Universidad de los Andes

Análisis espacial de datos y sus aplicaciones en Python

Profesores: Germán González

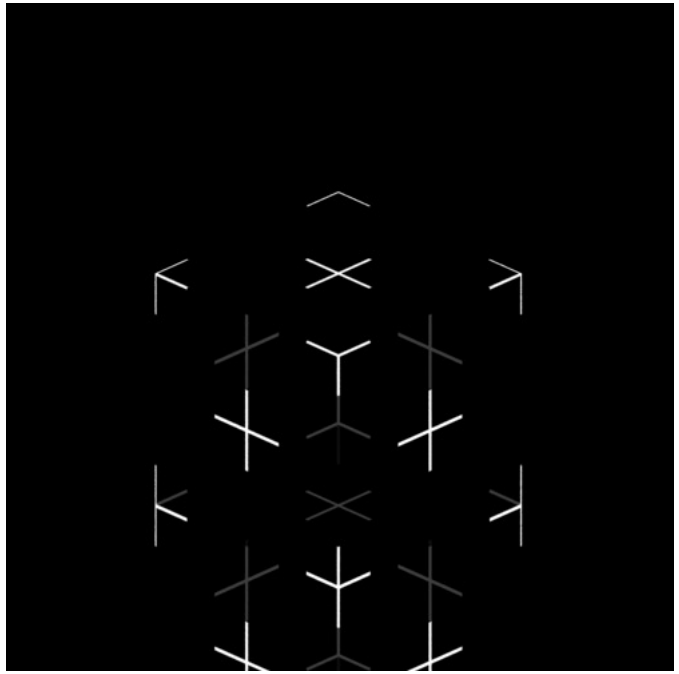
Sesión 9: Dimensionalidad

Universidad de los Andes

Índice

Problema

- Número de dimensiones se puede equiparar al número de variables o características (features) que estemos utilizando.
- A medida que aumenta el número de dimensiones, los modelos requiere de más capacidad de computo para calibrar.
- La variabilidad de la distancia disminuye exponencialmente con el número de dimensiones.



Soluciones:

1) **Selección de variables**: Selección de variables en un gran conjunto de posibles variables explicativas

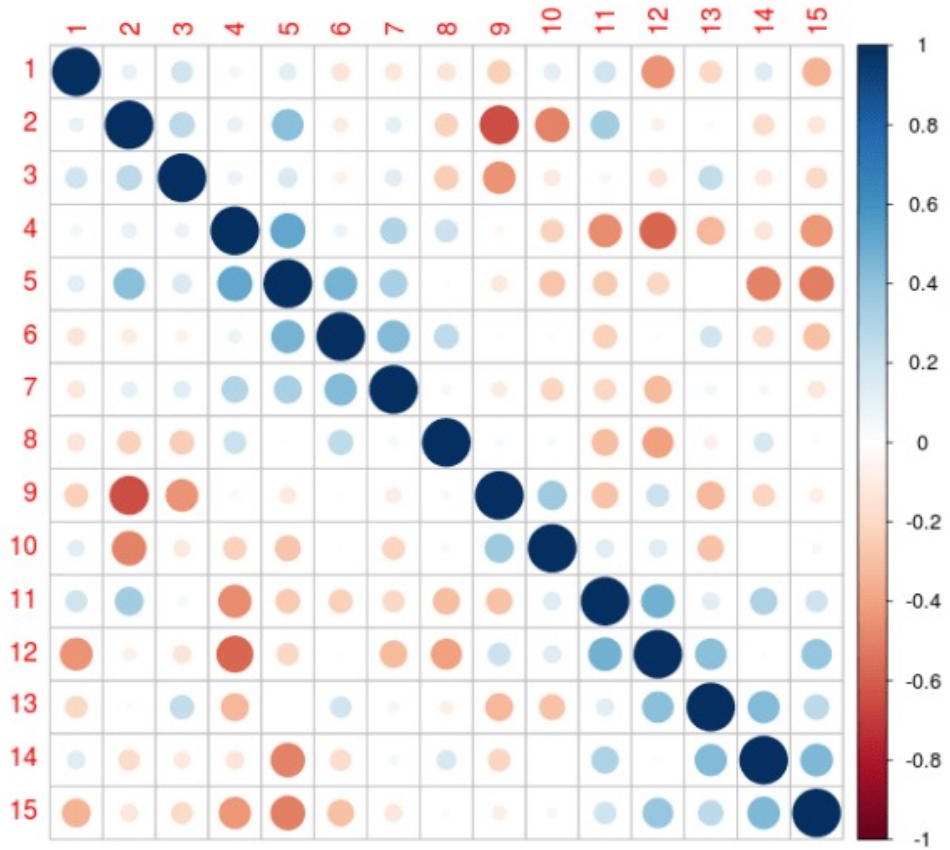
2) **Reducción de dimensionalidad:** generar variables independientes que expliquen lo mejor posible un conjunto de variables dependientes

Selección de variables

Objetivo: Identificar y seleccionar, entre un conjunto de variables predictoras, aquellas que están más relacionados con la variable respuesta y así crear el mejor modelo

¿Cómo saber cuáles variables debo agregar y cuáles debo quitar?

¿Qué tipo de modelos me pueden ayudar para abordar este tipo de problemas?



Índice

Stepwise Selection

La selección de un modelo en un subconjuntos de modelos, ya sea utilizando o Validación Cruzada, es una tarea dispendiosa pues requiere 2^k subconjuntos de k regresores.

Stepwise es un algoritmo que a partir de iteraciones acota los posibles subconjuntos de modelos y no explora todas las combinaciones de estos. (Es útil en casos específicos)

I. Forward
II. Backward

Forward

i. Iniciar con una regresión que solo contenga el **intercepto**.
ii. Incluya un primer regresor con el p-valor más pequeño.
iii. Incluya un segundo regresor con el p-valor más pequeño.
iv. Realice este procedimiento hasta que la inclusión de una nueva variable explicativa ya no sea significativa.

Generalmente, se utilizan criterios de información o Validación Cruzada para seleccionar el mejor modelo a partir de la secuencia escalonada de modelos.

Backward

i. Iniciar con una regresión que incluya todas las variables
ii. Pase a un modelo de K-1 variables, eliminando el regresor con el p-valor más grande.
iii. Pase a un modelo de K-2 variables, eliminando el regresor con el p-valor más grande.
iv. Realice este procedimiento hasta que el p-valor de la variable excluida sea estadísticamente significativo.

Generalmente, se utilizan criterios de información o validación cruzada para seleccionar el mejor modelo a partir de la secuencia escalonada de modelos.

Stepwise Selection

Ventajas:

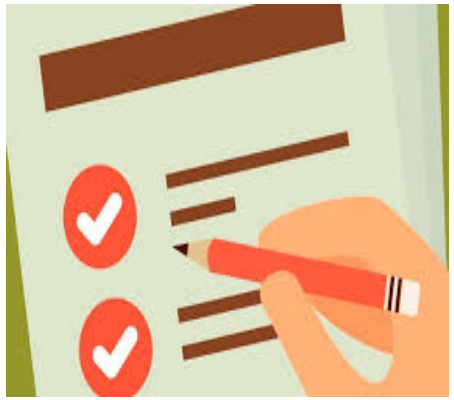
i. Buena aproximación en casos muy específicos.
ii. Reduce tiempos y costos computacionales.

Desventajas:

I. No garantiza que se seleccione el mejor modelo de entre todos los posibles ya que no se evalúan todas las posibles combinaciones.

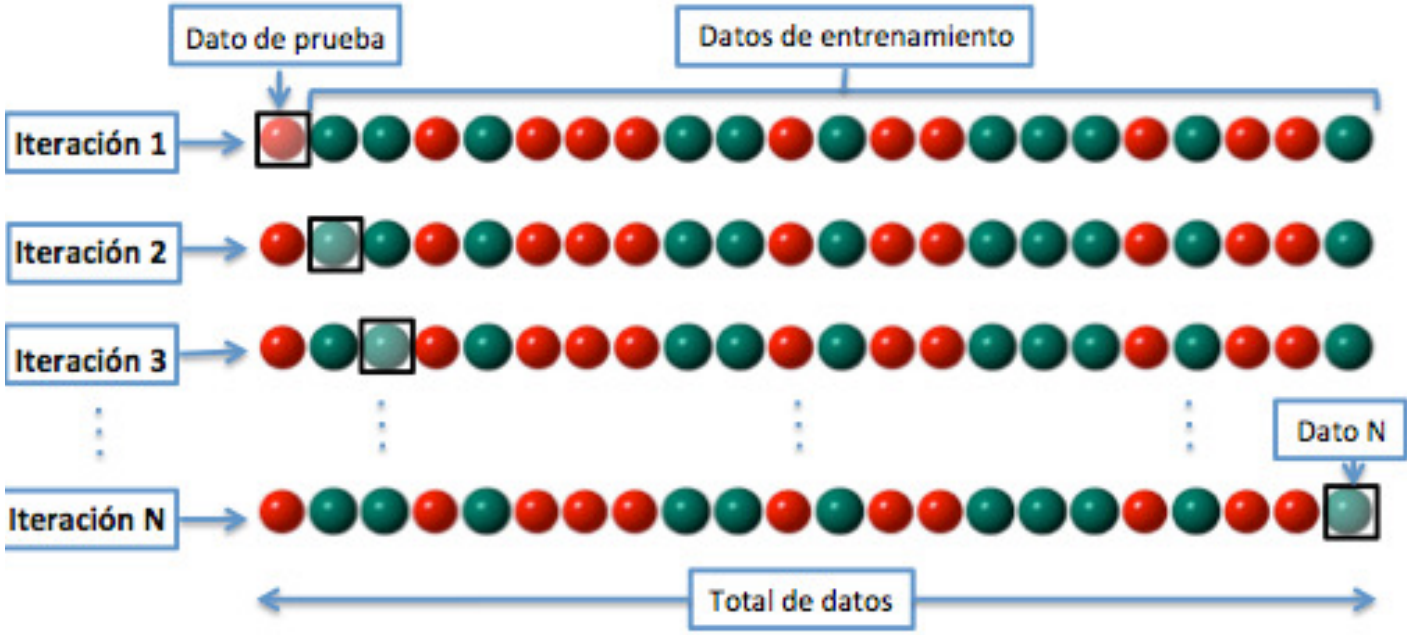
Índice

Validación Cruzada



Técnica para evaluar los resultados de un modelo estadístico y garantizar que el análisis es independiente de la partición entre datos de entrenamiento y prueba.

1. Validación cruzada dejando un dato por fuera
2. Validación cruzada T−Fold
3. Validación cruzada VS SIC

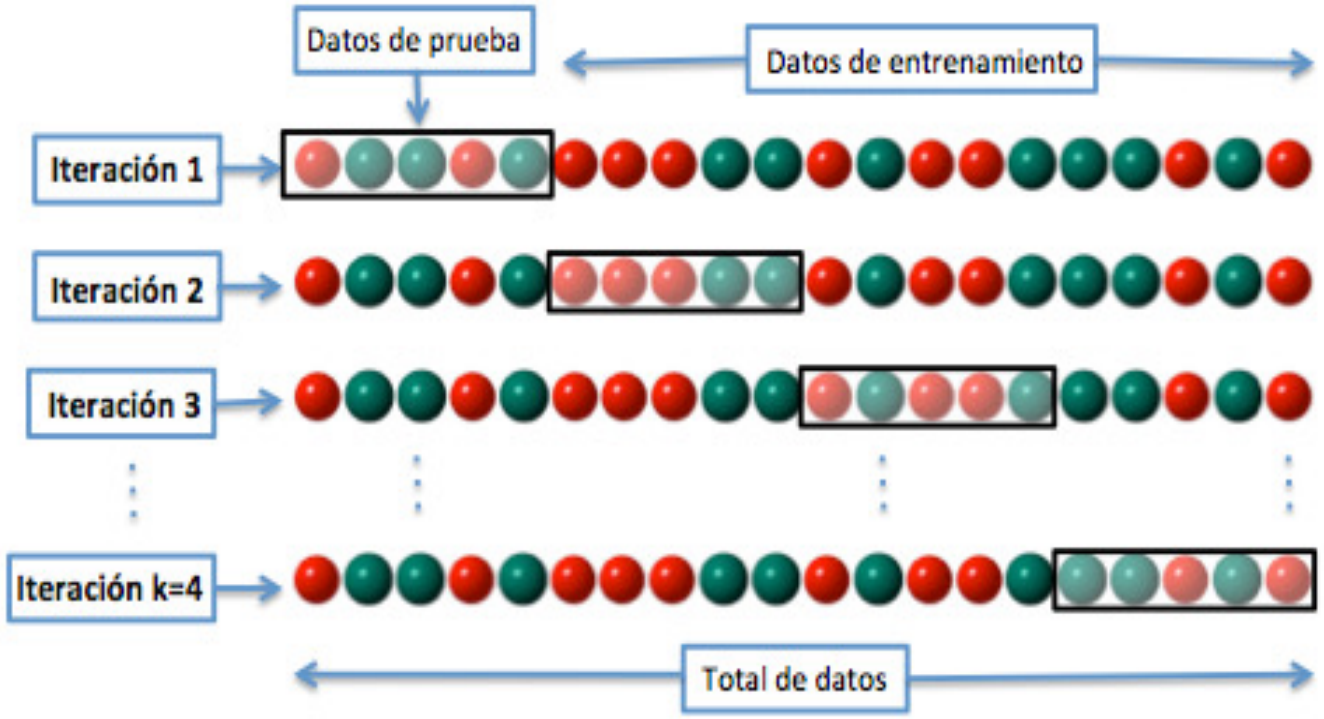


Validación cruzada dejando un dato por fuera

i. Estime el primer modelo usando todas las observaciones, a excepción de la primera. Utilice el modelo para predecir la observación eliminada y calcule el error de predicción al cuadrado.

ii. Repita este proceso para cada observación, y promedie los errores al cuadrado al predecir cada una observaciones eliminadas.

iii. Realice este procedimiento con todos los modelos y escoja el que dé menor promedio de errores cuadráticos



Validación cruzada T −fold

I. Divida en M muestras aleatorias la muestra original de tamaño T (M<T). Dada la partición M entrene el modelo sin los datos de esta muestra, y extraiga el error de predicción de esa partición.

II. Repita este proceso para cada división M y promedie los errores al cuadrado.

III. Realice este procedimiento con todos los modelos y escoja el que dé menor promedio de errores cuadráticos

(El óptimo de divisiones depende del número de datos. Un M grande con pocos datos sobre estima el error de prueba (varianza alta). Un M bajo subestima el error (sesgo alto))

* El autor recomienda M=10

Regularización – Contracción (Shrinkage)

Modificación a MCO, incluye todos las variables predictoras, pero contiene una penalización que fuerza a que las estimaciones de los coeficientes de regresión tiendan a cero. Es decir, tiene a minimizar la influencia de los predictores menos importantes.

- ***Ridge***: aproxima a cero los coeficientes de los predictores pero sin llegar a excluir ninguno.
- **Lasso**: aproxima a cero los coeficientes, llegando a excluir predictores.

Regularización – Contracción (Shrinkage)

Busca estimadores que pueden ser sesgados, pero con menor error cuadrático medio.

$$\hat{\beta} = (X'X)^{-1}X'Y \qquad \hat{\beta}_{Ridge} = (X'X + \lambda I)^{-1} X'Y$$

I. Si $\lambda = 0$ el estimador Ridge es idéntico al MCO

II. Si $\lambda = \infty$ el estimador de Ridge se contrae a 0.

λ no puede ser elegido por criterios de información, debido a que los regresores (K) se incluyen independientemente de λ. Para ello se selecciona un rango de valores de λ y se estima el cross-validation error resultante para cada uno, finalmente se selecciona el valor de λ para el que el error es menor y se ajusta de nuevo el modelo,

Regularización – Contracción (Shrinkage)

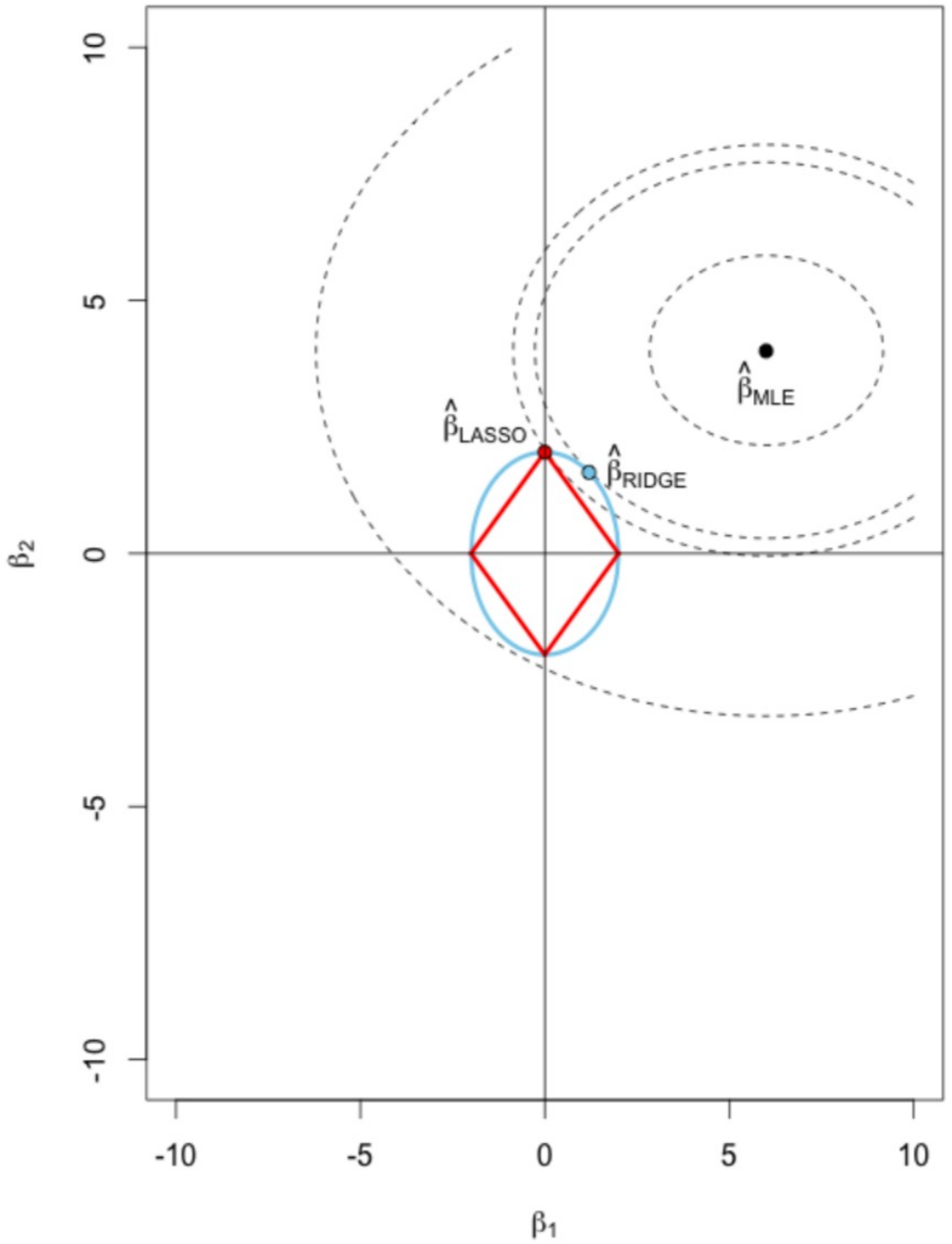
$$\hat{\beta}_{Lasso} = argmin_{\beta} \sum_{t=1}^{T} \left(y_t - \sum_{i} \beta_i x_{it} \right)^2$$

Sujeto a $\sum_{i=1}^{K} |\beta_i| \leq c$

Lasso al igual Ridge es una técnica de regresión lineal regularizada, con una diferencia en la penalización (norma) que trae consecuencias en la selección.

i. Ridge contrae
ii. Lasso contrae y selecciona.

Comparación de Lasso y Ridge



Índice

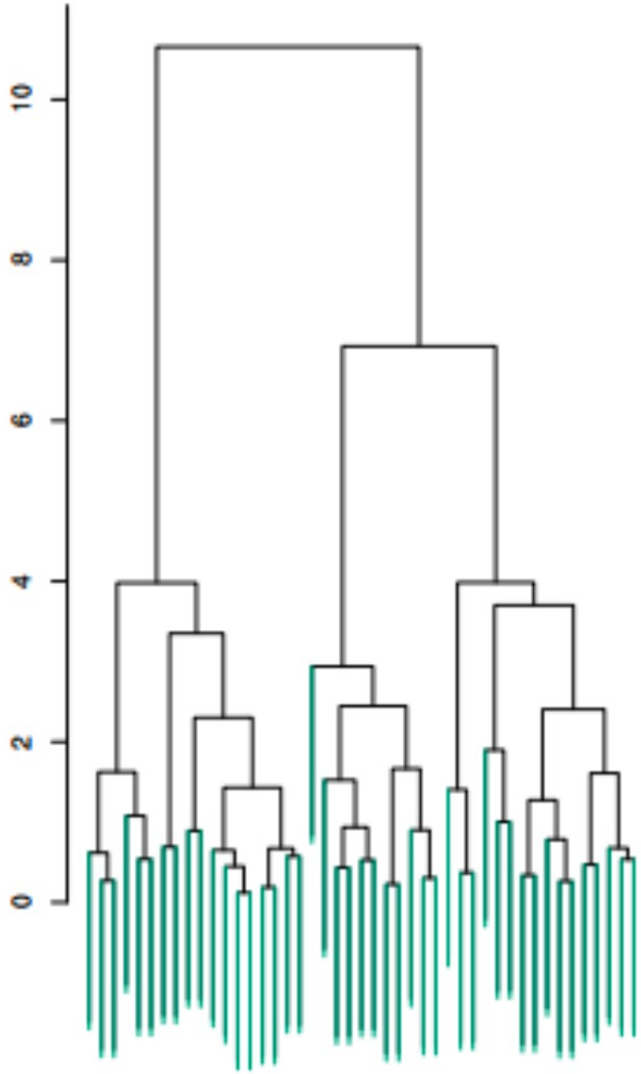
Universidad de los Andes

Arboles de decisión

- Las observaciones se representan como una hoja y se van fusionando en ramas hasta llegar al tronco del árbol

- La parte inferior del árbol lo componen las hojas, que representan cada uno de las observaciones en la base de datos.

- A medida que subimos por el árbol, las hojas se van fusionando formando ramas, que corresponden a observaciones similares entre sí

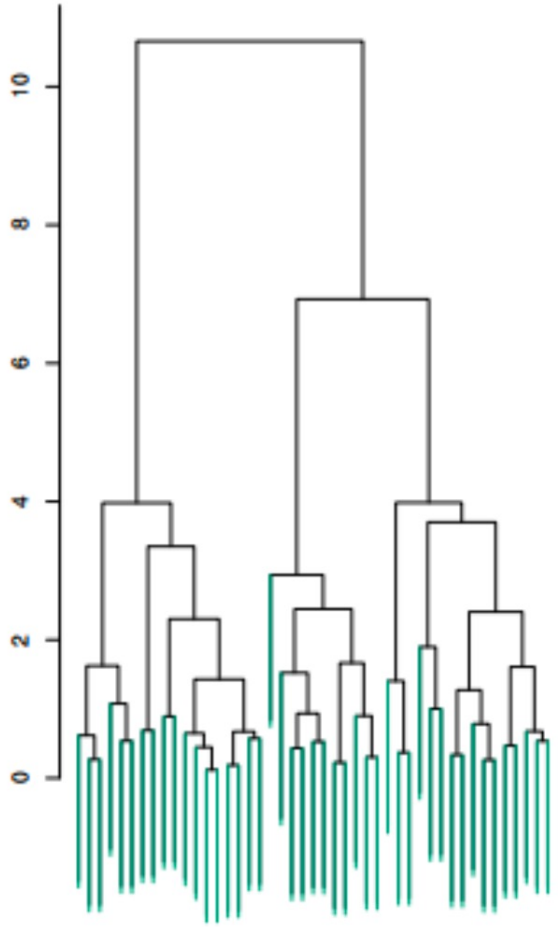


Universidad de los Andes

Arboles de decisión

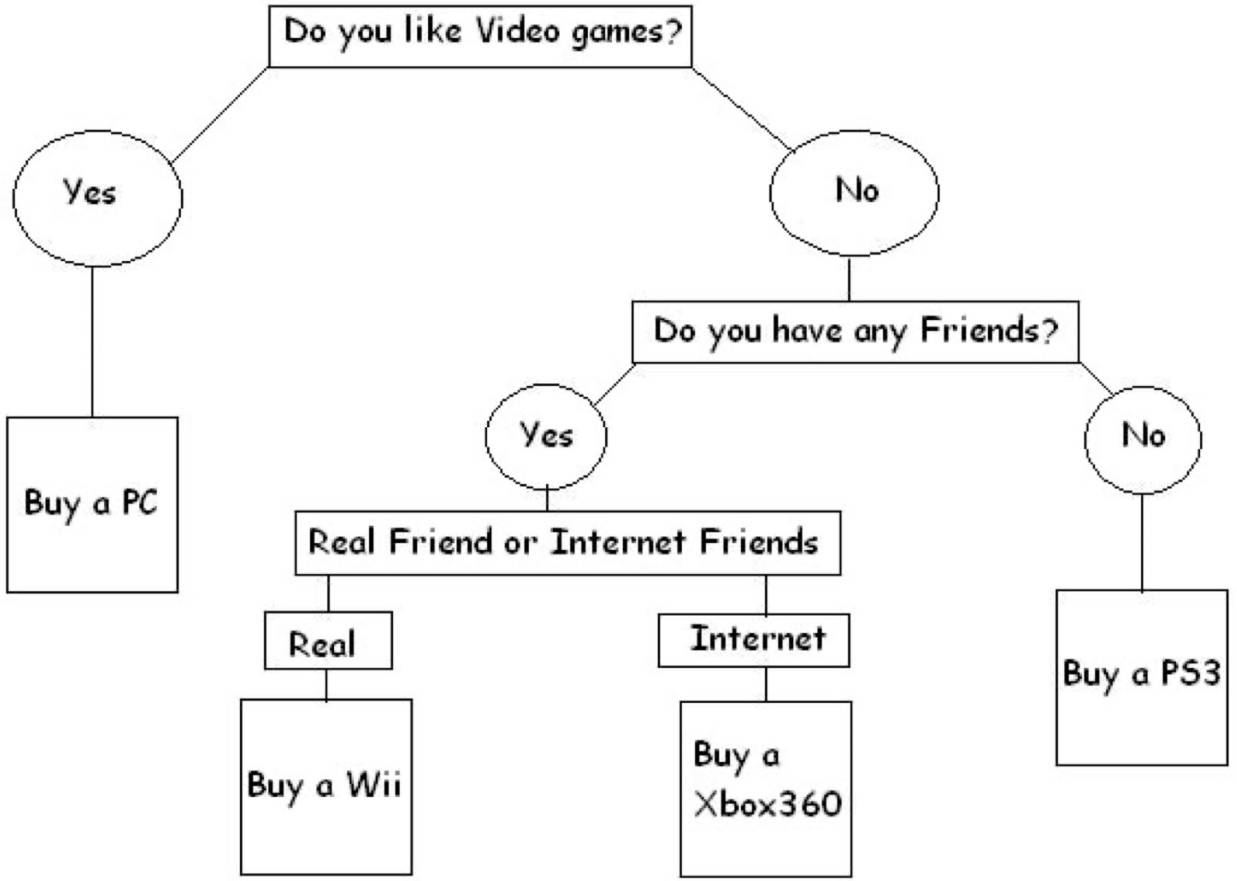
Al seguir subiendo por el dendograma, las mismas ramas se empiezan a fusionar o bien con otras ramas o con hojas

Entre más rapido ocurra una fusión (más abajo en el árbol) más parecidas seran los grupos de observaciones.



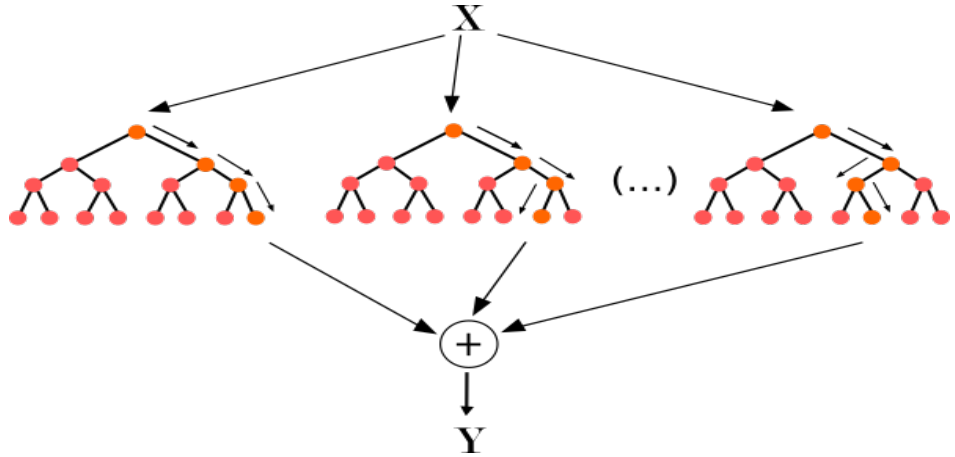
Universidad de los Andes

Ejemplo:



Gini

- Método aplicable únicamente para clasificadores de árboles.
- **Índice de Gini:** Sirve para entender la pureza de los nodos.



$$G = 1 - \sum_{k=1}^{k} \hat{P}_k{}^2$$

- $\hat{P}_k$ es la proporción de observaciones de entrenamiento que pertenecen a la clase k.

Un valor bajo indica que el nodo contiene predominantemente observaciones de una sola clase.

Aplicación: Encontrar cuáles son las características más importantes que mejoran la predicción en un determinado modelo.

Ejemplo

¿cuál variable explica mejor si un estudiante paso el examen? El grado o las horas estudiadas?

	Grado	Horas estudiadas	Paso el examen
0	Primero	>2h	Si
1	Primero	>2h	Si
2	Segundo	>2h	Si
3	Sexto	<2h	Si
4	Primero	>2h	No
5	Segundo	<2h	No

$$I_{\text{Gini}}(a) = \sum_{k \in M} P_{k,a} \cdot \text{Gini}(k)$$

- a is the feature
- M be the list of all categories in feature a
- $P_{k,a}$ is the fraction of category k in feature a

Grado

$$\text{Gini(Primero)} = 1 - \left(P_{No}^2 + P_{si}^2\right) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = 4/9$$

$$\text{Gini(Segundo)} = 1 - \left(P_{No}^2 + P_{si}^2\right) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 1/2$$

$$\text{Gini(Sexto)} = 1 - \left(P_{No}^2 + P_{si}^2\right) = 1 - (0)^2 - (1)^2 = 0$$

Gini(Grado) = (3/6)(4/9) + (2/6) (1/2) + (1/6)(0)= **0,38**

Horas estudiadas

$$\text{Gini(>2h)} = 1 - \left(P_{No}^2 + P_{si}^2\right) = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0{,}375$$

$$\text{Gini(<2h)} = 1 - \left(P_{No}^2 + P_{si}^2\right) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5$$

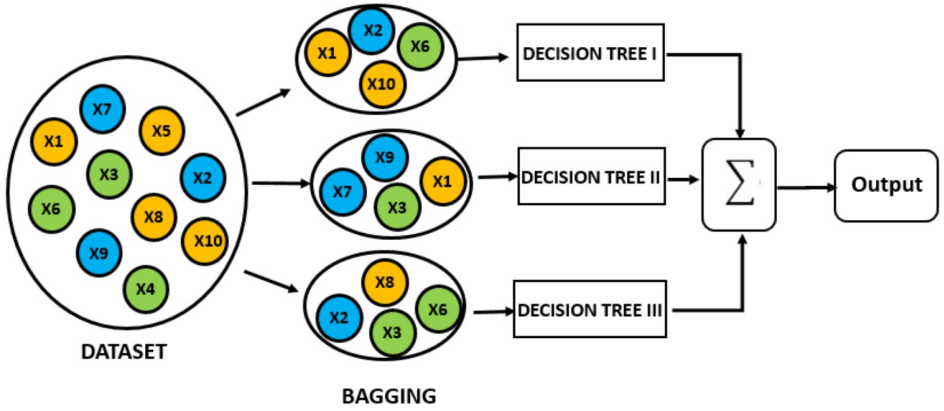
Gini(Horas) = (4/6)(0,375) + (2/6) (0,5) = 0,416

Random forest

Metodología que busca dividir la muestra de entrenamiento en diferentes **sub-muestras** con las cuales luego se construyen diferentes estimadores a partir de árboles de decisión a esas sub-muestras.

Un árbol inicia desde la raíz, y se extiende en diferentes particiones que tienen como objetivo **dividir la muestra** en dos o más conjunto de datos dependiendo de las **características de cada observación.**

Objetivo: crear **grupos homogéneos** que se separan a partir de características relevantes de la muestra.

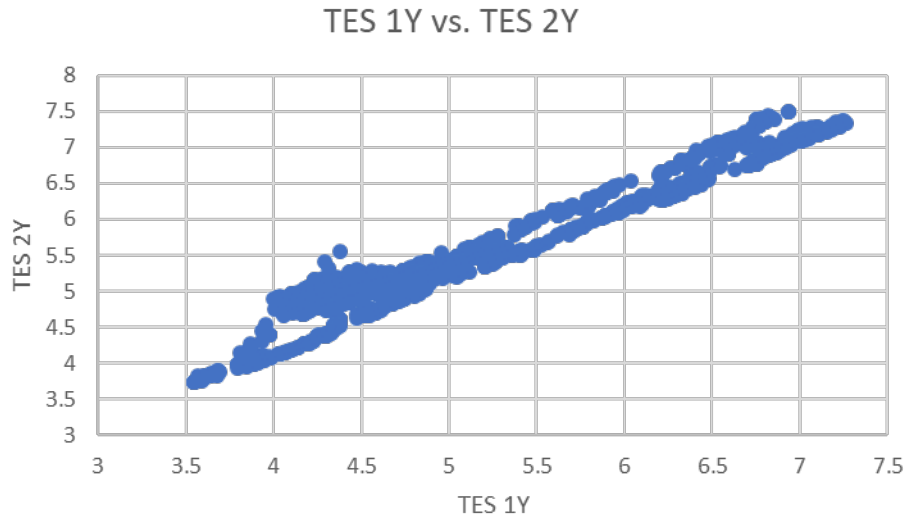


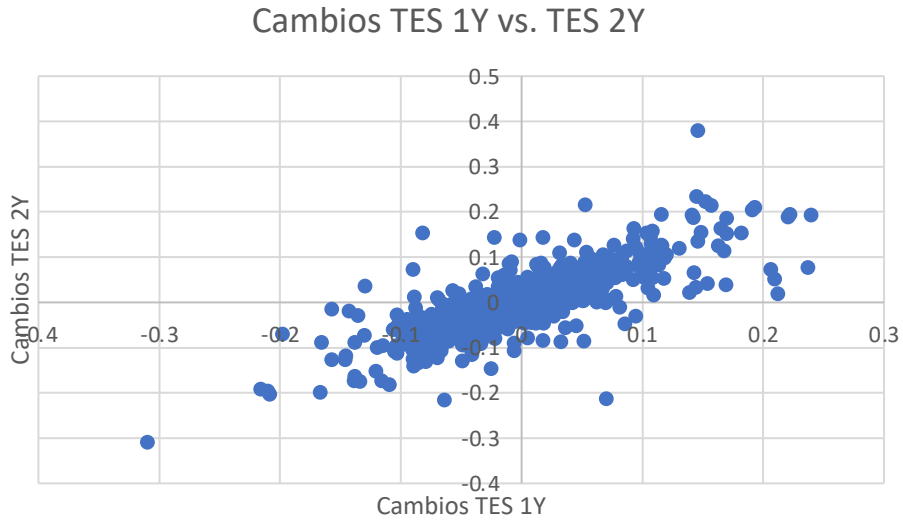
Entre más correlación existe entre los modelos: menor es el poder predictivo.

Índice

Componentes Principales

- Tomemos las siguientes dos variables:
  - Nodo 1Y de la curva TESCOP.
  - Nodo 2Y de la curva TESCOP.
- Es evidente que los dos nodos se parecen mucho, tanto en nivel como en cambios.





Componentes Principales

- Nuestro objetivo es hacer un modelo para predecir el comportamiento conjunto de los dos nodos.
  - Podríamos hacer un modelo conjunto (VAR, VEC, DCC – Garch, etc.).

    O…
  - Podríamos hacer un modelo que correlacione de alguna manera ambas variables.

- Se observa que hay una alta correlación entre ambas variables.
- ¿Y si generamos una variable que represente a ambas?

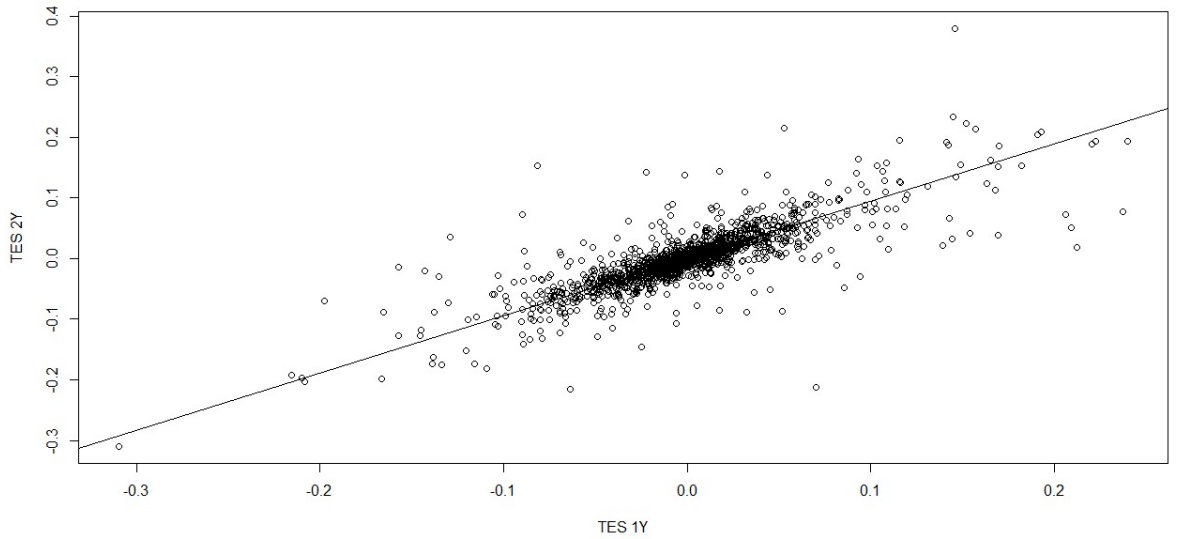
Componentes Principales

- El análisis de componentes principales (ACP o PCA) consiste en generar variables independientes que expliquen lo mejor posible un conjunto de variables dependientes. Por ejemplo, generemos una variable.
- $C = 0.6873812 * \text{Nodo}_{1Y} + 0.7262968 * \text{Nodo}_{2Y}$

- ¿Por qué esos números?

Se ve bien:

- ¿Qué % de la varianza explica?



Componentes Principales

- Pero de donde salieron esos números:

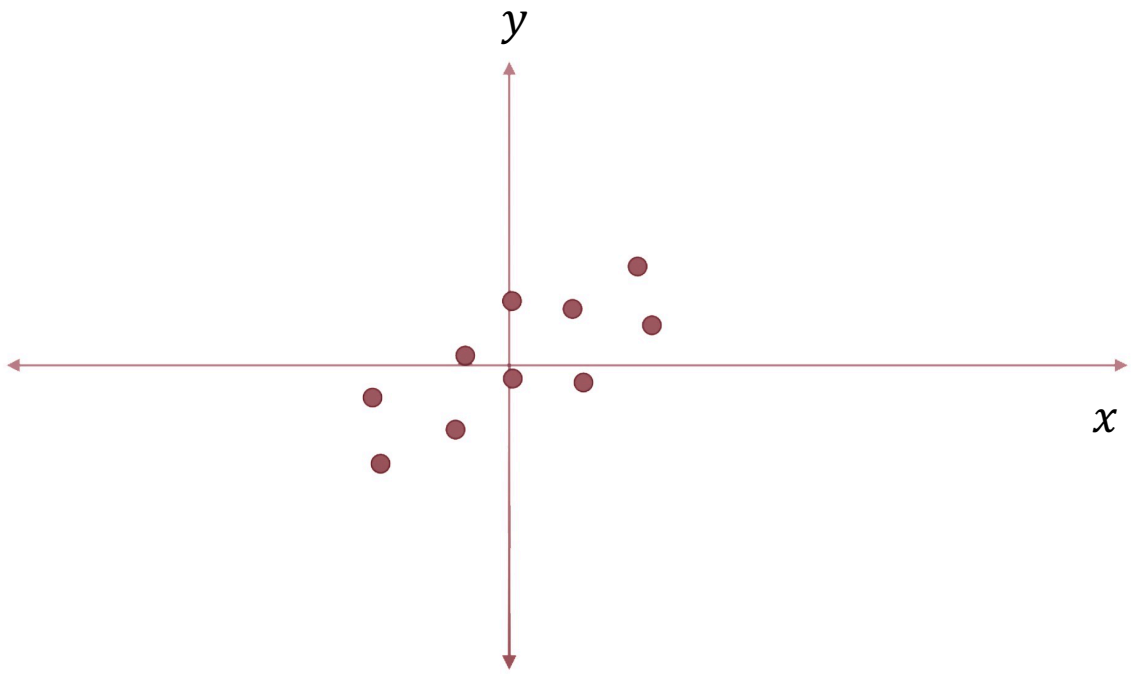
$$w = \underset{\|w\|=1}{\operatorname{argmax}} \|Xw\| = \underset{\|w\|=1}{\operatorname{argmax}} w^T * X^T * X^T * w$$

- ¿Y esto con qué se come?

- ¡Esos de ahí son los vectores propios!

- Recordemos lo que es un vector propio y un valor propio:

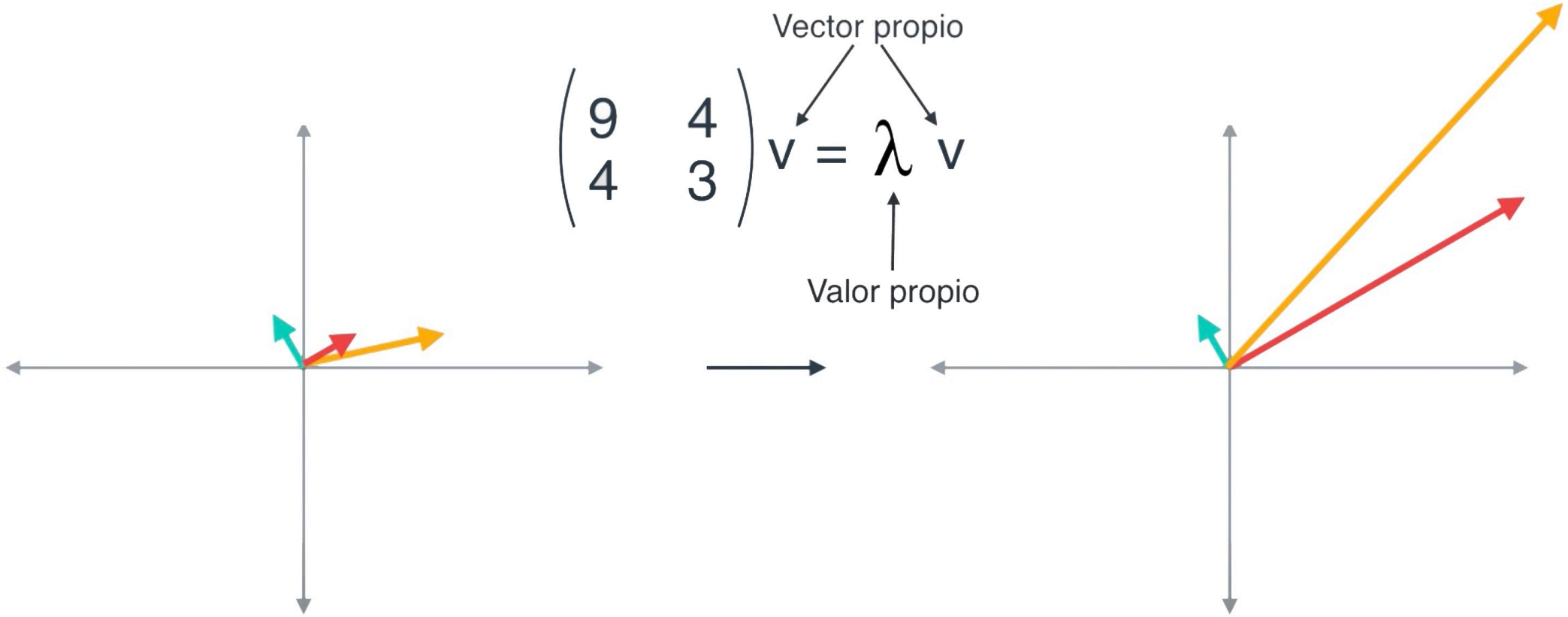
Componentes Principales



$$VarCov = \begin{bmatrix} var(x) & cov(x,y) \\ cov(x,y) & var(y) \end{bmatrix}$$

$$VarCov = \begin{bmatrix} 9 & 4 \\ 4 & 3 \end{bmatrix}$$

Componentes Principales



Componentes Principales

- Tenemos nuestra matriz de Varianza-Covarianza: A

$$A = \begin{bmatrix} 9 & 4 \\ 4 & 3 \end{bmatrix}$$

- Para obtener vectores y valores propios, necesitamos cumplir:

$$|A - \lambda * I| = 0$$

$$\left|\begin{bmatrix} 9 & 4 \\ 4 & 3 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right| = \left|\begin{bmatrix} 9-\lambda & 4 \\ 4 & 3-\lambda \end{bmatrix}\right|$$

- El determinante de esa matriz es:

$$(9-\lambda)(3-\lambda) - 16 = 27 - 9\lambda - 3\lambda + \lambda^2 - 16 = 0$$

$$\lambda^2 - 12\lambda + 11 = (\lambda - 11)(\lambda - 1)$$

- Por lo que las dos posibles soluciones son:

$$\lambda_1 = 11\,;\lambda_2 = 1$$

Componentes Principales

- Ahora faltan encontrar los vectores de esta nueva matriz:

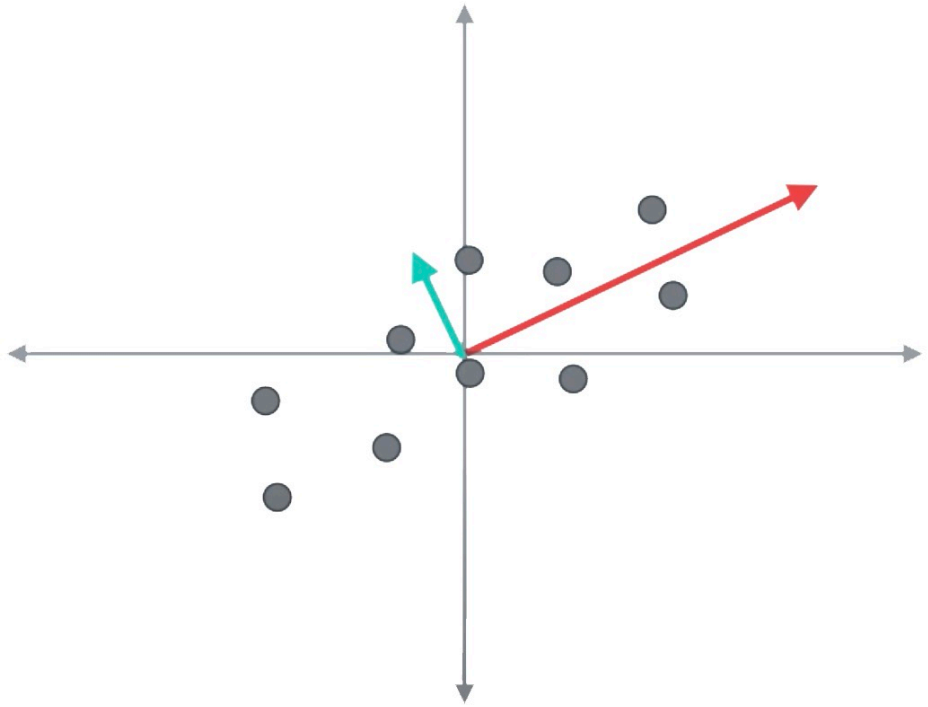
$$A * v = \lambda * v$$

Donde $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$

$$\begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = 11\begin{pmatrix} u \\ v \end{pmatrix} \qquad \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} = 1\begin{pmatrix} u \\ v \end{pmatrix}$$

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \qquad \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$

Componentes Principales

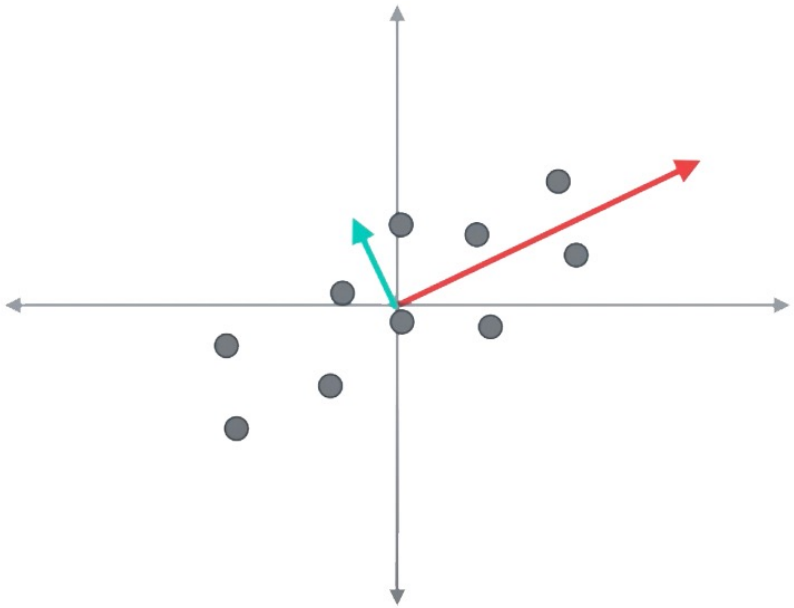


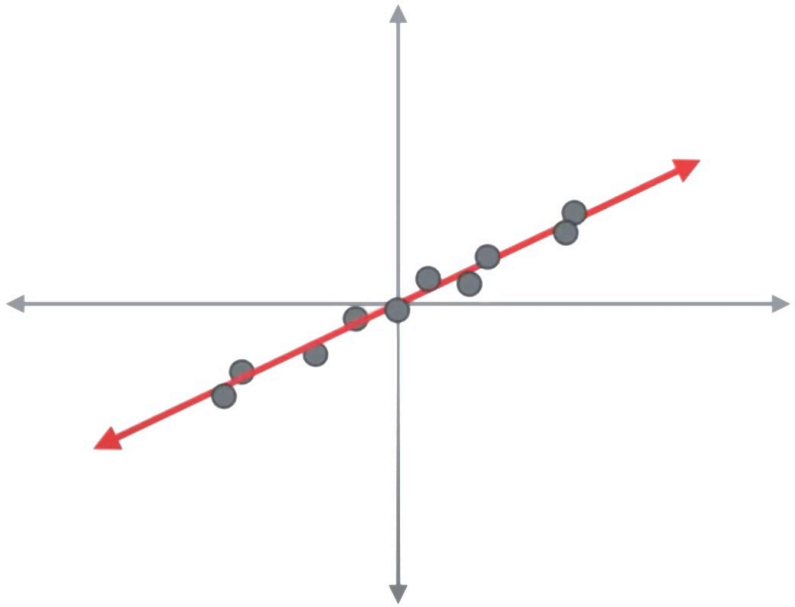
$$\Sigma = \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$ Vectores propios (dirección)

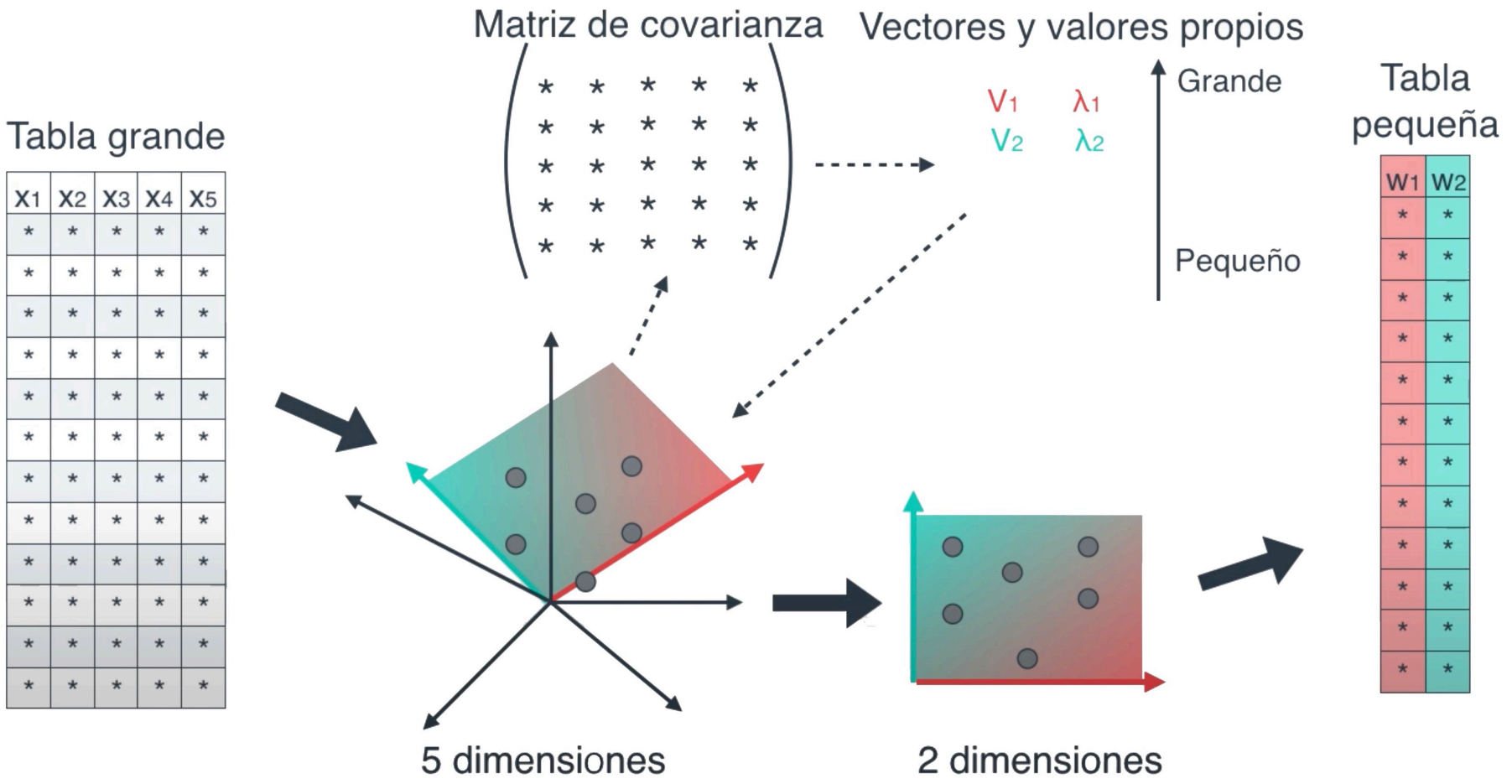
11 1 Valores propios (magnitud)

Componentes Principales





N-dimensiones



Componentes Principales

Resumamos:

- Generamos nuestra matriz de varianza-covarianza.
- Encontramos los valores propios de la matriz cuadrada.
- Tomamos los valores propios de mayor a menor y obtenemos los vectores propios.
- Los valores propios representan la varianza explicada.
- Los vectores propios son los componentes principales – los pesos de nuestras nuevas variables.
- La suma de los valores propios son iguales a la suma de la varianza (diagonal de var-cov): la traza de la matriz es igual a la suma de sus valores propios.

Componentes Principales

Algunas consideraciones:

- Ventajas:
  - Las variables son linealmente independientes: podemos hacer modelos separados.
  - Es fácil de entender. Devolverse a las variables originales es trivial.
  - Usualmente permite reducir la dimensionalidad (esto es necesario!).
- Desventajas
  - Los pesos de nuestras variables son estáticos.
    - Esto hace que la correlación también lo sea.
  - El verdadero provecho se ve solo en variables correlacionadas.

Componentes principales

Los componentes principales son, por construcción, ortogonales entre si mismos.

Cada componente explica una proporción de la varianza. Para reducir dimensionalidad se puede seleccionar un número de componentes (N) que expliquen un umbral de varianza (generalmente el 95%).

Con menos componentes se logra un modelo más parsimonioso y a cambio de una pequeña perdida de información.

Componentes principales

Es posible realizar econometría (Regresiones, ARIMA, GARCH, VAR) sobre los auto-valores y después devolverse a la escala original (Ejemplo: Curva TES).

La regresión Ridge y la PRC (*Principal Regression Components)* son procedimientos de contracción que involucran Componentes principales.

Ridge incluye de manera efectiva todas los componentes y los reduce de acuerdo con el tamaño de los auto-valores asociados con los componentes.

La regresión por componentes principales reduce efectivamente algunos componentes a cero (no incluidos) y no reduce otros (incluidas).

Castro, S. (10 de junio de 2013). Estimación y selección de variables en grandes dimensiones. [Diapositivas]. Recuperado de http://www.iesta.edu.uy/wpcontent/uploads/2014/05/CursoPosgrado_Aprendizaje_Automatico_SCastro_2013.pdf

Diebold, F. (2017). *Forecasting in Economics, Bussines, Finance and Beyond*. Pensilvania, Estados Unidos: University of Pennsylvania.

Lejarza, I. (s.f.). Introducción a la inferencia bayesiana. [Diapositivas]. Recuperado de https://www.uv.es/mlejarza/actuariales/iibayes.pdf

Riascos, A. (Marzo de 2018). Selección, Riesgo Esperado y Validación de Modelos. [Diapositivas] Recuperado de https://www.uv.es/mlejarza/actuariales/iibayes.pdf

http://www.dm.uba.ar/materias/estadistica_teorica_Mae/2006/2/practicas/bayes.PDF

https://ocw.ehu.eus/pluginfile.php/3145/mod_resource/content/1/estadistica/tema-10-regresion-sesgada.pdf

Gracias

matemáticas aplicadas