Optimization of Biosensor Waveform Preprocessing

Siemens Healthineers

Report Week 6

Dylan Longert

Jinxin (Jessie) Wang

Nan Tang

Nayeli Montiel Rodríguez

Han (Eden) Chen

Project Progress Summary

This week, we had modified the way we visualized the extracted window data and FPCA scores as per the client's suggestions. Originally, we split the data into two main categories: blood and aqueous to visualize the FPCA scores by binning four important features: small fluid type, ambient temperature, fluid temperature, and test card age. However, the bins for each feature are not balanced, which makes it difficult to capture the true insights from the data. Therefore, we decided to perform the Functional Data Analysis (FDA) separately for each feature and to balance each feature by resampling for each analysis (We may show some plots as an example). In other words, for each FDA analysis to compare the characteristics of waveforms between system 1 and system 2, we included the necessary preprocessing steps, window visualization, FPCA components, and FPCA scores visualization, and plotted confidence intervals by bootstrapping the first component. Similarly, we also decided to complete the slope (coefficient of regressing timestamp on sensor values) comparison of the aggregated mean function for all features because clients thought this approach was very straightforward to characterize the differences between the two systems. At this stage, the Siemens team is quite satisfied with the jobs we did for the main task of Waveforms characterizations and advised us that we can start looking at our secondary task: window optimization.

Additionally, we ended up landing on two different approaches to scaling the data. One is subtracting the column mean from the raw sensor data, and the other method is bringing the sensor value of the starting point to the zero value and aligning the remaining data within each specific window. Clients find it might be interesting that we keep more than one approach for scaling data, so they can replicate them both for all the other sensors. As per discussed last week, we are also providing some interpretations of our results for the FDA and Bootstrap after we scaled the data.

Functional Data Analysis high level interpretation

The waveforms under study consist of electrical signal measurements in millivolts (mV) taken during two-time windows, the calibration window and sample window. These windows are periods where the signal values are relatively stable. These waveforms include several features intrinsic to the process, with the following four the most important characteristics:

- Type of fluid: the sample injected into the card read by the system.
- Card age: the age of the card in days at the time of the test.

- Ambient temperature: the temperature of the entire system (card, reader, host and fluid).
- Fluid temperature: the temperature of the sample fluid when available, otherwise the ambient temperature if the sample temperature is not provided.

Due to the complexity of the data, we employed Functional Principal Component Analysis (FPCA) to provide a low-dimensional representation of the patterns of the multiple time series. Similar to how standard PCA identifies components that capture dominant patterns of variability in data within ordinary Euclidean space, FPCA attempts to find component functions capturing the dominant patterns of variability in functional data [1].

The waveform datasets were first balanced using sampling with replacement. This ensured an equal number of waveforms for each additional feature and bin, preventing the dominance of waveforms with certain characteristics (e.g. we considered 72 waveforms in each bin for fluid temperature: below 20 °C, 20°C-26°C, and above 26°C). After balancing, we ran FPCA on four datasets, each balanced by fluid type, card age, ambient temperature and fluid temperature. This process was performed on the raw data centered and the zero-aligned centered data. Centering was done within the FPCA() object from the 'scikit-fda' library, which subtracts the column-wise mean (mean function) [2]. This resulted in a total of eight separate Jupyter notebooks.

The raw data represents each waveform at its original level. For example, at the beginning of a window (timestamp 0), the signal values range between -160mV to 100mV and remain “apparently” constant over the time, resulting in flat curves when plotting various waveforms together. By contrast, zero-aligning the data involves subtracting the values in the first column (timestamp 0) from all other columns, bringing all waveforms to the same starting point. This preserves the shape of the waveforms and allows us to visualize them as a trending time series, with some showing upward trends and others showing downward trends. This can be interpreted as an increase or decrease from the electrical signal over the time.

In both cases, centering is a common preprocessing step before applying FPCA. It removes the average level of each signal value at each time point, shifting the data so that each column has a mean of zero. This can be visualized as shrinking the data closer to the mean function. In some cases, centering can change the shape of the waveforms because it highlights the fluctuations or deviations from the average behaviour at each time point. Once centered, the waveform might show more pronounced peaks and troughs relative to the new baseline, effectively altering the visual impression of its shape.

Based on the variance explained by each component, we retained one component for analysis, which describes around 99% of the variance across all waveforms within each dataset. However, we display component two (FPC2) which captures less than 1% of the variation in both cases because together they facilitate the representation of the scores in 2 dimensions.

The FPCA technique decomposes the waveforms into a mean component and functional principal components(FPC1 and FPC2), each weighted by their contribution to the waveform. The FPCs and their weights (scores) together capture the waveform fluctuations [1]. FPC1 represents variations around the mean function, while FPC2 is the trend in case a) and noise in the waveforms in case b). This can be illustrated in Figure 1.

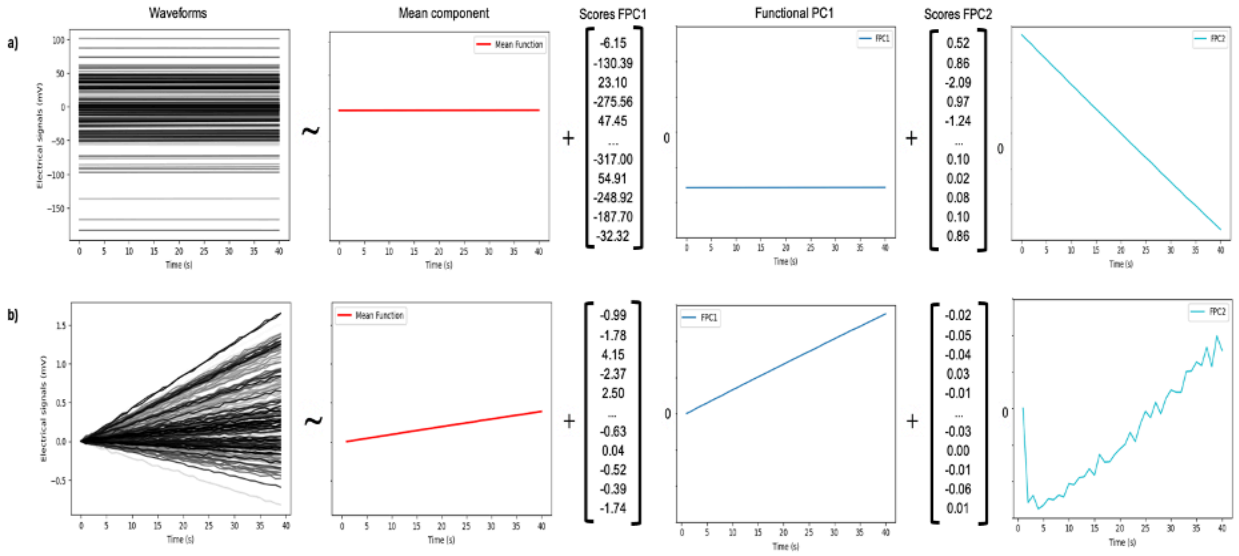

Figure 1. Waveform decomposition a) for the raw data centered and b) for the zero-aligned data [3].

The scores for the dataset balanced by fluid temperature within the cal window are displayed in Figure 2. It is evident that for the raw centered data, the contribution to the first functional principal component (FPC1) is relatively larger than to the second (FPC2). This is because FPC1 captures the major variability in the data. Additionally, large positive scores on the y-axis, such as point (a), correspond to waveforms with high positive electrical signals, while low positive scores (d) correspond to waveforms with low negative electrical signals. Scores around 0 on the y-axis represent waveforms with a signal value level around zero, such as points (b) and (c).

In the case of the zero-aligned data representation, we can identify an outlier labeled (c), which shows a decline at timestamp 25. Large negative scores (a) correspond to

waveforms with an evident upward trend, and large positive scores (d) correspond to waveforms with an evident downward trend. Scores around 0 on the y-axis are close to the mean function (b).

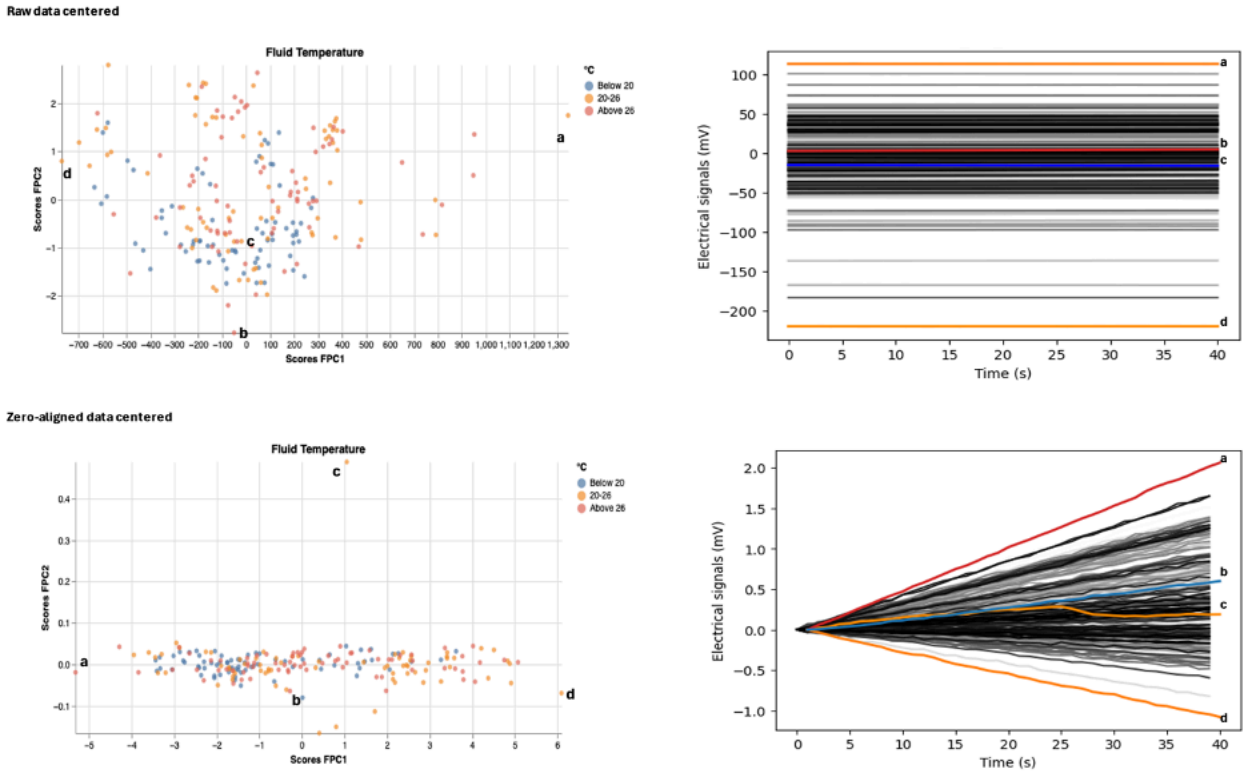

Figure 2. Scatterplots of the scores from the first two components, with FPC1 scores on the y-axis and FPC2 scores on the x-axis. The dataset balanced by fluid temperature in the cal window is used, with raw data centered at the top and zero-aligned data at the bottom. Each point represents a waveform and is color-coded by binned temperature. The letters a, b, c, and d indicate specific points that correspond to the original waveforms in the second column plots.

Bootstrap

After balancing the data, we want to assess the robustness of the first component across different sub-groups. As depicted in the plots below, the first plot represents the 95% confidence interval, while the second and third plots show the functional boxplots after balancing the data by Fluid Temperature (the red dashed line represents outliers and the red area represents the box region), there are some interesting findings:

- The first component varies across different sub-groups.

- In the Cal window, the first component can be considered similar as indicated by the presence of outliers among the lines.
- However, in the Sample windows, the first component appears to be less stable in both systems with the red box region. But it shows similar patterns in both systems.

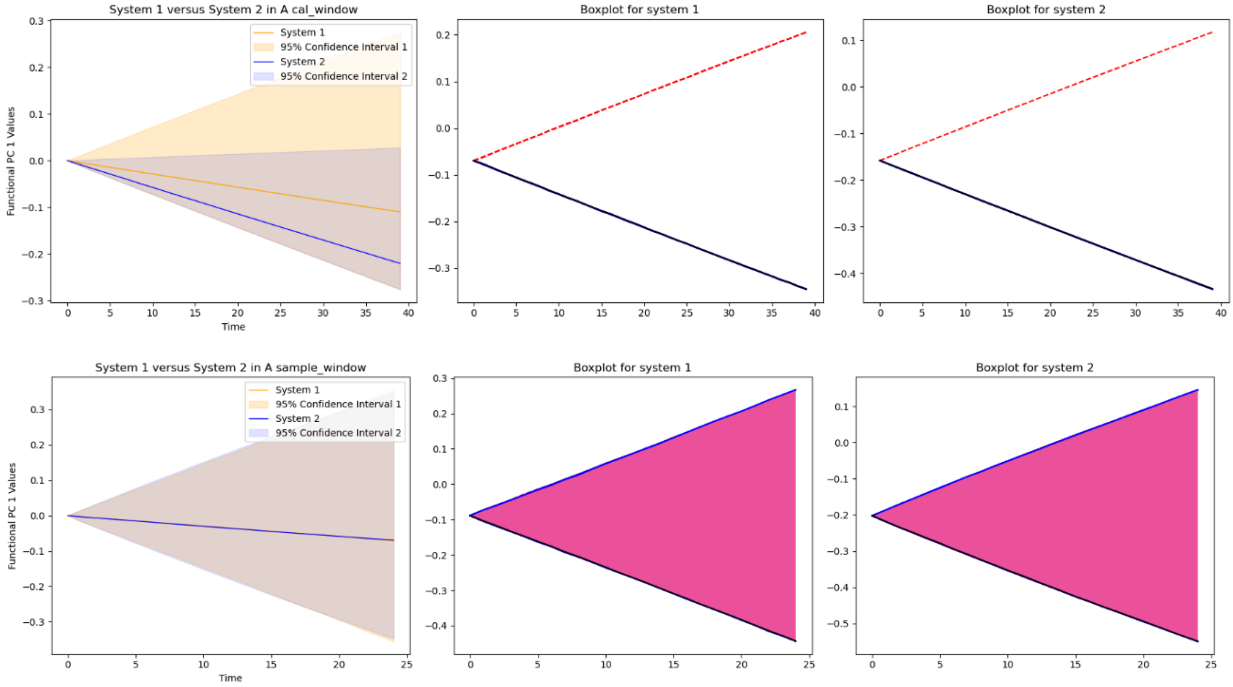

Figure 3. Confidence interval and boxplots for the first component. (The plots above are in the Cal window; The plots below are in the Sample window)

Next Steps

Overall, for the remaining weeks, we will separate into two subgroups.

One subgroup of three will still focus on the main task and complete all Python scripts to make sure all information and codes are in good order and with high readability. We will set up a folder structure with pipeline documentation to organize all coding files and upload this file into Siemens cloud drive for them to review first once the draft pipeline folder is done.

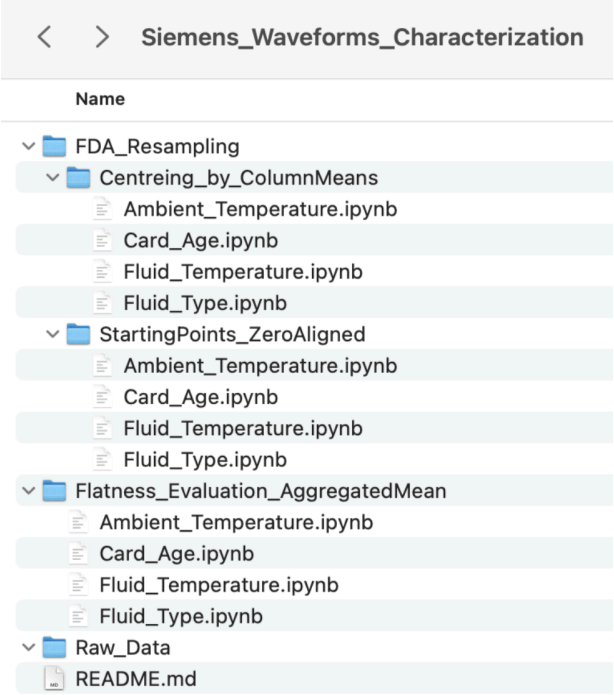

Figure 4. FD

The other subgroup of the two will try to optimize the cal window and the sample window of Sensor A and Sensor B in System 2. Firstly, we will create a series of functions that could calculate the difference in regression slopes (shown by Dylan last week) between System 1 and System 2 by changing the window placement positions in System 2. Then, we will use some optimization algorithms based on the results we get in the previous step to find the calDelimit/sampleDelimit which can provide the slope of the improved system 2 window curves closest to the one for system 1. Finally, we will use experiment design methodologies to divide data into separate groups to see the applicability of the optimal calDelimit/sampleDelimit.

References

[1] Corson Areshenkoff, Daniel J Gale, Dominic Standage, Joseph Y Nashed, J Randall Flanagan, Jason P Gallivan (2022) Neural excursions from manifold structure explain patterns of learning during human sensorimotor adaptation eLife 11:e74591

https://doi.org/

[2] GAA-UAM, "skfda/preprocessing/dim_reduction/_fpca.py," [Online]. Available: https://github.com/GAA-UAM/scikit-fda/blob/develop/skfda/preprocessing/dim_reduction/_fpca.py. [Accessed: June 1, 2024].

[3] Zhu, R.J.B., Wei, XX. Unsupervised approach to decomposing neural tuning variability. *Nat Commun* **14**, 2298 (2023). https://doi.org/10.1038/s41467-023-37982-z