Identifying the Van Hove singularity in overdoped Bi-2201 using k-means clustering

Horst, Marijn van der

Identifying the Van Hove singularity in overdoped Bi-2201 using k-means clustering

THESIS

submitted in partial fulfillment of the
requirements for the degree of

BACHELOR OF SCIENCE
in
PHYSICS

Author :	Marijn van der Horst
Student ID :	s2051907
Supervisor :	Dr. M.P. Allan
2^{nd} corrector :	Dr. W. Löffler

Leiden, The Netherlands, July 2, 2022

Identifying the Van Hove singularity in overdoped Bi-2201 using k-means clustering

Marijn van der Horst

Huygens-Kamerlingh Onnes Laboratory, Leiden University
P.O. Box 9500, 2300 RA Leiden, The Netherlands

July 2, 2022

Abstract

Cuprates are mysterious materials with many not yet understood properties. One of these properties is the Van Hove singularity, which enhances the electron-electron interactions where it occurs. This Van Hove singularity would be expected to appear in all cuprates, but has only been shown to appear in a single cuprate: Bi-2201.[2]
The Van Hove singularity presents as a peak in the graph of the density of states. To identify the locations in the material featuring the Van Hove singularity the dimension was reduced by using principal component analysis and calculating the mean squared distance to the mean of intervals to reduce the dimension of the data. After this dimension reduction a k-means clustering algorithm was used to find relations in the data. When analysing the clusters found by the algorithm it was found that the Van Hove singularities are localised a striped pattern on the sample and that the singularity varies a lot in the energy spectra. This variation in the energy spectra can be explained by putting the Van Hove singularity on the Fermi level and then displacing it because of a combination of inhomogeneity of doping and the formation of a superconducting gap.

Chapter 1

Introduction

A lot of current advanced technologies use the phenomenon of superconductors: currently powerful magnets for MRI scans[7] are powered by superconductors, quantum computers can use super conductors for qubits[8] and it is even used to make maglev trains[9] work. Crucial to each of these applications is the fact that superconductors have an electrical resistance of 0 Ohm. Traditionally superconductivity occurs when a material is cooled to very low temperatures. Currently one of the best explanations we have of this traditional superconductivity the Bardeen Cooper Schrieffer theory or BCS theory[10]. This will be discussed further in the theory section. This theory however does not hold up in a specific class of superconductors: the high-temperature superconductors. These superconductors can hold superconductivity for higher temperatures than is predicted by BCS theory. In this thesis we will look at a specific subclass of high-temperature superconductors: the cuprates. These cuprates are copper oxides which hold superconductivity up to high temperatures and exhibit unusual properties at high temperatures.[5]

A property we are interested in is the Van Hove singularity. This is a peak in the density of states at points where the tight-binding model has a local maximum or minimum. This peak in the density of states increases the electron interactions in this specific point. Understanding this Van Hove singularity could lead to a better understanding of how cuprates actually exhibit superconductivity.

This project will try to find the Van Hove singularity using data mining methods. It will do so following the master thesis written by Stolte[1]. This thesis used fitting, PCA and UMAP to reduce the dimensions of the

data and then analysed it by hand. This project will use PCA and a new method to reduce the dimension of the data, but will use k-means clustering to analyse the resulting data.

Chapter 2

Theory

2.1 BCS theory

The currently used theory to explain superconductivity is the Bardeen-Cooper-Schrieffer theory[10]. This theory states that the main mechanism behind superconductivity is the forming of Cooper pairs. Cooper pairs form as pairs of electrons after electron-phonon interactions, phonons are vibrations in a crystal lattice which behave as particles[4]. Because these cooper pairs are a bond between two electrons, which are fermions, these pairs themselves behave as bosons. This is important because normally only two fermions are allowed to occupy the same energy level due to the Pauli exclusion principle. When the pairs bond and behave as bosons they can, at low temperatures, condense into a Bose-Einstein condensate. This condensation means all the bosons can occupy the same energy level. The second, and more important, thing this condensation achieves is the effect that the energy of the entire condensate is changed when a single pair is broken. This means that the energy to break a single pair is related to the energy to break all pairs. So when the temperature is low enough the thermal energy in the material is not high enough to break all of the pairs and as such it is not high enough to break any of the pairs. This means that the electrons can move through the material without interacting with the lattice in the material and as such will have a resistance of 0. This temperature where the condensate is formed is called the critical temperature. The classically predicted highest temperature the T_c could reach was 30K, this was later edited up to 39K with the discovery of MgB_2.[5]
when the temperature of the system goes up the strength of the superconductivity reduces. This effect is best seen in the graph of the density of states of a superconductor. As certain energies are not able to break the

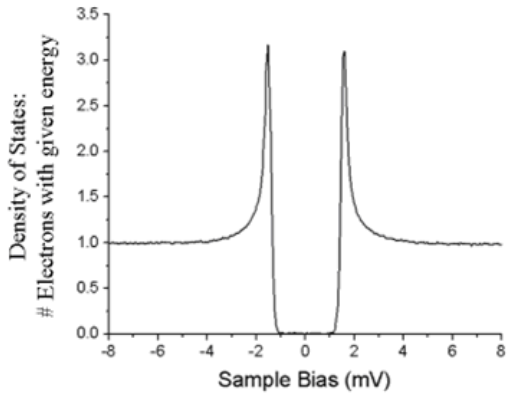


Figure 2.1: *The graph of a superconducting gap. When the temperature increases the gap grows more narrow until it vanishes. If the temperature is decreased the gap grows wider. Image adapted from:*`https://physics.stackexchange.com/questions/348501/physical-interpretation-of-density-of-states`

pairs in the condensate the states below the required energy to break a pair will not be occupied. The stronger the bonds in the condensate are, the greater the superconducting gap.

This superconducting gap does not just remove the existing density of states in the location it is formed. Instead it pushes the states to the side of the gap, forming a coherence peak on either or both sides of the gap. The amount of states represented by the graph should be the same as before the gap was formed.

2.2 High-Temperature Superconductors

While most superconductors behave as explained above, not all superconductors work like this. There is a class of superconductors called the high-temperature superconductors which function as superconductors well above the critical temperatures predicted by BCS theory. Currently it is not known what makes these superconductors function above the predicted temperatures.

2.2.1 Cuprates

A specific instance of these high-temperature superconductors is the group of superconductors known as the cuprates. Cuprates function like other high-temperature superconductors with some properties setting them apart

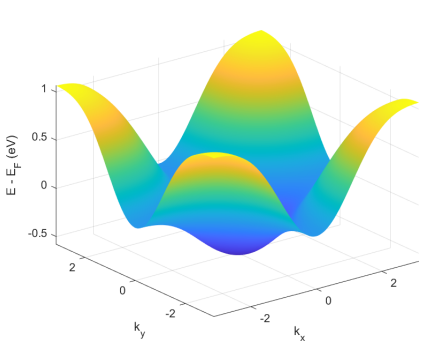
from the others. The following properties are discussed by Keimer et al[5] in their paper. When at no to very low doping the cuprates behave as a Mott insulator, creating an insulator rather than the expected metal. While within certain ranges of doping and sufficiently low temperature the cuprates exhibit superconductivity. When at lower doping the cuprate shows a pseudogap, which is a gap in which states are allowed but still heavily suppressed. When at higher doping the material enters the strange metal phase, where the cuprates resistivity scales linearly with temperature. When at high doping the cuprate behaves as a Fermi liquid, which is like normal metal.

It should however be noted that the cuprates studied in this project do not possess the incredibly high T_c seen in other cuprates. The feature we are looking for in this project should however still be present as it is not a feature which is restricted to high-temperature superconductors.
The doping used for the samples studied in this project was high enough that it should be in the superconducting phase with a transistion to a Fermi liquid. This high level of doping means the pseudogap and strange metal phases are not present in the studied samples.

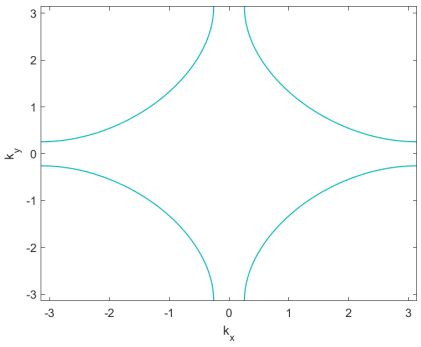
Van Hove Singularity

This feature we are looking for is the Van Hove singularity. The Van Hove singularity occurs when the band structure of the material has a local maximum, minimum or saddle point. The local density of states goes as one over the band structure so to identify the Van Hove singularity it can be found by finding a peak in the density of states. The Van Hove singularity has many interesting properties because it enhances the electron-electron interaction where it occurs. One can use the tight binding model to make a plot of the dispersion relation of material. The tight binding model is a tool used to easier represent the orbits around the atoms combined with a term to represent the chance for an electron to jump to another atom (the hopping term)[4]. The dispersion relation of the tight binding model of $Bi_2Sr_2CuO_{6+\delta}$ (Bi-2201) can be seen in figure 2.2 as calculated by He et al.[3]. Here the four saddle points can be seen at the boundaries of the plot signifying four expected Van Hove singularities.. These expected singularities are however very hard to detect and have only been directly detected in a single cuprate: Bi-2201 (the one used for this project).[2]

The Van Hove singularity is a material property, this can be seen because the temperature does not play a direct role in the tight binding model. The



(a) *A plot of the tight binding model as of Bi-2201 for a sample with $T_c = 15K$.[3] The saddle points causing the Van Hove singularities are clear.*



(b) *The cross-section of the tight binding model at the fermi level.*

Figure 2.2: *Plots of the tight binding model of Bi-2201 with a $T_c = 15K$. The code used to make the plots was made by W. Tromp.*

tight binding model is a result of the lattice it is calculated for, so the effects would be expected to be uniform across the lattice.
Because it is a property of the material the singularity should be present regardless of the temperature. When a superconducting gap opens on a spot where a Van Hove singularity is present we would expect the states to be pushed into the formed coherence peaks. If the singularity is on the left or right side of the superconducting gap it would be expected to be pushed into the corresponding coherence peak.

Unstable Fermi Surface

The Fermi surface of the cuprates has been shown to be unstable. This instability is because of local differences in the level of doping of the material. This inhomogeneity in doping affects the local electronic structure of the cuprates. As was shown by Wise et al[6] this inhomogeneity can have many effects on the material, including the shifting of the band structure by up to 20mV for the pseudogap state. While the superconducting state is less affected by this inhomogeneity of the doping, it is still affected.

Chapter 3

Methods

3.1 Goal

The aim of this research was to find structures in the measurements of Bi-2201. The main properties of interest are the peaks in the spectra which could be caused by multiple sources, the most relevant for this project are the Van Hove singularity and the peaks around the superconducting gap. Normally to classify spectra one would use supervised data mining algorithms, but these algorithms need to be trained on data. This training process requires (labeled) data to work. This constraint raises two problems for this project: the first is that the data is not labeled and quite a few spectra are ambiguous in how many peaks they have and what causes the peak. The second constraint is the amount of available data. To train a supervised learning algorithm one needs to feed it as much data as possible. When starting with the project there was concern about the amount of data available to train.

Because of these constraints the decision was reached to make use of unsupervised learning. Where supervised learning uses labeled data to predict the label of new data fed into the algorithm, unsupervised learning looks for correlations in the data to predict which instances belong together, but it will not explain these correlations.

The chosen unsupervised learning algorithm was k-means clustering. K-mean was chosen as it is a relatively easy algorithm and it makes the interpretation of the results much easier. the data for this experiment was made by measuring the energy at a point for different bias values, resulting in a ”data cube” where the x- and y-coordinates correspond to the spatial com-

ponents on the surface and the z-axis corresponds to the measured values for the measured energies. These measured values will be in dI/dV which corresponds to the density of states. The k-means algorithm requires a 2 dimensional matrix. This means that first the data needs to be put in the proper shape, this is done by creating a 2 dimensional matrix using the rows as being the separate measurements and the columns being the parameters. The first two columns are the spatial coordinates and the rest of the columns are the measured dI/dV.

For this thesis two data sets will be considered. One is a cuprate with a critical temperature of 12K or OD12K and the other is OD3K. The OD in the shorthand names stands for the fact that the cuprate has been overdoped before performing the measurements on it. The samples used to make the data sets supplied for this project were part of a larger series of measurement which also contained underdoped cuprates.

3.2 K-means clustering

For this project k-means clustering was used to find the correlations in the data. The motivation for using k-means will be explained more in the methods section, but here the method itself will be explained.
K-means clustering is a clustering algorithm which means it will classify the data into clusters, the algorithm will not attempt to give any information about the clusters other than them being the best fits for the cluster.
The k-means clustering algorithm works through starting by designating the number k to be the number of clusters, then k random points within the boundaries of the data are randomly chosen. The data is then assigned to a cluster corresponding to the closest point. The averages of these clusters are taken and then the partition is repeated with the averages as the new points (”centroids”). This process is repeated until the centroids remain stable between iterations or if the algorithm reaches a threshold of acceptable change.

3.2.1 Silhouette Score

A way to analyze the result of the clustering algorithm is to look at the silhouette scores. A silhouette score is calculated by

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{3.1}$$

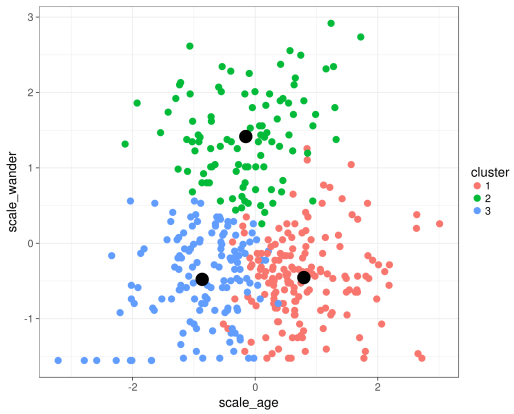


Figure 3.1: *result of a k-means clustering algorithm on a sample set of data. The clusters are denoted by different colors and the centroids are the bolded points. Image adapted from:* `https://rpubs.com/cyobero/k-means`

where a is the average distance from the point to all other points in the same cluster and b(i) is the average distance to points to another cluster for the cluster with lowest average distance to the point.
this score is between -1 and 1 with lower silhouette scores corresponding to poorly clustered points. When clustering the spatial coordinates are disregarded as the aim of the project is to find peaks in the spectra.

3.3 Dimensionality Reduction

Clustering on the 2 dimensional matrix would still do very little, as the euclidean distance loses most of its meaning at higher dimensions. Because the algorithm uses distance as the metric to analyze the data this is a problem. To deal with this problem two methods were applied. The first is PCA.

3.3.1 PCA

Principal component analysis(PCA) is a way to change the axes of the data along the lines of most variation. Principal component analysis requires a N x D matrix, where the N rows represent the repetitions of the experiment, or in the case of this experiment the different points on the surface. The D columns represent the different parameters of the experiment, or

the different measured energy values. From this matrix the principal components can be found using the following steps:

1. shift the mean of each column to 0.
2. make the D x D covariance matrix, consisting of the covariances of the parameters.
3. find the eigenvalues and corresponding eigenvectors of the covariance matrix
4. order the eigenvectors from highest to lowest eigenvalues. These are the new axes of the coordinate system. This order is important because the first eigenvector, which is also the first principal component, is the axis with the most variance.
5. make the projection from the old coordinate system to the new one. This transformation is 1:1 as it can be reversed if one has all the principal components.

While PCA space has as many dimensions as the original space, most of the variance (and thus information) is stored in the first few components. After calculating the principal components the lower principal components can be discarded to reduce the dimensions in the system. It should be noted that this means that an amount of information is lost and the transformation back stops being 1:1. This is not a problem if the matrix with the data is never reordered as the calculated labels can be applied to the original matrix. The centroids found did not lose information as they are calculated in the PCA space and can be transformed into the original space via the components.

3.3.2 Intervals

The second method of dimensionality reduction applied was splitting the measurement energies into intervals. For each interval the mean squared distance is calculated for the points to the average spectrum for the dataset. This mean squared distance is then used as the parameters for clustering. This means that if the number of intervals is set to 4 the dimensions of the output is 4.

3.4 Performing The Experiment

The code used to run the above operations can be found via the following link: `https://drive.google.com/drive/folders/1t7jaDhOuLkvFrRpmm4Hu1PMOF-s_f09Y?usp=sharing`. It also needs code from the Allen Lab shared Github, most notably read Nanonis and the data sets themselves. Read Nanonis turns the .3ds files containing the data into a usable matlab struct.

To check the results of the clusters multiple methods were used. The first among these is constant in both the PCA and interval models. This is the silhouette score, which can be calculated using an inbuilt function in MATLAB.

For the PCA model another way to check the results is by plotting the cluster centroids, which yield the middle for each cluster and see what is identifiable in those spectra. An important way of checking whether cluster is applicable after the transformation of the data is by plotting the points for the first few principal components.

For the interval method it is impossible to transform the found values back to the original dimensions so the chosen way to present the findings is by plotting the average of each cluster. By doing this we can get some insight into what spectra make up each cluster.

Version of July 2, 2022– Created July 2, 2022 - 02:06

Chapter 4

Results

In this chapter I will present the varying results of the methods on different datasets.

4.1 OD12K

In this section the results of the PCA method and the interval method on the OD12K sample are presented. After turning it into a 2D-array and labeling it the array is a 29929x86 matrix. This corresponds to 83 measurements being performed on a grid of 173x173 points. The 29929 is 173 times 173. The 86 represents the two coordinates, the measurements and the labels.

4.1.1 Interval Method

In figure 4.1a-c the cluster averages are plotted for the data set of OD12K. In 4.1d-f the silhouette values for the same clusters of the same data set are plotted. In 4.1g-i the spatial distributions of the data are plotted.
In figure 4.1 three spectra can be seen. One has two clear symmetrical coherence peaks around a (superconducting) gap. Another is a smaller gap with asymmetrical peaks, as it has only one peak. The last spectrum shows a gap but does not have any evident peaks.
In figure 4.1b two of the same spectra remain unchanged: the gap with no peaks and the two sharp coherence peaks. The single peak with a gap has been split into two spectra, each with a gap on the other side of the gap. When looking at figure 4.1c the two asymmetrical peaks remain but another spectrum appears which behaves similarly to the no peak spectrum

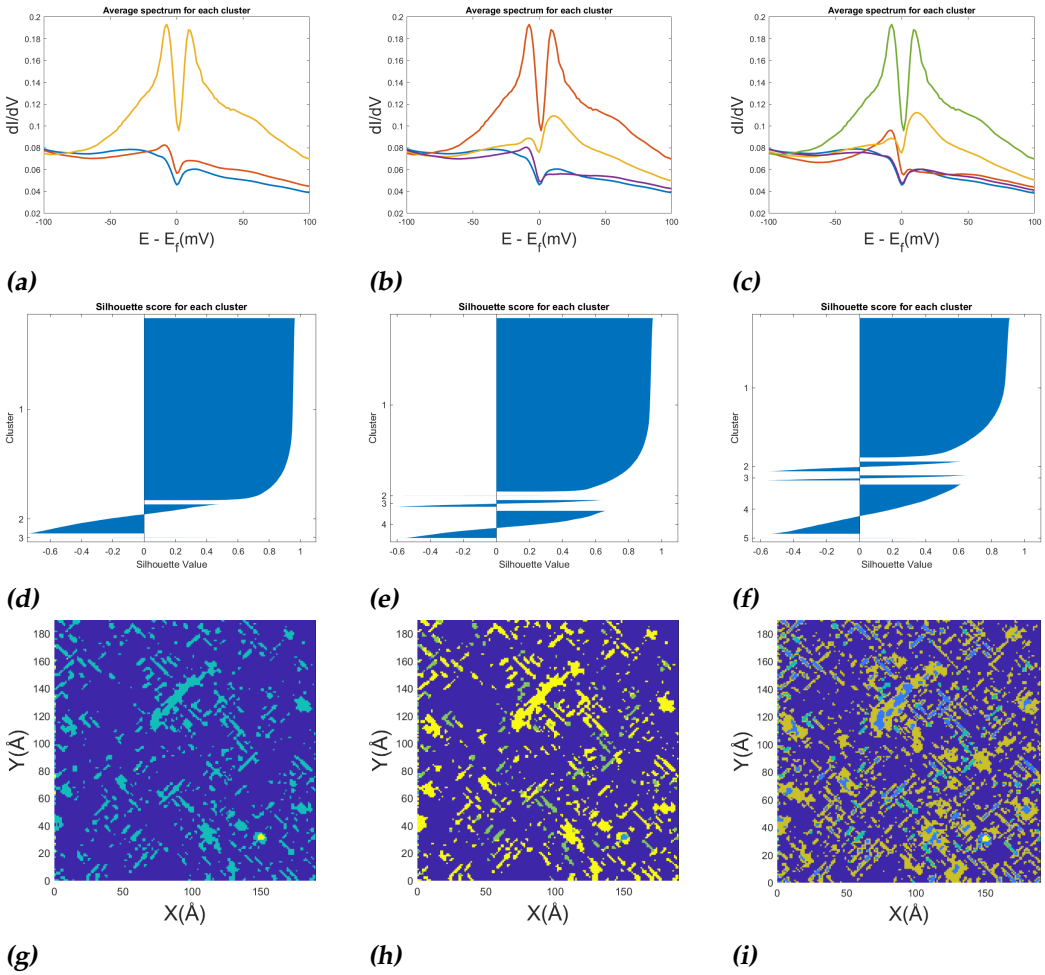


Figure 4.1: *The results of the interval method on the OD12K sample with k-means clustering for k=3,4,5 presented in a-c are the averaged spectra of the clusters with 3,4 and 5 centroids respectively. a shows spectra with 0,1 and 2 peaks. b shows spectra with 0,1 and 2 peaks. There are two spectra with one peak, each on the other side of the Fermi level. c shows spectra with 0,1 and 2 peaks. There are now two spectra with 0 peaks, showing no clear difference. d-f show the silhouette scores of the clusters for 3,4 and 5 means respectively. g-i show the spatial distributions of the clusters with 3,4 and 5 means respectively.*

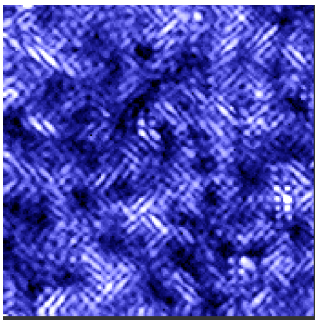


Figure 4.2: *The raw data as measured at the fermi level. The striped pattern is really clear here.*

from before.
This fifth cluster being almost the same to one already present in the original data points to 4 clusters being the correct amount of clusters, as no new information was gained with the new cluster.
This idea is supported by the silhouette scores shown in figures 4.1d-f. For the 3 clusters cluster 3 is quite poorly defined, as it consists of more poorly assigned points than properly assigned points. When moving to 4 clusters all clusters are more properly defined than they are poorly defined, although some points are still not assigned entirely properly. This ratio of properly assigned points goes down, however, when looking at 5 clusters. This also seems to support the idea of 4 clusters being optimal.
When looking at the spatial distributions both figures 4.1g and h seem to support the same striped structure, which is muddled quite a bit by adding a fifth cluster as is seen in figure 4.1i. When looking at the raw data using the aw(g) function available in the Allan lab Github the presence of these stripes in the data is evident, so all three ways of looking at the resulting clusters point to 4 clusters being applicable here. This raw data is pictured in figure 4.2.

When looking for a possible Van Hove singularity in the four cluster assignment three spectra stand out as the spectrum with no peaks does not have a clear enough peak to be called a Van Hove singularity. When looking at the symmetrical peaks we see the spectrum peaks far above the other spectra. This can be explained through a Van Hove singularity right in the middle of the superconducting gap, pushing the states equally to both sides. The two asymmetrical gaps show the superconducting gap as well, but also show a clear asymmetry in the peaks with either side being pushed up by a phenomenon other than the superconducting gap.

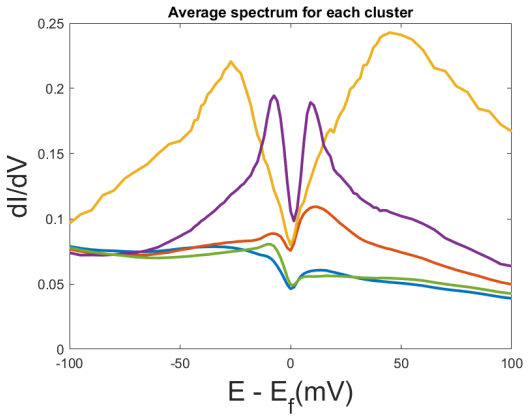


Figure 4.3: *The rarely occuring result of the clustering. For this run of the code it was run for 5 means, but it can also occur for 4 means. I have not yet observed it in 3 means, but that is likely due to the fact that 3 means has been tried the least. When found in 4 means it absorbs the coherence peaks, but in 5 means it apparently is found as a separate cluster.*

Drawbacks

The interval method will usually give the results as given above. Sometimes however the code will run and find a different structure, usually resulting in the two clear coherence going up and down in small steps, resulting in a graph resembling stairs. This outcome of the experiment is quite uncommon, but when it occurs the silhouette value is worse. It can be seen in figure 4.3.
I believe this occurs due to the random starting conditions of the k-means algorithm. When starting in random spots the algorithm will usually converge to the same stable result, however rarely it will find another stable solution leading to the structure discussed above.

4.1.2 PCA

To run the PCA clustering, I first wrote a code which also set the standard deviation to 1. This resulted in unusable plots as the clusters were all almost equal. While setting the deviation to 1 is useful when comparing

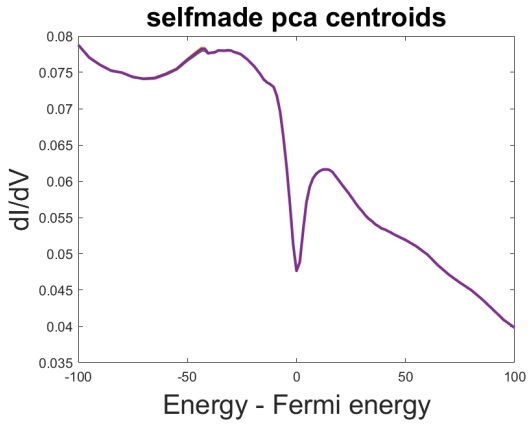


Figure 4.4: *The cluster centroids of of the clustering ran with self made code which set the standard deviation to 1, as is visible the spectra are indiscernible*

unrelated quantities to make sure one does not have the higher weight just because their values are measured in larger units, this is not applicable in this situation as all values are energy measured by the same device. This self made code is thus not useful for this experiment and the rest of this segment was done using the inbuilt pca() function in MATLAB.

All components

First the experiment was performed using all of the principal components and k=3,4,5 for k-means.
As can be seen in figures 4.5a-c the spectra are more two-peaked or zero-peaked spectra. Some spectra can be seen as either one-peaked or two peaked. This analysis of the number of peaks is further complicated by the implication in all of the spectra that there is a linearly declining background present. Some of the spectra curve down immediately after the gap, but it is unclear if this is due to a present peak or due to the apparently present background.
The present two-peaked spectra, especially in k=4 and k=5 example show a clear asymmetry, even larger than what seems to be the trend as given by the apparently present background. These spectra seem to be the most likely to have a Van Hove singularity as the zero-peaked spectra have no

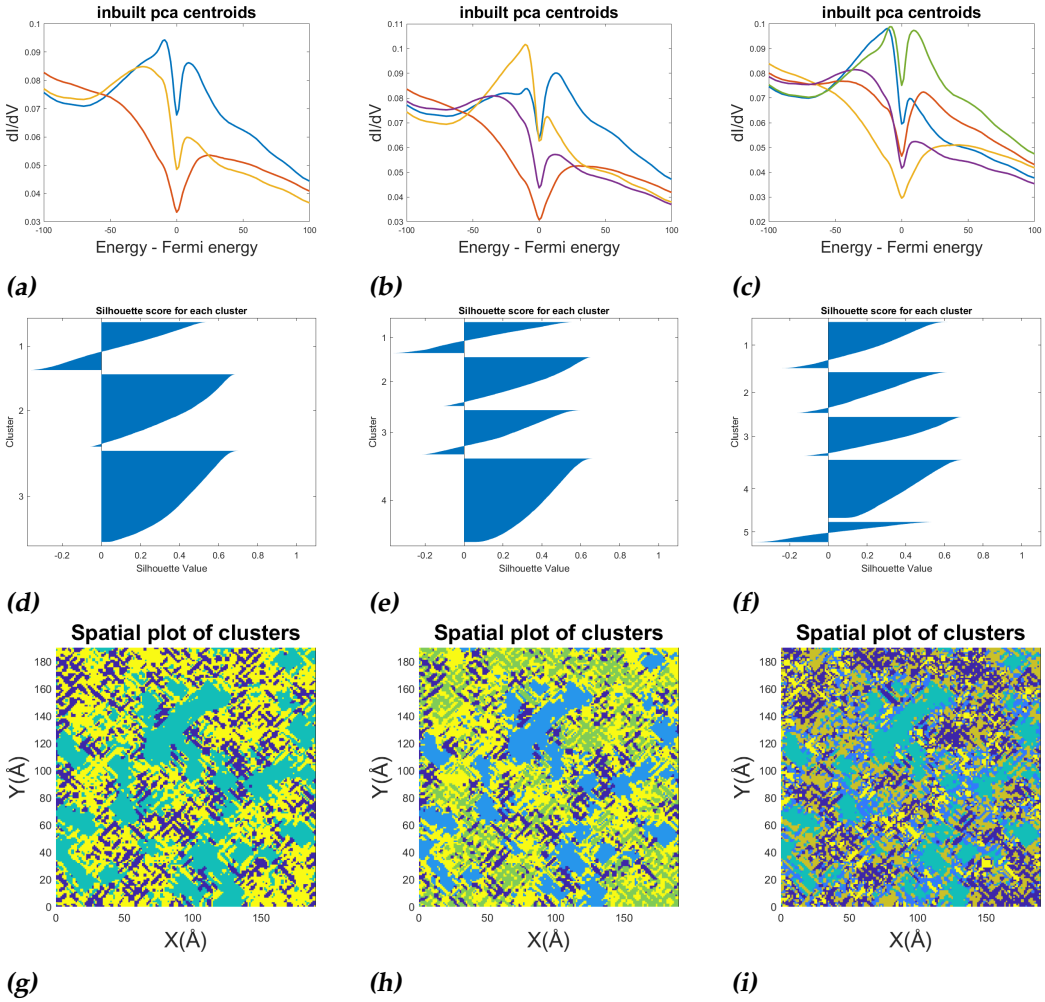


Figure 4.5: *Results of the k-means clustering on the OD12K data set using the PCA method with k=3,4,5. a-c show the centroids of the clusters. a shows spectra with 0 and 2 peaks, with two of the spectra having 2 peaks. b shows spectra with 0 and 2 peaks, with two of the spectra having 2 peaks. These spectra have highly asymmetrical peaks. c shows spectra with 0,2 and an unclear number of peaks. One of the spectra with two peaks has symmetrical peaks. d-f are the silhouette scores for 3,4 and 5 clusters respectively. g-i are the spatial distributions of the clusters for 3,4 and 5 clusters respectively.*

apparent peaks and the symmetrical spectra seem to be a direct consequence of the presence of the superconducting gap.

The silhouette scores for the assigned cluster are given in figure 4.5d-f. The silhouette scores seem to be overall very positive although for a lot of clusters the silhouette scores decline very quickly from the highest score to around 0. This implies that the clusters are overall assigned to the correct cluster, but the clusters are not particularly well defined.
The spatial distribution is presented in 4.5g-i. The spatial distribution is intricate and follows the striped pattern seen in the raw data as seen in figure 4.2. This implies the stripes have the same peaks within the same stripe.

The outcomes for this form of clustering were also noticeably less stable than the outcomes of the interval method. In the interval method the only spectra I noticed were the presented ones and the one talked about in the drawbacks section. In the PCA method the spectra changed more between different runs, and I have tried to present the ones which were the most common results.

Reduced Dimension

The above discussed result is quite promising so the amount of dimensions was reduced to 5 before clustering. The results are presented in figure 4.6. 4.6a-c are the centroids of the clusters as found by the algorithm. 4.6d-f are the silhouette scores for each of the clusters and figures 4.6g-i are the spatial distributions of the clusters on the actual sample.

When comparing the spectra in figure 4.5a-c with figure 4.6a-c the found centroids look much different to the average spectra found using clustering over all of the principal components, the reason for this is discussed later in this paragraph. Because the interval method also does not employ the centroids to visualise the spectra, but rather the average spectra I decided to plot these as well. The results for this are presented in figure 4.7a-c.

These spectra are much closer to what was presented in the previous figures outside of the dimension reduced centroids. When looking at the centroids presented in figure 4.6a-c the general trend of all the data seems to be followed somewhat while the details are not there or much less pronounced. The most likely explanation for this is that the first principal

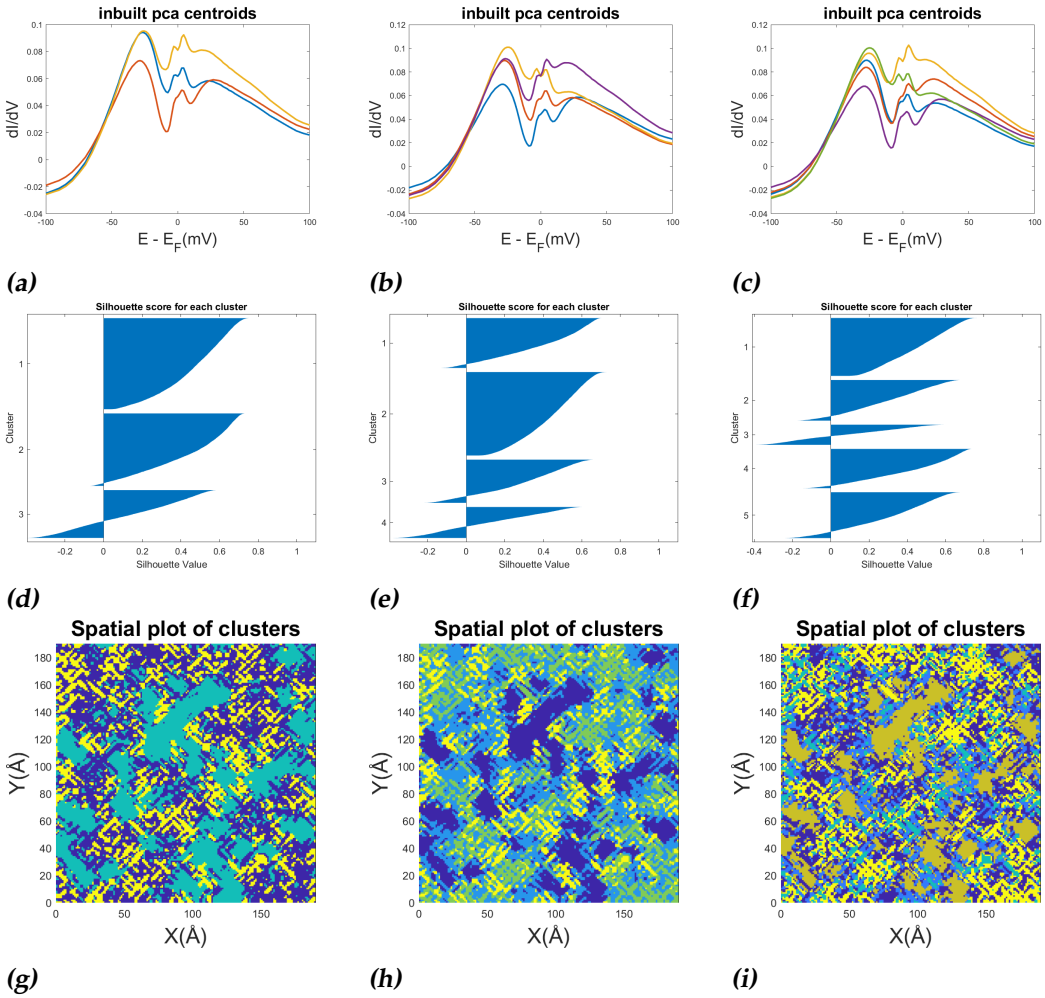
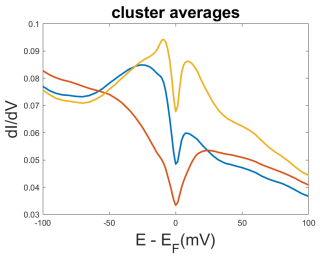
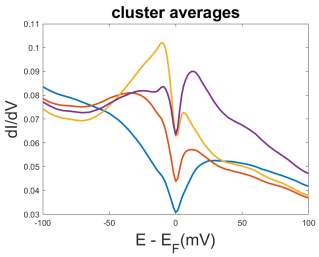


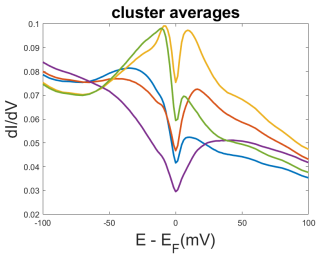
Figure 4.6: *Results of the k-means clustering on the OD12K data set using the PCA method, leaving out all principal components except for the first five. a-c show the centroids of the clusters for 3,4 and 5 clusters respectively. d-f show the silhouette score for 3,4 and 5 clusters respectively. g-i show the spatial distributions for 3,4 and 5 clusters respectively*



(a) *Average spectra of 3-means clustering of the OD12K sample using the PCA method using only the first five components*



(b) *Average spectra of 4-means clustering of the OD12K sample using the PCA method using only the first five components*

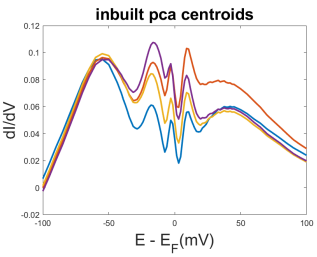


(c) *Average spectra of 5-means clustering of the OD12K sample using the PCA method using only the first five components*

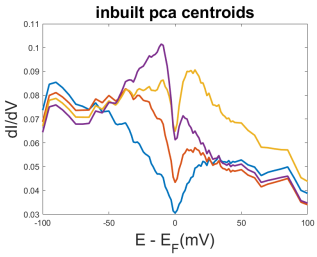
Figure 4.7: *The average spectrum for each cluster presented in figure 4.6*

components have the most variance in them and thus the most information. This means that the first principal components are chosen in such a way that they carry the general information but not the specific trends for each spectrum. I would therefore expect that the centroids move closer to the ones presented in figure 4.5a-c with each principal component added.

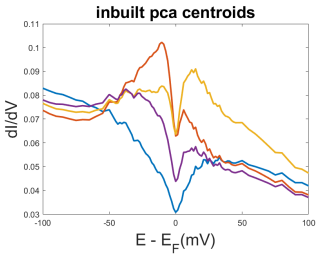
While the centroids are presented in the full 83 dimensions this is not how they are found. They are found using the first five principal components but then transformed back using the definition of the principal components. The resulting transformation will not add any new information. Because the original five components did not hold the information needed to identify the principal components, the centroids will not hold it either.
This theory was tested by increasing the dimensions and plotting the centroid spectra for each amount of dimension. The result of this can be seen in figure 4.8a-d. As can be seen in figure 4.8a-d the structures first become more pronounced, from first being almost the same spectrum to having more distinct features and finally the curves become smoother.
This shows that the centroids are not the correct way to represent the results for the PCA method and instead for this one the average spectrum should be used as well.



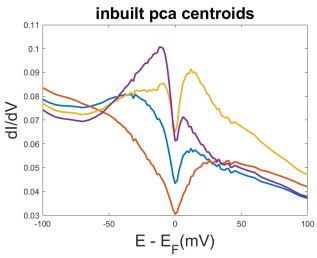
(a) *Average spectra of 4-means clustering of the OD12K sample using the PCA method using only the first ten components, this still looks quite a bit like the spectrum for the 5 dimensions*



(b) *Average spectra of 4-means clustering of the OD12K sample using the PCA method using only the first twenty components. General structures of the spectra start to arise.*



(c) *Average spectra of 4-means clustering of the OD12K sample using the PCA method using only the first fourty components. The structures from the previous example are more nuanced.*



(d) *Average spectra of 4-means clustering of the OD12K sample using the PCA method using only the first sixty components. The structures become a little smoother.*

Figure 4.8: *Centroid spectra plotted for a 4-means clustering using the PCA method, varying the amount of dimensions. Dimensions used are 10,20,40 and 60.*

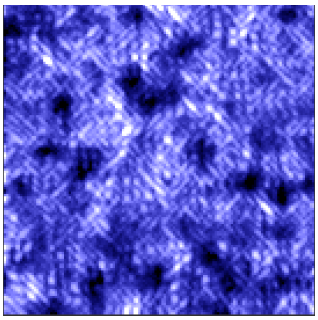


Figure 4.9: *The raw data of the measurements of dI/dV at an energy level close to the fermi level (-8.2mV) of the OD3K sample*

4.2 OD3K

In this section the results of the interval method and the PCA method on the OD3K sample are presented. After turning it into a 2D-array and labeling it the array is a 18225x122 matrix. This corresponds to 119 measurements being performed on a grid of 135x135 points. The 29929 is 135 times 135. The 122 represents the two coordinates, the measurements and the labels.
The raw data of the measurements at an energy close to the Fermi level is presented in figure 4.9. Once again a striped pattern is visible.

4.2.1 Interval Method

The results of the clustering using the interval method on the OD3K sample have been presented in figure 4.10. Figure 4.10a-c show the average spectra of each cluster, figure 4.10d-f show the silhouette score of each cluster and figure 4.10g-i show the spatial distribution of the clusters on the sample.

In figure 4.10a the average spectra for the clusters of the k=3 clustering have been presented. It can be observed that there is a spectrum with two asymmetrical peaks, a spectrum with one peak and a spectrum without clear peaks.
Figure 4.10b shows the average spectra for k=4. A new spectrum has appeared which is almost identical to the spectrum with no clear peaks from 4.10a. This implies that no new information was gained with the new spectrum.
Figure 4.10c shows the spectra for k=5. The previous spectra remain, but the two-peaked spectrum is more asymmetrical than before and a new,

symmetrical spectrum has been added.

The silhouette scores are presented in figure 4.10d-f. When going from k=3 to k=4 a new cluster is added, this cluster shows up as having a poor performance on the silhouette plot, having worse performance than the previously present clusters with a large portion of it being assigned poorly and the properly assigned part of the silhouette plot quickly dropping to 0. When going from k=4 to k=5 this happens again.

Both of the addition of a spectrum with no new information when going from k=3 to k=4 and the silhouette score for the new clusters performing poorly signals that the proper choice of clusters for this data set would be 3 when using the interval method. Each of the spectra in figure 4.10a could hide a Van Hove singularity, but the two spectra with clearly defined peaks are more likely to be interesting for this purpose, as they clearly show a peak.

While more clearly present in the k=4 and k=5 examples, the striped pattern is present in the spatial distribution of the k=3 clusters. As this pattern is also clearly available in the raw measurements it is good to see this pattern reflected in the spatial distribution of the clusters.

4.2.2 PCA method

As the functionality of the PCA method has been explored in the previous segment, the same steps will not be taken again. The clustering will again be performed on all components and on just five components.

all dimensions

In figure 4.11 the results of the PCA method using all components are presented. 4.11a-c are the average spectra, 4.11d-f are the silhouette scores and 4.11g-i are the spatial distributions on the real sample.

In figure 4.11a the average spectra of the clusters for 3-means are plotted. The gaps of these spectra are noticeably small. There is a spectrum with one peak, there is a spectrum which is unclear whether it has a peak or not and there is a spectrum without peak.
When looking at the next plot in 4.11b, the small gaps make way for a spectrum with no gap at all. The spectrum with one peak has also gained an additional peak on the other side of the gap, which has grown.

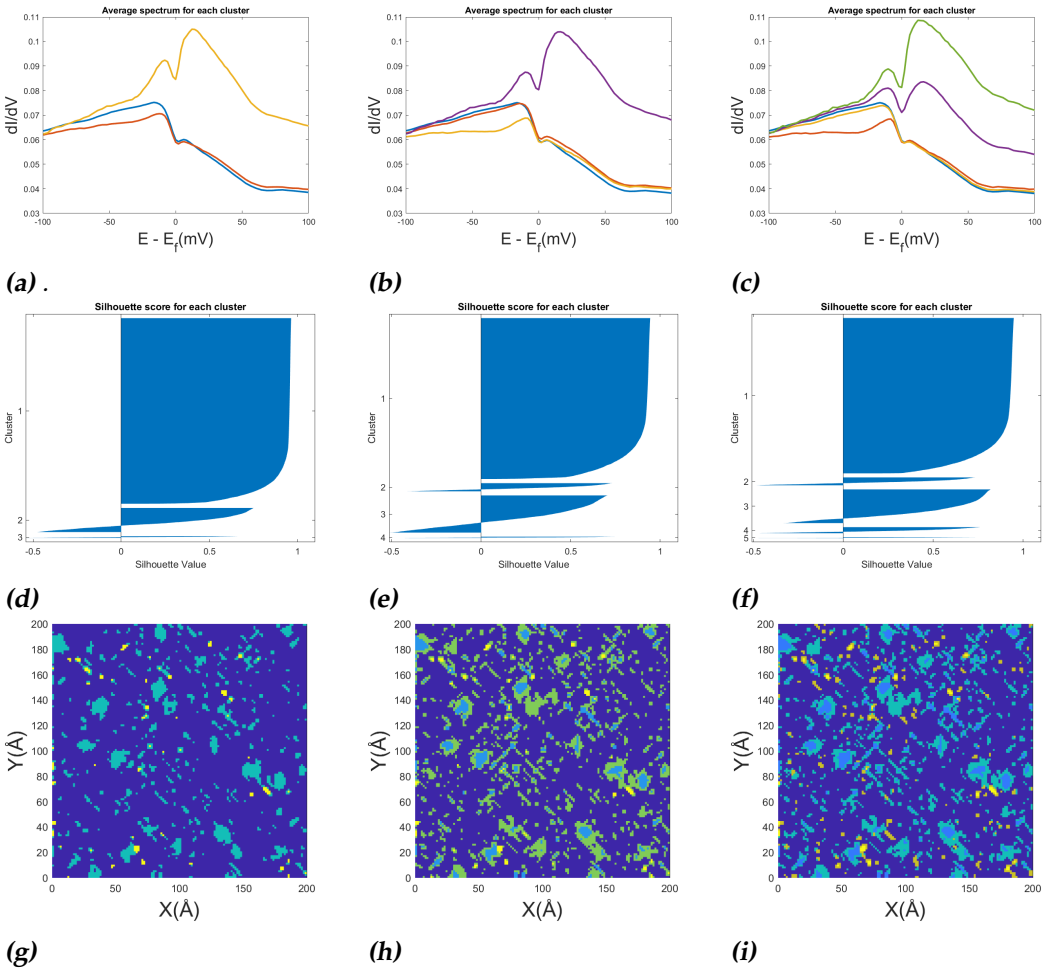


Figure 4.10: *Results of the k-means clustering on the OD3K data set using the interval method for k=3,4 and 5. a-c present the averaged spectra of the clusters. a shows spectra with 0,1 and 2 peaks. b shows spectra with 0,1 and 2 peaks. Two spectra with 1 peak can be seen but both are very similar. c shows a result from 5 means clustering, where the number of peaks is quite unclear for a lot of the spectra. d-f show the silhouette scores for 3,4 and 5 clusters respectively. g-i show the spatial distributions of 3,4 and 5 clusters respectively.*

Looking at 4.11c it becomes hard to decipher which spectra actually hold a peak and which do not.

When looking at the silhouette scores presented figure 4.11d shows a large cluster with reasonably well defined points, when moving to figure 4.11e this cluster appears to have been broken up into smaller clusters with worse scores. This happens once again in figure 4.11f.

When looking at the spectra there is no clear better amount of clusters, but when looking at the silhouette scores the favored amount of clusters is 3. When looking at the spectra presented here, the most interesting spectrum to look at is the one with the clearly defined peak, this might be due to a Van Hove singularity.

When looking at the spatial distributions presented in 4.11g-i the striped pattern is visible quite clearly indicating that each stripe is a spectrum with an amount of peaks consistent with the rest of the stripe.

5 Dimensions

The results of the PCA method using only the first five principal components are presented in figure 4.12.
Figure 4.12a shows the average spectra for the k=3 clustering on the first five principal components of the OD3K sample. The spectra in this figure are identical to ones presented in figure 4.11a. This means that the clusters found by the first five principal components and all of them are the same. In figure 4.12b the average spectra are presented for the k=4 clustering. The spectra are once again identical to the ones presented in figure 4.11b. Figure 4.12c is once again identical to figure 4.11c.

In figure 4.12d-f the silhouette scores are presented. While the clusters are identical, the silhouette scores are not. This is due to the fact that the data set used to find the clusters was different. The same clusters appear, but they have lower silhouette scores than the ones presented in figure 4.11d-f.

The spatial distributions are once again identical because the clusters are as well.

The lower silhouette scores mean that the certainty of the algorithm is higher when using all principal components than when using only the first five components. This is however also more computationally expen-

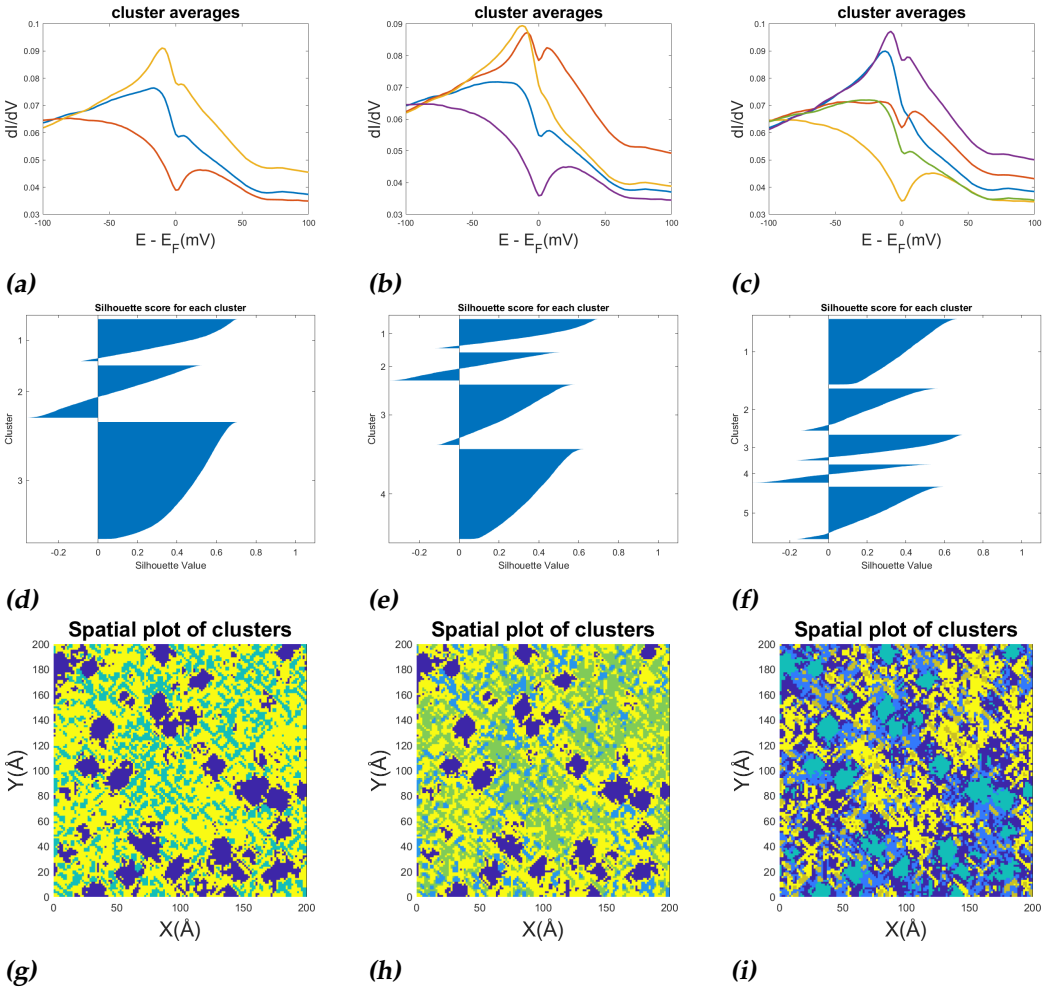


Figure 4.11: *Results of the k-means clustering on the OD3K data set using the PCA method for k=3,4 and 5. a-c show the cluster averages for 3,4 and 5 clusters respectively. a shows a spectrum with a clear peak, and two spectra with an unclear number of peaks. b shows a spectrum with 2 peaks, a spectrum without peak and two spectra without clear peaks. c shows a spectrum with 2 peaks, a spectrum with no gap and three spectra with 0 peaks. d-f show the silhouette scores of 3,4 and 5 clusters respectively. g-i show the spatial distributions of 3,4 and 5 clusters respectively*

sive while the shape of the silhouette scores remains the same.

4.2.3 The tails

The spectra presented in this section had more data attached on the tail-end, but this data has been left out because the research is most interested in the behaviour around the Fermi level and because the tails are all quite similar. For completeness the spectra presented above are presented in figure 4.13 with the tail-end.

Version of July 2, 2022– Created July 2, 2022 - 02:06

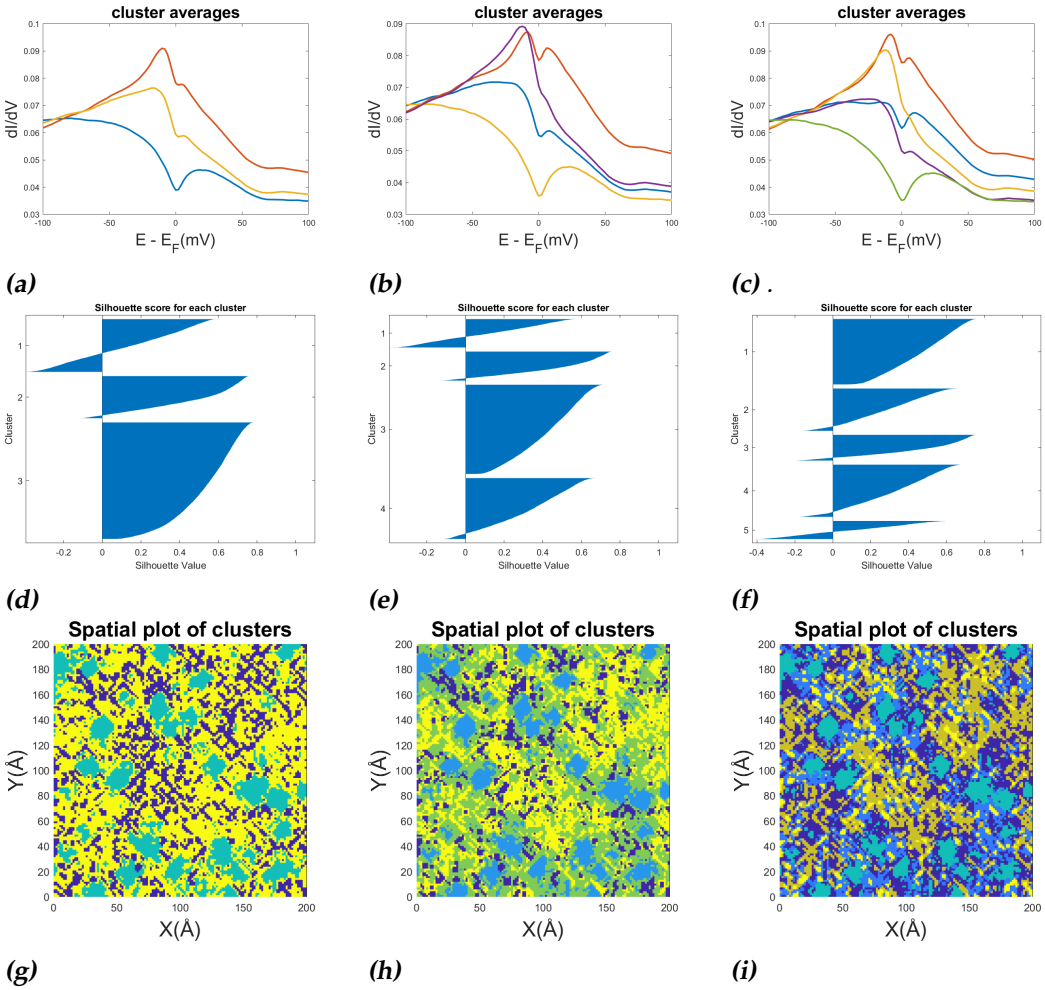


Figure 4.12: *Results of the k-means clustering on the OD3K data set using the PCA method using only the first five components for k=3,4,5. a-c show the average spectra for each cluster. The found clusters for this method are the same as with the PCA method with all components. d-f show the silhouette scores for 3,4 and 5 clusters respectively. These are different compared to the previously found values due to the loss of information in going from all components to 5. g-i show the spatial distribution which are the same as for the PCA method with all components*

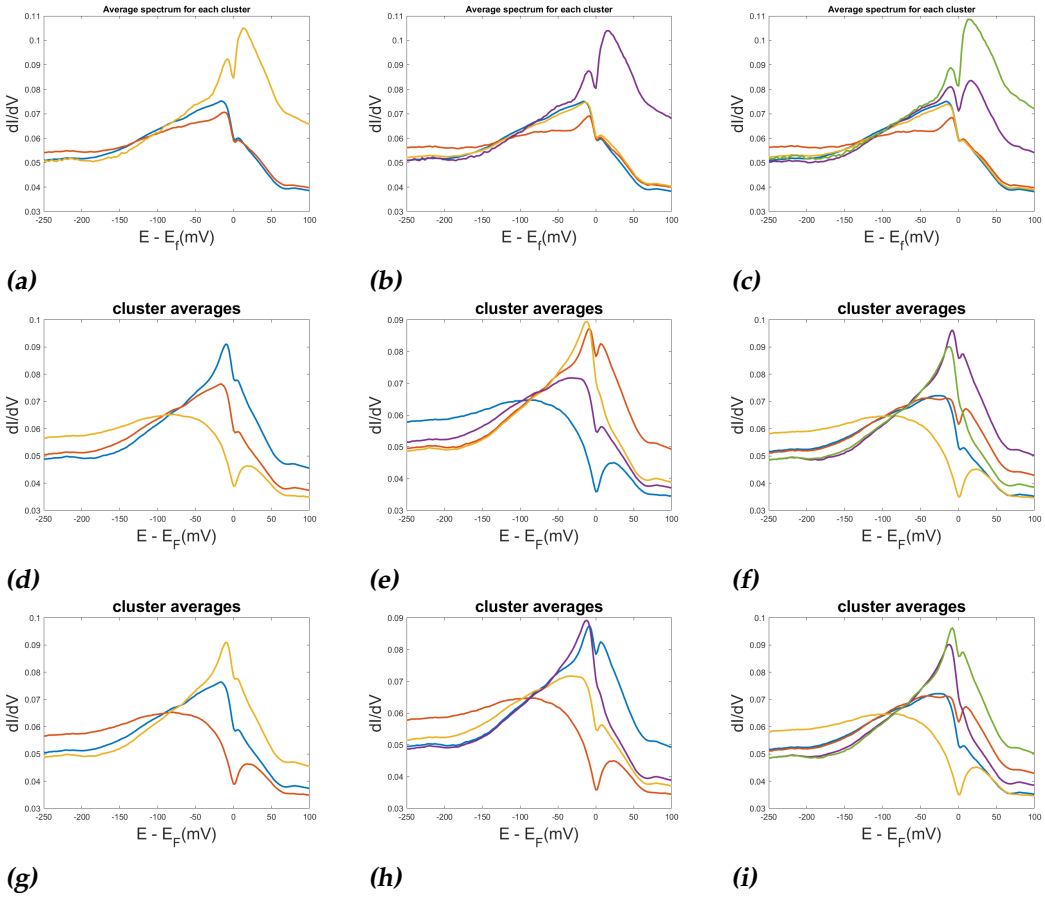


Figure 4.13: *The full spectra of the earlier presented results. a-c represent the results for the interval method, d-f represent the PCA with all dimensions and g-i represent the PCA method with 5 dimensions*

Chapter 5

Discussion

5.1 Methods

5.1.1 PCA Method

The hypothesis at the beginning that the centroids would be a proper way to look at the resulting clusters for the PCA method with reduced dimensions has been proven wrong. Replacing this with averaging the spectra proved a good replacement to find meaningful spectra plots of the clusters.

Due to the nature of PCA discussed in the methods section, the clustering of all principal components holds the same information as clustering over the entire data set.

5.1.2 Interval Method

The interval method mostly performed as expected. Because of the way it was defined it was more sensitive to the peaks in spectra than the PCA method as it functioned as a mean squared deviation from the mean over which was then clustered rather than raw data.
While this was not done for this project an interesting thing to look at in the future is the impact of increasing the amount of intervals, which is already built into the code used for this project but was not done due to time constraints.

5.1.3 Comparing The Methods

The first thing which is evident when looking at the results of both methods is the different cluster sizes. The interval method tends to find a large, well defined cluster and some smaller ones next to that while the PCA method finds more evenly divided clusters, which are usually defined slightly worse than the large, well defined cluster of the interval method. The smaller clusters found by the interval method usually score worse than the PCA method clusters, mostly due to a high percentage of poorly assigned points in each of the clusters.

The choice for which of the two methods to use eventually depends on the purpose of the research. For this project the interval method provides easier to interpret answers because of the earlier discussed tendency to identify peaks.
This tendency is most clear when looking at the OD12K sample. This sample exhibits a large gap with clear peaks in the interval method, while the peaks stand out far less in the PCA method. When looking at the spatial distribution of both of these clusterings one thing stands out: in the interval method there is a small, localised cluster on the lower right at around (150,35). This cluster is not present in the PCA method. When using the aw(g) function to look at the data directly, this small area of the data does indeed show these two clear peaks around a large gap.
This illustrates how the interval method is better at finding small structures with clear peak.

Comparing the silhouette scores between the different methods is not a proper way of looking at the data as the data sets are different for each of the methods. When looking at the silhouette scores they indicate how well defined the clusters are within the given data set, so the silhouette score also depends on the data set given. Because the data set used to cluster is different for each of the methods the comparison cannot be made based on the silhouette scores.

The spatial distribution is not useable to identify the best clusters as there is not a direct assumption as to how these are supposed to look or a way to verify it directly.

Because this project was aimed at finding the Van Hove singularity in the data which presents as a peak in the spectra, and the silhouette scores not being 1:1 comparable the best method for it seems to be the interval

method because of the clarity of the peaks.

5.2 Structure Of The Materials

When looking for the location of the Van Hove singularity in the sample the location seems to be given by the striped pattern as is shown by the spatial plots of all the plotted clusters as they all show the clusters at least loosely follow the striped pattern, with most spectra having a separate form.

This indicates that when identifying a striped pattern in the Bi-2201 samples the structure of the spectra follows the stripes, so when looking for a spectrum the first step would be to identify a point with the desired spectrum and then following the striped pattern to find more points with the desired spectrum.

To compare the results of the clustering the spatial distributions of each cluster is plotted separately. This can be seen in figure 5.1-5.4. Here we see the four clusters as found by the interval method and the PCA method for the OD12K sample as presented in figure 5.1 and figure 5.2. The three clusters found for the OD3K sample using the interval method and the PCA method can be seen in figure 5.3 and figure 5.4.

The OD12K sample with the interval method shown in figure 5.1 shows a large cluster with no peaks, the spectrum which features the two peaks is highly localised in a spot in the lower left quadrant and the two spectra with peaks follow the earlier observed striped pattern, with a larger portion of it being the peak on the left side of the Fermi level.

The OD12K sample with the PCA method shown in figure 5.2 shows two spectra with no clear peaks being clusters 1 and 2, together having a large portion of the sample between them clusters 3 and 4 show a spectrum with a peak below the Fermi level and a peak above the Fermi level respectively. The spectra with peaks can once again be seen to travel along the striped pattern. The spectra without peaks seem to follow the same pattern, but when taken together this pattern shows much less, this can be seen in figure 5.5.

The OD3K sample with the interval method shown in figure 5.3 shows a spectrum with no clear peak in cluster 1 and two spectra with peaks in

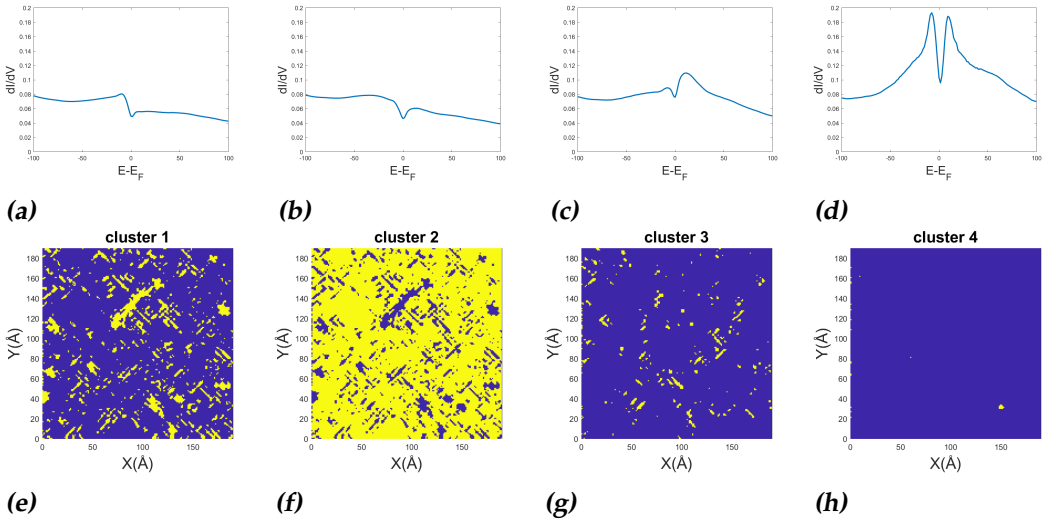


Figure 5.1: *The separate clusters of the OD12K sample with the interval method. a-d are the spectra of the cluster averages. e-h are the spatial distribution for each cluster showing the cluster in yellow*

clusters 2 and 3. Cluster 3 seems to once again behave in a sort of striped pattern while the spectrum with no peaks is by far the largest portion of the sample.

The OD3K sample with the interval method shown in figure 5.4 shows two spectra with no clear peaks and one spectrum showing a clear peak in cluster 2. Here the cluster with the peak is once again distributed along the striped pattern in the sample.

The Van Hove singularity seems to travel along a striped pattern in all of the examples, but what is also striking is the size of the spectra without a clear Van Hove singularity. As was discussed before the Van Hove singularity is expected to be uniform across the surface of the material and is expected to be present in the spectra due to the extrema in the dispersion relation.

5.3 The Van Hove Singularity

The result found in the previous section is very intriguing, as the result does not comply to the general understanding of the Van Hove singularity. This singularity was expected to arise as a result of the band structure of the crystal lattice but it does not behave as a material property. The fact

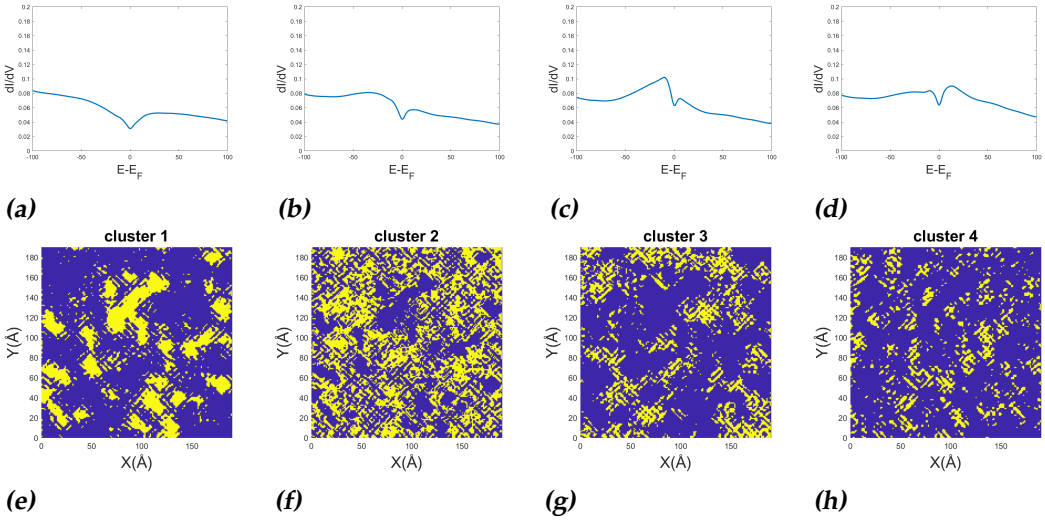


Figure 5.2: *The separate clusters of the OD12K sample with the PCA method. a-d are the spectra of the cluster averages. e-h are the spatial distribution for each cluster showing the cluster in yellow*

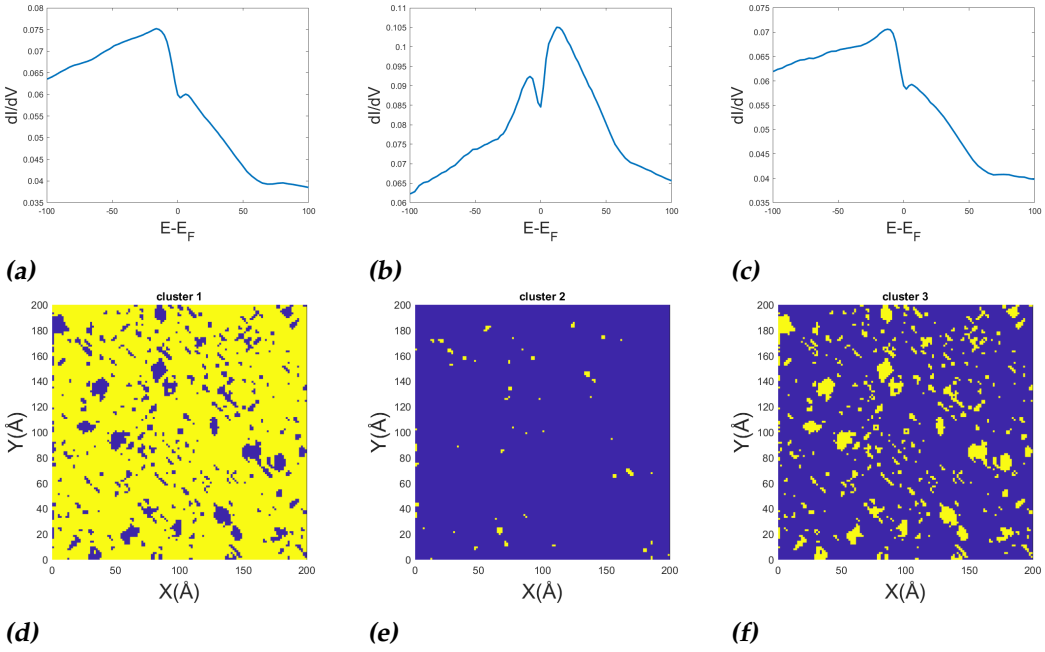


Figure 5.3: *The separate clusters of the OD3K sample with the interval method. a-c are the spectra of the cluster averages. d-f are the spatial distribution for each cluster showing the cluster in yellow*

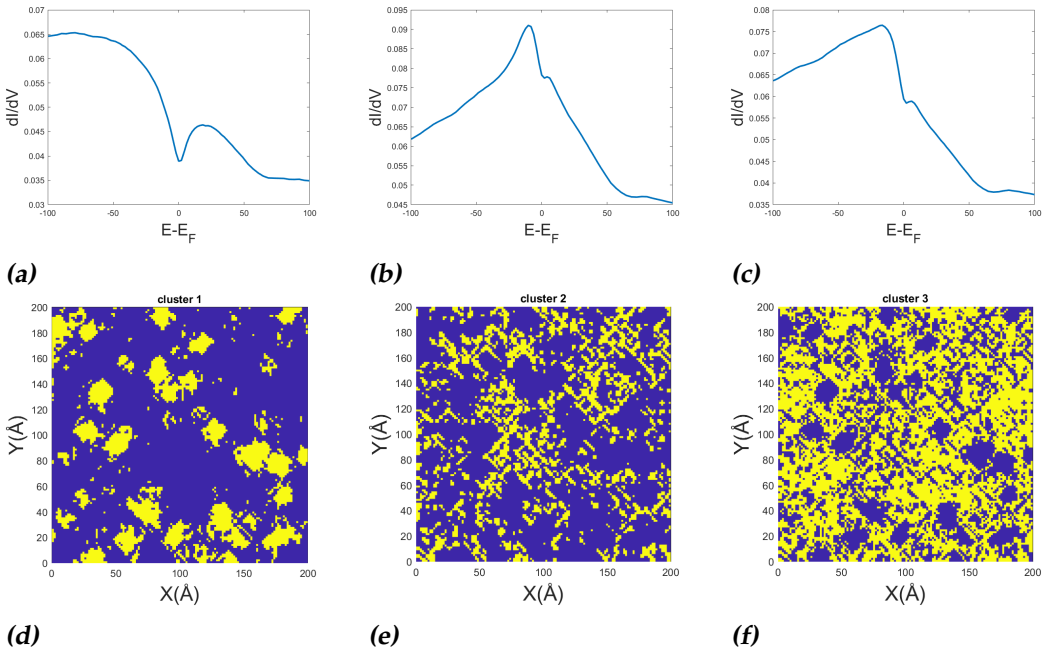


Figure 5.4: *The separate clusters of the OD3K sample with the PCA method. a-c are the spectra of the cluster averages. d-f are the spatial distribution for each cluster showing the cluster in yellow*

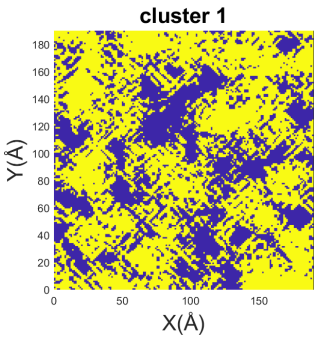


Figure 5.5: *The pattern of the two spectra without peaks from the OD12K with PCA, to show the absence of a striped pattern in itself*

that the clusters representing the different peaked spectra travel along the observed striped pattern is quite unexpected as this behaviour means that it varies very quickly in the direction perpendicular to the stripes.
This uniformity is also not preserved in the other expected property of the Van Hove singularity. Because it is expected to be a property of the material it is expected to be a peak at the same energy level, but the peaks vary in energy level from below the Fermi level to the Fermi level to above the Fermi level. This fluctuation could be explained by the earlier theory if it varied slowly in space, but as was discussed above the locations of the peak vary quickly in the space when moving in certain directions and stay the same when moving along the stripes.

To explain this unexpected result the discussed inhomogeneity of the material can be used. As was discussed the band structure is shifted because of the differences in doping. This would mean that the Fermi level as represented by the graphs in the earlier chapters are incorrect, as the x-axis is made in terms of the energy minus the Fermi energy. A changing Fermi energy would change the x-axis for the affected spectra. There is one problem with this theory compared to the presented data. All superconducting gaps as presented in chapter 4 are centered at 0 or close to 0. This means that the shift in axis cannot be large enough to account for the large changes in location of the Van Hove singularity, unless it is usually centered at almost exactly 0 and are pushed to the side by the shifted axis and then pushed to the side of the formed superconducting gap.
This proposed solution is supported by the fact that there were no identifiable peaks away from the superconducting gaps.

This could also explain the rapid spatial variation as the actual variation of the Van Hove singularity would not have to be large but just enough to cross to the other side of the Fermi level. Thus the best explanation for the Van Hove singularity with the information available is the idea that the Van Hove singularity is located extremely close to or on the Fermi level.
This solution however does not deal with the absence of the Van Hove singularity in a large portion of the spectra.

5.4 Moving Forward

For future research on this topic there are three directions to go:

1. Conclusively identifying the Van Hove singularity

2. Optimizing the currently used methods
3. Using a new form of data mining or machine learning to identify the peaks

For option 1, the found spectra will need to be analysed against the already known facts about the Van Hove singularity to prove the presence of the Van Hove singularity in the found spectra. This is hard to do as the current reason these peaks are believed to be Van Hove singularities is because no other theory fits the resulting data. Currently there is no clear way to prove the peak as a Van Hove singularity.
This means that the peak in the spectra could also be due to a yet unknown mechanism rather than an actual Van Hove singularity. It is however exceedingly hard to find the other mechanism used to produce the peak as currently we do not have a good explanation for it.
For option 2, the currently employed methods have variables not yet varied too much during this project, most notably the amount of dimensions of the PCA method and the amount of intervals used in the interval methods. A future project could look into optimizing these parameters to find better optimized clusters.
For option 3, a possible option would be training a neural network or one-shot learning algorithm on the data to identify the peaks more conclusively. This could be trained either on the raw data, data with reduced dimension or it could even be used with the graph using an image learning algorithm. Using this method would possibly lead to more exact ways to identify the spectra at the cost of explainability of the algorithm as most neural networks function as black box algorithms which have hidden layers and as such their results are not as easy to explain as the ones found here.

The image based algorithm could be interesting because a peak is sometimes more easily visible in a graph than when looking at data because the height at the base of the graph can vary and the height of the peak can also vary quite a bit. This is by no means a certainty that it would work but it is an interesting avenue of research. It could also grant insight into the difference between image based recognition algorithms versus the raw data used to make the images.
The raw data could be better than the image recognition algorithm because it is the rawer form of the data and computers are less visually oriented than humans and may be better at finding peaks in the raw data. It could run into the problem that the different peaks are at a different level to the point where some peaks are lower than the base of the other peaks. This

might be able to be overcome by a clever set-up of the network, but that would have to remain to be seen by a following researcher.

5.5 Conclusion

In conclusion, the variation of the location of the Van Hove singularity along the energy axis can be explained by the Van Hove singularity appearing at the Fermi level and being pushed to the side of it because of the inhomogeneity of the material. This slight variation is then amplified by the formation of the superconducting gap, pushing the Van Hove singularity to the side of the superconducting gap. This theory does not however explain the absence of the Van Hove singularity in the majority of the spectra but it provides a new starting point for following research.

Version of July 2, 2022– Created July 2, 2022 - 02:06

Bibliography

[1] Stolte, E. (2021, februari). Local interplay between the superconducting gap and the Van Hove singularity in overdoped Bi2201.

[2] Piriou, A. et al (2011). First direct observation of the Van Hove singularity in the tunnelling spectra of cuprates. Nature Communications.

[3] He, Y., Yin, Y., Zech, M., Soumyanarayanan, A., Yee, M. M., Williams, T., Boyer, M. C., Chatterjee, K., Wise, W. D., Zeljkovic, I., Kondo, T., Takeuchi, T., Ikuta, H., Mistark, P., Markiewicz, R. S., Bansil, A., Sachdev, S., Hudson, E. W. Hoffman, J. E. (2014). Fermi Surface and Pseudogap Evolution in a Cuprate Superconductor. Science, 344(6184), 608â611. https://doi.org/10.1126/science.1248221

[4] Simon, S. H. (2013). The Oxford Solid State Basics. Oxford University Press.

[5] Keimer, B., Kivelson, S. A., Norman, M. R., Uchida, S., Zaanen, J. (2015). From quantum matter to high-temperature superconductivity in copper oxides. Nature, 518(7538), 179â186. https://doi.org/10.1038/nature14165

[6] Wise, W. D., Chatterjee, K., Boyer, M. C., Kondo, T., Takeuchi, T., Ikuta, H., Xu, Z., Wen, J., Gu, G. D., Wang, Y., Hudson, E. W. (2009). Imaging nanoscale Fermi-surface variations in an inhomogeneous superconductor. Nature Physics, 5(3), 213â216. https://doi.org/10.1038/nphys1197

[7] Aarnink, R., Overweg, J. (2012). Magnetic Resonance Imaging, a success story for superconductivity. Europhysics News, 43(4), 26â29. https://doi.org/10.1051/epn/2012404

[8] Sparkes, M. (2021, November 18). IBM creates largest ever superconducting quantum computer. New Scientist. https://www.newscientist.com/article/2297583-ibm-creates-largest-ever-superconducting-quantum-computer/

[9] Principles of the Superconducting Maglev system — SCMAGLEV — Central Japan Railway Company. (n.d.). SCMAGLEV. https://scmaglev.jr-central-global.com/about/

[10] Bardeen, J., Cooper, L. N., Schrieffer, J. R. (1957). Theory of Superconductivity. Physical Review, 108(5), 1175â1204. https://doi.org/10.1103/physrev.108.1175