

# **RootProf**

## **TUTORIAL 9**

### **Analysis of single crystal X-ray data**

#### **Contents**

Chapter 1: The data set.....	pag.2
Chapter 1: Analysis of reciprocal lattice.....	pag.3
Chapter 2: Principal component analysis.....	pag.11

# Chapter 1

## The data set

3D patterns from X-ray diffraction measurements on membrane protein crystals compose our dataset. All patterns have been acquired at the Diamond synchrotron, by using crystals of Reaction Center from *R. Sphaeroides*. Many crystals have been obtained by slightly varying crystallization conditions. All patterns have the same crystallographic symmetry ( $P 3_12_12$ ) and nearly equal crystal cells. They have all been referred to the same asymmetric unit. Samples with subscript number come from the same well, so they have same crystallization conditions. The composition of the dataset is reported in Table 1. The corresponding files are included as demo files. They are formed by four columns: the three Miller indexes and the corresponding values of intensity.

**Table 1:** Samples used for reciprocal lattice analysis.

Number	Name
0	xtal1_1.dat
1	xtal1_2.dat
2	xtal3_2.dat
3	xtal9.dat
4	xtal10_1.dat
5	xtal10_2.dat
6	xtal7.dat
7	xtal11.dat
8	xtal12.dat
9	xtal14.dat
10	xtal13_1.dat

## Chapter 2

### Analysis of reciprocal lattice

#### Motivation

Obtaining a quick and joint view of the data taken by single crystal diffraction experiments.  
Grouping datasets according to their intensity patterns.

#### The command file

The list of commands for qualitative analysis of such dataset is the following.

```
whichanalysis 0
  dataType 3
  figpaper 1
  PreProcess 0 3 0 0
  file xtal1_1.hkl
  file xtal1_2.hkl
  file xtal3_2.hkl
  file xtal9.hkl
  file xtal10_1.hkl
  file xtal10_2.hkl
  file xtal7.hkl
  file xtal11.hkl
  file xtal12.hkl
  file xtal14.hkl
  file xtal3_1.hkl
```

The commands have been included in the demo file named *fileInputHKLFirstSight*. See the user guide for an explanation of their meaning.

#### Running RootProf

Start ROOT by clicking on his icon, or by typing “root” on a terminal window. Then write the root command:

```
Root> .x RootProf.C("fileInputHKLFirstSight")
```

or

```
Root> .> outputHKLFirstSight
```

```
.x RootProf.C("fileInputHKLFirstSight")
```

```
.>
```

After some seconds, graphic windows will start appearing on your screen, while text output will appear on the terminal window, or redirected in the file named *outputHKLFirstSight*. When the run

ends, the root prompt will appear again on the ROOT terminal, and you will be able to edit each single graphic window and read the output file by your text editor.

### The graphic output

The graphic windows in Figs. 1-4 show the 3D diffraction patterns of four datasets. Each point represent a node of the reciprocal lattice measured by the detector during the diffraction experiment. The width of the point is proportional to the measured intensity of that node. The 3D plots report the three Miller indices H, K, L on their three axis. In all cases the same asymmetric unit has been sampled during the experiment. The shape of the filled region of the plots depends on the strategy used to acquire data.

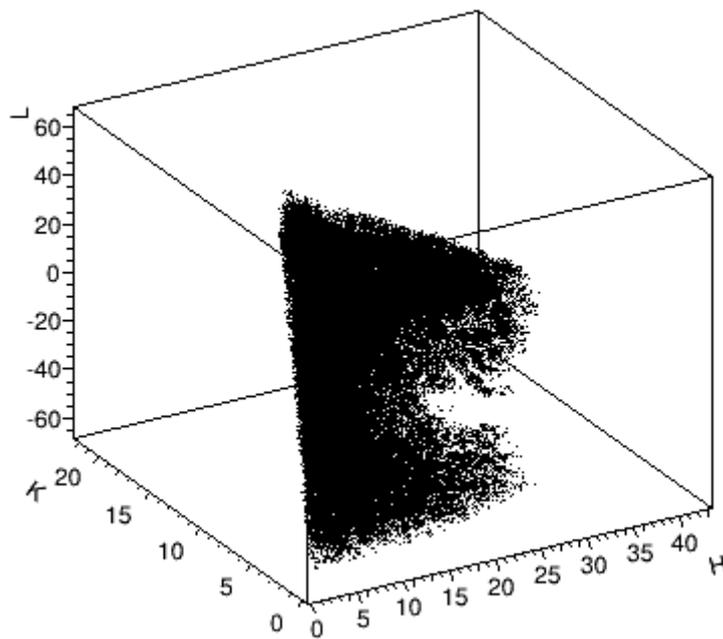


Fig. 1 Reciprocal lattice of sample 2

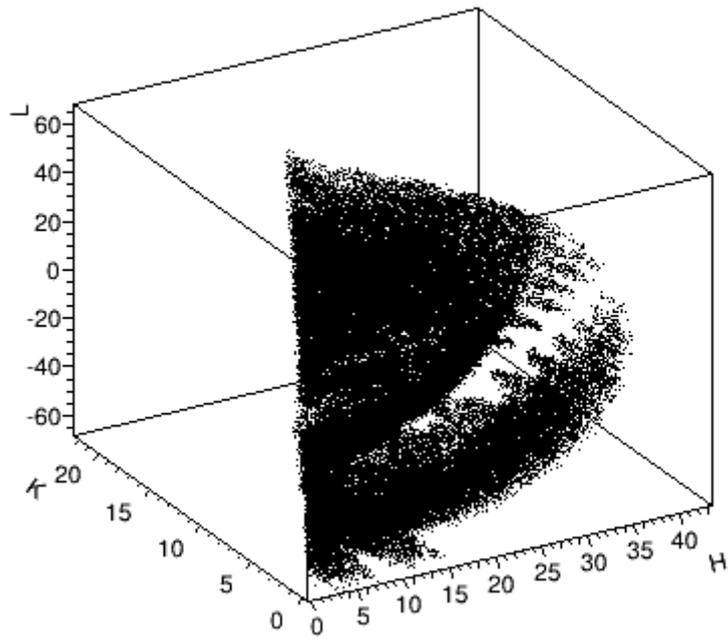


Fig. 2 Reciprocal lattice of sample 4

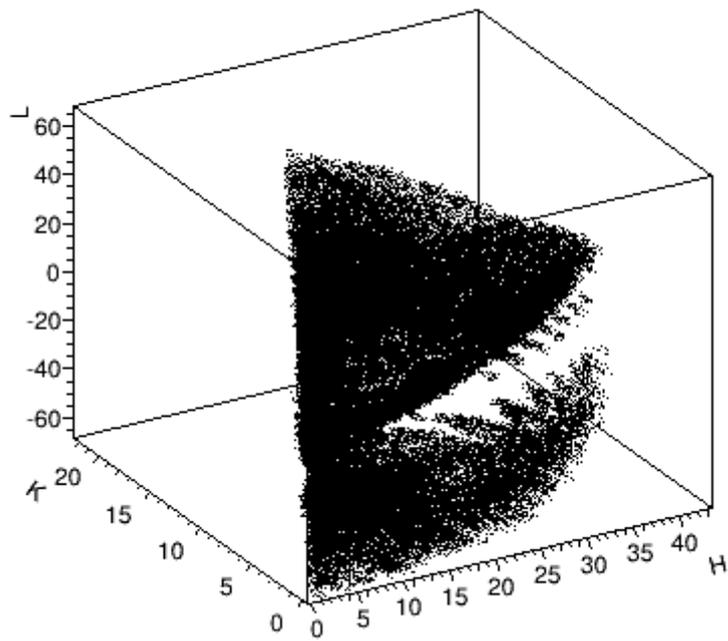


Fig. 3 Reciprocal lattice of sample 5

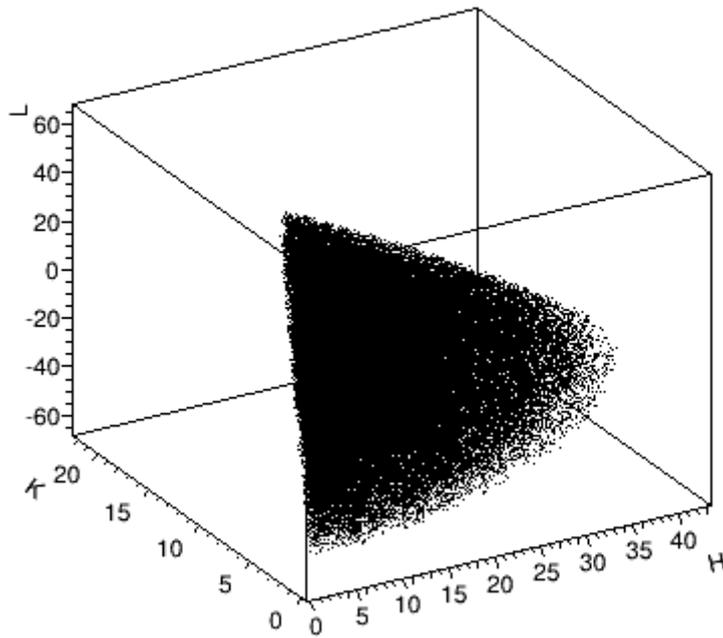


Fig. 4 Reciprocal lattice of sample 9

The matrix of superpositions (Fig.5) has been obtained by calculating for each couple of patterns the number of points which are contemporary present in the two patterns. Thus higher superpositions (red squares in Fig.5) indicate closer datasets.

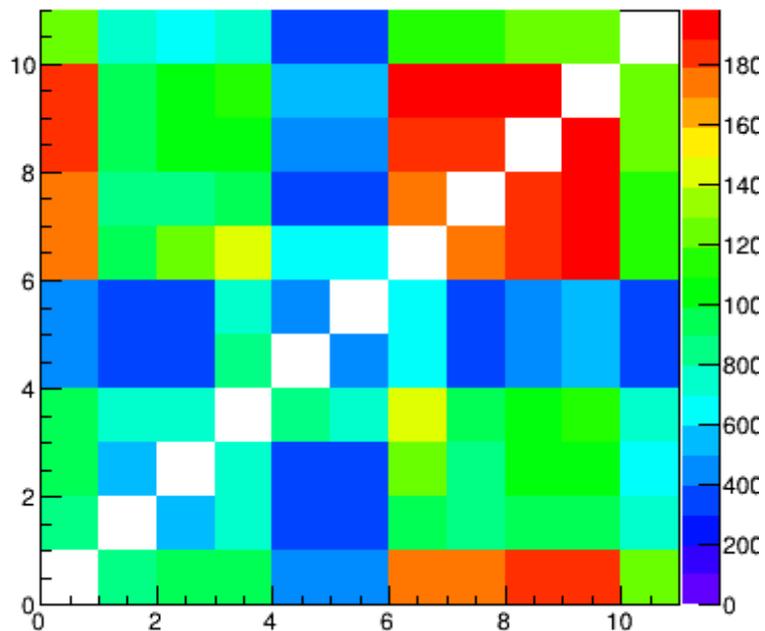


Fig. 5 Matrix of superpositions

The plot in Fig.6 represent the projection of the matrix of the superpositions on one axis. It is very important to decide which dataset to include in the qualitative analysis of single crystals patterns (see Chapter 3). In fact, patterns with low bin content in this figure have very low superposition with most of the others, so they should be removed from the qualitative analysis, as it is performed by using point in common to all the patterns.

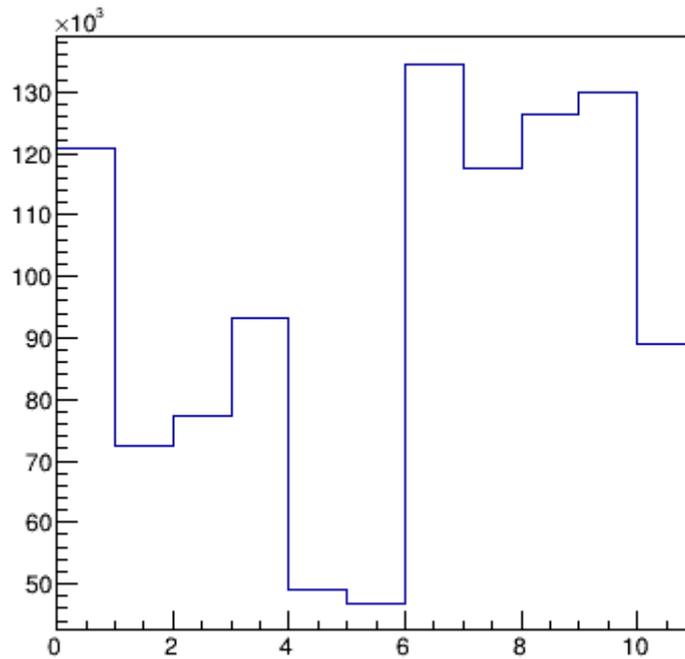


Fig. 6 Projection of the matrix of superpositions

Fig. 7 shows the matrix of superpositions rearranged after having grouped the patterns. A matrix of distances is calculated, by using the metrics N-S, where N is the largest number of nodes of the reciprocal lattice among the input data, and S is the number of superpositions between two datasets. Thus the distance is inversely proportional to the number of superpositions. As a result, the clustering algorithm groups datasets with similar sampling of reciprocal lattice.

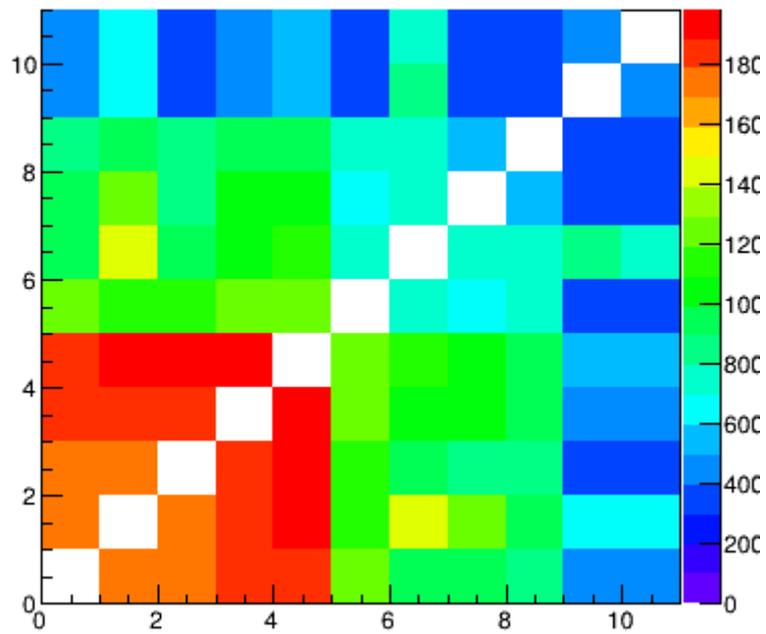


Fig. 7 Matrix of superpositions after clustering

## Output file

The content of the output file named *outputHKLFIRSTSIGHT* is reported below, with comments added.

```
Input from file: fileInputHKLFIRSTSIGHT
```

```
-----
whichanalysis 0
```

```
dataType 3
```

```
figpaper 1
```

```
PreProcess 0 3 0 0
```

```
file xtal1_1.hkl
```

```
file xtal1_2.hkl
```

```
file xtal3_2.hkl
```

```
file xtal9.hkl
```

```
file xtal10_1.hkl
```

```
file xtal10_2.hkl
```

```
file xtal7.hkl
```

```
file xtal11.hkl
```

```
file xtall2.hkl
```

```
file xtall4.hkl
```

```
file xtal3_1.hkl
```

The section above shows the commands read from the command file. It should be checked to ensure that they are interpreted correctly.

```
Sample 0 -> file xtall 1.hkl
          Found 20707 points
Sample 1 -> file xtall_2.hkl
          Found 13734 points
Sample 2 -> file xtal3_2.hkl
          Found 17445 points
Sample 3 -> file xtal9.hkl
          Found 25851 points
Sample 4 -> file xtall0_1.hkl
          Found 13032 points
Sample 5 -> file xtall0_2.hkl
          Found 12679 points
Sample 6 -> file xtal7.hkl
          Found 33573 points
Sample 7 -> file xtall1.hkl
          Found 20304 points
Sample 8 -> file xtall2.hkl
          Found 22811 points
Sample 9 -> file xtall4.hkl
          Found 23881 points
Sample 10 -> file xtal3_1.hkl
          Found 14003 points
```

The section above reports the number of data points read within each input file.

```
Miller Index Limits: H=[0 43] K=[0 22] L=[-68 68]
```

The section above reports the minimum and maximum Miller indices found in the input files.

```
===== Dendrogram =====
Step      Dist      Sample 1      Sample 2
  10         28887.20         0         5
   9         28636.44         0         4
   8         25338.38         0         1
   7         24343.14         0         2
   6         23203.17         0         3
   5         21437.40         0        10
   4         15617.50         0         6
   3         15340.33         6         7
   2         14685.50         6         8
   1         13755.00         8         9
=====
Normalized Cluster threshold: 0.200000 (0.315384)
Normalized Cluster threshold redefined: (0.200000) 0.315384
Cluster Threshold 18527.449
```

The section above shows the dendrogram resulting from the hierarchical clustering. The value of the threshold distance chosen to define the number of clusters is reported.

#### Cluster analysis

```
Cluster 1 5) 0 6 7 8 9
Cluster 2 1) 10
Cluster 3 1) 3
Cluster 4 1) 2
Cluster 5 1) 1
Cluster 6 1) 4
Cluster 7 1) 5
```

```
Cluster: 1
Member: 1 Number: 0 File: xtal1_1.hkl
Member: 2 Number: 6 File: xtal7.hkl
Member: 3 Number: 7 File: xtal11.hkl
Member: 4 Number: 8 File: xtal12.hkl
Member: 5 Number: 9 File: xtal14.hkl
```

```
Cluster: 2
Member: 1 Number: 10 File: xtal3_1.hkl
```

```
Cluster: 3
Member: 1 Number: 3 File: xtal9.hkl
```

```
Cluster: 4
Member: 1 Number: 2 File: xtal3_2.hkl
```

```
Cluster: 5
Member: 1 Number: 1 File: xtal1_2.hkl
```

```
Cluster: 6
Member: 1 Number: 4 File: xtal10_1.hkl
```

```
Cluster: 7
Member: 1 Number: 5 File: xtal10_2.hkl
```

The section above analyzes the formed clusters, specifically the content of each cluster in terms of samples and file names, its center and distance calculated by using the number of superpositions among patterns.

## Chapter 3

### Principal component analysis

#### Motivation

Classify single crystal datasets according to their intensities, by using common nodes of the reciprocal space.

#### The command file

The list of commands is the following.

```
whichanalysis 1
dataType 3
figpaper 1
PreProcess 0 3 0 0
file xtal1_1.hkl
file xtal1_2.hkl
file xtal3_2.hkl
file xtal9.hkl
file xtal10_1.hkl
file xtal10_2.hkl
file xtal7.hkl
file xtal11.hkl
file xtal12.hkl
file xtal14.hkl
file xtal3_1.hkl
```

They have been included in the demo file named *fileInputHKLQualitative*. See the user guide for an explanation of each command.

#### Running RootProf

Start ROOT by clicking on his icon, or by typing “root” on a terminal window. Then write the root command:

```
Root> .x RootProf.C("fileInputHKLQualitative")
```

or

```
Root> .> outputHKLQualitative
```

```
.x RootProf.C("fileInputHKLQualitative")
```

```
.>
```

After some seconds, graphic windows will start appearing on your screen, while text output will appear on the terminal window, or redirected in the file named *outputHKLQualitative*. When the

run ends, the root prompt will appear again on the ROOT terminal, and you will be able to edit each single graphic window and read the output file by your text editor.

### Data matrix representation

Graphic windows in Figs. 1 and 2 show the input diffraction patterns. They have been reduced to unidimensional profiles by using the technique of the super-index: The three Miller indices (h,k,l) are put on the variable  $ind=h*A+k*B+l$ , where  $B=2*l+1$  and  $A=B*(2*k+1)$ , so that each spot of the diffraction pattern has a unambiguous correspondence to a ind value. Here only the spots which are in common with all the diffraction pattern are considered for further analysis.

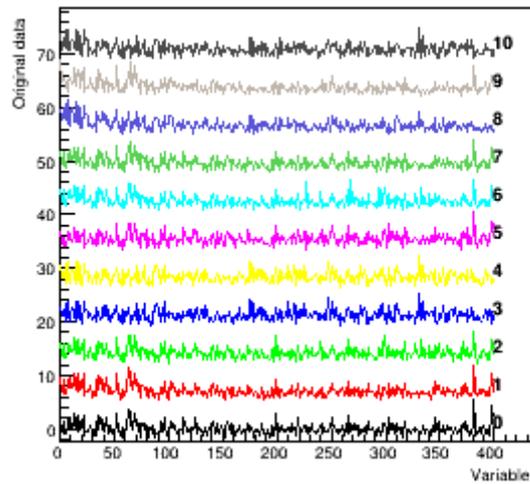


Fig.1 Original data shifted

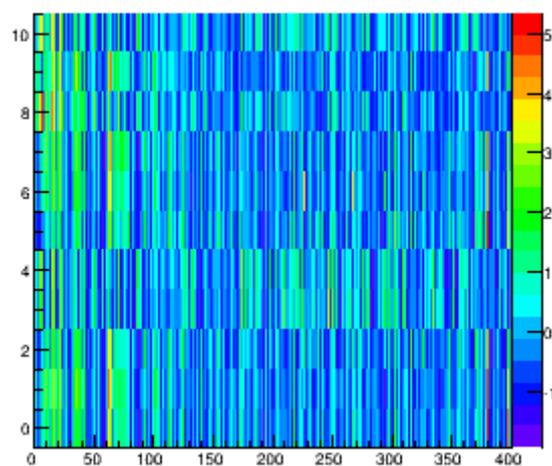


Fig.2 Data Matrix

PCA can thus be applied to unidimensional profiles so obtained. The scree plot, scores and loadings are reported in Figs. 3, 4 and 5, respectively. It can be noted that PC1, which explains more than 70% of total variance, produce a sharp separation between data points.

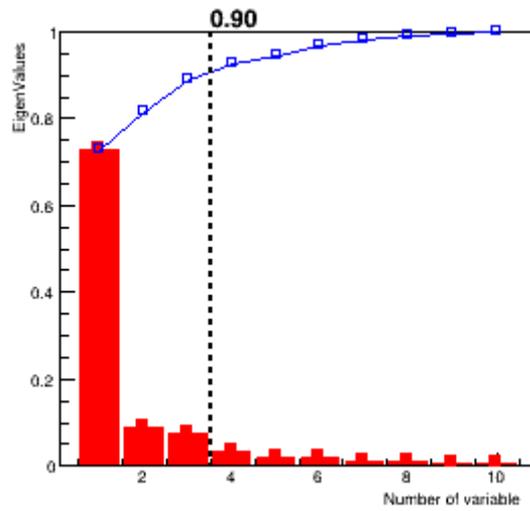


Fig.3 Scree plot

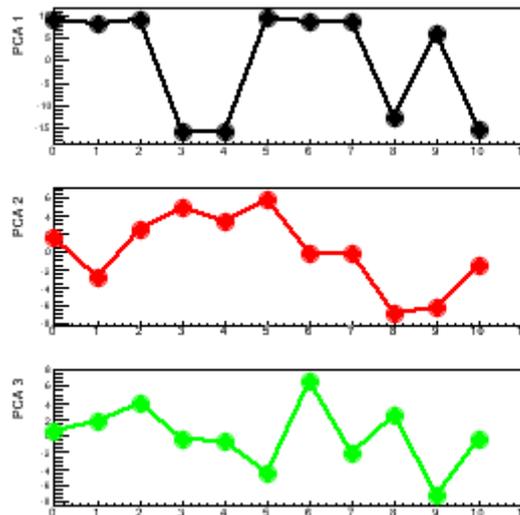


Fig.5 Scores

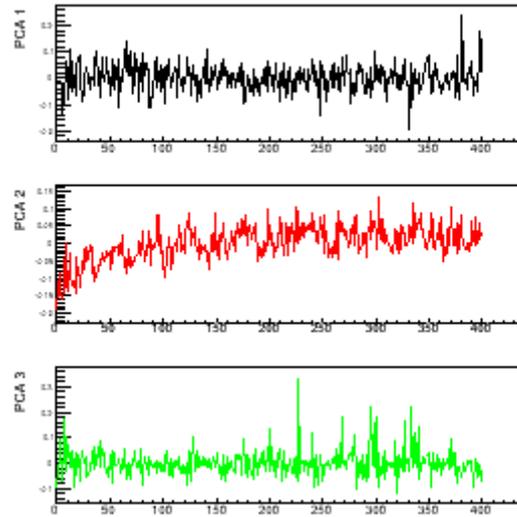


Fig.6 Loadings

Score plots in PC1-PC2 (Fig.7) and PC1-PC3 (Fig.8) show a clear separation between data points, caused by PC1. 95% confidence level ellipses indicate how the two groups are separated. A structural analysis pointed out that this separation is due to two non-equivalent lattice arrangement of the same membrane protein.

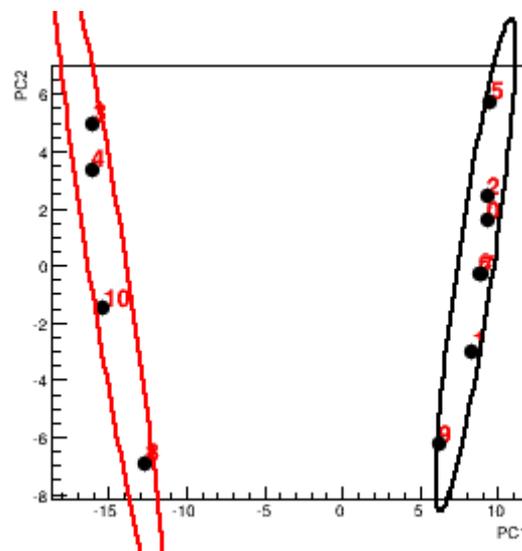


Fig.8 Score plot PCA1-PC2

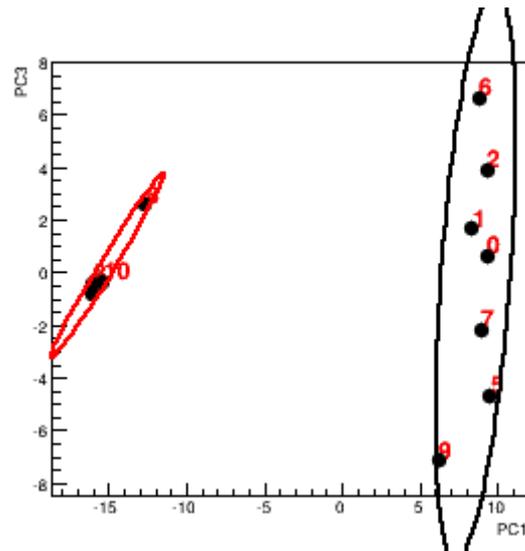


Fig.8 Score plot PCA1-PC3

Unidimensional profiles belonging to diffraction pattern have been superimposed separately for the two groups in Fig.9. It can be clearly seen the difference in the intensity values for the same spots between the two groups.

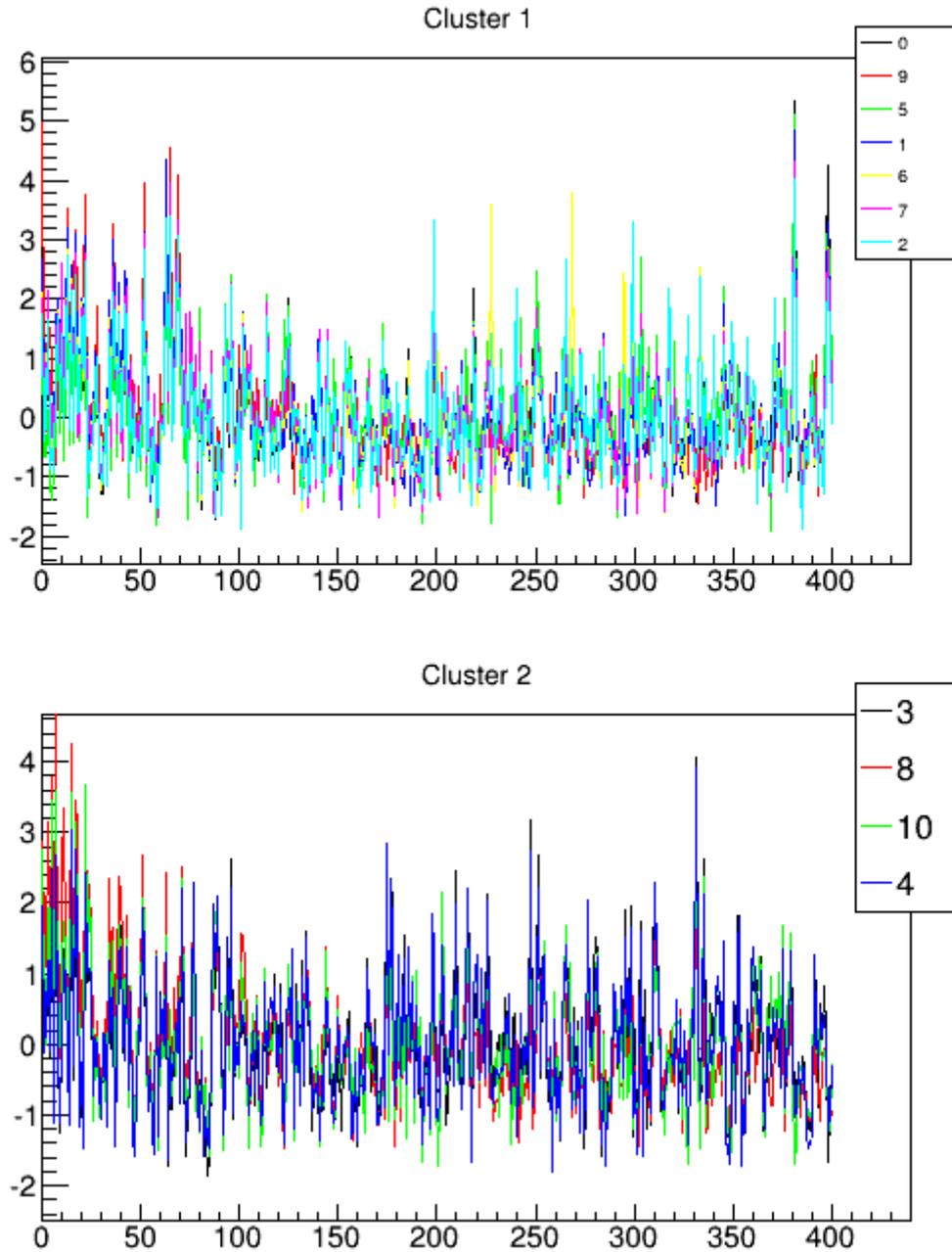


Fig. 15 Profiles in clusters

### Output file

The content of the output file named *outputHKLQualitative* is reported below, with comments added.

```
Input from file: fileInputHKLQualitative
-----
whichanalysis 1

dataType 3

figpaper 1
```



The section above reports the number of data points read within each input file.

#### Starting Qualitative analysis

```
n. points 401
Eigenvalues: 1 --> 72.71% (72.7%)
Eigenvalues: 2 --> 8.86% (81.6%)
Eigenvalues: 3 --> 7.43% (89.0%)
Eigenvalues: 4 --> 3.40% (92.4%)
Eigenvalues: 5 --> 2.19% (94.6%)
Eigenvalues: 6 --> 2.17% (96.8%)
Eigenvalues: 7 --> 1.14% (97.9%)
Eigenvalues: 8 --> 1.07% (99.0%)
Eigenvalues: 9 --> 0.61% (99.6%)
Eigenvalues: 10 --> 0.43% (100.0%)
```

```
Chosen value of k=3: ratio=0.92 error=0.038
```

The section above shows the results of the PCA analysis. The first eigenvalues are listed as a function of their value, and the number of eigenvalues selected for PCA analysis is reported (k), together with the values of the threshold on the cumulative eigenvalue distribution (ratio), and an estimate of the corresponding error between original and reconstructed data (error).

```
===== Dendrogram =====
Step      Dist      Sample 1      Sample 2
  10         24.86         0           3
   9         11.94         0           9
   8          9.23         3           8
   7          5.89         0           5
   6          5.73         3          10
   5          5.08         0           6
   4          4.74         1           7
   3          3.39         0           2
   2          1.64         3           4
=====
Normalized Cluster threshold: 0.200000 (0.721744)
Normalized Cluster threshold redefined: (0.200000) 0.721744
Cluster Threshold 18.401
```

The section above shows the dendrogram resulting from the hierarchical clustering. The value of the threshold distance chosen to define the number of clusters is reported.

#### Cluster analysis

```
Cluster 1 7) 0 9 5 1 6 7 2
Cluster 2 4) 3 8 10 4
Cluster 1 PC0 center=8.61
Cluster 1 PC1 center=0.01
Cluster 1 PC2 center=-0.15
Cluster 2 PC0 center=-15.07
Cluster 2 PC1 center=-0.01
Cluster 2 PC2 center=0.25
```

```
Distances among clusters
```

```
Cluster 1 Cluster 2 --> dist=23.68
```

```
Mahalanobis Distances among clusters
```

```
Cluster 1 Cluster 2 --> dist=20.62 pval=1.18e-07
```

```
Mean pval for Cluster 1 --> pval=1.18e-07
```

```
Mean pval for Cluster 2 --> pval=1.18e-07
```

```
Cluster: 1
```

```
Member: 1 Number: 0 File: xtal1_1.hkl
```

```
Member: 2 Number: 9 File: xtal14.hkl
```

```
Member: 3 Number: 5 File: xtal10_2.hkl
```

```
Member: 4 Number: 1 File: xtal1_2.hkl
```

```
Member: 5 Number: 6 File: xtal7.hkl
```

```
Member: 6 Number: 7 File: xtal11.hkl
```

```
Member: 7 Number: 2 File: xtal3_2.hkl
```

```
Cluster: 2
```

```
Member: 1 Number: 3 File: xtal9.hkl
```

```
Member: 2 Number: 8 File: xtal12.hkl
```

```
Member: 3 Number: 10 File: xtal3_1.hkl
```

```
Member: 4 Number: 4 File: xtal10_1.hkl
```

```
Cluster 1: Representative spectrum: 0
```

```
Cluster 2: Representative spectrum: 10
```

```
Cluster 1: Cluster population: 7 Representative spectrum: 0
```

```
Cluster 2: Cluster population: 4 Representative spectrum: 10
```

```
Cluster 1 Radius (8.94, 1.14)
```

```
Cluster 2 Radius (12.49, 1.19)
```

```
Cluster 1 Radius (10.87, 2.22)
```

```
Cluster 2 Radius (5.02, 0.40)
```

```
Cluster 1 Radius (11.11, 8.22)
```

```
Cluster 2 Radius (12.40, 1.63)
```

The section above analyzes the formed clusters. The content of each cluster in terms of samples and file names, its center and Euclidean distance calculated in the PCA space, and the representative profiles of each cluster, corresponding to those nearest to its center, are listed. The cluster radius is calculated by using the Mahalanobis distance, and it is used to draw the 95% confidence ellipse.