Introducción a la Ciencia de Datos

Maestría en Probabilidad y Estadística

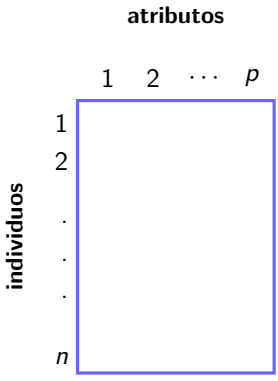
Dr. Marco Antonio Aquino López

Centro de Investigación en Matemáticas

Agosto-Diciembre 2025

Matrices y datos

- Los arreglos matriciales son el objeto natural para guardar información cuantitativa en Ciencias de Datos.



individuos	atributos
clientes	características del cliente
clientes	preferencias y niveles de compra
imágenes	niveles de color en pixeles
individuos	ratings
individuos	niveles de expresión de genes
textos	frecuencias de palabras
moléculas	descriptores
regiones	abundancia de especies

- En las aplicaciones se encuentran todos los casos:

$$n \gg p, \quad n > p, \quad n \approx p, \quad n < p, \quad n \ll p$$

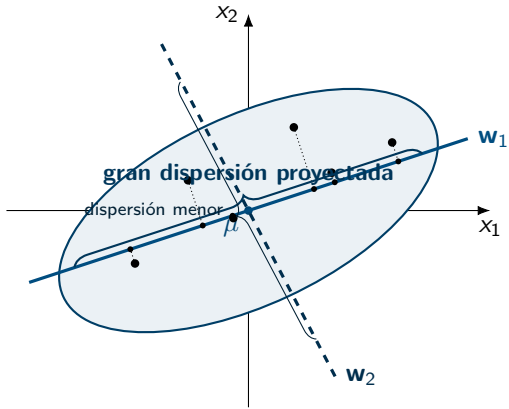
Motivación: ¿Por qué reducir la dimensionalidad?

- **Curse of dimensionality:** en alta dimensión, distancias se vuelven menos informativas; la intuición geométrica falla.
- **Redundancia:** variables correlacionadas aportan información repetida.
- **Ruido:** direcciones de baja varianza pueden distorsionar clustering.
- **Visualización:** requerimos proyecciones 2D/3D para inspección de clusters.

Idea central: encontrar un subespacio k-dimensional ($k \ll p$) que *preserve* la estructura esencial (variabilidad/vecindades) con *mínima pérdida* de información.

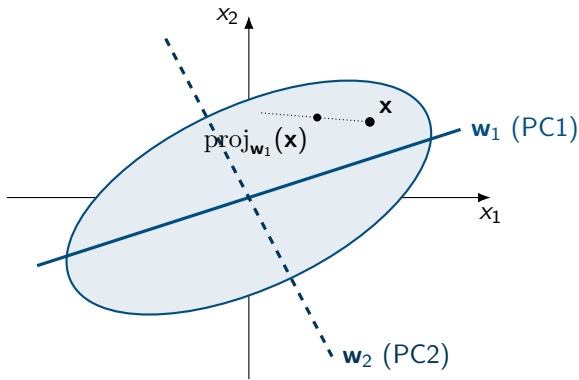
ver 3d

Intuición: direcciones de máxima varianza



- $\mathbf{w}_1$ maximiza esa dispersión (intuición de PCA).

Geometría: direcciones de máxima varianza



$\mathbf{w}_1$: dirección que *maximiza* la varianza proyectada; $\mathbf{w}_2$: siguiente dirección ortogonal con máxima varianza residual.

Hacia el formalismo

Idea clave: elegir $\mathbf{w}_1$ para *maximizar* la varianza de las proyecciones de datos centrados X_c sobre $\mathbf{w}_1$.

- Datos centrados: $X_c = X - \mathbf{1}\mu^\top$.
- Proyección: $t_i = \mathbf{w}^\top \mathbf{x}_{c,i}$.
- Varianza (muestral):

$$\hat{\sigma}^2(\mathbf{w}) = \frac{1}{n-1}\sum_{i=1}^{n}(t_i - \bar{t})^2.$$

- Con centrado, $\bar{t} = 0$, y

$$\hat{\sigma}^2(\mathbf{w}) = \mathbf{w}^\top \left(\tfrac{1}{n-1} X_c^\top X_c\right)\mathbf{w}.$$

Problema (preliminar)

$$\max_{\|\mathbf{w}\|=1} \mathbf{w}^\top S \mathbf{w}, \qquad S = \tfrac{1}{n-1} X_c^\top X_c.$$

Siguiente: mostrar que la solución es el *eigenvector principal*.

Fundamento matemático de PCA

Planteamiento

Sea $X_c \in \mathbb{R}^{n \times p}$ el conjunto de datos centrados y $S = \frac{1}{n} X_c^\top X_c$ su matriz de covarianzas muestral.

Theorem (Dirección de máxima varianza)

Sea $\mathbf{w} \in \mathbb{R}^p$ *con* $\|\mathbf{w}\| = 1$. *La varianza de las proyecciones* $t_i = \mathbf{w}^\top \mathbf{x}_{c,i}$ *es* $\sigma^2(\mathbf{w}) = \mathbf{w}^\top S \mathbf{w}$. *La dirección que maximiza* $\sigma^2(\mathbf{w})$ *es el* ***autovector principal*** *de* S *asociado a su mayor autovalor* λ_1.

Esquema.

Resolvemos

$$\max_{\mathbf{w}} \mathbf{w}^\top S \mathbf{w} \quad \text{sujeto a } \|\mathbf{w}\|^2 = 1.$$

Formamos el Lagrangiano

$$\mathcal{L}(\mathbf{w}, \lambda) = \mathbf{w}^\top S \mathbf{w} - \lambda(\mathbf{w}^\top \mathbf{w} - 1),$$

y derivando respecto a $\mathbf{w}$:

$$\nabla_{\mathbf{w}} \mathcal{L} = 2S\mathbf{w} - 2\lambda \mathbf{w} = 0 \implies S\mathbf{w} = \lambda \mathbf{w}.$$

λ es la varianza proyectada y la dirección óptima $\mathbf{w}_1$ corresponde al mayor λ. □

Subespacio de dimensión k y ortogonalidad

Proposición (Bases ortogonales de componentes principales)

Sea $W_k = [\mathbf{w}_1, \ldots, \mathbf{w}_k] \in \mathbb{R}^{p \times k}$ *con* $\mathbf{w}_i^\top \mathbf{w}_j = \delta_{ij}$. *El problema generalizado de PCA es*

$$\max_{W_k^\top W_k = I_k} \operatorname{Tr}(W_k^\top S W_k).$$

Resultado

Por el teorema espectral, la solución está dada por los k eigenvectores principales de S:

$$S = V \Lambda V^\top, \quad W_k = [\mathbf{v}_1, \ldots, \mathbf{v}_k],$$

donde $\lambda_1 \geq \cdots \geq \lambda_p$ son los eigenvalores.

Varianza explicada y su interpretación

Autovalores y varianza proyectada

Para cada componente principal $\mathbf{w}_i$:

$$S\mathbf{w}_i = \lambda_i \mathbf{w}_i, \quad \|\mathbf{w}_i\| = 1.$$

La varianza de las proyecciones sobre $\mathbf{w}_i$ es

$$\mathrm{Var}(\mathbf{w}_i^\top X_c) = \mathbf{w}_i^\top S \mathbf{w}_i = \lambda_i.$$

Varianza total

La varianza total de los datos centrados es la traza de S:

$$\mathrm{tr}(S) = \sum_{j=1}^{p} S_{jj} = \sum_{i=1}^{p} \lambda_i.$$

Fracción de varianza explicada (EVR)

$$\mathrm{EVR}(k) = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i}, \quad \text{con } \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p.$$

Interpretación: mide qué proporción de la dispersión total se conserva.

Motivación: del máximo de varianza al mínimo de error

Hasta ahora, PCA se formuló como el problema de encontrar las direcciones que *maximizan la varianza proyectada*.

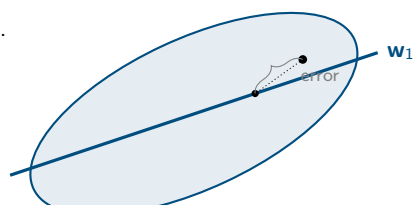
Otra visión equivalente —y útil desde un punto de vista geométrico y de compresión— es:

Idea principal

Buscar el subespacio lineal de dimensión k que **minimiza el error cuadrático medio de reconstrucción**.

- Los datos se "comprimen" al proyectarse en un plano o subespacio.
- Luego se "reconstruyen" desde ese subespacio.
- La mejor proyección es la que minimiza la pérdida de información:

$$\min_{\hat{X}} \ \|X_c - \hat{X}\|_F^2.$$



Formulación matemática del problema

Sea $X_c \in \mathbb{R}^{n \times p}$ el conjunto de datos centrado. Buscamos matrices $W \in \mathbb{R}^{p \times k}$ y $Z \in \mathbb{R}^{n \times k}$ que minimicen:

$$\min_{Z,W} \|X_c - ZW^\top\|_F^2 \quad \text{sujeto a } W^\top W = I_k.$$

- W: define las **direcciones ortonormales** del subespacio (loadings).
- Z: contiene las **coordenadas proyectadas** (scores).

Interpretación

$\hat{X} = ZW^\top$ es una reconstrucción de rango k del conjunto de datos. El objetivo es minimizar el error cuadrático medio de reconstrucción.

Proyección óptima sobre un subespacio

Dado un subespacio generado por las columnas de W, la mejor reconstrucción de los datos se obtiene al *proyectarlos ortogonalmente* sobre dicho subespacio.

Problema

$$\min_{Z,W} \|X_c - ZW^\top\|_F^2 \quad \text{sujeto a } W^\top W = I_k.$$

- Para un W fijo, la solución óptima para Z es:

$$Z = X_c W.$$

- Intuitivamente: cada fila de Z son las coordenadas de un punto proyectado sobre el subespacio generado por W.

Conclusión: PCA proyecta los datos de manera ortogonal sobre el subespacio de menor dimensión.

Reducción a un problema en W

Sustituyendo $Z = X_c W$ en la función de error, obtenemos:

$$J(W) = \|X_c - X_c W W^\top\|_F^2.$$

Interpretación geométrica

- $WW^\top$ es el proyector ortogonal sobre el subespacio generado por W.
- $I - WW^\top$ mide la distancia de cada punto a su proyección.

Resultado clave

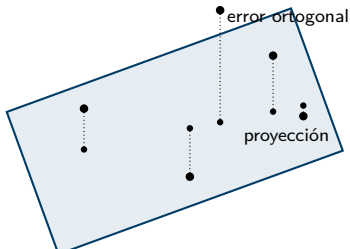
Minimizar el error equivale a maximizar la varianza retenida:

$$\min_W J(W) \iff \max_{W^\top W = I_k} \operatorname{tr}(W^\top X_c^\top X_c W).$$

Interpretación geométrica y error de reconstrucción

- PCA busca el subespacio de dimensión k que *minimiza la distancia ortogonal promedio* entre los puntos y el plano de proyección.
- La reconstrucción es $\hat{X} = X_c W W^\top$.
- El error cuadrático mínimo se relaciona con los eigenvalores omitidos:

$$\|X_c - \hat{X}\|_F^2 = (n-1) \sum_{i=k+1}^{p} \lambda_i.$$



Conclusión: PCA proporciona la mejor aproximación lineal (en MSE) de los datos en un subespacio de rango k.

Code

Motivación: de la covarianza a la SVD

La formulación clásica de PCA usa la matriz de covarianza

$$S = \frac{1}{n-1} X_c^\top X_c.$$

Sin embargo, calcular sus eigenvectores puede ser costoso cuando p es grande o $n < p$.

Idea

Usar la **descomposición en valores singulares (SVD)** del conjunto de datos centrado:

$$X_c = U \Sigma V^\top.$$

- U: direcciones ortogonales de las **observaciones**.
- V: direcciones ortogonales de las **variables**.
- Σ: valores singulares (miden la escala de variación).

Ventaja: Evita calcular S directamente y es numéricamente estable.

Conexión entre SVD y PCA

Multiplicando $X_c = U\Sigma V^\top$ por $X_c^\top$:

$$S = \frac{1}{n-1} X_c^\top X_c = V \left(\frac{\Sigma^\top \Sigma}{n-1} \right) V^\top.$$

Consecuencias

- Los **autovectores de** S son las columnas de V.
- Los **autovalores** son $\lambda_i = \sigma_i^2/(n-1)$.
- Los **componentes principales (scores)** son:

$$Z = X_c V_k = U_k \Sigma_k.$$

Interpretación: la SVD reexpresa los datos como una rotación ortogonal + escalamiento según su varianza.

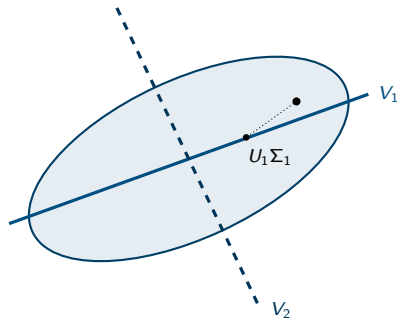
Interpretación geométrica de la SVD

- V_k: ejes principales (direcciones de máxima varianza).
- $U_k\Sigma_k$: coordenadas proyectadas de cada punto en esos ejes.
- La reconstrucción de rango k es:

$$\hat{X}_k = U_k \Sigma_k V_k^\top .$$

- El error de reconstrucción:

$$\|X_c - \hat{X}_k\|_F^2 = (n-1) \sum_{i=k+1}^{p} \lambda_i.$$



Conclusión: La SVD muestra que PCA es una *rotación ortogonal* que ordena las direcciones según la varianza de los datos.

Pausa

Pausa

Reducción de dimensión: un recordatorio

En aprendizaje no supervisado, la reducción de dimensión busca representar los datos $X \in \mathbb{R}^{n \times p}$ en un espacio de menor dimensión $r \ll p$.

- **PCA/SVD:** encuentra combinaciones lineales ortogonales que maximizan la varianza.
- **Problema:** las combinaciones pueden ser negativas, lo cual dificulta la interpretación.

Ejemplo: matriz de imágenes o conteos

Si cada pixel o palabra tiene un valor no negativo, ¿cómo interpretamos una combinación lineal con pesos negativos?

Necesitamos una descomposición "aditiva", no sustractiva.

Idea central de NMF

La **Factorización No Negativa de Matrices (NMF)** busca representar los datos como

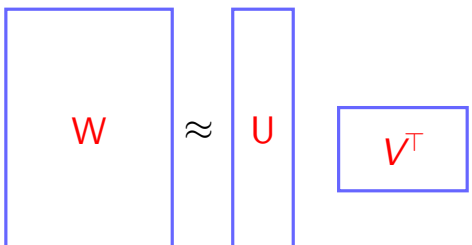
$$V \approx WH, \quad \text{con } W, H \geq 0.$$

- Cada columna de V (una observación) se aproxima como combinación *aditiva* de las columnas de W.
- Las columnas de W representan **partes básicas** o **temas latentes**.
- Las filas de H indican **cuánto contribuye cada parte** en cada observación.

Interpretación: NMF descubre estructuras interpretables, con componentes positivos que "se suman" para formar el todo.

Factorización no negativa de matrices

NMF aproxima los datos, en los renglones de $W_{n\times p}$ (W con entradas no negativas), como combinaciones lineales de una base $V_{p\times k}$ y pesos $U_{n\times k}$:



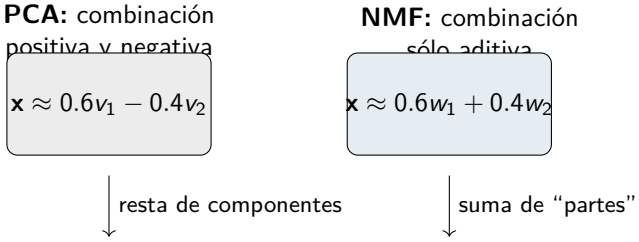
$$W = \begin{bmatrix} w_1^\top \\ \vdots \\ w_n^\top \end{bmatrix} \approx \begin{bmatrix} u_1^\top \\ \vdots \\ u_n^\top \end{bmatrix} V^\top$$

y para cada fila i : $\quad w_i \approx V u_i = u_{i1} v_1 + \cdots + u_{ik} v_k.$

$$W \approx UV^\top, \qquad U, V \geq 0$$

donde el rango k de la factorización es definido por el usuario.

Comparación conceptual con PCA



Ejemplo

En análisis de texto, cada documento se representa como mezcla de temas; en imágenes, cada rostro como combinación de partes (ojos, nariz, boca).

Aplicaciones típicas

- **Minería de textos:** detección de temas latentes.
- **Reconocimiento facial:** representación por "partes" en lugar de vectores globales.
- **Recomendadores:** descubrimiento de grupos de usuarios o productos.

$$V \approx WH$$

$\Rightarrow$ datos = combinación positiva de factores interpretables

Definición del problema

Sea $V \in \mathbb{R}^{m \times n}$ una matriz de datos con $V_{ij} \geq 0$. Buscamos dos matrices:

$$W \in \mathbb{R}^{m \times r}, \quad H \in \mathbb{R}^{r \times n},$$

tales que:

$$W_{ik} \geq 0, \quad H_{kj} \geq 0, \quad \text{y} \quad V \approx WH.$$

Dimensiones

$$\underbrace{V}_{m \times n} \approx \underbrace{W}_{m \times r} \underbrace{H}_{r \times n}$$

- r: número de factores latentes o "componentes".
- Columnas de W: patrones o partes básicas.
- Filas de H: pesos no negativos para cada observación.

Función objetivo general

El problema de NMF se formula como:

$$\min_{W,H\geq 0} f(V,W,H) \quad \text{donde} \quad f(V,W,H) = D(V\|WH)$$

y $D(\cdot\|\cdot)$ es una **divergencia** o medida de error entre matrices.

Dos funciones de costo más comunes

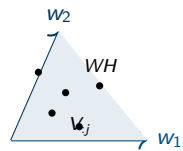
1. **Norma de Frobenius:**
$$D_{\mathrm{F}}(V\|WH) = \tfrac{1}{2}\|V-WH\|_F^2 = \tfrac{1}{2}\sum_{ij}(V_{ij}-(WH)_{ij})^2.$$

2. **Divergencia de Kullback–Leibler:**
$$D_{\mathrm{KL}}(V\|WH) = \sum_{ij}\left(V_{ij}\log\frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}\right).$$

Ambas funciones son convexas en W o en H individualmente, pero no en el par (W,H).

Interpretación geométrica

- Cada columna de V se representa como combinación no negativa de las columnas de W.
- El conjunto de todas las combinaciones posibles de W forma un **cono convexo**.
- NMF busca el cono convexo de dimensión r que mejor aproxima a los datos.



La factorización busca una representación aditiva dentro de un cono positivo.

Naturaleza no convexa del problema

El problema

$$\min_{W,H\geq 0} D(V\|WH)$$

no es convexo en W y H simultáneamente.

Consecuencias prácticas

- No existe garantía de encontrar el mínimo global.
- Se usan métodos iterativos (alternar entre W y H).
- La calidad del resultado depende de la inicialización.

Idea clave: aunque es no convexo, suele producir soluciones significativas e interpretables.

Panorama del problema

Recordatorio: dado $V \in \mathbb{R}_+^{m\times n}$, buscamos $W \in \mathbb{R}_+^{m\times r}$, $H \in \mathbb{R}_+^{r\times n}$ con

$$\min_{W,H\geq 0} D(V\|WH),$$

donde típicamente

$$D_{\mathrm{F}}(V\|WH) = \tfrac{1}{2}\|V - WH\|_F^2, \qquad D_{\mathrm{KL}}(V\|WH) = \sum_{ij}\left(V_{ij}\log\tfrac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}\right).$$

- Problema no convexo en (W, H) conjunto; convexo en cada bloque por separado.
- Estrategia: **alternar** actualizando H con W fijo, y viceversa.

Actualizaciones multiplicativas (MU) para Frobenius

Para $D_{\mathrm{F}}(V\|WH) = \frac{1}{2}\|V - WH\|_F^2$, las reglas MU clásicas (Lee & Seung) son:

$$H \leftarrow H \odot \frac{W^\top V}{W^\top WH + \varepsilon} \qquad W \leftarrow W \odot \frac{VH^\top}{WHH^\top + \varepsilon}$$

donde $\odot$ y la fracción son operaciones *elemento a elemento*; $\varepsilon > 0$ evita divisiones por cero.

- Se derivan imponiendo KKT y usando *majorization-minimization* para garantizar $\Delta D_{\mathrm{F}} \leq 0$.
- Mantienen la no negatividad (multiplican por factores ≥ 0).
- Sencillas de implementar; convergencia sublineal; sensibles a escalado.

Esquema alternado

1. Con W fijo, actualiza H con la regla MU.
2. Con H fijo, actualiza W con la regla MU.
3. Repite hasta parada (tolerancia o máximo de iteraciones).

Actualizaciones multiplicativas (MU) para KL

Para la divergencia de Kullback–Leibler:

$$D_{\mathrm{KL}}(V\|WH) = \sum_{ij}\left(V_{ij}\log\frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}\right),$$

las reglas MU son:

$$\boxed{H \leftarrow H \odot \frac{W^{\top}\left(\frac{V}{WH}\right)}{W^{\top}\mathbf{1}}} \qquad \boxed{W \leftarrow W \odot \frac{\left(\frac{V}{WH}\right)H^{\top}}{\mathbf{1}H^{\top}}}$$

donde $\frac{V}{WH}$ es elemento a elemento y $\mathbf{1}$ es una matriz/tensor de unos con dimensiones compatibles.

- Muy usada en *text mining* y conteos tipo Poisson.
- También garantiza no incremento de D_{KL} bajo condiciones estándar.

Referencias


Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.


Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of Machine Learning* (2nd ed.). MIT Press.


Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer.


Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.