# Worksheet 1 – Linear Discriminant Analysis

**High-Dimensional Data Analysis and Machine Learning**

Camille Mondon

January 12, 2026

## 1 LDA on Insurance Data

**Exercise 1.** The data set we consider comes from an insurance company and contains information about **2220 young drivers** who have a car insurance contract. The available variables are:

- TYPE: the type of insurance contract (A: good, B: medium, C: minimum)
- VALUE: the value of the insured vehicle (1 = very low, 2 = low, 3 = medium, 4 = high)
- SEX: the sex of the insured (1 = man, 0 = woman)
- AGEV: the age of the vehicle
- AGEI: the age of the insured

You may use either **R** or **Python** to answer the following questions.

1. Load the data `insurance.csv`[1] and provide a short description of the variables. You should:
    - Display the first few rows of the dataset.
    - Compute summary statistics for each variable.
    - Provide counts for each category of the TYPE variable.

*Hint:* You may use `summary()` and `table()` in R, or `describe()` and `value_counts()` in Python.

2. Randomly split the dataset into a **training sample** and a **test sample**.
    - Use approximately 70% of the observations for training and 30% for testing.
    - Ensure that all levels of the TYPE variable are represented in both samples.

*Hint:* In R, you can use `sample()` or the `caret` package. In Python, you can use `train_test_split` from `sklearn.model_selection`.

3. Using the training sample, predict the type of insurance contract (TYPE) using the explanatory variables VALUE, SEX, AGEV, and AGEI.

[1] You can download it here.

- Fit a **Linear Discriminant Analysis (LDA)** model on the training sample.

- Compute predicted classes for the test sample.

- Evaluate the model's accuracy using a confusion matrix.

*Hint:* In R, use `MASS::lda()`. In Python, use `sklearn.discriminant_analysis.LinearDiscriminantAnalysis`.

💡 Remark

*Remark* 1.1.

- Ensure that categorical variables have the correct type before applying LDA.

- Standardizing numerical variables is not required for LDA but may affect the interpretation of coefficients.

4. Interpret the results obtained from LDA:

- Comment on which variables appear most influential in discriminating between the insurance contract types (using the $R^2$ equivalent in LDA).
- Interpret the variables in the discriminant subspace using the correlations.
- Discuss any misclassification patterns observed in the test sample.

- Reflect on the overall accuracy of LDA and its usefulness for predicting `TYPE`.

*Hint:* Use the confusion matrix to identify which classes are most often misclassified. You may also inspect the LDA coefficients to assess variable importance.

5. Why isn't it possible to draw a ROC curve? How can you modify the insurance variables so that it becomes possible?