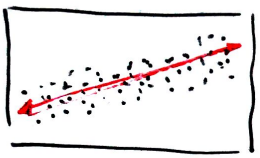
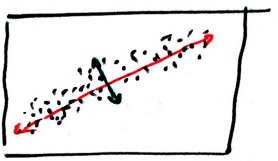
## Interpretacion Geometrica

Supongase un conjunto de datos definido por dos variables $(X_1, X_2)$ el vector que define la primera componente principal $(Z_1)$ es aquel en la direccion de mayor varianza, la direccion en la que las observaciones presentan mayor varianza. la proyeccion de cada observacion sobre esa direccion es el valor de la primera componente





la segunda componente es la de mayor varianza tambien y que no tenga correlacion con la primera, es decir, ortogonal

## Calculo de componentes

Cada Z se obtiene por combinacion lineal de las variables originales $(X_1, X_2, X_3, \cdots X_p)$

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \cdots + \phi_{p1} X_p$$

$$\sum_{j=1}^{p} \underbrace{\phi_{j1}^2}_{=1} = 1 \quad \text{. combinacion lineal normalizada}$$

↕ loadings

Proceso:

- Centrar las variables: Se resta a cada valor la media de la variable $\Rightarrow$ media = 0 . std = 1
- Resolver problema optimizacion para loadings para maximizar varianza calculo de eigenvalues y eigenvectores de la matriz de covarianza

$\Rightarrow$ Una vez se calcula la primera $Z_1$, se calcula la segunda agregando la condicion de que $Z_2$ y $Z_1$ no pueden estar correlacionadas - Ortogonales. el orden de importancia lo da el eigenvalue

## # PCA_ALGORITHM ##

Este sirve para reducir la dimension del dataset basado en la relacion existente entre los datos

Evaluacion previa

correlograma $$S(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x)\,Var(y)}}$$

Varianza

$$Var(x) = \sum_{x \in R_e} (x - \mu)^2$$

covarianza

$$cov(x,y) = \frac{1}{2n^2} \sum_{i=2}^{n} \sum_{j=1}^{n} (x_i - x_j)(y_i - y_j)$$

- Permite la reduccion de dimensiones bajando la complejidad del dataset

Supongo un espacio muestral de n muestras y p variables
PCA permite encontrar unos factores subyacentes z ($z < p$) que explican aproximadamente lo mismo que p. estas z variables reciben el nombre de componentes principales

- Cabe resaltar que es importante contar con las variables originales

PCA se aplica para visualizacion y el preprocesado de predictores previo al ajuste de modelos supervisados

Conceptos base (Algebra lineal)

- eigenvalues
es el factor de multiplicacion del eigenvectors.

- eigenvectors
vectores que al multiplicarlos por la matriz dan el mismo vector o un multiplo entero

- los eigenvectores son ortogonales

- Se recomienda remover outlayers

Varianza explicada

Cuanta informacion se pierde al realizar PCA

Varianza total en el set de datos $\sum_{J=1}^{p} Var(X_j) = \sum_{J=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$ (1)

Varianza explicada por la componente m $\frac{1}{n} \sum_{i=1}^{n} z_{im}^2 = \frac{1}{n} \sum_{i=1}^{n} \left( \sum_{J=1}^{p} \phi_{jm} x_{ij} \right)^2$ (2)

la proporcion de varianza explicada es el ratio (2)/(1)

Numero de componentes principales

Se estima con la varianza explicada acumulada mirando a partir de que componente deja de aumentar sustancialmente.