Pearl Ayem -34404140

EOSC 510/410 Assignment 2

Problem 1:

a. Timeseries of each variable

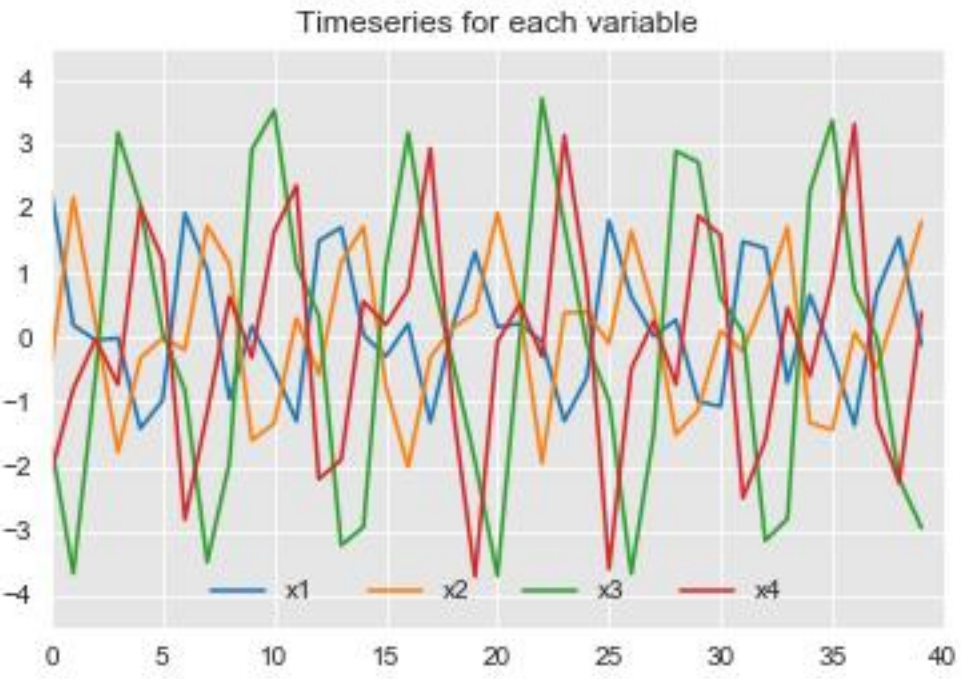

Figure 1: Timeseris of all variables in dataset 1 (PCA.csv)

b. PCA results. Lines 10 and 11 check how much of the variance is explained by the first 'n' modes. In this case n is 2 modes. The output shows that the first 2 modes explain 97.61% of the variance. The lines 15 to 22 check the expected output of PCA and compare it with the actual output to verify is PCA has been carried out correctly. In this case it did.

```python
#normalize data and check it out
data_norm = (df - df.mean())/df.std()
## We want to run PCA
n_modes = np.min(np.shape(data_norm))
pca = PCA(n_components = n_modes)
PCs = pca.fit_transform(data_norm)
eigvecs = pca.components_
fracVar = pca.explained_variance_ratio_
n=2
print(np.sum(fracVar[:n])*100)  #sum of the first n modes = total percent variance explained by the first neigvecs
print(np.shape(eigvecs))
#investigate: did PCA work as we expected?  What size of variables do we expect?
nObservations = np.shape(data_norm)[0]
nVariables = np.shape(data_norm)[1]
print('Expected sizes:')
print('\t' + str(nVariables) + ' eigenvectors, each of length ' + str(nVariables))
print('\t' + str(nVariables) + ' eigenvalues, one for each eigenvector')
print('\t' + str(nVariables) + ' PCs, each of length ' + str(nObservations))
print('Actual sizes:')
print('\t' + str(np.shape(eigvecs)[0]) + ' eigenvectors, each of length ' + str(np.shape(eigvecs)[1]))
print('\t' + str(len(fracVar)) + ' eigenvalues')
print('\t' + str(np.shape(PCs)[1]) + ' PCs, each of length ' + str(np.shape(PCs)[0]))

```

```
97.61264720571651
(4, 4)
Expected sizes:
        4 eigenvectors, each of length 4
        4 eigenvalues, one for each eigenvector
        4 PCs, each of length 40
Actual sizes:
        4 eigenvectors, each of length 4
        4 eigenvalues
        4 PCs, each of length 40
```

Figure 2: Screenshot of the code that performed normalization and PCA on dataset 1.

c. I want to keep the first 2 modes because they explain 97.61% of all the variance as seen in Figure 3. Even though there is a drop off after the first mode, the reason I kept the first two was because just the first one explained only 59.79% of the variance. Figure 3 compares all the modes, with just the chosen ones.

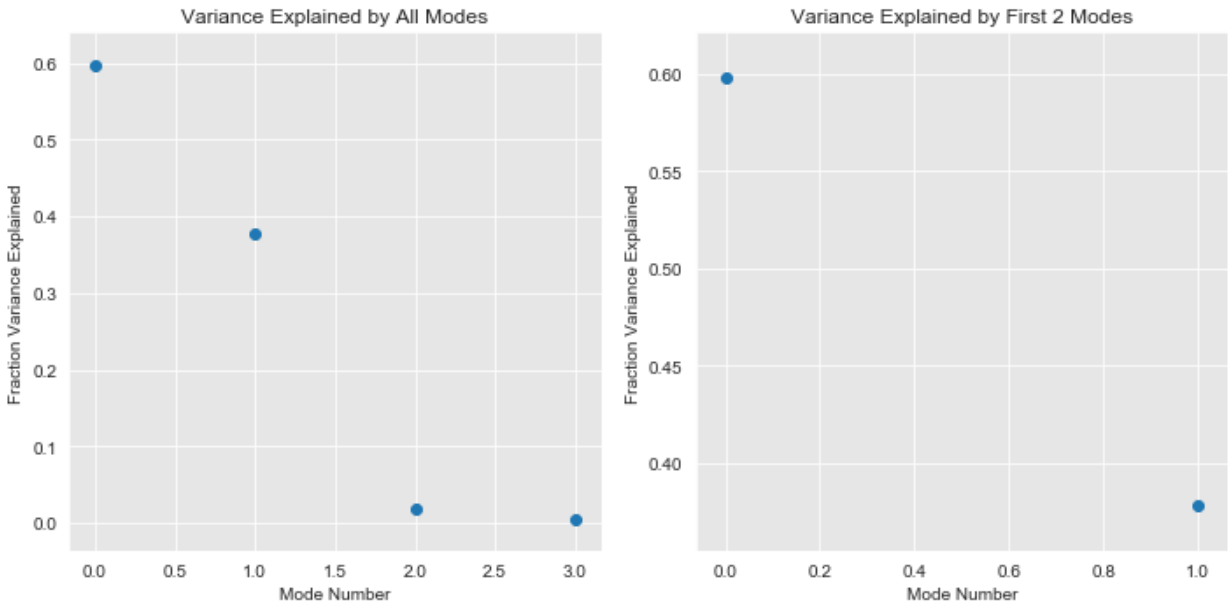

Figure 3: Fraction of variance explained by the first 2 modes after PCA

d. Eigenvectors show the spatial patterns and PCs show the temporal patterns. So the first eigenvector plot for mode 1 which accounts for the highest variance, shows the change of variance over the spatial x domain explained by that mode. The first PC plot for MOde 1 shows the temporal variance explained by that mode. Similarly, is true for Mode 2, but for the second most significant mode. Overall Eigenvector 1 and 2 are orthogonal, and the PCs are uncorrelated. In this example the PCs show an oscillation through time.

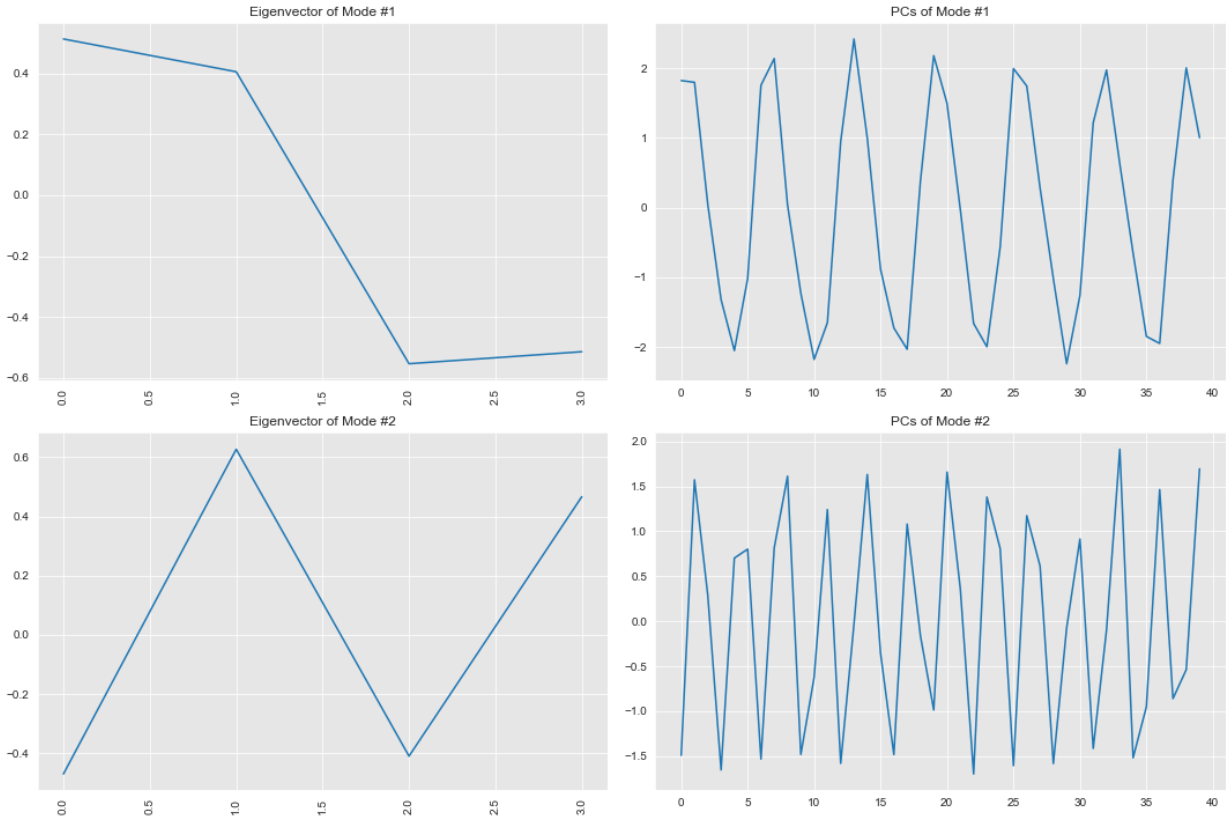

Figure 4: Eigenvectors and PCs of every PCA mode

e. PC1 vs PC2

There are no defined clusters in this plot, and lowest spread or variance is in the centre. There is however a pattern of datapoints aligned towards the left and right corners (kind of stacked vertically). These edges have most variance too as seen in Figure 5. The points prove that the PCs are uncorrelated, however they are aligned in a bimodal cluster as seen in the distribution above the main plot. Similarly, there are pointy ends in almost all the 4 corners if looked closely.

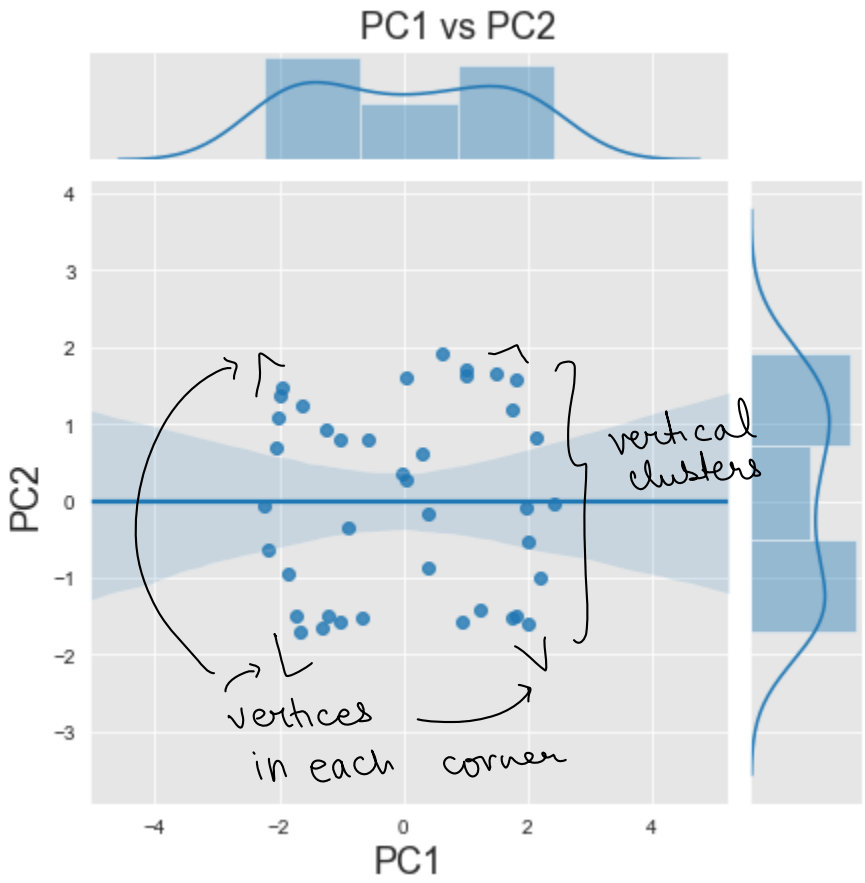

Figure 5: Correlation between PC1 and PC2

Problem 2

a. Timeseries plots

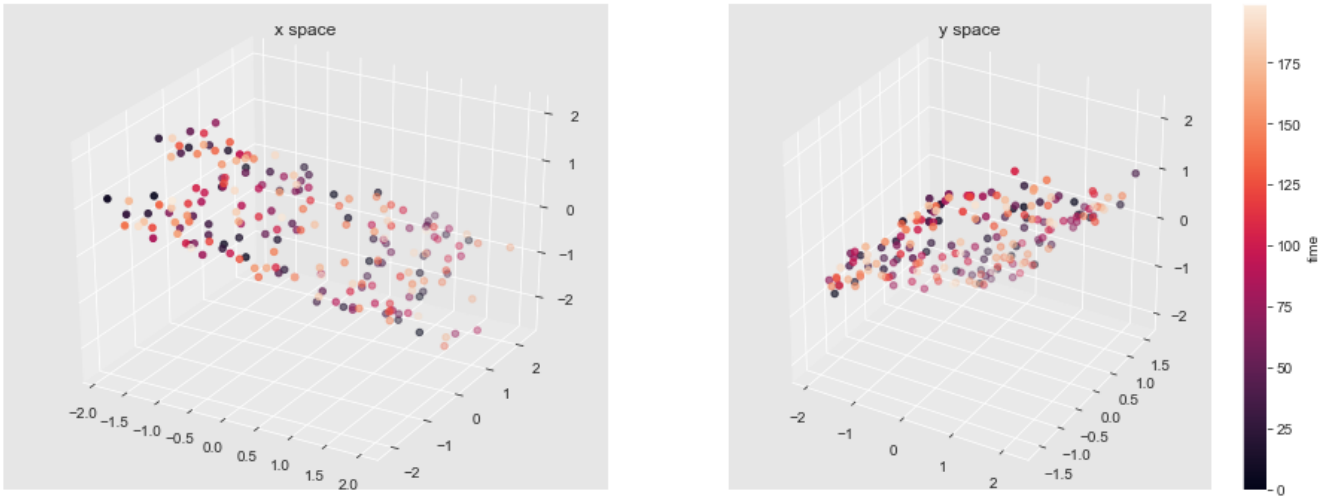

Figure 6: Timeseries plot of all the variables in Dataset 2 (CCA.csv)

b. Figure 7 shows the code for the CCA and Figure 8 is the result of the CCA.

```
n_modes = 3 #modes to keep
cca = CCA(n_components=n_modes,max_iter = 10000)
U,V = cca.fit_transform(xdata,ydata)
A = cca.x_weights_
B = cca.y_weights_
F = np.cov(xdata.T)@A
G = np.cov(ydata.T)@B
r = [np.corrcoef(U[:,ii],V[:,ii]) for ii in range(n_modes)]
```

Figure 7: Screenshot of the code written to perform CCA on Dataset 2

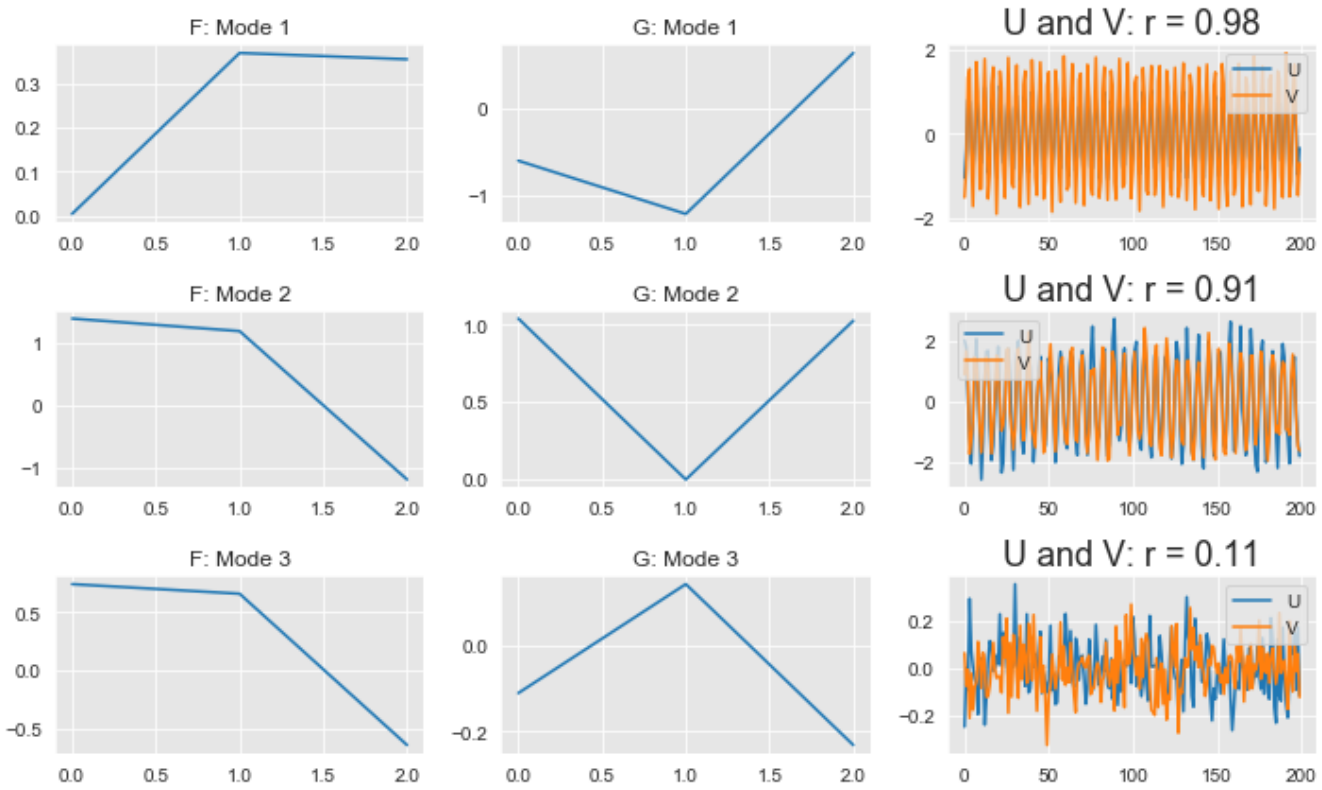

Figure 8: Results of CCA performed on dataset 2

c. F1 is black, F2 is blue. G1 is black, G2 is blue.

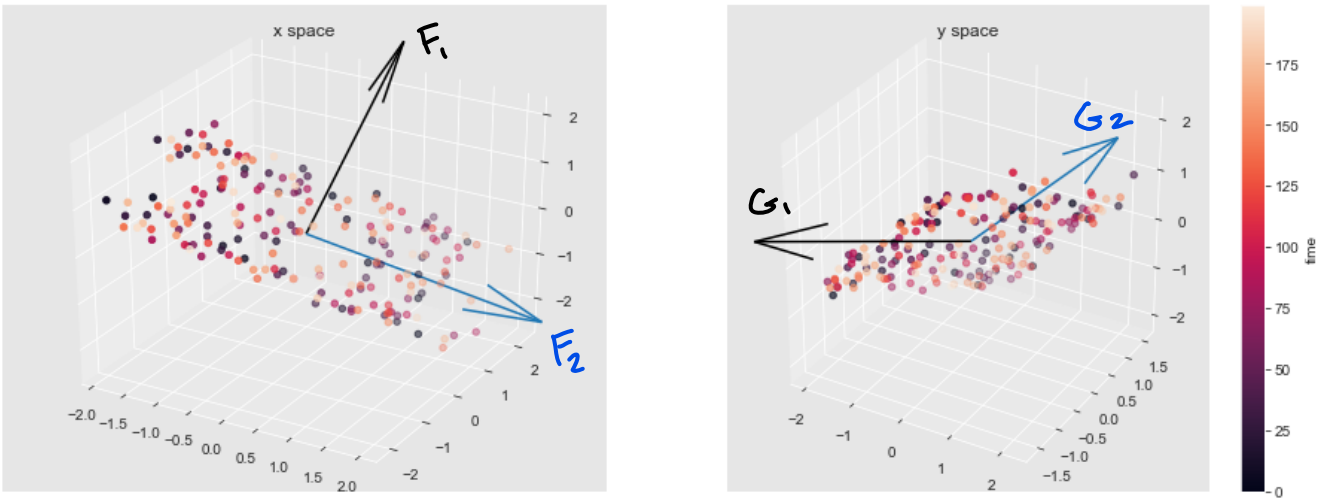

Figure 9: CCA vectors F and G in x-space and y-space. x space has F vectors, and y space has G vectors

d. Mode 1 has the highest r value of 0.98 and hence the highest correlation. This is seen in the plot below (grey diamonds), where the data points for Mode 1 show the smallest deviation and follow a narrow positively increasing trend. The second significant mode is Mode 2 (r=0.91). It has some more spread than Mode 1 (the + markers) but still shows a positive correlation between the U and V markers. Mode 3 is just a cluster in one spot, and shows lowest correlation as predicted by the r value of 0.11. It does not follow any linear trends.

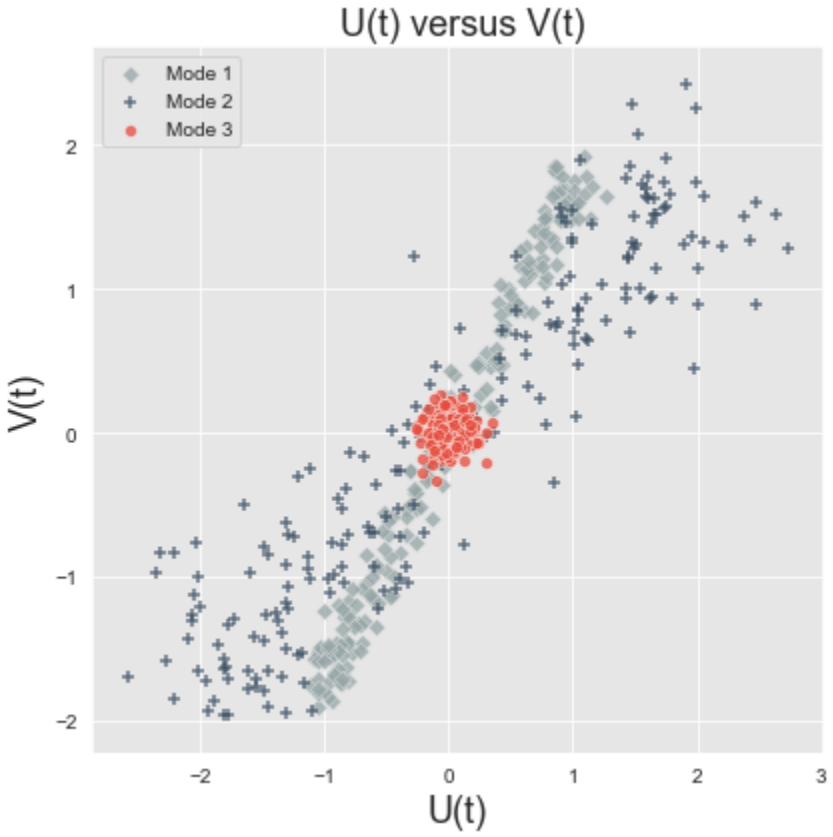

Figure 10: Correlation between U and V

e. The plots in the x-space have the vectors lining up perfectly, which indicates that they explain a lot of the variance in the x-space and are highly correlated. The vectors in the y-space are less correlated. Even though some remain in the direction of the variance, there is less correlation between the modes. This would suggest that overall there is larger variance in the x space.

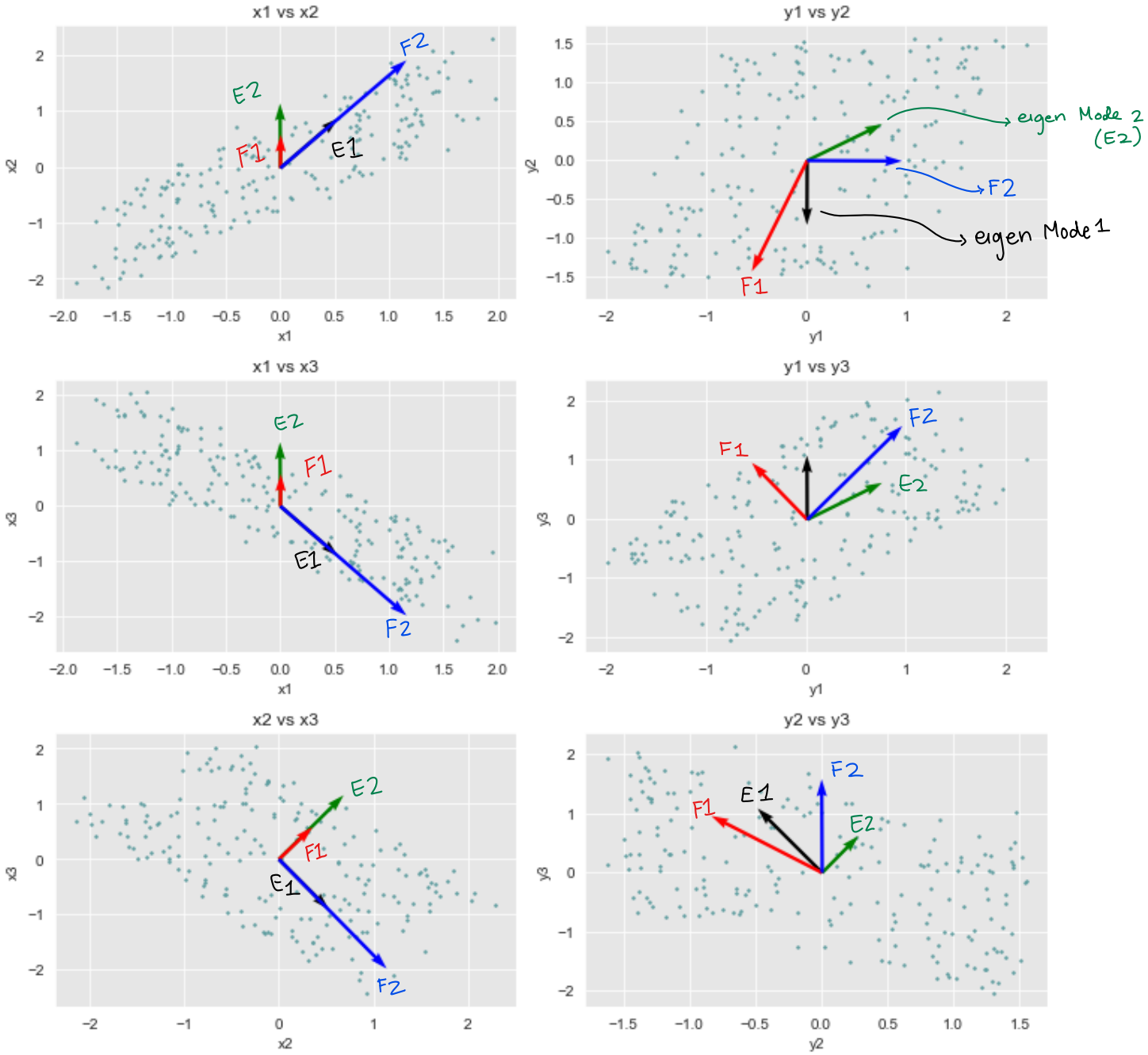

Figure 11: CCA vectors F and G, compared to the eigenvectors from PCA.