Modelos Avanzados para Análisis de Datos I - MIIA4201

Periodo Vacaciones - Módulo 2

Maestría en Inteligencia Analítica para Toma de Decisiones

Profesor: Carlos Valencia

Departamento de Ingeniería Industrial
Universidad de Los Andes, Bogotá, Colombia

Clase 1 - Junio 26, 2018

Table of Contents

Introducción a la Reducción de Dimensiones

Reducción de Dimensiones

En muchas aplicaciones los datos que se deben analizar contienen muchas variables de interés, esto es, presentan alta dimensionalidad.

Recuerde que consideramos el problema en el cual tenemos datos estructurados, es decir, que se tiene una observación (dato o individuo) con p registros (variables).

Ejemplos:

1. Una observación es una persona y los registros corresponden a ciertas mediciones antropométricas (altura, peso, diámetro de la cabeza, etc.)
2. Un individuo es un estudiante con información de desempeño académico en diferentes evaluaciones de diferentes materias.
3. Una observación es una imagen digital monocromática y los registros son las 256 mediciones de intensidad de gris en cada pixel.
4. Una obervación en el tiempo t contiene varios indicadores económicos (inflación, desempleo, trm, etc.). Note que en este cado las observaciones no son independientes !

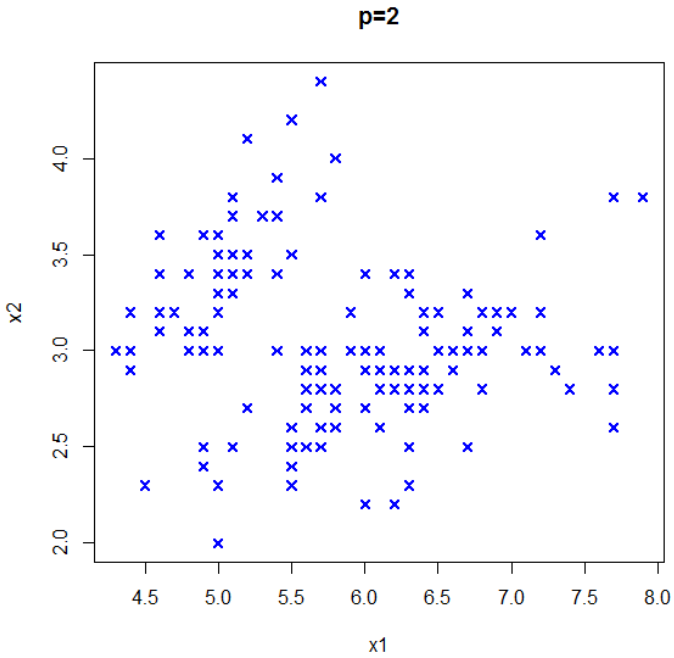
Introducción a la Reducción de Dimensiones

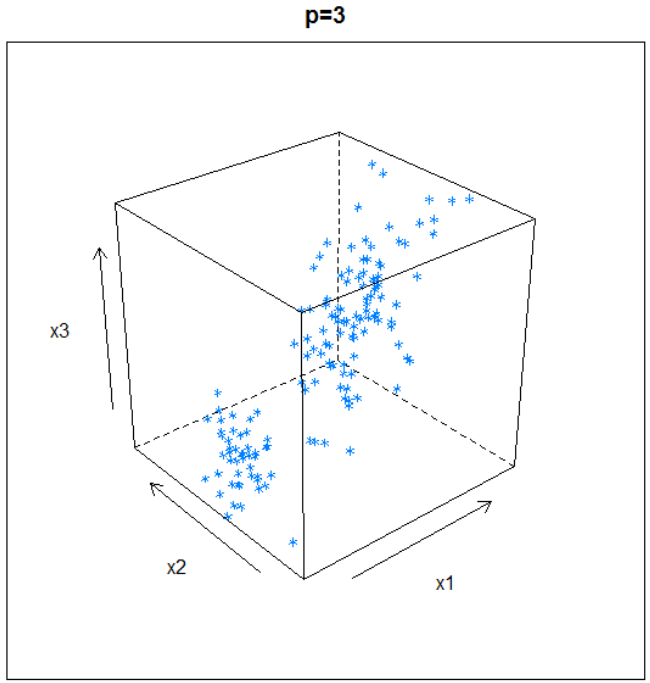
Reducción de Dimensiones

Estas estructuras de datos permiten representarse en forma matricial, donde cada observación es un fila: $x_i^T = (x_{i1}, x_{i2}, \cdots, x_{ip})$

$$\mathbf{X} = \in \mathbb{R}^{n \times p}$$

Los datos se pueden visualizar como puntos en el espacio $\mathbb{R}^p$:





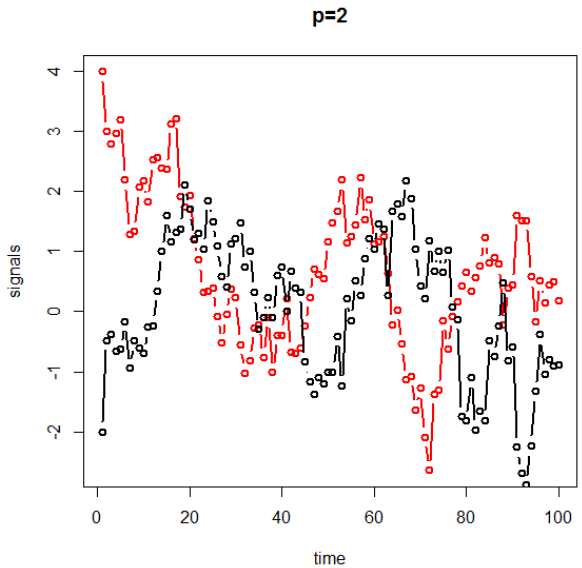
Introducción a la Reducción de Dimensiones

Reducción de Dimensiones

Se asume que las observaciones son realizaciones de vectores aleatorios con a misma distribución.
En algunos casos es conveniente asumir que son realizaciones independientes:

$$x_i \sim_{iid} f(x_i) : \mathbb{R}^p \rightarrow \mathbb{R}^{\geq 0}$$

En algunos casos, cada observación es una medición del grupo de variables en el tiempo t: $x_t^T = (x_{t1}, x_{t2}, \cdots, x_{tp})$, con lo que cada columna de $\mathbf{X}$ es una series de tiempo (señal). En este caso, los datos no son independientes y tiene más sentido visualizar $\mathbf{X}$ como:



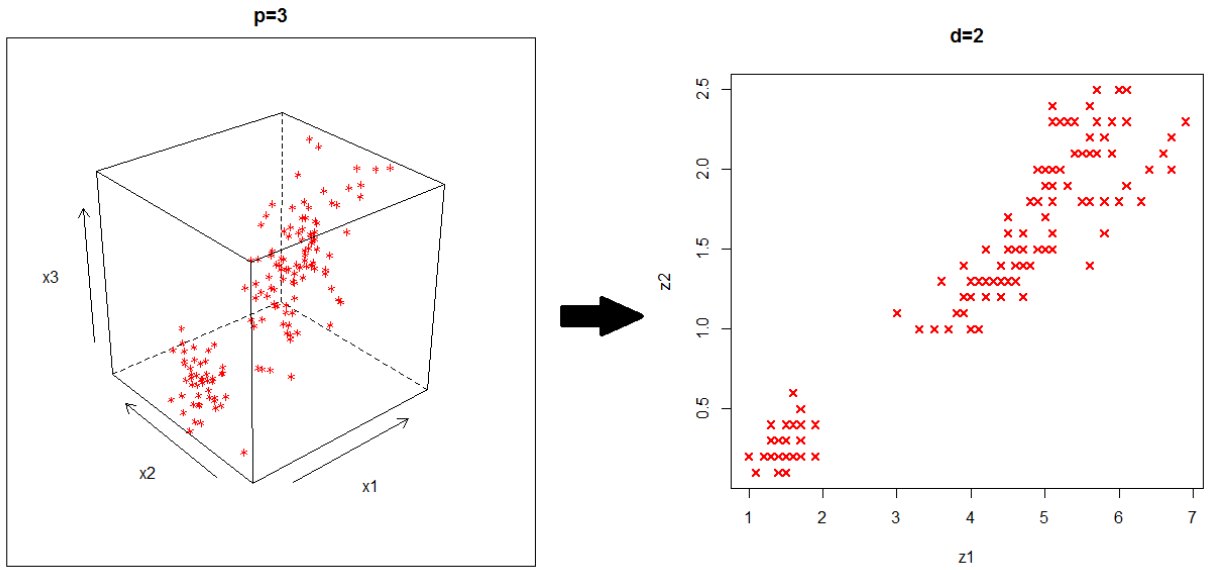
Introducción a la Reducción de Dimensiones

Reducción de Dimensiones

En muchos casos p es demasiado grande para el propósito del análisis.

El objetivo de la reducción de dimensiones es representar el conjunto de datos $\mathbf{X}$ lo mejor posible en un número menor de dimensiones $d < p$, manteniendo tanta información como sea posible. Esto generalmente implica transformar los datos encontrando nuevas variables:

$$\mathbf{X} \in \mathbb{R}^{n \times p} \rightarrow \mathbf{Z} \in \mathbb{R}^{n \times d}$$



Introducción a la Reducción de Dimensiones

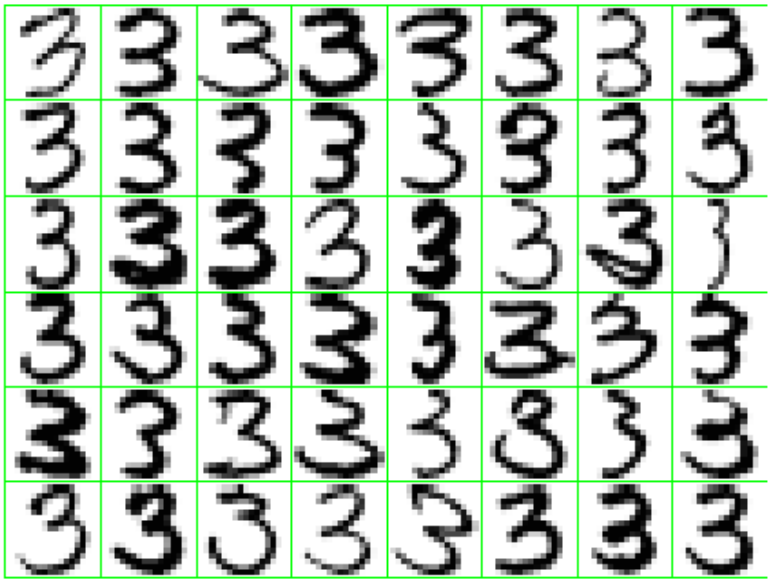
Por qué reducir la dimensionalidad?

En general se busca al menos uno de los siguientes objetivos:

1. Manipulación de Datos: Al tener menos dimensiones se pueden manipular los datos más fácilmente y hacer cómputos eficientes (*High Dimensionality*).
2. Visualización: Al tener menos variables, los datos se pueden visuallizar más fácil. Por ejemplo, para 2 o 3 dimensiones se pueden realizar gráficos de dispersión (*scatterplot*).
3. Compresión de archivos (matrices): Archivos de gran tamaño se pueden comprimir. Por ejemplo imágenes o videos.
4. Interpretación: Si se tiene muchas variables, es dificil comprender las características generales de un dato.
5. Identificación de variables independientes generadoras de las variables observadas (*unmixing*)
6. Generación de variables para usar en modelos predictivos (*feature generation*). Por ejemplo, regresión por componentes principales.

Introducción a la Reducción de Dimensiones

Ejemplos



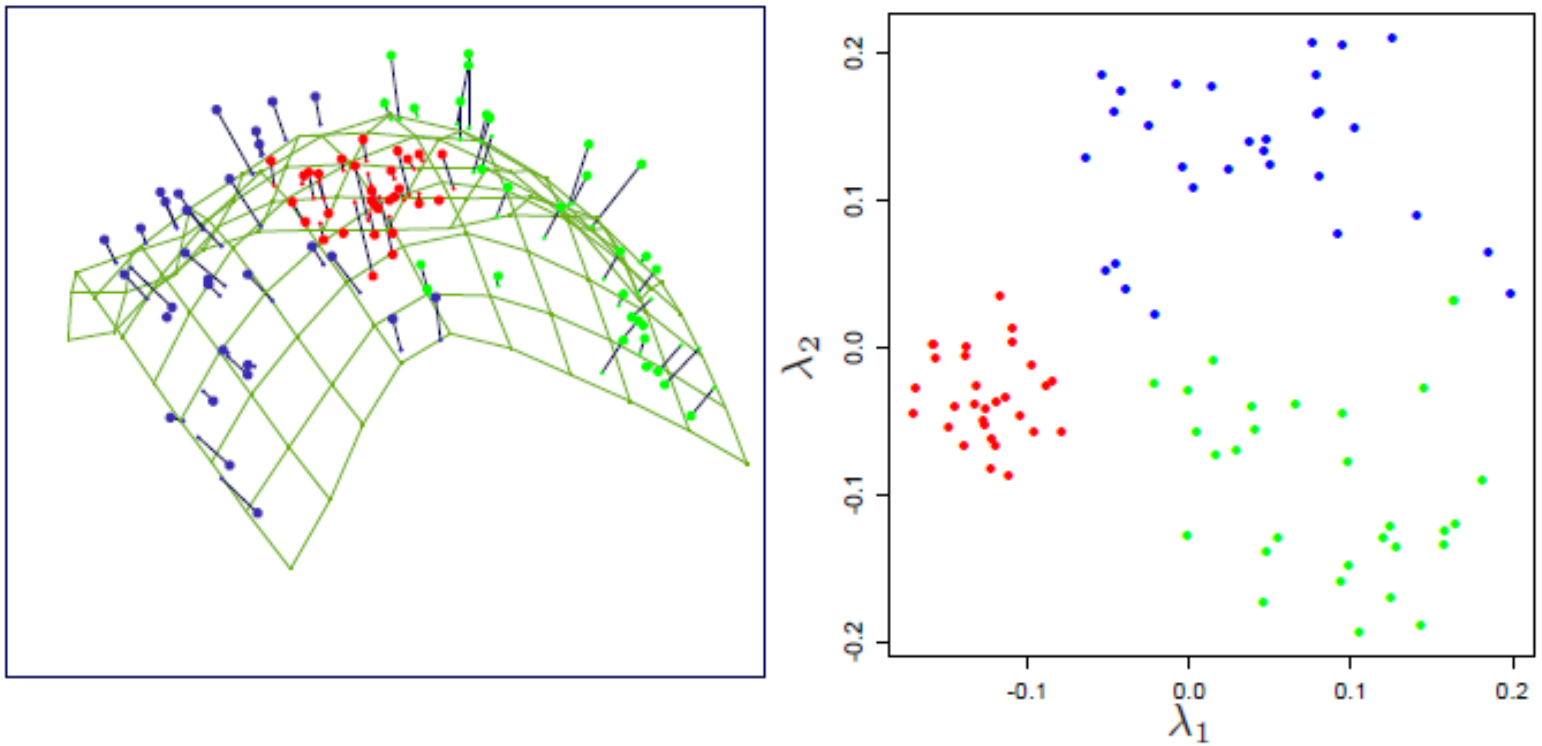
$$\begin{aligned} \hat{f}(\lambda) &= \bar{x} + \lambda_1 v_1 + \lambda_2 v_2 \\ &= \boxed{} + \lambda_1 \cdot \boxed{} + \lambda_2 \cdot \boxed{}. \end{aligned}$$

Tomado de The Elements of Statistical Learning. Hastie, et al.

Introducción a la Reducción de Dimensiones

Principio Fundamental

La dimensionalidad puede ser reducida cuando las variables producen información redundante. Una forma más formal de definir este fenómeno es que los datos tienden a caer en un *manifold* en $\mathbb{R}^p$.

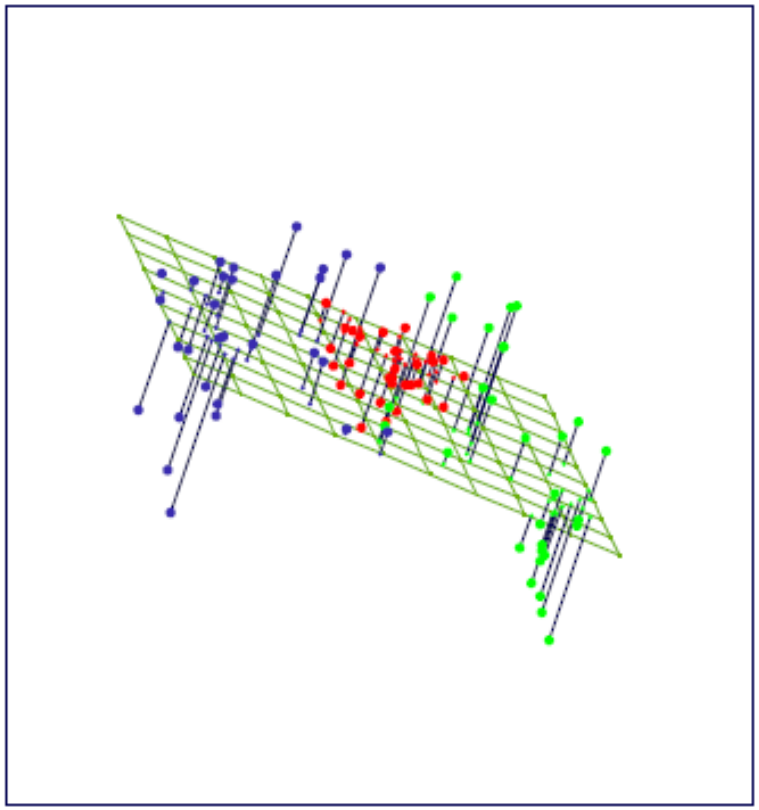


Tomado de: The Elements of Statistical Learning. Hastie, et al.

Introducción a la Reducción de Dimensiones

Principio Fundamental

Esto separa los métodos en lineales y no-lineales, dependiendo de la estructura del *manifold*.



Análisis de Componentes Principales: PCA

PCA es el método lineal más conocido y busca transformaciones a d nuevas variables no correlacionadas que se parezcan lo más posible a los datos originales (en p dimensiones).

Definición geométrica: Pasando de $\mathbb{R}^2$ a $\mathbb{R}$.
De dos variables (X_1, X_2) se pasa a la nueva variave Z, la cual es una transformación lineal:

$$Z = w_1 X_1 + w_2 X_2$$

Escribiendo $\mathbf{w} = (w_1, w_2)$, entonces los datos transformados se escriben como:

$$\mathbf{Z} = \mathbf{X}\mathbf{w} \in \mathbb{R}^n$$

Análisis de Componentes Principales: PCA

PCA es el método lineal más conocido y busca transformaciones a d nuevas variables no correlacionadas que se parezcan lo más posible a los datos originales (en p dimensiones).

Definición geométrica: Pasando de $\mathbb{R}^2$ a $\mathbb{R}$.
De dos variables (X_1, X_2) se pasa a la nueva variave Z, la cual es una transformación lineal:

$$Z = w_1 X_1 + w_2 X_2$$

Escribiendo $\mathbf{w} = (w_1, w_2)$, entonces los datos transformados se escriben como:

$$\mathbf{Z} = \mathbf{X}\mathbf{w} \in \mathbb{R}^n$$

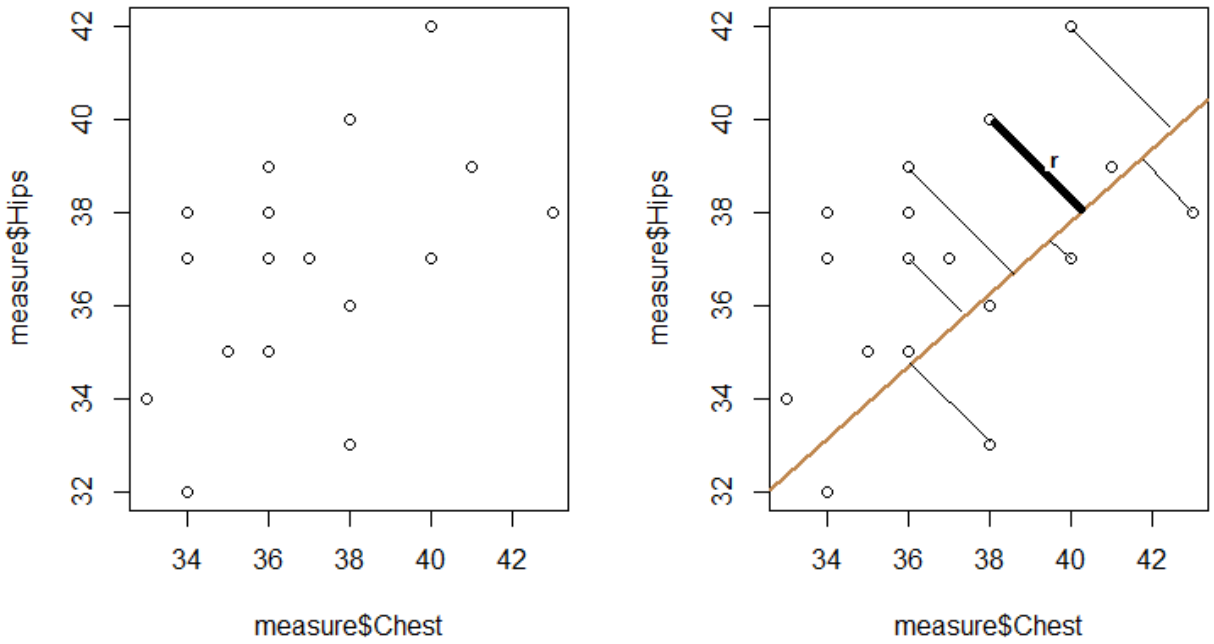
Nota: Si $X_1 = 0$ y $X_2 = 0$ entonces $Z = 0$, por lo que para facilitar el problema se suponene que los datos están centrados. Esto no afecta el resultado.

Análisis de Componentes Principales: PCA

Se busca el vector **w** que permita poner los datos en una sóla línea, y que al poner estos puntos en el espacio original $\mathbb{R}^p$, se pierde la menor información posible.

Geométricamente: Encontrar **w** tal que

$$\mathbf{w} = \arg\min_{w_1, w_2} \sum_{i=1}^{n} r_i^2$$

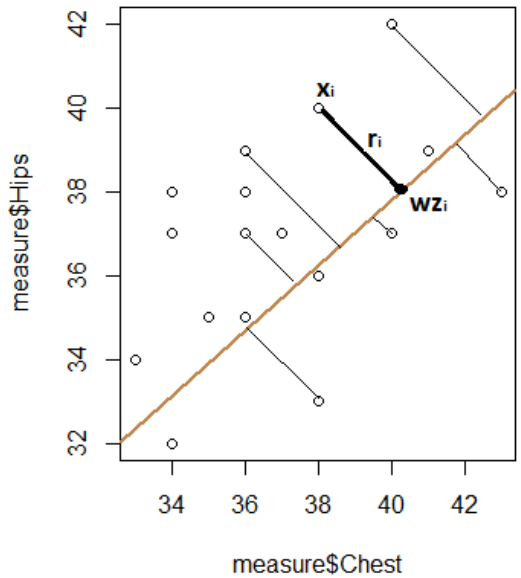


Análisis de Componentes Principales: PCA

Es claro que los puntos más cercanos son las proyeccciones sobre la recta $h\mathbf{w}$, para cualquier constante h. Con esto, los nuevos puntos son:

$$x_i \rightarrow \frac{\mathbf{w}x_i^T\mathbf{w}}{\|w\|^2} = \frac{\mathbf{w}z_i}{\|w\|^2}$$

Para que el problema quede bien definido, se impone la restricción $\|w\|^2 = 1$.



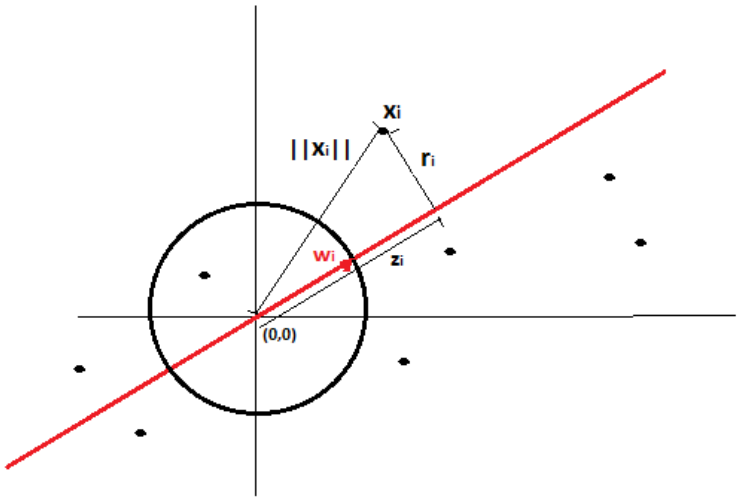
Análisis de Componentes Principales: PCA

La reformulación del problema queda:

$$\mathbf{w} = \arg\min_{\|w\|^2=1} \sum_{i=1}^{n} r_i^2$$

Sin embargo, por el teorema de Pitágoras, $\sum_{i=1}^{n} \|x_i\|^2 = \sum_{i=1}^{n} z_i^2 + r_i^2$, por lo que:

$$\mathbf{w} = \arg\max_{\|w\|^2=1} \sum_{i=1}^{n} (\mathbf{w}^t x_i)^2 = \arg\max_{\|w\|^2=1} \sum_{i=1}^{n} z_i^2$$



Análisis de Componentes Principales: PCA

Dado que el promedio de los $z_i = \bar{Z} = 0$, maximizar $\sum_{i=1}^{n} z_i^2$ es equivalente a maximizar $\hat{\mathrm{Var}}(Z) = s_z^2$:

$$\mathbf{w} = \arg\max_{\|w\|^2=1} \frac{1}{n} \sum_{i=1}^{n} z_i^2 = \arg\max_{\|w\|^2=1} s_z^2$$

Análisis de Componentes Principales: PCA

Dado que el promedio de los $z_i = \bar{Z} = 0$, maximizar $\sum_{i=1}^{n} z_i^2$ es equivalente a maximizar $\hat{\mathrm{Var}}(Z) = s_z^2$:

$$\mathbf{w} = \arg\max_{\|w\|^2=1} \frac{1}{n} \sum_{i=1}^{n} z_i^2 = \arg\max_{\|w\|^2=1} s_z^2$$

Formulación algebráica y estadística:

Para un vector aleatorio $X = (X_1, \cdots, , X_p)$, la media y matriz de varianza-covarianza se definen como:

$$\begin{aligned} \mathbb{E}(X) &= \mu = (\mu_1, \cdots, \mu_p) \\ \mathrm{Var}(X) &= \Sigma = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_1, X_p) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_1, X_p) & \cdots & \cdots & \mathrm{Var}(X_p) \end{pmatrix} \end{aligned}$$

Análisis de Componentes Principales: PCA

Formulación algebráica y estadística:

Las versiones estimadas con los datos de la media y la varianza son:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S_x = \begin{pmatrix} s_1^2 & s_{1,2} & \cdots & s_{1,p} \\ s_{1,2} & s_2^2 & \cdots & s_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1,p} & \cdots & \cdots & s_p^2 \end{pmatrix} = \frac{1}{n-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}$$

Donde $\tilde{\mathbf{X}}$ es la matriz de datos centrados. Es fácil determinar que si $\mathbf{Z} = \mathbf{X}\mathbf{w}$,

$$s_z^2 = \mathbf{w}^T S_x \mathbf{w}$$

Con lo que los pesos (coordenadas) del primer componente principal se encuentran resolviendo:

$$\mathbf{w} = \underset{\|w\|^2=1}{\arg\max}\, s_z^2 = \underset{\|w\|^2=1}{\arg\max} = \mathbf{w}^T S_x \mathbf{w}$$

Análisis de Componentes Principales: PCA

Formulación algebráica y estadística:

Las condiciones de optimlidad de este problema implican que:

$$S_x \mathbf{w} = \lambda \mathbf{w} \quad \Rightarrow \quad s_z^2 = \lambda$$

Análisis de Componentes Principales: PCA

Formulación algebráica y estadística:

Las condiciones de optimlidad de este problema implican que:

$$S_x \mathbf{w} = \lambda \mathbf{w} \quad \Rightarrow \quad s_z^2 = \lambda$$

Solución al problema:

Como se quiere maximizar la varianza de Z, la solución a nuestro problema es que $\mathbf{w}$ es el primer vector propio (*eigenvector*) de la matriz S_x, y λ es el valor propio (*eigenvalue*) asociado.

Encontrar los componentes principales, implica encontrar la descompocisión espectral de la matriz S_x (o también de $\tilde{\mathbf{X}}$).

Análisis de Componentes Principales: PCA

Formulación algebráica y estadística:

Las condiciones de optimlidad de este problema implican que:

$$S_x \mathbf{w} = \lambda \mathbf{w} \quad \Rightarrow \quad s_z^2 = \lambda$$

Solución al problema:

Como se quiere maximizar la varianza de Z, la solución a nuestro problema es que $\mathbf{w}$ es el primer vector propio (*eigenvector*) de la matriz S_x, y λ es el valor propio (*eigenvalue*) asociado.

Encontrar los componentes principales, implica encontrar la descompocisión espectral de la matriz S_x (o también de $\tilde{\mathbf{X}}$).

Ejemplo en R:

Código adjunto en archivo `ejemplo1.txt`

Análisis de Componentes Principales: PCA

Reducción a más de un Componente Principal

Si se quiere pasar de $\mathbb{R}^p$ a $\mathbb{R}^d$ con $d \geq 2$, entonces, dado que ya se tiene el primer componente, el segundo debe ser ortogonal al primero.

Haciendo procedimientos imilares, es fácil determinar que los componentes siguientes están asociados a los *eigenvectors*(*values*) que siguen en magnitud.

Análisis de Componentes Principales: PCA

Reducción a más de un Componente Principal

Si se quiere pasar de $\mathbb{R}^p$ a $\mathbb{R}^d$ con $d \geq 2$, entonces, dado que ya se tiene el primer componente, el segundo debe ser ortogonal al primero.

Haciendo procedimientos imilares, es fácil determinar que los componentes siguientes están asociados a los *eigenvectors*(*values*) que siguen en magnitud.

Descompocisión espectral:
Se pueden encontrar tods los p vectores y valores propios, de forma tal que:

$$S_x = \lambda_1 \mathbf{w}_1 \mathbf{w}_1^T + \cdots + \lambda_p \mathbf{w}_p \mathbf{w}_p^T$$

Con $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Los respectivos componentes son:

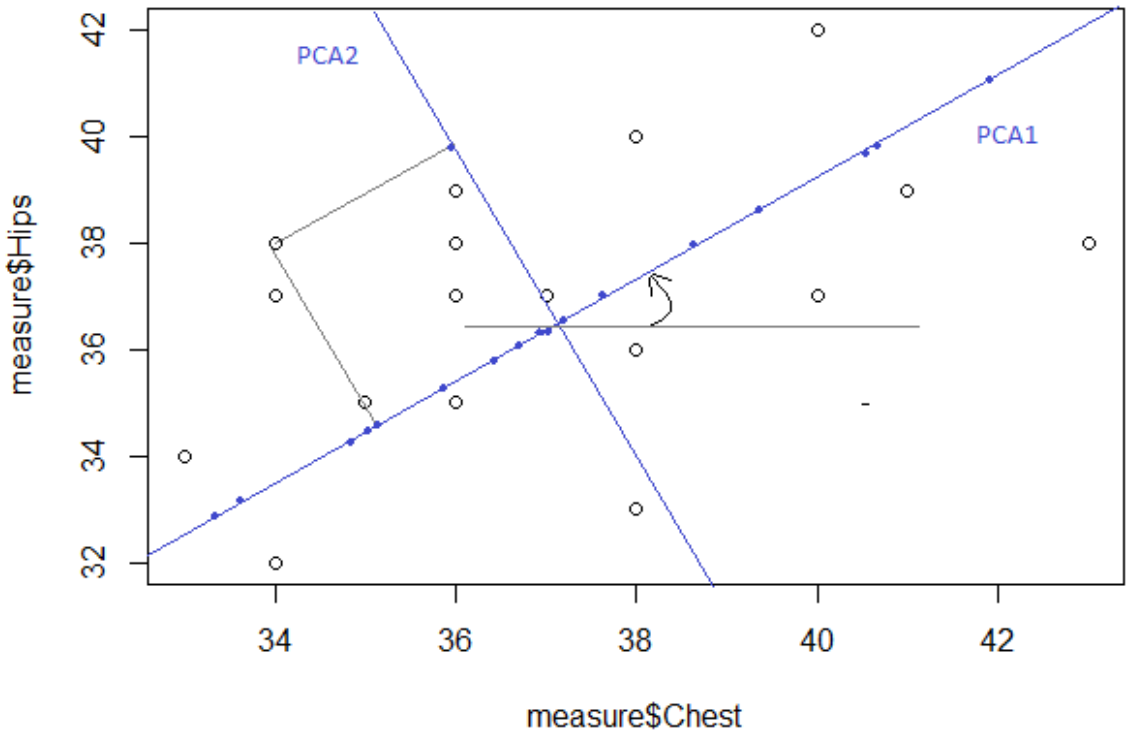
$$Z_j = \mathbf{X}\mathbf{w}_j \quad \Rightarrow \quad \mathbf{Z} = \mathbf{X}\mathbf{W}$$

Donde $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_d]$.

Análisis de Componentes Principales: PCA

Reducción a más de un Componente Principal

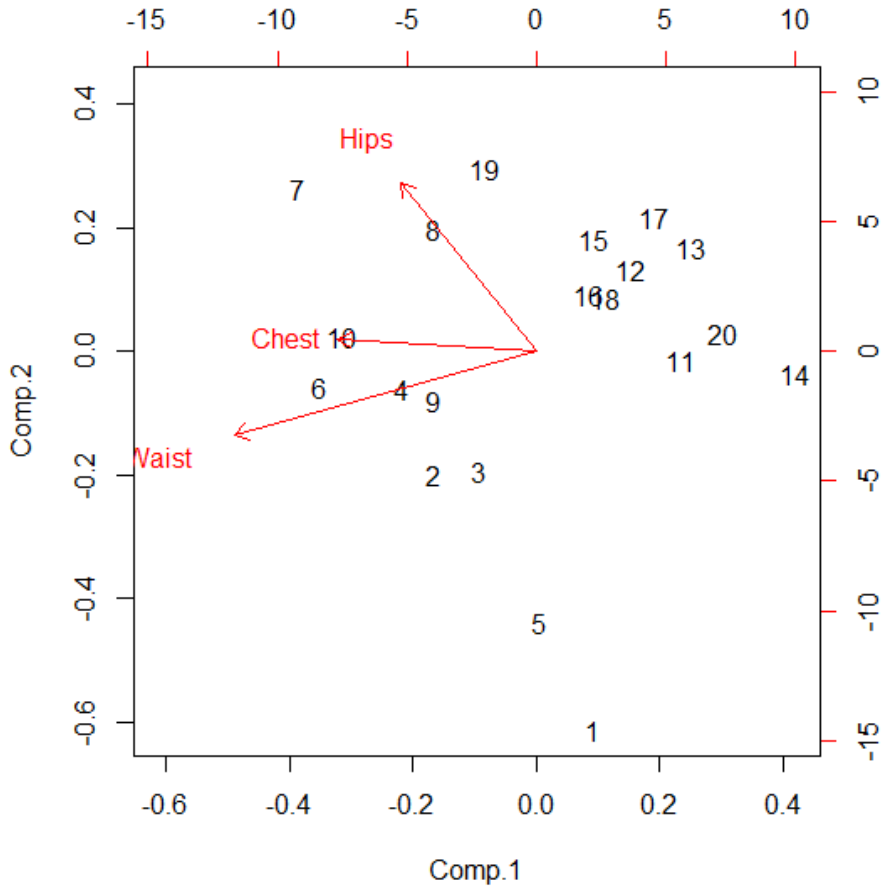
Cuando $d = p$, la solución de los componentes principales se convierte en una rotación del espacio, es decir, encontrar unas nuevas coordenadas óptimas. (Ejemplo 2).



Análisis de Componentes Principales: PCA

Biplot

Si se visualiza en dos dimensiones, es interesante ver el peso de las variables originales



Análisis de Componentes Principales: PCA

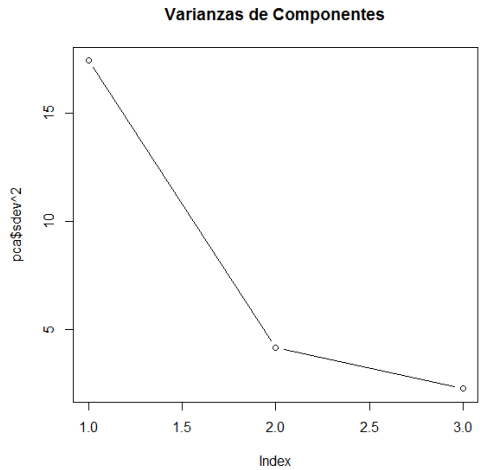
Número de Componentes

Para determinar cuántos componentes usar, se debe tener una mediad de qué tanta información se está perdiendo. Para esto se puede usar la *Varianza total*:

$$VT = trace(S_x) = s_1^2 + \cdots + s_p^2 = \lambda_1 + \cdots + \lambda_p$$

De esta forma, el porcentaje de vaarianza total explicada por los primeros d componentes es:

$$P^d = \frac{\sum_{j=1}^{d} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$$



Análisis Factorial Exploratorio: FA

Análisis Factorial: FA

En este caso no se busca una recomposición geométrica de las observaciones, sino que se permite error de medición bajo un modelo estadístico.

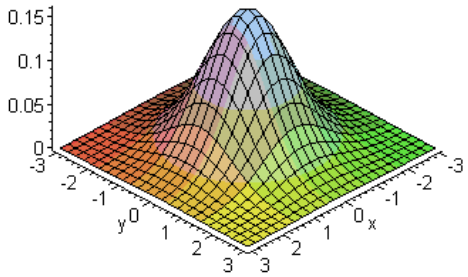
Modelo Estadístico: Se asume que cada observación $x_i \in \mathbb{R}^p$ es una realización de una variable aleatoria

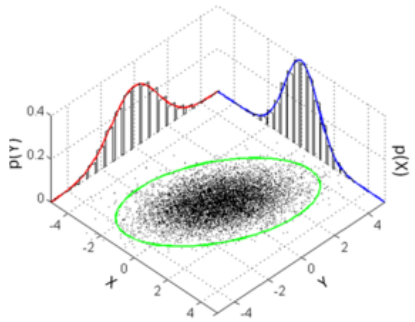
$$\mathbf{x} \sim Normal_p(\mu, \Sigma)$$

Esto quiere decir que $\mathbf{x} \sim g(\mathbf{x})$, donde:

$$g_j(\mathbf{x}_i) = (2\pi)^{-\frac{p}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \mu_j)^T \Sigma^{-1} (\mathbf{x}_i - \mu_j)\right)$$

Que es una normal multivariada. En el caso bivariado se ve asi:





Análisis Factorial Exploratorio: FA

Análisis Factorial: FA

Modelo Estadístico: $\mathbf{x} \sim Normal_p(\mu, \Sigma)$

Para simplificar el problema, se asume que $\mu = 0$ (esto no cambia los resultados).

La variación común entre las varaibles (dimensiones) se explica por la existencia de variables latentes (factores) que afectan el resultado de las observaciones. El modelo se puede representar como:

$$\mathbf{x} = \mu + \Lambda f + u = \Lambda f + u$$

Análisis Factorial Exploratorio: FA

Análisis Factorial: FA

Modelo Estadístico: $\mathbf{x} \sim Normal_p(\mu, \Sigma)$

Para simplificar el problema, se asume que $\mu = 0$ (esto no cambia los resultados).

La variación común entre las varaibles (dimensiones) se explica por la existencia de variables latentes (factores) que afectan el resultado de las observaciones. El modelo se puede representar como:

$$\mathbf{x} = \mu + \Lambda f + u = \Lambda f + u$$

Ejemplos:

- El atributo de inteligencia es difícil de medir directamente. Sin embargo, hay indicios (mediciones indirectas) que estn conrrelacionadas con inteligencia, por ejemplo, el resultado en pruebas.
- El desarrollo de los pases. No se puede medir directamente, pero creemos que hay unos más desarrollados que otros por diferentes variables: educación, ingreso, institucionalidad, etc.

Análisis Factorial Exploratorio: FA

Análisis Factorial: FA

Supongamos el caso más sencillo en donde sólo existe un factor y $p = 3$. Por ejemplo, el resultado de una persona en tres diferentes pruebas: verbal (X_1), matemática (X_2) y abstracta (X_3). Según esto, existe un factor común que mueve las tres variables:

$$X_1 = \lambda_1 f + u_1$$
$$X_2 = \lambda_2 f + u_2$$
$$X_3 = \lambda_3 f + u_3$$

Advertencia: Para resolver el problema, se asume que $f \sim Normal(0, 1)$, de otra forma el factor no sería identificable. Consecuencias?

Además, se asume que cada $u_j \sim Normal(0, \phi_j)$ independientes entre sí.

Esto implica que $X_j \sim Normal(0, \sigma_j^2)$ y

$$\sigma_j^2 = \lambda_j^2 + \phi_j$$

Análisis Factorial Exploratorio: FA

Análisis Factorial: FA

$$X_1 = \lambda_1 f + u_1$$
$$X_2 = \lambda_2 f + u_2$$
$$X_3 = \lambda_3 f + u_3$$

Además, se sabe que:

$$\text{Cov}(X_i, X_j) = \sigma_{ij} = \lambda_i \lambda_j$$

Con estas condiciones, conociendo las varianzas σ_i^2 y las covaianzas σ_{ij}, es posible aproximar el valor de las cargas (λ) y las varianzas de los errores (ϕ).

Ejemplo sencillo: En el problema de las 3 pruebas y con $n = 100$ individuos:

$$\hat{\Sigma} = \begin{pmatrix} 1 & & \\ 0.4 & 1 & \\ 0.32 & 0.2 & 1 \end{pmatrix}$$

Encuentre las cargas (comunalidades) y los errores específicos.

Análisis Factorial Exploratorio: FA

Análisis Factorial: FA

En el ejemplo anterior, el modelo implicaría que un sólo facor explica todas las covarianzas. Qué pasaría si todas las covarianzas son cero? El resultado sería que no hay personas más inteligentes que otras.

En general es bueno extender a más factores simpre y cuando se puedan estimar las cargas λ_{jk} (comunaldades) y las varianzas específicas (ϕ_j). Si se tienen m factores ($m < p$), entonces:

$$\Lambda = [\lambda_{jk}] = \begin{pmatrix} \lambda_{1,1} & \cdots & \lambda_{1,m} \\ \vdots & \ddots & \vdots \\ \lambda_{p,1} & \cdots & \lambda_{p,m} \end{pmatrix}$$

Y por lo tanto:

$$\mathbf{X} = \Lambda \mathbf{f} + \mathbf{U}$$

donde tanto $\mathbf{X}$ com $\mathbf{U}$ son vectores aleatorios de longitud p.

$\mathbf{f} = (f_1, \ldots, f_m)^T$ es el vector de factores, donde $f_j \sim Normal(0, 1)$.

Análisis Factorial Exploratorio: FA

Análisis Factorial: FA

$$\Sigma = \Lambda\Lambda^T + \Phi$$

Los parámetros de interés Λ y Φ se pueden estimar. el método más común es **máxima verosimilitud**, dado que permite hacer pruebas de hipótesis.

Cada factor se pude interpretar a partir de las cargas corespondientes (como en PCA).

Análisis Factorial Exploratorio: FA

Análisis Factorial: FA

$$\Sigma = \Lambda\Lambda^T + \Phi$$

Los parámetros de interés Λ y Φ se pueden estimar. el método más común es **máxima verosimilitud**, dado que permite hacer pruebas de hipótesis.

Cada factor se pude interpretar a partir de las cargas corespondientes (como en PCA).

Rotaciones de factores:

Como el modeo es insensible a las rotaciones de los m factores, es posible jugar con estas rotaciones para que sean más fáciles de interpretar. Esto es, si cada factor afecta un grupo de variables específico, entonces es más fácil decir qué significa cada factor. Existen dos formas de encontrar las rotaciones:

1. Ortogonales (Varimax, Quartimax)
2. Oblicuas (Promax, Oblimin)