MSEN660 HW3

Akshay Rao

Nov 2018

1 Data Preprocessing

- First we read the spreadsheet into python
- Next we discard all features that don't have at least 5% non zero values.
- Then we discard rows that don't have a coercivity value.
- Now we add zero mean Gaussian Noise to all the features.
- Finally we normalized all the features to have zero mean and unit normality.

2 Run PCA

We run PCA on the 12 features.

3 Explained Variance

- Here we obtain 'scree' plots.
- We plot the explained variance accounted for by each PC as a function of PC number.

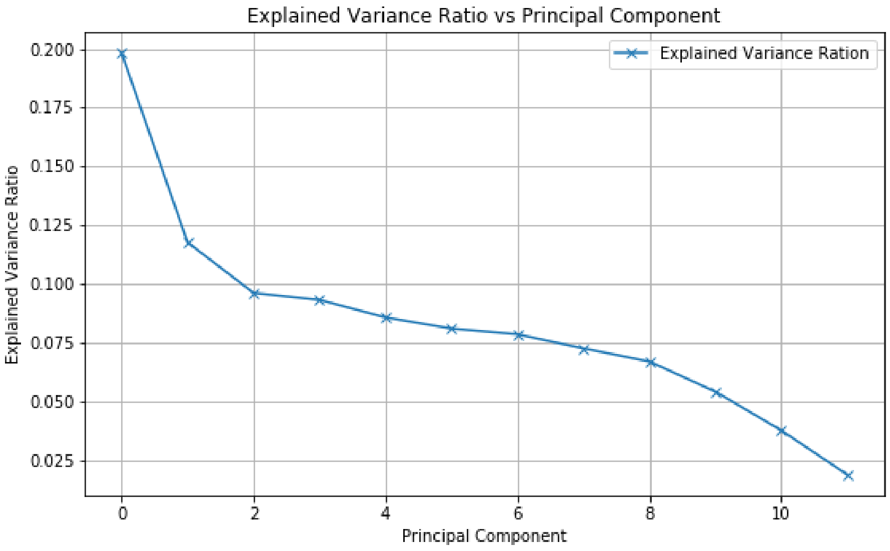


Figure 1: Explained Variance vs PC number σ=1

- We plot the cumulative explained variance accounted for as a function of PC number. From the plot it appears that 10 PCs are needed to explain 95% of the variance.

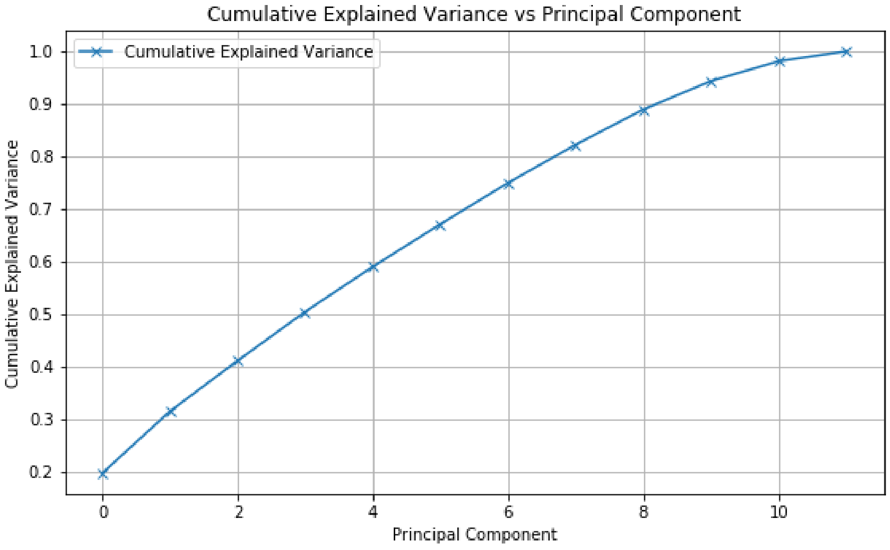


Figure 2: Cumulative Explained Variance vs PC number

4 PC Scatter Plots

- We plot PC1, PC2, PC3 against each other. The coercivity values are categorized into 'Low'(Red), 'Medium' (Green) and 'High' (Blue).
- From the plots it appears that PC1 does the best job of discriminating amongst the coercivity classes.
- In both the PC1 vs PC2 as well as the PC1 vs PC3 we can see that the data is being separated along the PC1 axis. The red 'low ' is forming a cluster to the left while the blue 'high' is forming a cluter to the right. The green 'medium' on the other hand is a little more difficult to differentiate from the other two classes.

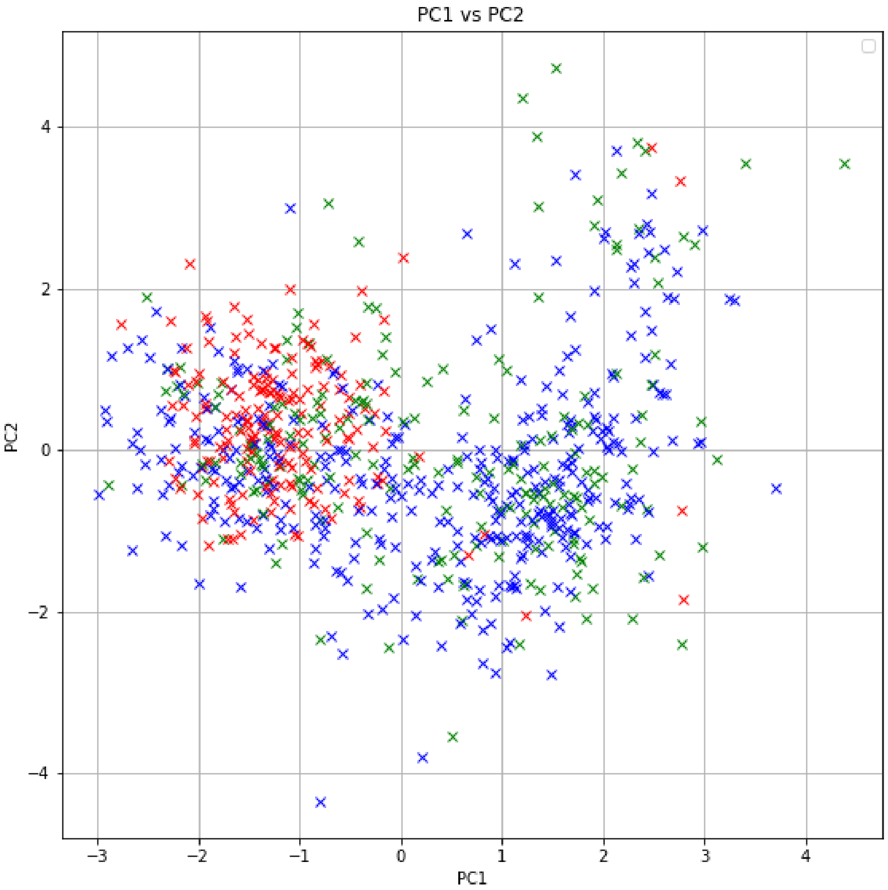


Figure 3: PC1 vs PC2

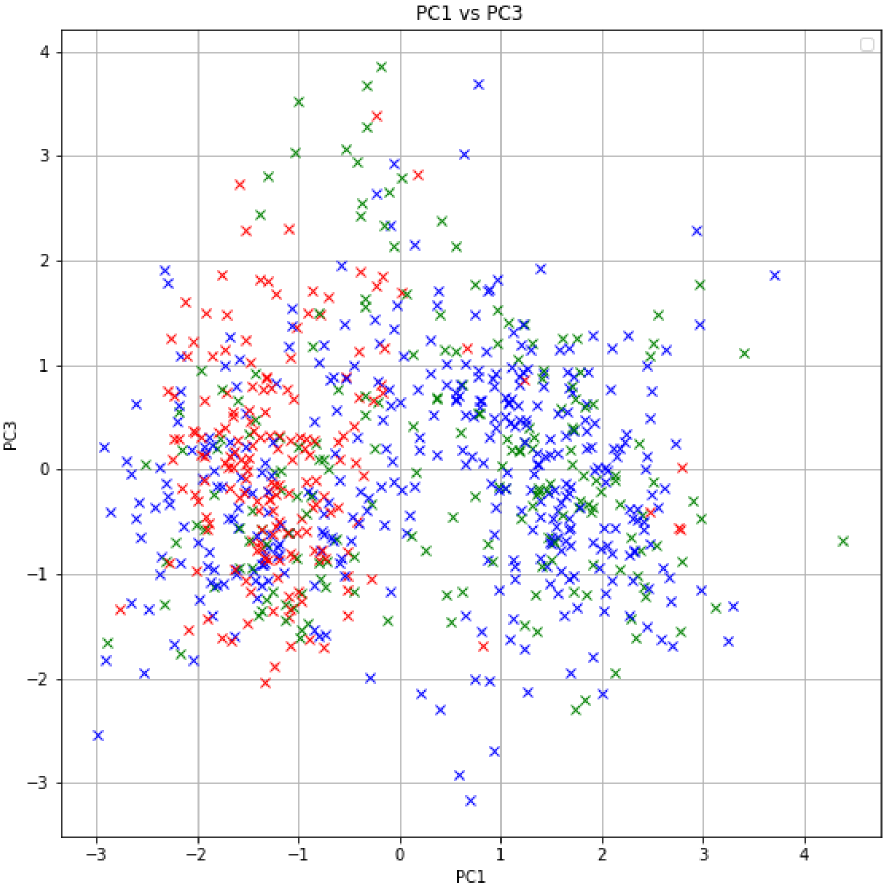


Figure 4: PC1 vs PC3

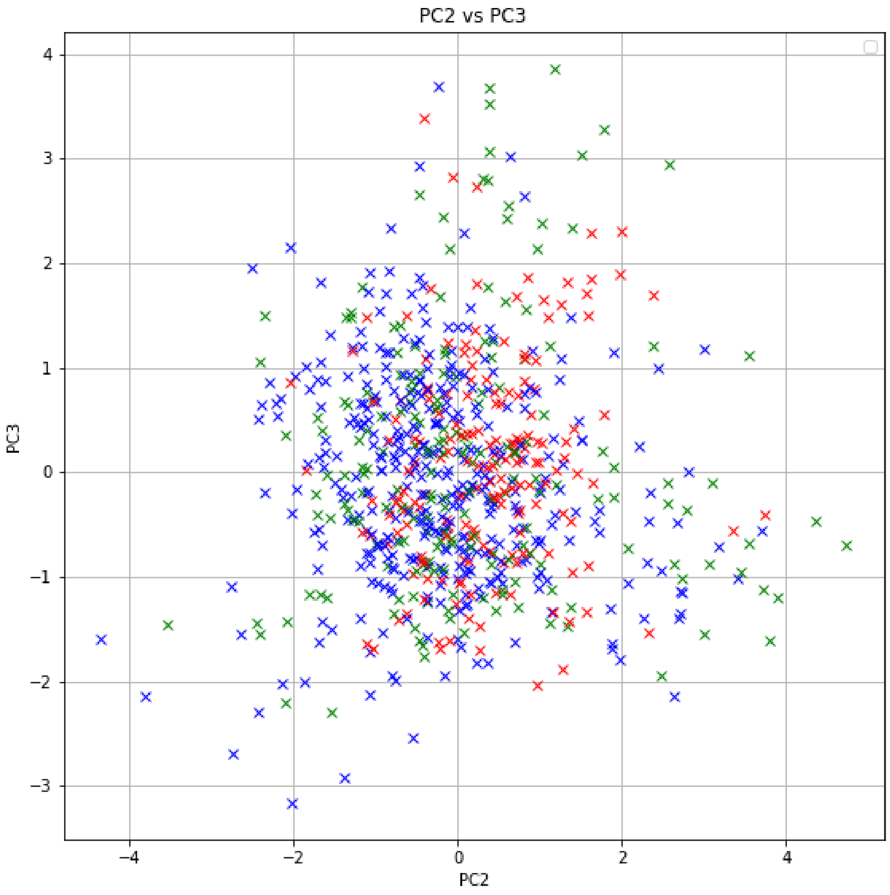


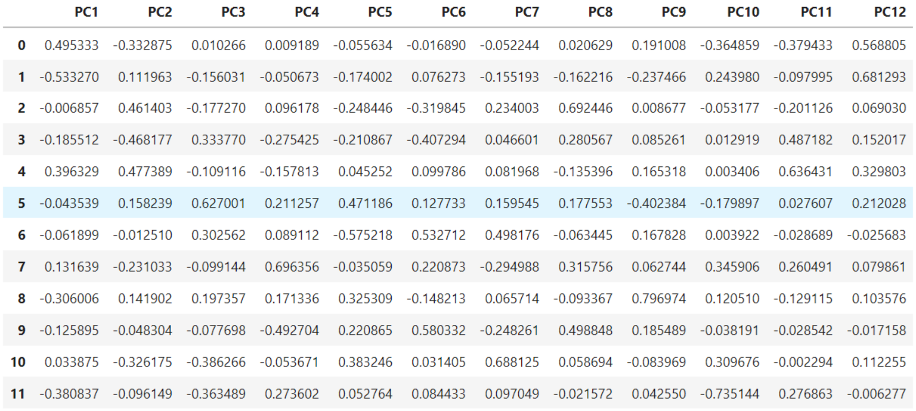
Figure 5: PC2 vs PC3

5 Loading Matrix W

- We look at the matrix W of eigen vectors. Here based on the discriminating PC- PC1 from earlier we want to select the two most important features. We can determine this by observing the coefficients of the PCs for each feature. Features are each row.

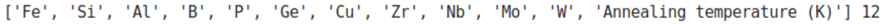
- From this W matrix it appears that 'Fe' and 'Si' are the two most important features. They have the largest magnitude coefficients for PC1.

- A large positive coefficient indicates a strong increasing relationship between the PC and the feature, while a large negative coefficient indicates a strong decreasing/reverse relationship.



	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
0	0.495333	-0.332875	0.010266	0.009189	-0.055634	-0.016890	-0.052244	0.020629	0.191008	-0.364859	-0.379433	0.568805
1	-0.533270	0.111963	-0.156031	-0.050673	-0.174002	0.076273	-0.155193	-0.162216	-0.237466	0.243980	-0.097995	0.681293
2	-0.006857	0.461403	-0.177270	0.096178	-0.248446	-0.319845	0.234003	0.692446	0.008677	-0.053177	-0.201126	0.069030
3	-0.185512	-0.468177	0.333770	-0.275425	-0.210867	-0.407294	0.046601	0.280567	0.085261	0.012919	0.487182	0.152017
4	0.396329	0.477389	-0.109116	-0.157813	0.045252	0.099786	0.081968	-0.135396	0.165318	0.003406	0.636431	0.329803
5	-0.043539	0.158239	0.627001	0.211257	0.471186	0.127733	0.159545	0.177553	-0.402384	-0.179897	0.027607	0.212028
6	-0.061899	-0.012510	0.302562	0.089112	-0.575218	0.532712	0.498176	-0.063445	0.167828	0.003922	-0.028689	-0.025683
7	0.131639	-0.231033	-0.099144	0.696356	-0.035059	0.220873	-0.294988	0.315756	0.062744	0.345906	0.260491	0.079861
8	-0.306006	0.141902	0.197357	0.171336	0.325309	-0.148213	0.065714	-0.093367	0.796974	0.120510	-0.129115	0.103576
9	-0.125895	-0.048304	-0.077698	-0.492704	0.220865	0.580332	-0.248261	0.498848	0.185489	-0.038191	-0.028542	-0.017158
10	0.033875	-0.326175	-0.386266	-0.053671	0.383246	0.031405	0.688125	0.058694	-0.083969	0.309676	-0.002294	0.112255
11	-0.380837	-0.096149	-0.363489	0.273602	0.052764	0.084433	0.097049	-0.021572	0.042550	-0.735144	0.276863	-0.006277

Figure 6: Loading Matrix 'W'



```
['Fe', 'Si', 'Al', 'B', 'P', 'Ge', 'Cu', 'Zr', 'Nb', 'Mo', 'W', 'Annealing temperature (K)'] 12
```

Figure 7: Features

6 Hierarchical Clustering

- Now we cluster the deterministically chosen samples (0,12,..132) according to the two features we decided on earlier.
- We perform clustering based three different types of linkage- 'Single', 'Complete' and 'Average'.
- The dendrograms for each of these are included below. The leaves are labelled 'Low', 'Medium' or 'High' based on coercivity value for that sample.
- We can see that two major clusters form for each of the linkage methods.

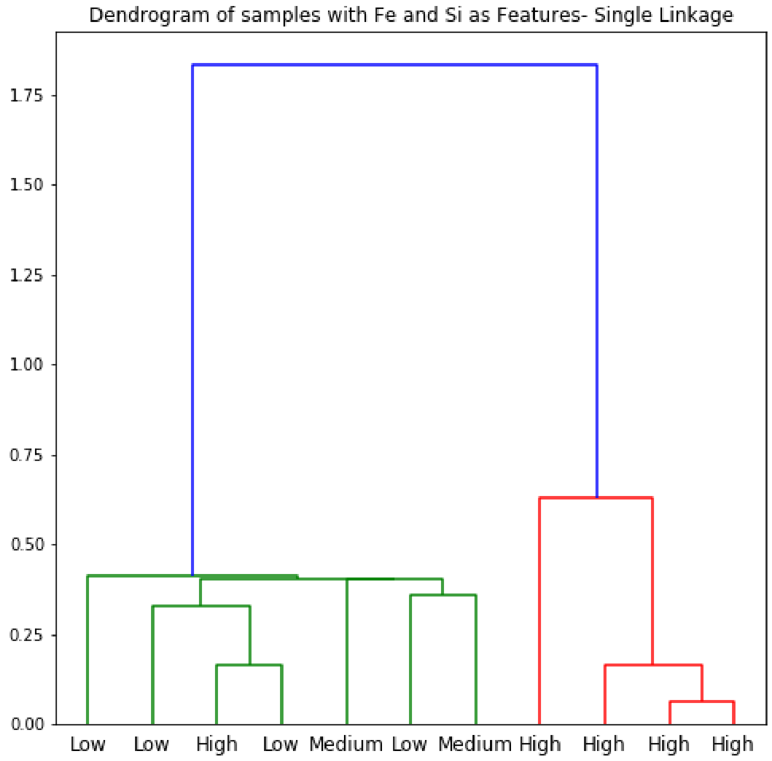


Figure 8: Hierarchical Clustering with Single Linkage'

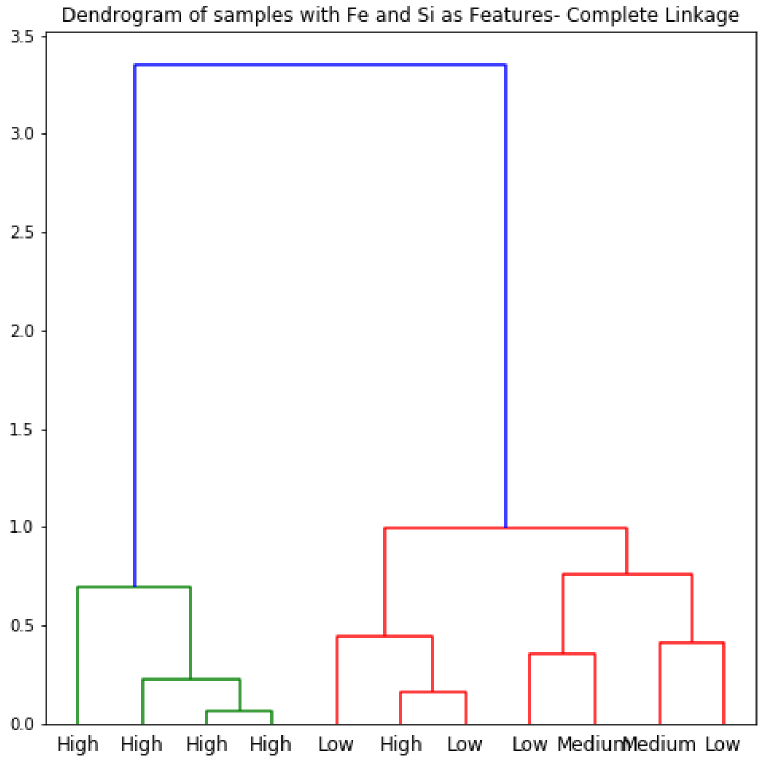


Figure 9: Hierarchical Clustering with Complete Linkage

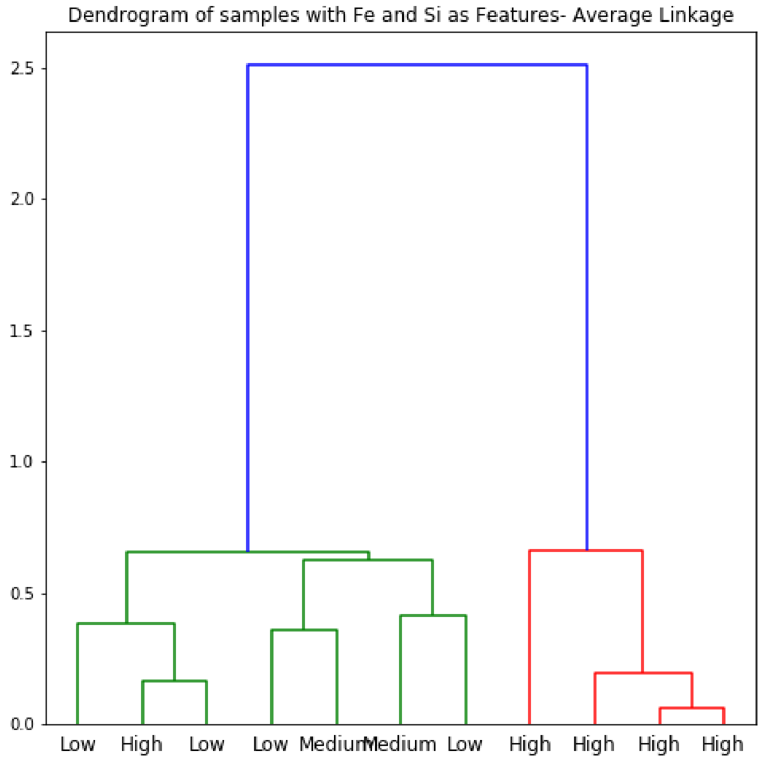


Figure 10: Hierarchical Clustering with Average Linkage

- The hierarchical clustering algorithm begins with all n samples as their own clusters. Then it iteratively groups together the closest pair of clusters based upon a chosen dissimilarity measure, usually euclidean. This dissimilarity between clusters in this case is by single, average and complete linkage. For single we have minimal intercluster dissimilarity, average we consider mean intercluster dissimilarity and for complete we look at the maximal intercluster dissimilarity.
- We can see the formation of two primary clusters. With the exception of one outlier, all the 'high' coercivity samples are close together in the dendrogram. The 'low' and 'medium' coercivity samples form their own cluster.
- This implies that the 'high' samples are close together in euclidian distance considering the two features 'Fe' and 'Si'. In addition for all linkage meth-

ods; ie single, average as well as complete linkage we get similar results for the observed clusters.