REDUCCIÓN DE DIMENSIONALIDAD

Dr. Jorge Hermosillo
Laboratorio de Semántica Computacional

Centro de Investigación en Ciencias

Universidad Autónoma del Estado de Morelos

VALORES PROPIOS Y VECTORES PROPIOS

VALORES PROPIOS DE UNA MATRIZ

1. Definición: Un escalar λ es llamado un valor propio (*eigenvalue*) de la matriz A $n \times n$ si existe una solución no trivial de

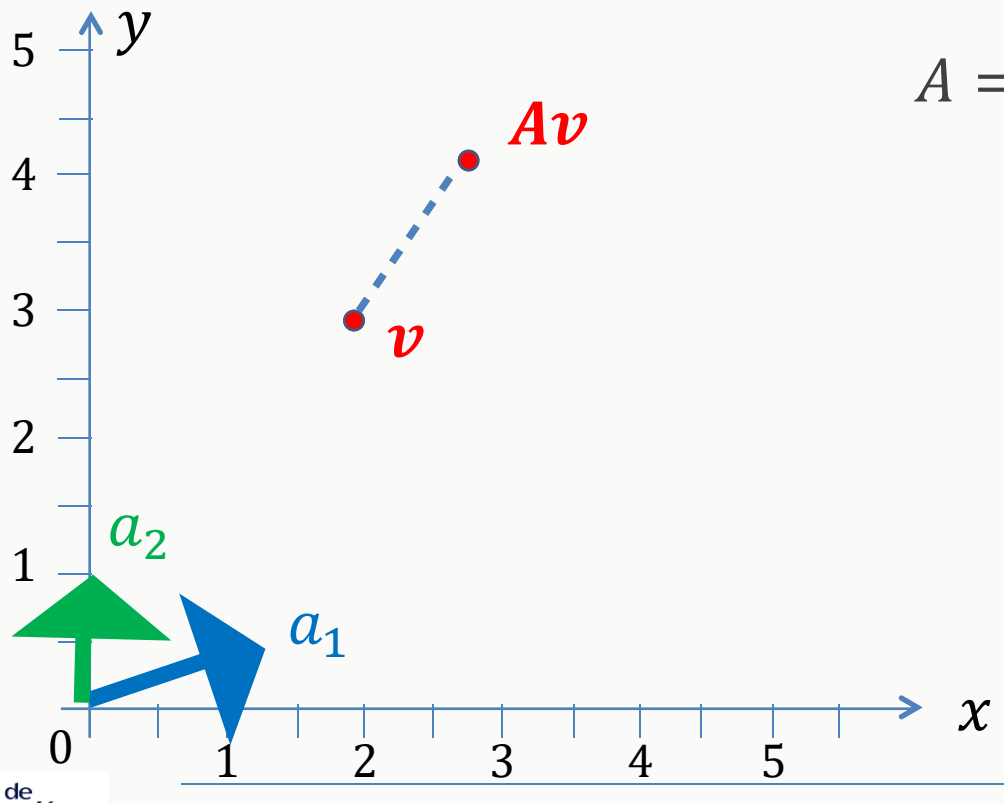
$$A\mathbf{x} = \lambda\mathbf{x}$$

tal que $\mathbf{x}$ es llamado un ***vector propio*** (*eigenvector*) correspondiente al ***valor propio*** λ.

¿A qué corresponde esto geométricamente?

INTERPRETACIÓN GEOMÉTRICA DE VALORES Y VECTORES PROPIOS

- Sea v un vector y A una matriz con columnas a_1 y a_2 (mostradas como flechas). Si multiplicamos v por A, entonces A envía v a un nuevo vector Av.



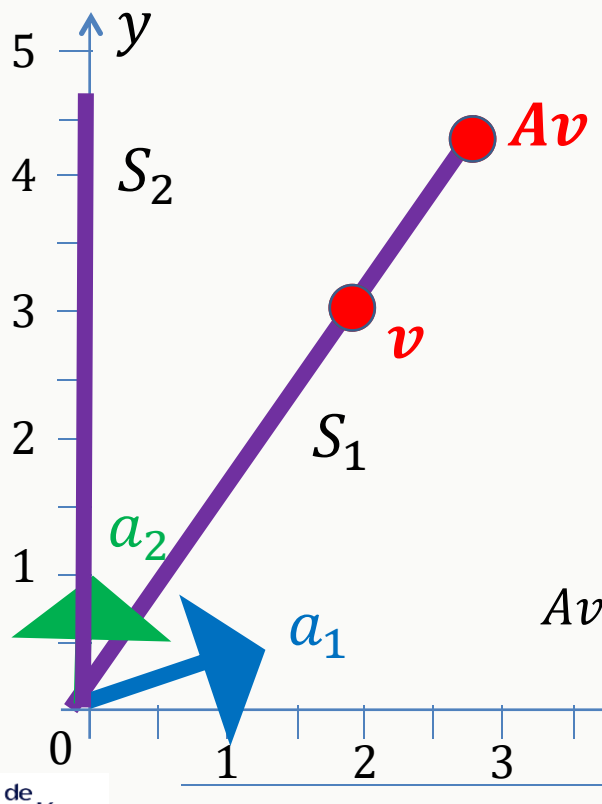
$$A = \begin{bmatrix} a_1, x & a_2, x \\ a_1, y & a_2, y \end{bmatrix} = \begin{bmatrix} 1.39 & 0.00 \\ 0.54 & 1.02 \end{bmatrix}$$

$$v = \begin{bmatrix} 2.00 \\ 2.91 \end{bmatrix}$$

$$Av = \begin{bmatrix} 2.78 \\ 4.04 \end{bmatrix}$$

http://setosa.io/ev/eigenvectors-and-eigenvalues/

INTERPRETACIÓN GEOMÉTRICA DE VALORES Y VECTORES PROPIOS



- Si podemos dibujar una línea que pase por los tres puntos $(0,0)$, v y Av, entonces Av es simplemente v multiplicado por un número λ; es decir, $Av = \lambda v$.
- En este caso, llamamos λ un valor propio (*eigenvalue*) y v un vector propio (*eigenvector*).

$$Av = \begin{bmatrix} 1.39 & 0.00 \\ 0.54 & 1.02 \end{bmatrix} \cdot \begin{bmatrix} 2.00 \\ 2.91 \end{bmatrix} = \begin{bmatrix} 2.78 \\ 4.04 \end{bmatrix} = 1.39 \begin{bmatrix} 2.00 \\ 2.91 \end{bmatrix} \quad \begin{cases} \lambda_1 = 1.39 \\ \lambda_2 = 1.02 \end{cases}$$

CONDICIÓN NECESARIA Y SUFICIENTE

- Si λ es un valor propio de la matriz A $n \times n$, entonces existe una solución no trivial x de la ecuación

$$Ax = \lambda x$$
$$Ax - \lambda x = 0$$
$$Ax - \lambda Ix = 0$$
$$(A - \lambda I)x = 0$$

- Entonces x es un vector propio de A correspondiente al valor propio λ *ssí* x y λ satisfacen

$$(A - \lambda I)x = 0$$

ECUACIÓN CARACTERÍSTICA

- La ecuación $(A - \lambda I)x = 0$ tiene una solución no trivial ($x \neq 0$) ssí la matriz $A - \lambda I$ es no invertible (Teorema de Matriz Invertible). ¿Por qué? Si la matriz es invertible entonces al multiplicar ambos lados por $(A - \lambda I)^{-1}$ la única solución posible es $x = 0$.
- **Teorema**: El determinante de una matriz A $n \times n$ es 0 *ssí la matriz* A es no invertible.
- **Teorema**: Un escalar λ es un valor propio de una matriz A $n \times n$ *ssí* λ satisface la ecuación característica:

$$\det(A - \lambda I) = 0$$

CARACTERÍSTICAS DE VALORES Y VECTORES PROPIOS

1. El espacio propio de una matriz A $n \times n$ correspondiente al valor propio λ de A es el conjunto de todos los vectores propios de A correspondientes a λ.
2. **Teorema**: si $e_1, \dots, e_n$ son vectores propios que corresponden a distintos valores propios (i.e. multiplicidad = 1) de una matriz A $n \times n$, entonces el conjunto $\{e_1, \dots, e_n\}$ es linealmente independiente.
3. Si $\mathrm{P} = (e_1, \dots, e_n)$ es la matriz de vectores propios que corresponden a valores propios $\lambda_1, \dots, \lambda_n$ entonces:

$$e_{\mathrm{i}}^{\mathrm{T}} \mathrm{e_j} = 0, \ \lambda_i \neq \lambda_j \ (\text{vectores ortogonales})$$

$$\mathrm{P}^{\mathrm{T}} P = I$$

$$\mathrm{P}^{\mathrm{T}} = \mathrm{P}^{-1}$$

DIAGONALIZACIÓN DE UNA MATRIZ USANDO VECTORES PROPIOS

- Si A es una matriz $n \times n$, cuyos vectores propios son $\mathrm{P} = (e_1, \ldots, e_n)$ y valores propios distintos son $\{\lambda_1, \cdots, \lambda_n\}$.

Entonces:

$$\mathrm{A} = \mathrm{P}\Lambda\mathrm{P}^{\mathrm{T}} \Longrightarrow \mathrm{P}^{\mathrm{T}}\mathrm{AP} = \Lambda$$

donde Λ es la matriz diagonal:

$$\Lambda = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix}$$

MATRIZ DE COVARIANZA Y VALORES PROPIOS

- Sea $\boldsymbol{\Sigma}$ la matriz de covarianza, $\mathrm{Tr}(\boldsymbol{\Sigma})$ –la traza de $\boldsymbol{\Sigma}$– es la suma de los términos de la diagonal de $\boldsymbol{\Sigma}$
- Si $\lambda_i, i = 1,2, \ldots d$ son los valores propios de $\boldsymbol{\Sigma}$, entonces

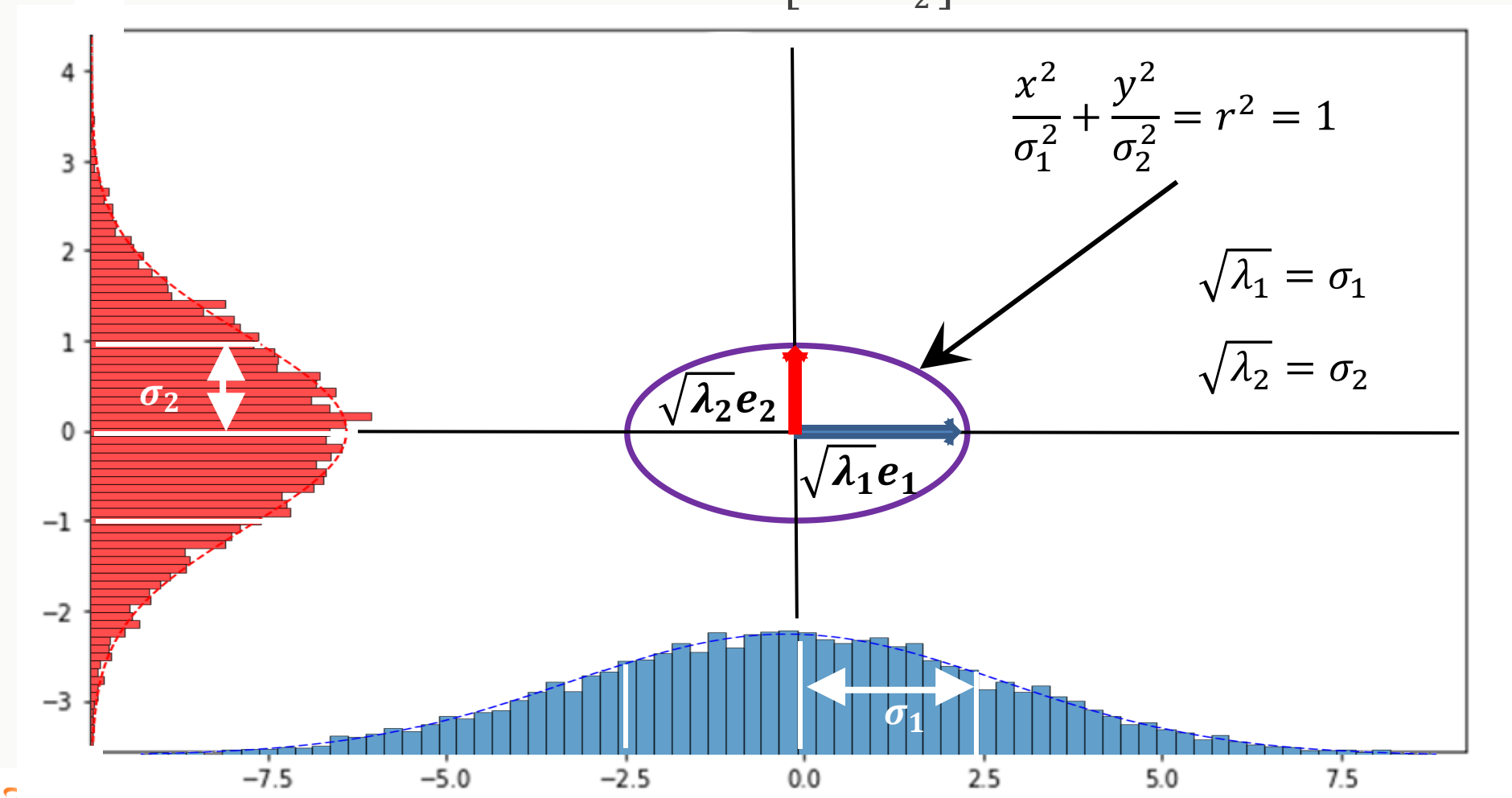
$$\sum_i \lambda_i = \mathrm{Tr}(\boldsymbol{\Sigma})$$

i.e. los valores propios de $\boldsymbol{\Sigma}$ corresponden a las varianzas.

- $\det(\boldsymbol{\Sigma}) = \prod_i \lambda_i$

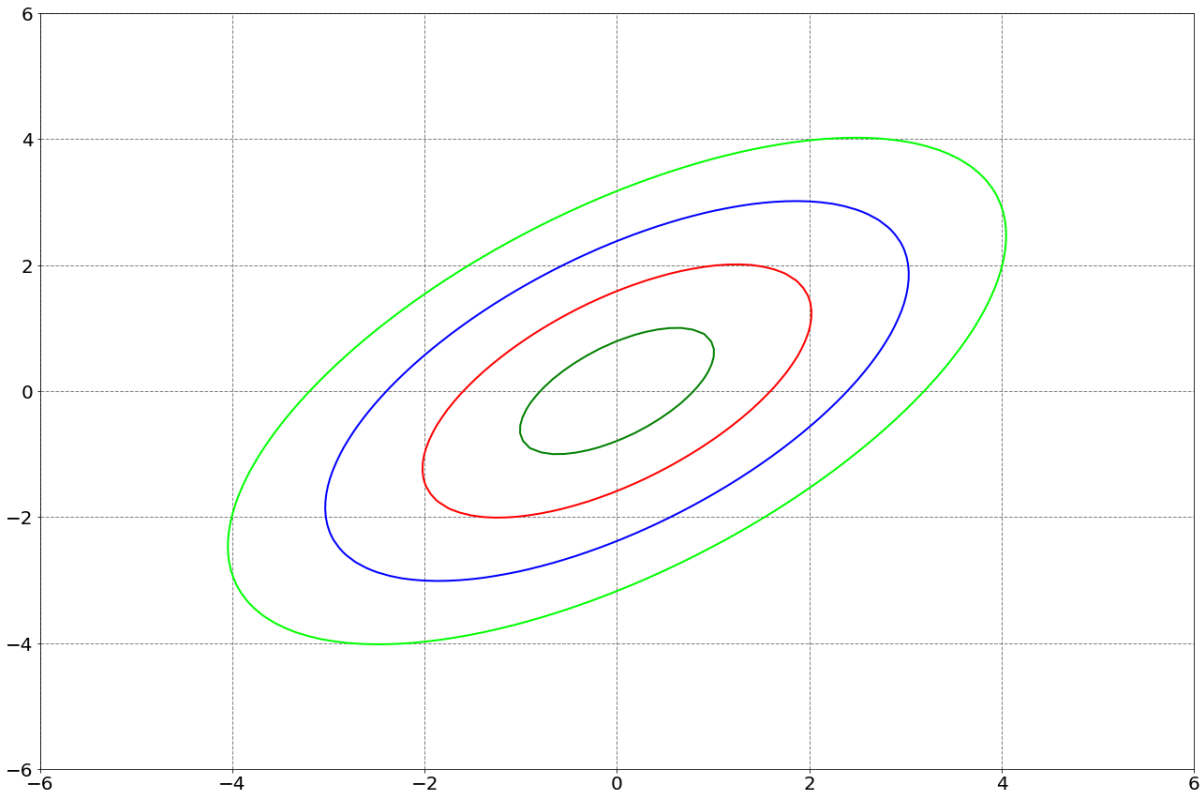
INTERPRETACIÓN GEOMÉTRICA SIN CORRELACIÓN

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} (2.5)^2 & 0 \\ 0 & 1 \end{bmatrix}$$



INTERPRETACIÓN GEOMÉTRICA CON CORRELACIÓN

$$\boldsymbol{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}$$



DIAGONALIZACIÓN MATRIZ DE TRANSFORMACIÓN

- Diagonalización de una matriz (2D) mediante valores y vectores propios
  - *Supongamos la elipse es de la forma:* $\mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x} = r^2$
  - *Pongamos* **M** *en su forma diagonal usando sus vectores y valores propios:*

$$\mathbf{M} = [\boldsymbol{e}_1\ \boldsymbol{e}_2]\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}\begin{bmatrix}\boldsymbol{e}_1^{\mathrm{T}}\\ \boldsymbol{e}_2^{\mathrm{T}}\end{bmatrix}$$

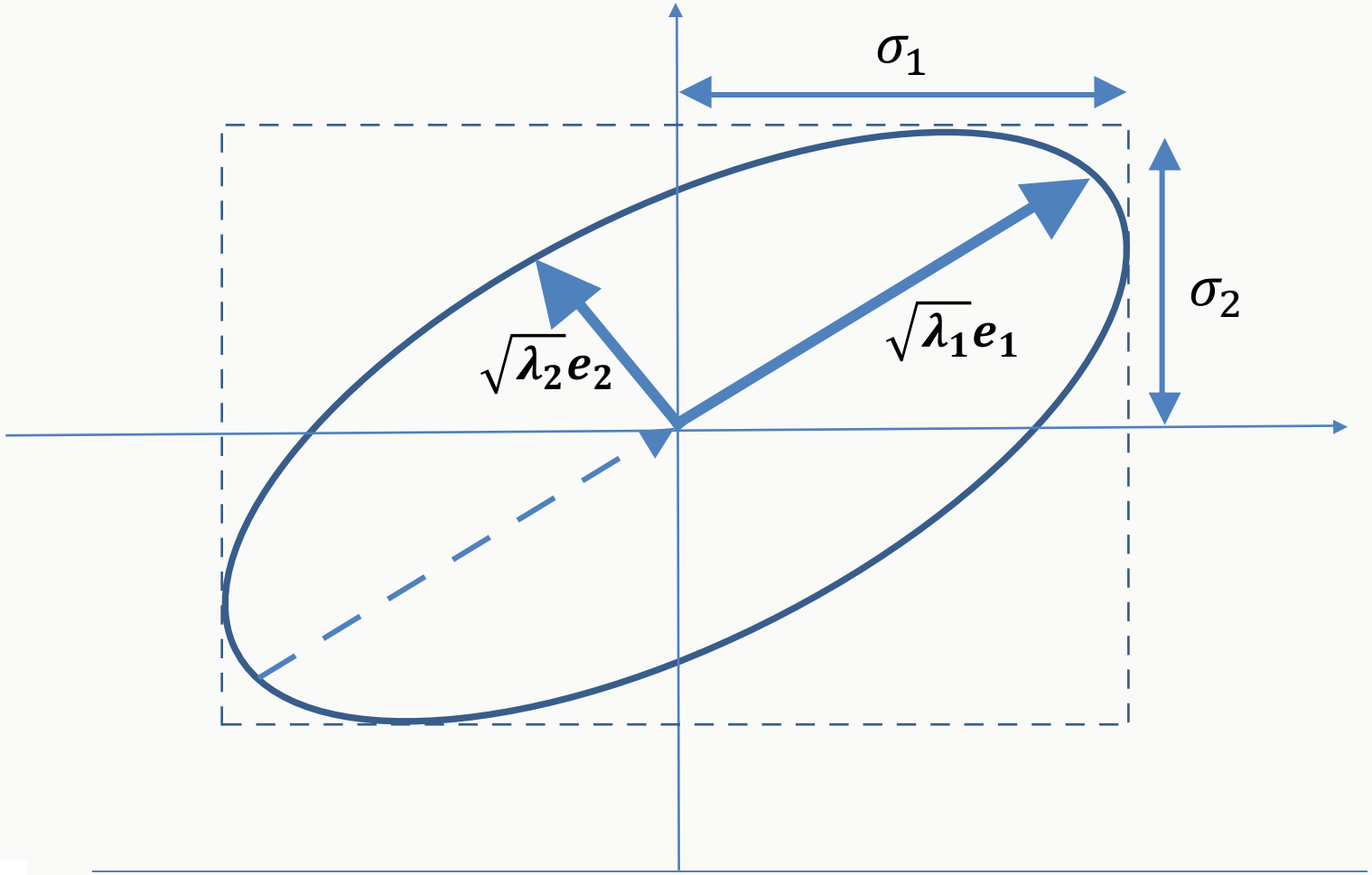
  - *La ecuación de la elipse es ahora:*

$$\mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x} = \mathbf{x}^{\mathrm{T}}[\boldsymbol{e}_1\ \boldsymbol{e}_2]\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix}\begin{bmatrix}\boldsymbol{e}_1^{\mathrm{T}}\\ \boldsymbol{e}_2^{\mathrm{T}}\end{bmatrix}\mathbf{x} = \mathrm{r}^2$$

$$\lambda_1\mathbf{x}^{\mathrm{T}}\mathbf{e}_1\boldsymbol{e}_1^{\mathrm{T}}\mathbf{x} + \lambda_2\mathbf{x}^{\mathrm{T}}\mathbf{e}_2\boldsymbol{e}_2^{\mathrm{T}}\mathbf{x} = \mathrm{r}^2$$

$$\frac{(\boldsymbol{e}_1^{\mathrm{T}}\mathbf{x})^2}{\left(1/\sqrt{\lambda_1}\right)^2} + \frac{(\boldsymbol{e}_2^{\mathrm{T}}\mathbf{x})^2}{\left(1/\sqrt{\lambda_2}\right)^2} = r^2$$ *donde* λ_1 *es el valor propio más pequeño*

INTERPRETACIÓN GEOMÉTRICA CON CORRELACIÓN



PRINCIPAL COMPONENT ANALYSIS (PCA)

REDUCCIÓN DE DIMENSIONALIDAD

- La reducción de dimensionalidad es muy importante en aplicaciones de ciencia de datos, bioinformática, recuperación de información, machine learning y reconocimiento de patrones.
- Existen dos enfoques: supervisado (e.g. Mixture Discriminant Analysis (MDA); Linear Discriminant Analysis (LDA)), y no supervisado (e.g. Independent Component Analysis (ICA); Principal Component Analysis (PCA))

¿PARA QUÉ USAR PCA?

- PCA se usa para:
  - *Encontrar relaciones entre observaciones.*
  - *Extraer la información más importante de los datos.*
  - *Detectar y remover de valores atípicos (“ouliers”).*
  - *Reducir la dimensionalidad para fines de visualización.*
  - *Obtener features para fines de clasificación.*

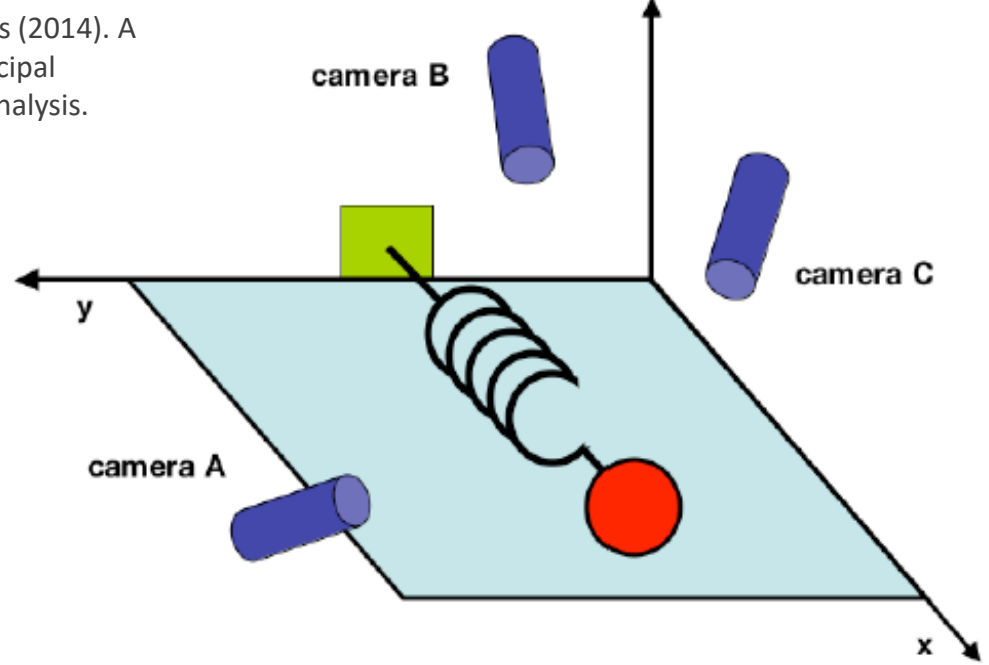
Objetivo de PCA

- Encontrar un espacio de dimensión reducida que se usa para transformar datos de un espacio de dimensión más alta a un espacio de dimensión más pequeña.
- Encontrar factores comunes, los llamados componentes principales, en forma de combinaciones lineales de las variables bajo investigación, y ordenarlas.
- **Identificar la base (vectorial) más significativa para volver a expresar un conjunto de datos**.

EJEMPLO

- El objetivo es obtener una simple ecuación de x. Pero…
- Los datos están referenciados a los ejes de coordendas de cada cámara y los deconocemos.
- Hay ruido en las mediciones
- No sabemos cuál de las señales refleja mejor las propiedades del sistema.
- PCA será útil para conocer x.

Jonathon Shlens (2014). A tutorial on Principal Componennt Analysis.



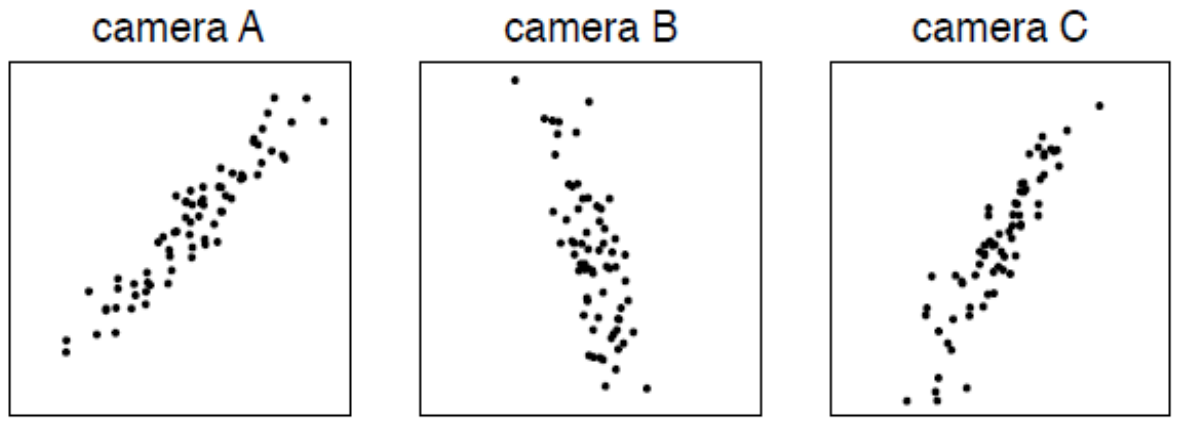
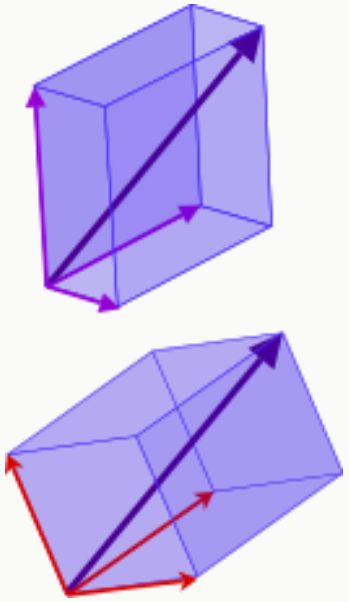


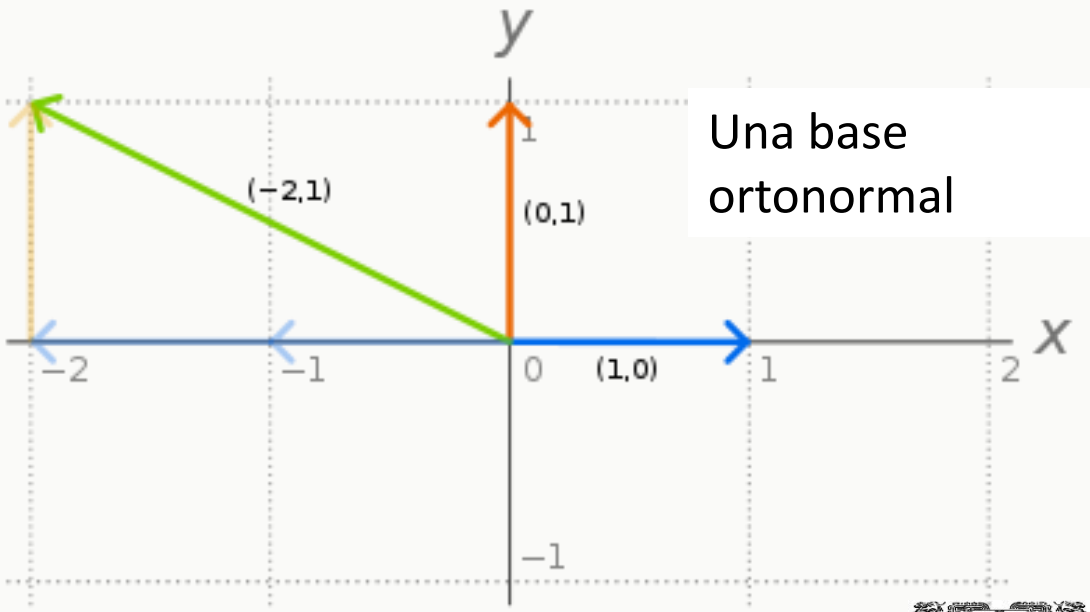
FIG. 1 A toy example. The position of a ball attached to an oscillating spring is recorded using three cameras A, B and C. The position of the ball tracked by each camera is depicted in each panel below.

PCA → CAMBIO DE BASE

- La idea es que PCA nos ayudará a cambiar de base y que en esta nueva base filtraremos el ruido y mostraremos la estructura interna; para el ejemplo: “la dinámica ocurre sobre el eje x”, es decir, la dimensión importante es el eje x.

Dos bases para un mismo vector





Una base ortonormal

PUNTO DE PARTIDA: UNA BASE "INGENUA" PARA LOS DATOS

- Supongamos que nuestros datos:

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix}$$

- Están referenciados a una base:

$$B = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{bmatrix} = \mathbf{I}$$

- Toda observación X sería una combinación lineal de $(b_1, \ldots, b_m)$.
- Esta base refleja el método de medición de nuestros datos.

CAMBIO DE BASE

- La pregunta es ahora:

*¿existe otra base, que sea una **combinación lineal** de la base original, que re-exprese mejor nuestro conjunto de datos?*

- PCA supone **linealidad**: está limitado a re-expresar los datos como una ***combinación lineal*** de sus vectores base.

PRINCIPIO BÁSICO

- Sea $\mathbf{X}$ la matriz $m \times n$ de datos originales, donde cada columna es un vector dimensión m; por lo tanto, n es el número de observaciones.
- Sea $\mathbf{Y}$ otra matriz $m \times n$ relacionada a $\mathbf{X}$ por la transformación $\mathbf{P}$.
- $\mathbf{X}$ es la matriz de datos originales, y $\mathbf{Y}$ es la nueva representación del conjunto de datos:

$$\mathbf{PX} = \mathbf{Y} \qquad (1)$$

INTERPRETACIONES DE PCA

- La ecuación (1) representa un cambio de base y como tal puede tener muchas interpretaciones:
  - $\mathbf{P}$ es una matriz que transforma $\mathbf{X}$ en $\mathbf{Y}$.
  - Geométricamente, $\mathbf{P}$ es una rotación y escalamiento que, nuevamente, transforma $\mathbf{X}$ en $\mathbf{Y}$.
  - Los renglones de $\mathbf{P}$, $\{\mathbf{p_1}, \ldots, \mathbf{p_m}\}$, son un conjunto de nuevos vectores base para expresar las columnas de $\mathbf{X}$.

LOS RENGLONES DE $\mathbf{P}$ SON UN CONJUNTO DE NUEVOS VECTORES BASE PARA EXPRESAR LAS COLUMNAS DE $\mathbf{X}$

$$\mathbf{PX} = \begin{bmatrix} \mathbf{p_1} \\ \vdots \\ \mathbf{p_m} \end{bmatrix} \begin{bmatrix} \mathbf{x_1} & \cdots & \mathbf{x_n} \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} \mathbf{p_1} \cdot \mathbf{x_1} & \cdots & \mathbf{p_1} \cdot \mathbf{x_n} \\ \vdots & \ddots & \vdots \\ \mathbf{p_m} \cdot \mathbf{x_1} & \cdots & \mathbf{p}_m \cdot \mathbf{x}_n \end{bmatrix}$$

Cada coeficiente $\mathbf{y_i}$ de $\mathbf{Y}$ es un producto punto de $\mathbf{x_i}$ con el renglón correspondiente en $\mathbf{P}$.

En otras palabras, el coeficiente j de $\mathbf{y_i}$ es una proyección sobre el renglón j de $\mathbf{P}$.

LO QUE QUEDA POR RESOLVER

- El problema se reduce a encontrar el apropiado cambio de base.
- Los vectores fila $\{\mathbf{p_1}, \ldots, \mathbf{p_m}\}$ en esta transformación serán los **componentes principales** de $\mathbf{X}$.
  - *¿Cuál es la mejor manera de volver a expresar* $\mathbf{X}$*?*
  - *¿Cuál es una buena opción de base* $\mathbf{P}$*?*
- Estas preguntas se responden preguntándonos qué características nos gustaría que Y exhibiera.

RUIDO Y ROTACIÓN

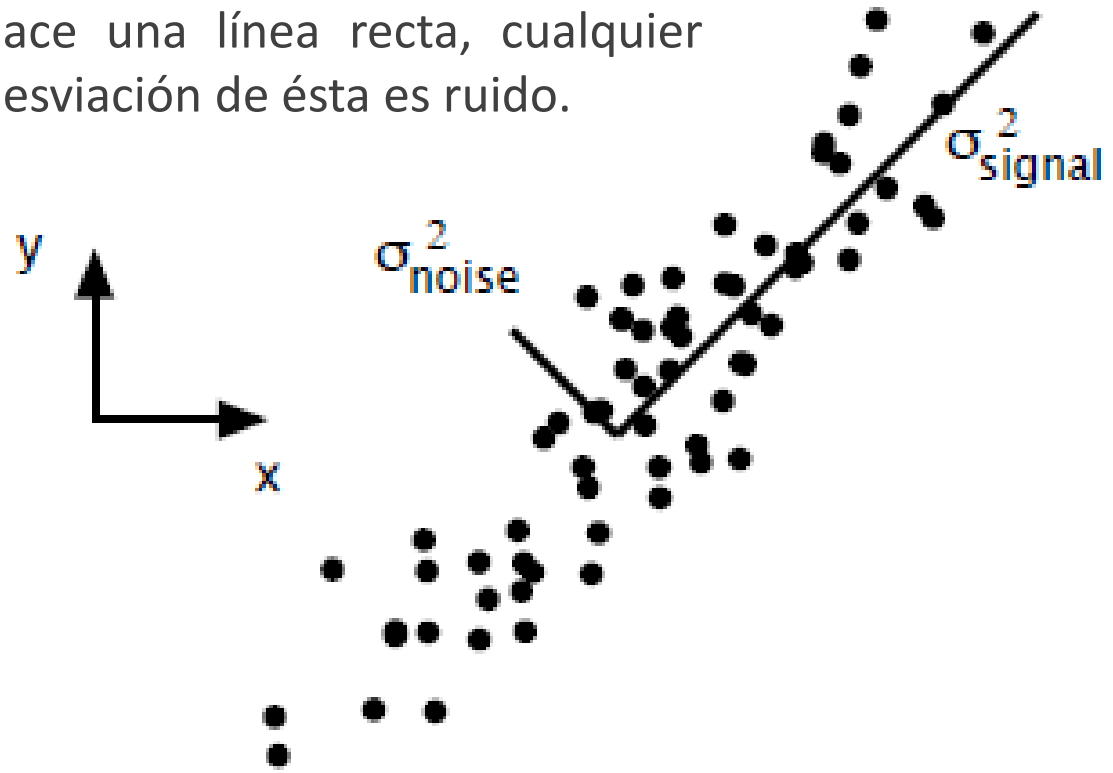
- En Teoría de la Información una medida común de la calidad de una señal es la *Relación Señal a Ruido* (SNR por sus siglas en inglés).

$$SNR = \frac{\sigma^2_{señal}}{\sigma^2_{ruido}}$$

- Un alto valor de SNR (>>1) indica una medición de precisión.

RUIDO Y ROTACIÓN: EL EJEMPLO DE LA CÁMARA

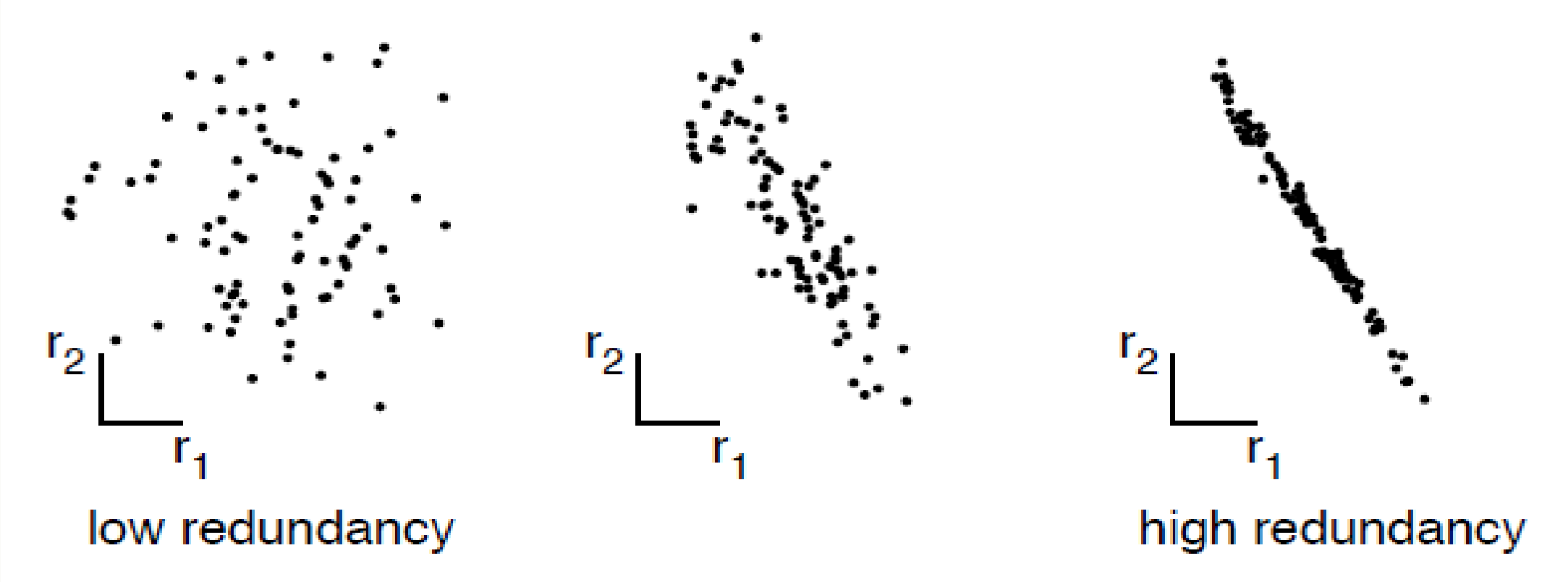
Si suponemos que el resorte hace una línea recta, cualquier desviación de ésta es ruido.



Jonathon Shlens (2014). A tutorial on Principal Componennt Analysis.

- Nota que la dirección de la varianza más grande no está en la dirección de la base del registro (x_A, y_A) sino más bien sobre la línea de mejor ajuste; i.e. la información importante está ahí
- → rotar la base original (ingenua) hacia la línea de mejor ajuste revelaría la dirección del movimiento del resorte en 2-D.

REDUNDANCIA



Jonathon Shlens (2014). A tutorial on Principal Componennt Analysis.

MATRIZ DE COVARIANZA

- La matriz de covarianza permite generalizar la noción de redundancia a un número arbitrario de dimensiones.

$$\mathbf{C_X} = \frac{1}{N}\mathbf{X}\mathbf{X}^{\mathrm{T}}$$

  - *En los términos de la diagonal, como supuesto, valores altos corresponden a estructura interesante.*
  - *En los términos fuera de la diagonal, magnitudes grandes corresponden a alta redundancia.*

DIAGONALIZACIÓN DE LA MATRIZ DE COVARIANZA

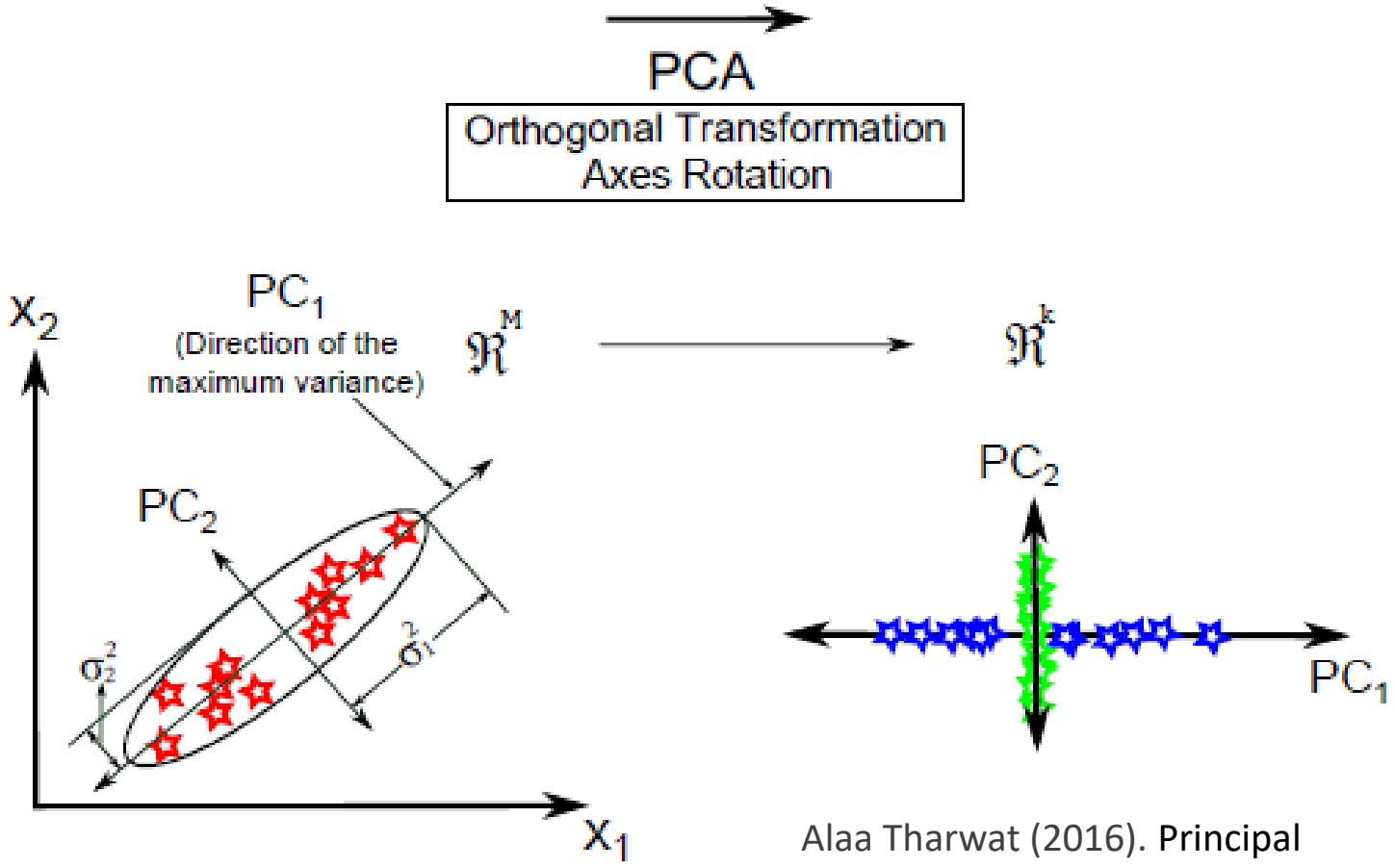
- De acuerdo con lo anterior, para nuestra matriz $\mathbf{C_Y}$ deseamos: (1) minimizar la redundancia, medida por la magnitud de la covarianza, y (2) maximizar la señal, medida por la varianza

¡¡Hay que diagonalizar $\mathbf{C_Y}$*!!*

SUPUESTOS DE PCA

- *Linealidad*: esto enmarca el problema como un cambio de base.
- *Varianza grande tiene estructura importante*: esto abarca la suposición de que los datos tienen una alta SNR. Por lo tanto, las componentes principales con mayores asociadas varianzas representan estructura interesante, mientras que, aquellas con poca varianza representan ruido.
- *Las componentes principales son ortogonales*: una suposición que facilita el cálculo mediante técnicas de álgebra lineal.

SOLUCIÓN UTILIZANDO VECTORES PROPIOS



Alaa Tharwat (2016). **Principal component analysis - a tutorial**

OBJETIVO:

Encuentra una matriz ortonormal $\mathbf{P}$ en $\mathbf{Y} = \mathbf{PX}$ tal que $\mathbf{C_Y} = \frac{\mathbf{1}}{\boldsymbol{n}}\mathbf{YY}^{\mathrm{T}}$ es una matriz diagonal.

Las filas de $\mathbf{P}$ son las componentes principales de $\mathbf{X}$.

EXPRESIÓN DE $\mathbf{C_Y}$ EN FUNCIÓN DE $\mathbf{C_X}$

$$\mathbf{C_Y} = \frac{1}{n}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$$

$$= \frac{1}{n}(\mathbf{P}\mathbf{X})(\mathbf{P}\mathbf{X})^{\mathrm{T}}$$

$$= \frac{1}{n}\mathbf{P}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}$$

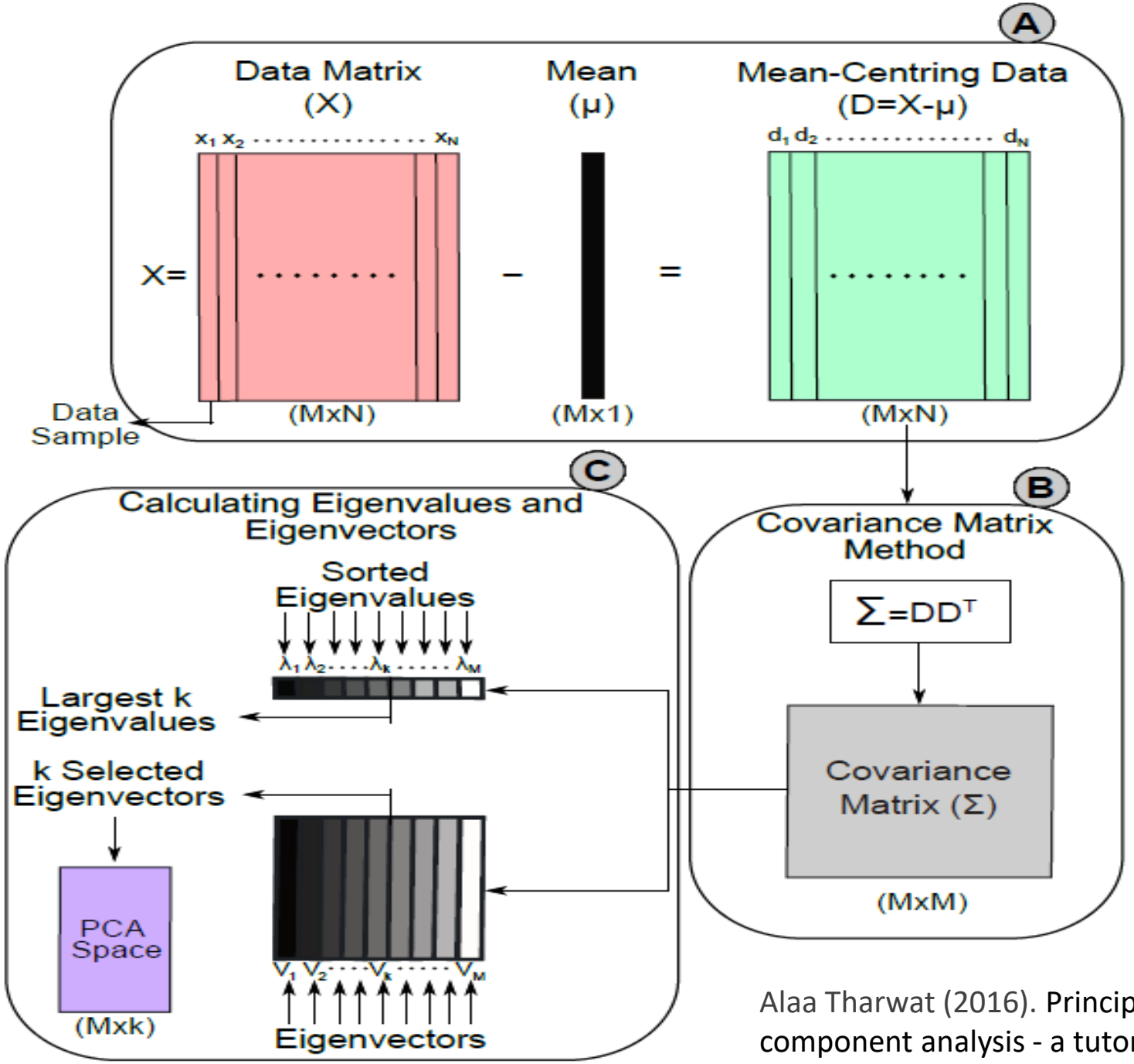
$$= \mathbf{P}\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}$$

$$\mathbf{C_Y} = \mathbf{P}\mathbf{C_X}\mathbf{P}^{\mathrm{T}}$$

DESCOMPOSICIÓN DE $\mathbf{C_X}$ EN VECTORES PROPIOS $\mathbf{E}$ (ASUMIMOS QUE $\mathbf{P} \equiv \mathbf{E}$)

$$\mathbf{C_Y} = \mathbf{P C_X P^T}$$
$$= \mathbf{P(E^T D E)P^T}$$
$$= \mathbf{P(P^T D P)P^T}$$
$$= \mathbf{P P^{-1} D P P^{-1}}$$
$$\mathbf{C_Y} = \mathbf{D}$$

- Las componentes principales de $\mathbf{X}$ son los vectores propios de $\mathbf{C_X} = \frac{1}{n}\mathbf{X X^T}$.
- El valor *i-ésimo* de la diagonal de $\mathbf{C_Y}$ es la varianza de $\mathbf{X}$ a lo largo de $\mathbf{p_i}$



Alaa Tharwat (2016). Principal component analysis - a tutorial

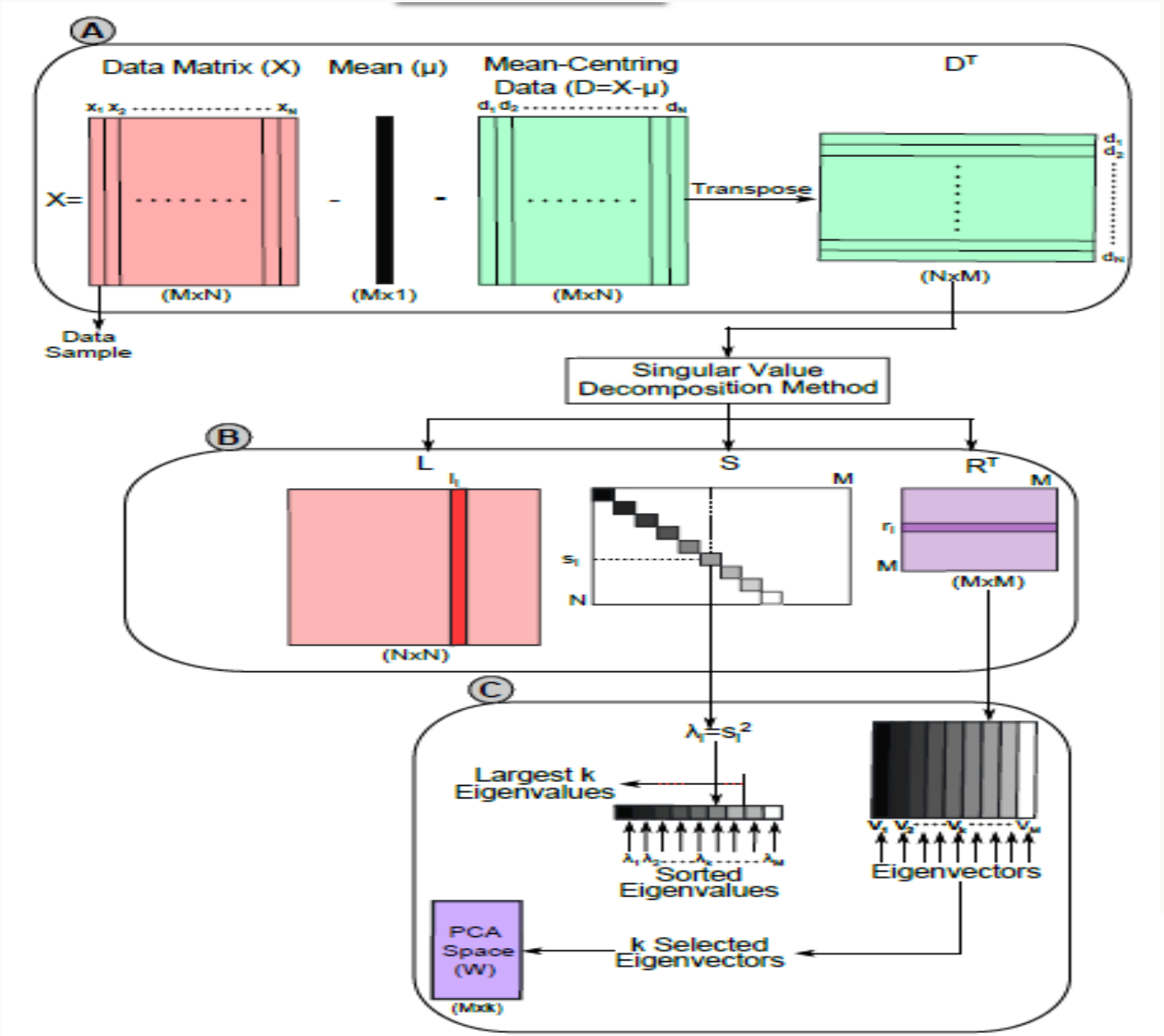
SOLUCIÓN UTILIZANDO DESCOMPOSICIÓN EN VALORES SINGULARES (SVD)

OBJETIVO DE SVD:

Descomponer $\mathbf{X}$ en tres matrices

$$X = LSR^T = \begin{bmatrix} l_1 \cdots l_p \end{bmatrix} \begin{bmatrix} s_1 & 0 & 0 & 0 \\ 0 & s_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & s_q \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -r_1^T - \\ -r_2^T - \\ \vdots \\ -r_q^T - \end{bmatrix}$$

Alaa Tharwat (2016). Principal component analysis - a tutorial



Alaa Tharwat (2016). Principal component analysis - a tutorial