Projects

Statistical Analysis of Networks and Systems (SANS-MIRI)

October 7, 2023

1 Project. Improve the calibrated data using temporal patterns in Air Quality Sensor Monitoring Networks.

The objective of this project is to perform denoising for LCSs used in IoT monitoring platforms. Sensors are first calibrated in-situ using linear or nonlinear machine learning models that only take into account instantaneous measurements. We give you calibrated data using a linear model (multiple linear regression, MLR) and a nonlinear model (support vector regression, SVR), Figure 1.

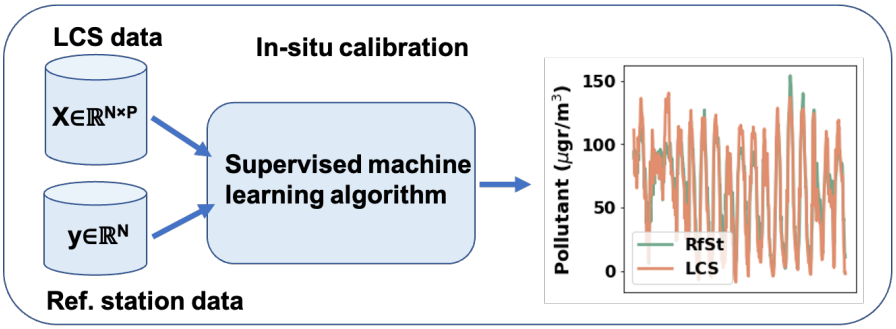

Figure 1: Calibration using a supervised machine learning mechanism.

To improve the estimation, i.e., to decrease the RMSE value, we propose in this second stage to use the temporal daily correlations of reference instrumentation. To do this, calibrated sensor data are taken at a temporal granularity, in this case daily, to take advantage of the temporal relationships of the data. The temporal pattern-based denoising (TPB-D) approach, Figure 2, consists of a denoising step based on a singular value decomposition generated subspace, in which the in-situ calibrated LCS data are projected onto the left-singular vector space obtained from a database of data collected by reference instruments on a daily basis.

The denoised data can be further improved if the projected sensor signal is recalibrated by a linear mapping onto the left singular vector space of the reference data projection, TPB-C mechanism in the Figure, as explained in class. However, we will not do it in this homework, and we will only perform the denoising step.

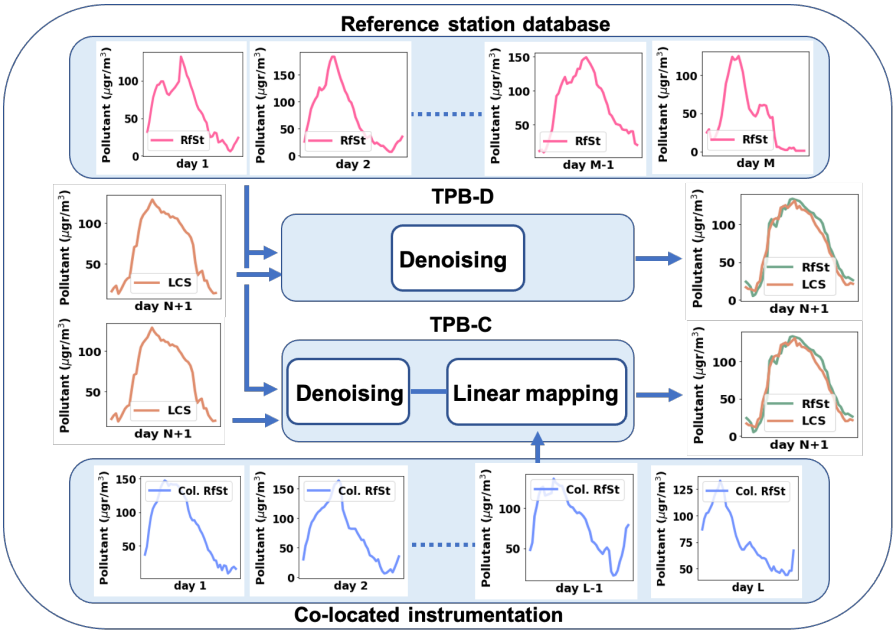

Figure 2: Denoising using a singular value decomposition (SVD).

2 Project Realization.

The data consists off a CSV file called ”data-HW2.csv”, the format being as follows:

date; Sensor_MLR_O3, Sensor_SVR_O3, RefSt

where it can be seen that the first row is a header that describes the data:

- **date:** Timestamp (UTC) for each measurement,
- **RefSt:** Reference Station O_3 concentrations, in μgr/m^3.
- **Sensor_MLR_O3:** Calibrated O_3 concentrations using MLR, in μgr/m^3,
- **Sensor_SVR_O3:** Calibrated O_3 concentrations using SVR, in μgr/m^3

You have to explain your conclusions along the homework process. Plot curves, draw tables and explain the process you follow, as if you write a technical report or a paper. Observe that in the report we are not interested in code.

2.1 Visualize your data.

The first step consists on understanding the data. For that purpose, the best approach is to plot several curves to see the temporal behaviour of your sensor data. I recommend using PANDAS as tool to handle the data.

Second, you can compare the in-situ calibration quality by calculating the RMSE and R^2 of the MLR and SVR estimated data against the reference station data.

Finally, organize your data in days (remove those days that are not complete). In the report mention how many days do you get. Again try to visualize your daily data, and plot several days in the report comparing the daily reference station data and the daily sensor estimated data (both with MLR and SVR).

2.2 Denoise the data.

Now you have to denoise the data using the eigenfaces methodology as explained in class. Here, daily data are equivalent to the faces. The database of faces now is a database of **reference station** daily data, and the sensor data are the faces that you want to denoise. You have to obtain the best **reference station** subspace using the RMSE as metric to see which is the best dimension, e.g. plot a curve with x-axes the rank of the subspace and y-axes the RMSE obtained reconstructing the LCS daily signals with each rank. Compare the RMSE with the one obtained in the calibration. For example, for each $\hat{\mathbf{x}}_i$ LCS daily data, you can reconstruct it using

$$\tilde{\mathbf{x}}_i = \mathbf{\Psi} + \mathbf{U}_\kappa \mathbf{U}_\kappa^T \hat{\mathbf{x}}_i$$

where $\hat{\mathbf{x}}_i$=$\mathbf{x}_i - \mathbf{\Psi}$, $\mathbf{\Psi}$ is the average of reference station daily data, and $\mathbf{U}$ comes from the SVD of $\mathbf{Y}$ after substracting the average reference station daily data.

Repeat the process using the Gavish *et al.* method. This method consists of obtaining an approximation for the low rank value κ in the absence of a reference instrument in place. Gavish *et al.* propose the recovery of low-rank matrices by introducing singular value thresholds. Assuming that we want to recover data using the centered matrix $\hat{\mathbf{Y}}_R$=$[\hat{\mathbf{y}}_{R_1}, \ldots, \hat{\mathbf{y}}_{R_M}] \in \mathbb{R}^{D \times M}$, where each $\hat{\mathbf{y}}_{R_i}$ is a day of reference station data, the threshold singular value is given as:

$$\tilde{\sigma} = w(\beta) * \sigma_{med} \tag{1}$$

where σ_{med} is the median of the singular values, while $w(\beta)$ is obtained as:

$$w(\beta) \approx 0.56\beta^3 - 0.95\beta^2 + 1.82\beta + 1.43 \tag{2}$$

where β=D/M, with D the dimension of daily data and M the number of days used by nearby reference stations. Finally, κ corresponds to the number of singular values that are greater than the threshold $\tilde{\sigma}$.

Draw a curve with the singular values, and which is the κ value obtained with the formula and compare it with the one using the optimization you used previously. Repeat the result of the denoising step obtaining the RMSE using the Gavish *et al.* κ value and comparing with the previous RMSE obtained with the experimental κ value.