IRootLab Tutorial

Data analysis with

Difference-between-mean spectra

and

Cross-calculated LDA

Julio Trevisan – juliotrevisan@gmail.com

Updated on 1st/Dec/2012



Introduction 2

Loading and checking the dataset 2

Differences between mean spectra 4

Running the LDAs 6

Standardizing dataset 6

Direct LDA 6

Cross-calculated LDA 9

Discussion 11

# Introduction

This tutorial shows how to

1) Use differences between mean spectra as a simple way to check for biochemical alterations

2) Use leave-one-out cross-validation to calculate LDA scores (“cross-calculation” of scores)

Difference between means consists of choosing one class to be the reference, and subtracting the mean spectrum from this class from all the spectra in the dataset. This allows one to find which classes have higher or lower absorption for each wavenumber, when compared with a reference class.

As for the cross-calculated LDA[1], the principle is:

1) Use a dataset containing all, except one, samples to train the LDA model

2) Use the model with the left-out sample to calculate the scores for this sample

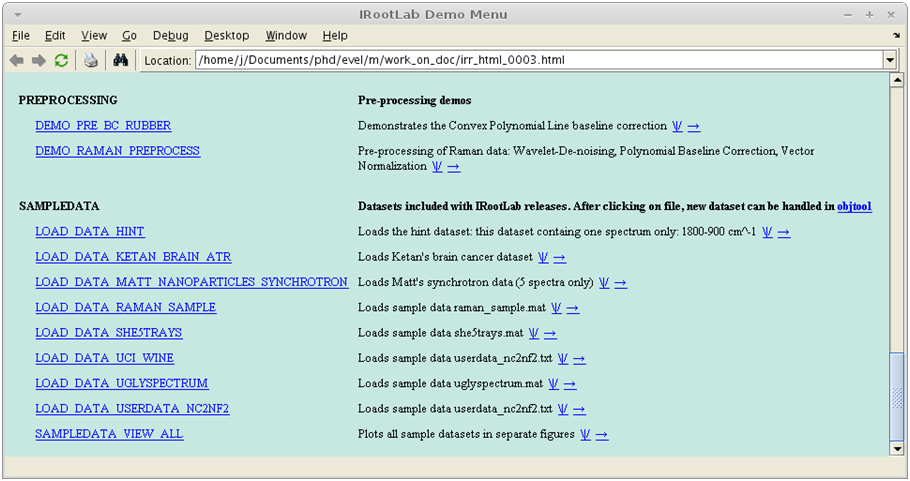
3) Repeat steps 1) and 2) to all the samples until all scores are calculated.

Sample = patient, slide etc (depending on the experiment). All spectra from one sample need to be kept together.

# Loading and checking the dataset

This tutorial uses Ketan’s Brain data[2], which is shipped with IRootLab.

1. At MATLAB command line, enter browse\_demos
2. Click on “LOAD\_DATA\_KETAN\_BRAIN\_ATR”
3. Click on “objtool” to launch objtool

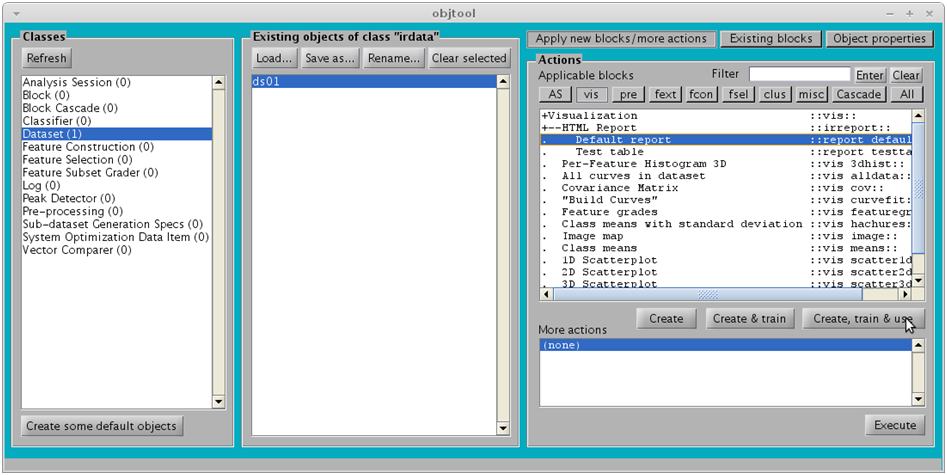


**2**

**3**

The next step will generate a report on the dataset.

1. Click on Apply new blocks/more actions
2. Click on vis
3. Click on Default report
4. Click on Create, train & use



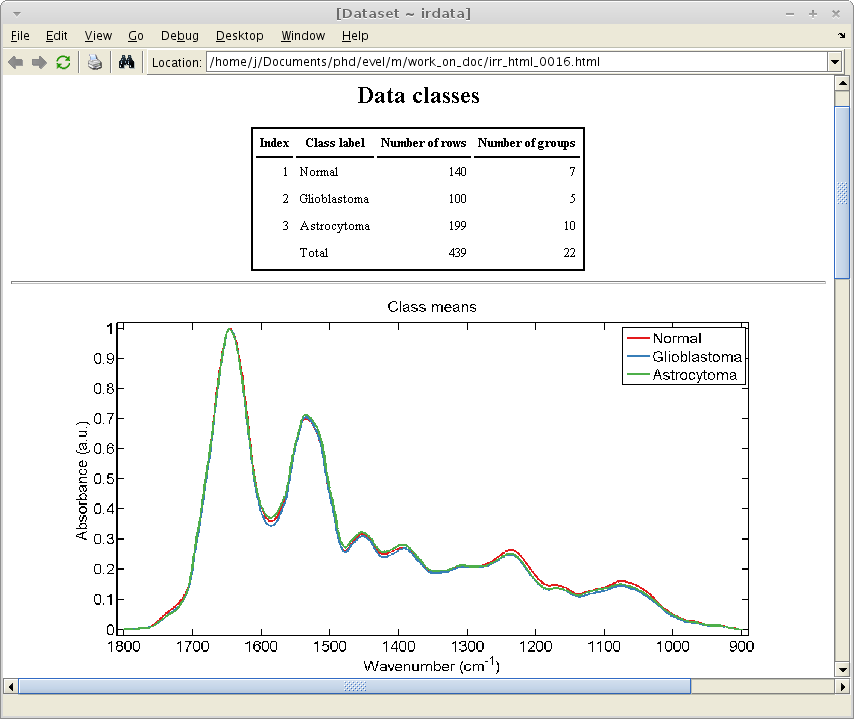
**4**

**6**

**7**

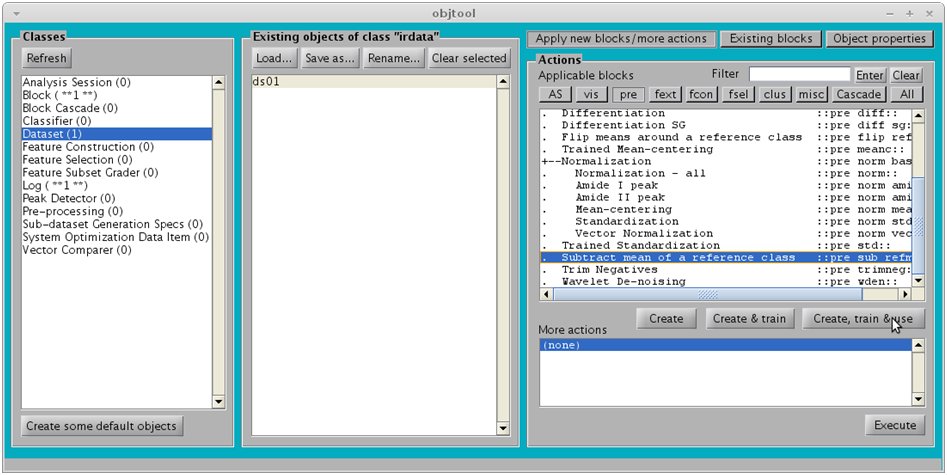
**5**

A window should open displaying the following. As seen, the “Normal” class, which will be the reference class, is the first class (it has index 1).



# Differences between mean spectra

1. Click on pre
2. Click on Subtract mean of a reference class
3. Click on Create, train & use

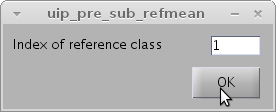


**8**

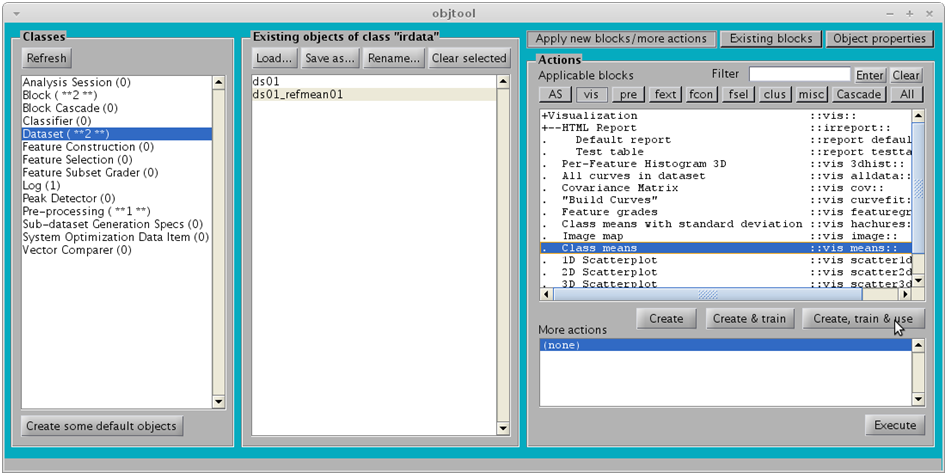
**10**

**9**

1. Accept the value **1** (refers to the first class (“Normal”))



1. Click on ds01\_refmean01
2. Click on Class means
3. Click on Create, train & use



**12**

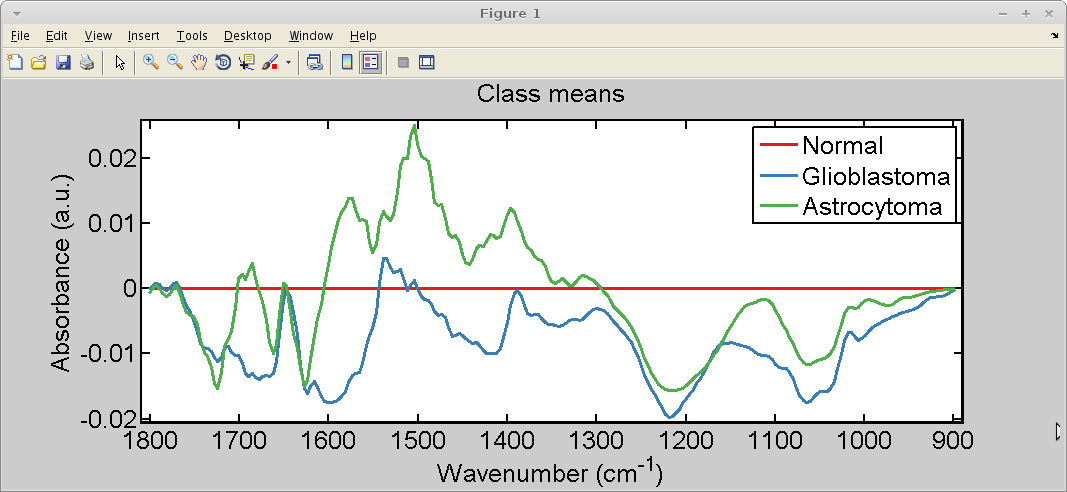
**13**

**14**

The following figure should appear.

**Note** –For example, the average absorbance of the “Astrocytoma” samples is higher than “Normal” between 1600 and 1300 cm-1, peaking at 1500 cm-1, and is lower than “Normal” between 1300 and 900 cm-1.

The “Normal” curve is a flat line as a consequence of the Subtract mean of a reference class operation.



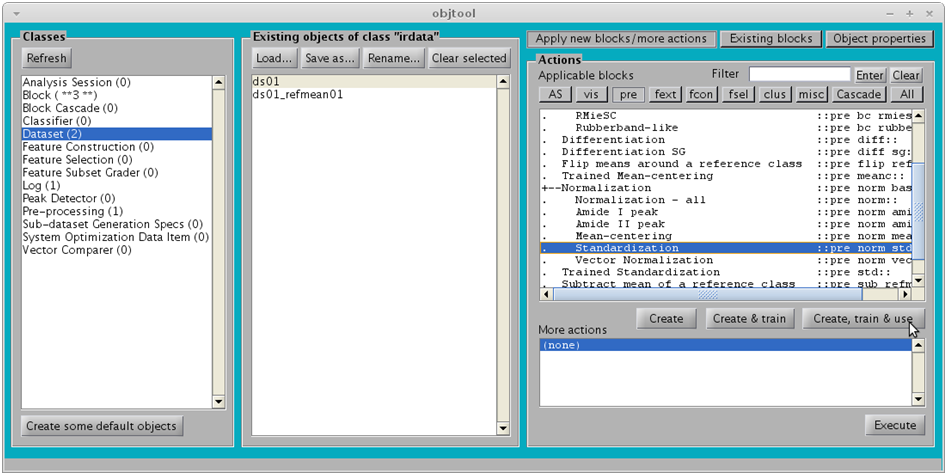
# Running the LDAs

## Standardizing dataset

1. Click on ds01
2. Click on pre
3. Click on Standardization
4. Click on Create, train & use

**Note** –The dataset must be either mean-centered or standardized before cross-calculated LDA. Standardization provides more numerical stability.

Standardization[3] is mean-centering followed by scaling of each variable so that their standard deviations become 1.



**15**

**16**

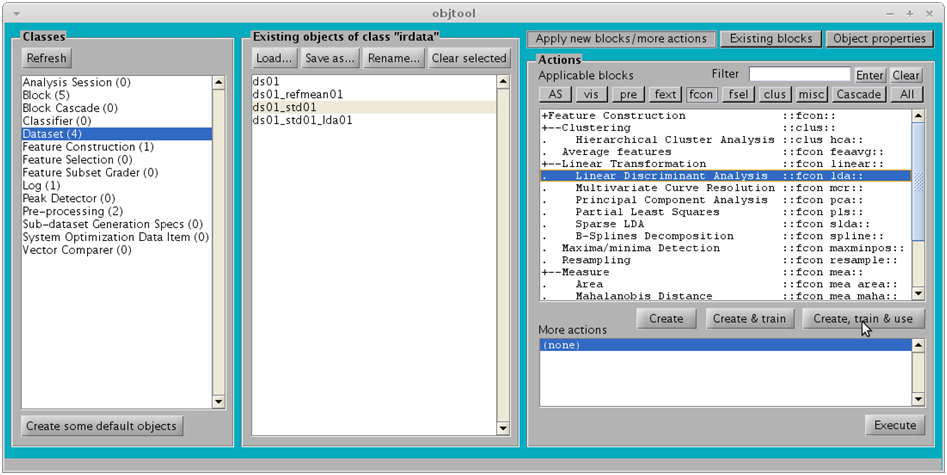
**17**

**18**

## Direct LDA

First we will apply LDA to later compare with the cross-calculated LDA.

1. Click on ds01\_std01
2. Click on fcon
3. Click on Linear Discriminant Analysis
4. Click on Create, train & use



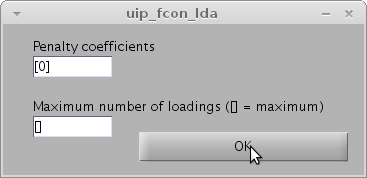
**19**

**20**

**22**

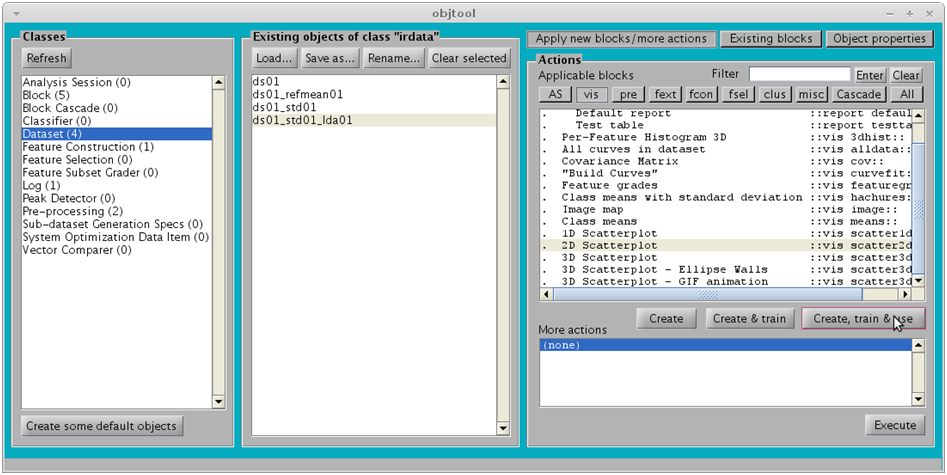
**21**

1. Click on OK



The next step will generate a scores plot

1. Click on ds01\_std01\_lda01
2. Click on vis
3. Click on 2D Scatterplot
4. Click on Create, train & use



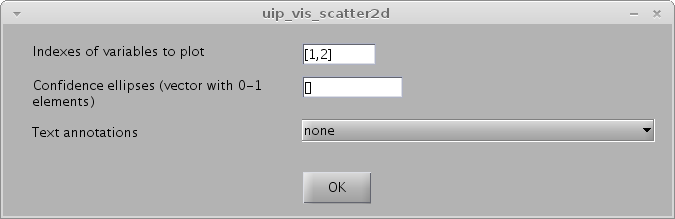
**24**

**25**

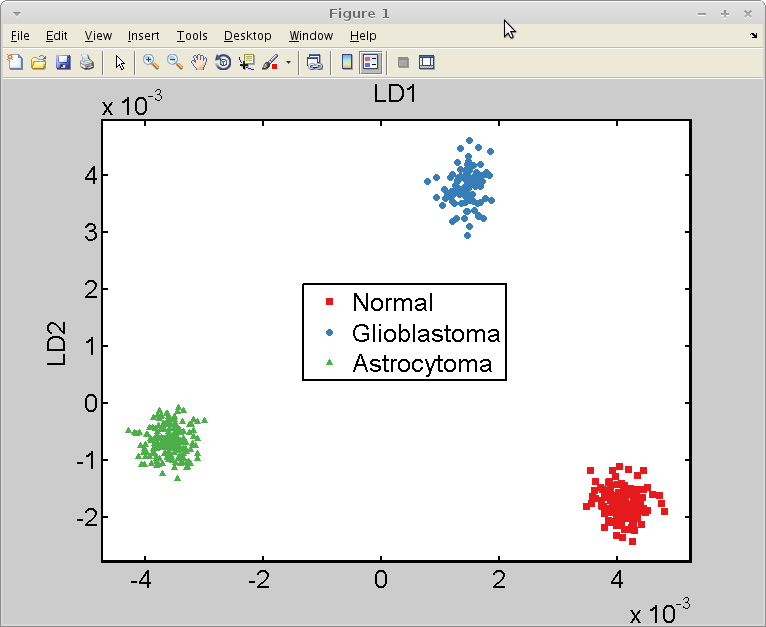
**26**

**27**

1. Click on OK



The following figure should appear:

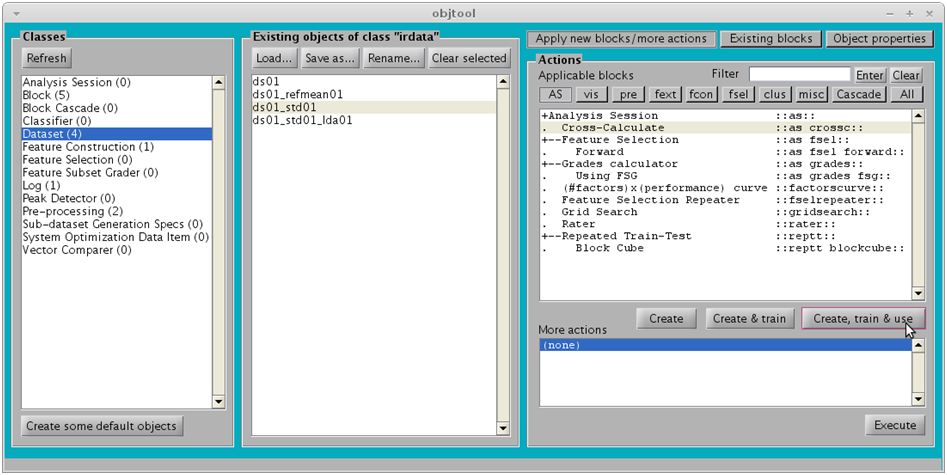


**But … are the classes so well separated because LDA overfits the data???**

**We will do the cross-calculated LDA to find this out.**

## Cross-calculated LDA

1. Click on ds01\_std01
2. Click on AS
3. Click on Cross-calculate
4. Click on Create, train & use



**29**

**30**

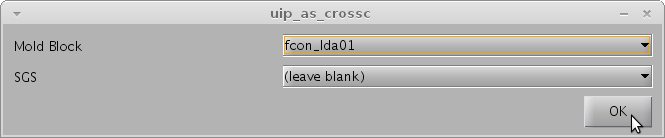
**31**

**32**

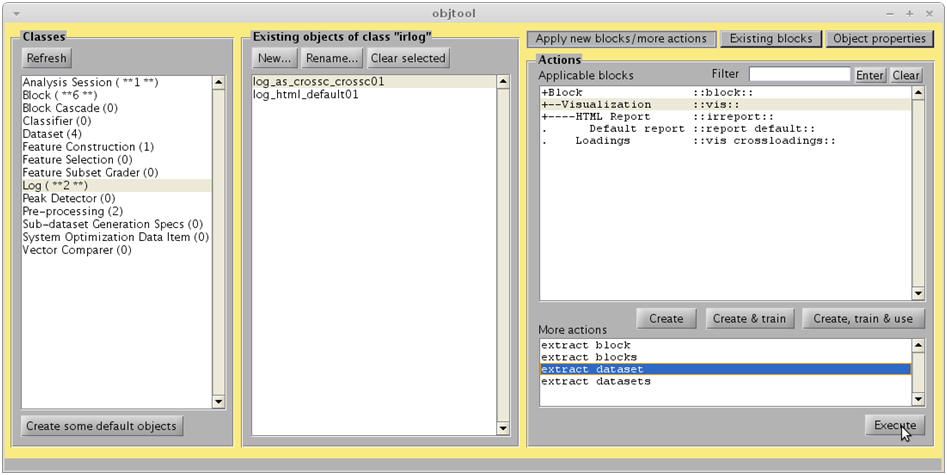
1. Click on OK (a new Log will be created, may take a few seconds)

**Note** –The SGS can be left blank because the cross-calculation block (which we are creating now) will automatically create one. The default SGS is a leave-one-out (LOO) cross-validation that keeps together all the spectra from the same group (patient in this case).

In our case, LOO cross-validation is equivalent to 22-fold cross-validation because the dataset has 22 groups (check report generated on step 7). This implies that to calculate the scores for each group, the 21 other groups will be used to train an LDA block that will be then used on the spectra from that group.



1. Click on Log
2. Click on log\_as\_crossc\_crossc01
3. Click on extract dataset
4. Click on Execute (will create a dataset)



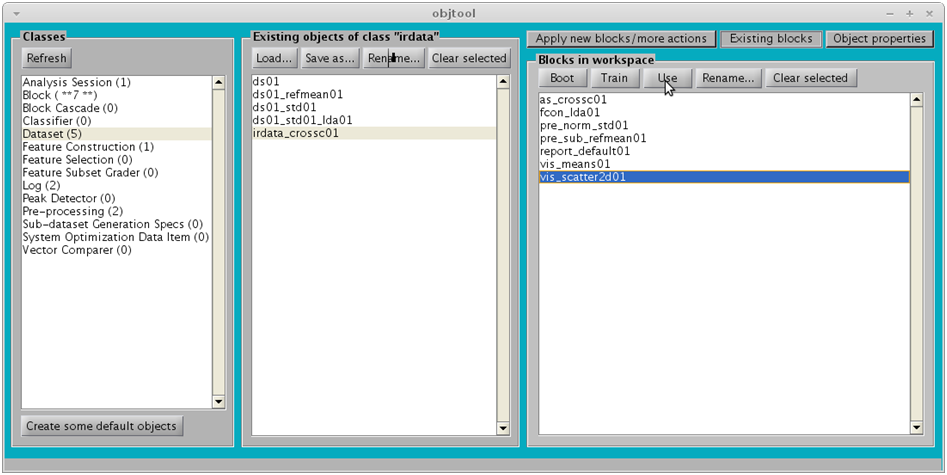
**34**

**35**

**36**

**37**

1. Click on Dataset
2. Click on irdata\_crossc01
3. Click on Existing blocks (we are going to re-use the 2D Scatterplot block)
4. Click on vis\_scatter2d01
5. Click on Use



**38**

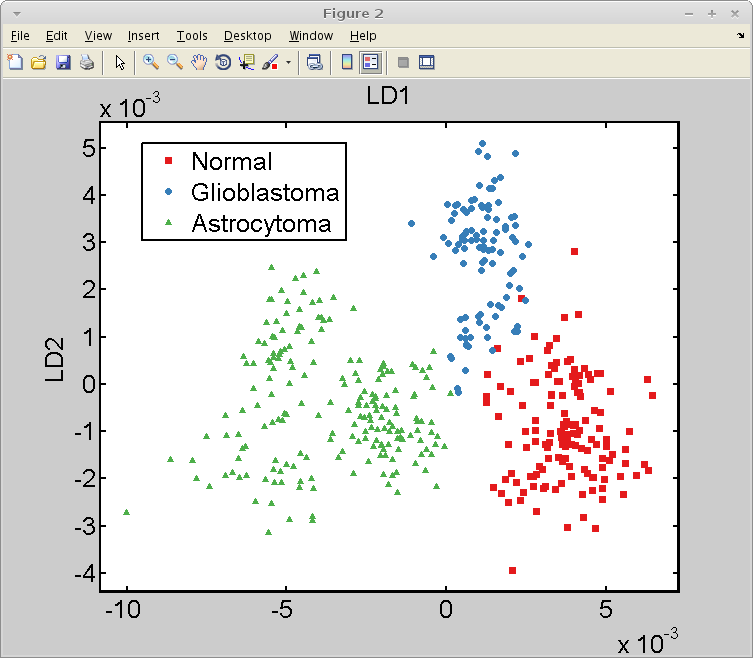
**39**

**41**

**40**

**42**

The following figure should appear:



## Discussion

The classes are not as well separated as with direct LDA. The segregation seen before with direct LDA was unrealistic because of overfitting.

Nevertheless, the results from cross-calculated LDA are still excellent, because the classes are nearly completely separated; just a small overlap is seen between “Normal” and “Glioblastoma”.

# References

[1] M. J. Riding, F. L. Martin, J. Trevisan, V. Llabjani, I. I. Patel, K. C. Jones, and K. T. Semple, “Concentration-dependent effects of carbon nanoparticles in gram-negative bacteria determined by infrared spectroscopy with multivariate analysis.,” *Environ. Poll.*, vol. 163C, pp. 226–234, Jan. 2012.

[2] K. Gajjar, L. Heppenstall, W. Pang, K. M. Ashton, J. Trevisan, I. I. Patel, V. Llabjani, H. F. Stringfellow, P. L. Martin-Hirsch, T. Dawson, and F. L. Martin, “Diagnostic segregation of human brain tumours using Fourier-transform infrared and/or Raman spectroscopy coupled with discriminant analysis,” *Analytical Methods*, vol. 44, no. 0, pp. 2–41, 2012.

[3] T. Hastie, J. H. Friedman, and R. Tibshirani, *The Elements of Statistical Learning*, 2nd ed. New York: Springer, 2007.