RootProf

TUTORIAL 1

Qualitative analysis

Contents

Chapter 1

The data set

Unidimensional patterns from X-ray diffraction measurements on polycrystalline mixtures form our dataset. Experimental samples have been produced by crystallization processes aiming at obtaining co-crystals formed by an active pharmaceutical ingredient (API) and a co-former. In our case polymorph III of carbamazepine constitutes the API, saccharine the co-former. The experimentally determined weight fractions of carbamazepine (CBZ III), saccharine (SAC) and co-crystal (CBZ-SAC) are reported in Table 1. The corresponding files are included as demo files. They are formed by two columns, the first containing the 2θ values, the second the corresponding values of diffracted intensity.

Table 1: Weight fractions of prepared mixtures. Samples 6-8 (shadowed) are composed by pure phases.

Sample n.	CBZ III	SAC	CBZ-SAC	File name
0	0	0.565	0.435	Rocco_S3_mac.txt
1	0.500	0.500	0	Rocco_S5_mac.txt
2	0.500	0	0.500	Rocco_S7_Como.txt
3	0.347	0.334	0.319	Rocco_S11_mac.txt
4	0.263	0.482	0.255	Rocco_S21_mac.txt
5	0.238	0.364	0.399	Rocco_S22_mac.txt
6	1	0	0	Rocco_CBZ_III_nomac.txt
7	0	1	0	Rocco_SAC_pura_nomac.txt
8	0	0	1	Rocco_CBZSAC_90511_n.txt

Chapter 2

First sight analysis

Motivation

Obtaining a quick and joint view of all input profiles, comparing and inspecting their features, and testing the effect of pre-processing.

The command file

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

```
whichanalysis 0
figpaper 1
dataType 2
range 10 50
preprocess 0 2 100
file  Rocco_S3_mac.txt
file  Rocco_S5_mac.txt
file  Rocco_S7_Como.txt
file  Rocco_S11_mac.txt
file  Rocco_S21_mac.txt
file  Rocco_S22_mac.txt
file  Rocco_CBZ_III_nomac.txt
file  Rocco_SAC_pura_nomac.txt
file  Rocco_CBZSAC_90511_n.txt
```

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

Running RootProf

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

```
Root> .x RootProf_v15.C(“fileInputFirstSight”)
```

After some seconds, graphical windows will start appearing on your screen, while text output will appear on the text window of ROOT.

If you want to store the text output on an external file, just replace the previous command with the following ones:

```
Root> .> outputFirstSight
.x RootProf_v15.C(“fileInputFirstSight”)
.>
```

Then the text output will be redirected in the file named *outputFirstSight*. 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

Input profiles are plotted shifted (Fig.1) and as a data matrix (Fig.2) as they are read. The same plots are repeated after application of pre-processing (Figs 3 and 4, respectively).

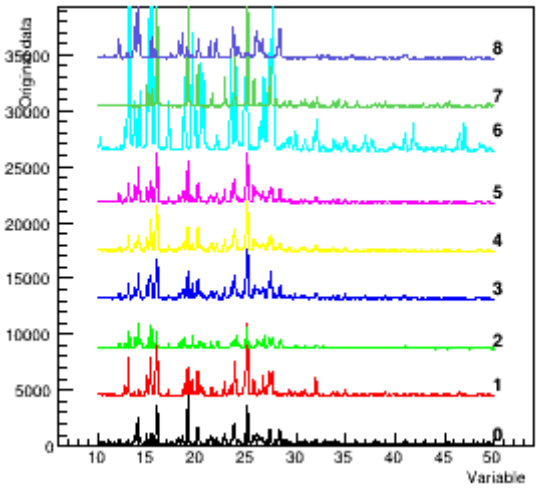

Fig. 1 Original data shifted (before pre-processing)

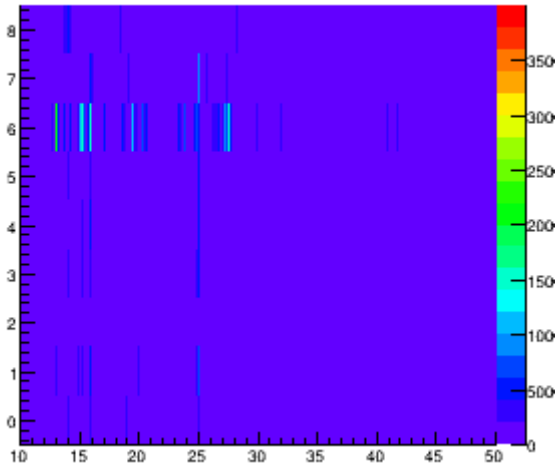

Fig. 2 Data Matrix (before pre-processing)

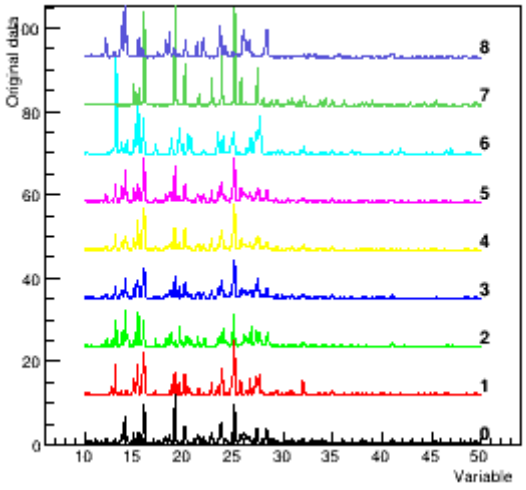

Fig. 3 Original data shifted after pre-processing)

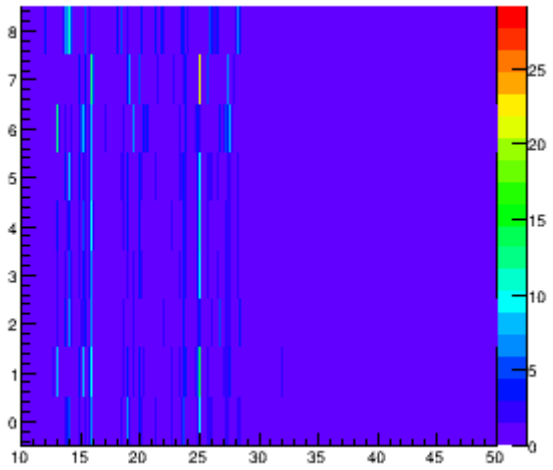

Fig. 4 Data Matrix (after pre-processing)

Output file

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

```
 Input from file: fileInputFirstSight
-------------------------------
  whichanalysis 0

  figpaper 1

  dataType 2

  range 10 50

  preprocess 0 2 100
```

```
  file  Rocco_S3_mac.txt

  file  Rocco_S5_mac.txt

  file  Rocco_S7_Como.txt

  file  Rocco_S11_mac.txt

  file  Rocco_S21_mac.txt

  file  Rocco_S22_mac.txt

  file  Rocco_CBZ_III_nomac.txt

  file  Rocco_SAC_pura_nomac.txt

  file  Rocco_CBZSAC_90511_n.txt
```

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

```
Reading input files:
 ----------------------
Sample 0 -> file Rocco_S3_mac.txt
         Found 1999  points
Sample 1 -> file Rocco_S5_mac.txt
         Found 1999  points
Sample 2 -> file Rocco_S7_Como.txt
         Found 1999  points
Sample 3 -> file Rocco_S11_mac.txt
         Found 1999  points
Sample 4 -> file Rocco_S21_mac.txt
         Found 1999  points
Sample 5 -> file Rocco_S22_mac.txt
         Found 1999  points
Sample 6 -> file Rocco_CBZ_III_nomac.txt
         Found 1999  points
Sample 7 -> file Rocco_SAC_pura_nomac.txt
         Found 1999  points
Sample 8 -> file Rocco_CBZSAC_90511_n.txt
         Found 1999  points
```

The section above reports the number of data points read within each input file, as determined by the command *range*.

Chapter 3

Principal component analysis

Motivation

This analysis allows classifying input profiles, by using a common multivariate analysis method which identifies directions of maximum variability in data.

The command file

The list of commands is the following.

```
whichanalysis 1
figpaper 1
dataType 2
range 10 50
preprocess 0 2 100
skipdata 3
file  Rocco_S3_mac.txt
file  Rocco_S5_mac.txt
file  Rocco_S7_Como.txt
file  Rocco_S11_mac.txt
file  Rocco_S21_mac.txt
file  Rocco_S22_mac.txt
file  Rocco_CBZ_III_nomac.txt
file  Rocco_SAC_pura_nomac.txt
file  Rocco_CBZSAC_90511_n.txt
```

They have been included in the demo file named *fileInputQualitative*. See the user guide for an explanation of each command. The command *skipdata* is optional and has been added uniquely to speed up the analysis.

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_v15.C(“fileInputQualitative”)
```

or

```
Root> .> outputQualitative
.x RootProf_v15.C(“fileInputQualitative”)
.>
```

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 *outputQualitative*. 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

The first graphic windows produced give different representation of the input profiles. They are shown in the figures below, with captions indicating the title of each graphic window, which can be read on the screen. The profiles are all shown after application of pre-processing, which in this case is done by the command *preprocess 0 2 100*, which means normalizing to unity the area of each diffraction pattern and subtracting the background estimated by the SNIP algorithm with a clipping window of 100 channels (see user guide). The profiles are shown superimposed, with different colors, in the graphic window named “Original data” (Fig.1), and vertically shifted, as a function of the 2θ variable in the graphic window named “Original data shifted” (Fig.2).

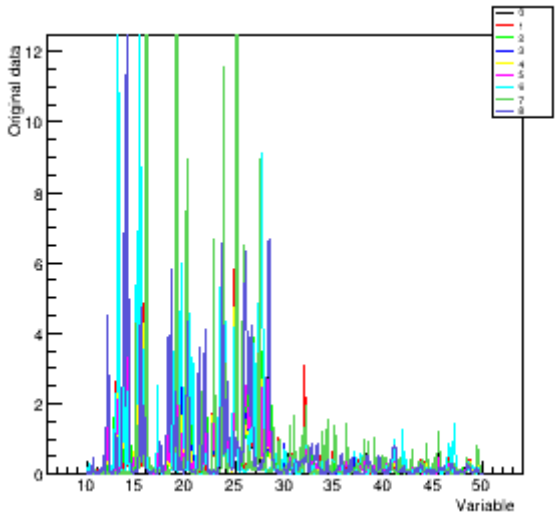

Fig.1 Original data

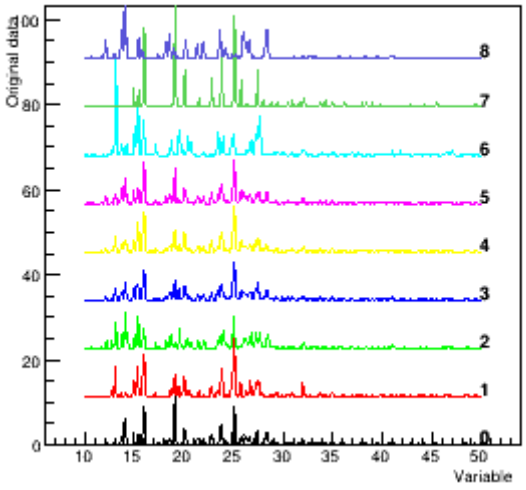

Fig.2 Original data shifted

A different representation is given in the graphic window “Data Matrix” (Fig.3), where the 2θ values are reported along rows and the intensities of the different samples are reported along columns, numbered as in Table 1. A color bar no the right indicates the colors used to represent the intensity values.

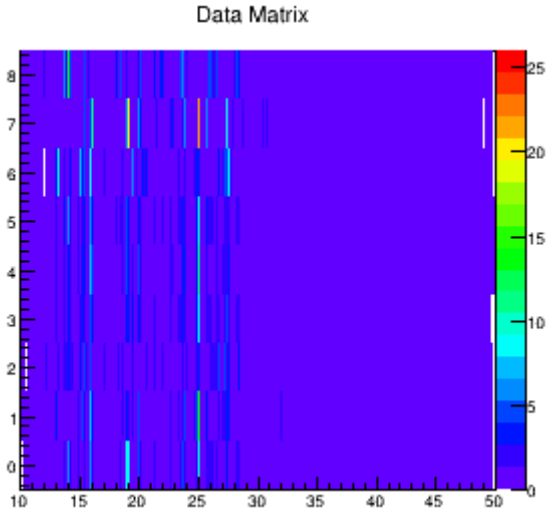

Fig.3 Data Matrix

Eigenvectors and Eigenvalues

A PCA analysis is performed on the data matrix given in input. A full description of the method can be found in the documentation of the TPrincipal class of ROOT, available at root.cern.ch. The

normalized eigenvalues calculated for the covariance matrix are reported in Fig.4, with a vertical dashed line indicating the threshold chosen for dimensionality reduction. It has been set by the command *threshold 0.7*, which operates on the cumulative distribution of the eigenvalues (blue curve in Fig.1) . The cumulative value nearest to 0.7 is 0.84, representing the sum of the first two eivenvalues. Thus the original space of 666 2θ values is reduced to only two variables: the first two principal components PCA 1 and PCA 2.

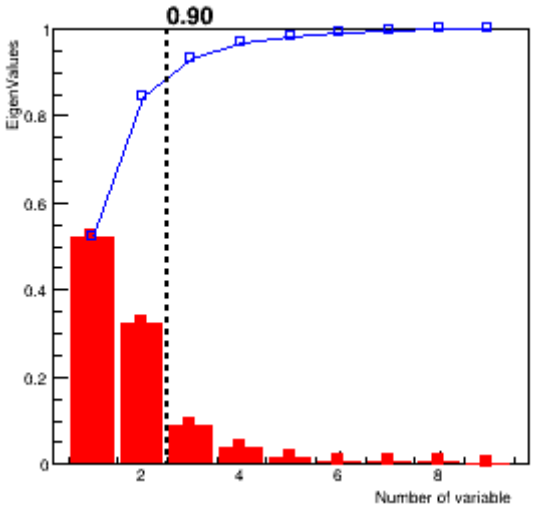

Fig.4 Scree plot

The scores and the loadings for the selected principal components are shown in Fig.5 and Fig.6, respectively. Scores represent the contribution of each sample to the principal components, while loadings indicate the contribution of each 2θ value.

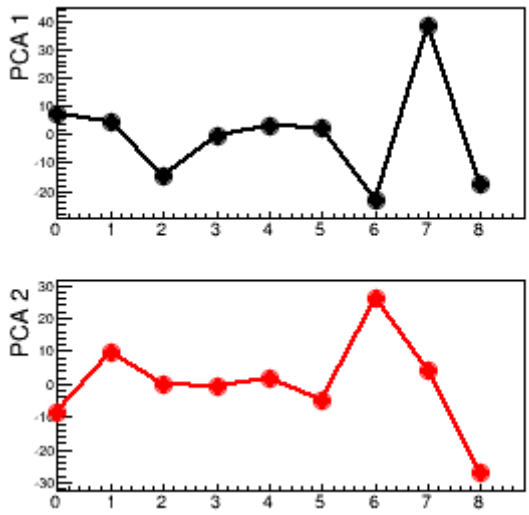

Fig.5 Scores

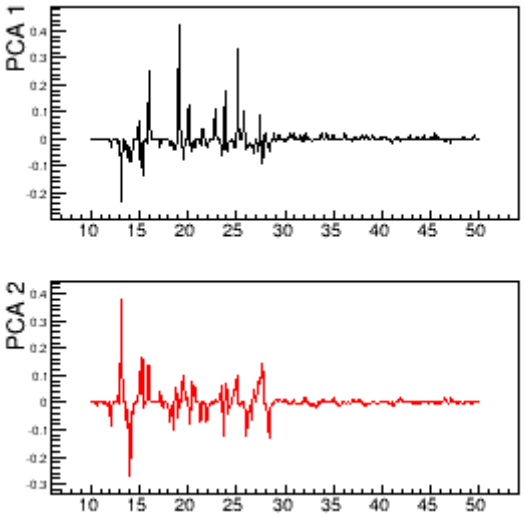

Fig.6 Loadings

Profiles reconstructed by using only the selected two principal components are shown superimposed in Fig.7, which should be compared with Fig.1. They are also shown shifted in Fig.8, to be compared with Fig.2. Most of the features of the original data set are reproduced by using only two variables (PCA 1 and PCA 2) out of the original 666 2θ values! The following PCA analysis is based on the profiles shown in Fig.7 and Fig.8.

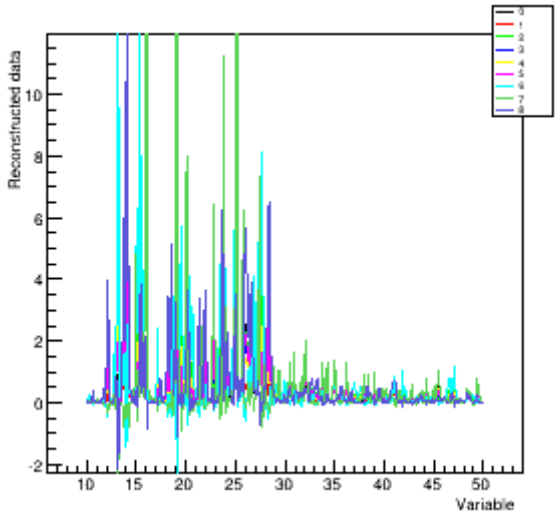

Fig.7 Reconstructed data

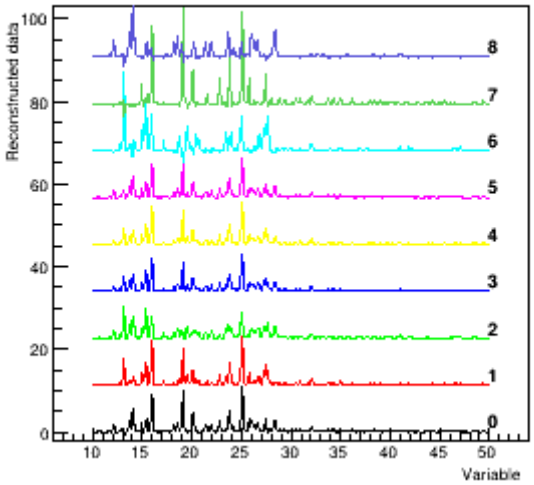

Fig.8 Reconstructed data shifted

PCA representation

Data can be projected in the space of the PCA latent variables by using the score plots. Fig.9 shows the sole score plot that can be formed by only two principal components: the scatter plot of the PCA2 scores versus the PCA1 ones. The sample number is put in red near each representative point. It can be noted that the score plot reveals a ternary diagram trend, with pure phases 6, 7, 8 at the edges and the mixture points within the triangle. The mixtures 0, 1, 2, containing binary mixtures, are roughly placed at the sides of the triangle, while mixtures 3, 4, 5, containing nearly equal weight fractions of the pure phases, are at the center of the triangle. A 95% confidence level ellipse is shown, enclosing data point grouped by the hierarchical clustering algorithm (See next paragraph). This is an example how looking at data in the latent variable space can help in classifying samples.

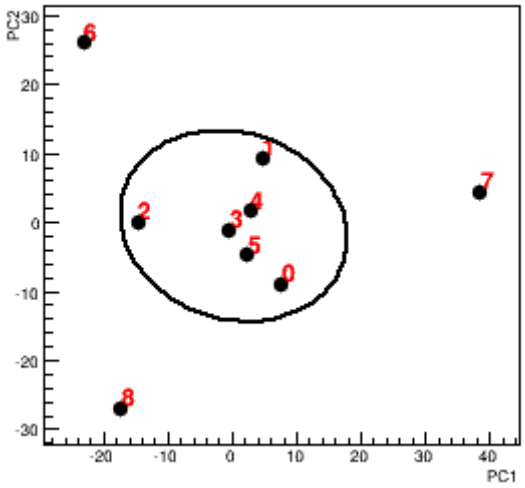

Fig.9 Score plot PCA1-PC2

The loadings plot in Fig.10 is the scatter plot of the PCA2 loadings versus the PCA1 ones. Here each point represent a 2θ value (the 2θ values are indicated by blue numbers). It can be used to reckon the role of each peak of the powder pattern to the overall sample classification. For example, the point (0.4, 0) in Fig.9 is responsible for the differentiation of sample 7 (SAC) from the others. By zooming the window named "Loading plot PC1-PC2 on the screen, it can be seen that this point correspond to a 2θ value around 19.1°. By checking on the loading plots (Fig. 6) it can be inferred that the higher peak of the SAC spectrum occurs at 2θ=19.1°. Thus PCA1 mainly differentiates SAC from CBZ and CBZ-SAC, thanks to the huge contribution from such peak.

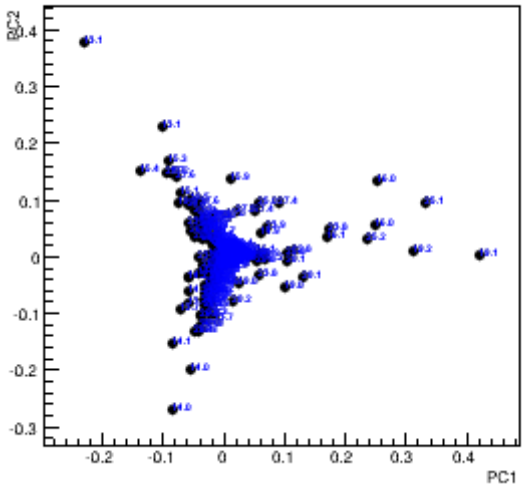

Fig.10 Loading plot PC1-PC2

Hierarchic clustering in the PCA space

Data represented in the PCA latent variable space are arranged so that to optimize their classification. Therefore it is appropriate grouping them in such a space. RootProf adopts a hierarchical clustering algorithm, based on the nearest neighbouring metrics calculated in the PCA space. The matrix of distances among samples is plotted in Fig. 11. It represents the starting point for the clustering algorithm. As output, a dendrogram is produced, which can be read on the output file. In addition, the distance matrix after clusterization is plotted in Fig. 12, where numbers on X and Y axes do not represent sample numbers anymore, but number of sample in clusters. For example, in the output file it can be seen that samples are arranged in clusters as following:

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

the first cluster is formed by samples. Than the first 6 elements in the matrix of Fig.11 are samples 0,2,1,3,5,4, forming the first cluster. The remaining elements are samples 8,6,7, forming three separate clusters.

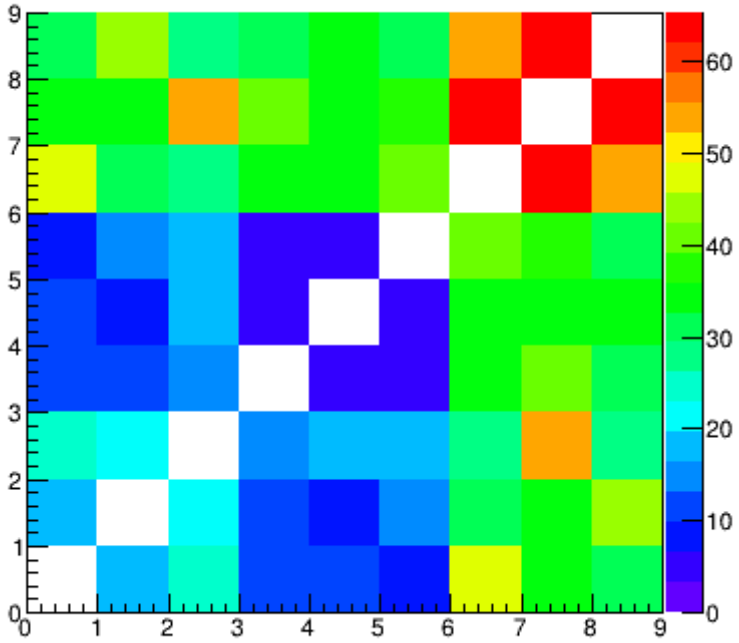

Fig.11 Matrix of distances

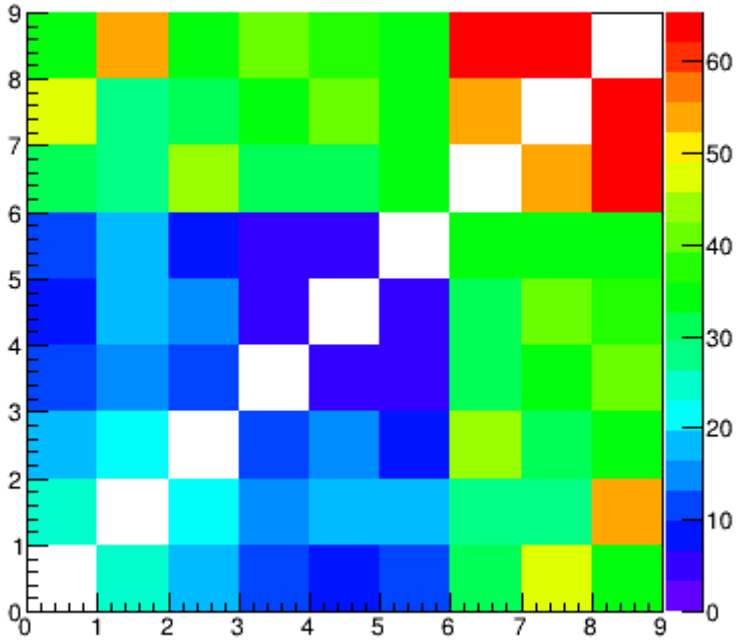

Fig.12 Matrix of distances after clustering

Fig.13 shows the cluster size distribution, i.e. the distribution of the number of clusters (X-axis) versus the nearest neighbouring distance (Y-axis). The dashed line indicates the threshold distance: that used to define the actual number of clusters. It is determined automatically, by considering the derivative cluster size distribution (Fig.14). In fact the threshold is settled where the cluster size distribution have a discontinuity, indicating a data driven separation among clusters. The cluster size distribution as a function of the normalized nearest neighbouring distance is also given (Fig.15). Note that plots in Figs. 13 and 14 are only produced if the *verbose* command is given in input, with value greater than 1. The user can override the automatic threshold determination by giving in the input file the value of the normalized distance threshold, through the command *sogdiff*.

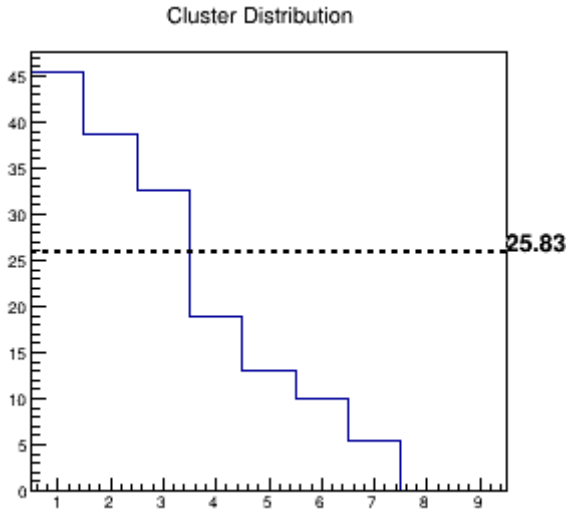

Fig.13 Cluster size distribution

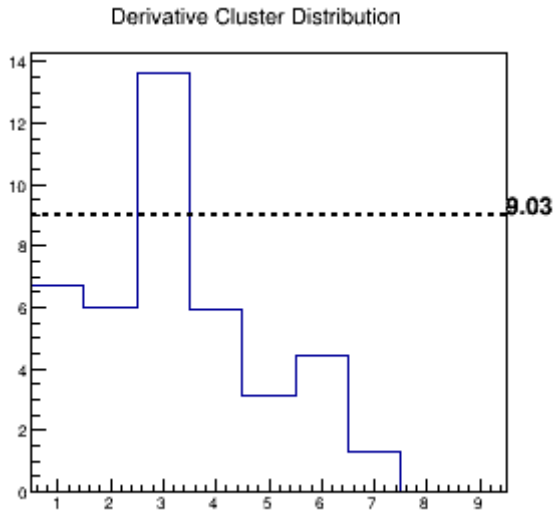

Fig.14 Derivative cluster size distribution

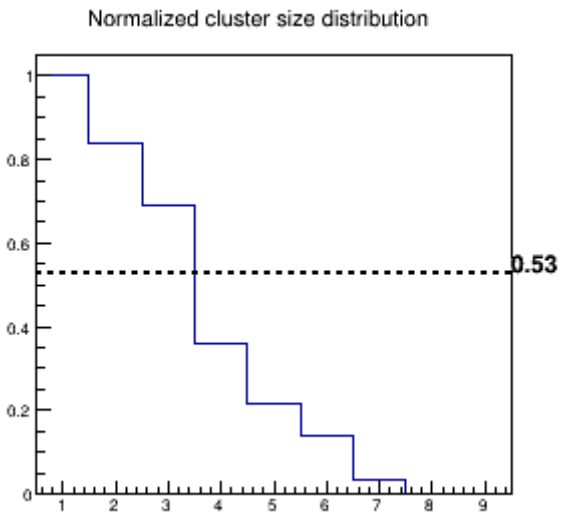

Fig.15 Normalized cluster size distribution

Input profiles grouped in clusters are shown in Fig.16, where profiles belonging to the same cluster are superimposed with different colors.

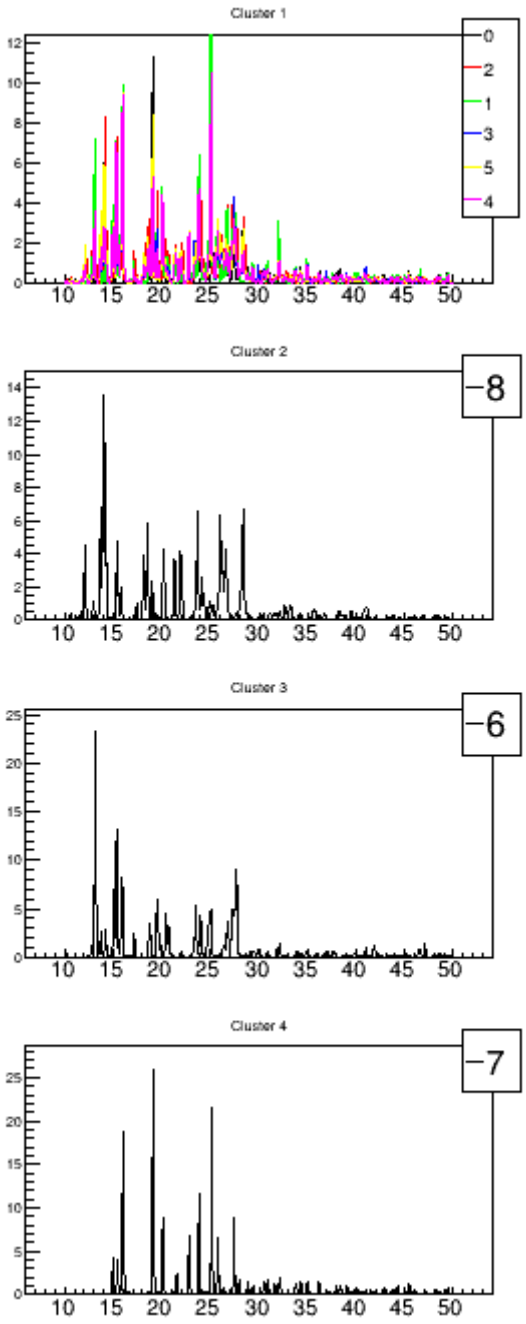

Fig. 16 Profiles in clusters

Output file

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

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

  figpaper 1

  dataType 2

  range 10 50
```

```
  threshold 0.7

  preprocess 0 2 100

  skipdata 3

  file  Rocco_S3_mac.txt

  file  Rocco_S5_mac.txt

  file  Rocco_S7_Como.txt

  file  Rocco_S11_mac.txt

  file  Rocco_S21_mac.txt

  file  Rocco_S22_mac.txt

  file  Rocco_CBZ_III_nomac.txt

  file  Rocco_SAC_pura_nomac.txt

  file  Rocco_CBZSAC_90511_n.txt
```

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

```
Reading input files:
 ----------------------
Sample 0 -> file Rocco_S3_mac.txt
         Found 666  points
Sample 1 -> file Rocco_S5_mac.txt
         Found 666  points
Sample 2 -> file Rocco_S7_Como.txt
         Found 666  points
Sample 3 -> file Rocco_S11_mac.txt
         Found 666  points
Sample 4 -> file Rocco_S21_mac.txt
         Found 666  points
Sample 5 -> file Rocco_S22_mac.txt
         Found 666  points
Sample 6 -> file Rocco_CBZ_III_nomac.txt
         Found 666  points
Sample 7 -> file Rocco_SAC_pura_nomac.txt
         Found 666  points
Sample 8 -> file Rocco_CBZSAC_90511_n.txt
         Found 666  points
```

The section above reports the number of data points read within each input file, as determined by the commands *range* and *skipdata*.

```
Starting Qualitative analysis

 n. points 666
Eigenvalues: 1 --> 52.20% (52.2%)
Eigenvalues: 2 --> 32.26% (84.5%)
Eigenvalues: 3 --> 8.80% (93.3%)
```

```
Eigenvalues: 4 --> 3.59% (96.9%)
Eigenvalues: 5 --> 1.43% (98.3%)
Eigenvalues: 6 --> 0.64% (98.9%)
Eigenvalues: 7 --> 0.60% (99.5%)
Eigenvalues: 8 --> 0.48% (100.0%)
Eigenvalues: 9 --> 0.00% (100.0%)

Chosen value of k=2: ratio=0.93 error=0.034
```

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). The threshold value is chosen on the basis of the command *threshold*.

```
 ============ Dendrogram =============
  Step      Dist     Sample 1    Sample 2
   8        45.35       0           7
   7        38.65       0           6
   6        32.63       0           8
   5        19.02       0           2
   4        13.10       0           1
   3        9.97       0           3
   2        5.55       3           5
   1        4.23       3           4
 ====================================
 Normalized Cluster threshold: 0.200000 (0.525116)
 Normalized Cluster threshold redefined: (0.200000) 0.525116
 Cluster Threshold 25.825
```

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 6)  0  2  1  3  5  4
 Cluster 2 1)  8
 Cluster 3 1)  6
 Cluster 4 1)  7
Cluster  1 PC0 center=0.33
Cluster  1 PC1 center=-0.60
Cluster  2 PC0 center=-17.33
Cluster  2 PC1 center=-26.93
Cluster  3 PC0 center=-23.10
Cluster  3 PC1 center=26.12
Cluster  4 PC0 center=38.44
Cluster  4 PC1 center=4.39

 Distances among clusters

 Cluster 1 Cluster 2 --> dist=31.71
 Cluster 1 Cluster 3 --> dist=35.54
 Cluster 1 Cluster 4 --> dist=38.44
 Cluster 2 Cluster 3 --> dist=53.37
 Cluster 2 Cluster 4 --> dist=63.97
 Cluster 3 Cluster 4 --> dist=65.27


 Cluster: 1
```

```
Member: 1  Number: 0  File: Rocco_S3_mac.txt
Member: 2  Number: 2  File: Rocco_S7_Como.txt
Member: 3  Number: 1  File: Rocco_S5_mac.txt
Member: 4  Number: 3  File: Rocco_S11_mac.txt
Member: 5  Number: 5  File: Rocco_S22_mac.txt
Member: 6  Number: 4  File: Rocco_S21_mac.txt

Cluster: 2
Member: 1  Number: 8  File: Rocco_CBZSAC_90511_n.txt

Cluster: 3
Member: 1  Number: 6  File: Rocco_CBZ_III_nomac.txt

Cluster: 4
Member: 1  Number: 7  File: Rocco_SAC_pura_nomac.txt

Cluster 1:  Representative spectrum: 3
Cluster 2:  Representative spectrum: 8
Cluster 3:  Representative spectrum: 6
Cluster 4:  Representative spectrum: 7

Cluster 1: Cluster population: 6  Representative spectrum: 3
Cluster 2: Cluster population: 1  Representative spectrum: 8
Cluster 3: Cluster population: 1  Representative spectrum: 6
Cluster 4: Cluster population: 1  Representative spectrum: 7

Cluster 1  Radius (17.82, 13.56)
```

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.

Chapter 3

Correlation analysis

Motivation

Classifying profiles according to a different method with respect to PCA analysis. Clustering is performed by adopting a metrics based on the Pearson's correlation coefficient, calculated by considering the corresponding intensities of pairs of profiles.

The command file

The list of commands is the following.

```
whichanalysis 2
figpaper 1
dataType 2
range 10 50
preprocess 0 2 100
file  Rocco_S3_mac.txt
file  Rocco_S5_mac.txt
file  Rocco_S7_Como.txt
file  Rocco_S11_mac.txt
file  Rocco_S21_mac.txt
file  Rocco_S22_mac.txt
file  Rocco_CBZ_III_nomac.txt
file  Rocco_SAC_pura_nomac.txt
file  Rocco_CBZSAC_90511_n.txt
```

They have been included in the demo file named *fileInputCorrel*. 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_v15.C("fileInputCorrel")

or

Root> .> outputCorrel

.x RootProf_v15.C("fileInputCorrel")

.>

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 *outputCorrel*. 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

Input profiles after application of pre-processing are plotted shifted (Fig.1) and as a data matrix (Fig.2).

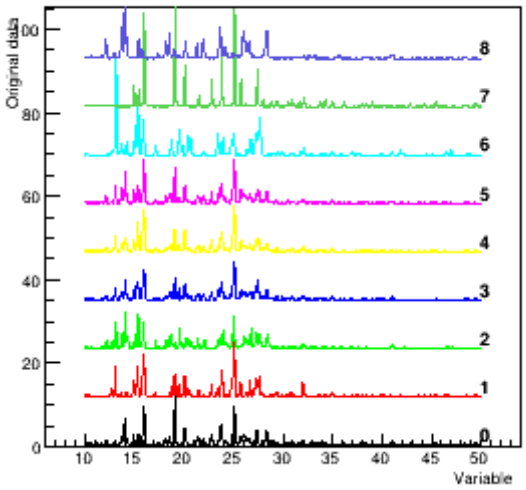

Fig. 1 Original data shifted

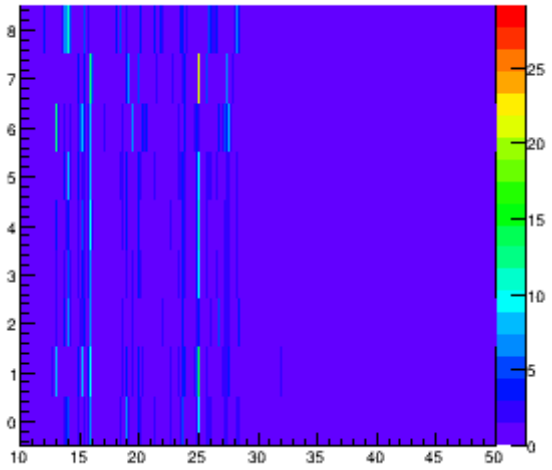

Fig. 2 Data Matrix

The clustering algorithm is applied by using the Pearson's correlation coefficient of their intensities as distance among profiles. The distance matrix so obtained is plotted in Fig. 3, while in Fig.4 is reported the same matrix after application of the clustering algorithm. The corresponding cluster size distribution is reported in Fig.5, and the profiles grouped in cluster are reported in Fig.6. By comparing Fig.6 with Fig.16 of chapter 2 one can note that different distance matrices (PCA or correlation coefficient) produce different clustering of the same profiles.

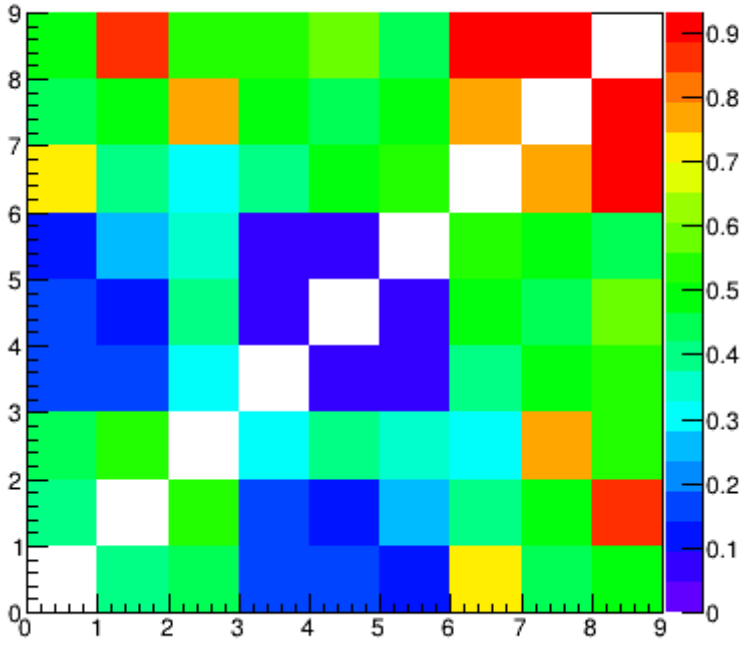

Fig. 3 Matrix of distances

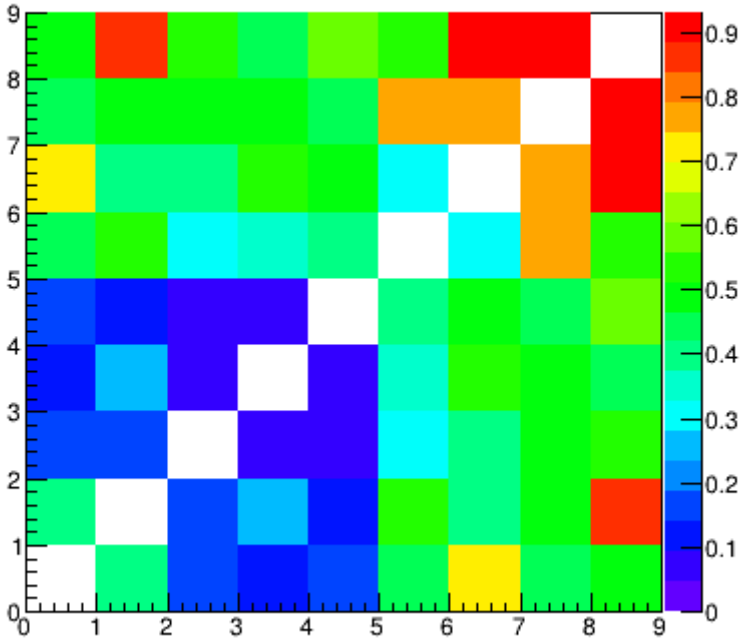

Fig. 4 Matrix of distances after clustering

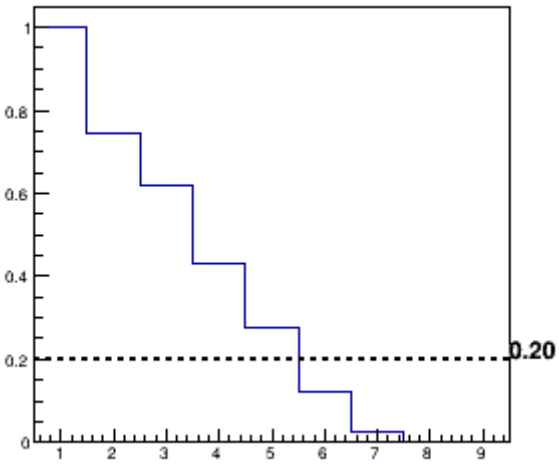

Fig. 5 Normalized cluster size distribution.

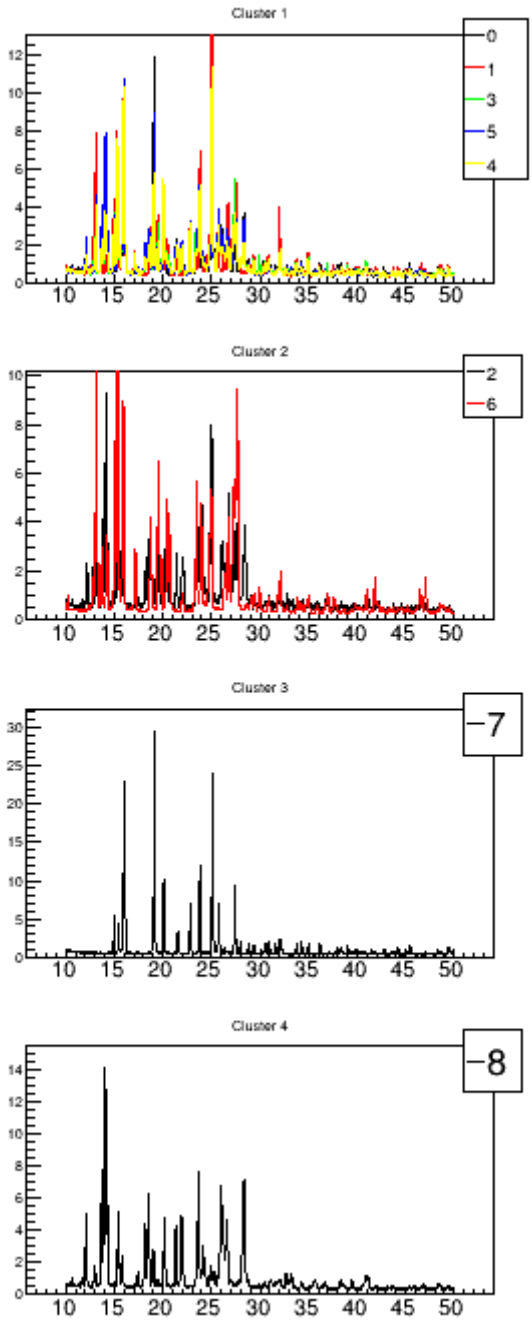

Fig. 6 Profiles in clusters

Output file

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

```
 Input from file: fileInputCorrel
-------------------------------
  whichanalysis 2

  figpaper 1

  dataType 2

  range 10 50
```

```
  preprocess 0 2 100

  file  Rocco_S3_mac.txt

  file  Rocco_S5_mac.txt

  file  Rocco_S7_Como.txt

  file  Rocco_S11_mac.txt

  file  Rocco_S21_mac.txt

  file  Rocco_S22_mac.txt

  file  Rocco_CBZ_III_nomac.txt

  file  Rocco_SAC_pura_nomac.txt

  file  Rocco_CBZSAC_90511_n.txt
```

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

```
Reading input files:
 ----------------------
Sample 0 -> file Rocco_S3_mac.txt
         Found 1999  points
Sample 1 -> file Rocco_S5_mac.txt
         Found 1999  points
Sample 2 -> file Rocco_S7_Como.txt
         Found 1999  points
Sample 3 -> file Rocco_S11_mac.txt
         Found 1999  points
Sample 4 -> file Rocco_S21_mac.txt
         Found 1999  points
Sample 5 -> file Rocco_S22_mac.txt
         Found 1999  points
Sample 6 -> file Rocco_CBZ_III_nomac.txt
         Found 1999  points
Sample 7 -> file Rocco_SAC_pura_nomac.txt
         Found 1999  points
Sample 8 -> file Rocco_CBZSAC_90511_n.txt
         Found 1999  points
```

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

```
 ============ Dendrogram =============
  Step       Dist      Sample 1    Sample 2
   8         0.70        0           8
   7         0.54        0           2
   6         0.46        0           7
   5         0.34        2           6
   4         0.24        0           1
   3         0.14        0           3
   2         0.07        3           5
   1         0.06        3           4
 ====================================
```

```
Normalized Cluster threshold: 0.200000 (0.872177)
Cluster Threshold 0.188

Cluster analysis

 Cluster 1 4)  0  3  5  4
Cluster 2 1)  1
Cluster 3 1)  2
Cluster 4 1)  6
Cluster 5 1)  7
Cluster 6 1)  8


Cluster: 1
Member: 1  Number: 0  File: Rocco_S3_mac.txt
Member: 2  Number: 3  File: Rocco_S11_mac.txt
Member: 3  Number: 5  File: Rocco_S22_mac.txt
Member: 4  Number: 4  File: Rocco_S21_mac.txt

Cluster: 2
Member: 1  Number: 1  File: Rocco_S5_mac.txt

Cluster: 3
Member: 1  Number: 2  File: Rocco_S7_Como.txt

Cluster: 4
Member: 1  Number: 6  File: Rocco_CBZ_III_nomac.txt

Cluster: 5
Member: 1  Number: 7  File: Rocco_SAC_pura_nomac.txt

Cluster: 6
Member: 1  Number: 8  File: Rocco_CBZSAC_90511_n.txt


Cluster: 1
Member: 1  Number: 0  File: Rocco_S3_mac.txt
Member: 2  Number: 3  File: Rocco_S11_mac.txt
Member: 3  Number: 5  File: Rocco_S22_mac.txt
Member: 4  Number: 4  File: Rocco_S21_mac.txt

Cluster: 2
Member: 1  Number: 1  File: Rocco_S5_mac.txt

Cluster: 3
Member: 1  Number: 2  File: Rocco_S7_Como.txt

Cluster: 4
Member: 1  Number: 6  File: Rocco_CBZ_III_nomac.txt

Cluster: 5
Member: 1  Number: 7  File: Rocco_SAC_pura_nomac.txt

Cluster: 6
Member: 1  Number: 8  File: Rocco_CBZSAC_90511_n.txt

Cluster 1:  Representative spectrum: 0
Cluster 2:  Representative spectrum: 1
Cluster 3:  Representative spectrum: 2
Cluster 4:  Representative spectrum: 6
```

```
Cluster 5:  Representative spectrum: 7
Cluster 6:  Representative spectrum: 8
```

```
Cluster population: 4
Cluster population: 1
Cluster population: 1
Cluster population: 1
Cluster population: 1
Cluster population: 1
```

The section above shows the results of the clustering analysis applied by using the metrix defined by the correlation coefficient. The distance is taken as 1-corr, where corr is the Pearson's correlation coefficient between a pair of profiles. The dendrogram resulting from the hierarchical clustering is shown. The values of the threshold distance chosen to define the number of clusters are reported.

Chapter 4

Testing user-defined classification

Description

In case the classification of the samples is known in advance, the user can use the program to test it on input samples.

The command file

The list of commands is the following.

```
whichanalysis 1
figpaper 1
dataType 2
range 10 50
preprocess 0 2 100
skipdata 3
clusterswitch 2
file  Rocco_S3_mac.txt
myclust 1
file  Rocco_S5_mac.txt
myclust 3
file  Rocco_S7_Como.txt
myclust 3
file  Rocco_S11_mac.txt
myclust 2
file  Rocco_S21_mac.txt
myclust 2
file  Rocco_S22_mac.txt
myclust 2
file  Rocco_CBZ_III_nomac.txt
myclust 3
file  Rocco_SAC_pura_nomac.txt
myclust 1
file  Rocco_CBZSAC_90511_n.txt
myclust 1
```

They have been included in the demo file named *fileInputUserClustering*. 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_v15.C(“fileInputUserClustering”)
```

or

```
Root> .> outputUserClustering
.x RootProf_v15.C("fileInputUserClustering")
.>
```

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 *outputUserClustering*. 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.

Score plot with ellipses

The PCA analysis is performed on input profiles, and data points in the PCA space are projected in the score plots. The subsequent clustering of data points is inhibited, but the user-defined classification is taken instead. The score plot in the first two principal components is plotted in Fig.1, where data points are colored according to the user-defined classification: black for the cluster n.1, comprising points 0,7,8, red for the cluster n.2, comprising points 3,4,5 and green for the cluster n.3, comprising points 1,2,6. These assignments are given in input to each file through the command *myclust*. The separation among data points in the score plot is quantified by drawing 95% confidence ellipses. They are calculated from the probability distribution of the Mahalanobis distance among data points, and define the statistical significance of class separation.

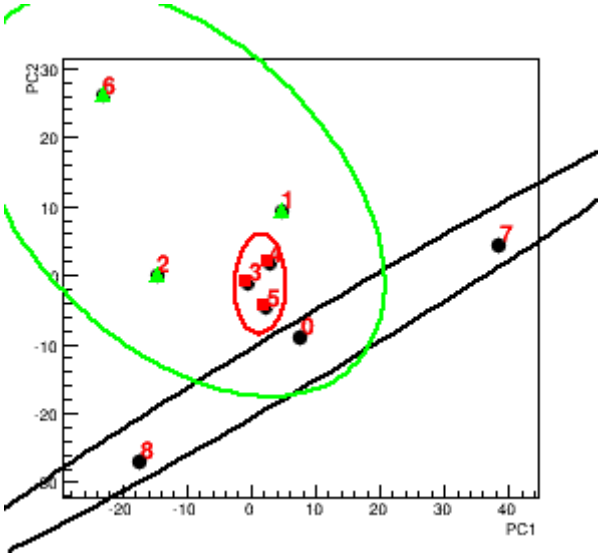

Fig.1 Score plot PC1-PC2

In addition, *p* values are calculated for every pair of user-defined clusters in the PCA space, by using the Mahalanobis distance. They represent the probability for accepting the null hypothesis

that the two clusters are drawn from the same multivariate normal distribution. The p-values are shown in a matrix, rescaled between 0 and 1, to highlight their difference (Fig.2), and written in the output file. In our case, for example, the p values are the following:

```
Mahalanobis Distances among clusters

 Cluster 1 Cluster 2 --> dist=4.79 pval=3.36e-02
 Cluster 1 Cluster 3 --> dist=2.61 pval=1.49e-01
 Cluster 2 Cluster 3 --> dist=1.55 pval=3.81e-01

 Mean pval for Cluster 1 --> pval=9.11e-02
 Mean pval for Cluster 2 --> pval=2.07e-01
 Mean pval for Cluster 3 --> pval=2.65e-01
```

Therefore, the lowest p-value is between cluster 1 and 2: their ellipses are disjointed, therefore the probability that they are drawn from the same normal distribution is low. Clusters 1 and 3 have higher p-values, since their ellipses have a non-null intersection, while clusters 2 and 3 have the highest p values, since their ellipses are completely intersected. An average p-value is assigned to each cluster, by considering those of all the pairs in which it is involved. Lower this value, more separated is this cluster from all the others.

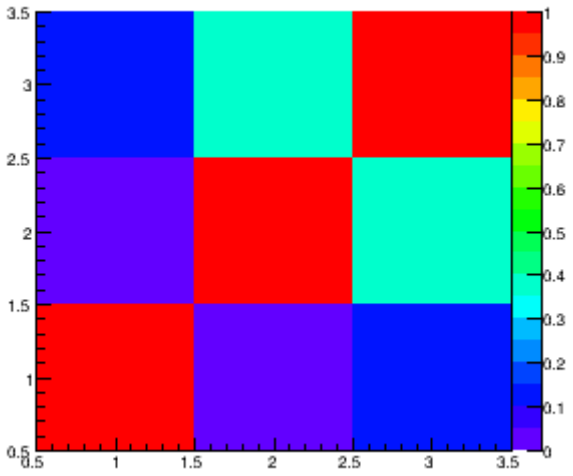

Fig.2 Matrix of P-values

Profiles classified according to the user-defined criterion are plotted in Fig.3.

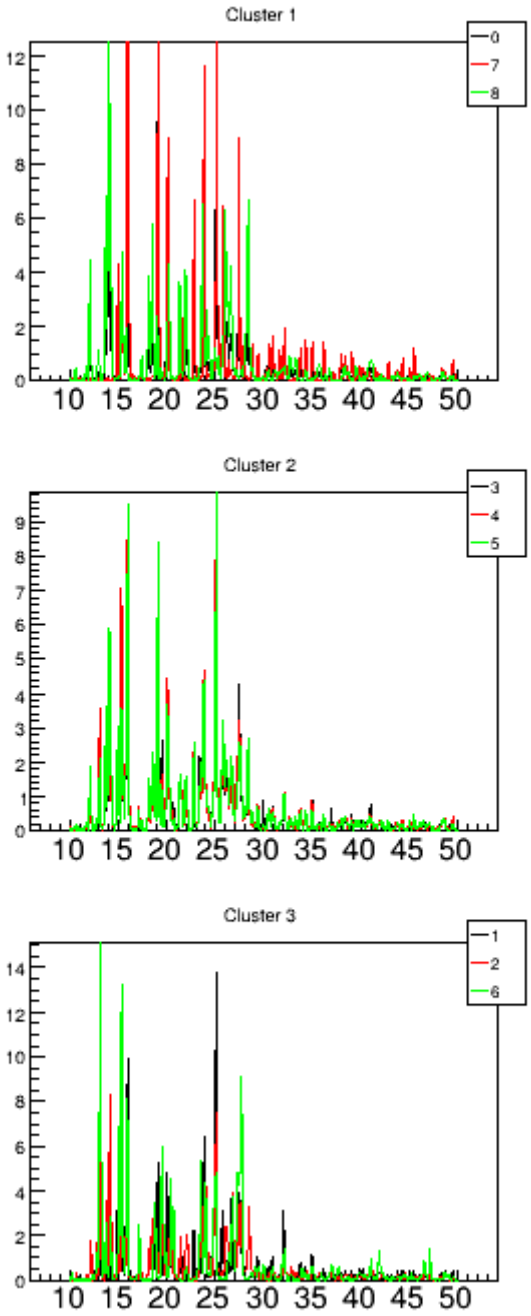

Fig.3 Profiles in clusters

Output file

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

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

  figpaper 1

  dataType 2
```

```
  range 10 50
  preprocess 0 2 100
  skipdata 3
  clusterswitch 2
  file  Rocco_S3_mac.txt
  myclust 1
  file  Rocco_S5_mac.txt
  myclust 3
  file  Rocco_S7_Como.txt
  myclust 3
  file  Rocco_S11_mac.txt
  myclust 2
  file  Rocco_S21_mac.txt
  myclust 2
  file  Rocco_S22_mac.txt
  myclust 2
  file  Rocco_CBZ_III_nomac.txt
  myclust 3
  file  Rocco_SAC_pura_nomac.txt
  myclust 1
  file  Rocco_CBZSAC_90511_n.txt
  myclust 1
```

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

```
  Reading input files:
 ----------------------
Sample 0 -> file Rocco_S3_mac.txt
         Found 666  points
Sample 1 -> file Rocco_S5_mac.txt
         Found 666  points
Sample 2 -> file Rocco_S7_Como.txt
         Found 666  points
Sample 3 -> file Rocco_S11_mac.txt
         Found 666  points
Sample 4 -> file Rocco_S21_mac.txt
```

```
         Found 666  points
Sample 5 -> file Rocco_S22_mac.txt
         Found 666  points
Sample 6 -> file Rocco_CBZ_III_nomac.txt
         Found 666  points
Sample 7 -> file Rocco_SAC_pura_nomac.txt
         Found 666  points
Sample 8 -> file Rocco_CBZSAC_90511_n.txt
         Found 666  points
```

The section above reports the number of data points read within each input file, and their initial and final number. They correspond to the range of the profiles variable chosen by the command *range*.

```
Starting Qualitative analysis

 n. points 666
Eigenvalues: 1 --> 52.20% (52.2%)
Eigenvalues: 2 --> 32.26% (84.5%)
Eigenvalues: 3 --> 8.80% (93.3%)
Eigenvalues: 4 --> 3.59% (96.9%)
Eigenvalues: 5 --> 1.43% (98.3%)
Eigenvalues: 6 --> 0.64% (98.9%)
Eigenvalues: 7 --> 0.60% (99.5%)
Eigenvalues: 8 --> 0.48% (100.0%)
Eigenvalues: 9 --> 0.00% (100.0%)

 Chosen value of k=2: ratio=0.93 error=0.034
```

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). The threshold value is chosen on the basis of the command *threshold*.

```
Cluster  1 PC0 center=9.59
Cluster  1 PC1 center=-10.49
Cluster  2 PC0 center=1.47
Cluster  2 PC1 center=-1.37
Cluster  3 PC0 center=-11.06
Cluster  3 PC1 center=11.87

 Distances among clusters

 Cluster 1 Cluster 2 --> dist=12.21
 Cluster 1 Cluster 3 --> dist=30.43
 Cluster 2 Cluster 3 --> dist=18.23
```

The section above analyze the user-defined clusters. The center of each cluster and the Euclidean distance among clusters are calculated in the PCA space.

```
 Mahalanobis Distances among clusters

 Cluster 1 Cluster 2 --> dist=4.79 pval=3.36e-02
 Cluster 1 Cluster 3 --> dist=2.61 pval=1.49e-01
 Cluster 2 Cluster 3 --> dist=1.55 pval=3.81e-01

 Mean pval for Cluster 1 --> pval=9.11e-02
```

```
Mean pval for Cluster 2 --> pval=2.07e-01
Mean pval for Cluster 3 --> pval=2.65e-01
```

The section above reports the Mahalanobis distance in the PCA space, and the calculated p values for each pair of clusters. An average p-value for each cluster is also calculated, by considering the p values of all the pairs in which the cluster is included.

```
Cluster: 1
Member: 1  Number: 0  File: Rocco_S3_mac.txt
Member: 2  Number: 7  File: Rocco_SAC_pura_nomac.txt
Member: 3  Number: 8  File: Rocco_CBZSAC_90511_n.txt

Cluster: 2
Member: 1  Number: 3  File: Rocco_S11_mac.txt
Member: 2  Number: 4  File: Rocco_S21_mac.txt
Member: 3  Number: 5  File: Rocco_S22_mac.txt

Cluster: 3
Member: 1  Number: 1  File: Rocco_S5_mac.txt
Member: 2  Number: 2  File: Rocco_S7_Como.txt
Member: 3  Number: 6  File: Rocco_CBZ_III_nomac.txt

Cluster 1:  Representative spectrum: 0
Cluster 2:  Representative spectrum: 3
Cluster 3:  Representative spectrum: 2

Cluster 1: Cluster population: 3  Representative spectrum: 0
Cluster 2: Cluster population: 3  Representative spectrum: 3
Cluster 3: Cluster population: 3  Representative spectrum: 2

Cluster 1  Radius (71.54, 4.47)
Cluster 2  Radius (7.19, 3.98)
Cluster 3  Radius (36.98, 22.72)
```

The section above lists the content of each cluster in terms of samples and file names, and the representative profiles of each cluster, corresponding to those nearest to its center.