Imputación múltiple y análisis de componentes principales

Nombre1 Apellido1[1], Nombre2 Apellido2[1], Nombre3 Apellido3[1], Mario Pacheco[2]

[1] Universidad Santo Tomás, Pregrado en Estadística [2] Profesor AEDF

Contexto

Análisis de componentes principales (ACP). Método estadístico ampliamente utilizado para reducir la dimensionalidad de variables numéricas en conjuntos de datos multivariados, preservando la mayor parte de la variabilidad original. Se aplica en áreas como análisis exploratorio, compresión de datos, visualización y construcción de indicadores.

Imputación múltiple. La ausencia de datos es un problema frecuente en el análisis estadístico multivariado. La imputación múltiple es una técnica robusta para abordarlo: consiste en generar varias versiones completas del conjunto de datos, imputando valores plausibles bajo un modelo estadístico. Luego se analizan por separado y se combinan los resultados para reflejar la incertidumbre debida a los valores faltantes.

Propuesta

Para abordar el problema de datos faltantes en el contexto del análisis multivariado, se propone combinar imputación múltiple con análisis de componentes principales (ACP), permitiendo explorar la estructura latente de los datos incorporando la incertidumbre asociada a la imputación.

Metodología

1. Realizar un proceso de imputación con *m* imputaciones y un número máximo de iteraciones, *maxit*, empleando un modelo de imputación como *pmm*.
2. Para cada conjunto de datos imputado, realizar un ACP, guardando los valores propios asociados, las coordenadas de las variables e individuos, así como los cosenos cuadrados y las contribuciones.
3. Combinar los resultados de cada uno de los *m* análisis en un único ACP combinado.
4. Realizar la visualización del análisis combinado, en conjunto con el ACP de cada conjunto de datos imputado.

Aplicación

- **Datos.** Se empleó un conjunto de datos con cinco variables relacionadas con los factores de la personalidad, medidas en 1788 individuos:
  - `Nscore`: neuroticismo
  - `Escore`: extraversión
  - `Oscore`: apertura a la experiencia
  - `Ascore`: amabilidad
  - `Cscore`: escrupulosidad
- **Implementación.** Se desarrolló un algoritmo en R (versión 4.4.3) en el entorno RStudio (versión 2025.05.0+496). El algoritmo permite la imputación múltiple mediante la función `mice` de la librería `mice` (versión 3.17.0), luego realiza un ACP sobre cada conjunto de datos imputado empleando la función PCA de la librería `FactoMineR` (versión 2.11) y finalmente, con la función propuesta `comb.acp()`, se combinan los resultados. Se empleó la librería `ggplot2` (versión 3.5.1) para las representaciones gráficas del análisis propuesto.

Resultados

Algunos de los resultados del ACP para cada imputación y combinados se presentan a continuación.

Porcentajes de inercia

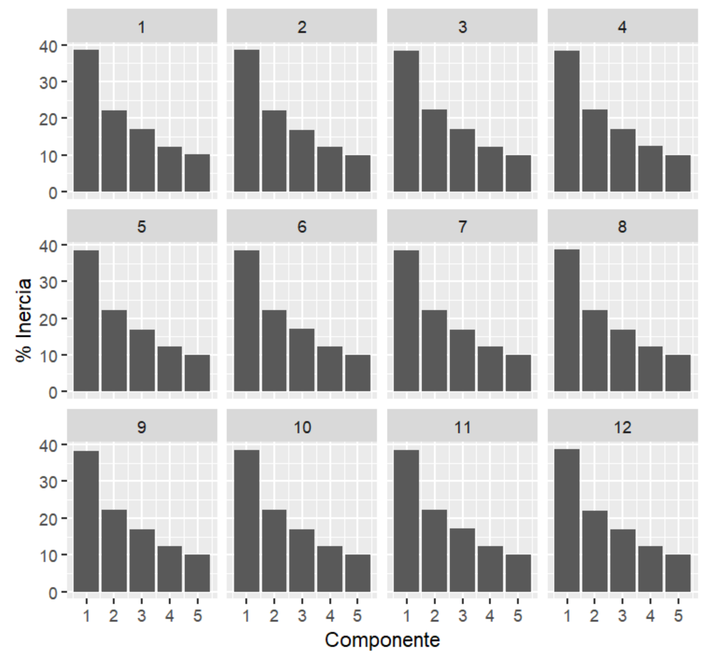


Figura 1. Inercia en cada imputación

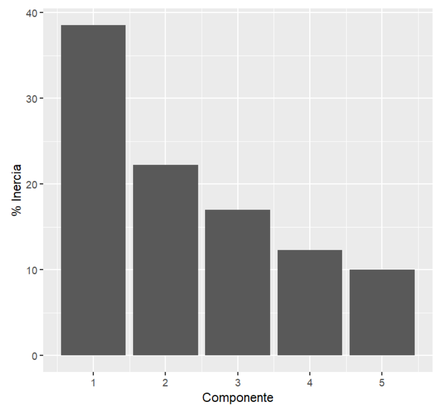


Figura 2. Inercia combinada

Representación de las variables y de los individuos

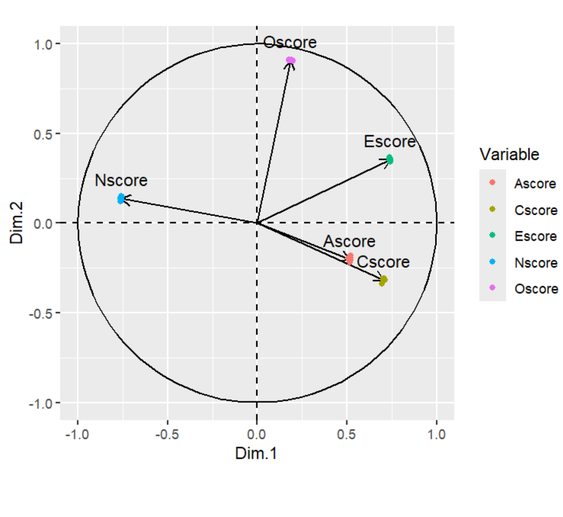


Figura 3. Variables en el primer plano factorial

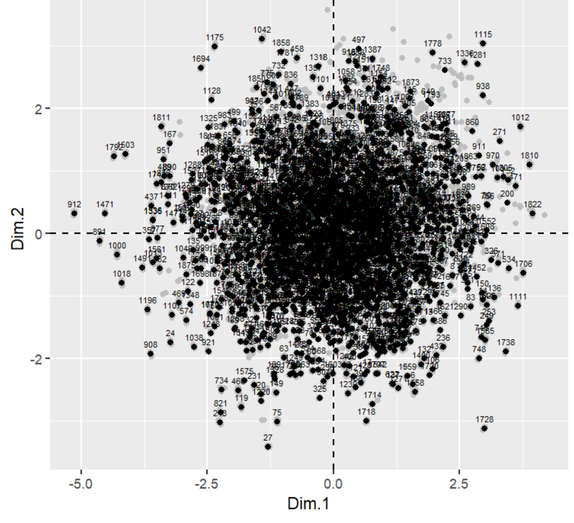


Figura 4. Individuos en el primer plano factorial

Conclusiones

- Es posible adaptar la imputación múltiple al análisis de componentes principales.
- La representación de las variables para cada imputación y para los resultados combinados permite examinar la incertidumbre debida a los datos faltantes.
- La representación de los individuos para cada imputación y para los resultados combinados permite examinar la incertidumbre debida a los datos faltantes.

Referencias

- Husson, F., Lê, S. y Pagès, J. (2011). *Exploratory Multivariate Analysis by Example Using R*. Chapman & Hall/CRC.
- van Buuren, S. (2018). *Flexible Imputation of Missing Data* (2.ª ed.). Chapman & Hall/CRC.

CONTACTO

nombre.apellido@usantotomas.edu.co