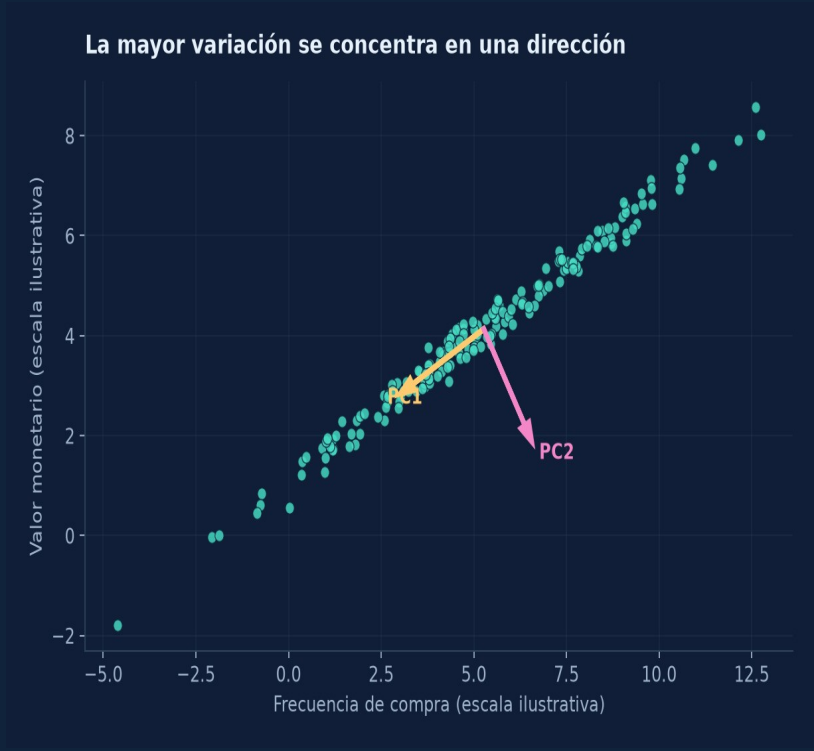
CIENCIA DE DATOS PARA DECISIONES DE NEGOCIO

MÓDULO 10

PCA · Análisis de Componentes Principales

Desde la geometría de la varianza hasta una representación defendible para el negocio.

Clases conceptuales · Matemáticas · Evidencia visual · Decisión de negocio



Sergio Gevatschnaider · Material de cátedra

Antes de reducir dimensiones: ¿qué queremos preservar?

IDEAS CLAVE

- Conservar variación útil sin confundir variación con relevancia empresarial.
- Sintetizar correlaciones y visualizar datos de muchas variables.
- Preparar una representación para explorar o modelar, sin etiquetas.
- Evaluar qué información e interpretabilidad se pierden al comprimir.

PREGUNTAS DE DECISIÓN

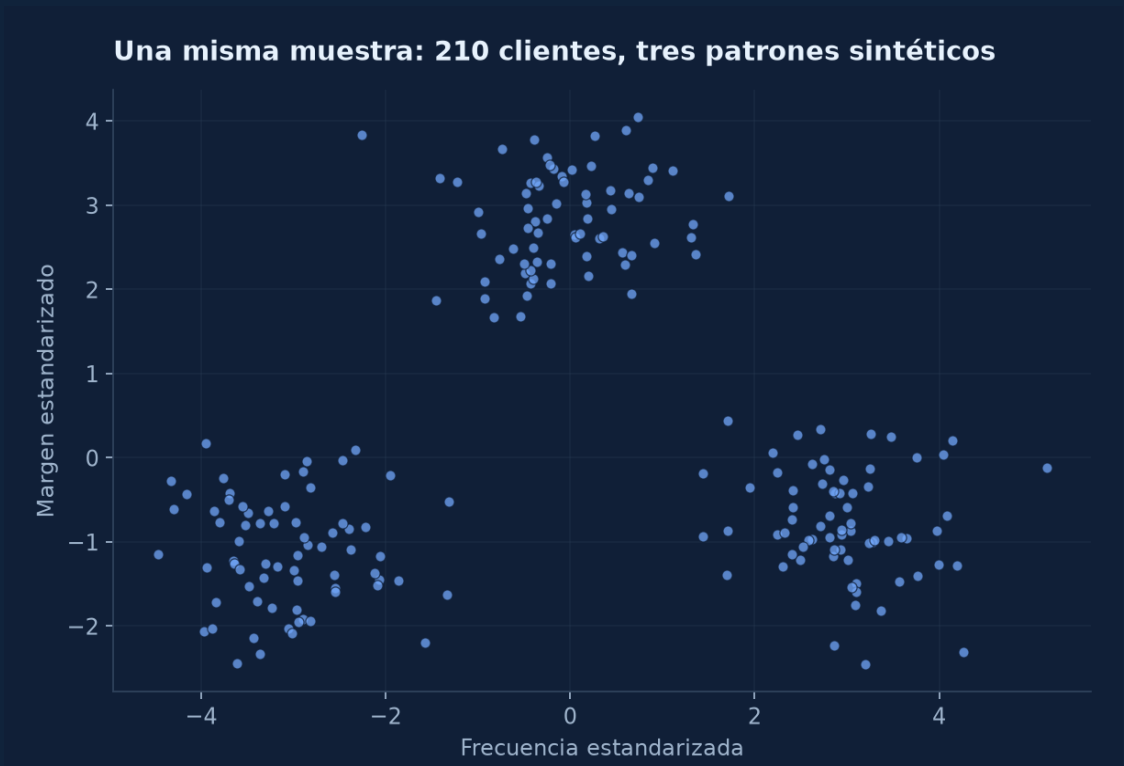
1. ¿Qué decisión necesita el mapa de datos?
2. ¿Qué variables pueden dominar por su escala?
3. ¿Qué estructura conservaría la proyección?
4. ¿Cómo evaluaríamos pérdida y utilidad?

IDEA QUE DEBE QUEDAR **PCA maximiza varianza retenida; no maximiza valor de negocio.**

Caso conductor: clientes con métricas heterogéneas

IDEAS CLAVE

- Una fila representa un cliente observado durante 90 días.
- Variables: recencia, frecuencia, gasto, margen, uso digital.
- Las escalas mezclan días, cantidades y unidades monetarias.
- Buscamos estructura para explorar clientes, no crear etiquetas de “buenos” o “malos”.

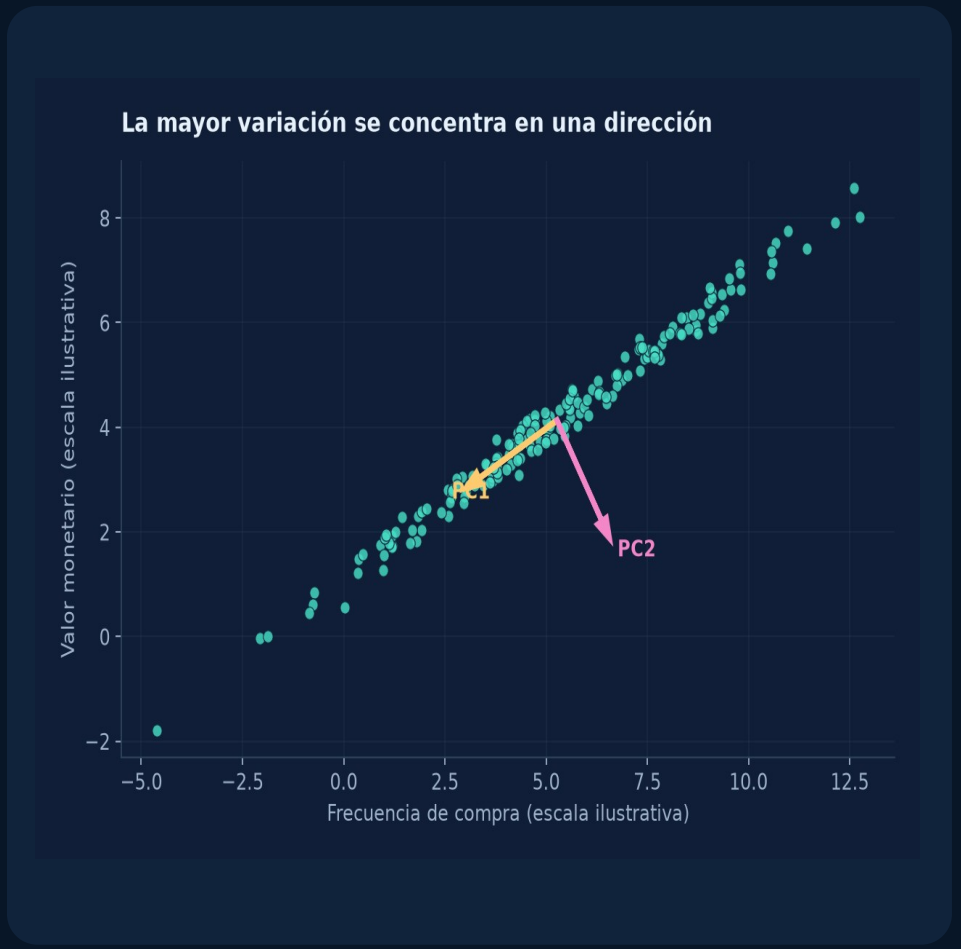


ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **Primero definir unidad de análisis, ventana temporal y variable de acción.**

01 · Intuición geométrica

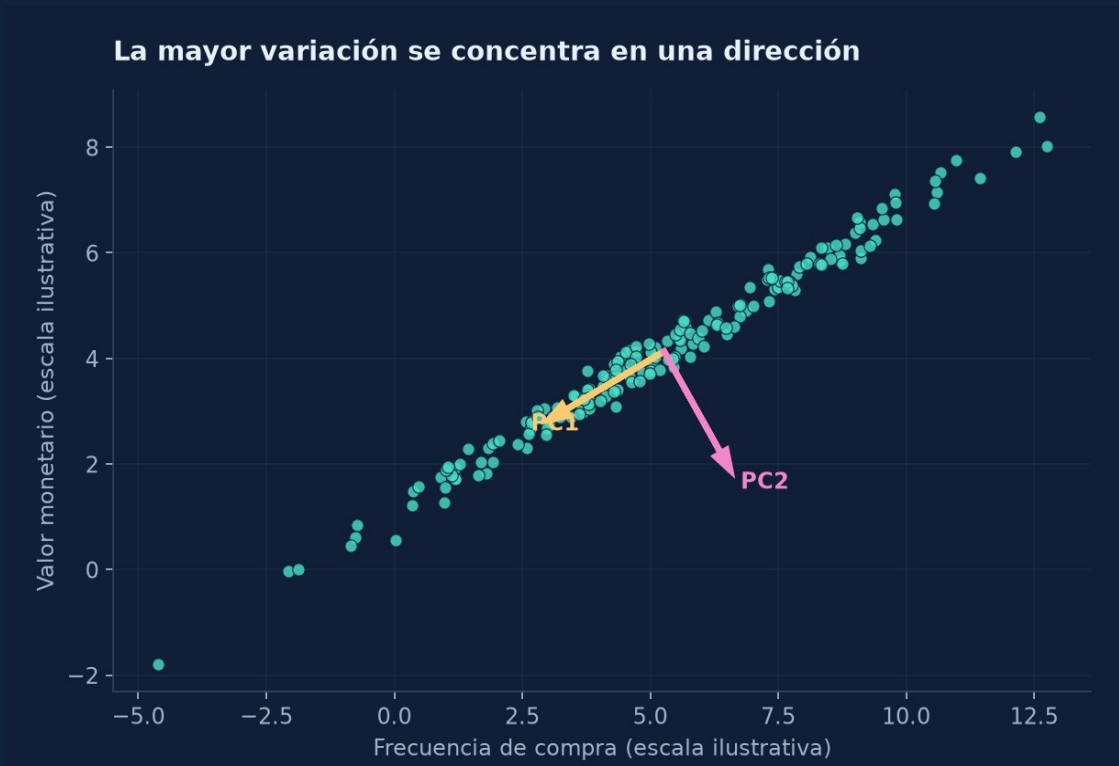
Una rotación de ejes encuentra las direcciones que contienen más variabilidad.



PCA no borra variables: construye nuevos ejes

IDEAS CLAVE

- PC1 es la dirección que captura la mayor varianza de la nube centrada.
- PC2 es ortogonal a PC1 y explica la mayor varianza residual.
- Las componentes son combinaciones lineales de las variables.
- Con todas las componentes, el cambio de base no pierde información.



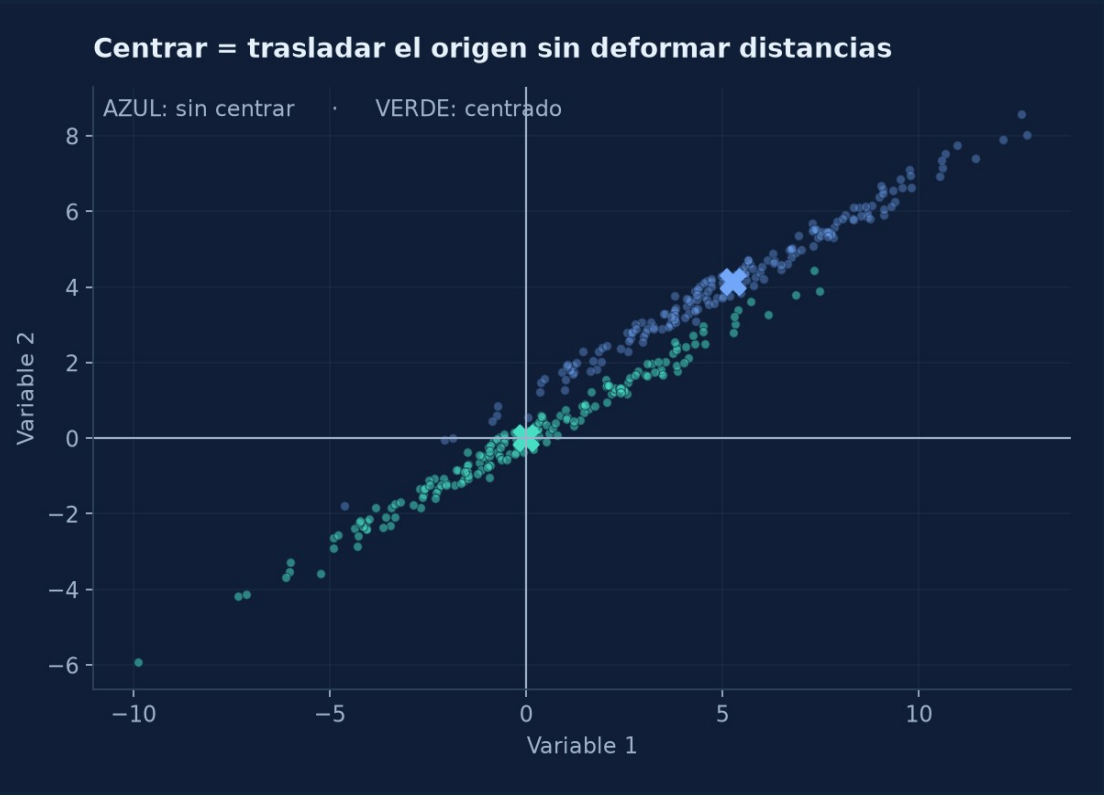
ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **Reducir dimensión significa descartar componentes, no solo rotar ejes.**

Centrar los datos: transformar X en X − μ

IDEAS CLAVE

- Restar la media de cada variable sitúa el centroide de la muestra en cero.
- Las distancias relativas entre observaciones no cambian.
- Sin centrado, los ejes pueden reflejar la localización respecto del origen.
- Una transformación aprendida en train debe aplicarse igual a test.



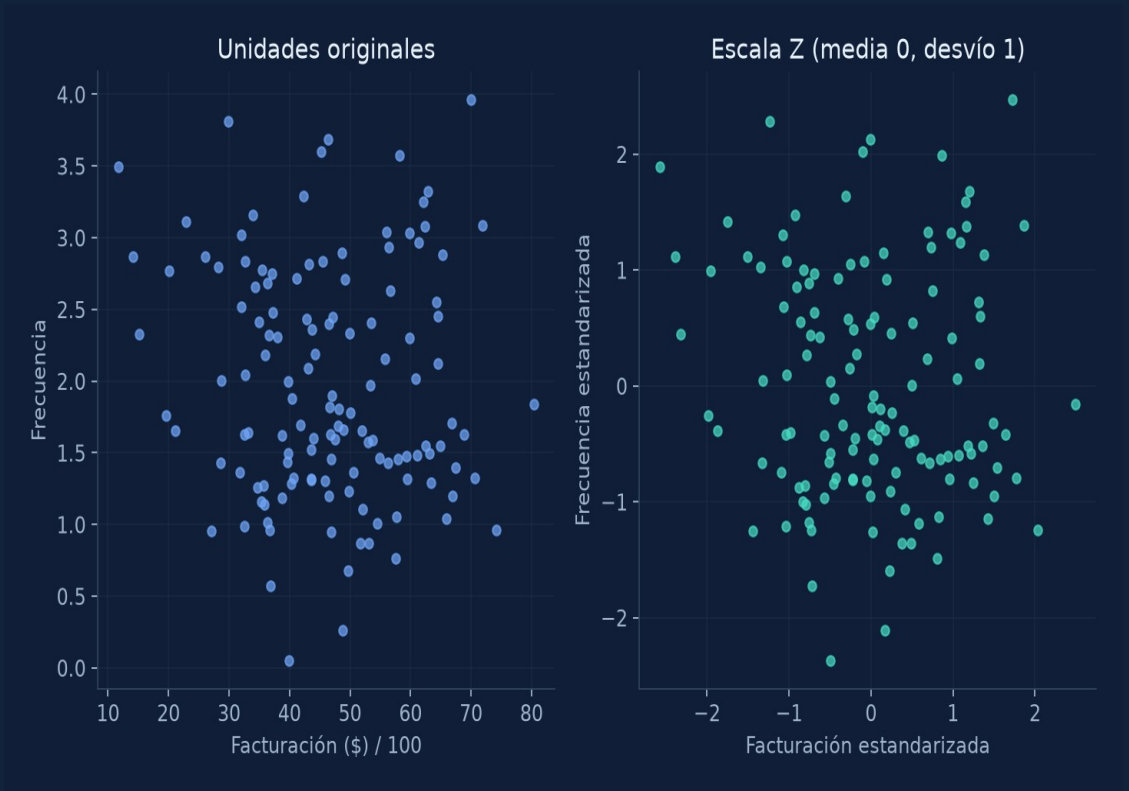
ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **Centrar no equivale a escalar; son decisiones distintas.**

Escalar o no escalar: el peso de las unidades

IDEAS CLAVE

- PCA sobre covarianzas conserva el peso de la dispersión original.
- PCA sobre variables Z equivale a operar sobre correlaciones.
- Si las unidades difieren mucho, una variable puede dominar por escala.
- La estandarización también puede amplificar variables muy ruidosas.



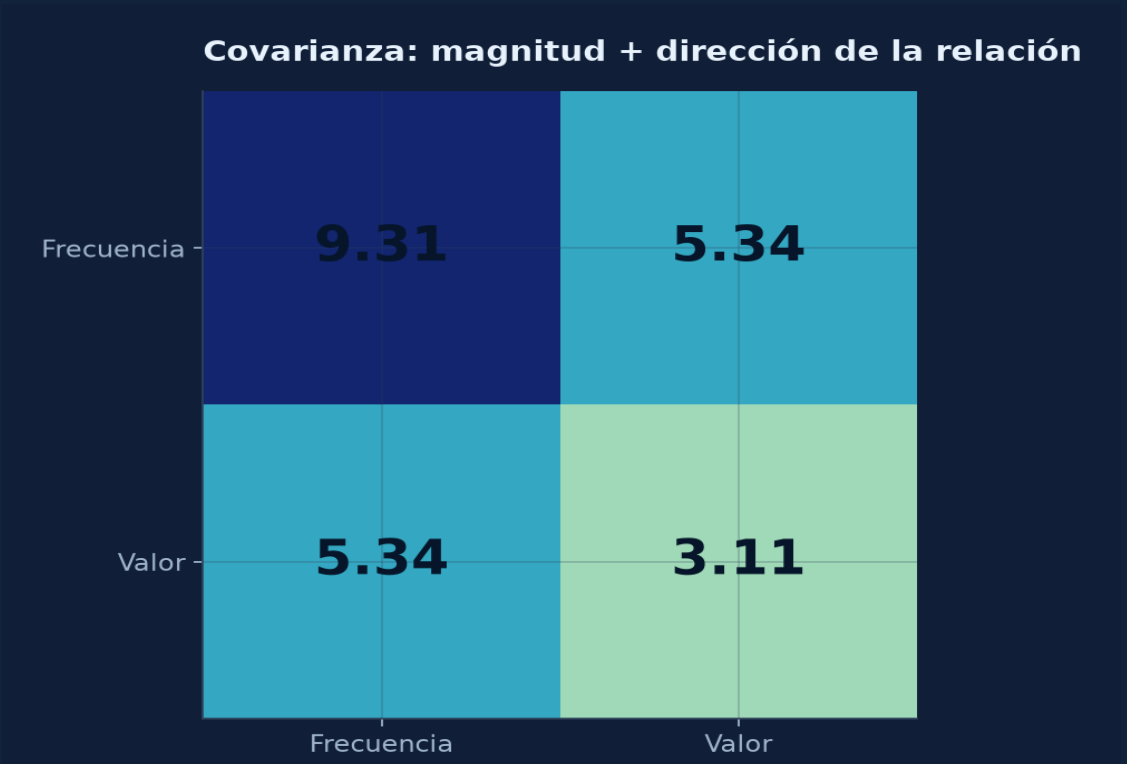
ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **Elegir la escala a partir del significado de las variables, no por automatismo.**

La matriz que describe dispersión y dependencia

IDEAS CLAVE

- La diagonal de Σ contiene las varianzas de cada variable.
- Fuera de la diagonal aparecen covarianzas entre pares.
- Σ es simétrica y semidefinida positiva.
- Una covarianza grande puede reflejar unidades, no necesariamente relación fuerte.

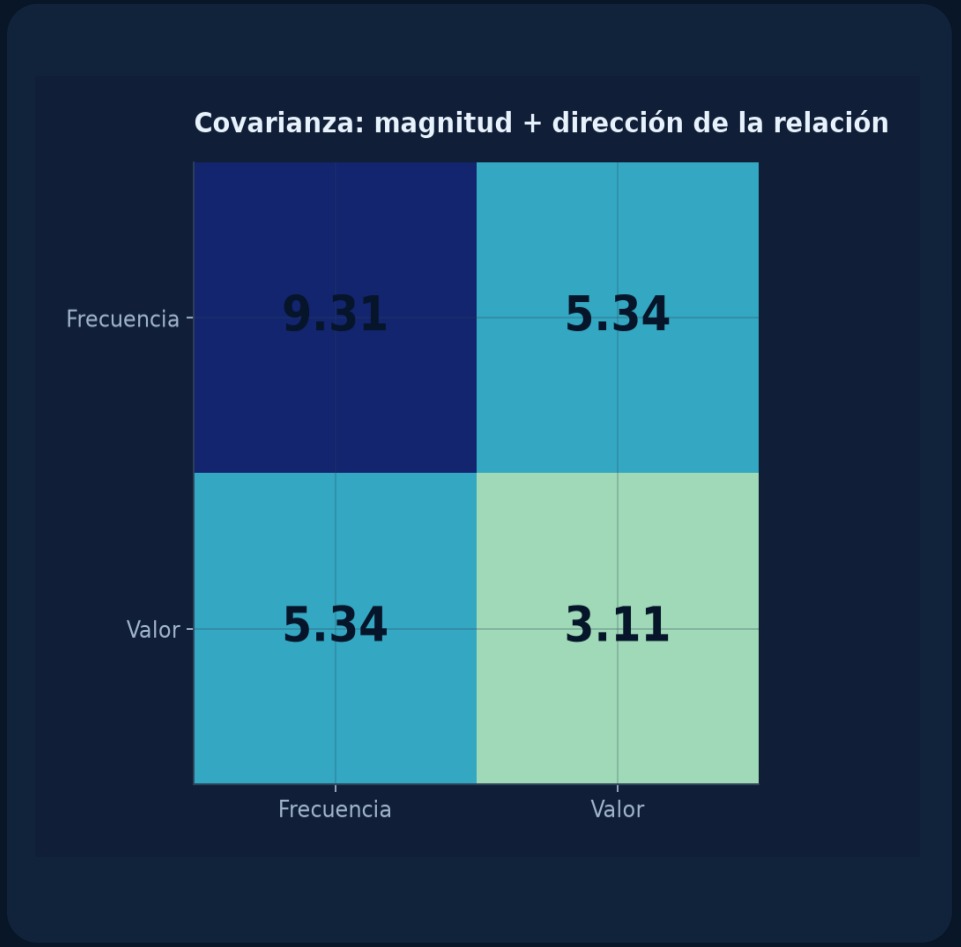


ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **$\Sigma = X_c^T X_c / (n - 1)$ para la covarianza muestral.**

02 · Fundamento matemático

Autovalores, autovectores y descomposición en valores singulares traducen la geometría a cálculo.



De la covarianza a las componentes principales

IDEAS CLAVE

- Cada autovector es una dirección de variación de la muestra.
- Cada autovalor mide cuánta varianza existe en esa dirección.
- Ordenar autovalores de mayor a menor determina PC1, PC2, etc.
- Las direcciones son ortogonales cuando la matriz es simétrica.

MODELO MATEMÁTICO

$$\Sigma v_j = \lambda_j v_j$$

- Σ: covarianza de las variables centradas.
- v_j: eje principal de norma 1.
- $\lambda_j \geq 0$: varianza de la componente j.
- $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$.

IDEA QUE DEBE QUEDAR **Los autovectores dan direcciones; los autovalores dan varianzas.**

PCA resuelve un problema de optimización

IDEAS CLAVE

- Buscamos una dirección unitaria v para proyectar cada fila de Xc.
- La varianza proyectada es $v^T\Sigma v$.
- La solución de máxima varianza es el autovector principal de Σ.
- Para más componentes, imponemos ortogonalidad entre ejes.

MODELO MATEMÁTICO

$$\max_v \; v^T\Sigma v \quad \text{sujeto a} \quad v^Tv = 1$$

- La solución es v_1, autovector de λ_1.
- PC1 captura λ_1 de varianza muestral.
- PC2 maximiza con $v_2 \perp v_1$.

IDEA QUE DEBE QUEDAR **Una componente es una dirección óptima según un objetivo muy específico.**

SVD: la forma numéricamente robusta de PCA

IDEAS CLAVE

- Descomponemos $Xc = U\,S\,V^T$, sin necesidad de construir Σ explícitamente.
- Las filas de V^T aportan direcciones principales.
- Los scores aparecen como U S; las varianzas como $s_j^2/(n-1)$.
- En práctica, librerías usan variantes de SVD por eficiencia y precisión.

MODELO MATEMÁTICO

$$Xc = U\,S\,V^T$$

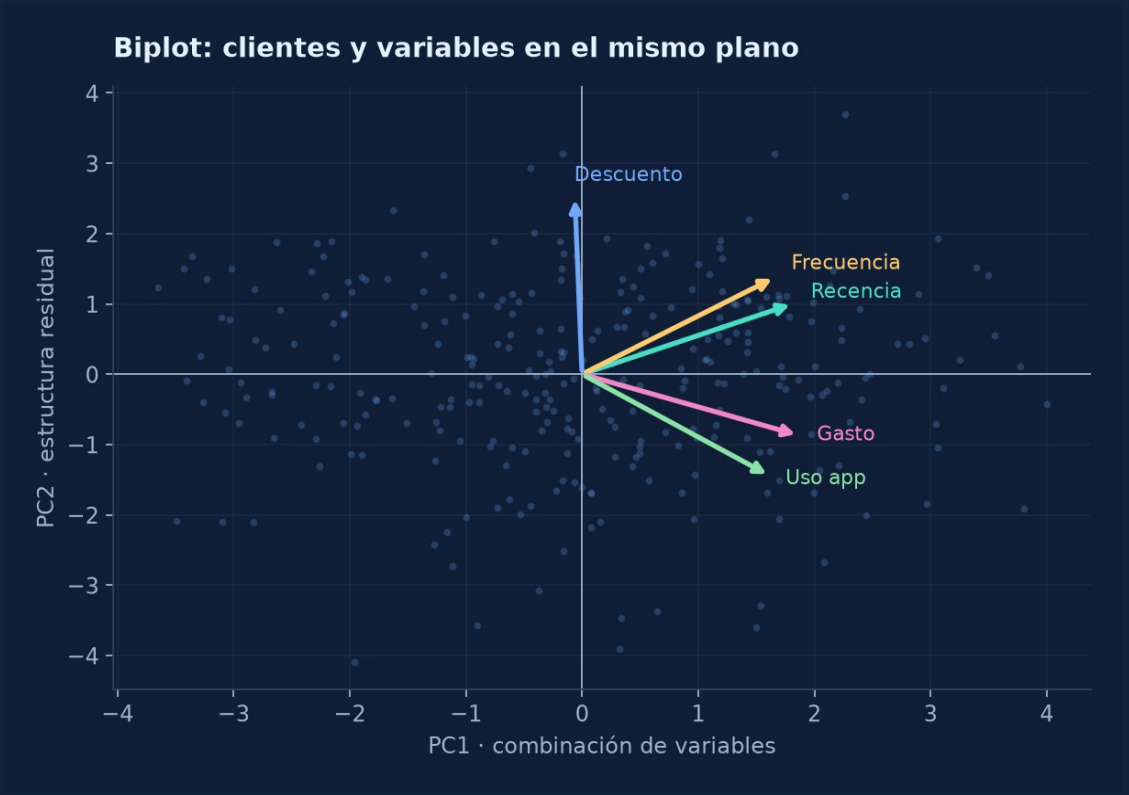
- Scores: $T = XcV = U\,S$.
- Loadings: columnas de V.
- Varianzas: $\lambda_j = s_j^2 / (n-1)$.
- Para $n \ll p$ pueden convenir métodos truncados.

IDEA QUE DEBE QUEDAR **SVD conecta la explicación geométrica con la implementación reproducible.**

Scores y loadings no son lo mismo

IDEAS CLAVE

- Score: coordenada de cada observación en el nuevo sistema de ejes.
- Loading: peso de la variable original en una componente.
- Un valor alto en PC1 no implica automáticamente un “buen” cliente.
- El signo de toda componente puede invertirse sin cambiar la solución.



ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **Interpretar cargas requiere contexto, escala y convención de signos.**

Ejemplo numérico pequeño para resolver en clase

IDEAS CLAVE

- Cuatro observaciones: (1,2), (2,3), (3,5), (4,6).
- Las medias son $\bar{x} = 2,5$ e $\bar{y} = 4$; se resta cada media.
- La covarianza muestral resulta aproximadamente [[1,67; 2,33],[2,33; 3,33]].
- El primer autovalor concentra alrededor de 99,6 % de la variabilidad.

MODELO MATEMÁTICO

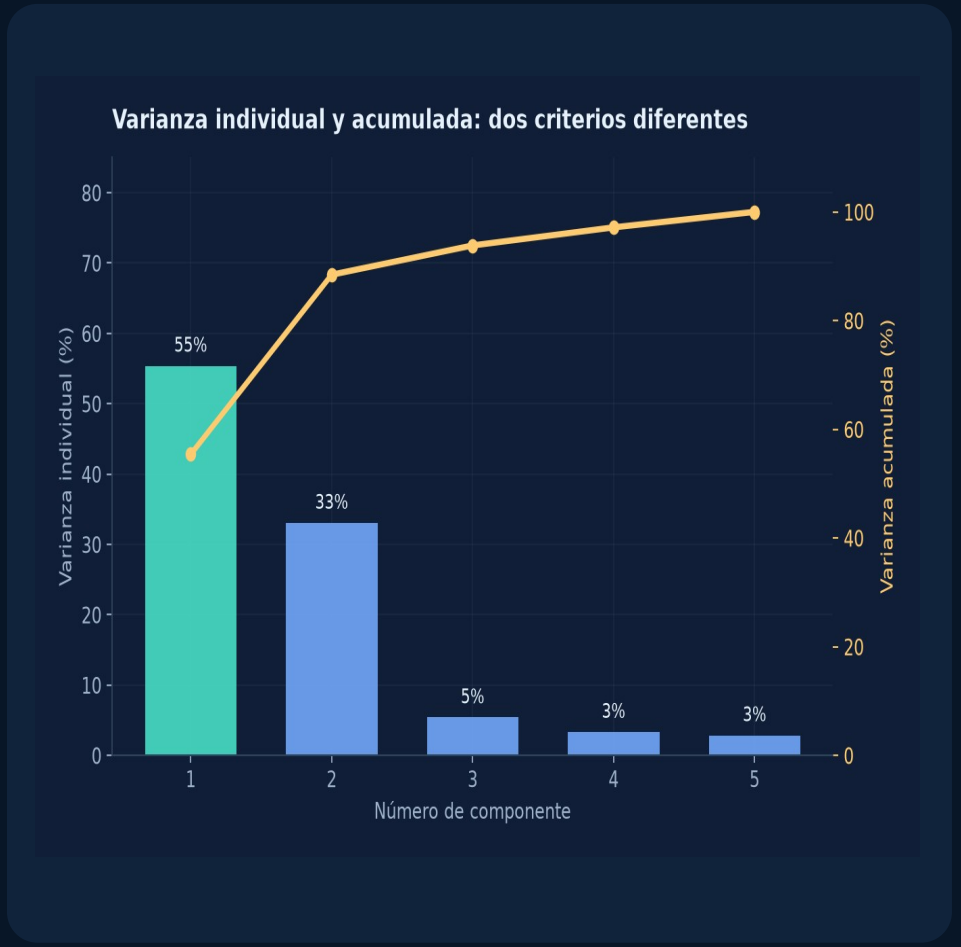
$$\Sigma \approx [1,67 \;\; 2,33 \; ; \; 2,33 \;\; 3,33]$$

- Calcular media de cada columna.
- Construir Xc y $Xc^{T}Xc$ / 3.
- Ordenar autovalores y normalizar vectores.
- Proyectar cada observación sobre PC1.

IDEA QUE DEBE QUEDAR **Un ejemplo mínimo vincula datos, matriz, ejes y proyección.**

03 · ¿Cuántas componentes conservar?

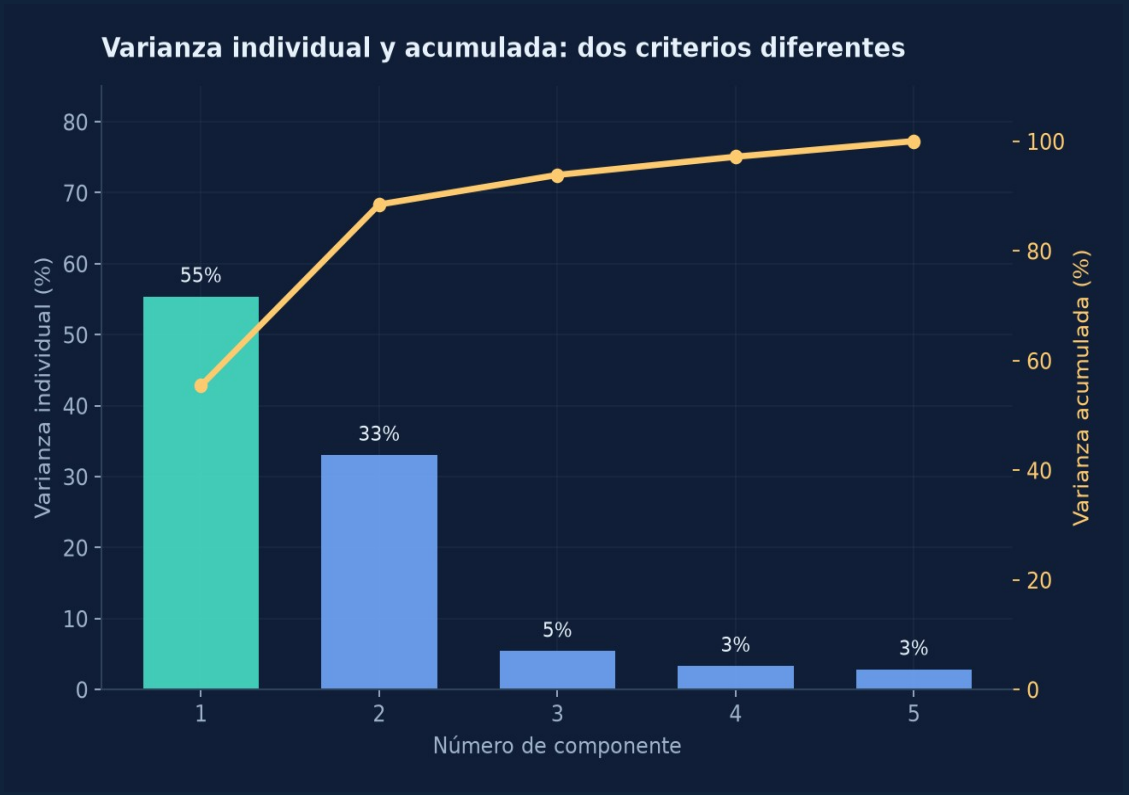
Cada componente adicional retiene variabilidad, pero también aumenta complejidad y puede reintroducir ruido.



Varianza explicada: individual y acumulada

IDEAS CLAVE

- Proporción individual: $EVR_j = \lambda_j / \Sigma \lambda_i$.
- Varianza acumulada: suma de las proporciones hasta k.
- Scree plot identifica cambios de pendiente; no prueba un “K correcto”.
- Los umbrales 80 % o 95 % son convenciones, no garantías de utilidad.



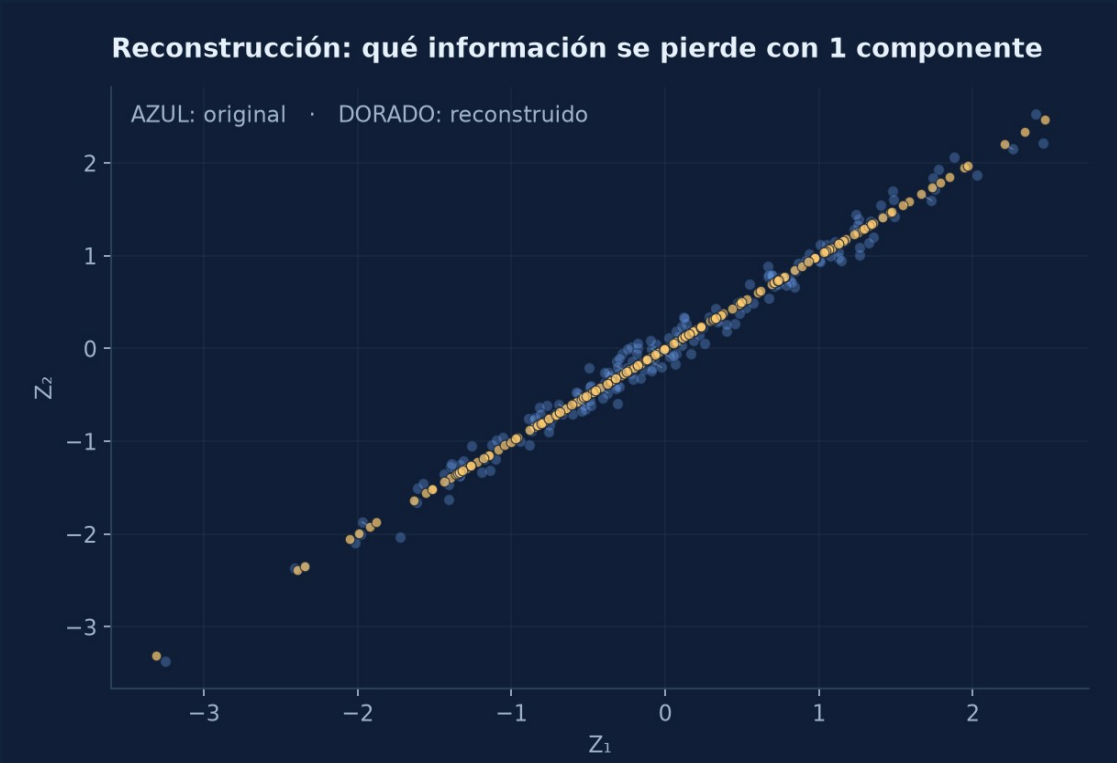
ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **Retener varianza no garantiza preservar segmentos, anomalías ni predicción.**

Reconstruir los datos: el costo medible de comprimir

IDEAS CLAVE

- Con k componentes, $\hat{X}c = T_k V_k^T$ aproxima la matriz centrada.
- La diferencia $Xc - \hat{X}c$ cuantifica pérdida de información cuadrática.
- PCA de rango k minimiza el error de reconstrucción de Frobenius.
- La suma de autovalores omitidos mide la varianza descartada.



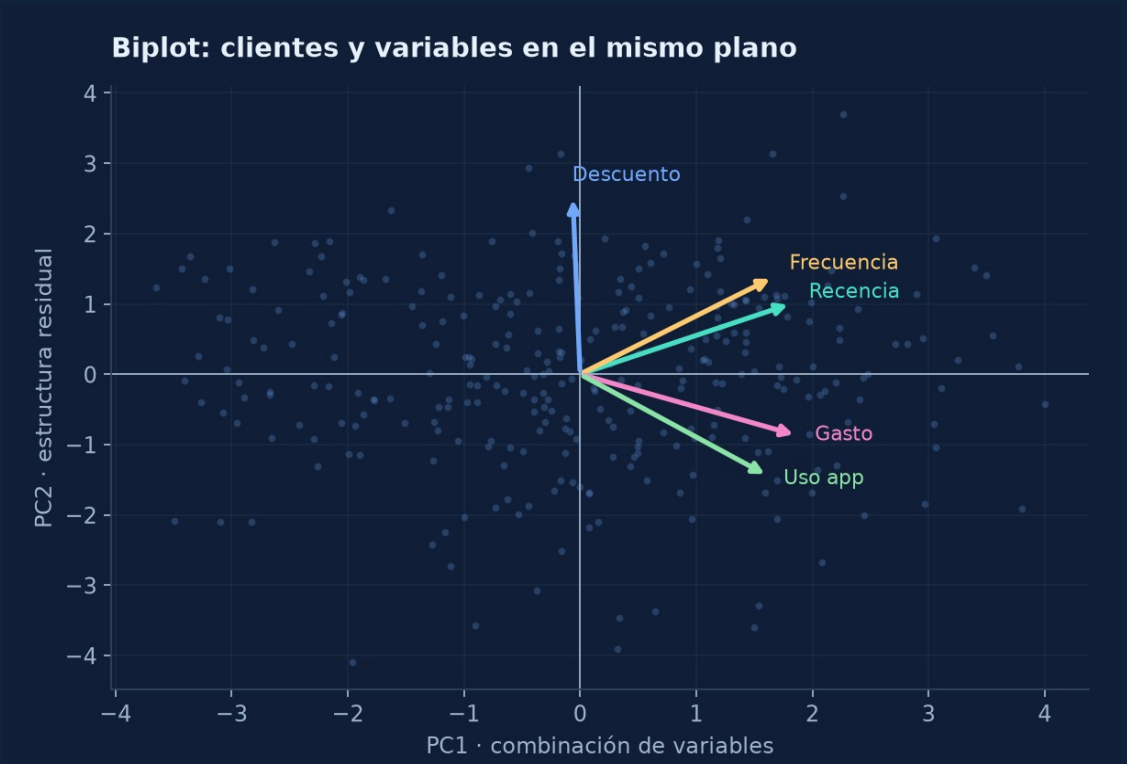
ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **Compresión útil = pérdida aceptable para una decisión concreta.**

Biplot: observaciones y relaciones entre variables

IDEAS CLAVE

- Puntos cercanos en PC1–PC2 tienen scores similares, no necesariamente distancia original idéntica.
- Las flechas muestran la relación de cada variable con el plano proyectado.
- El ángulo entre vectores sugiere asociaciones si la representación es adecuada.
- Las longitudes deben leerse según la escala elegida para el biplot.



ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **El biplot es una aproximación de dos dimensiones, no un mapa exhaustivo.**

PCA frente a seleccionar variables

IDEAS CLAVE

- Selección: conservar variables originales; maximiza interpretabilidad.
- PCA: generar factores lineales; reduce redundancia entre medidas.
- Variables con baja varianza pueden ser útiles si predicen un evento raro.
- PCA no selecciona atributos por relevancia para una etiqueta futura.

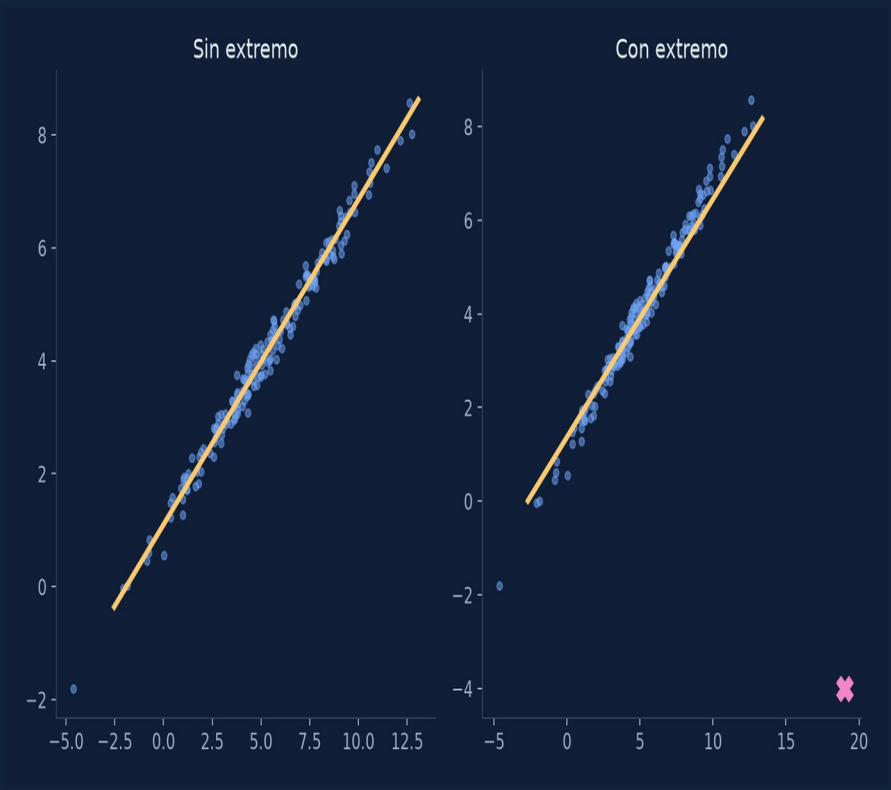
COMPARACIÓN

SELECCIÓN	Retiene columnas originales.
PCA	Combina columnas en ejes nuevos.
PREDICCIÓN	Validar ambos caminos con métricas externas.
RIESGO	Confundir varianza con información útil.

IDEA QUE DEBE QUEDAR **Conservar interpretabilidad puede ser más valioso que una compresión extrema.**

04 · Riesgos y decisiones

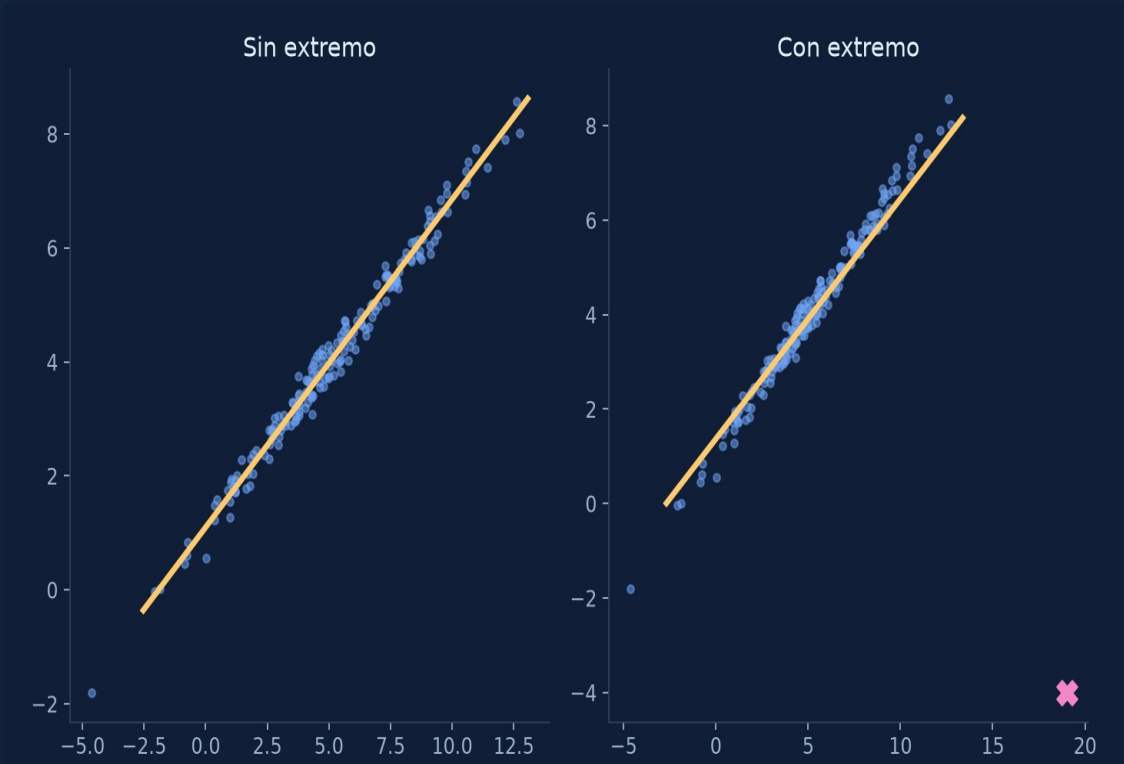
El procedimiento puede ser matemáticamente correcto y analíticamente poco útil si se ignora el contexto.



Sensibilidad a extremos, ruido y escalas

IDEAS CLAVE

- Una sola observación muy extrema puede rotar PC1 al cambiar Σ.
- Centrado y escalado clásicos también reaccionan a outliers.
- Comparar con y sin extremos; evaluar alternativas robustas.
- No excluir clientes excepcionales sin una justificación de negocio.



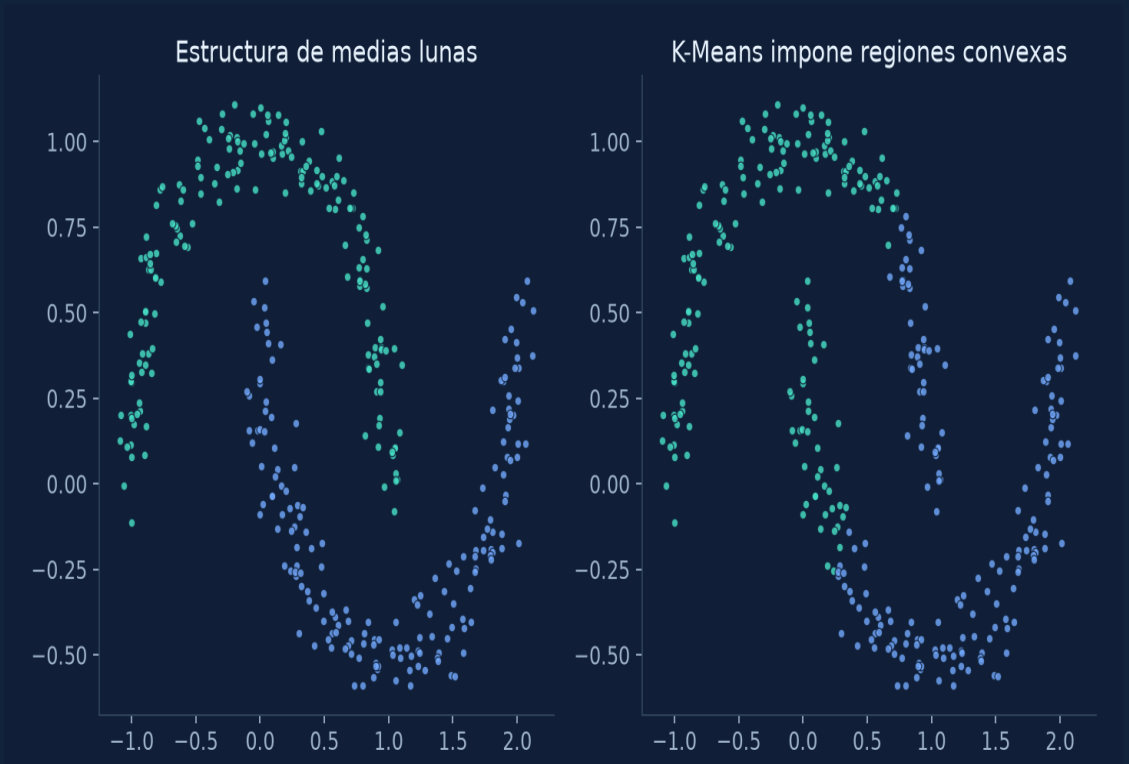
ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **El análisis de sensibilidad es obligatorio cuando la estructura depende de pocos casos.**

PCA es lineal; algunas formas no caben en un plano lineal

IDEAS CLAVE

- PCA busca ejes rectos y ortogonales, no variedades curvas.
- Los métodos no lineales pueden representar estructuras curvas.
- t-SNE y UMAP son útiles para exploración, pero las distancias pueden distorsionarse.
- Nunca inferir clusters fiables solo por manchas en una visualización.



ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **Un gráfico atractivo en 2D no demuestra que existan segmentos reales.**

PCA dentro del pipeline: evitar fuga de información

IDEAS CLAVE

- Separar train y test antes de ajustar imputación, escalado y PCA.
- Ajustar medias, desvíos y componentes únicamente en train.
- Aplicar transformaciones aprendidas sobre validación y test.
- Si se evalúan clusters, medir estabilidad sin usar datos futuros.

PIPELINE CORRECTO

1. Dividir por cliente o período
2. Imputar / escalar con entrenamiento
3. Ajustar PCA solamente en train
4. Proyectar train y test
5. **Validar estabilidad y efecto en decisión**

IDEA QUE DEBE QUEDAR **Todo preprocesamiento aprendido del dataset pertenece dentro del pipeline.**

¿PCA antes de K-Means? Beneficios y costos

IDEAS CLAVE

- Puede atenuar ruido y reducir el costo del cálculo de distancias.
- Puede eliminar direcciones de baja varianza que sí separan grupos.
- Se compara clustering en variables originales escaladas versus PCA(k).
- La comparación requiere criterios de estabilidad e interpretación externa.

COMPARACIÓN

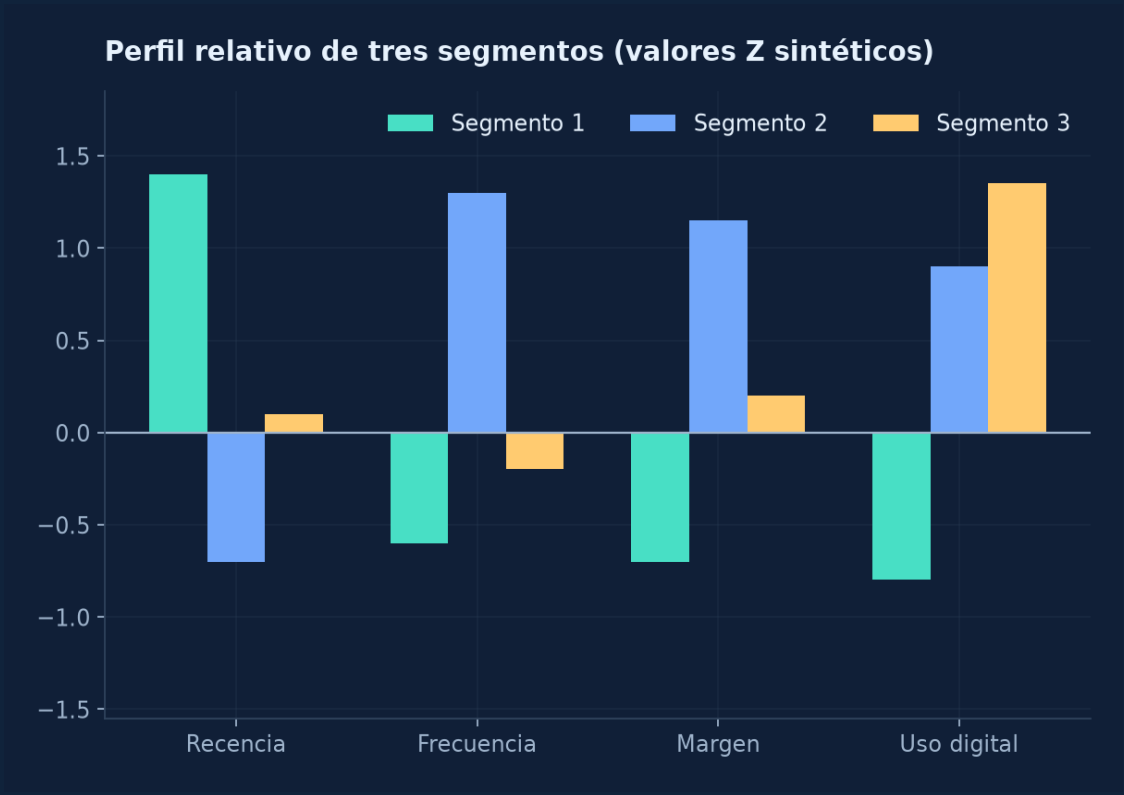
ORIGINAL	Métricas transparentes y accionables.
PCA(k)	Compresión, posible reducción de ruido.
VALIDAR	Silhouette, estabilidad, perfiles, utilidad.
DECIDIR	Preferir solución reproducible e interpretable.

IDEA QUE DEBE QUEDAR **PCA y clustering son etapas diferentes; no siempre conviene combinarlas.**

De la gráfica a una decisión concreta

IDEAS CLAVE

- PC1 puede capturar intensidad de relación comercial; PC2, sensibilidad promocional.
- Esas etiquetas solo valen si los loadings respaldan la interpretación.
- Construir perfiles con métricas originales, no únicamente con componentes.
- Proponer una acción medible: retención, activación o costo de contacto.



ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **Toda recomendación debe recuperar las variables originales y su impacto esperado.**

Laboratorios vinculados a cada idea

IDEAS CLAVE

- S01: rotación de PC1–PC2 y nube con correlación ajustable.
- S02: centrar, estandarizar y modificar pesos de variables.
- S03: scree plot, varianza acumulada y error de reconstrucción.
- S04: outliers, biplot y comparación PCA → K-Means.

SECUENCIA DE ANÁLISIS

S01 · Geometría en tiempo real

↓

S02 · Escala / covarianza / correlación

↓

S03 · Compresión / pérdida

↓

S04 · Sensibilidad / segmentación

IDEA QUE DEBE QUEDAR **Cada laboratorio cambiará una hipótesis y mostrará evidencia observable.**

Ejercicio de transferencia y evaluación

IDEAS CLAVE

- Dos variables están muy correlacionadas, pero una mide montos 1.000 veces mayores.
- Se obtiene 92 % de varianza con dos componentes de cinco.
- Un outlier cambia PC1 y la silhouette mejora después de comprimir.
- Explicar qué decisiones sí se pueden justificar y qué evidencia falta.

PREGUNTAS DE DECISIÓN

1. ¿Qué escalado usarías y por qué?
2. ¿Retendrías dos componentes?
3. ¿Qué experimento mide sensibilidad?
4. ¿Qué indicador externo probaría utilidad?

IDEA QUE DEBE QUEDAR **La evidencia de varianza es necesaria, pero no suficiente.**

Mapa de errores frecuentes

IDEAS CLAVE

- “PC1 es la variable más importante” → es una combinación, no una columna.
- “95 % de varianza = 95 % de precisión” → son magnitudes diferentes.
- “PCA identifica clusters” → PCA proyecta; no asigna grupos.
- “La misma proyección sirve para siempre” → debe vigilarse el drift.

COMPARACIÓN

NO	Varianza = capacidad predictiva.
NO	Componente = variable original.
NO	Biplot = distancias fieles para todo.
SÍ	Validación, sensibilidad y contexto.

IDEA QUE DEBE QUEDAR **Las interpretaciones fuertes necesitan pruebas externas.**

Bibliografía esencial y cierre de PCA

IDEAS CLAVE

- Jolliffe, I. T. & Cadima, J. (2016). Principal component analysis: a review.
- Hastie, Tibshirani & Friedman (2009). The Elements of Statistical Learning, cap. 14.
- James et al. (2023). An Introduction to Statistical Learning with Applications in Python.
- scikit-learn: User Guide → Decomposition / PCA; documentación de Pipeline.

SECUENCIA DE ANÁLISIS

1. Comprender → proyectar
2. Cuantificar → reconstruir
3. Validar → interpretar
4. **Decidir → monitorear**

IDEA QUE DEBE QUEDAR **En la próxima presentación convertimos la geometría en una asignación de grupos.**