Comparación de propuestas para el análisis de componentes principales en matrices con datos faltantes

Nicolás Murrone[1] y Alejandra M. Martínez[2]

[1]Comisión Nacional de Energía Atómica, Argentina
nicomurrone@gmail.com
[2] Universidad Nacional de Luján, Argentina
ale_m_martinez@hotmail.com

Palabras Claves: Análisis de Componentes Principales, Datos Faltantes, Lenguaje R, Algoritmo NIPALS.

El análisis de componentes principales es una poderosa herramienta exploratoria, utilizada en diversas disciplinas tales como la biología, arqueología, entre otras, que tiene como principal objetivo reducir la dimensionalidad del conjunto de datos perdiendo la menor cantidad de información posible, lo cual facilita la interpretación y puede utilizarse como paso intermedio en un análisis de datos más complejo.

Las matrices provenientes de análisis de muestras ambientales suelen contener un importante número de datos faltantes lo cual es un problema para la aplicación clásica de análisis de componentes principales. La mayoría de los procedimientos clásicos consisten en imputar los datos con técnicas que se basan en la distribución o naturaleza de las variables y luego calcular las componentes principales. El algoritmo NIPALS (Non-linear Iterative Partial Least Squares) es un procedimiento iterativo que en cada iteración calcula una componente principal resolviendo un problema de regresión de mínimos cuadrados parciales (PLS, por sus siglas en inglés) y que, en caso de haber presencia de datos faltantes, no requiere de imputación de datos para su cómputo.

En esta presentación se realiza un estudio de simulación en R para comparar el algoritmo NIPALS con otros procedimientos clásicos de imputación de datos y, de esta manera, elegir la técnica a utilizar sobre un conjunto de datos reales de particulado atmosférico.