NASA Turbofan Level B2: Understand your data and your modelling goal

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

The work addresses missing sensor data in the aircraft engine degradation dataset. The dataset contains 25 variables, including three operational settings and 21 sensor measurements, recorded over multiple operational cycles. Within a single operational setting (the first one was selected), the Operational Setting variables are considered constant and therefore not used in 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 (Sensor 17 was selected), and the goal is to estimate its missing values using multivariate regression. The observations from individual sensors do not have a known physical representation. Key aspects include determining the most appropriate number of latent variables for regression and identifying the minimum number of required sensors.

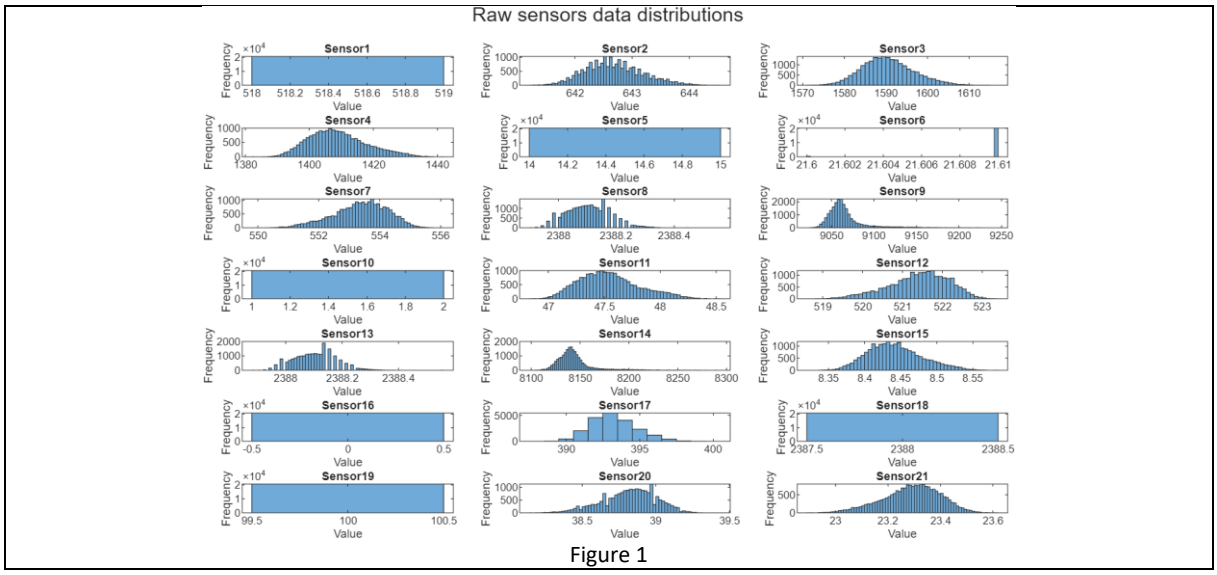

Figure 1

The goal is to estimate the values of the faulty Sensor 17 based on measurements from the remaining sensors, where each unit's measurements form a time series of varying length. Missing sensor values will be predicted via regression over a defined sliding time window, capturing the temporal context within each sequence. The sliding window (size will be the regression model hyperparameter) approach handles the problem of sequences with different lengths. For initial PCA analysis, all measurements were consolidated into a single matrix of 20,631 observations by 21 sensor features.

Figure 1 shows distributions of sensors observations. It is visible that sensors 1, 5, 6, 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. The target sensor 17 is correlated with sensors 2, 3, 4, 8, 11, 13, 15. 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 with 2 PCs. Sensors 3 and 17 (target) are closely aligned, indicating that they convey similar information along the directions of maximum variance. It can be expected that sensor 3 will be crucial for the regression. Sensor 6 has lowest importance. Sensors 2, 4, 11 and 15 have high importance in the 1st PC and low in the 2nd. Sensors 7, 12, 20 and 21 show opposite loadings on the 1st PC. This suggests that the 1st PC captures contrasting behaviour between the two sensor groups across the samples. These findings are also visible in Figure 5, which presents the loading sizes according to PCs.

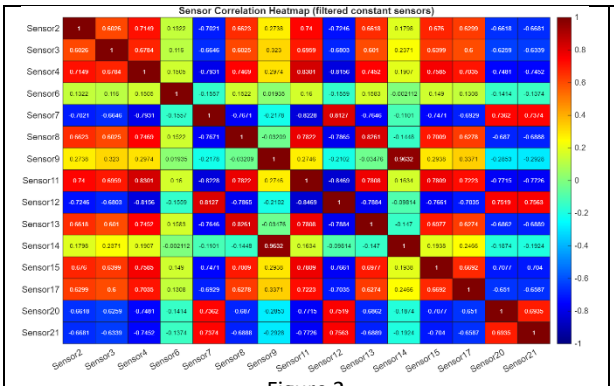

Figure 2

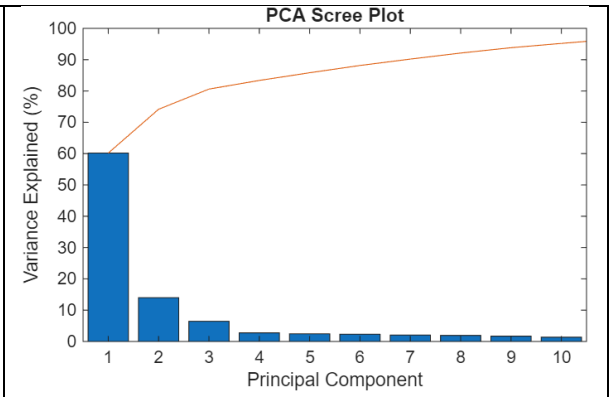

Figure 3

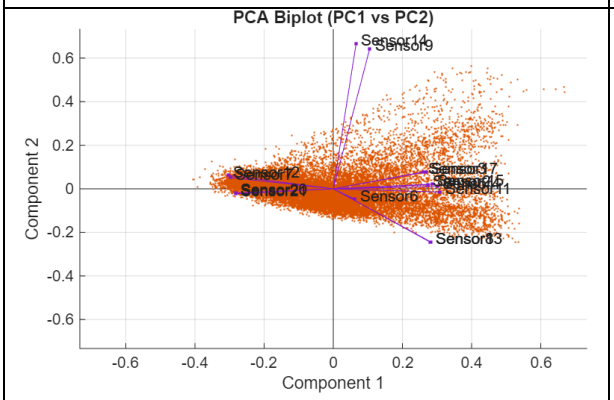

Figure 4

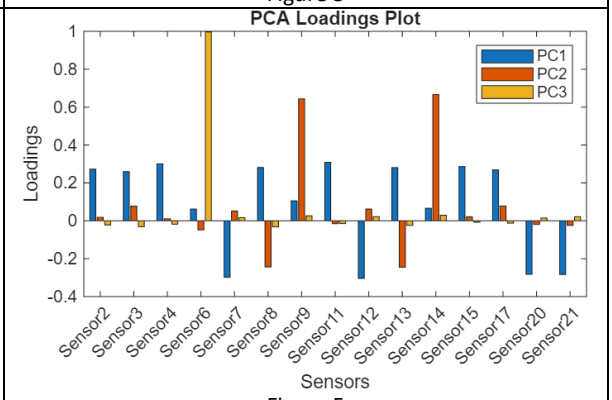

Figure 5

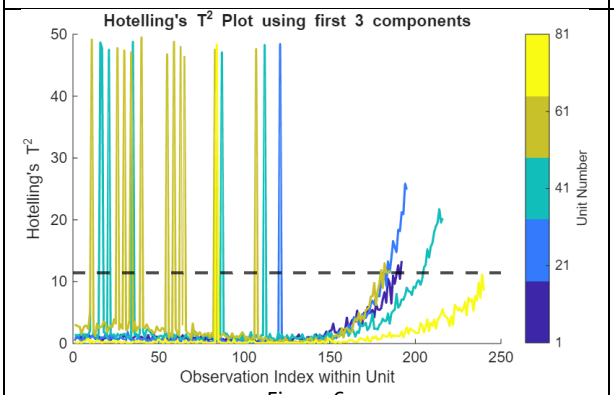

Figure 6

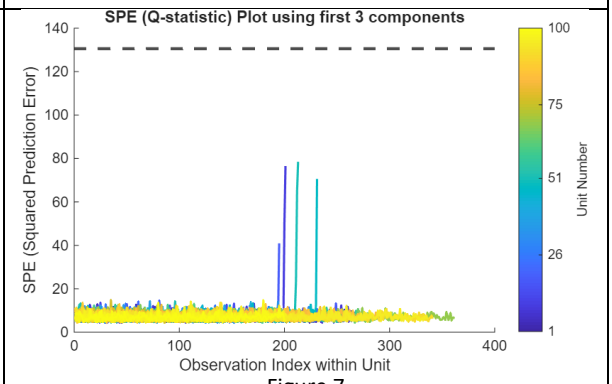

Figure 7

The T^2 control chart shows the temporal evolution of Hotelling's T^2 for individual units. Five representative unit trajectories were selected. In the first half of the trajectories, isolated observations exceed the control limit, but values quickly return below it. In the second half, T^2 values increase gradually and systematically, eventually surpassing the critical threshold, determined according to Hadian & Rahimifard [1]. Figure 7 shows the Squared Prediction Error (SPE). Despite 4 peaks around 200th observation, all measurements remain below the SPE critical limit (Jackson & Mudholkar [2]).

In conclusion, 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 there are constant sensors. The regression will be performed in a window-based approach. One potential follow-up approach is to develop two separate models: one for windows containing only T^2 values below the control limit and another for windows that include measurements exceeding it.

[1] https://doi.org/10.1016/j.cie.2019.03.021

[2] https://doi.org/10.1080/00401706.1979.10489779