_i\|^2 = \text{cte} = \sum_{i=0}^{n-1} \left( d^2_{\mathbf{v},i} + d^2_{\perp\mathbf{v},i} \right) = \left( \sum_{i=0}^{n-1} d^2_{\mathbf{v},i} \right) + \left( \sum_{i=0}^{n-1} d^2_{\perp\mathbf{v},i} \right) \tag{2.9}$$

La función objetivo (2.7) corresponde a minimizar el segundo término en la última expresión de (2.9). Pero, tal como vimos en el ejemplo de los dos puntos, la suma de los dos términos es una constante, por lo que esto es equivalente a maximizar el primer término del lado derecho de (2.9). Es decir, es equivalente a maximizar la suma del cuadrado de las proyecciones ortogonales de los vectores $\mathbf{y}_i$ sobre $\mathbf{v}$. A dicha cantidad la vamos a llamar *varianza* de los datos en la dirección determinada por $\mathbf{v}$ (más sobre esto en sección 3):

$$\operatorname{Var}\left(\{\mathbf{y}_i\}, \mathbf{v}\right) = \frac{1}{n} \sum_{i=0}^{n-1} d^2_{\mathbf{v},i} \tag{2.10}$$

y el objetivo (2.7) se puede entonces plantear como

$$\underset{\mathbf{v}:\|\mathbf{v}\|=1}{\operatorname{argmax}} \left[ \operatorname{Var}\left(\{\mathbf{y}_i\}, \mathbf{v}\right) \right] \tag{2.11}$$

En la sección 3 vamos a ver en detalle que el modo (2.11) de plantear el objetivo (2.7), tal como lo describimos en el ejemplo de los dos puntos, tiene una interpretación estadística correspondiente a encontrar la dirección de máxima varianza de los datos.

Observando (2.8) vemos que $d_{\mathbf{v},i} = \mathbf{y}_i^\top \mathbf{v}$ (producto escalar entre $\mathbf{y}_i$ y $\mathbf{v}$), por lo que con (2.10) el objetivo (2.11) es encontrar el vector unitario $\mathbf{v}$ que maximice la función:

$$\operatorname{Var}\left(\{\mathbf{y}_i\}, \mathbf{v}\right) = \frac{1}{n} \sum_{i=0}^{n-1} \left(\mathbf{y}_i^\top \mathbf{v}\right)^2 = \frac{1}{n} \sum_{i=0}^{n-1} \left(\mathbf{v}^\top \mathbf{y}_i\right)\left(\mathbf{y}_i^\top \mathbf{v}\right) = \frac{1}{n} \sum_{i=0}^{n-1} \mathbf{v}^\top \left(\mathbf{y}_i \mathbf{y}_i^\top\right) \mathbf{v} \tag{2.12}$$

[3] Notar que, dados los vectores $\mathbf{y}_i$ y $\mathbf{v}$, la combinación lineal (2.8) es única y determina un único plano independientemente de la dimensión $d$

Pero esto es una forma cuadrática asociada a la matriz simétrica $\in \mathbb{R}^{d\times d}$, semidefinida positiva:

$$A = \frac{1}{n}\sum_{i=0}^{n-1} \mathbf{y}_i \mathbf{y}_i^\top, \qquad \mathrm{Var}\left(\{\mathbf{y}_i\}, \mathbf{v}\right) = \mathbf{v}^\top A \mathbf{v} \tag{2.13}$$

El $i$-ésimo vector $\mathbf{y}_i$ tiene componentes $y_{i,j}$, $j = 0, \ldots, d-1$:

$$\mathbf{y}_i = \begin{pmatrix} y_{i,0} \\ y_{i,1} \\ \vdots \\ y_{i,d-1} \end{pmatrix} \tag{2.14}$$

Construyamos la matriz:

$$\underset{n\times d}{D} = \begin{pmatrix} -\mathbf{y}_0^\top - \\ -\mathbf{y}_1^\top - \\ \vdots \\ -\mathbf{y}_{n-1}^\top - \end{pmatrix} = \begin{pmatrix} y_{0,0} & y_{0,1} & \cdots & y_{0,d-1} \\ y_{1,0} & y_{1,1} & \cdots & y_{1,d-1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n-1,0} & y_{n-1,1} & \cdots & y_{n-1,d-1} \end{pmatrix} \tag{2.15}$$

es decir, cada fila de $D$ corresponde a un punto de la nube de puntos que queremos aproximar. $y_{i,j}$ corresponde al valor de la $j$–ésima coordenada ($j = 0, \ldots, d-1$) del $i$–ésimo punto ($i = 0, \ldots, n-1$) en el sistema de referencia en el centro de la nube.

Vemos que la matriz de (2.13) es

$$A = \frac{1}{n} D^\top D \tag{2.16}$$

Entonces, el objetivo (2.11), equivalente al objetivo original (2.7), es:

$$\underset{\mathbf{v}:\|\mathbf{v}\|=1}{\mathrm{argmax}} \left[\mathrm{Var}\left(\{\mathbf{y}_i\}, \mathbf{v}\right)\right] = \underset{\mathbf{v}:\|\mathbf{v}\|=1}{\mathrm{argmax}} \left[\mathbf{v}^\top A \mathbf{v}\right] = \underset{\mathbf{v}:\|\mathbf{v}\|=1}{\mathrm{argmax}} \left[\frac{1}{n}\mathbf{v}^\top \left(D^\top D\right)\mathbf{v}\right] \tag{2.17}$$

La matriz $D$ en (2.15) es $\mathbb{R}^{n\times j}$, por lo que la matriz $A = (1/n)D^\top D \in \mathbb{R}^{d\times d}$ es simétrica $A^\top = A$ y semidefinida positiva.

Que es simétrica se ve así: $A^\top = \left(D^\top D\right)^\top = D^\top \left(D^\top\right)^\top = D^\top D = A$.

Una matriz cuadrada $A \in \mathbb{R}^{d\times d}$ es semidefinida positiva si y sólo si, para todo vector $\mathbf{v} \in \mathbb{R}^{d\times 1}$, $\mathbf{v}^\top A \mathbf{v} \geq 0$. La matriz $A$ en (2.16) cumple con esta condición, ya que para todo vector $\mathbf{v}$, el cuadrado de la norma $\ell_2$ del vector $\mathbf{z} = D\mathbf{v} \in \mathbb{R}^{n\times 1}$ es $\|\mathbf{z}\|_2^2 = \mathbf{z}^\top \mathbf{z} = \mathbf{v}^\top D^\top D \mathbf{v} = \mathbf{v}^\top A \mathbf{v}$. Como sabemos, $\|\mathbf{z}\|_2 \geq 0$, e $\|\mathbf{z}\|_2 = 0 \Leftrightarrow \mathbf{z} = \mathbf{0}$.

Sabemos que toda matriz simétrica de $d\times d$ es "diagonalizable" en la base de autovectores. Tiene $d$ autovalores reales (contando cada uno tantas veces como sea su multiplicidad) y $d$ autovectores reales ortonormales. Además, como es semidefinida positiva, sus autovalores son positivos o cero, pero nunca negativos.

Recordemos que una matriz simétrica se puede descomponer como una suma productos "outer" de sus autovectores ortonormales, multiplicados por el correspondiente autovalor:

$$A = \sum_{i=0}^{d-1} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top \tag{2.18}$$

Dado cualquier vector $\mathbf{w}$ de modulo 1,

$$\mathbf{w}^\top A \mathbf{w} = \sum_{i=0}^{d-1} \lambda_i \mathbf{w}^\top (\mathbf{v}_i \mathbf{v}_i^\top) \mathbf{w} = \sum_{i=0}^{d-1} \lambda_i (\mathbf{w}^\top \mathbf{v}_i)(\mathbf{v}_i^\top \mathbf{w}) = \sum_{i=0}^{d-1} \lambda_i (\mathbf{w}^\top \mathbf{v}_i)^2 \tag{2.19}$$

Como $A$ es semidefinida positiva, sus autovalores son positivos o cero. Por otro lado $0 \leq (\mathbf{w}^\top \mathbf{v}_i)^2 \leq 1$, dado que ambos vectores tienen módulo unidad. Por lo tanto, (2.19) es un promedio ponderado de los autovalores de $A$. Dicho promedio obviamente se maximiza cuando $\mathbf{w} = \mathbf{v}_{\max}$, donde $\mathbf{v}_{\max}$ es el autovector de $A$ con máximo autovalor.

Llegamos así al resultado fundamental que estábamos buscando: el vector $\mathbf{v}$ que genera una recta que maximiza el objetivo (2.11) es entonces el autovector con máximo autovalor de la matriz simétrica y semi definida positiva $A = D^\top D$.✓

Ya identificamos la recta óptima, ahora queremos el valor de las coordenadas de los puntos en esta recta. Esto es simplemente la proyección ortogonal de los puntos sobre la recta. Si llamamos $\mathbf{v}_0$ al autovector identificado antes con norma 1 (Python automáticamente devuelve los autovectores normalizados a 1), entonces la coordenada del punto $i$-ésimo (2.14) en la recta óptima es:

$$c_{i,0} = \mathbf{y}_i^\top \mathbf{v}_0 \tag{2.20}$$

el nombre "$c$" viene de "coordenada", y el subíndice "$i, 0$" significa que es la proyección del punto $i$ sobre el autovector vector 0 (correspondiente al máximo autovalor).

Para tener un vector de $\mathbb{R}^n$ que contenga todas las coordenadas de los $n$ puntos $\mathbf{y}_i$, $i = 0, \ldots, n-1$, observando la matriz $D$ en (2.15) vemos que tal vector es:

$$\underset{n\times 1}{\mathbf{c}_0} = \underset{n\times d}{D}\ \underset{d\times 1}{\mathbf{v}_0} \tag{2.21}$$

este vector es la mejor "reducción dimensional", o "compresión" de nuestros $n$ puntos desde $\mathbb{R}^d$ a $\mathbb{R}$ (hay un solo número por punto, en vez de $d$ como había originalmente).

Si queremos la posición de esta reducción dimensional de los puntos en el espacio original $\mathbb{R}^d$, volvemos a nuestro sistema de coordenadas original así:

$$\mathbf{x}_{i,\text{red1D}} = \bar{\mathbf{x}} + c_{i,0}\mathbf{v}_0 = \bar{\mathbf{x}} + (\mathbf{y}_i^\top \mathbf{v}_0)\,\mathbf{v}_0 \in \mathbb{R}^{d\times 1}, \quad i = 0, \ldots, n-1 \tag{2.22}$$

esta es la mejor aproximación 1-D de nuestros puntos en $\mathbb{R}^d$:

$$\mathbf{x}_{i,\text{red1D}} \approx \mathbf{x}_i \tag{2.23}$$

el símbolo $\approx$ no debe interpretarse como que el lado izquierdo es necesariamente una buena aproximación del lado derecho. En algunos problemas lo será y en otros no, dependiendo de la distribución de puntos $\mathbf{x}_i$. Pero es la mejor aproximación 1-D de nuestros puntos bajo el criterio de minimización de la distancia Euclidiana.

Si trasponemos la ecuación (2.22), $\mathbf{x}_{i,\text{red1D}}^\top = \bar{\mathbf{x}}^\top + c_{i,0}\mathbf{v}_0^\top \in \mathbb{R}^{1\times d}$. Teniendo en cuenta que el vector (2.21) contiene todos los $c_{i,0}$, y que la componente $ij$ del producto outer entre dos vectores $\mathbf{u}\mathbf{w}^\top$ es $u_i w_j$, vemos que

$$X_{\text{red1D}} = \mathbf{one}_{n\times 1}\bar{\mathbf{x}}^\top + \mathbf{c}_0\mathbf{v}_0^\top = \mathbf{one}_{n\times 1}\bar{\mathbf{x}}^\top + (D\mathbf{v}_0)\,\mathbf{v}_0^\top = \mathbf{one}_{n\times 1}\bar{\mathbf{x}}^\top + D\,(\mathbf{v}_0\mathbf{v}_0^\top) \tag{2.24}$$

donde $\mathbf{one}_{n\times 1}$ es un vector de $n \times 1$ con todos los elementos iguales a 1[4]. La fila $i$-ésima de la matrix $X_{\text{red1D}}$ (de $n \times d$ ) son las coordenadas en $\mathbb{R}^d$ de la reducción 1-D del $i$-ésimo punto.

Notemos que en el segundo término de la última expresión de (2.24) aparece la matriz $\mathbf{v}_0\mathbf{v}_0^\top$, que es el proyector sobre el subespacio generado por $\mathbf{v}_0$. El hecho de multiplicar a dicho proyector por $D$ por izquierda, significa que estamos proyectando las filas de $D$ sobre dicho subespacio. Pero las filas de $D$ son los vectores de nuestros puntos en el sistema de referencia en el centro de los puntos, ver (2.15). Es decir que las filas del segundo término de la última expresión de (2.24) son las proyecciones de los vectores de nuestros puntos en el subespacio spaneado por $\mathbf{v}_0$.

El primer término de la última expresión de (2.24) es una matriz cuyas filas son toda iguales y corresponden a la posición del centro de la nube de puntos en el sistema de referencia original. Entonces las $n$ filas de $X_{\text{red1D}}$ son los vectores de $\mathbb{R}^d$ correspondientes a la proyección de los $n$ puntos. Con entrenamiento, una ecuación como (2.24) se puede escribir directamente sin ninguna derivación ya que (con entrenamiento) resulta "obvia".

## 2.2 El mejor $k$-plano

El análisis anterior trivialmente se extiende al mejor $k$-plano, y la solución es que dicho $k$-plano, en el sistema en el centro de la nube de puntos, es el subespacio generado por combinaciones lineales de los $k$ autovectores con los $k$ mayores autovalores.

La ecuación (2.20) es ahora

$$c_{i,j} = \mathbf{y}_i^\top \mathbf{v}_j, \quad j = 0, \ldots, k-1 \tag{2.25}$$

donde $c_{i,j}$ es la coordenada del $i$-ésimo punto en el subespacio spaneado por el $j$-ésimo autovector (el autovector cuyo autovalor es el $j$-ésimo en tamaño). (2.21) es

$$\underset{n\times 1}{\mathbf{c}_j} = \underset{n\times d}{D}\ \underset{d\times 1}{\mathbf{v}_j}, \quad j = 0, \ldots, k-1 \tag{2.26}$$

La matriz (2.24) generaliza a

$$X_{\text{redkD}} = \mathbf{one}_{n\times 1}\bar{\mathbf{x}}^\top + D\left(\sum_{j=0}^{k-1} \mathbf{v}_j\mathbf{v}_j^\top\right) \tag{2.27}$$

[4]En numpy este vector es np.ones(n,1), pero en realidad no es necesario hacer la operación $\mathbf{one}_{n\times 1}\bar{\mathbf{x}}^\top$ explícitamente en Python porque $\bar{\mathbf{x}}^\top$, de "shape" $(1, d)$, es "broadcasted" automáticamente a $\mathbf{one}_{n\times 1}\bar{\mathbf{x}}^\top$, de shape $(n, d)$. en (2.24).

El proyector sobre el "mejor" $k$-plano es la suma de los proyectores sobre los 1-planos (o rectas) de cada autovector porque estos son ortogonales entre sí.

En Python, dada la matriz $A = (1/n)D^\top D$ en (2.16), la función **eigh** de la library linalg de numpy nos da la matriz:

$$U_k = \begin{pmatrix} | & | & & | & | \\ \mathbf{v}_{k-1} & \mathbf{v}_{k-2} & \cdots & \mathbf{v}_1 & \mathbf{v}_0 \\ | & | & & | & | \end{pmatrix} \tag{2.28}$$

donde la última columna es el "mayor" autovector, la anteúltima el siguiente, etc. Con $U_k$, por la manera "conjunto de productos matriz-vector" de ver el producto de matrices, los vectores proyección de los $n$ puntos sobre los $k$ autovectores (2.26) se pueden ordenar como las columnas de la matriz

$$\underset{n\times k}{C} = \underset{n\times d}{D}\ \underset{d\times k}{U_k} \tag{2.29}$$

que se pueden graficar para $k = 1, 2$ o $3$ y son la mejor representación en $k$ dimensiones de nuestros puntos de $\mathbb{R}^d$.

Por la manera "producto outer" de ver el producto de matrices, (2.27) se puede escribir de manera compacta así:

$$X_{\text{redkD}} = \mathbf{one}_{n\times 1}\bar{\mathbf{x}}^\top + DU_kU_k^\top \tag{2.30}$$

que corresponden al mejor $k$-plano en $\mathbb{R}^d$.

En la figura 8 vemos en acción a las ecuaciones (2.29) y (2.30) para una proyección 2-D de puntos en 3-D.

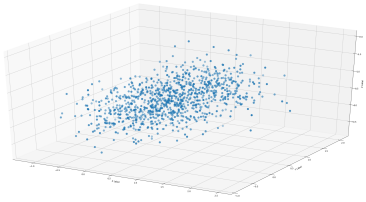
## 2.3 Los datos y la cuestión del pre-procesamiento

Pensemos en una tabla con una gran cantidad de datos sobre todos los países: PBI, PBI per cápita, nivel de desarrollo humano, índice de Gini, etc. Para esos datos, $n$ = número de países, y $d$ = número de tipos de datos que hay en la tabla. Pero los diferentes datos se miden en unidades completamente diferentes, cómo los hacemos comparables? En un primer intento, los datos de cada dimensión se dividen por la desviación estándar de dichos datos de esa dimension.

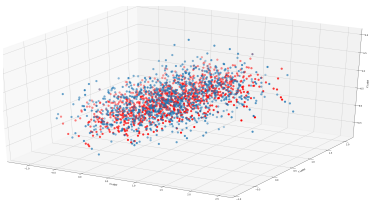
O supongamos que en una dimensión tenemos datos medidos en centímetros y en otra tenemos datos medidos en kilómetros, dividiendo cada dato por la desviación estándar se vuelven comparables.

Pero hay que tener cuidado, ya que aplicando estos procedimientos a ciegas, corremos el riesgo de eliminar varianzas genuinas de los datos que con este pre-procesamiento parcialmente se pierden. Por eso el pre-procesamiento de los mismos, al menos en el contexto del uso de PCA para ciencia de datos, requiere de cierto conocimiento previo de los mismos y es parte ciencia parte arte[5].

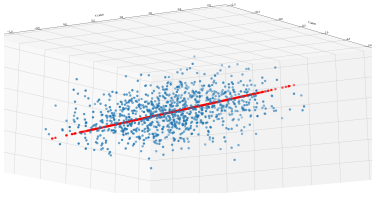
[5]En un contexto puro de machine learning no supervisado, hay en principio mecanismos para que la máquina encuentre el "mejor" pre-procesamiento de los datos.



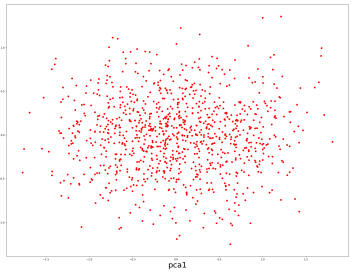
(a) Nube original de puntos 3-D.



(b) Los puntos rojos son la proyección de los puntos de 8a al mejor plano (2-D), generados con la ecuación (2.30).



(c) Desde esta perspectiva se percibe que los puntos rojos de 8b yacen en un plano.




(d) La proyección 2-D de las figuras 8b y 8c, sin el espacio 3-D en el que está inmerso, generado con la ecuación (2.29).

Figure 8: Reducción dimensional de 3-D a 2-D.

# 3 La estadística y su relación con el análisis de PCA

En la sección 2 mencionamos en repetidas oportunidades que la función objetivo de PCA se puede pensar como la maximización de la suma de los cuadrados de las proyecciones sobre la recta óptima, y que, a su vez, esto tiene la interpretación estadística de encontrar la dirección en la que los datos, pensados como variables aleatorias de dimension $d$, tienen máxima varianza (2.11). En esta sección profundizamos y generalizamos dicha interpretación.

Asumamos que $x_0, x_1 \ldots, x_{d-1}$ son $d$ variables aleatorias que ordenamos en un vector $\mathbf{x} \in \mathbb{R}^{d\times 1}$. Hay subyacente una distribución de probabilidades multivariada $p(\mathbf{x})$ que suponemos que no conocemos. Empíricamente obtenemos $n$ samples de esta variable aleatoria vectorial. Ordenamos nuestra data así:

$$D_{\underset{n\times j}{\text{or}}} = \begin{pmatrix} -\mathbf{x}_0^\top - \\ -\mathbf{x}_1^\top - \\ \vdots \\ -\mathbf{x}_{n-1}^\top - \end{pmatrix} = \begin{pmatrix} x_{0,0} & x_{0,1} & \cdots & x_{0,d-1} \\ x_{1,0} & x_{1,1} & \cdots & x_{1,d-1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n-1,0} & x_{n-1,1} & \cdots & x_{n-1,d-1} \end{pmatrix} \tag{3.1}$$

el subíndice "or" refiere a que son los datos empíricos originales. Cada fila de $D_{\text{or}}$ corresponde a un sample de nuestra variable aleatoria. $x_{i,j}$ corresponde al valor de la $j-$ésima variable aleatoria ($j = 0, \ldots, d-1$) del $i-$ésimo sample ($i = 0, \ldots, n-1$).

La estimación empírica de la media de la variable aleatoria $j$ es:

$$\bar{x}_j = \frac{1}{n}\sum_{i=0}^{n-1} x_{i,j} \tag{3.2}$$

podemos vectorizar esta expresión, capturando en una misma ecuación vectorial la media de todas nuestras variables aleatorias:

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=0}^{n-1} \mathbf{x}_i \tag{3.3}$$

La estimación empírica de la varianza de la variable aleatoria $j$ es:

$$\text{Var}\left(x_j\right) = \frac{1}{n-1}\sum_{i=0}^{n-1}\left(x_{i,j} - \bar{x}_j\right)^2 \tag{3.4}$$

y la estimación empírica de la covarianza entre la variable aleatoria $j$ y la $k$ es:

$$\text{Cov}\left(x_j, x_k\right) = \frac{1}{n-1}\sum_{i=0}^{n-1}(x_{i,j} - \bar{x}_j)(x_{i,k} - \bar{x}_k) \tag{3.5}$$

El $n-1$ en el denominador de (3.4) y (3.5) hace a esos estimadores "unbiased". De todos modos típicamente se trabaja con una cantidad de datos tal que $n$ es lo suficientemente grande como para que $1/(n-1) \approx 1/n$, por lo que de aquí en más pasamos a reemplazar ese denominador por $1/n$.

Mirando las ecuaciones (3.2-3.5), es natural "trasladarnos", en el espacio de nuestras variables, de modo que el origen coincida con $\bar{\mathbf{x}}$ y la media en las nuevas variables sea cero. Es decir, haciendo el cambio de variables

$$\mathbf{y} = \mathbf{x} - \bar{\mathbf{x}}, \quad \Rightarrow \quad \bar{\mathbf{y}} = \bar{\mathbf{x}} - \bar{\mathbf{x}} = \mathbf{0} \tag{3.6}$$

tal como hicimos en la figura 3.

En estas nuevas variables, ordenamos nuestros samples como en la matriz $D_{\text{or}}$ de (3.1), pero sin el subíndice "or":

$$\underset{n\times j}{D} = \begin{pmatrix} -\mathbf{y}_0^\top - \\ -\mathbf{y}_1^\top - \\ \vdots \\ -\mathbf{y}_{n-1}^\top - \end{pmatrix} = \begin{pmatrix} y_{0,0} & y_{0,1} & \cdots & y_{0,d-1} \\ y_{1,0} & y_{1,1} & \cdots & y_{1,d-1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n-1,0} & y_{n-1,1} & \cdots & y_{n-1,d-1} \end{pmatrix} \tag{3.7}$$

Notemos que esta matriz es idéntica a (2.15) si identificamos a los samples de nuestras variables aleatorias con las coordenadas de los puntos que estudiamos en la sección 2.

En las nuevas variables, las estimaciones empíricas de las varianzas y covarianzas toman la forma

$$\text{Var}\left(y_j\right) = \frac{1}{n}\sum_{i=0}^{n-1} y_{i,j}^2 \tag{3.8}$$

$$\text{Cov}\left(y_j, y_k\right) = \frac{1}{n}\sum_{i=0}^{n-1} y_{i,j}y_{i,k} \tag{3.9}$$

donde recordamos que simplificamos el denominador de $n-1$ a $n$ como fue explicado antes.

Teniendo en cuenta que el elemento $ik$ de la matriz $D$ en (3.7) es $y_{i,k}$ y que el elemento $ji$ de la matriz transpuesta $D^\top$ es $y_{i,j}$, reconocemos en (3.8-3.9) los elementos del producto de matrices $D^\top$ por $D$, por lo que llegamos a la siguiente expresión vectorizada:

$$\underset{d\times d}{\Sigma} = \frac{1}{n} \underset{d\times n}{D}^{\top} \underset{n\times d}{D} \tag{3.10}$$

Los elementos diagonales de $\Sigma$ son las varianzas (3.8) de las variables $y_i$, y los elementos no diagonales son las respectivas covarianzas (3.9).

Notamos que la matriz varianza-covarianza (3.10) es exactamente igual a la matriz (2.16), cuyos máximos autovectores era la solución de nuestro problema.

Con la identificación mencionada antes entre los samples de nuestras variables aleatorias con las coordenadas de los puntos que estudiamos en la sección 2 vemos que la solución de PCA coincide con encontrar la dirección en el espacio de nuestras variables aleatorias que maximizan la varianza en el sentido estadístico.

# 4 PCA vs. regresiones

En economía, y en general en ciencias sociales, el uso de regresiones es muy generalizado en trabajos empíricos. PCA es mucho menos utilizado, aunque con creciente existencia de bases de datos enormes de alta dimensionalidad su uso esta creciendo mucho en los últimos años. La creciente necesidad de usar PCA, y el entrenamiento previo de los investigadores en regresiones suele confundir a los mismos respecto de las diferencias y semejanzas entre estas dos técnicas. En esta sección tratamos de clarificar este punto.

En la figura 9 comparamos, para $k = 1$, el problema al que nos enfrentamos cuando hacemos regresiones (en azul) con el problema planteado aquí (en rojo).

En regresiones hay una variable que se diferencia de las otras, sirve para problemas en los que tenemos razones para suponer que las variables dependientes (la dimension horizontal en la figura 9) "explican" a la variable independiente (la dimension vertical en la figura 9). Por lo tanto queremos minimizar el cuadrado del error de la predicción de nuestro modelo (recta azul). En la figura 9, minimizamos la suma de los cuadrados de las longitudes de los segmentos azules punteados. Notar que dichos segmentos son verticales, y su longitud mide el error de predicción para valores dados de la variable independiente, que se presume conocida y sin errores de medición.

Por el contrario, en PCA no hacemos diferenciación entre variables explicativas y variables a ser explicadas. Todas son equivalentes a priori, y dejamos que los datos nos indiquen si existen o no unas pocas variables cuyo efecto es suficiente para aproximar con precisión los datos empíricos. La respuesta por sí o por no es objetiva, y depende, como vimos, de la magnitud relativa de los autovalores de la matriz varianza-covarianza[6].

[6]Sin embargo, en la práctica, el uso productivo de PCA requiere de un pre-procesamiento de los datos, que a su

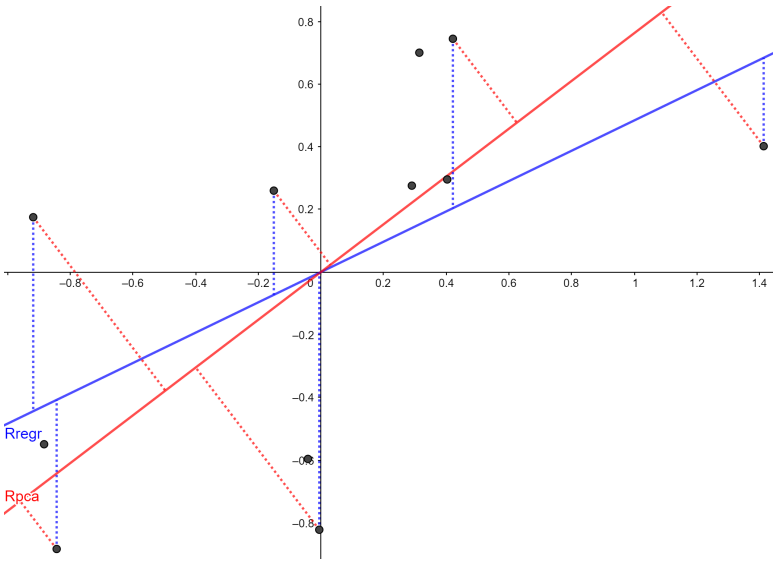



Figure 9: Mismos datos, diferentes rectas óptimas con PCA y regresiones. Las líneas punteadas rojas muestran lo que minimiza PCA y las azules lo que minimizan las regresiones.

# 5 Conclusiones

PCA es una de las metodologías mas poderosas de reducción dimensional (lineal). Esto es fundamental en la época del Big Data, donde en todo tipo de ciencias hay cada vez más datos de alta dimensionalidad. Economía, finanzas y otras ciencias sociales, no son ajenas a esta tendencia y, como vimos en la introducción, algunos de los trabajos empíricos más interesantes de los últimos quince años utilizan métodos basados en PCA y en general métodos asociados a "machine learning". La tecnología y bases de datos de cada vez más alta dimensionalidad aseguran que esta tendencia sólo puede crecer. Sin embargo, dichos métodos siguen ausentes en la curricula clásica de formación cuantitativa de los científicos sociales.

En este artículo intentamos aportar a reducción de esa carencia, en el convencimiento de que dicha formación es fundamental para realizar trabajos empíricos a la altura de los datos actualmente existentes. Presentamos el método PCA de manera simultáneamente intuitiva y rigurosa, mostrando sus dos "caras", la geométrica y la estadística, y la equivalencia entre ellas. Alertamos de la usual necesidad del pre-procesamiento de los datos en la práctica. Y enfatizamos la diferencia entre PCA y regresiones: mientras que las regresiones asumen de parte del investigador un pre-concepto acerca de qué variables son explicativas y cuáles son explicadas, PCA no asume nada, el método, cuando funciona bien, descubre estructura intrínseca en los datos (módulo el pre-procesamiento de los mismos). En machine learning esa diferencia ubica a las regresiones entre los métodos de aprendizaje "supervisado" y a PCA entre los métodos de aprendizaje "no-supervisado"

Para terminar, es importante advertir al lector que hay casos relevantes en la practica en los

vez típicamente presupone cierto conocimientos parcial de los mismos, ver sección 2.3.

que el método PCA "puro" no genera los resultados esperados, y se requieren modificaciones al mismo. El método PCA asume que la variación en los datos corresponde a información interesante. Sin embargo grandes variaciones en los datos pueden también representar grandes errores de medición. Esto es relativamente frecuente en machine learning, y también en aplicaciones en economía y ciencias sociales en casos en los que efectos dependientes de factores estructuralmente débiles (que PCA debería descartar) aparecen inflados en los datos por un gran ruido en la medición de los mismos. Afortunadamente modificaciones a PCA relativamente sencillas suelen ser útiles para eliminar dichas distorsiones, ver por ejemplo [Bai-Ng (2017)].

También puede ocurrir que un factor estructuralmente importante tenga naturalmente pequeña varianza. Por ejemplo, este suele ser el caso en la estimación de factores importantes para la valuación de activos en mercados financieros (si dichos factores tuvieran gran varianza el mercado ya los hubiera detectado y arbitrado). En ese caso también, modificaciones a PCA pueden dar los resultados deseados, ver por ejemplo [Lettau-Pelger (2020)].

# References


[Athey-Imbens (2019)] Susan Athey and Guido W. Imbens, "Machine Learning Methods That Economists Should Know About", Annual Review of Economics 2019 11:1, 685-725. https://www.annualreviews.org/doi/abs/10.1146/annurev-economics-080217-053433.

[Bai-Ng (2004)] Bai, J. and Ng, S. "A PANIC Attack on Unit Roots and Cointegration". Econometrica, 72: 1127-1177. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0262.2004.00528.x.

[Bai-Ng (2007)] Bai, Jushan, and Serena Ng. "Determining the Number of Primitive Shocks in Factor Models." Journal of Business & Economic Statistics 25, no. 1 (2007): 52–60. http://www.jstor.org/stable/27638906.

[Bai-Ng (2017)] Bai, J. and Ng, S.. "Principal components and regularized estimation of factor models", 2017. https://arxiv.org/abs/1708.08137.

[Bates-PLAGBORG-MoLLER-Stock-Watson (2013)] Bates, B. J., M. PLAGBORG-MoLLER-Stock-Watson, J. H. Stock, AND M. W. Watson. "Consistent Factor Estimation in Dynamic Factor Models with Structural Instability," Journal of Econometrics, V. 177 (2), p. 289-304, 2013. https://www.sciencedirect.com/science/article/abs/pii/S0304407613000912?via%3Dihub.

[Bernanke-Boivin-Eliasz (2005)] Ben S. Bernanke, Jean Boivin, Piotr Eliasz, "Measuring the Effects of Monetary Policy: A Factor-Augmented Vector Autoregressive (FAVAR) Approach", The Quarterly Journal of Economics, Volume 120, Issue 1, February 2005, Pages 387–422. https://doi.org/10.1162/0033553053327452.

[Cheng-Liao-Schorfheide (2016)] X. Cheng, Z. Liao, F. Schorfheide, "Shrinkage Estimation of High-Dimensional Factor Models with Structural Instabilities", The Review of Economic Studies, Volume 83, Issue 4, October 2016, Pages 1511–1543, https://doi.org/10.1093/restud/rdw005.

[Chen-Dolado-Gonzalo (2021)] Chen, L., Dolado, J.J. and Gonzalo, J. (2021), "Quantile Factor Models". Econometrica, 89: 875-910. https://doi.org/10.3982/ECTA15746.

[Lettau-Pelger (2020)] Martin Lettau, Markus Pelger. "Estimating latent asset-pricing factors", *Journal of Econometrics*, ISSN: 0304-4076, Vol: 218, Issue: 1, Page: 1-31, 2020. https://doi.org/10.1016/j.jeconom.2019.08.012.

[Mol-Giannone-Reichlin (2007)] Mol, C., Giannone, D., Reichlin, L. "Forecasting Using a Large Number of Predictors: Is Bayesian Regression a Valid Alternative to Principal Components?". Journal of Econometrics, Volume 146, Issue 2, Pages 318-328. https://www.sciencedirect.com/science/article/abs/pii/S0304407608001103?via%3Dihub.

[Stock-Watson (2002)] Stock, J. H., and M. W. Watson. "Forecasting Using Principal Components from a Large Number of Predictors." Journal of the American Statistical Association 97, no. 460 (2002): 1167–79. http://www.jstor.org/stable/3085839.

[Stock-Watson (2012)] Stock, J. H., and M. W. Watson, "Disentangling the Channels of the 2007-09 Recession," Brookings Papers on Economic Activity, pp. 81-156. https://www.brookings.edu/bpea-articles/disentangling-the-channels-of-the-2007-2009-recession/.

[Vyas-Kumaranayake (2006)] S. Vyas, L. Kumaranayake, "Constructing socio-economic status indices: how to use principal components analysis", Health Policy and Planning, Volume 21, Issue 6, November 2006, Pages 459–468. https://doi.org/10.1093/heapol/czl029.