Clustering (ab-)normal pressure signals of pipelines

CUI Yiran
Student ID. 20789781
ycuiat@connect.ust.hk
Room 3588
Department of Civil and Environmental Engineering, HKUST

Abstract

Leakages of urban water supply systems cost considerable economic damage and labor costs per year and many methods are developed to dress problems. Transient-based methods are based on human-controllable transient waves inside the pipeline to detect the location of leaks and they have many advantages such as fast, low-cost, and accurate. However, to implement these methods preliminary knowledge of pipelines is necessary, including the number of leaks, and the existence of leaks. in order to facilitate preliminary knowledge gathering of transient-based methods, the author tried to cluster data from background noises of intact pipes, background noises of leaking pipes, transient waves generated on intact pipes, and that on leaking pipes. To explore the effects of datasets, four datasets are created and they tailored measurements, response functions of the system from measurements, sorted measurements, and Truncated Fast Fourier Transform (FFT) of measurements. The dataset of truncated FFT is more efficient than others in clustering since high-frequency noises are truncated. PCA, LTSA, Laplacian LLE, and MNcut are applied here and it turns out that four clusters can be found correctly by properly choosing the dataset and method. Noise signals are significantly different from transient signals in all four datasets, but two different noises are easier to cluster by the truncated FFT dataset and they are more entangled in the dataset of response functions.

1 Introduction

With the rapid growth of urban scale and populations, water demand in urban areas is expanding. Many new supply pipes are built or connected to existing distribution networks. However, due to improper design, aging, unexpected demands and relatives, pipes may keep leaking undiscovered. The leak rate in many developed cities is around 20% to 30%, and it will cost considerable economic costs. Since most of the distribution pipes are buried in the soil, labor costs of replacement and repair are high, especially if the exact location of leakage is unknown. To locate leaks, many methods are developed, such as Smart Ball, accelerometers, transient-based methods, and so on. The transient-based methods (also known as water hammer) use the property that leaking points will reflect transient waves that can be generated by a valve closure or mechanical devices. By monitoring the reflected waves and the wave speed of the pipe, leakages can be located properly. However, due to noises inside pipes, the reflected signals are easily buried by noises. To find the reflection, it is usually implemented the cross-correlations to find the most mismatch of measured (containing leakage) signal with the intact signal. Unfortunately, in the case of unsure the existence of leakages, the cross-correlation technique can still provide a meaningless point. Therefore, it is important to study the existence of

Submitted to CSIC-5011 (2022) Final Project. Do not distribute. Any comments are welcomed

leakages to gather more preliminary knowledge about the system before implementing leak-locating techniques.

In this article, the author tried to explore the dataset to make clusters to separate signals from cases with and without leaks, so the main question this report trying to answer is whether that is a way to cluster signals with and without leaks. It is starting from classical principle component analysis (PCA), to manifold methods such as local tangent space alignment (LTSA), Laplacian locally linear embedding (Laplacian LLE), and then to NCut methods. It is found that different datasets will influence the performances of the above-mentioned methods significantly. Besides, by choosing a dataset carefully with the proper embedding method, good clusters can be found.

The second section introduces datasets used in this article with an experimental description of collecting them. Besides, a brief introduction of the methods applied is given. Then section three will show the results with discussions. The fourth section provides conclusions and further questions. Besides, all data and codes used in this report can be found on the link at the end of this report.

2 Data description and methodology

The data used is sampled from Water-Lab in HKUST [1] and then they are modified and arranged into four datasets. Four datasets are named as 'Tailored dataset $\mathbb{M}_T$', 'Response function dataset $\mathbb{M}_R$', 'Sorted dataset $\mathbb{M}_S$', and 'Truncated FFT dataset $\mathbb{M}_{FFT}$'. Expect the 'Truncated FFT dataset $\mathbb{M}_{FFT}$', all other three are of dimensions of $\mathbb{R}^{1001\times 80}$, but the 'Truncated FFT dataset $\mathbb{M}_{FFT}$' is of $\mathbb{R}^{300\times 80}$.

2.1 Data description

All of the four datasets consist of four types of data with different modifications. They are background noises inside an intact pipe, background noises inside a leaking pipe, transient wave signals of an intake pipe, and that of a leaking pipe. Transient signals are generated by a manual value closure. All four datasets have 80 columns. The first 20 columns are time series recordings of noises of an intact pipe, the second 20 columns are noises of a leaking pipe, the third 20 columns are transient signals of intact pipes, and the last 20 columns are transient signals of leaking pipes. The sampling frequency is 1000 Hz.

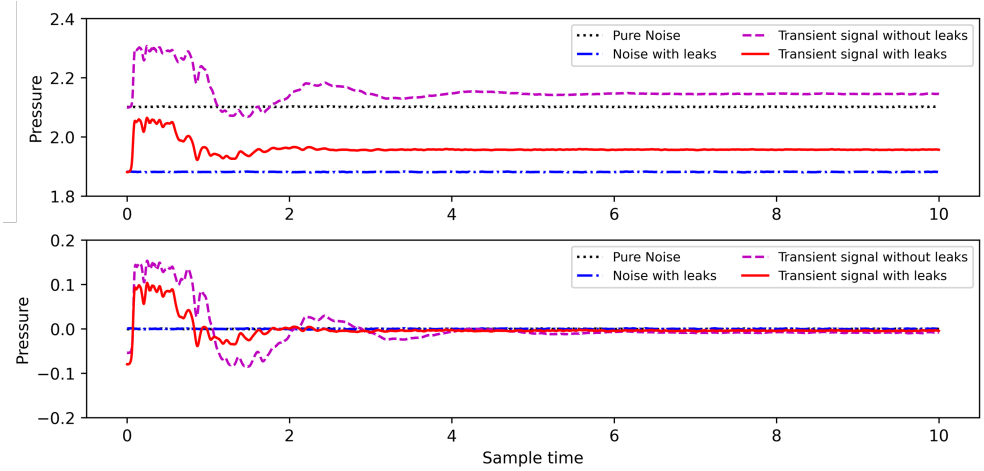

Figure 1: Example of four types of data, upper: un-normalized tailored samplings, lower: normalized tailored samplings. Sample time in unit of $10^{-1}s$

Tailored dataset $\mathbb{M}_T$: Since the valve closure is done by a friend of the author, the reaching time of the transient signal (the time from the starting recording to the transient wave reaches the sensor) will be flexible. This means that the same transient can be represented by different time series just

by different reaching times. Therefore, the maximum pressure value is found in each time series (transient waves will produce a pressure jump increasingly), and then tailor the time series from the record that is 40 records before the maximum value point to the record that is 1000 behind the starting record. Therefore each time series is tailored to a length of 1001 records and the 41^{th} point is the maximum value. Noise records were just randomly sampled for 1 second. Figure 1 upper one shows an example of four different types of tailored data. However, it is noticeable that the static state pressure (equilibrium state value) of the four cases is significantly different from each other. Even this feature will make the cluster much easier, but these differences often do not exist in the real system. Therefore, all records are normalized by subtracting the average value, and the lower part of figure 1 shows examples of the final shape of four types of data making up the tailored dataset.

Response function dataset $\mathbb{M}_R$: The response function is a name borrowed from the match field processing method since the same sampling modification method is applied. To obtain each modified time series, the original time series is 'shifted' downward and then subtracted by the original time series. To reduce the effects of different transient reaching times, the same procedures are done as introduced in $\mathbb{M}_T$ after the subtraction. Figure 2 gives an example of shapes of response function in the dataset $\mathbb{M}_R$. However, 40 columns of noise are the same as in the tailored dataset.

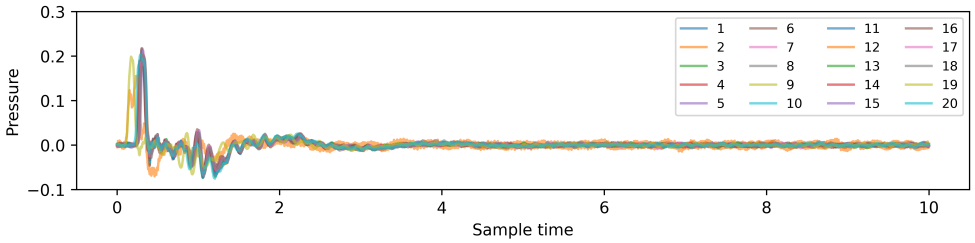

Figure 2: Example of modified transient time series in the dataset named response function

Sorted dataset $\mathbb{M}_S$: The sorted dataset $\mathbb{M}_S$ is generated from Tailored dataset $\mathbb{M}_T$. Since the maximum value of a transient wave will be influenced by noise and manual valve closure is also a part of noise influencing the shape of signals, the tailored data still contain much noise. However, it is sure that the maximum value is reached at the transient wave reaching, the minimum value is reached at the first end-reflected transient wave reaching, and pressure goes to the equilibrium state with fluctuating between the maximum and minimum values. Therefore, the tailored records are sorted from small to large to make the new sorted dataset $\mathbb{M}_S$. In other words, the $\mathbb{M}_S$ and $\mathbb{M}_T$ contain the exactly same values but the $\mathbb{M}_S$ is the ordered version of $\mathbb{M}_T$.

Truncated FFT dataset $\mathbb{M}_{FFT}$: The Truncated FFT dataset is generated by performing Fast Fourier Transformation onto the original sampled data (Sampling frequency: 1000 Hz, Sampling time: 10 s) and make truncation to keep only the 1-300 Hz amplitudes, the 0-frequency amplitude representing the constant value is dropped.

2.2 Experiment description

Experiments were contacted at Water-Lab HKUST. Figure 3 shows a brief sketch of setups of pipes. A transient generation valve is near the buffer tank and can be used to generate transient waves manually. A controllable leak labeled as LKV and at 98.42 m away from the transient generation valve. In this report, data are collected at MV4. Transient waves are generated by a friend of the author manually. In cases of a leak that exists, the leak point is fully open otherwise that point is fully closed. Once starting recording, the transient is generated manually by a valve closure. More details about the lab can be found on the Smart-UWSS website [1] and more details of experimental procedures can be found in [2].

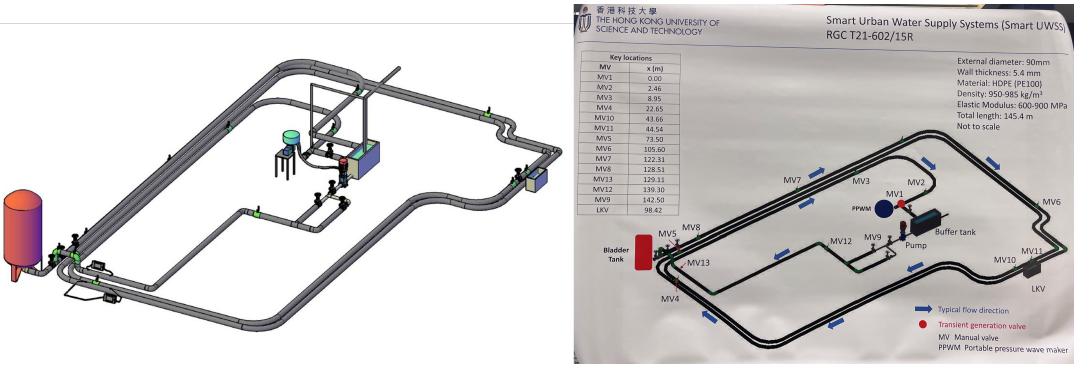

Figure 3: Pipes setups of Water-Lab in HKUST

2.3 Study questions

This report is trying to answer one question: Is it a way to separate/cluster signals from a dataset containing signals with and without leaks?

2.4 Methodology of analysis

PCA, LTSA, Laplacian-LLE, the second eigenvalue analysis (Fiedler's theorem) and eigenvector analysis (Cheeger's Inequality) are applied in this report. Here just brief introductions of them are shown, more details can be found in [3]. PCA is a classical but powerful tool in data dimension reduction.

Principle component analysis (PCA) looks for the variation direction of a dataset. In general, the first component represents the direction in which series of data have the most variation and the second component represents the direction that has the second largest variation. Besides, PCA is also a tool that can be used to find the local basis of high dimension data, since PCA captures the variation information of data.

Local tangent space alignment (LTSA) is a method of data dimension reduction by manifold learning. It uses local PCA to find the local approximate tangent space and then alignment high dimension data into lower dimension while keeping as much tangent space information of each point as possible. Since it is a manifold method, different clusters of data should be far away from each other in order to make them have different local tangent spaces to separate them. In mathematical words, the local PCA of different clusters is expected to be different from each other.

Laplacian locally linear embedding (Laplacian LLE) is a method that defines a normalized or un-normalized graph Laplacian to find the (approximately) optimum cut among the dataset. It can be seen that the two types of graph Laplacian are related and the normalized one is also related to row Markov matrix which is related to the random walk on a graph. Through the Fiedler's theorem the connectivity of a graph can be evaluated. It gives us a way to find the number of components in a graph (notice that four different types of data are expected as four different components). Besides, we can make an approximately optimal graph cut by Cheeger's Inequality, so it gives us a way to separate a graph even when they are fully connected.

3 Results and discussions

In this section, the results of the analysis are shown with discussions, starting from PCA to LTSA to Laplacian LLE. Analysis methods are applied to the four aforementioned datasets. The main conclusion that can be drawn is that there is a way to cluster different types of data but the difficulty is highly influenced by the dataset applied. Truncated FFT dataset truncated FFT dataset $\mathbb{M}_{FFT}$ has the best performance but the response function dataset $\mathbb{M}_R$ performs the worst.

3.1 PCA

Figure 4 shows the PCA analysis of the four datasets. It is clear that noise signals can be separated from transient signals in all datasets. However, the two types of noise signals are more similar to each other than the two types of transient signals. It is because the reflected transient signals are stronger than noises induced by leaks in magnitudes. However, the performance of the dataset $\mathbb{M}_R$ is worst than others, but it is expected because the subtraction will reduce both noises and needed information.

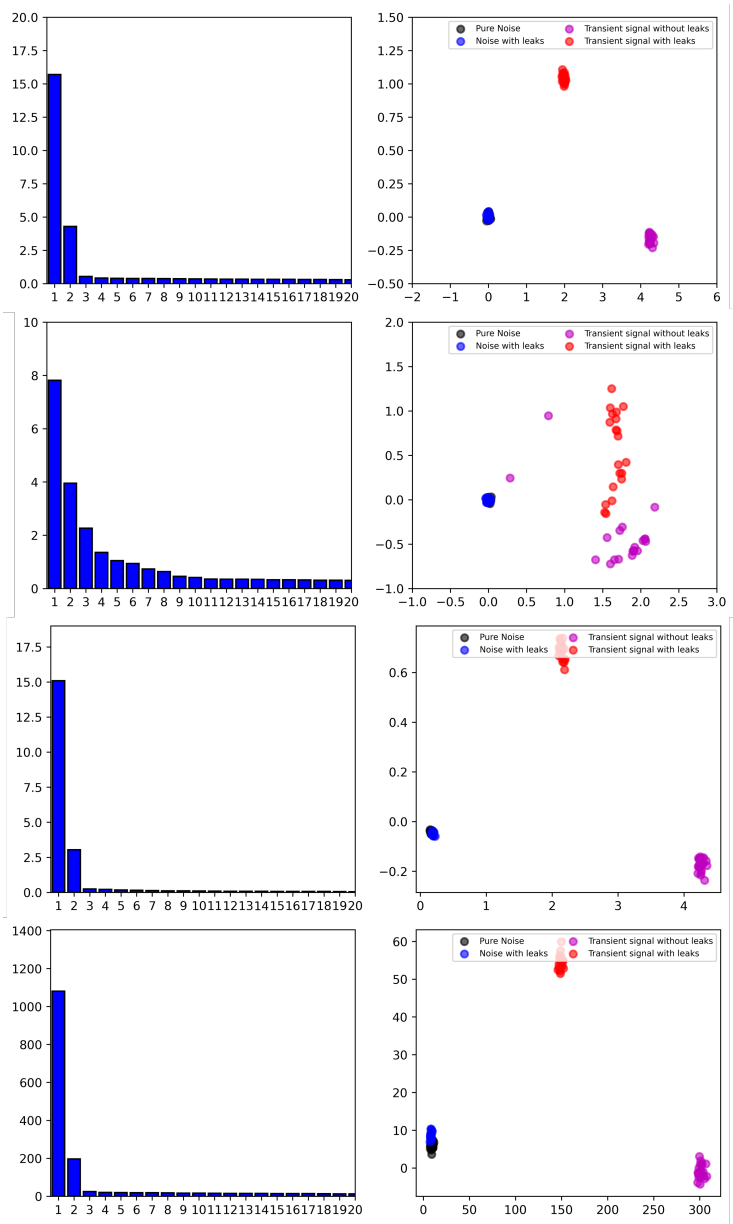

Figure 4: Principle eigenvalues of each dataset with the embedding by the first two eigenvectors. The first one on $\mathbb{M}_T$, the second one on $\mathbb{M}_R$, the third one on $\mathbb{M}_S$ and the last one on $\mathbb{M}_{FFT}$

Besides, it can be noticed that in the $\mathbb{M}_{FFT}$ dataset, two types of noise are separated. Therefore, noise data are extracted from the $\mathbb{M}_{FFT}$ dataset to perform PCA on them, as shown in figure 5. Noise data in other datasets are also extracted and analyzed by PCA, but two types of noise are entangled

with each other, so their results are not omitted here. Since the $\mathbb{M}_{FFT}$ is a truncated dataset where high-frequency information is deleted, this implies that leak noise mainly exists in the low-frequency range (in this case the frequency below 300 Hz), and it is also reasonable because high-frequency waves die out faster than low-frequency waves.

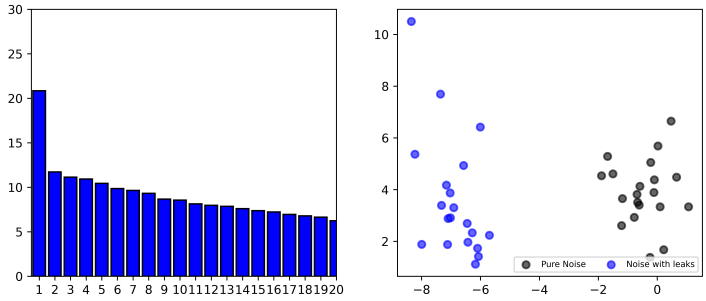

Figure 5: Principle eigenvalues of 40 noise data series from $\mathbb{M}_{FFT}$

3.2 LTSA

In general, PCA results imply that two types of transient data are distinguished from each other and also distinguished from noise data. However, noise data are still entangled. Therefore, LTSA is performed mainly to explore the ability of the manifold method to separate noise data. Actually, LTSA is also performed on not only noise data but on the whole dataset due to curiosities.

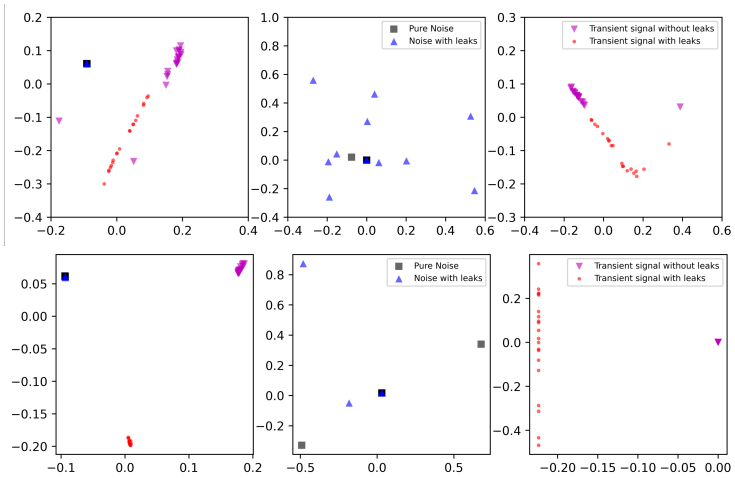

Figure 6: LTSA embedding of the whole dataset, separated noise data and separated transient data, the upper one is based on $\mathbb{M}_R$ and the lower one is based on $\mathbb{M}_{FFT}$

Only LTSA results based on $\mathbb{M}_R$ (noise data is actually the same as $\mathbb{M}_T$) and on $\mathbb{M}_{FFT}$ are shown here in figure 6, the upper one and the lower one, respectively. It can be found that LTSA can separate noise signals from transient signals as well, but two types of noise signals are still entangled. This is because the variation inside the transient data group is higher than that inside the noise data group, hence when performing the global alignment small variations of noises data are dropped, as we only use the smallest 2 eigenvectors in this case. Besides, the results of LTSA on noise data separated from $\mathbb{M}_R$ show that pure noise (the case without leak) is more concentrated than of with leak, but in the case of $\mathbb{M}_{FFT}$ both two types of noises are concentrated into one point expect four outliers. This means that with the leak existing background noise has more variations than that without leak, and those variations mainly exist in high frequency that is truncated in the $\mathbb{M}_{FFT}$. The last thing

that should be mentioned is that the embedding of transient signals without leaks from $\mathbb{M}_{FFT}$ is also shrinking into a point. It means that variations inside transient signals with leak are much higher than those inside transient signals without leak.

3.3 Laplacian LLE

The Laplacian LLE is performed and the second eigenvalues of normalized graph Laplacian and eigenvectors are also calculated to further explore the Laplacian LLE results. Figure 7 shows the results of Laplacian embedding results of a simple connected graph defined by sampled data with different settings of the number of neighbors. It can be found that two types of transient data can be well separated, except in the case of the graph generated by response function (upper-right figure) one data series of transient signals without leaks is embedded closer to data series of transient signals with leaks. This may be because of an improper subtraction of the original data. Except for noise data from $\mathbb{M}_{FFT}$, all other embeddings of noise data show that two types of noise are entangled. The successful embedding of noise data from $\mathbb{M}_{FFT}$ implies that low-frequency parts of two types of noise are distinguished but high-frequency parts are similar. Besides, results of embeddings of the whole dataset show that two types of noises are more similar to each other than two types of transient signals.

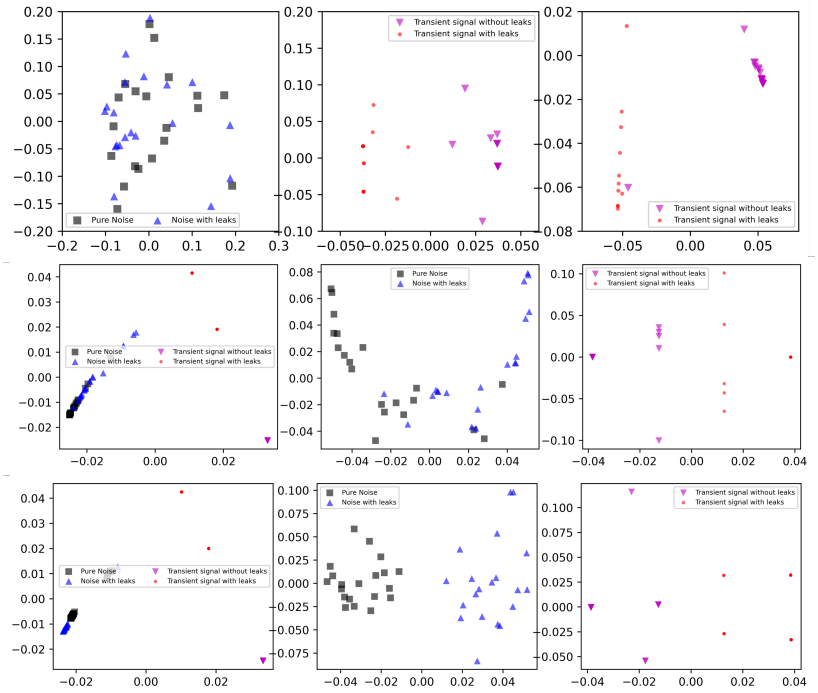

Figure 7: Laplacian LLE results of simple connected graph generated from different datasets. The first one is based on noise data from $\mathbb{M}_T$ with 5 neighbors (left), transient data from $\mathbb{M}_T$ with 22 neighbors (middle), and transient data from $\mathbb{M}_R$ with 22 neighbors (right). The second one is based on the whole data from $\mathbb{M}_S$ with 25 neighbors (left), noise data from $\mathbb{M}_S$ with 25 neighbors (middle), and transient data from $\mathbb{M}_S$ with 25 neighbors (right). The last one is based on the whole data from $\mathbb{M}_{FFT}$ with 25 neighbors (left), noise data from $\mathbb{M}_{FFT}$ with 25 neighbors (middle), and transient data from $\mathbb{M}_{FFT}$ with 25 neighbors (right)

Since the above results are generated from simple connected graphs, they can be further improved by adding weights to generate weighted graphs. By assigning different denominators to the power of exponential, neighbors who are far away from each other can be further distinguished. Parts of the results of weighted Laplacian graphs LLE are shown in figure 8. It can be seen that in this case the noise data from $\mathbb{M}_S$ are separated better than in the case of simple connected graph. However, noise

data from $\mathbb{M}_T$ are still entangled. Adding weights to neighbors can make non-similar neighbors has less weight but if different data are mixed up with each other (a point with both the same type and different types of neighbors with similar distance), this method will fail. By rearranging the number of neighbors and carefully setting γ, there do has a hope to cluster them properly.

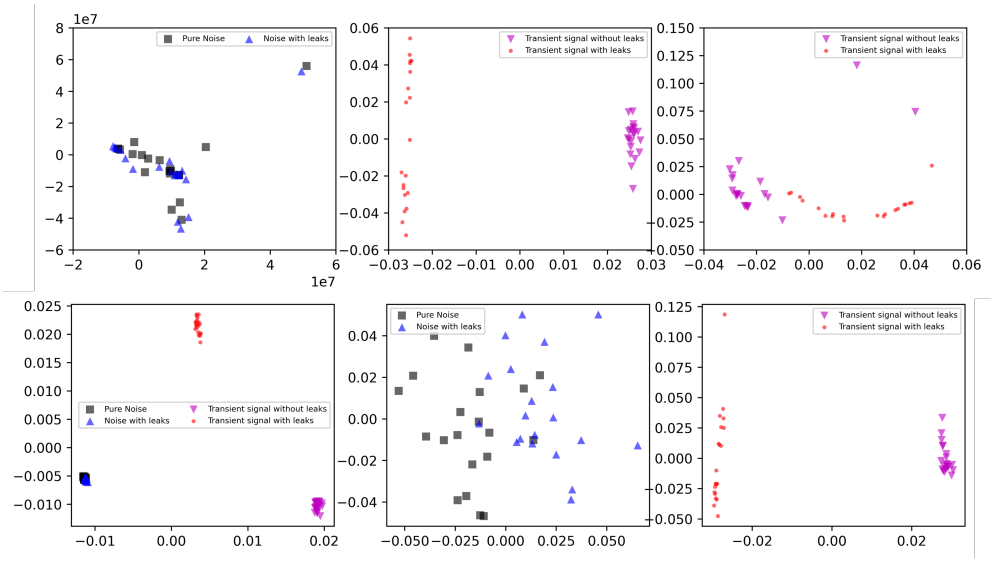

Figure 8: Laplacian LLE results of weighted connected graph generated from different datasets. The upper left: noise data from $\mathbb{M}_T$ with $\gamma = 520$, the upper middle: transient data from $\mathbb{M}_T$ with $\gamma = 0.01$, the upper right: transient data from $\mathbb{M}_R$ with $\gamma = 0.01$, the lower left: the whole data from $\mathbb{M}_S$ with $\gamma = 10^{-10}$, the lower middle: noise data from $\mathbb{M}_S$ with $\gamma = 10^{-5}$, the lower right: transient data from $\mathbb{M}_S$ with $\gamma = 0.1$, where γ is a parameter of calculating weights $e^{-\gamma\|x_1-x_2\|^2}$

To further explore the ability of graph Laplacian, eigenvalues of unnormalized graph Laplacian (simple connected) are calculated. Based on Fiedler's theorem, the number of zeros is equal to the number of components on a graph. Therefore, it is clear that in both $\mathbb{M}_T$, $\mathbb{M}_S$ and $\mathbb{M}_{FFT}$, two types of noise are connected. In four datasets transient signals generated unnormalized graph Laplacians imply that there are at least two components on the graph. However, as shown in figure 7 and 8, these components may consists of data with different types.

To say components of graphs, the normalized graph Laplacians are calculated and the first three smallest eigenvectors (the eigenvector corresponding to eigenvalue zero is dropped hence it is excluded here) are plotted. The results based on dataset $\mathbb{M}_S$ and $\mathbb{M}_{FFT}$ are shown as an example in figure 10. It is clear that if play with the whole dataset, two types of transient data are seen as two components which is what we want but two types of noise data are mixed up as one component. If the normalized graph Laplacian is built on transient types of data, two components with correct clusters can still be found. In the case of noise in $\mathbb{M}_S$, two components are found in the graph but it can be seen that each component consists no longer only one type of noise. Since all settings of them are the same but the noise in $\mathbb{M}_{FFT}$ can be clustered nicely, it suggests that the high-frequency part of background noise will negatively influence the performance of Laplacian LLE.

To further illustrate the point, figure 11 shows the approximate optimal NCut calculated from noise data from $\mathbb{M}_S$ and $\mathbb{M}_{FFT}$. Clearly, noise from $\mathbb{M}_{FFT}$ can give a nice cut where each component contains only one type of data but that from $\mathbb{M}_S$ generates a relatively poor cut even though it still cut data into two clusters. To illustrate the meaning of Ncut of noise data from $\mathbb{M}_{FFT}$, the Cheeger's constant calculated from the reordered node set ($f_1 \leq f_2 \leq \cdots \leq f_n$, $\vec{f} = D^{-1/2}v$, where D is the degree matrix and v is the eigenvector corresponding to the second smallest eigenvalue) is shown in figure 12. A clear sharp corner is found in that figure and this implies that the NCut does separate two different types of data otherwise the curve of Cheeger constant will be flattened.

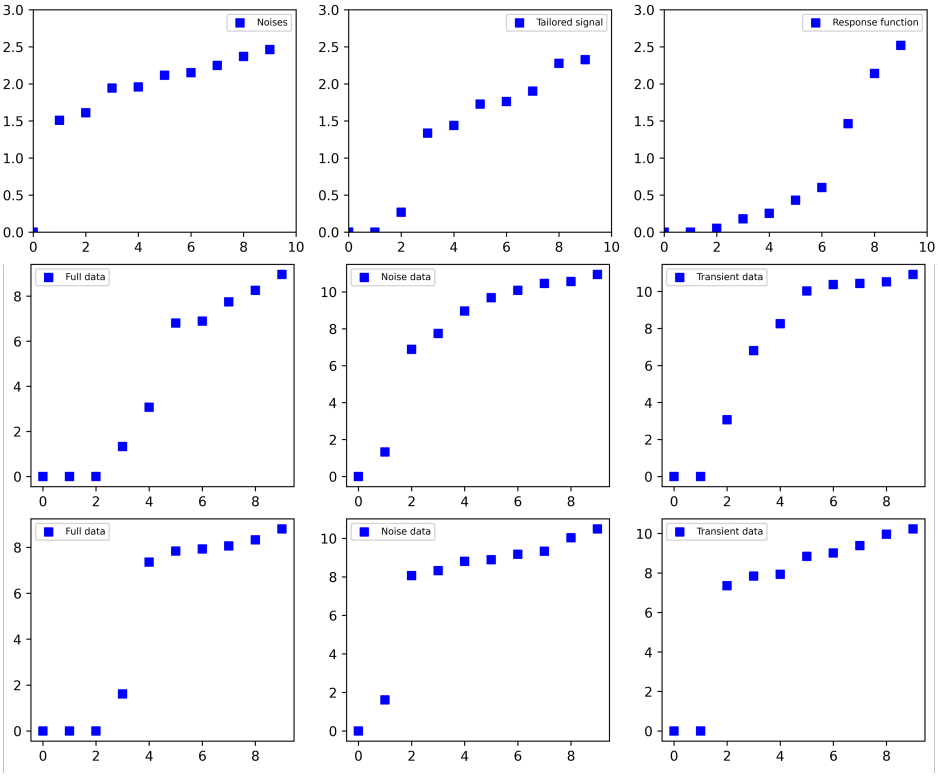

Figure 9: The ascending ordered eigenvalues of unnormalized Laplacian graph (to apply Fiedler theory). They are generated based on search neighbors with the number of 10 and calculated based on: noise data from $\mathbb{M}_T$ (the upper left), the transient data from $\mathbb{M}_T$ (the upper middle), the transient data from $\mathbb{M}_R$ (the upper right), the whole dataset of $\mathbb{M}_S$ (the middle left), the noise data from $\mathbb{M}_S$ (the middle central), the transient data from $\mathbb{M}_S$ (the middle right), the whole dataset of $\mathbb{M}_{FFT}$ (the lower left), the noise data from $\mathbb{M}_{FFT}$ (the lower middle) and the transient data from $\mathbb{M}_{FFT}$ (the lower right)

4 Conclusions

This report is trying to answer the question 'Is it a way to separate/cluster signals from a dataset containing signals with and without leaks?'. Pressure signals are sampled from Water-Lab in HKUST, and four different datasets are generated based on them. Both PCA, LTSA, and Laplacian LLE are done to answer the question. It turns out that there do have a way to cluster different types of signals, but the final cluster highly depends on the quality of the dataset applied and the pre-knowledge of the processed data. With a proper understanding of the data (at least the magnitude of different phenomena), graph Laplacian provides a way to achieve the job.

Supplement

All data are collected by the author and his friends, so it is decided to publish the data for educational purposes. All data and codes can be found here: `https://drive.google.com/drive/folders/1ygImku3AsiVB9mamjkzqyPsLHnm9X9nt?usp=share_link`

References

[1] Smart UWSS. (2023, May 17). *Controlled facility 1: HKUST lab pipeline.* Smart Urban Water Supply System. URL:`https://sites.google.com/view/smartuwss/facilities/Facilities_1`

[2] Wang, X., Lin, J., Keramat, A., Ghidaoui, M.S., Meniconi, S., & Brunone, B.(2019c). Matched-Field Processing for Leak Detection in a Viscoelasticity Pipe: An Experimental Study. Mechanical Systems and Signal Processing, 124, 459-478

[3] Yao Yuan (2023). *Geometric and Topological Data Reduction: A Mathematical Introduction to Data Science*. URL:https://yao-lab.github.io/bookdatasci/

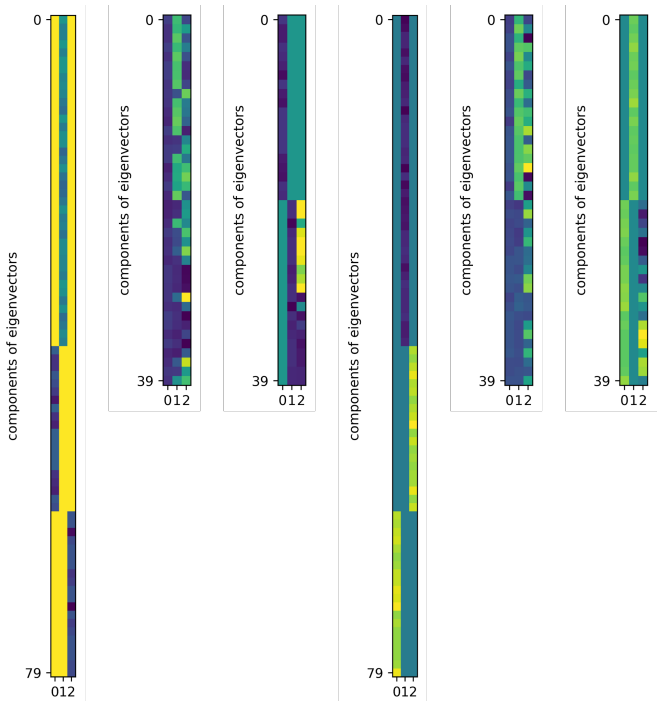

Figure 10: The first three smallest eigenvectors of normalized graph Laplacian with different colors representing values, positive number are marked as green, negative number is marked as blue and zero is marked as yellow, darker color means higher absolute value. The leftest (first) based on the whole dataset of $\mathbb{M}_S$, the second one from noise data from $\mathbb{M}_S$, the third one from transient data from $\mathbb{M}_S$, the fourth one from the whole dataset of $\mathbb{M}_{FFT}$, the fifth one from noise data from $\mathbb{M}_{FFT}$, and the last (sixth) one from transient data from $\mathbb{M}_{FFT}$

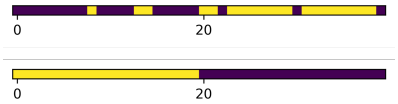

Figure 11: Approximate optimal NCut based on Cheeger's constant, calculated from the noise data from $\mathbb{M}_S$ (upper), and noise data from $\mathbb{M}_{FFT}$ (lower), two colors represents two different clusters of data after the cut.

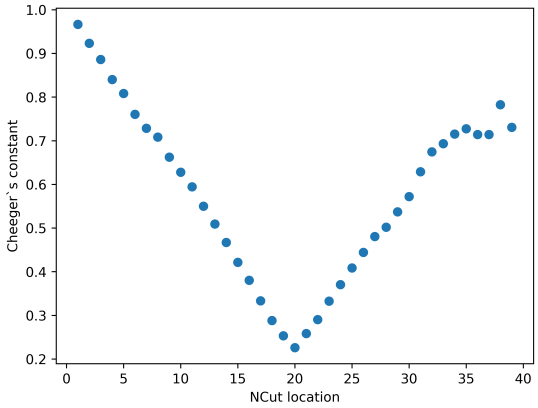

Figure 12: The Cheeger constant calculated from the reordered node set of noise data in dataset $\mathbb{M}_{FFT}$