Componentes principales

Santiago Laplagne

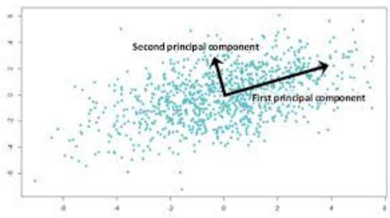
29 de mayo de 2024

1. Introducción

En estas notas desarrollamos los conceptos teóricos de álgebra lineal detrás del análisis de componentes principales.

1.1. Qué es el análisis de comonentes principales (PCA)

- PCA es un método de *aprendizaje no supervisado* para reducir la dimensionalidad de un conjunto de datos.
- Identifica las direcciones (componentes principales) que explican la mayor parte de la variabilidad en los datos.
- Las componentes principales son combinaciones lineales de las variables originales.



1.2. Ejemplo

- Si tenemos un conjunto de datos con 3 variables pero hay dependencia lineal entre las variables (es decir, $ax_1 + bx_2 + cx_3 \approx 0$), al representar a los puntos en el espacio, todos los puntos quedaran ubicados cerca de un plano (el plano $ax_1 + bx_2 + cx_3 = 0$).
- Si proyectamos los puntos sobre ese plano, y consideramos nuevas variables definidas por las coordenadas de las proyecciones en ese plano, estaremos representando las 3 variables originales utilizando solo 2 variables, sin perder mucha información.

2. Varianza, covarianza y correlación

2.1. Varianza.

Ya vimos la fórmula de varianza para una variable $x = (x_1, x_2, \dots, x_n)$:

$$\mathrm{var}(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n},$$

donde $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ es el promedio o la media de los valores de x.
Recordemos la definición de producto interno entre vectores:

$$\langle x, y \rangle = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = x^T y,$$

donde usamos la convención de representar a los vectores como matrices columna.
Usando productos internos podemos escribir

$$\mathrm{var}(x) = \frac{\langle x - \bar{x}, x - \bar{x} \rangle}{n} = \frac{\|x - \bar{x}\|^2}{n}.$$

2.2. Covarianza.

Definimos ahora la covarianza entre dos variables $x = (x_1, x_2, \dots, x_n)$ e $y = (y_1, y_2, \dots, y_n)$ que mide la relación entre ellas:

$$\mathrm{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\langle x - \bar{x}, y - \bar{y} \rangle}{n}$$

Interpretación:

- Si $(y_i - \bar{y})$ es siempre positivo cuando $(x_i - \bar{x})$ es positivo, la covarianza será grande positiva.
- Si uno es siempre positivo cuando el otro es negativo, la covarianza será grande negativa.

2.3. Correlación.

Dividiendo la fórmula de covarianza por los desvíos de cada variable obtenemos la fórmula de correlación:

$$\mathrm{corr}(x, y) = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x)}\sqrt{\mathrm{var}(y)}} = \frac{\mathrm{cov}(x, y)}{\sigma(x)\sigma(y)}$$

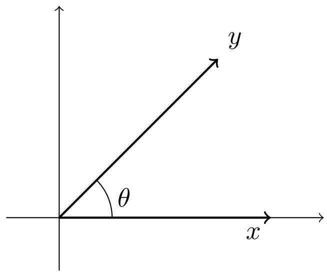
Podemos reescribir esta fórmula:

$$\mathrm{corr}(x, y) = \frac{\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}}{\frac{\|x - \bar{x}\|}{\sqrt{n}} \frac{\|y - \bar{y}\|}{\sqrt{n}}} = \frac{\langle x - \bar{x}, y - \bar{y} \rangle}{\|x - \bar{x}\| \|y - \bar{y}\|},$$

2.4. Ángulos y correlación

Recordemos la fórmula para el ángulo entre dos vectores:

$$\cos(\theta) = \frac{\langle x, y\rangle}{\|x\|\|y\|}$$



La correlación

$$\operatorname{corr}(x, y) = \frac{\langle x - \bar{x}, y - \bar{y}\rangle}{\|x - \bar{x}\|\|y - \bar{y}\|}$$

es el coseno del ángulo entre $x - \bar{x}$ e $y - \bar{y}$.

- Si el ángulo es 0°, la correlación es 1.
- Si el ángulo es 180°, la correlación es -1.
- Si el ángulo es 90°, la correlación es 0.

2.5. Matriz de Covarianza

Consideramos ahora una matriz de datos $X \in \mathbb{R}^{N\times p}$, donde las p columnas representan distintas variables y las N filas, observaciones.

- La matriz de covarianza mide la relación lineal entre dos o más variables.
- Dada una matriz de datos X de N observaciones y p variables, la matriz de covarianza Σ se define como:

  $$\Sigma = \frac{(X - \bar{X})^T(X - \bar{X})}{N} \in \mathbb{R}^{p\times p},$$

  donde la matriz $\bar{X}$ tiene en cada columna el valor medio de la columna correspondiente de X.
- Si llamamos $X^\star$ a la matriz con datos normalizados $X^\star = X - \bar{X}$,

  $$\boxed{\Sigma = \frac{(X^\star)^T X^\star}{N}}$$

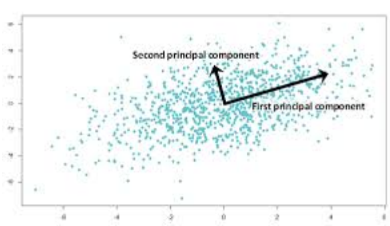
- Las columnas de $X^\star$ son las columnas originales de X normalizadas a media 0.
- Cada elemento Σ_{ij} en la matriz de covarianza representa la covarianza entre las variables x^i y x^j,

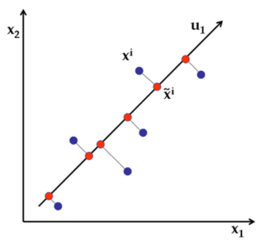
  $$\operatorname{cov}(x^i, x^j) = \frac{\sum_{k=1}^N (x_k^i - \overline{x^i})(x_k^j - \overline{x^j})}{N} = \frac{\langle x^i - \overline{x^i}, x^j - \overline{x^j}\rangle}{N} = \Sigma_{ij}.$$

3. Aspectos geométricos

3.1. Proyección de un punto sobre una recta

Para construir las componentes principales, proyectamos los datos sobre rectas y buscamos la recta para la cual se maximiza la varianza.

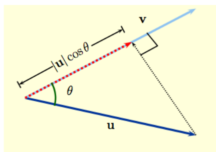




3.2. Longitud de la proyección

La longitud de la proyección de un vector u sobre la recta generada por un vector v es

$$longitud = \|u\| \cos(\theta) = \frac{\langle u, v \rangle}{\|v\|}.$$



Si $\|v\| = 1$ (vector unitario), obtenemos

$$longitud = \langle u, v \rangle.$$

3.3. Proyección de un conjunto de datos

Si tenemos una matriz de datos $X \in \mathbb{R}^{N \times p}$, el vector de proyecciones sobre $v = (v_1, v_2)$ (unitario) será simplemente Xv.

Si X tiene dos columnas,

$$Xv = \begin{pmatrix} x_1^1 & x_1^2 \\ x_2^1 & x_2^2 \\ \vdots & \vdots \\ x_n^1 & x_n^2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} x_1^1 \cdot v_1 + x_1^2 \cdot v_2 \\ x_2^1 \cdot v_1 + x_2^2 \cdot v_2 \\ \vdots \\ x_n^1 \cdot v_1 + x_n^2 \cdot v_2 \end{pmatrix} = \begin{pmatrix} \langle (x_1^1, x_1^2), (v_1, v_2) \rangle \\ \langle (x_2^1, x_2^2), (v_1, v_2) \rangle \\ \vdots \\ \langle (x_n^1, x_n^2), (v_1, v_2) \rangle \end{pmatrix}.$$

Ejemplo. Para proyectar los datos sobre el eje x_1, multiplicamos la matrix X por el vector canónico $(1, 0)$ y obtenemos la primera columna de la matriz X.

3.4. Varianza de las proyecciones

Si trabajamos con $X^\star$ (datos normalizados), las variables tienen media 0, y las proyecciones sobre una recta también tendrán media 0 (ejercicio).

Por lo tanto, podemos calcular la varianza de la proyección sobre la recta generada por v por la fórmula:

$$\operatorname{var}(X^\star v) = \frac{\langle X^\star v, X^\star v\rangle}{n} = \frac{(X^\star v)^T(X^\star v)}{n} = \frac{v^T(X^\star)^T X^\star v}{n} = v^T\left(\frac{(X^\star)^T X^\star}{n}\right)v.$$

¡Apareció la matriz de covarianza!

$$\operatorname{var}(X^\star v) = v^T\Sigma v$$

4. Componentes principales

4.1. Primera componente

Para encontrar la dirección en la que las proyecciones tienen mayor varianza, debemos encontrar el vector v que maximiza

$$v^T\Sigma v.$$

Teoremas de Álgebra Lineal. La matriz Σ es simétrica y definida positiva, por lo tanto satisface las siguientes propiedades:

- Es diagonalizable.
- Todos los autovalores son reales no-negativos.
- Existe una base de autovectores ortogonales.

El vector v que maximiza $v^T\Sigma v$ es el autovector de Σ correspondiente al mayor autovalor.

Este vector v es la dirección de la primera componente principal.

4.2. ¿Cómo calculamos las demás componentes?

Elegimos las direcciones de forma tal que

- la proyección en la primer dirección tenga la mayor varianza,
- la proyección en la segunda dirección tenga la segunda mayor varianza,
- y así siguiendo,

con la condición adicional de que las direcciones sean perpendiculares. Esto implica que la información contenida en cada proyección sea independiente de las demás proyecciones.

Obtenemos: Las siguientes direcciones corresponden a los autovectores de la matriz de covarianza correspondientes a los demás autovalores ordenados de mayor a menor.

4.3. Cálculo de PCA paso a paso

A partir de la matriz de covarianza, podemos calcular las componentes principales de la siguiente forma.

1. Calculamos la matriz $X^\star = X - \bar{X}$ de datos normalizados.
2. Calculamos la matriz de covarianza $\Sigma = \frac{(X^\star)^T X^\star}{n} \in \mathbb{R}^{p\times p}$.
3. Calculamos los autovalores de Σ: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ y los correspondientes autovectores $u_1, \ldots, u_p$ (de norma 1).
4. Definimos las nuevas variables $z_i = X^\star u_i$ (las proyecciones de los datos sobre las direcciones principales).
5. Si U es la matriz de autovectores, el nuevo conjunto de datos es $Z = X^\star U$ (las nuevas variables son las columnas de Z).

4.4. Relación entre autovectores y componentes principales

- Los autovectores de la matriz de covarianza definen las direcciones de las componentes principales.
- Los autovalores indican la cantidad de varianza explicada por cada componente principal.
- Las componentes principales se obtienen proyectando los datos sobre estos autovectores:
$$Z = X^\star U$$
donde Z son las componentes principales, X es la matriz de datos y U es la matriz de autovectores.

4.5. Varianza explicada

- Los autovalores indican la cantidad de varianza explicada por cada componente principal.

Ejemplo: Si los autovalores de la matriz de covarianza son

$$\lambda_1 = 30, \lambda_2 = 8, \lambda_3 = 2,$$

definimos la varianza total como la suma $\lambda_1 + \lambda_2 + \lambda_3 = 30 + 8 + 2 = 40$.

El porcentaje explicado por cada componente sera el valor de cada autovalor dividido por la varianza total.

Los porcentajes de varianza explicada son

$$\frac{30}{40} = 0.75, \frac{8}{40} = 0.2 \text{ y } \frac{2}{40} = 0.05.$$

4.6. Varianza explicada acumulada

Los porcentajes de varianza explicada son

$$\frac{30}{40} = 0.75, \frac{8}{40} = 0.2 \text{ y } \frac{2}{40} = 0.05.$$

Si consideramos las dos primeras componentes, tendremos un total de varianza explicada

$$0.75 + 0.2 = 0.95.$$

Es decir, que si utilizamos las dos primeras componentes, perdemos solo un 5 % de información.

De esta forma podemos reducir la dimensión de los datos, podemos quedarnos con solo estas dos componentes sin perder mucha información.

4.7. Ventajas de PCA

- Reducción de dimensionalidad: Permite trabajar con menos variables sin perder mucha información.
- Eliminación de redundancia: Los componentes principales son ortogonales entre sí, eliminando la multicolinealidad.
- Mejora del rendimiento de los algoritmos: Menos variables pueden resultar en tiempos de computación más rápidos.