

UNIVERSIDAD NACIONAL DE COLOMBIA
DEPARTAMENTO DE ESTADÍSTICA
TEORÍA ESTADÍSTICA MULTIVARIADA, 2025-I
TALLER 2: PCA, CLASIFICACIÓN, CLUSTERING y MONITOREO DE PROCESOS

METODOS

1. Con relación al artículo de Zou, H., Hastie, T., & Tibshirani, R. (2006).
 - a. Explique en qué consiste la propuesta de los autores
 - b. Replique el ejemplo PITPROPS DATA
2. Con relación al artículo de Hubert, M., Rousseeuw, P. J., & Vanden Branden, K. (2005).
 - a. Explique en qué consiste la propuesta de los autores
 - b. Replique el ejemplo car data
 - c. Replique el ejemplo octane data
3. Con relación al artículo de Hubert, M., Reynkens, T., Schmitt, E., & Verdonck, T. (2016):
 - a. Explique en qué consiste la propuesta de los autores
 - b. Replique el ejemplo glass data
4. Con relación al artículo de Sanusi, R. A., Ajadi, J. O., Abbasi, S. A., Dauda, T. O., & Adegoke, N. A. (2024):
 - a. Explique en qué consiste la metodología de proyecciones aleatorias?
 - b. Qué papel juega en este método el teorema de Johnson-Lindenstrauss?
 - c. Indique los tipos de matrices aleatorias consideradas por los autores.
5. Siguiendo el artículo de Cardoso, Â., & Wichert, A. (2012):
 - a. Explique cómo puede identificar los clusters presentes en un conjunto de datos?
 - b. Indique posibles deficiencias de la metodología propuesta y posibles soluciones

APLICACIONES

6. Usando el conjunto de datos car data, monitoree el proceso
 - a. usando cartas de control T2
 - b. usando cartas de control basadas en PCA
7. Usando el conjunto de datos octane data, monitoree el proceso
 - a. usando cartas de control T2
 - b. usando cartas basada en PCA
8. Usando el conjunto de datos glass data, monitoree el proceso
 - a. usando cartas de control T^2
 - b. usando cartas basada en PCA
 - c. implemente una carta de control basada en proyecciones aleatorias.
9. Considere el conjunto de datos wine-dataset Suponga que no conoce las diferentes clases de vino consideradas en el dataset.
 - a. Implemente varios procesos de clustering y compare su eficiencia.

- b. Realice un PCA y tome un número de componentes adecuados. Sobre los componentes seleccionados, Implemente varios métodos de clustering. Compare la eficiencia de los métodos, y compare los resultados obtenidos con los resultados del punto anterior.
 - c. Implemente varios métodos de clasificación y compare su eficiencia.
 - d. Realice un PCA y tome un número de componentes adecuados. Sobre los componentes seleccionados, Implemente varios métodos de clasificación. Compare la eficiencia de los métodos, y compare los resultados obtenidos con los resultados del punto anterior.
10. Usando el conjunto de datos glass data, identifique los clusters presentes en este conjunto de datos.

REFERENCIAS (orden de lectura sugerido)

- Peña, D. (2013). *Análisis de datos multivariantes*. Cambridge: McGraw-Hill España. Capítulo 5: Análisis de componentes principales.
- Hubert, M., Rousseeuw, P. J., & Vanden Branden, K. (2005). ROBPCA: a new approach to robust principal component analysis. *Technometrics*, 47(1), 64-79.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2020). *An Introduction to Statistical Learning with Applications in R*. Springer. Del capítulo 6 estudiar:
 - a. Introduction: 225-226
 - b. 6.2 Shrinkage methods: 237-250
 - c. 6.5 Lab: Ridge Regression and the Lasso: 274-279
- Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2), 265-286.
- Hubert, M., Reynkens, T., Schmitt, E., & Verdonck, T. (2016). Sparse PCA for high-dimensional data with outliers. *Technometrics*, 58(4), 424-434.
- Sanusi, R. A., Ajadi, J. O., Abbasi, S. A., Dauda, T. O., & Adegoke, N. A. (2024). Multivariate technique for detecting variations in high-dimensional imagery. *IEEE Access*.
- Cardoso, Â., & Wichert, A. (2012). Iterative random projections for high-dimensional data clustering. *Pattern Recognition Letters*, 33(13), 1749-1755.