Assignment 1 Report

DATA MINING

CSE 572: Fall 2019

Submitted to:

Professor Ayan Banerjee
Ira A. Fulton School of Engineering
Arizona State University

Avinash Sivaraman (1215054529 - asivara6@asu.edu)
Kaviya Kalaichelvan(121512889 - kkalaich@asu.edu)
Ashwin Karthik Ambalavanan (1215105307 - aambalav@asu.edu)
Aishwarya Sankararaman(125102941- asanka17@asu.edu)

October 8, 2019

1. FEATURE EXTRACTION

We have split the data for each day (150 mins) into 5 intervals of 6 values each (every 30 mins) and extracted features for each interval separately.

1.1 Interval-wise Mean / Shifted Mean: Take the mean of the CGM values in each interval. This gives 5 feature columns each representing the mean for the corresponding interval. We also create 5 new features by shifting the 5 mean's by 1 such that every row (except the first row) also contains the mean of each interval from the previous day.

1.2 Count Num Local Maxima / Minima: A Local Maxima in a sequence of values is a value with lower values on either side and a Local Minima is a value with higher values on each side. For each interval in a day, we count the number of local maximum values and local minimum values in that interval. Thus, we'll get 2 counts for each interval and 10 counts per day.

1.3 CGM Displacement and Velocity: Take the difference of CGM values between all pairs of consecutive timestamps. Calculate the total sum and mean of all differences in an interval which corresponds to CGM displacement and CGM velocity respectively. This gives ten feature columns each representing CGM displacement and velocities for the corresponding interval.

1.4 Global Maximum : Global maximum refers to the maximum value of high among all the low-high-low series of CGM values. After calculating the local maxima of all intervals for a day we can calculate the global maximum by obtaining the maximum value from this list of local maximas and its corresponding index. We then determine the interval in which this index is located. This gives two feature columns representing the value of global maximum and the interval of its occurrence.

2. INTUITION FOR EACH FEATURE

2.1 Interval-wise Mean / Shifted Mean: The intuition behind choosing this feature is that the patient typically consumes lunch at the same 30 minute time frame every day and hence the model would be able to take into account the current mean and previous mean of each interval and successfully predict the interval at which the meal was consumed. Moreover, increase in mean CGM level directly correlates with our goal of finding the interval in which lunch is consumed i.e. more the mean CGM of an interval during a day, more the probability that the patient consumed lunch in the current or previous time interval.

2.2. Count Num Local Maxima / Minima: The intuition behind choosing this feature is that once the patient consumes food, the direction of CGM will change by going up

(increasing) but it might not steadily increase throughout and might incur local maximum and minimum along the way to attain the Global Maximum. This would induce multiple local maximum and minimum values and counting it would help find the interval in which lunch is typically consumed i.e. the interval in which there are multiple occurrences of local maxima and minima is probably the interval in which lunch was consumed.

2.3 CGM Displacement and Velocity: The intuition for choosing this feature is that generally CGM level is moderate and it goes below normal when insulin is injected. It then spikes up by a large value when food is taken. So the difference between these extremes is high. So the interval where these extremes belong have the highest sum of differences translating to highest CGM displacement. The mean of differences of CGM values in an interval gives the CGM velocity for that corresponding interval. Thus the interval with highest CGM displacement and velocity denotes the interval of food consumption which is our objective.

2.4 Global Maximum: CGM values occur in trends of low-high-low. The intuition for choosing this feature is that one of these low-high-low set corresponds to the CGM values of insulin intake, food consumption and regulation of glucose level by insulin respectively. Finding the interval of this set gives the time of food intake by the patient which satisfies our goal.

3. FEATURE VALIDATION

For the exact values and for all patients, refer the excel sheet attached. Here we show values for Patient 1 alone.

3.1 Interval-wise Mean / Shifted Mean:

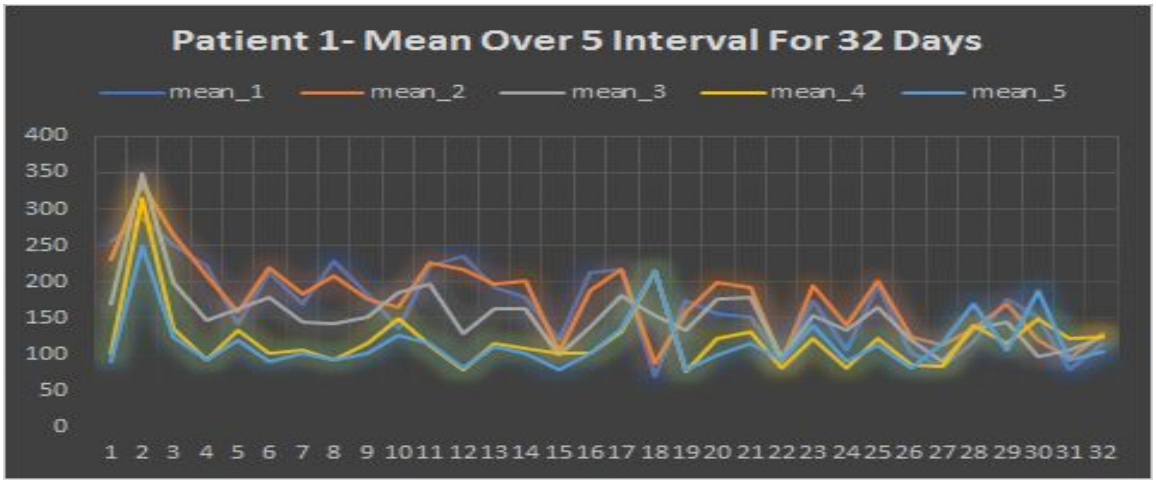

Fig 1: Mean over 5 intervals for 32 days for Patient 1

By seeing the mean over the 5 intervals for each of the 32 days as plotted below, we cannot identify one particular interval that the patient has continuously consumed food in. This goes to show that either the patient's eating habits are irregular everyday or his CGM levels are highly sensitive to even small amounts of food or in the worst case, a defective sensor. But ignoring the worst case scenario, the data does give us an insight that the mean of interval 5 is mostly lower than other intervals which means that the person probably consumed insulin before this interval or did not consume any food all day before interval 5.

3.2 Count Number of Local Maxima / Minima:

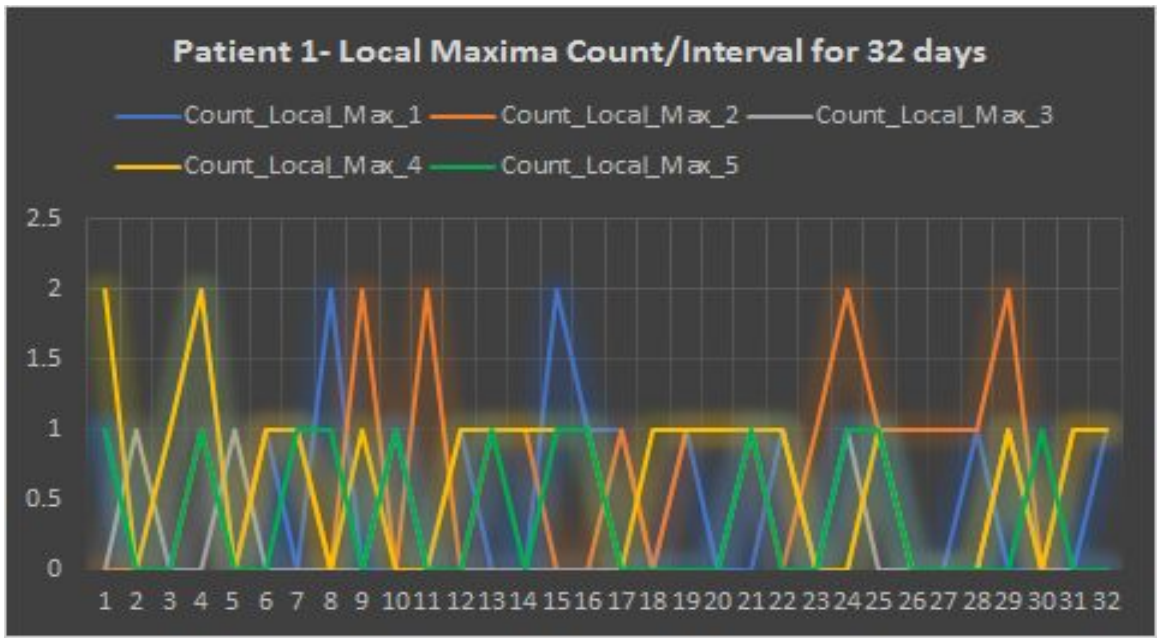

Fig 2: Local Maxima Count over 5 intervals for 32 days for Patient 1

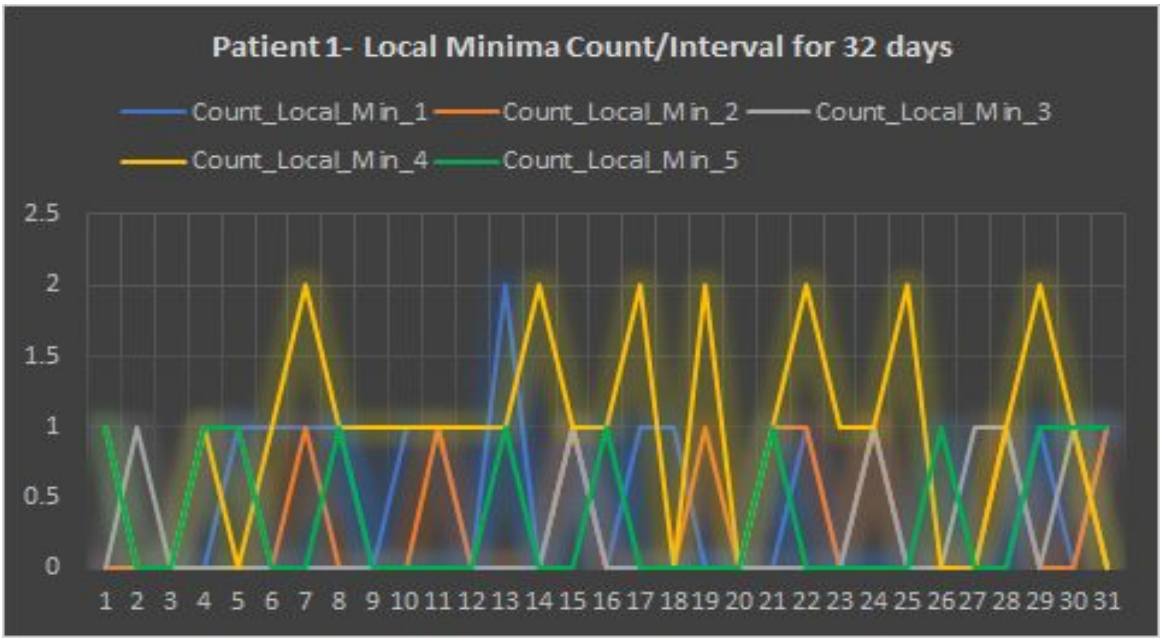

Fig 3: Local Minima Count over 5 intervals for 32 days for Patient 1

From the two plots we can see that there exists at least one interval every day which has both local maxima and minima. Moreover, there exist some days where there are only local maximums in an interval but no local minimum and vice versa. By just looking at the plots it is not obvious to choose just one interval as the interval in which food was consumed in but the varying data for count in each interval can be a useful factor in resolving confusion between two intervals when there is an equal probability of lunch being consumed in either interval. Thus the model would assign weights appropriately to the count in order to resolve conflicts.

3.3 CGM Displacement and Velocity:

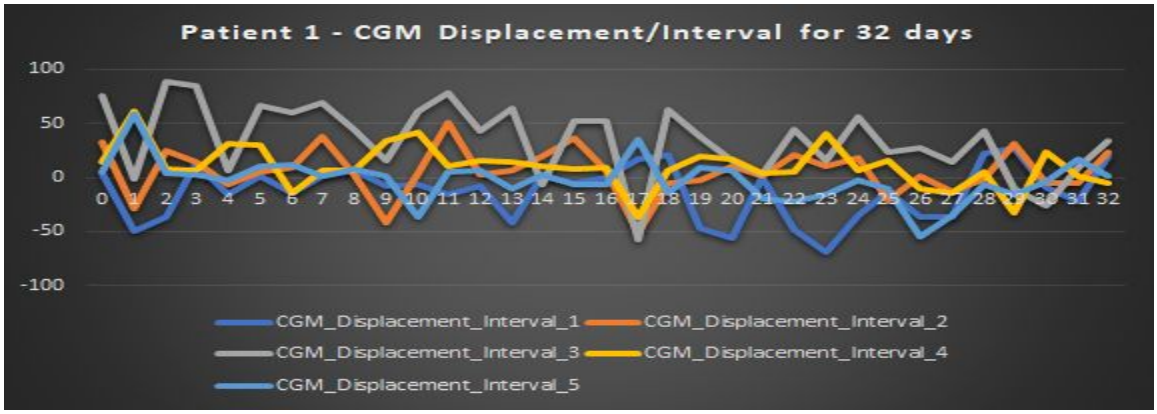

Fig 4: CGM Displacement over 5 intervals for 32 days for Patient 1

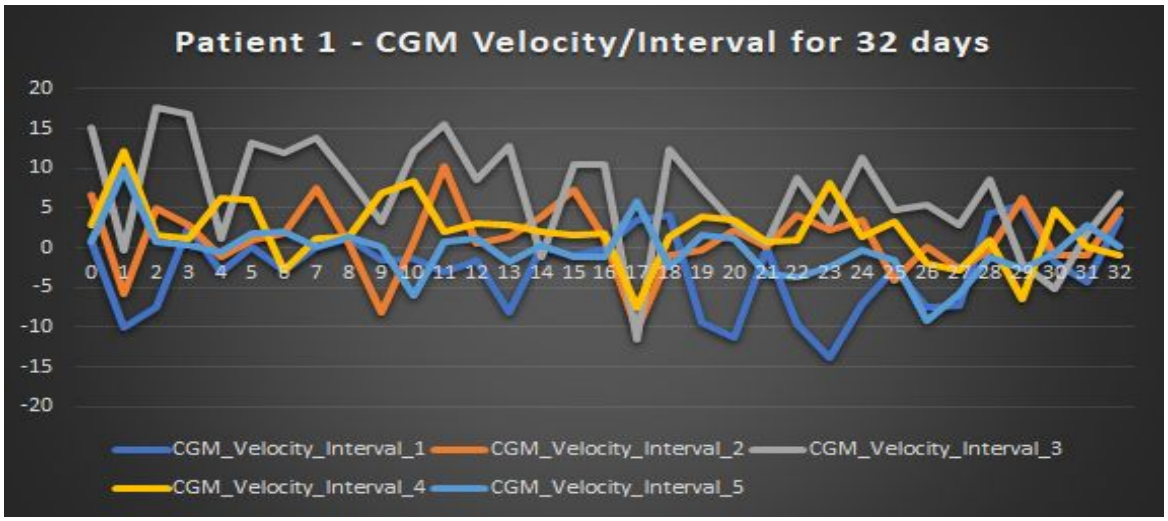

Fig 5: CGM Velocity over 5 intervals for 32 days for Patient 1

From the two plots we can see that there is one interval with a large variation in CGM values. This means that variation is accounted by the sum and mean of the differences of consecutive CGM values is high in this interval. Thus the probability that the food was consumed by the patient in this interval is high compared to the other intervals.

3.4 - Global Maximum:

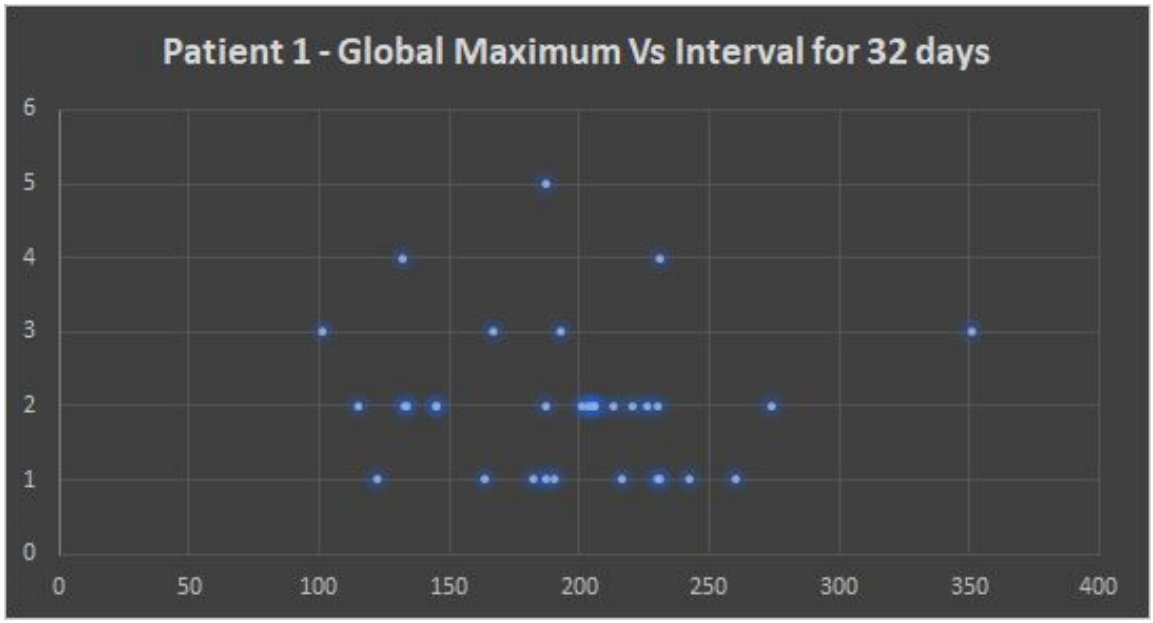

Fig 6: Count of Global Maximum per Interval for 32 days for Patient 1

From the above plot, we can infer that the Interval 2 has the maximum number of global maximums. This indirectly translates to the fact that the CGM value was highest in this interval and this occurred in a low-high-low pattern where the former low refers to insulin intake and the later low value infers to insulin regulated CGM value. Thus from this we can determine the interval in which the patient has consumed food for the majority of days.

We have also considered difference between local maxima and minima as a feature to feed as an input to PCA along with the other features.

For each interval, we find the Local Maxima and the immediate Minima before it and take the difference between the two values. The intuition behind this is to capture the maximum increase in CGM level per interval in each time series. The CGM levels should typically increase the most in the interval in which food is consumed and hence this feature would play an important role in identifying that interval distinctly.

4. FEATURE MATRIX CREATION:

A feature matrix is created from the features extracted from the data. We extracted 33 features from the data and we have 210 time series when you combine all the patients data. We created the feature matrix of size 210 x 33 which will be used for the feature selection.

5. FEATURE SELECTION

Principal Component Analysis (PCA) is used as the feature selection to select top 5 principal components which are ranked according to importance through their variance and each variable contributes with varying degree to each component. Once we extract the top features which contributes the most, we can use it as our new features instead of the original ones.

We used sklearn for performing the PCA operations. Before starting the PCA process, we need to normalize the data. We used the MinMaxScalar to normalize the data. We normalized all the feature to fit in range of 0 and 1 such that all the features are given equal importance. Once we normalize the features, we pass the normalized data and the k values of 5 to the PCA function in sklearn which returns principal components for each time series.

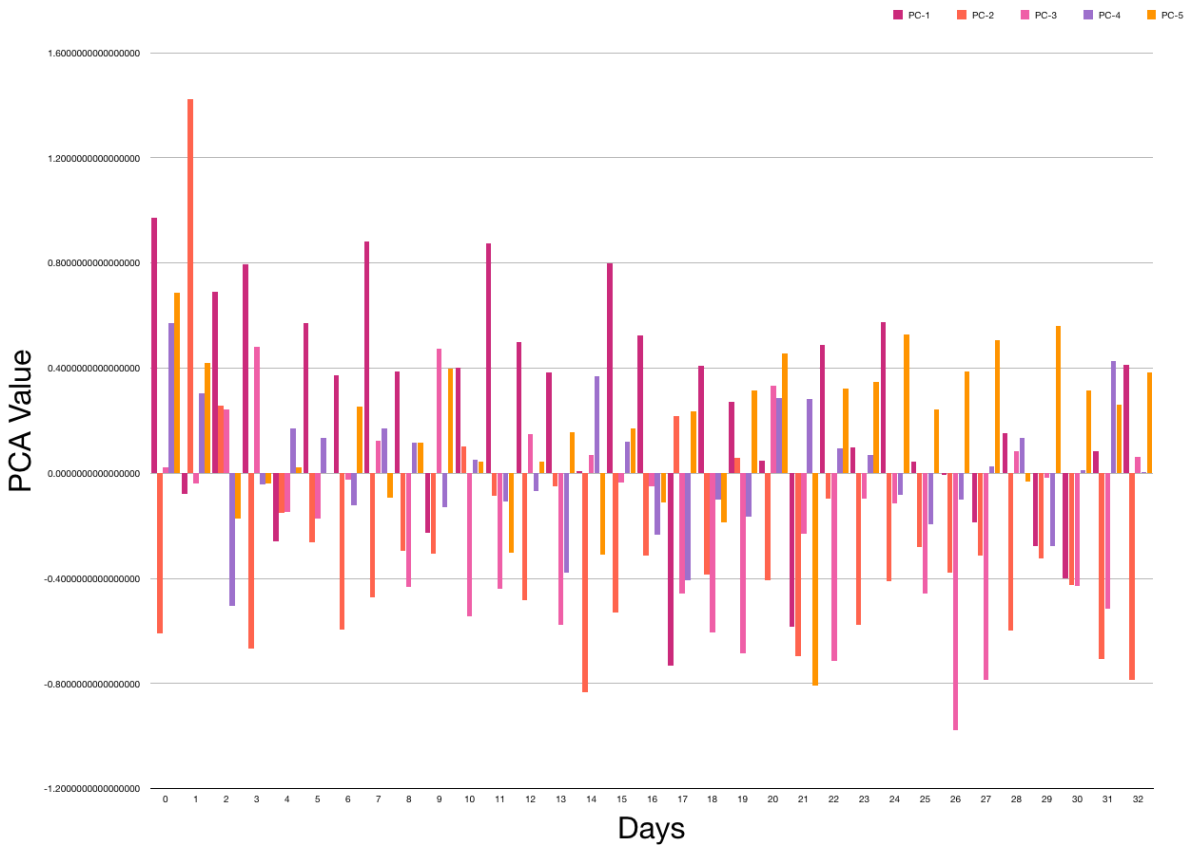

Fig 7: PCA Values for each component over 32 das

PCA class in sklearn also provides you the variable called components_. It is gives the contribution of each features in the 5 principal components which is selected the PCA. We make use of this variable to find the top 5 important features among the 33 features we extracted from the data. The variable components_ is of size 5 x 33 with principal components as rows and features as columns. The weight of each features are calculated by summing up each column to get the overall weight of each features. Once the overall weight is calculated, we sort the features in descending order to select the top 5 features.

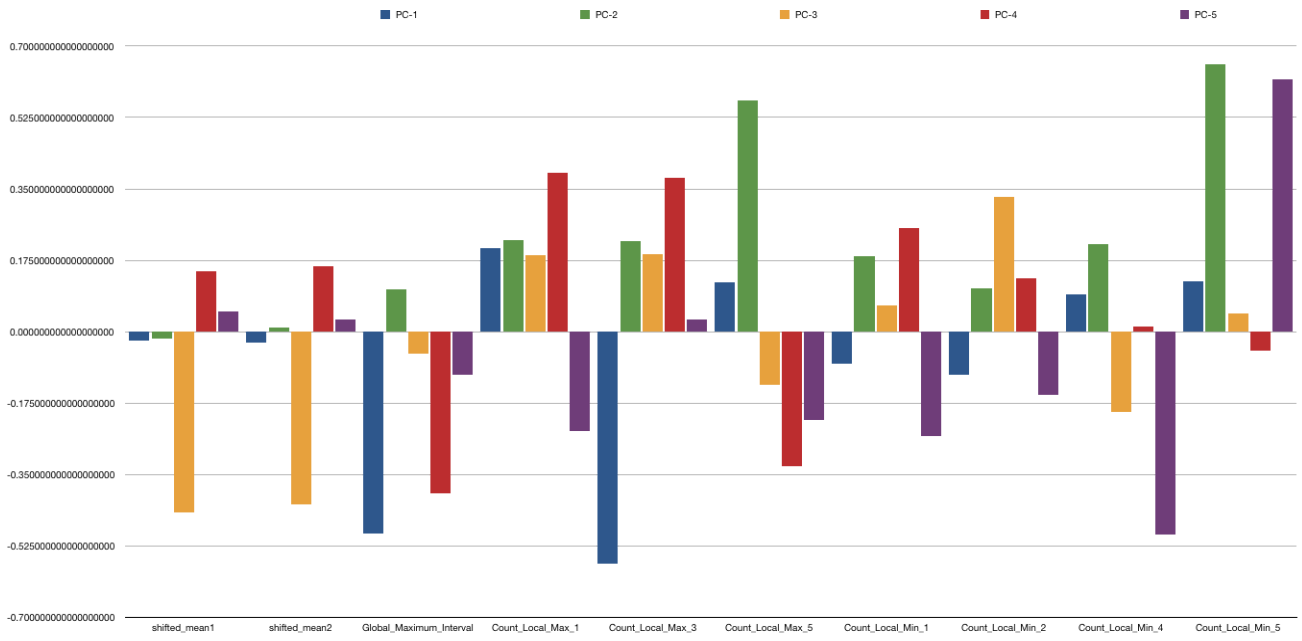

Fig 8: Importance / Contribution of the top 10 features to each of the Principal Component

The top 5 features that are chosen by PCA are -

Count_Local_Min_5	1.489639
Count_Local_Max_3	1.387756
Count_Local_Max_5	1.364914
Count_Local_Max_1	1.249771
Global_Maximum_Interval	1.152316

The total contribution of the top 5 principal components is 82% which represents the features

Count Local Min_5:

In the graph above, Count_Local_Min_5 contributes most to PC-5 and PC-2 compared to the other top 10 features returned by PCA. This means that this feature contributes more to these two directions / components compared to other features. Furthermore, since this feature contributes to two different components, the variance and importance of the feature is quite high. In an intuitive sense, the count of the local minima tends to be high in intervals farther from the interval in which lunch is consumed. Hence, PCA might be implying the same in a mathematical sense.

Count Local Max 3:

From the graph, We notice that Count_Local_Max 3 has high variance and contributes the most to the Principal Component PC-1 which has the total variance of 28% of the data. This feature is skewed a lot in the direction of the PC-1 component. Hence, It is chosen by PCA as one of the top 5 features.

Count Local Max_5:

Count of local max in an interval is directly proportional to the count of local mins in an interval. In the graph above, Count_Local_Max_5 contributes most to PC-2 compared to the other top 10 features returned by PCA. Which is proportional in terms of contribution of count of Count_Local Min_5. This means that this feature contributes more to this direction/components compared to other features. Furthermore, since this feature contributes only to one component common to Count_Local Min_5, as the number of local maximums in an interval is half the number of local minimums in an interval. In an intuitive sense, the count of the local maxima tends to be high in interval farther from the interval in which lunch is consumed. Hence, PCA might be implying the same in a mathematical sense.

Count Local Max_1:

This feature contributes equally in all directions to all components magnitude wise and it contributes most to PC-4. Being the end of the 2.5 hour interval, the CGM levels seem to climb higher and hence there seems to be a higher variation in Local Maximum Count.

Global Maximum Interval:

Since the number of global maximums in each interval varies by a large amount. For example, in the fourth and fifth, the number of global maximums tends to 0 for most days, but whereas for interval two, it tends to larger value. Thus this feature is chosen by PCA among the top 5 features.