

# Introducción a PCA

# Maldición de la dimensionalidad

- ▶ Un problema muy importante en Estadística Multivariante es la **maldición de la dimensionalidad**: si el ratio  $n/p$  no es suficientemente grande, algunos problemas no se pueden abordar tan fácilmente.
- ▶ Que el ratio  $n/p$  sea grande significa que  $n$  tiene que ser mucho mayor que  $p$ :  
 $n \gg p$ .
- ▶ Por ejemplo, supongamos que tenemos  $n$  datos de una Normal  $p$  –dimensional:  
$$\mathbf{x} \sim N(\mu, \Sigma)$$
- ▶ En este caso, el número de parámetros a estimar es  $p + p(p + 1)/2$
- ▶ Si  $p = 5$  o  $p = 10$  hay 20 o 65 parámetros respectivamente.
- ▶ Mientras más alto sea  $p$ , mucho mayor tiene que ser el número de observaciones para poder obtener estimaciones fiables.

# Reducción de la dimensión

- Hay varias técnicas de reducción de la dimensión que tratan de contestar la misma pregunta:

¿Es posible describir con precisión los valores de las  $p$  variables mediante un número menor de variables  $r < p$ ?

La respuesta es sí, y con ello se habrá reducido la dimensión del problema a costa de una pequeña pérdida de información.

- Vamos a ver dos técnicas:

1. Análisis de Componentes Principales (PCA)
2. Análisis Factorial (FA), En algunos libros aparece como AFE (análisis factorial exploratorio).

# Objetivo

- ▶ Dadas  $n$  observaciones de  $p$  variables, se analiza si es posible representar adecuadamente esta información con un número **menor** de variables construidas como **combinaciones lineales** de las originales.
- ▶ Por ejemplo, con variables con **alta dependencia** es frecuente que un pequeño número de nuevas variables (menos del 20% de las originales) expliquen la mayor parte (más del 80%) de la variabilidad original.

# Utilidad

Su utilidad es doble:

1. Permite **representar óptimamente en un espacio de dimensión pequeña**, observaciones de un espacio general  $p$  –dimensional. En este sentido componentes principales es el primer paso para identificar posibles variables latentes no observadas, que están generando la variabilidad de los datos.
2. Permite transformar las variables originales, en general correladas, en **nuevas variables incorreladas**, facilitando la interpretación de los datos.

# PCA y FA

- ▶ En esta Sección presentaremos PCA únicamente como una **herramienta exploratoria** para facilitar la descripción e interpretación de los datos.
- ▶ El problema de inferir si las propiedades de reducción de la dimensión encontradas en los datos puede extenderse a una **población** se estudiara en el capítulo de análisis factorial (FA).

# El problema

- ▶ El problema que se desea resolver es cómo encontrar un espacio de dimensión más reducida que represente adecuadamente los datos.
- ▶ El problema puede abordarse desde tres perspectivas equivalentes:
  1. Enfoque descriptivo
  2. Enfoque estadístico
  3. Enfoque geométrico

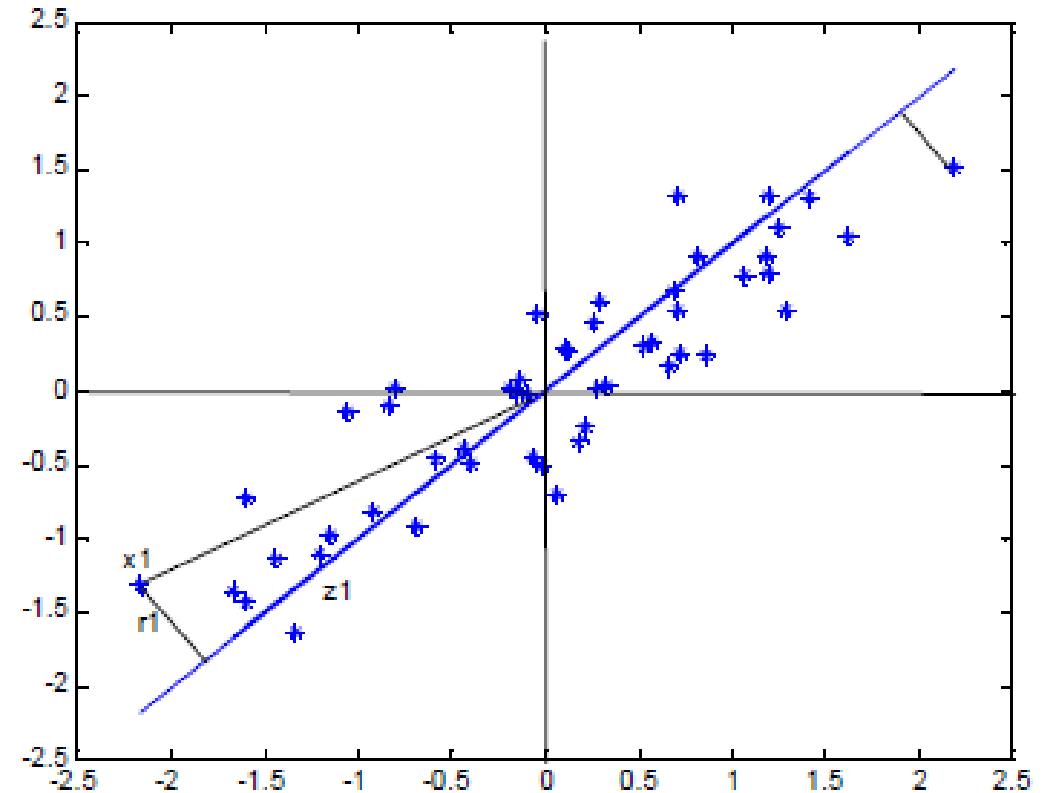
# Enfoque descriptivo

- ▶ Se desea encontrar un subespacio de dimensión menor que  $p$ , tal que, al proyectar los puntos sobre él, estos conserven su estructura con la menor distorsión posible.
- ▶ Veamos cómo convertir esta noción intuitiva en un criterio matemático operativo.
- ▶ Consideremos primero el caso de dos dimensiones ( $p = 2$ ) y un subespacio de dimensión uno, una recta.
- ▶ Se desea que las proyecciones de los puntos bidimensionales sobre esta recta mantengan, lo mejor posible, sus posiciones relativas.



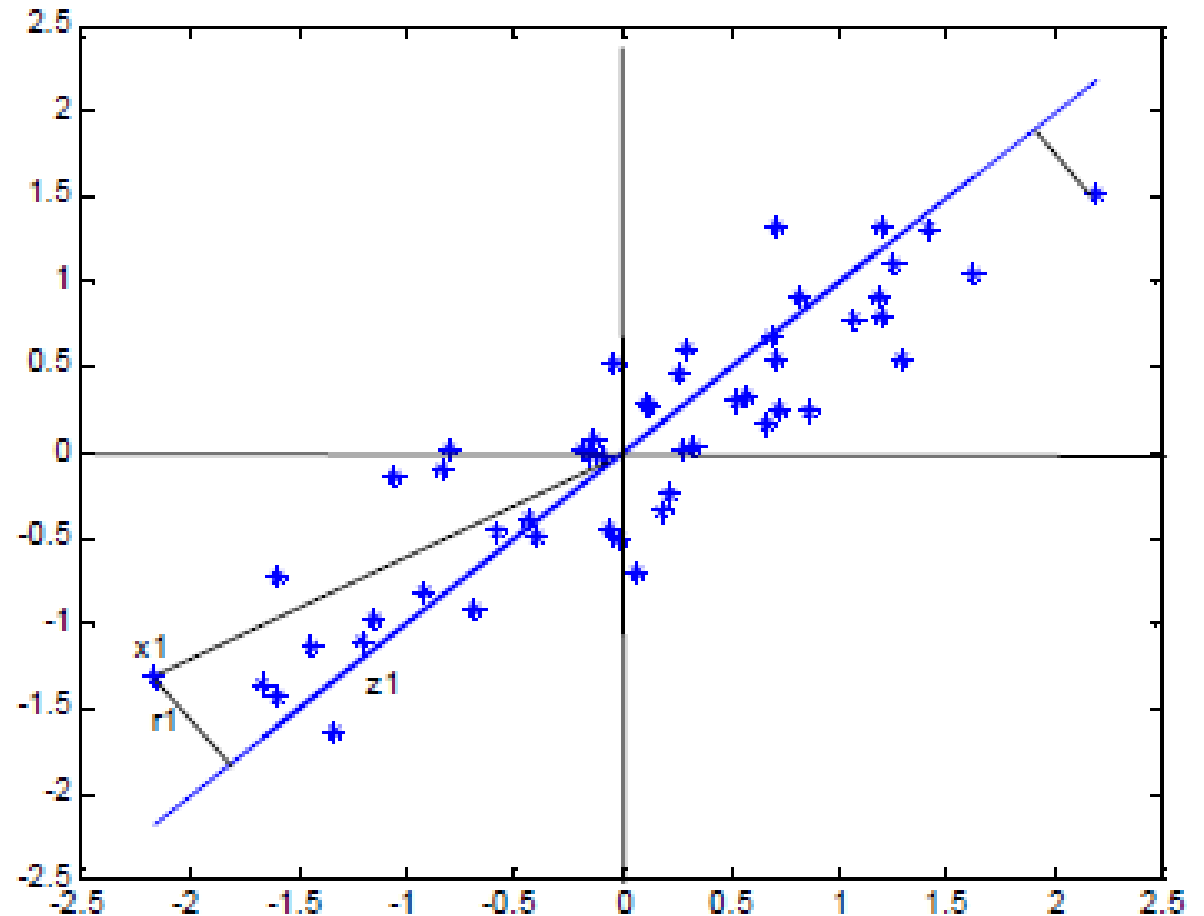
# Enfoque descriptivo

- ▶ La figura muestra los puntos en un diagrama de dispersión, y una recta.
- ▶ Esta recta, intuitivamente, proporciona un buen resumen de los datos, ya que las proyecciones de los puntos sobre ella indican aproximadamente la situación de los puntos en el plano.
- ▶ La representación es buena porque la recta pasa cerca de todos los puntos y estos se deforman poco al proyectarlos.



# Enfoque descriptivo

- ▶ Al proyectar cada punto sobre la recta se forma **un triángulo rectángulo**
- ▶ La **hipotenusa** es la distancia del punto al origen  $(\mathbf{x}_i^t \mathbf{x}_i)^{1/2}$
- ▶ Los **catetos** son la proyección del punto sobre la recta ( $z_i$ ) y la distancia entre el punto y su proyección ( $r_i$ ).



# Enfoque descriptivo

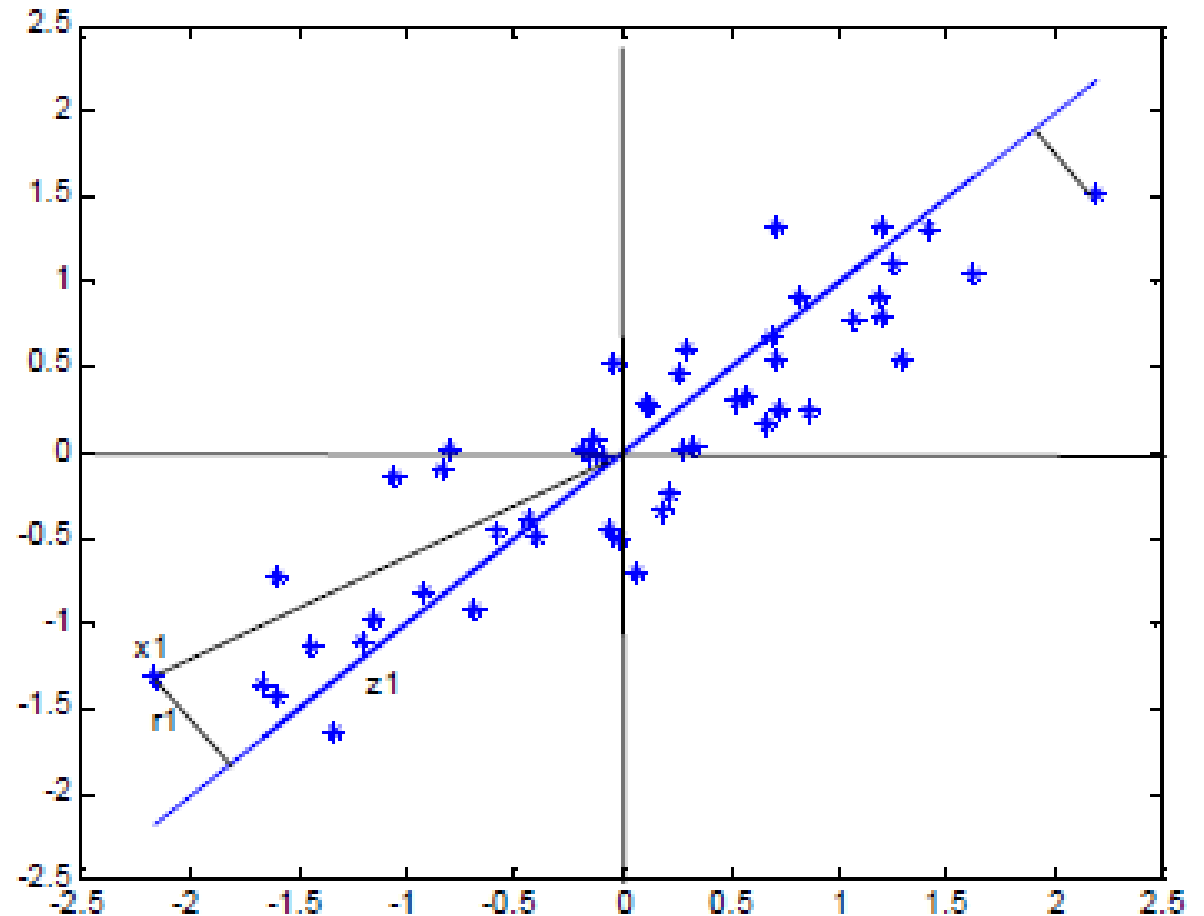
- ▶ Por el Teorema de Pitágoras, podemos escribir:

$$\mathbf{x}_i^t \mathbf{x}_i = z_i^2 + r_i^2$$

- ▶ Sumando esta expresión para todos los puntos, se obtiene:

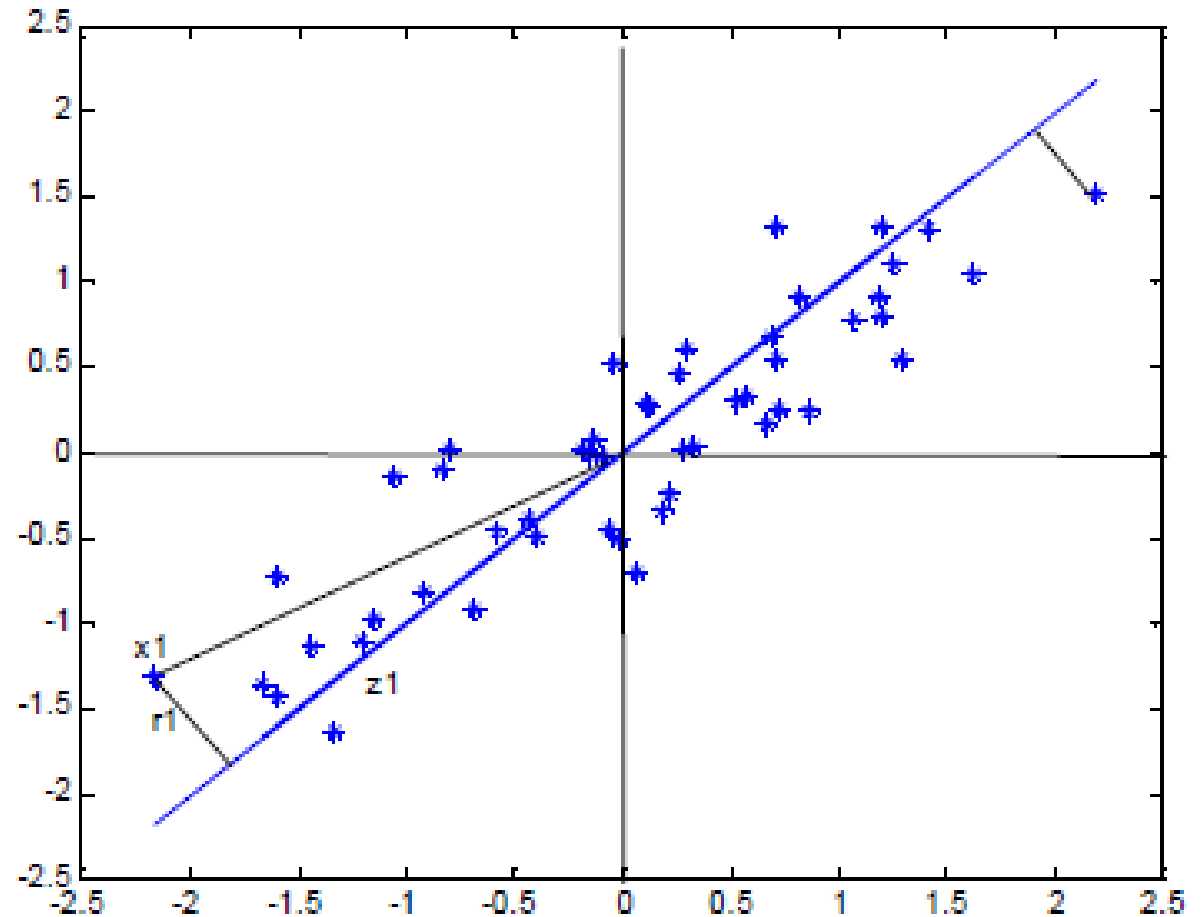
$$\sum_{i=1}^n \mathbf{x}_i^t \mathbf{x}_i = \sum_{i=1}^n z_i^2 + \sum_{i=1}^n r_i^2$$

- ▶ Minimizar  $\sum_{i=1}^n r_i^2$ , la suma de las distancias a la recta de todos los puntos, es equivalente a maximizar  $\sum_{i=1}^n z_i^2$ , la suma al cuadrado de los valores de las proyecciones.



# Enfoque descriptivo

- ▶ Como las proyecciones  $z_i$  son variables de media cero, maximizar la suma de sus cuadrados equivale a maximizar su **varianza**.
- ▶ Este resultado es intuitivo: la recta parece adecuada porque conserva lo más posible la **variabilidad** original de los puntos.
- ▶ Si consideramos una dirección de proyección **perpendicular** a la de la recta en esta figura, los puntos tendrían muy poca variabilidad y perderíamos la información sobre sus distancias en el espacio.



# Enfoque estadístico

- ▶ Representar puntos  $p$  –dimensionales con la mínima pérdida de información en un espacio de **dimensión uno** es equivalente a sustituir las  $p$  variables originales por una nueva variable,  $z_1$ , que resuma óptimamente la información.
- ▶ Esto supone que la nueva variable debe tener globalmente **máxima correlación** con las originales o, en otros términos, debe permitir prever las variables originales con la máxima precisión.
- ▶ Esto no será posible si la nueva variable toma un valor semejante en todos los elementos, es decir, usaremos la variable de **máxima variabilidad**.

# Enfoque estadístico

- ▶ En la figura anteriormente vista, la recta no es la línea de regresión de ninguna de las variables con respecto a la otra, que se obtienen minimizando las distancias verticales u horizontales, sino que al minimizar las distancias ortogonales o de proyección se encuentra **entre ambas rectas de regresión**.
- ▶ Este enfoque puede extenderse para obtener el mejor subespacio resumen de los datos de dimensión 2. Para ello calcularemos **el plano que mejor aproxima a los puntos**.
- ▶ Estadísticamente esto equivale a encontrar una segunda variable  $z_2$ , incorrelada con la anterior, y que tenga **varianza máxima**.
- ▶ En general, la componente  $z_r$  ( $r < p$ ) tendrá varianza máxima entre todas las combinaciones lineales de las  $p$  variables originales, con la condición de estar incorrelada con las  $z_1, \dots, z_{r-1}$  previamente obtenidas.

# Enfoque geométrico

- ▶ Si consideramos la nube de puntos de la figura vemos que los puntos se sitúan siguiendo una **elipse** y podemos describir su orientación dando la dirección del eje mayor de la elipse y la posición de los puntos por su proyección sobre esta dirección.
- ▶ Puede demostrarse que este eje es la recta que minimiza las distancias ortogonales y volvemos al problema que ya hemos resuelto.
- ▶ En mayores dimensiones tendremos **elipsoides** y la mejor aproximación a los datos es la proporcionada por el eje mayor del elipsoide.
- ▶ Considerar los ejes del elipsoide como nuevas variables originales supone pasar de variables correladas a variables ortogonales.

# Cálculo de los componentes



# Cálculo del primer componente

- ▶ El primer componente principal será la **combinación lineal de las variables originales que tenga varianza máxima**.
- ▶ Los valores de este primer componente en los  $n$  individuos se representarán por un vector  $\mathbf{z}_1$ , dado por:

$$\mathbf{z}_1 = \mathbf{x}\mathbf{a}_1$$

- ▶ Estamos suponiendo sin pérdida de generalidad que ya  $\mathbf{x}$  tiene los datos centrados.
- ▶ Como las variables originales tienen media cero también  $\mathbf{z}_1$  tendrá media nula.
- ▶ Su **varianza** será:

$$var(\mathbf{z}_1) = \frac{1}{n} \mathbf{z}_1^t \mathbf{z}_1 = \frac{1}{n} \mathbf{a}_1^t \mathbf{x}^t \mathbf{x} \mathbf{a}_1 = \mathbf{a}_1^t \mathbf{S} \mathbf{a}_1$$

donde  $\mathbf{S}$  es la matriz de covarianza de las observaciones.

# Cálculo del primer componente

- ▶ Es obvio que podemos maximizar la varianza sin limite aumentando el módulo del vector  $\mathbf{a}_1$ .
- ▶ Para que la maximización tenga solución debemos imponer una **restricción** al módulo del vector  $\mathbf{a}_1$ , y, sin pérdida de generalidad, impondremos que  $\mathbf{a}_1^t \mathbf{a}_1 = 1$ .
- ▶ Introduciremos esta restricción mediante el **multiplicador de Lagrange**:

$$M = \mathbf{a}_1^t S \mathbf{a}_1 - \lambda (\mathbf{a}_1^t \mathbf{a}_1 - 1)$$

donde  $S$  es la matriz de covarianza de las observaciones.

# Cálculo del primer componente

- ▶ **Maximizaremos** esta expresión de la forma habitual derivando respecto a los componentes de  $\mathbf{a}_1$  e igualando a cero.
- ▶ Entonces:

$$\frac{\partial M}{\partial \mathbf{a}_1} = 2S\mathbf{a}_1 - 2\lambda\mathbf{a}_1 = 0$$

- ▶ Cuya solución es:

$$S\mathbf{a}_1 = \lambda\mathbf{a}_1$$

- ▶ Esto implica que  $\mathbf{a}_1$  es un **vector propio** de la matriz  $S$ , y  $\lambda$  su correspondiente **valor propio**.

# Cálculo del primer componente

- ▶ Para determinar qué valor propio de  $S$  es la solución de la ecuación, tendremos en cuenta que, multiplicando por la izquierda por  $\mathbf{a}_1^t$  y usando la restricción  $\mathbf{a}_1^t \mathbf{a}_1 = 1$ , queda:

$$\mathbf{a}_1^t S \mathbf{a}_1 = \lambda \mathbf{a}_1^t \mathbf{a}_1 = \lambda$$

- ▶ Como vimos anteriormente,  $\text{var}(\mathbf{z}_1) = \mathbf{a}_1^t S \mathbf{a}_1$
- ▶ Entonces  $\lambda$  es la varianza de  $\mathbf{z}_1$ .
- ▶ Como esta es la cantidad que queremos maximizar,  $\lambda$  será el mayor valor propio de la matriz  $S$ .
- ▶ Su vector propio asociado,  $\mathbf{a}_1$ , define los coeficientes de cada variable en el primer componente principal.

# Cálculo del segundo componente

- ▶ Vamos a obtener el mejor **plano de proyección** de las variables originales.
- ▶ Lo calcularemos estableciendo como función objetivo que la **suma de las varianzas** de  $\mathbf{z}_1 = \mathbf{x}\mathbf{a}_1$  y  $\mathbf{z}_2 = \mathbf{x}\mathbf{a}_2$  sea máxima donde  $\mathbf{a}_1$  y  $\mathbf{a}_2$  son los vectores que definen el plano.
- ▶ La **función objetivo** será:

$$\phi = \mathbf{a}_1^t \mathbf{S} \mathbf{a}_1 + \mathbf{a}_2^t \mathbf{S} \mathbf{a}_2 - \lambda_1 (\mathbf{a}_1^t \mathbf{a}_1 - 1) - \lambda_2 (\mathbf{a}_2^t \mathbf{a}_2 - 1)$$

que incorpora las **restricciones** de que las direcciones deben de tener módulo unitario  $\mathbf{a}_i^t \mathbf{a}_i = 1$ , para  $i = 1, 2$ .

# Cálculo del segundo componente

- **Derivando** e igualando a cero:

$$\frac{\partial \phi}{\partial \mathbf{a}_1} = 2S\mathbf{a}_1 - 2\lambda_1\mathbf{a}_1 = 0$$

$$\frac{\partial \phi}{\partial \mathbf{a}_2} = 2S\mathbf{a}_2 - 2\lambda_2\mathbf{a}_2 = 0$$

- La **solución** de este sistema es:

$$\begin{aligned} S\mathbf{a}_1 &= \lambda_1\mathbf{a}_1 \\ S\mathbf{a}_2 &= \lambda_2\mathbf{a}_2 \end{aligned}$$

que indica que  $\mathbf{a}_1$  y  $\mathbf{a}_2$  deben ser vectores propios de  $S$ .

# Cálculo del segundo componente

- ▶ Tomando los vectores propios de norma uno y sustituyendo, se obtiene que, en el máximo, la función objetivo es:

$$\phi = \lambda_1 + \lambda_2$$

Entonces:

- ▶  $\lambda_1$  y  $\lambda_2$  deben ser los dos auto-valores mayores de la matriz  $S$
- ▶  $\mathbf{a}_1$  y  $\mathbf{a}_2$  deben ser sus correspondientes auto-vectores.

# Cálculo del segundo componente

- Observemos que la covarianza entre  $z_1$  y  $z_2$ , dada por  $\mathbf{a}_1^t \mathbf{S} \mathbf{a}_2$ , es cero ya que

$$\mathbf{a}_1^t \mathbf{a}_2 = 0$$

- Entonces, las variables  $z_1$  y  $z_2$  estarán incorreladas.
- Puede demostrarse que si en lugar de maximizar la suma de varianzas, que es la traza de la matriz de covarianzas de la proyección, se maximiza la varianza generalizada (el determinante de la matriz de covarianzas) se obtiene el mismo resultado.



# Generalización

- ▶ Puede demostrarse análogamente que el espacio de dimensión  $r$  que mejor representa a los puntos viene definido por los vectores propios asociados a los  $r$  mayores autovalores de  $S$ .
- ▶ Estas direcciones se denominan direcciones principales de los datos y a las nuevas variables definidas por esas direcciones se les llama componentes principales.
- ▶ En general, la matriz de datos  $\mathbf{x}$  (y por tanto la  $S$ ) tienen rango  $p$ , existiendo entonces tantas componentes principales como variables que se obtendrán calculando los valores propios  $\lambda_1, \dots, \lambda_p$  de la matriz de covarianzas  $S$ , mediante:

$$|S - \lambda I| = 0$$

- ▶ Y sus vectores asociados son:

$$(S - \lambda_i I) \mathbf{a}_i = 0$$

# Generalización

- ▶ Los términos  $\lambda_i$  son **reales**, al ser la matriz  $S$  simétrica.
- ▶ Serán además **positivos**, ya que  $S$  es definida positiva.
- ▶ Los vectores propios asociados a dos valores propios diferentes serán **ortogonales**.
- ▶ Si  $S$  fuese **semi-definida positiva** de rango menor que  $p$ , habrían algunos autovalores positivos y el resto serían ceros.

# Generalización

- ▶ Llamando  $\mathbf{z}$  a la matriz cuyas columnas son los valores de los  $p$  componentes en los  $n$  individuos, estas nuevas variables están relacionadas con las originales mediante:

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

donde la matriz  $A$  cumple:  $A^t A = I$ .

- ▶ Calcular los **componentes principales** equivale a aplicar una **transformación ortogonal**  $A$  a las variables  $\mathbf{x}$  (ejes originales) para obtener unas nuevas variables  $\mathbf{z}$  **incorreladas** entre sí.
- ▶ Esta operación puede interpretarse como **elegir unos nuevos ejes coordenados**, que coincidan con los ejes naturales de los datos.

# Generalización


- ▶ La transformación ortogonal  $A$  es de tamaño  $p \times r$  si se están hallando los  $r$  primeros componentes principales.
- ▶ Sus columnas serán los  $r$  primeros auto-vectores de la matriz de covarianzas de los datos  $S$ .
- ▶ La matriz de covarianza de  $z$  sería  $S_z$ , y es una matriz diagonal con elementos  $\lambda_1, \dots, \lambda_r$ , es decir, los primeros auto-valores de  $S$ .

$$S_z = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_r \end{pmatrix}$$

# Ventajas de PCA

Entonces, la utilidad de PCA se puede ver de las siguientes formas:

- ▶ Permite una representación óptima, en un espacio de **dimensiones reducidas**, de las observaciones originales.
- ▶ Permite que las variables correlacionadas originales se transformen en nuevas variables **no correlacionadas**, facilitando la interpretación de los datos.



# Propiedades de los componentes

# Propiedades de los componentes

## 1. Conservan la variabilidad inicial:

- La suma de las varianzas de los componentes es igual a la suma de las varianzas de las variables originales, y la varianza generalizada de los componentes es igual a la original.
- En efecto, la varianza del componente  $\mathbf{z}_h$  es:

$$var(\mathbf{z}_h) = \lambda_h$$

- La suma de los auto-valores es la traza de la matriz de covarianza:

$$traza(S) = var(\mathbf{x}_1) + \dots + var(\mathbf{x}_p) = \lambda_1 + \dots + \lambda_p$$

- Entonces:

$$\sum_{i=1}^p var(\mathbf{x}_i) = \sum_{i=1}^p \lambda_i = \sum_{i=1}^p var(\mathbf{z}_i)$$

# Propiedades de los componentes

## 1. Conservan la variabilidad inicial:

- Esto significa que las nuevas variables  $\mathbf{z}_i$  tienen conjuntamente la misma variabilidad que las variables originales, pero su distribución es muy distinta en los dos conjuntos.
- Para comprobar que también conservan la varianza generalizada, valor del determinante de  $S$ , como el determinante es el producto de los  $\lambda_i$ :

$$|S_{\mathbf{x}}| = \lambda_1 \dots \lambda_p = \prod_{i=1}^p \text{var}(\mathbf{z}_i) = |S_{\mathbf{z}}|$$



# Propiedades de los componentes

2. La proporción de variabilidad explicada por un componente es el cociente entre su varianza, el valor propio asociado al vector propio que lo define, y la suma de los valores propios de la matriz.
- La varianza del componente  $h$  es  $\lambda_h$ , el valor propio que define el componente.
  - La suma de todas las varianzas de las variables originales es  $\sum_{i=1}^p \lambda_i$ , que es igual a la suma de las varianzas de los componentes.
  - Entonces la proporción de variabilidad total explicada por el componente  $h$  es:

$$\frac{\lambda_h}{\sum_{i=1}^p \lambda_i}$$

# Propiedades de los componentes

3. Las covarianzas entre cada componente principal y las variables originales  $\mathbf{x}$  vienen dadas por el producto entre las coordenadas del vector propio que define al componente y el valor propio:

$$\text{cov}(\mathbf{z}_i; \mathbf{x}_1, \dots, \mathbf{x}_p) = \lambda_i \mathbf{a}_i = (\lambda_i a_{i1}, \dots, \lambda_i a_{ip})$$

donde  $\mathbf{a}_i$  es el vector de coeficientes de la componente  $\mathbf{z}_i$

- Para justificar este resultado, vamos a calcular la matriz  $p \times p$  de covarianzas entre los componentes y las variables originales. Esta matriz es:

$$\text{cov}(\mathbf{z}, \mathbf{x}) = \frac{1}{n} \mathbf{z}^t \mathbf{x}$$

# Propiedades de los componentes

3.

$$\text{cov}(\mathbf{z}_i; \mathbf{x}_1, \dots, \mathbf{x}_p) = \lambda_i \mathbf{a}_i = (\lambda_i a_{i1}, \dots, \lambda_i a_{ip})$$

- La primera fila de la matriz  $\text{cov}(\mathbf{z}, \mathbf{x})$  proporciona las covarianzas entre la primera componente principal y las  $p$  variables originales. Como  $\mathbf{z} = \mathbf{x}A$ , sustituyendo:

$$\text{cov}(\mathbf{z}, \mathbf{x}) = \frac{1}{n} \mathbf{A}^t \mathbf{x}^t \mathbf{x} = \mathbf{A}^t S = D \mathbf{A}^t$$

donde  $A$  contiene en sus columnas los vectores propios de  $S$ , y  $D$  es la matriz diagonal de los valores propios.

- En consecuencia, la covarianza entre, por ejemplo, el primer componente principal y las  $p$  variables vendrá dada por la primera fila de  $\mathbf{A}^t S$ , es decir,  $\mathbf{a}_1^t S$ , o también  $\lambda_1 \mathbf{a}_1^t$ , donde  $\mathbf{a}_1$  es el vector de coeficientes de la primera componente principal.

# Propiedades de los componentes

4. La correlación entre un componente principal y una variable  $x$  es proporcional al coeficiente de esa variable en la definición del componente, y el coeficiente de proporcionalidad es el cociente entre la desviación típica del componente y la desviación típica de la variable.
- Para comprobarlo:

$$\text{corr}(\mathbf{z}_i, \mathbf{x}_j) = \frac{\text{cov}(\mathbf{z}_i, \mathbf{x}_j)}{\sqrt{\text{var}(\mathbf{z}_i)\text{var}(\mathbf{x}_j)}} = \frac{\lambda_i a_{ij}}{\sqrt{\lambda_i s_j^2}} = a_{ij} \frac{\sqrt{\lambda_i}}{s_j}$$

# Propiedades de los componentes

- 5. Las  $r$  componentes principales ( $r < p$ ) proporcionan la predicción lineal óptima con  $r$  variables del conjunto de variables  $x$ .
  - Esta afirmación puede expresarse de dos formas. La primera demostrando que la mejor predicción lineal con  $r$  variables de las variables originales se obtiene utilizando las  $r$  primeras componentes principales.
  - La segunda demostrando que la mejor aproximación de la matriz de datos que puede construirse con una matriz de rango  $r$  se obtiene construyendo esta matriz con los valores de los  $r$  primeros componentes principales.

# Propiedades de los componentes

6. Si estandarizamos los componentes principales, dividiendo cada uno por su desviación típica, se obtiene la estandarización multivariante de los datos originales.

- Estandarizando los componentes  $\mathbf{z}$  por sus desviaciones típicas, se obtienen las nuevas variables:

$$\mathbf{y}_c = \mathbf{z}D^{-1/2} = \mathbf{x}AD^{-1/2}$$

donde  $D^{-1/2}$  es la matriz que contiene las inversas de las desviaciones típicas de las componentes.

- La estandarización multivariante de una matriz de variables  $\mathbf{x}$  de media cero viene dada por:

$$\mathbf{y}_s = \mathbf{x}AD^{-1/2}A^t$$

- Por tanto, la estandarización multivariante puede interpretarse como obtener los componentes principales y estandarizarlos para que tengan todos la misma varianza.



# PCA normado o con correlaciones

# PCA normado o con correlaciones

- ▶ Los componentes principales se obtienen **maximizando la varianza de la proyección**. En términos de las variables originales esto supone maximizar:

$$M = \sum_{i=1}^p a_i^2 s_i^2 + 2 \sum_{i=1}^p \sum_{j=i+1}^p a_i a_j s_{ij}$$

- ▶ Con la restricción  $\mathbf{a}^t \mathbf{a} = 1$
- ▶ Si alguna de las variables originales, por ejemplo  $\mathbf{x}_1$ , tiene una varianza  $s_1^2$  **mayor** que las demás, la manera de aumentar  $M$  es hacer tan grande como podamos la coordenada  $a_1$  asociada a esta variable.
- ▶ En el límite si una variable tiene una varianza mucho mayor que las demás el **primer componente principal** coincidirá muy aproximadamente con esta variable.



# PCA normado o con correlaciones

- ▶ Cuando las variables tienen **unidades distintas** esta propiedad no es conveniente: si **disminuimos la escala de medida** de una variable cualquiera, de manera que **aumenten en magnitud sus valores numéricos** (pasamos por ejemplo de medir en km a medir en metros), el peso de esa variable en el análisis aumentará, ya que:
  1. Su varianza será mayor y aumentará su coeficiente en el componente, ya que contribuye más a aumentar  $M$ .
  2. Sus covarianzas con todas las variables aumentarán, con el consiguiente efecto de incrementar  $a_i$ .
- ▶ En resumen, cuando las escalas de medida de las variables son muy distintas, la maximización de  $M$  **dependerá decisivamente de estas escalas de medida** y las variables con valores más grandes tendrán más peso en el análisis.
- ▶ Si queremos evitar este problema, conviene **estandarizar las variables antes de calcular los componentes**, de manera que las magnitudes de los valores numéricos de las variables sean similares.

# PCA normado o con correlaciones

- ▶ La **estandarización** resuelve otro posible problema.
- ▶ Si las variabilidades de las  $\mathbf{x}$  son muy distintas, las variables con mayor varianza van a influir más en la determinación de la primera componente.
- ▶ Este problema se evita al estandarizar las variables, ya que entonces **las varianzas son la unidad**, y las **covarianza son iguales a los coeficientes de correlación**.
- ▶ La ecuación a maximizar se transforma en:

$$M' = 1 + 2 \sum_{i=1}^p \sum_{j=i+1}^p a_i a_j r_{ij}$$

donde  $r_{ij}$  es el coeficiente de correlación lineal entre las variables  $\mathbf{x}_i$  y  $\mathbf{x}_j$ .

- ▶ En consecuencia la solución depende de las correlaciones y no de las varianzas.

# PCA normado o con correlaciones

- ▶ Los **componentes principales normados** se obtienen calculando los vectores y valores propios de la matriz  $R$  de correlaciones.
- ▶ Llamando  $\lambda_i^R$  a los valores propios de esa matriz, que suponemos no singular, se verifica que:

$$\sum_{i=1}^p \lambda_i^R = \text{traza}(R) = p$$

# Propiedades

Las propiedades de los **componentes** extraídos de  $R$  son:

1. La proporción de variación o variabilidad explicada por  $\lambda_p^R$  será:

$$\frac{\lambda_p^R}{p}$$

2. Las correlaciones entre cada componente  $\mathbf{z}_j$  y las variables  $\mathbf{x}$  originales vienen dadas directamente por  $\mathbf{a}_j^t \sqrt{\lambda_j}$  **siendo**  $\mathbf{z}_j = \mathbf{x} \mathbf{a}_j$

# Conclusiones

- ▶ Cuando las variables  $x$  originales están en **distintas unidades** conviene aplicar el análisis de la matriz de **correlaciones** o **análisis normado**.
- ▶ Cuando las variables tienen las **mismas unidades**, ambas alternativas son posibles.
- ▶ Si las diferencias entre las varianzas de las variables son **informativas** y queremos tenerlas en cuenta en el análisis, **no debemos estandarizar las variables**.
- ▶ **Por ejemplo**, supongamos dos índices con la misma base pero uno fluctúa mucho y el otro es casi constante. Este hecho es informativo, y para tenerlo en cuenta en el análisis, no se deben estandarizar las variables, de manera que el índice de mayor variabilidad tenga más peso.
- ▶ **Por el contrario**, si las diferencias de variabilidad no son relevantes podemos eliminarlas con el análisis normado. En caso de duda, conviene realizar ambos análisis, y seleccionar aquel que conduzca a conclusiones más informativas.