Ttitle Principal component analysis

Afdasdadf

# Introduccion y motivacion

En épocas modernas la data abunda y los análisis sobre ella han llevado a desarrollar técnicas y procedimientos para su entendimiento que permean a todas las ramas de la ciencia.

Dentro de dichos estudios, el análisis multivariado ha probado ser de mucha utilidad para el procesamiento de datos y el descubrimiento de patrones. De igual manera la cantidad abrumadora de datos que se presentan día a día requieren que los métodos para procesar datos sean eficientes y conserven propiedades fundamentales en el análisis, es para responder esta necesidad que surge el análisis de componentes principales.

El análisis de componentes principales surge como un método muy eficiente y poderoso para reducir la dimensionalidad de la data sin sacrificar mucha calidad informativa a la hora de ser aplicado.

Particularmente, el análisis de componentes principales o bien conocido por sus siglas en inglés (PCA), consiste en reducir la dimensión de la data en unas variables llamadas componentes principales (PC) mientras se retiene tanta variabilidad proveniente de la data como sea posible.

Este trabajo está apuntado a explicar el método de componentes principales, su estimación, su relación con el análisis matricial y finalmente una aplicación directa del mismo.

# Definición de los componentes principales

Suponga que *x* es un vector de *m* variables aleatorias y que se desea analizar la estructura de varianzas y covarianzas de las *m* variables. Si las *m* variables aleatorias son pocas, no es inconveniente alguno analizar las varianzas y covarianzas una a una. Lamentablemente cuando el número de variables escala la complejidad del análisis también lo hace ya que sería necesario analizar covarianzas.

El PCA se concentra en las varianzas de las variables para derivar los valores de sus PC.

El primer paso es entonces buscar una función lineal de los elementos de *x* que tengan la máxima varianza, donde es un vector de *m* constantes . Tal que:

Luego de encontrar esta función lineal se repite el mismo procedimiento para buscar una función que cumpla todas las propiedades antes mencionadas y que además esté no correlacionado con el anterior vector. Este procedimiento se sigue hasta llegar a la función que es una función que maximice la varianza dado que este vector no está correlacionado con ninguno de las anteriores *k-1* funciones. Dado que se necesitan vectores no correlacionados estos mismos tienen un número limitado que asciende a *m*, el número de variables.

Afortunadamente, en la mayoría de veces los primeros *w* PC van a explicar la mayoría de la varianza y se presenta que este *w << m* por lo que resulta en un método efectivo de reducción de dimensiones.

# Derivación de los componentes principales

Luego de haber definido los componentes principales se necesita un método de cálculo, que sea eficiente y que cumpla con todos los requerimientos antes mencionados para los PC.

Considere la matriz de covarianza conocida , ahora, aplicando la derivación de los PC anteriormente expuesta se obtiene que la función *k*-ésima es donde es un vector propio de la matriz , asociado a *k*-ésimo valor propio .