Universidad de SanAndrés

Clase 10:
Reducción de dimensionalidad (PCA)

Analítica de Datos

Maestría en Ciencias del Comportamiento

Primavera 2026

Medimos **ocho cosas** de cada estudiante.

¿Se pueden resumir en una sola?

Datos reales de **4.424 estudiantes** de educación superior: notas, materias inscriptas y aprobadas en dos semestres, nota de admisión y edad de ingreso.

Vamos a ver que la respuesta es **sí**, y que ese único número termina diciendo quién abandona.

De qué vamos a hablar hoy

Universidad de SanAndrés

Objetivo

Entender el **análisis de componentes principales (PCA)**: qué hace, cómo se lee, cuánta información conserva y dónde conviene usarlo en Ciencias del Comportamiento.

1. **Por qué reducir.** Muchas variables correlacionadas esconden pocas dimensiones.
2. **Qué es una componente.** De la intuición en 2D al caso de p variables.
3. **Leer el resultado.** El biplot y cómo se interpretan sus flechas.
4. **Cuánta información conservamos.** Reconstrucción, varianza explicada y scree plot.
5. **Decisiones y alcance.** Estandarizar, métodos no lineales y casos aplicados.
6. **Supuestos y límites.** Qué asume PCA y cuándo no corresponde.

Para la práctica vamos a **Colab**.

BLOQUE 1 DE 6

Por qué reducir

Cuando hay demasiadas variables para mirarlas.

Aprendizaje no supervisado: no hay qué predecir

- Hasta ahora tuvimos siempre una variable respuesta Y. Acá **no la hay**: solo un conjunto de features $X_1, \ldots, X_p$.
- El objetivo cambia: ya no es predecir, es descubrir estructura en los datos.
- Por eso el análisis no supervisado casi siempre es parte de un **análisis exploratorio**.
- Y por eso mismo es más **subjetivo**: no hay una respuesta correcta contra la cual validar, ni test set que nos diga si acertamos.

ISLP, Sección 12.1.

Agrupar clientes que nadie etiquetó

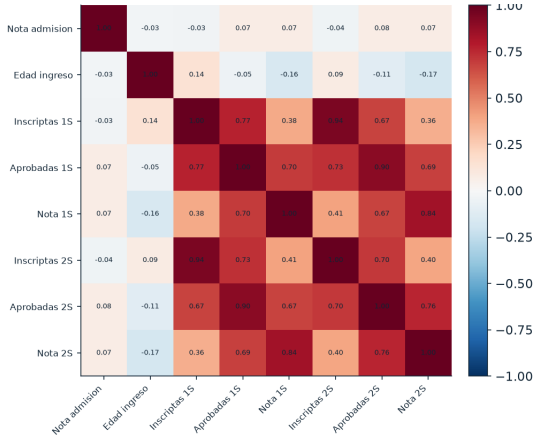
Universidad de SanAndrés

Un caso cotidiano

Un sitio de compras online quiere recomendarle a cada persona lo que le interesa. Agrupa a sus clientes según su **historial de navegación y compra**, sin que nadie haya dicho de antemano a qué grupo pertenece cada uno.

- Nadie definió los grupos: **emergen de los datos**.
- Un buscador hace lo mismo con historiales de clics para decidir qué resultados mostrar.
- PCA ataca una versión de ese problema: cuando hay **muchas variables correlacionadas**, resumirlas en unas pocas que conserven casi toda la variabilidad.

Varias variables dicen casi lo mismo



Correlaciones del bloque académico. Las materias **aprobadas en 1er y 2do semestre** correlacionan **0,90**: quien aprueba mucho en un semestre aprueba mucho en el otro. Hay redundancia para explotar.

Con 8 variables ya hay 28 gráficos posibles

- Para mirar las relaciones de a pares hacen falta $p(p-1)/2$ diagramas de dispersión.
- Con nuestras **8 variables** son **28** gráficos. Con las **24** columnas numéricas del dataset, 276.
- Ninguno muestra la foto completa: cada uno contiene solo una fracción de la información.

La pregunta que abre PCA

¿Existe una representación de **pocas dimensiones** que capture la mayor parte posible de la variabilidad de los datos?

De 4.500 adjetivos a cinco factores

- En 1936, Allport y Odbert listaron **4.500 términos** que describen diferencias de personalidad.
- Con análisis factorial (el pariente cercano de PCA), esos términos se redujeron primero a 16 y finalmente a cinco: el modelo **Big Five** (OCEAN).
- De Raad y Barelds (2008) aplicaron PCA a **2.365 ítems** del léxico holandés y recuperaron el Big Five más tres factores adicionales.

El principio, en una línea

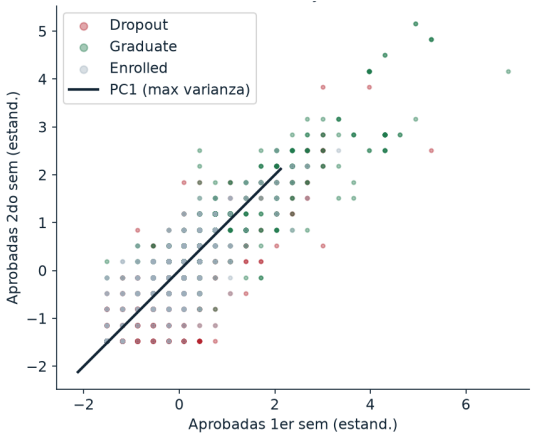
Con p enorme, unas pocas **dimensiones compuestas** pueden resumir una variedad inmensa. Es exactamente lo que vamos a hacer hoy, en menor escala.

BLOQUE 2 DE 6

Qué es una componente principal

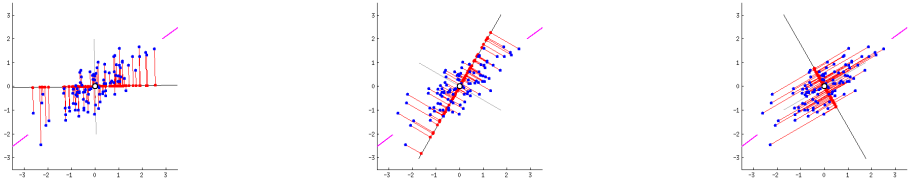
De la intuición en dos dimensiones al caso general.

En 2D: la dirección de máxima varianza



Materias aprobadas en 1er y 2do semestre (estandarizadas). La recta negra es la **primera componente**: la dirección a lo largo de la cual los puntos varían más. Ella sola concentra el 95,2 % de la variabilidad de las dos variables.

...y a la vez la recta más cercana a los puntos



La misma recta girando. Los **segmentos rojos** son los errores de proyección. Cuando la recta llega a la primera componente, esos errores alcanzan su **mínimo** y, en ese mismo instante, las proyecciones quedan lo más dispersas posible.

Dos lecturas del mismo problema

Máxima varianza

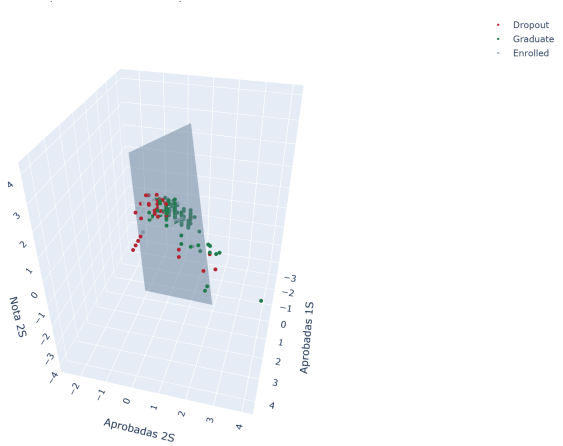
La dirección a lo largo de la cual los datos proyectados tienen la **mayor dispersión** posible.

Mínima distancia

La recta que **minimiza** la suma de distancias al cuadrado entre cada punto y su proyección.

- Las dos definiciones llevan a la misma solución matemática.
- No es casualidad: la varianza total es fija, así que todo lo que la proyección captura es exactamente lo que el error deja de perder.
- Volvemos sobre esto en el Bloque 4, ya con números.

En 3D: dos componentes definen un plano



Tres variables académicas. Las **dos primeras componentes** generan el plano que mejor se apoya sobre la nube. Con *p* variables la idea es la misma: un subespacio de pocas dimensiones, lo más cerca posible de los datos.

Una componente es una combinación lineal

Cada componente es una **variable nueva**, construida ponderando las originales:

$$Z_m = \phi_{1m} X_1 + \phi_{2m} X_2 + \cdots + \phi_{pm} X_p$$

donde:

- $X_1, \ldots, X_p$ son las variables originales, ya centradas en media cero;
- los ***loadings*** ϕ_{jm} son los pesos: cuánto aporta cada variable a la componente m;
- los ***scores*** Z_m son los valores resultantes: la posición de cada estudiante en esa nueva variable.

Los loadings describen **variables**; los scores describen **observaciones**. Es la distinción que hace legible un biplot.

Por qué los pesos se normalizan

- Se exige que la suma de los cuadrados de los pesos sea **igual a 1**.
- Sin esa restricción el problema no tendría solución: agrandando los pesos, la varianza crecería sin límite.
- Normalizar obliga a buscar una **dirección** en el espacio de las variables, no una magnitud arbitraria.

Cómo se calcula en la práctica

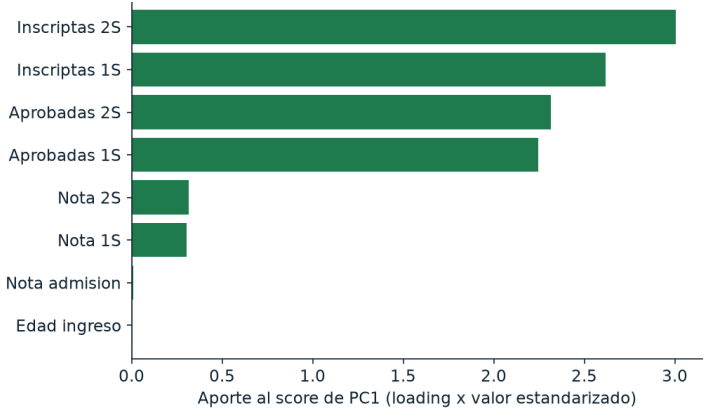
Los pesos salen de la **descomposición en autovectores y autovalores** de la matriz de covarianzas (o de correlaciones, si estandarizamos). Cada componente es un **autovector**; la varianza que explica es su **autovalor**. No hace falta resolverlo a mano: lo hace la función de PCA.

Cada componente es ortogonal a las anteriores

- La segunda componente se busca igual que la primera, maximizando varianza, pero con una condición extra: ser ortogonal (no correlacionada) con la primera.
- Lo mismo para la tercera respecto de las dos anteriores, y así sucesivamente.
- Cada una captura **información nueva**, no una repetición de la anterior.
- En total hay como máximo $\text{mín}(n-1,\ p)$ componentes. Con nuestras 8 variables, **8 componentes**.

Usar *todas* las componentes no pierde nada: es solo **rotar** el sistema de coordenadas. La compresión aparece cuando nos quedamos con algunas.

El score de un estudiante, término a término



Aporte de cada variable al score en PC1 de un estudiante concreto: su **valor estandarizado por el loading** de esa variable. Las cuatro variables de materias explican casi todo el score; la nota de admisión y la edad, prácticamente nada.

Las componentes están definidas salvo el signo

Un detalle que confunde

Dos programas pueden devolver la **misma componente con el signo invertido**, y ambos son correctos.

- Un vector de pesos define una **dirección**, y una dirección se extiende hacia los dos lados.
- Dar vuelta el signo de los loadings y el de los scores a la vez deja **todo igual**.
- Lo que importa es la **geometría relativa**: qué variables van juntas y cuáles se oponen.
- En la práctica conviene **orientar** cada componente para que se lea bien, como hicimos al poner el éxito académico hacia la derecha.

A la notebook

Calcular un score **a mano** y comprobar que coincide con el del algoritmo.

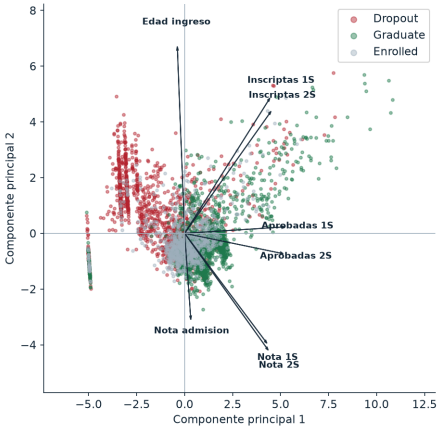
También: rotar la recta y ver cómo cambia la varianza proyectada.

BLOQUE 3 DE 6

Leer el resultado: el biplot

Observaciones y variables, en un mismo gráfico.

El biplot: scores y loadings a la vez



Dos gráficos superpuestos: los **puntos** son estudiantes (scores) y las **flechas** son variables (loadings), cada uno con su escala. Las dos primeras componentes resumen el 71,5 % de la variabilidad de las ocho variables.

PC1 resultó un eje de riesgo académico

El hallazgo

PCA **nunca vio** la etiqueta de deserción. Aun así, la primera componente separa a quienes abandonan de quienes se gradúan.

- Score medio en PC1: **+0,96** entre quienes se gradúan, **–1,43** entre quienes desertan.
- PC1 sola explica el 54,2 % de la variabilidad de las ocho variables.
- Las variables de **materias** son las que definen ese eje; la nota de admisión casi no participa.

Es el mismo tipo de lectura con la que el libro interpreta que la PC1 de `USArrests` mide criminalidad general.

Cómo se leen las flechas: dirección

- La **dirección** de una flecha dice qué mezcla de PC1 y PC2 representa a esa variable.
- Flecha casi **horizontal**: la variable carga sobre la primera componente. Casi **vertical**: sobre la segunda.
- Es lo que permite ponerle nombre sustantivo a cada eje.

En nuestro biplot

Las seis variables de cursada apuntan hacia la **derecha**, con pesos del mismo signo y magnitud parecida: por eso el eje horizontal se lee como **rendimiento académico**. La edad de ingreso y la nota de admisión viven sobre el eje vertical.

Ángulo y longitud: lo que agregan las flechas

Universidad de SanAndrés

Ángulo entre flechas

Aproxima la **correlación** entre las variables.

- Misma dirección: muy correlacionadas.
- Perpendiculares: sin correlación.
- Opuestas: correlación negativa.

Longitud de la flecha

Indica qué tan bien **representada** queda esa variable en el plano.

- Larga: el plano la captura bien.
- Corta: vive en las componentes que no estamos mirando.

- Las flechas de **inscriptas 1S y 2S** son casi paralelas, y su correlación real es alta: el gráfico no miente.
- La **nota de admisión** tiene la flecha más corta: sobre ella el biplot no autoriza ninguna lectura.

Nombrar las componentes: un caso de consumo

Lácteos locales (Apulia, Italia, 2022)

Un experimento de elección online con **543 encuestados** sobre productos lácteos locales: calidad, sustentabilidad y disponibilidad.

- Aplicaron PCA a las respuestas y quedaron **cuatro componentes**, que nombraron según qué ítems cargaban fuerte en cada una: sensibilidad a la calidad, “lo local es mejor”, “lo local es sustentable” y demanda de mayor disponibilidad.
- Después relacionaron esas dimensiones con variables **sociodemográficas**.

El flujo típico en Ciencias del Comportamiento: cuestionario con muchos ítems → PCA → nombrar cada componente → relacionarla con otras variables.

BLOQUE 4 DE 6

Cuánta información conservamos

Reconstruir, medir el error y decidir cuántas componentes.

Reconstruir un dato desde sus componentes

Si los scores se armaron combinando las variables, se puede hacer el **camino inverso**:

$$x_{ij} \approx \sum_{m=1}^{M} z_{im}\,\phi_{jm}$$

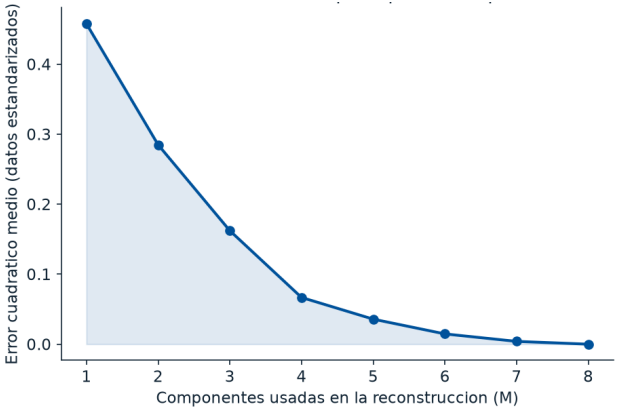
donde:

- x_{ij} es el dato original: la variable j del estudiante i;
- M es cuántas componentes usamos. Si $M = p$, la reconstrucción es **exacta**;
- z_{im} es el *score* del estudiante i en la componente m;
- ϕ_{jm} es el *loading* de la variable j en esa componente.

Por qué importa

Con M componentes, esta es la mejor aproximación posible en M dimensiones. Ninguna otra combinación lineal reconstruye mejor. Eso es lo que hace de PCA una técnica de **compresión**, no un resumen más.

El error cae rápido: pocas componentes bastan



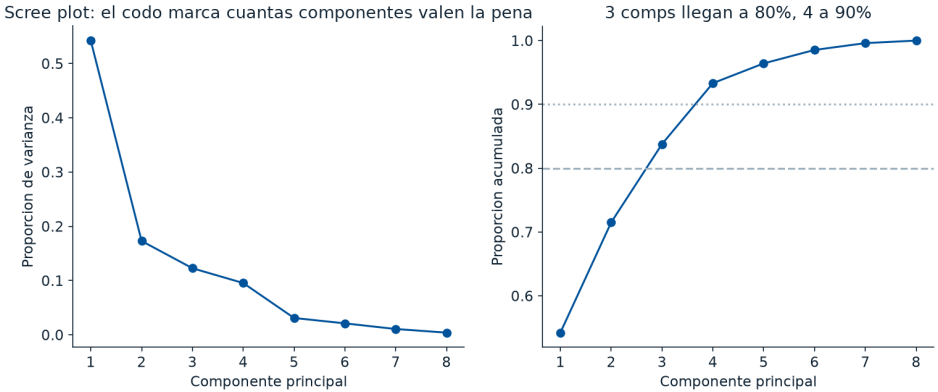
Error cuadrático medio al reconstruir las ocho variables de los 4.424 estudiantes. Con **1** componente es 0,46; con **4** baja a 0,07; con las **8** es exactamente cero: usar todas es solo rotar los ejes.

La varianza total se reparte entre las componentes

- Con variables **estandarizadas**, cada una aporta 1 unidad de varianza: la varianza total es exactamente p (acá, **8**).
- Cada componente explica una porción de ese total: su **autovalor**.
- La **PVE** de una componente es esa porción sobre el total. Sumando todas, siempre da 100 %.

Componente	1	2	3	4	5	6	7	8
PVE (%)	54,2	17,3	12,3	9,5	3,1	2,1	1,1	0,4
Acumulada (%)	54,2	71,5	83,8	93,3	96,4	98,5	99,6	100

El scree plot y el criterio del codo



Se busca el **codo**: el punto donde cada componente adicional deja de aportar. Acá, 3 componentes llegan al 80 % de la varianza y 4 superan el 90 %.

La PVE es el R^2 de la aproximación

Un gancho con lo que ya saben

La proporción de varianza explicada por las primeras M componentes se interpreta **exactamente** como el R^2 de la regresión lineal: qué fracción de la variabilidad total se logra explicar.

- Con 2 componentes explicamos el **71,5 %** de la variabilidad de las ocho variables.
- Por eso el biplot en dos dimensiones es una representación fiel del dataset completo.
- Y por eso retener componentes es quedarse con la señal y descartar, sobre todo, ruido.

Máxima varianza y mínimo error son lo mismo

- La varianza total de los datos se reparte en **dos partes**: lo que capturan las primeras M componentes, más el **error** de reconstruirlos con esas M.
- Como la varianza total es un número **fijo**, subir una es bajar la otra.

La equivalencia, cerrada

Maximizar la varianza explicada por las primeras M componentes es exactamente lo mismo que **minimizar** el error de reconstrucción.

Es la razón por la que las dos definiciones del Bloque 2, máxima varianza y mínima distancia, describen el mismo objeto.

Cuántas retener: no hay respuesta única

- No existe un criterio objetivo y universal. En análisis no supervisado, la decisión es en parte un **juicio**.
- Se mira el **codo** del scree plot, o se fija un umbral de varianza acumulada (80 % o 90 %).
- Regla práctica: si las primeras componentes no muestran nada interesante, las siguientes tampoco suelen hacerlo.

Cuando sí hay criterio objetivo

Si las componentes se usan como **predictores** de un modelo posterior, el número se elige como cualquier hiperparámetro: por **validación cruzada**.

A la notebook

Reconstruir un dato con 1, 2 y todas las componentes, y ver caer el error.

Ejercicio: cuántas componentes hacen falta para llegar al 90 %.

BLOQUE 5 DE 6

Decisiones prácticas y alcance

Estandarizar, métodos no lineales y casos aplicados.

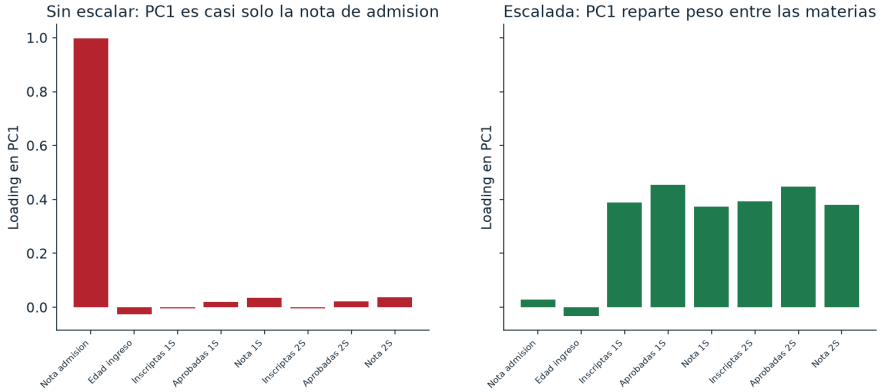
¿Qué pasa si no estandarizamos?

Antes de mirar el resultado

Vamos a correr PCA sobre las ocho variables **sin estandarizar**. ¿Qué variable creen que va a dominar la primera componente, y por qué?

- La **nota de admisión** va de 0 a 200; las materias aprobadas, de 0 a 10.
- Sus varianzas: **210** frente a **9,6**.
- Como PCA busca máxima varianza, sin estandarizar la escala de medición decide el resultado.

Sin escalar, una variable secuestra PC1



Pesos de PC1 sin escalar (izquierda) y escalando (derecha). Sin estandarizar, PC1 es **casi solo la nota de admisión**; estandarizando, reparte peso entre las materias y se vuelve interpretable. Salvo unidad común, se estandariza siempre.

Más allá de PCA: t-SNE y UMAP

PCA es **lineal**: solo captura estructura describible con direcciones rectas. t-SNE y UMAP construyen una vista 2D que preserva sobre todo la **vecindad local**, y por eso los grupos aparecen mucho más separados.

Lo que t-SNE y UMAP no garantizan

Tres cuidados

- **No preservan distancias globales**: la separación entre dos grupos lejanos en el dibujo no es interpretable.
- **Dependen de hiperparámetros**: cambiar la *perplexity* o el número de vecinos cambia el gráfico.
- **Sus ejes no significan nada**: no hay loadings que interpretar.

- No están en el capítulo 12 del libro: son un complemento moderno.
- Regla: usalos para **descubrir** grupos, nunca para **medir** distancias.

PC1 como índice: el Nivel Socioeconómico

Filmer y Pritchett (2001), *Demography*

Para estimar la riqueza del hogar sin datos de gasto, construyeron un **índice** a partir de indicadores de **posesión de bienes** del hogar, usando PCA para derivar los ponderadores.

- Los hogares se ordenan por su valor en **PC1** y se agrupan en **quintiles**, de más pobre a más rico.
- Vyas y Kumaranayake (2006) es la guía metodológica de referencia sobre cómo construirlo e interpretarlo.
- **Punto de discusión:** casi todos estos trabajos usan *solo* PC1. Hay literatura que propone combinar varias componentes, porque PC1 suele explicar una parte modesta de la varianza.

32 medidas de comportamiento, seis dimensiones

Arguello y Crescenzi (2019)

Analizaron con PCA hasta **32 medidas de comportamiento** por sesión (cantidad de búsquedas, clics, tiempos de permanencia, scrolls, movimientos del mouse) en estudios de búsqueda de información.

- Objetivo: entender qué **fenómenos latentes** capturan esas medidas, y cómo influyen en las percepciones posteriores (carga de trabajo, dificultad, *engagement*).
- Usaron la matriz de **correlación** y no la de covarianza, porque las medidas están en escalas muy distintas: es el mismo argumento del escalado que vimos recién.
- Para elegir cuántas componentes: autovalor **mayor a 1**, y que cada componente tenga al menos **dos variables** con pesos altos.

Del componente al modelo

- Aplicaron una **rotación varimax**: redistribuye los pesos para que cada variable cargue fuerte en una sola componente. Facilita nombrarlas, y es práctica estándar aplicada aunque el libro no la desarrolle.
- Así nombraron cada dimensión según qué medidas pesaban en ella, por ejemplo *abandono de búsquedas* o *ritmo de interacción*.
- Después usaron los **scores** de las componentes, en lugar de las 32 medidas originales, como predictores de un modelo de regresión.

El resultado

Seis componentes explicaron el **76 %** de la varianza de las medidas de comportamiento. El circuito completo: comportamiento crudo → dimensiones interpretables → efectos sobre lo que la gente percibe.

BLOQUE 6 DE 6

Supuestos y límites

Qué asume PCA, y cuándo conviene desconfiar.

Qué asume PCA

- PCA se apoya en la matriz de **covarianzas o correlaciones** (de Pearson).
- Eso trae tres supuestos implícitos sobre las variables:
  - que son **continuas**,
  - que las relaciones entre ellas son **lineales**,
  - que las distancias entre valores consecutivos son equiespaciadas.
- El tercero es el que más tensiona en Ciencias del Comportamiento: en una escala de acuerdo, ¿la distancia entre “en desacuerdo” y “neutral” es la misma que entre “neutral” y “de acuerdo”?

La práctica estándar en psicometría

Con escalas de **5 o más puntos** (idealmente 7) y una distribución no muy asimétrica, se aplica PCA directamente, tratando la variable ordinal como si fuera continua.

- El sesgo que introduce suele ser **chico y manejable**.
- Así se analiza la enorme mayoría de los cuestionarios de personalidad, actitudes y clima laboral.
- Es exactamente lo que hicieron los estudios de Big Five que vimos al principio.

Conviene igual **declarar** la decisión al reportar: es un supuesto que se asume, no un detalle técnico invisible.

Qué conviene evitar

El límite claro

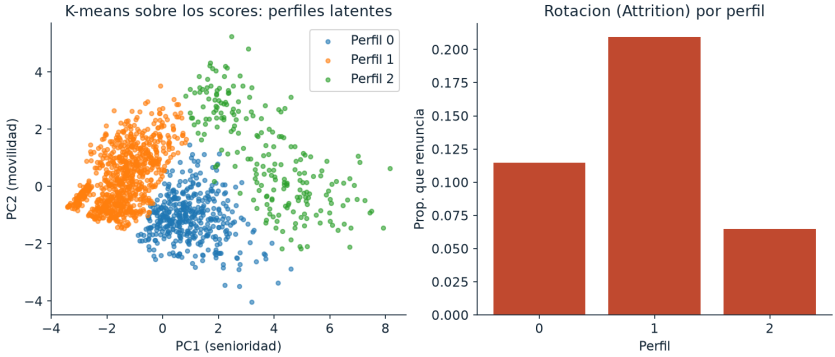
Aplicar PCA con correlaciones de Pearson a variables **nominales**, es decir, sin ningún orden: tipo de contrato, carrera, provincia.

- Si no hay orden, no hay nada que preservar: los números que codifican las categorías son etiquetas arbitrarias, y las correlaciones que se calculan sobre ellas no significan nada.
- El chequeo previo más útil sigue siendo el más simple: **mirar la matriz de correlaciones**. Si es casi diagonal, no hay redundancia que explotar y PCA no va a ayudar.

La práctica

Aplicar PCA **+ clustering** sobre el dataset de rotación de personal que ya conocen, para descubrir perfiles latentes de empleados.

Los perfiles son hipótesis, no verdades



Clusters sobre los primeros scores, y su tasa de rotación. El clustering **no es robusto**: cambia con la estandarización, con cuántas componentes se retienen y con k. Conviene repetirlo con distintas opciones y quedarse con lo que persiste.

Ya podés...

- Explicar por qué muchas variables correlacionadas se resumen en pocas dimensiones.
- Interpretar **loadings** y **scores**, y leer un **biplot** con sus tres criterios.
- Decidir cuándo **estandarizar** y cuántas componentes **retener**.
- Reconocer qué aportan y qué no los métodos no lineales, y qué asume PCA.

El hallazgo de hoy, en una frase

Sin haber visto nunca la etiqueta de deserción, la **primera componente** de ocho variables académicas resultó ser un eje de riesgo que separa a quienes abandonan de quienes se gradúan.

Lo que viene: representar textos en pocas dimensiones, con *embeddings* (Clases 11 y 12).