NASA Turbofan Level B2

Understand your data and your modelling goal

- codes: https://github.com/drvojtex/ADAML_NASATurbofanLevelB2

The objective of the work is to address missing data in the C-MAPSS engine degradation dataset, which simulates aircraft engine operations under various fault modes. The dataset contains 25 variables, including three operational settings and 21 sensor measurements, recorded over multiple operational cycles. Within a single operational setting (we selected the operational setting number 1), the variables Operational Setting 1, Operational Setting 2, and Operational Setting 3 are considered constant and therefore redundant; they are not used in further analysis. Measurements are performed on 100 individual units, each with a different number of time cycles. One of the sensors is assumed to have stopped emitting data (we selected Sensor 17), and the goal is to estimate its missing values using multivariate regression approaches. The observations from individual sensors do not have a known physical representation within the scope of this task. Key aspects include determining the most appropriate number of latent variables for accurate prediction and identifying the minimum number of other sensors required to reliably reconstruct the missing sensor measurements.

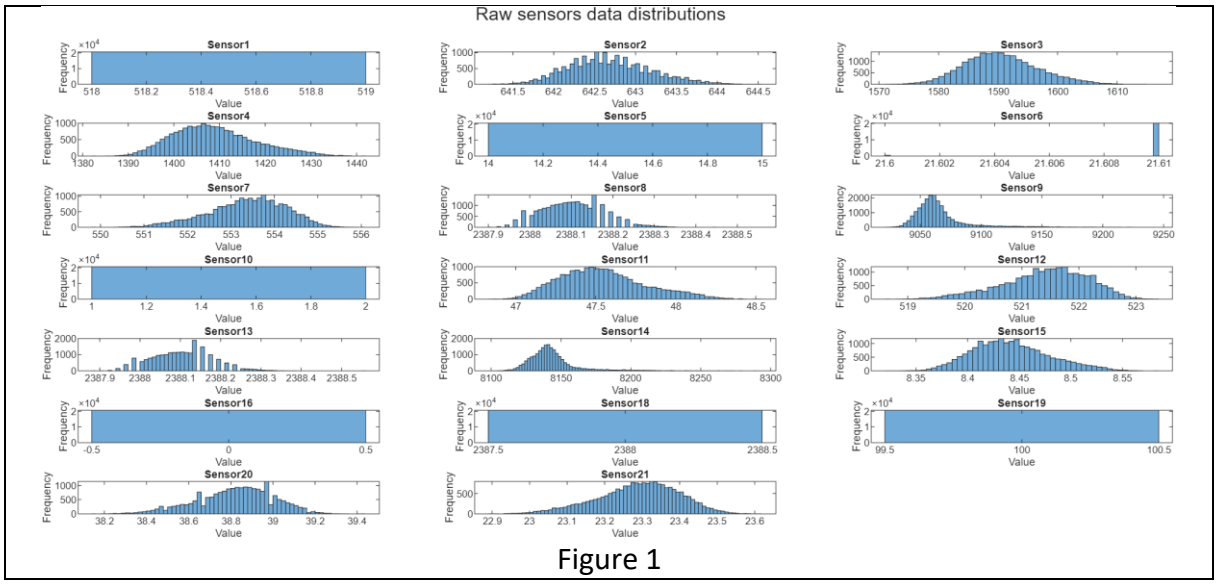

Figure 1

Since the goal is to estimate the value of a faulty sensor (Sensor 17) from the measurements of the other sensors within a single observation, this task is not treated as a time series problem. Each observation, corresponding to a snapshot of all sensor readings at a given time cycle, can be considered an independent sample. Therefore, the temporal order of measurements is not relevant for the estimation of the missing sensor values. The training data does not include any missing values. Matrix of samples has dimensions rows as observations (20631) and columns as sensors measurements (20).

Figure 1 shows distributions of sensors observations. It is visible that sensors 1, 5, 10, 16, 18, 19 are constant. Next, it is also visible that measurements are differently scaled and located (the standardization will be needed). Figure 2 shows a correlation heatmap (without constant sensors). It can be seen that there are mutually correlated sensors (e.g. 2 and 3 and 4). Furthermore, there are two highly correlated sensors 9 and 14. Figure 3 shows principal components explained variance. The PCA was computed after z-standardization and filtering constant sensors. It is visible that over 80% of

information is given by the first three components. Figure 4 shows PCA biplot. It is visible that some sensors are correlated (7 and 12, 20 and 21, 2 and 15 and 4). These findings are also visible in Figure 5, which presents the loading sizes according to PCs.

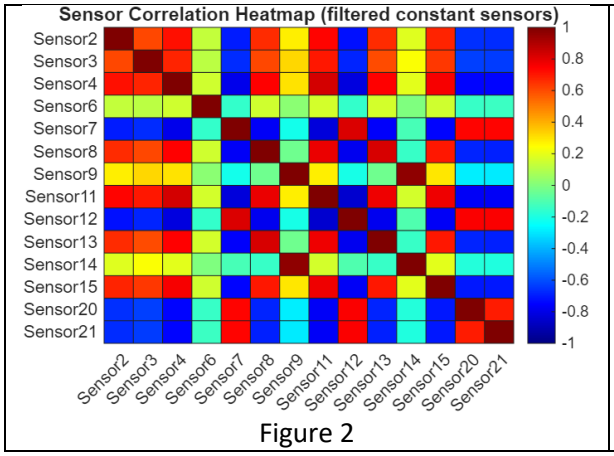

Figure 2

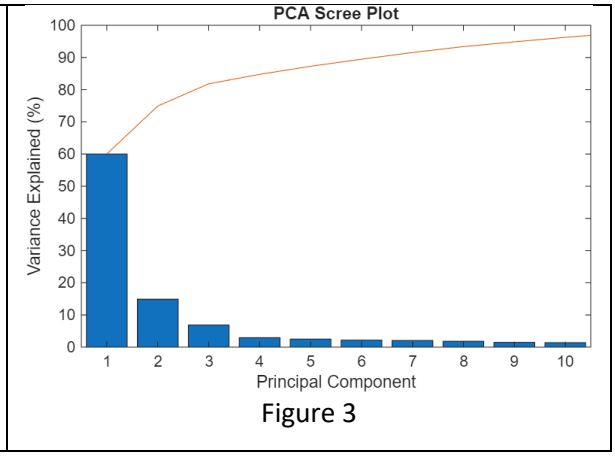

Figure 3

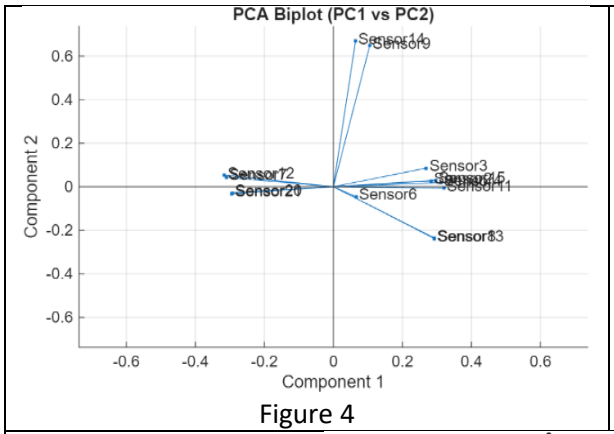

Figure 4

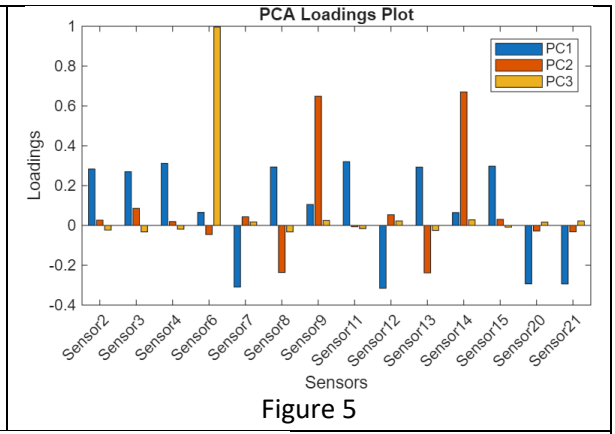

Figure 5

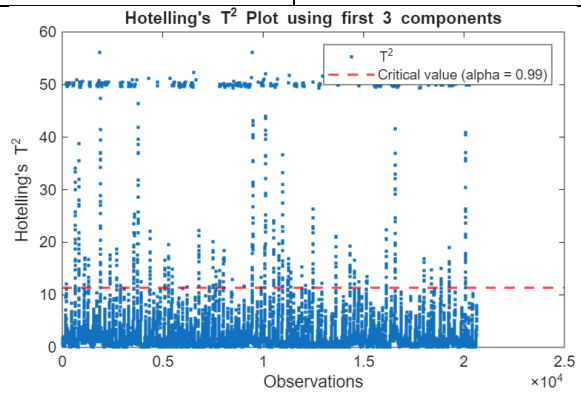

Figure 6

Figure 6 shows a systematic variation computed based on the first three PCs. It is visible that there are outliers not well captured by the PCA model. There are 4.25% of samples exceeding the critical value (F-distribution, alpha 0.99).

In conclusion, it can be said that due to the distributions of measured sensor values according to histograms, it will be necessary to standardize the data (zscore standardization). It can also be assumed that not all sensors will be necessary for the prediction of the missing sensor 17, since many sensors are correlated with each other and thus carry similar or the same information, or some sensors have constant measurement values. The anomaly identification method, such as PCA or Mahalanobis distance, can also be a further pre-treatment step.