Holger Cevallos Valdiviezo, Ph.D.
holgceva@espol.edu.ec
Profesor de Estadística
Facultad de Ciencias Naturales y Matemáticas (FCNM)

DETECCIÓN DE VALORES ATÍPICOS MEDIANTE ANÁLISIS DE COMPONENTES PRINCIPALES ROBUSTOS

PROBLEMA

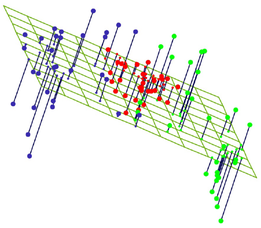
- Los valores atípicos son un problema muy común en el análisis de datos, sobre todo en Big Data.
- Ante la presencia de valores atípicos, los métodos clásicos de estadística obtienen resultados estadísticamente inválidos, lo cual puede conllevar a tomar decisiones erróneas.
- Es necesario considerar emplear métodos estadísticos robustos que traten de ajustar el modelo impuesto por los datos correctos. Este ajuste robusto puede ser usado también para detectar valores atípicos.
- Los métodos robustos para Análisis de Componentes Principales (ACP) pueden ser usados para detectar valores atípicos.

OBJETIVO GENERAL

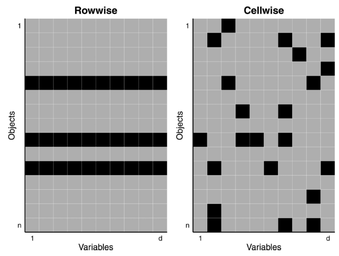
- Presentar ejemplos de detección de valores atípicos utilizando métodos robustos de Análisis de Componentes Principales (ACP) en datos reales de diferentes campos de la ciencia y mostrar cómo el usuario puede utilizar este resultado para identificar las causas de la atipicidad de estos valores.

PROPUESTA

- ACP encuentra una representación en menor dimensión que produce la mejor aproximación posible a los datos originales. Sin embargo, ACP clásico es sensible ante valores atípicos (i.e. obtiene una estimación errónea).

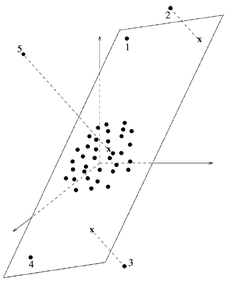


- Tipos de atipicidad: por filas, por celdas, por filas y celdas



- Se sugiere considerar métodos robustos para ACP con el fin de estimar el subespacio de menor dimensión.

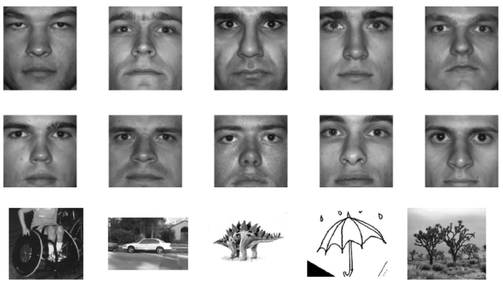
- El ajuste robusto de ACP puede ser usado para detectar valores atípicos.
- Los métodos robustos por filas para ACP (por ej. ROBPCA, estimador LTS, estimador S,…) permiten detectar diferentes tipos de observacioes atípicas en base a puntos críticos definidos por Hubert et al. (2005).



- Los métodos robustos por celda para ACP (por ej. ROCPCA, S-estimadores por coordenadas,…) son más recientes y corresponden con un paradigma mas realista de valores atípicos.
- Los métodos robustos por celdas pueden ofrecer mayor detalle sobre qué variables producen atipicidad en una observación.
- Hubert et al. (2019) propusieron el método MACRO-PCA, el cual es robusto por filas y por celdas.

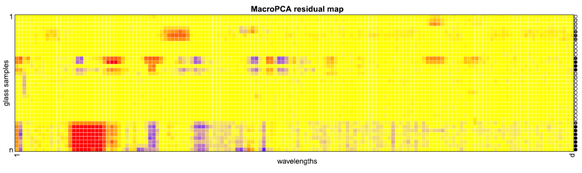
RESULTADOS

- Reconocimiento facial: usando ACP robusto por fila.



- Discriminar imágenes de rostros humanos de imágenes de objetos.

- Muestras arqueológicas de vidrios del siglo XVI y XVII, espectros con 750 longitudes de onda:



- Las muestras de vidrio 22-30 tienen una alta y atípica concentración de fósforo.
- Las muestras 57-63 y 74-76 tienen una alta y atípica concentración de calcio.
- El instrumento se limpió antes de medir los últimos 38 espectros.

CONCLUSIONES

- Se sugiere ajustar métodos robustos resistentes a filas y a celdas atípicas para estimar el subespacio de menor dimensión.
- Este ajuste robusto puede ser luego comparado con el ajuste clásico para evaluar la influencia de los valores atípicos.
- Los métodos robustos para ACP también pueden ser usados para detectar filas y celdas atípicas.
- La identificación de valores atípicos proporciona una alarma al investigador sobre posibles problemas en los datos, los cuales pueden ser luego investigados e interpretados.