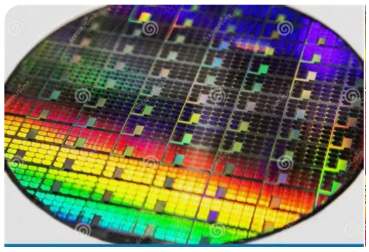
Problem Statement

A factory produces wafers https://en.wikipedia.org/wiki/Wafer_(electronics)

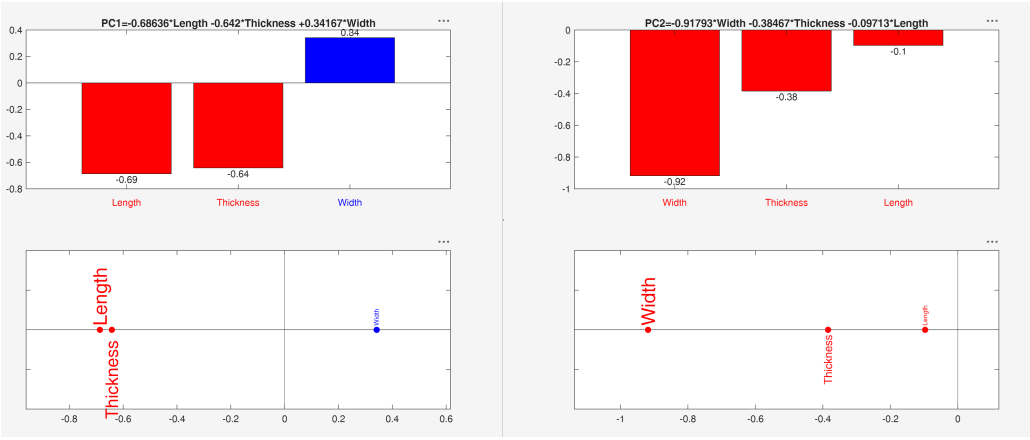


onto which electrical circuits will be printed. For each wafer type, the ratios among three dimensions—thickness, length, and width—are fixed. The process manager asks you to verify whether the production process complies with the prescribed ratios. The data are in the file **wafer.txt**. This file contains **VariableNames** but no **RowNames**. Add the RowNames **w1, …, w10**.

Remark: The file **wafer.txt** is included in **FSDA** version **8.7.10.8**. Make sure you have the latest version of FSDA installed.

QUESTIONS

- Represent the data using **parallel coordinates** to identify any anomalous wafers. The first column of the **parallel coordinates plot** must contain the row names. **Comment on the plot.**
- Construct the scatterplot matrix with univariate boxplots on the main diagonal and a background color for each scatterplot proportional to the corresponding correlation coefficient.
- Compute the correlation matrix (in table format). Discuss the correlations by explaining the angles you would expect between the different pairs of variables.
- Discuss the implications of the statement “for each type of wafer, the proportions among the three measurements are fixed” on the correlations we should expect among the different variables.
- Perform dimensionality reduction. Discuss whether, in this context, it is preferable to work with the covariance matrix or the correlation matrix.
- Write the condition and equation that the eigenvalues of the correlation matrix **R** must satisfy.
- Find the eigenvalues by computing the roots of the polynomial equation (use symbolic computations and function `vpa` to show just 3 decimal places).
- Show the percentage of variance (relative and cumulative) explained by the different components. Comment on the relationship between the first principal component (PC) and the total variance if the proportionality assumption stated above holds.
- Choose an appropriate number of principal components (PCs). See the criteria in **Chapter 11** of the book *IDS*.
- Compute and comment on the correlations between the principal components (PCs) and the original variables.
- Verify that the sum of the squared correlations equals the corresponding eigenvalue.
- Provide a graphical representation of these correlations together with the scores and comment on the plot (use the **biplotAPP** function).
- Create the two plots shown below.



- Interpret the extracted principal components (PCs). What characteristics do the wafers with high values on the first PC exhibit? What characteristics do the wafers with high values on both the first and the second PC exhibit?
- Show the biplot in which
- the row points represent standardized principal components sqrt(n-1)*U=ZV*\Gamma^-1
- Arrows represent correlations between variables and the principal components \Gamma*V
- The length of each arrow represents the **communality** of the corresponding variable.
- the unit circle is also shown
- Compute and display, using a bar plot, the **Score Distance** of each wafer. Programmatically identify the two units with the largest Score Distance. Show a table with their names and the associated distances.
- Show the points in the original space with the 3 PC axes (label the units)
- Show the points in the original space with the projection along the first principal line
- Show the points in the original 3D space together with the 2D spanned by the first two PCs
- Show the points projected in the 3D space of the PCs together with the corresponding axes.
- Show the ellipsoid both in the original space and in the transformed space of the PCs.
- Compute the volume of the ellipsoid (see HW 11.7 for the formula of the volume of the ellipsoid)
- Find the length of the second semiaxis of the ellipsoid
- Find the matrix which provides the best rank 1 approximation of the original matrix (`Zhat`). First find it with call to function `pcares` and then applying `svd`. Check the equality of both procedures
- Show the scatter3D of the matrix `Zhat`. What is the rank of `Zhat`? Comment the plot
- Compute the sum of squares of the differences with the best rank 1 approximation of the original (standardized) matrix.

Questions for the future (after the cluster analysis week)

1. Perform a cluster analysis using the k-means method (with k=2) and add the group membership information to the previous biplot. Also add the two centroids to the same plot.
2. Decompose the total variance (computed on the original, non-standardized variables) into within-group and between-group components and provide an assessment of the quality of the classification.