

CSE 572 Data mining Project 1

By Arun Deepak Chandrasekar

Task1 The four types of timeseries features extracted from CGM data and timestamp are as follows:

- 1) Welch's method
- 2) Fast Fourier Transform
- 3) Polyfit
- 4) Rolling Mean and Standard Deviation

Task 2 Reason for choosing the above features

1) Welch's Method

Welch's method, provides estimates of the spectral density of a signal, making use of a periodogram. The advantage of using this method consists in the reduction of noise in the power spectra although it leads to a worsening of the frequency resolution. As a result, the spectral curve obtained from this method is smooth and gives us a single peak. This is different from FFT due to the fact that FFT gives distinct slopes and has several peaks.

The proposed system extracts exactly one feature that is the largest peak value of the power spectral density for each data record. Reference [1] uses Welch's method to analyze CGM signals and the proposed system is motivated by their results and extracts the maximum peak value of the power spectral density.

2) FFT-

Fast Fourier transform decomposes a time-based function, which is the CGM series in this system, into a set of frequencies. The system uses the FFT function in the SciPy library of python to compute the discrete Fourier transform of a sequence.

The motivation behind using FFT is that it provides a curve with very pronounced slopes and composing of several signal peaks. These peaks and the frequency at which the peaks occur can be useful features to extract. The peaks represent the values of the power spectrum, in this case they represent glucose values. The proposed system selects the top five peak values and the frequency associated with each as the features.

3) Rolling Mean and Rolling Standard Deviation-

Rolling window calculation is an essential technique employed in timeseries data and signal processing. A fixed window size of length n is taken, then the window is slid over the data and certain mathematical operation is performed over the elements in the window. In this case, the system uses window mean and window standard deviation of window length 10 over the CGM timeseries data. All n values are weighted equally. The reason for choosing rolling window operations is that this operation smooths out short-term fluctuations in timeseries data and highlight long-term trends.

4) Polyfit

The idea is to fit a nonlinear function of n degree over the given CGM glucose timeseries data. The result is a vector consisting of coefficients of the n degree polynomial that minimize the squared error in the order from highest degree to lowest degree. The system uses a polynomial of degree 4. Further, the Polyfit function of the NumPy library was used to obtain the coefficients.

Task 3 Feature values and the intuition obtained

The figure below depicts the glucose levels of patient 1's record 10. All the figures below depict the same data record for simplicity of analysis.

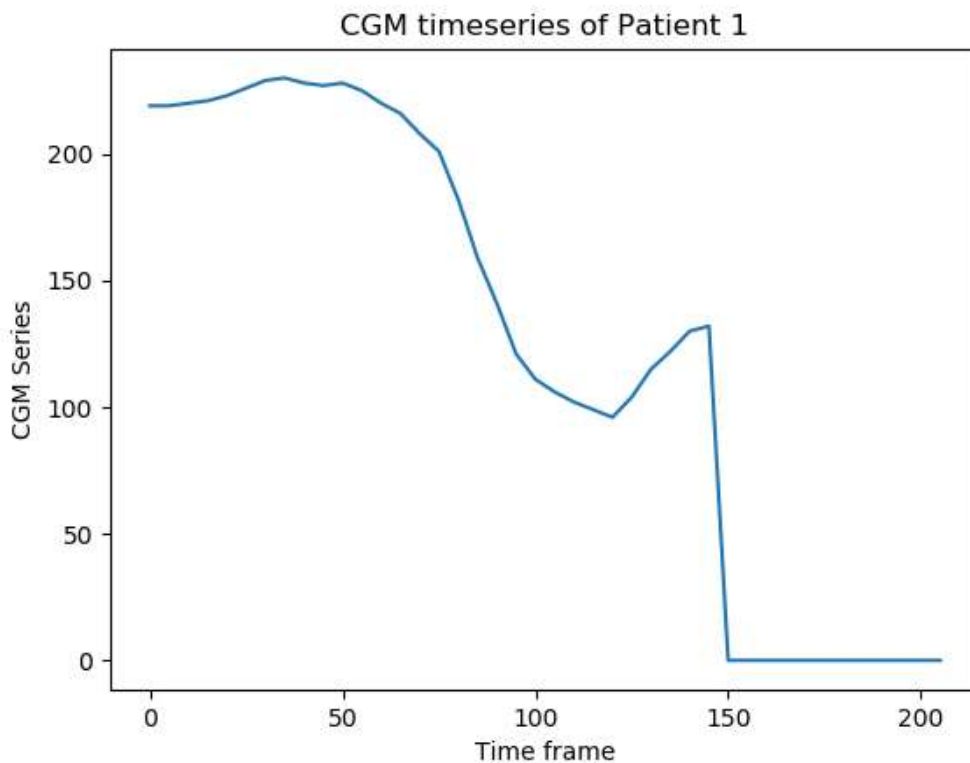


Fig 1 CGM Series data of patient 1 record 10

Based on the intuition and the reasoning from task 2 the following results were observed when they were practically implemented.

1) Welch's method

Expectation: The Welch's method is expected to return a smooth curve with a single peak. This peak can be used as a feature.

Observation: The program returned a single feature set consisting of the value of the maximum peak of each data record in the dataset. The image below depicts the peaks extracted. The Welch's method is applied to the combination of all the records of all the five patients thus totaling 216 data records.

0	84262.134880
1	364047.593730
2	112726.865101
3	56332.778788
4	66996.156523
...	...
211	47184.226339
212	110555.381041
213	170063.891987
214	93008.014531
215	59173.539023

216 rows x 1 columns

Fig 2 Features extracted from Welch's method

The graph obtained for patient 1 clearly validates the expectation obtained from the intuition from task 2, that a smooth curve with a single maximum peak representing the power spectral density is obtained.

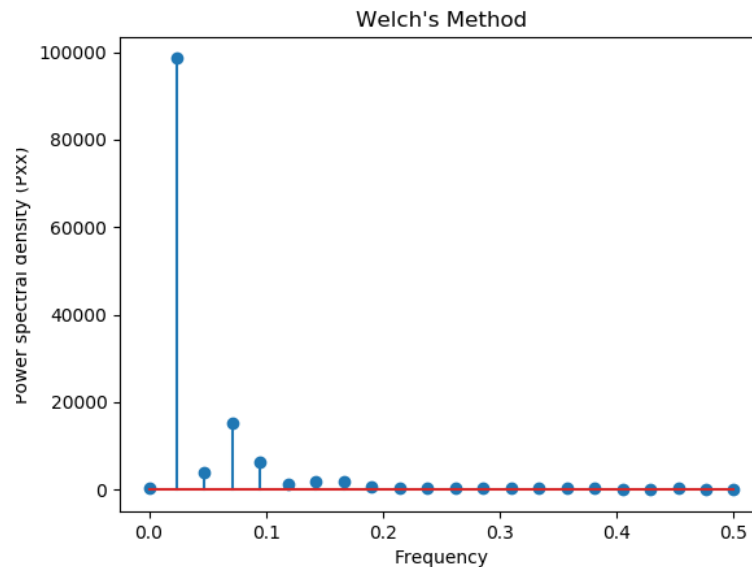


Fig 3 Graph containing from Welch's Method

2) FFT

Expectation- It provides a curve with very pronounced slopes and composing of several signal peaks. These peaks and the frequency at which the peaks occur can be useful features to extract.

Observation- Figure 4 shows the top 5 highest peaks (depicted by each column) out of the several peaks FFT returns. The function is applied to all patients and all records of each patient resulting in a 216x5 dimensional feature matrix.

	0	1	2	3	4
0	414.632630	550.141246	1247.850747	2246.983205	5196.0
1	394.575536	678.028862	1426.896447	3786.424944	9207.0
2	370.658125	417.644159	1310.895588	2539.290590	5856.0
3	381.167977	442.493722	1163.474466	1907.700671	4679.0
4	304.371198	330.747420	696.395327	1704.024370	4305.0
...
211	383.113955	507.684669	1008.513325	1391.121679	4580.0
212	547.439831	547.517338	1039.373045	2326.022600	5899.0
213	530.274194	748.902904	1371.348797	2998.542076	5901.0
214	570.492055	611.460387	1417.359756	2349.492491	6225.0
215	553.306127	806.511197	1351.326672	1859.463059	5339.0

216 rows x 5 columns

Fig 4 FFT top five peaks

Figure 5 shows the frequencies at which the peaks occur. This is the second type of feature obtained from FFT.

	0	1	2	3	4
0	0.0	0.02381	0.047619	0.071429	0.142857
1	0.0	0.02381	0.047619	0.095238	0.119048
2	0.0	0.02381	0.047619	0.095238	0.119048
3	0.0	0.02381	0.047619	0.071429	0.142857
4	0.0	0.02381	0.047619	0.071429	0.119048
...
211	0.0	0.02381	0.047619	0.071429	0.142857
212	0.0	0.02381	0.047619	0.071429	0.119048
213	0.0	0.02381	0.047619	0.071429	0.119048
214	0.0	0.02381	0.047619	0.071429	0.119048
215	0.0	0.02381	0.047619	0.071429	0.119048

216 rows x 5 columns

Fig 5 Frequencies at which the peaks occur

Finally, the graph depicted in Fig 6 validates the expectation by clearly showing the pronounced slopes and multiple peaks of FFT. The system just considers one half of the mirror image like graph to calculate top five peaks.

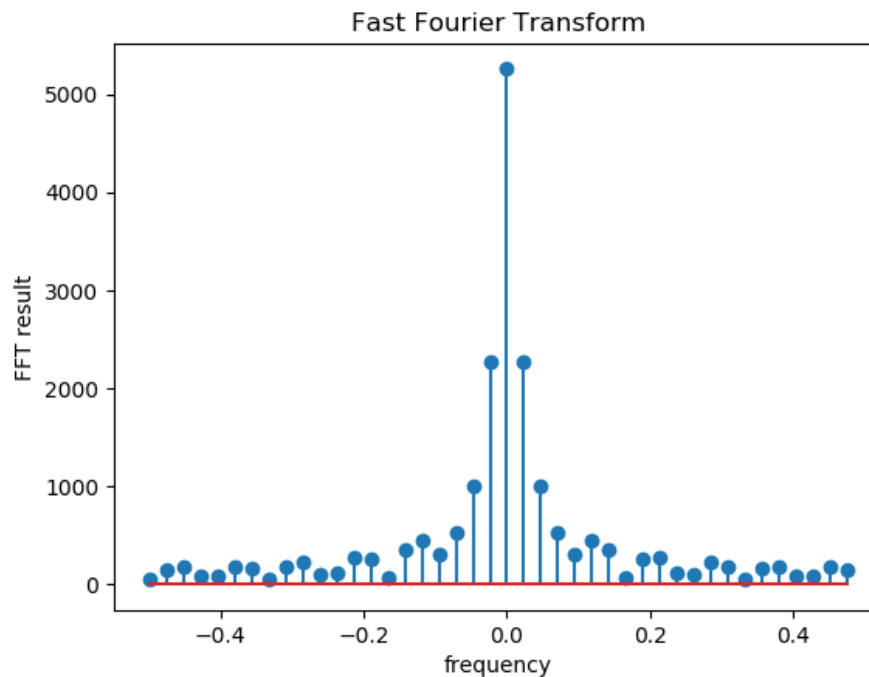


Fig 6 Graphical depiction of FFT

3) Rolling Mean and Rolling Standard Deviation

[illegible]

Fig 7 Features obtained using Rolling Mean

Expectation: This method smooths out the short time fluctuations in timeseries data and highlights long-term trends. Rolling Standard deviation depicts the measure of volatility in the data.

Observation: Fig7 depicts the rolling mean calculated and Fig 8 depicts the rolling mean plotted against the timeframe. It can be seen that Fig 8 is a smoother version of the original CGM dataseries depicted in Fig 1 implying that rolling window mean has successfully produced a smooth curve , eliminating the short term fluctuations while highlighting the long term features.

Fig 9 and 10 depict the rolling window standard deviation which is another important timeseries feature. It depicts the measurement of volatility in the timeseries data.

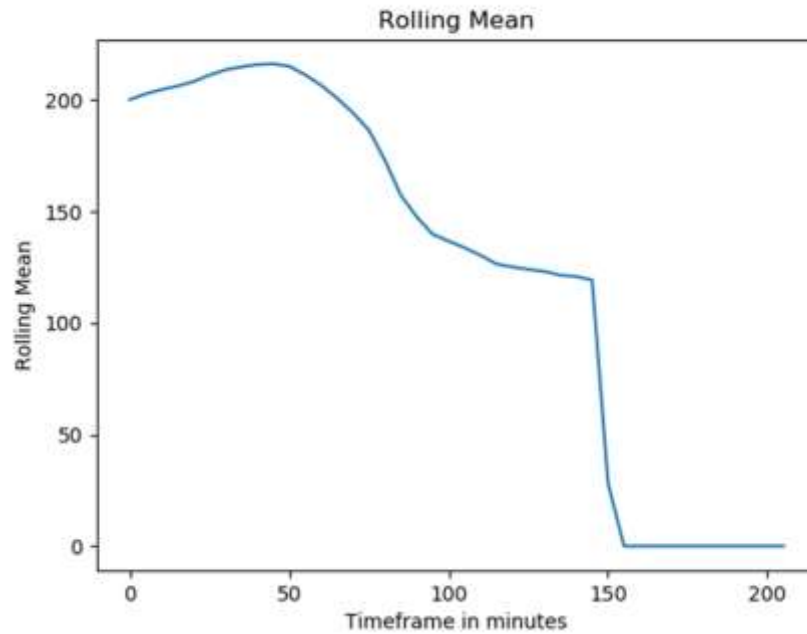


Fig 8 Graph depicting Rolling Mean over the timeframe

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
0	1.222421	11.008600	10.288880	24.061801	64.662322	61.052000	66.687173	70.222871	18.600000	68.067227	60.000000	60.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000	100.000000
1	10.167780	16.000000	17.778780	10.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000
2	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000
3	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000
4	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000	10.000000

Fig 9 Features obtained using Rolling Standard Deviation

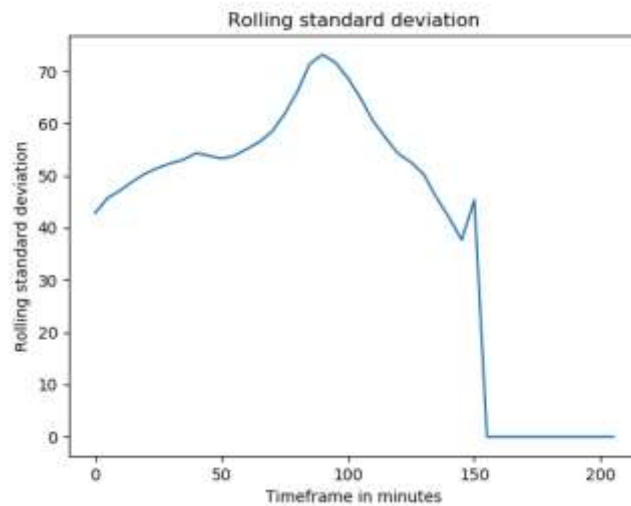


Fig 10 Graph depicting Rolling standard deviation against the timeframe

4) Polyfit

Expectation: Obtain the coefficients of a 4 degree polynomial which accurately depicts and fits the CGM data given.

Observation: Fig 11 depicts the coefficients of the polynomial from degree 4 term to the constant term as:

$\text{Feature}[0]*x^4 + \text{Feature}[1]*x^3 + \text{Feature}[2]*x^2 + \text{Feature}[3]*x + \text{Feature}[4]$

Figure 12 clearly shows how a 4 degree polynomial neatly fits the CGM curve of patient 1. This graph validates our expectation of a smooth curve fitting the timeseries data.

Thus, polyfit gives rise to efficient features.

	0	1	2	3	4
0	-3.234653e-07	0.000176	-0.029657	0.111290	262.140007
1	3.159587e-06	-0.001141	0.101999	-1.470619	284.743457
2	-2.036347e-07	0.000182	-0.044413	1.990058	237.650145
3	-2.242096e-07	0.000122	-0.021030	-0.064735	229.708270
4	1.236418e-06	-0.000446	0.039235	-0.651302	143.105051
...
211	1.723654e-06	-0.000745	0.093767	-3.659600	166.219253
212	1.136776e-06	-0.000388	0.029257	-0.550847	222.847917
213	-4.499352e-07	0.000264	-0.044035	0.167904	329.190646
214	1.106830e-06	-0.000419	0.044292	-2.461856	291.133025
215	1.783313e-06	-0.000745	0.096491	-5.242130	286.148140

216 rows x 5 columns

Fig 11 Features extracted from polyfit

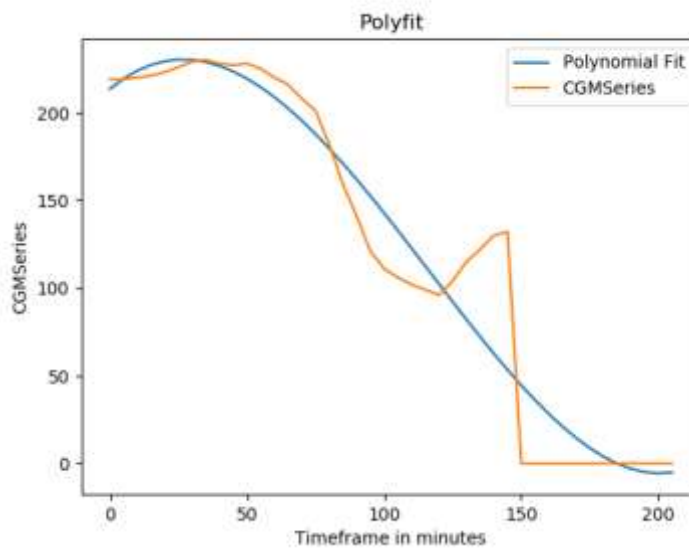


Fig 12 4-degree Polynomial fit into CGM timeseries data

Task 4 Consolidated feature matrix with all types of features

Fig 13 The consolidated feature matrix consisting of feature vectors obtained from the different timeseries features extracted above.

- Each row depicts one patient's one record. Combining all patient records, we get 216 rows.
- As far as the columns are concerned:
 - 42 features from rolling mean
 - 42 features from rolling standard deviation
 - 5 from FFT peaks
 - 5 from FFT frequency
 - 5 from Polyfit
 - 1 from Welch's method

This results in 100 features. Thus, the consolidated feature matrix is 216x100.

Task 5 PCA and new feature matrix

The above feature matrix is fed to the PCA algorithm and only the top five principal components i.e the eigen vectors corresponding to the top 5 highest valued eigen values are selected. Then the features are transformed, i.e projected over these 5 dimensions only.

Fig 14 depicts the new feature matrix obtained from PCA transformation with 5 features represented by columns. They were obtained by projecting the original features to the top 5 dimensions given by PCA.

	0	1	2	3	4
0	-4853.313183	3154.316416	7308.787796	-5566.545272	-15316.848279
1	-20741.667418	13025.813435	25415.217261	-22744.790947	-61383.230736
2	-6793.741419	4454.786976	9309.722911	-7352.903461	-19941.160851
3	-3639.168031	2507.119007	5523.574418	-3831.325283	-10513.992609
4	-4202.898477	2822.059557	5950.480924	-4359.107591	-12024.649087
...
211	-3097.551903	2088.002427	4727.275236	-3226.445477	-9032.296729
212	-6576.131386	4211.172775	8910.386990	-7184.300817	-19754.033905
213	-9827.576444	6220.963482	12768.913763	-10798.741746	-29577.968812
214	-5605.103899	3637.610077	7987.702457	-6174.353660	-17083.540328
215	-3732.000584	2485.542941	5730.591853	-4067.843294	-11339.495133
216 rows x 5 columns					

Fig 14 New feature matrix obtained from PCA

Feature versus timeframe plots along the top five Principal Components

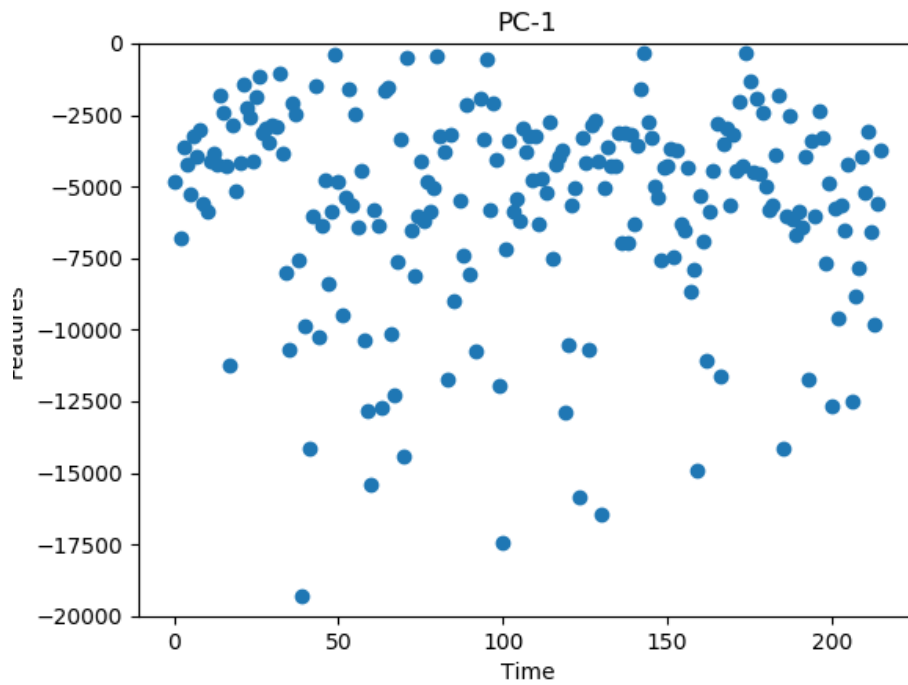


Fig 15 Features vs Timeframe along PC1

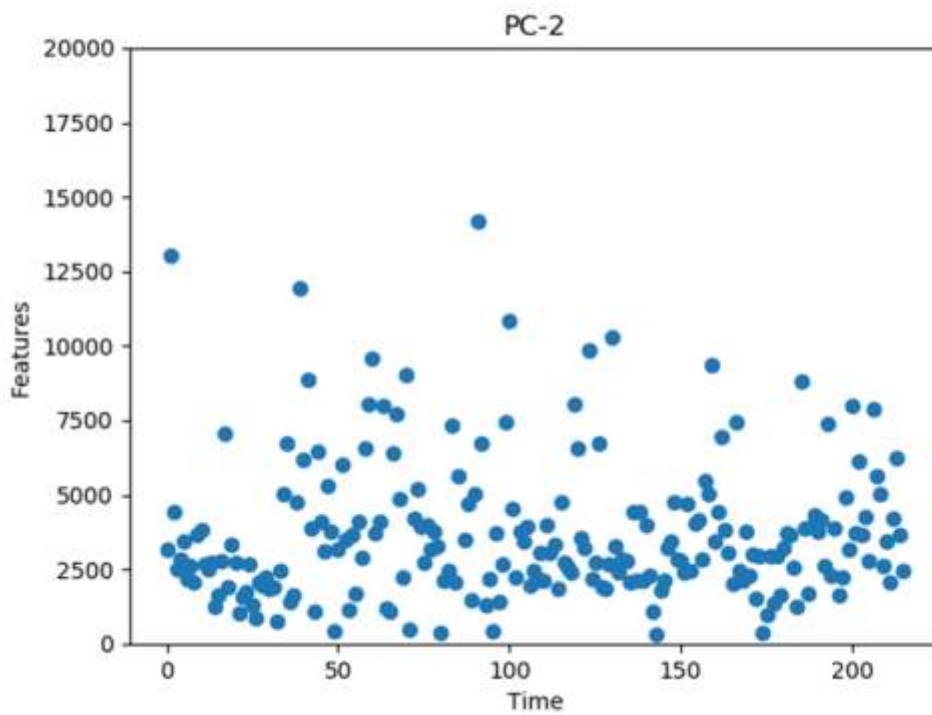


Fig 16 Features vs Timeframe along PC2

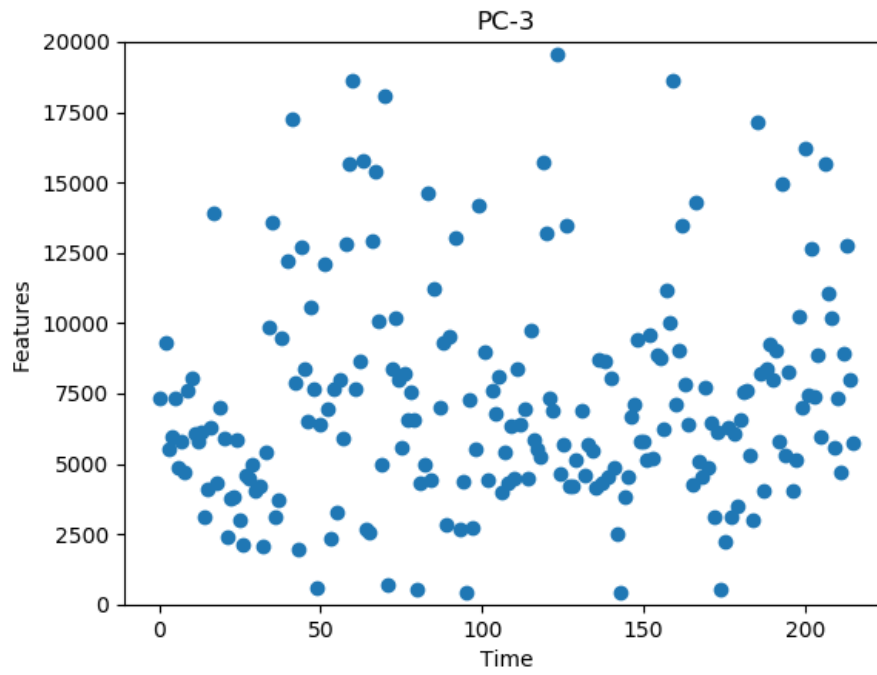


Fig 17 Features vs Timeframe along PC3

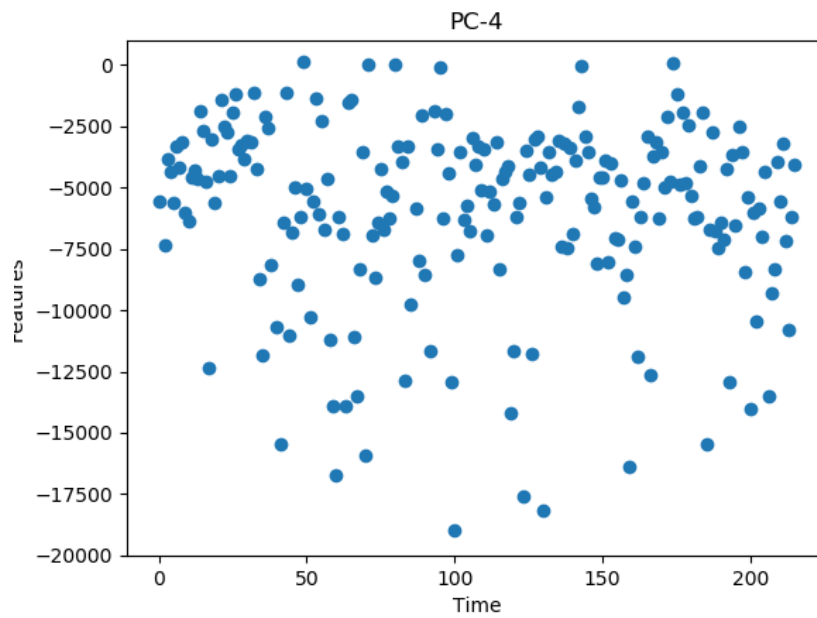


Fig 18 Features vs Timeframe along PC4

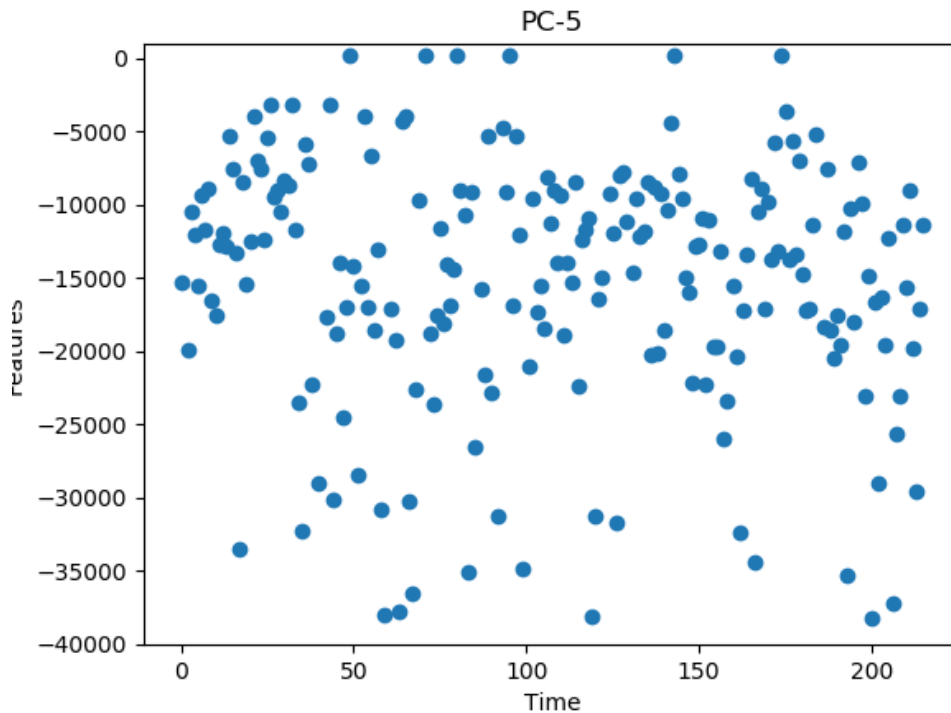


Fig 19 Features vs Timeframe along PC5

Task 6 Why PCA chose these features?

The major assumption involving PCA is that higher variance implies a better feature. Thus, PCA tries to obtain the direction in which when the data is projected, we get the maximum variance. This means that the principal components are obtained in descending order of variance. The task of the system is to select 5 features that represents the feature set as completely as possible.

Figure 20 displays that the principal component 1 displays the maximum variance among all other PC's with a variance of 27.26, principal component 2 displays a lower variance of 21.33, principal component 3 displays a lower variance of 17.39, principal component 4 displays a lower variance of 9.85, principal component 5 displays an even lower variance of 6.86. This trend is true for the entire set of principal components. This clearly tells us the fact that PC1,2,3,4 and 5 were selected because they have a higher variance than other principal components.

Figure 21 depicts the PCA ratio which is the amount of information retained after feature is reduced to 5 features. We can see that the top 5 PC's single handedly can represent 82.32% of the information of the original 100 featured data. And since higher dimensionality is a curse,

selecting these 5 PC's representing 82% of the information is wiser than selecting a 100 dimensioned dataset which just represents 18% extra information.

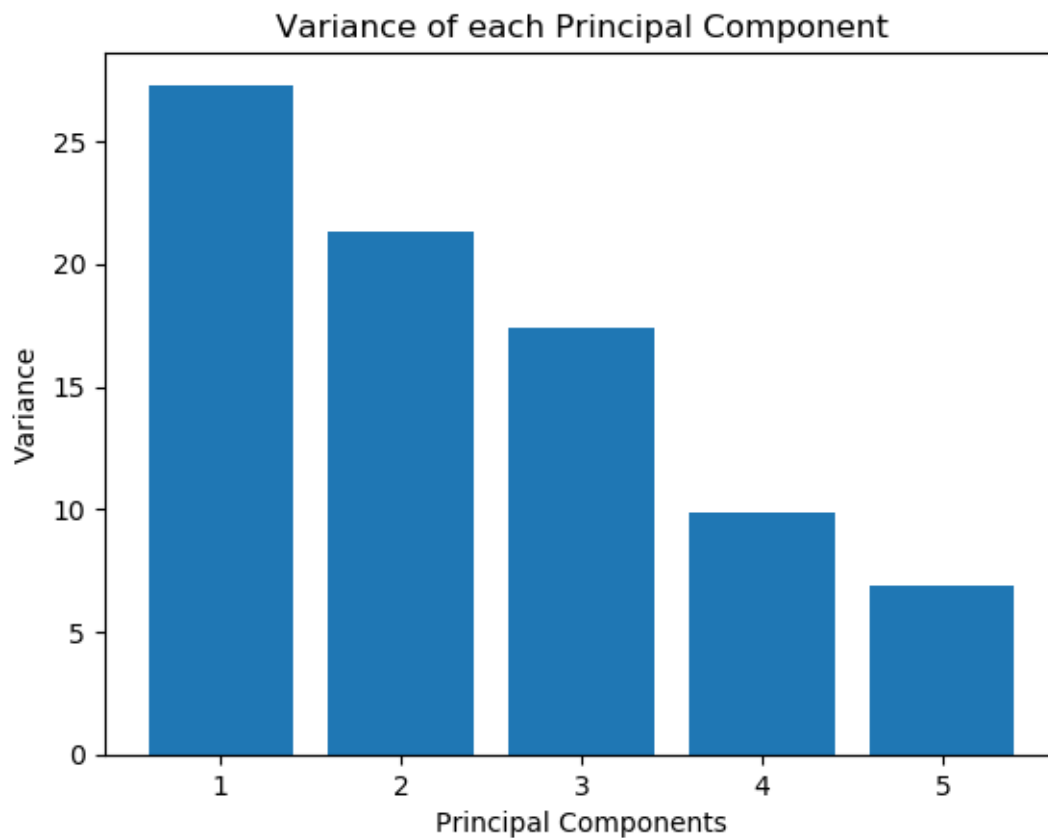
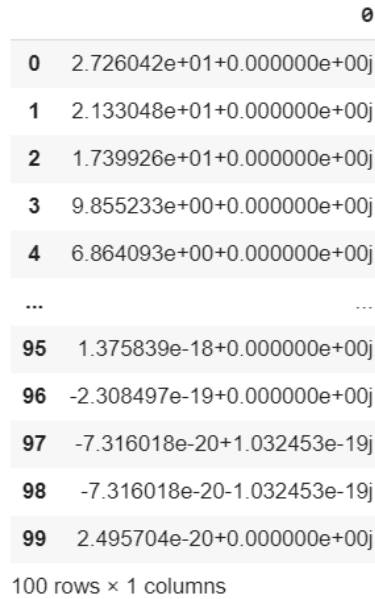


Fig 20 Variance displayed by each principal component

PCA ratio: 0.8232657135968147

Fig 21 PCA ratio of the five components

The principal components are chosen based on eigen values of the covariance matrix of the feature set. Figure 21 displays the eigen values of the feature matrix. The first five maximum eigen values correspond to the five principal components chosen as depicted in Figure 22



	0
0	2.726042e+01+0.000000e+00j
1	2.133048e+01+0.000000e+00j
2	1.739926e+01+0.000000e+00j
3	9.855233e+00+0.000000e+00j
4	6.864093e+00+0.000000e+00j
...	...
95	1.375839e-18+0.000000e+00j
96	-2.308497e-19+0.000000e+00j
97	-7.316018e-20+1.032453e-19j
98	-7.316018e-20-1.032453e-19j
99	2.495704e-20+0.000000e+00j

100 rows × 1 columns

Fig 21 Eigen values of feature matrix



	0
0	27.260421+0.000000j
1	21.330479+0.000000j
2	17.399259+0.000000j
3	9.855233+0.000000j
4	6.864093+0.000000j

Fig 22 Eigen Values corresponding to the top 5 PC's from PC1 to PC5

The direction along of which the variance is maximum is represented by the eigen vector corresponding the largest eigen value. The eigen vectors are depicted in Figure 23. The feature matrix is projected along such eigen vectors to get the transformed feature matrix depicted by figure 14.

Thus, due to the above concepts involved in PCA the above principal components and the transformed feature matrix were obtained.

	0	1	2	3	4	5	6	7	8	9	10	
0	-0.354677+0.000000i	0.323355+0.000000i	0.061367+0.000000i	0.060107+0.000000i	-0.160991+0.000000i	0.013506+0.000000i	-0.223334+0.000000i	0.110927+0.000000i	0.246545+0.000000i	0.190123+0.000000i	-0.043635+0.000000i	-0.513392+0.000000i
1	-0.011589+0.000000i	0.825795+0.000000i	0.069057+0.000000i	0.065717+0.000000i	-0.232986+0.000000i	0.064236+0.000000i	-0.125872+0.000000i	-0.065563+0.000000i	0.111032+0.000000i	0.060048+0.000000i	0.174345+0.000000i	0.141353+0.000000i
2	-0.032123+0.000000i	0.224982+0.000000i	0.090514+0.000000i	0.079764+0.000000i	-0.239359+0.000000i	0.005331+0.000000i	-0.114959+0.000000i	-0.125189+0.000000i	0.110793+0.000000i	0.090052+0.000000i	0.023645+0.000000i	0.125952+0.000000i
3	-0.012742+0.000000i	0.313267+0.000000i	0.100944+0.000000i	0.071922+0.000000i	-0.169529+0.000000i	-0.034890+0.000000i	-0.136834+0.000000i	-0.181019+0.000000i	0.187911+0.000000i	0.057965+0.000000i	0.066560+0.000000i	0.201990+0.000000i
4	-0.020982+0.000000i	0.019148+0.000000i	0.125482+0.000000i	0.032706+0.000000i	-0.286297+0.000000i	-0.094802+0.000000i	-0.106275+0.000000i	0.043181+0.000000i	0.271085+0.000000i	0.244152+0.000000i	0.021371+0.000000i	-0.323315+0.000000i
...
95	-0.074439+0.000000i	0.009629+0.000000i	-0.014655+0.000000i	0.123475+0.000000i	-0.179540+0.000000i	0.173391+0.000000i	-0.298782+0.000000i	0.046834+0.000000i	-0.160963+0.000000i	-0.071013+0.000000i	-0.127394+0.000000i	-0.182394+0.000000i
96	-0.075724+0.000000i	-0.009337+0.000000i	0.016298+0.000000i	-0.124323+0.000000i	0.174594+0.000000i	-0.168509+0.000000i	0.291944+0.000000i	-0.079837+0.000000i	0.154473+0.000000i	0.195296+0.000000i	0.129960+0.000000i	0.115130+0.000000i
97	-0.076089+0.000000i	0.008887+0.000000i	-0.022386+0.000000i	0.121436+0.000000i	-0.183391+0.000000i	0.268417+0.000000i	-0.299245+0.000000i	0.041529+0.000000i	-0.210297+0.000000i	-0.153206+0.000000i	-0.101431+0.000000i	0.082798+0.000000i
98	-0.055395+0.000000i	-0.007942+0.000000i	0.007363+0.000000i	-0.099038+0.000000i	0.153822+0.000000i	-0.218752+0.000000i	0.306231+0.000000i	0.043884+0.000000i	0.221342+0.000000i	0.202613+0.000000i	0.066965+0.000000i	-0.268913+0.000000i
99	-0.012844+0.000000i	0.005261+0.000000i	0.142003+0.000000i	0.083058+0.000000i	-0.157941+0.000000i	-0.030011+0.000000i	-0.160823+0.000000i	-0.134289+0.000000i	0.150157+0.000000i	0.113365+0.000000i	0.214606+0.000000i	0.508320+0.000000i

100 rows x 13 columns

Fig 23 Eigen Vectors

Reference

[1] Fico, G., Hernández, L., Cancela, J., Isabel, M.M., Facchinetti, A., Fabris, C., Gabriel, R., Cobelli, C. and Arredondo Waldmeyer, M.T., 2017. Exploring the frequency domain of continuous glucose monitoring signals to improve characterization of glucose variability and of diabetic profiles. *Journal of diabetes science and technology*, 11(4), pp.773-779.