Project 1

Statistical Methods for Artificial Intelligence
(MEIA, 1st Semester, 2nd Quarter, 2025/2026)

Handed out on Nov 27, 2025.

To be handed back on Dec 12, 23:59h, 2025.

Consider for the **machines** data frame, available in R library **rrcov**, the subset from machine **hp-3000/64** until machine **ibm-4331-2**.

1. Explore/describe the data applying methods toughs in this course, in particular using plots and summary statistics (e.g. mean, median, trimmed and winsorized mean, variance, mad, covariance, generalized/total variance and Mahalanobis distances) and discuss what you have learned from this preliminary analysis.

   (a) Apply principal components analyse (PCA):

      i. Considering the variables in original scale and the classical sample covariance estimate;

      ii. Considering the standardized variables.

   (b) With the aim of dimension reduction, but keeping at least 95% of total variance of the data, which of the two previous analyses do you recommend? Justify your choice based on the percentage of total variance explained by both sample principal components. Interpret the sample principal components you have chosen to retain and plot the data using principal components scores.

2. Introduce an outlier into the data set by changing observation **hp-3000/64** to:

   $\mathbf{x}^{\text{new}}_{\textbf{hp-3000/64}} = (75, 2000, 0.8, 80000, 300, 24, 62, 47)^{\text{T}}$. Apply to the new data set (without standardization):

   (a) the classical PCA;

   (b) the robust PCA based on the MCD estimate.

   Discuss the effect of one atypical observation in both analysis.

About the report:

The report should not exceed 10 pages (with Annexes). Do not forget to include: introduction, objectives of study, decisions, conclusions and bibliography. The R code and the report must be upload in fenix.