ECLIPSE tutorial

The ECLIPSE package provides a novel multivariate method to identify and characterize aberrant cells present in individuals out of homeostasis. ECLIPSE combines dimensionality reduction by Simultaneous Component Analysis with Kernel Density Estimates to eliminate cells in responder samples that overlap in marker expression with cells of controls. Subsequent data analyses focus on the immune response-specific cells, leading to more informative and focused models.

The package includes scripts and functions written for MATLAB, which is needed to apply the method. We suggest running the code in a Windows environment, as some functions may not work otherwise. ECLIPSE has been thoroughly tested using Windows 10, on MATLAB 2016.

Step by step

- Open MATLAB
- Add the directory containing the folder named "MATLAB" to the MATLAB path:

```
p = genpath('C:\Mydirectory\myfiles\ECLIPSE\MATLAB');
addpath (p)
```

- For convenience, set as MATLAB working directory the folder containing the cytometry data. You can use the LPS data included in the repository.
  (note the Asthma data used in the ECLIPSE manuscript is available at http://www.ru.nl/science/analyticalchemistry/research/data)

```
cd ('C:\mydata\MFCdata')
```

- To open the script with the method type:

```
open ECLIPSE_main_script.m
```

Click on each section, which will highlighted in yellow, to activate it. In the tab **EDITOR** in MATLAB, press **Run Section** to run the activated section.

Below it is shown how to run the ECLIPSE analysis, using the LPS as example of usage.

```
- SECTION A import/load data
```

Do you need to import the data or load the data from struct (import/load)?

Type **import** to import data which need to be in fcs, lmd, txt, csv format.

Excel sheet with filenames, Identifiers of the individuals (IDs) and Labels (group belonging, e.g. 0 for controls, 1 for responders) should be uploaded as well, if present. If this is not present, it needs to be created.

Type load if you already created the .mat file.

- SECTION B data selection

1. Paired data question: paired data=1 if the two groups consists of the same individual (e.g. before and after treatment), paired data=0 if the two groups NOT consists of the same individual
2. Data selection: to select the variables to use for the analysis and specify the labels associated to the control and responder group.
   Note: You should only use variables that do not have a large shifts due to sample preparations e.g. forward and sideward scatter. You should also remove variables that were only used as a 'dump' gate.

   For the LPS data:

```
Is the data paired? (yes/no) no

What variables should be used (give in [] or as 1:5)?
1:  FS
2:  SS
3:  SIRL
4:  Lair
5:  CD62L
6:  CD11c
7:  CD69
8:  CD32
9:  CBRM1/5
10:  CD11b
11:  CD64
12:  CD16
13: Time

Variables used: [5:8 10:12]
What is the label of the control in the data:    0
What is the label of the diseased in the data:   1
```

- SECTION C pre-processing

Different options of pre-processing are possible (see the Supplementary Material of the ECLIPSE manuscript for detail).

For the LPS data use `selection_mode=[1 5 2]`

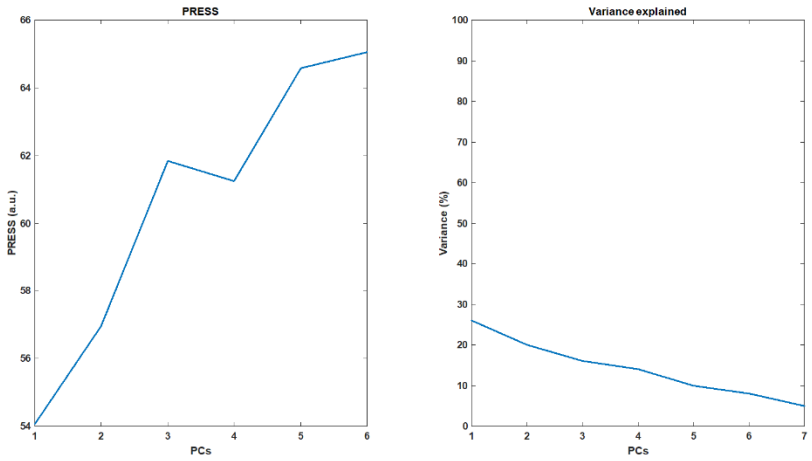
- Sub-SECTION C

  In this section, you can visually investigate the effect of the preprocessing by plotting histograms of single markers expression. You can choose to color the histograms either per individual sample or per class (in this last case, *red* for group with *label 1* and *blue* for group with *label 0*). (Note type close `close all` to close all the generated figures).

- SECTION D: Simultaneous Component Analysis (Control Model)

Section D is used to define the PCs to use for building the Control Model. Control Model is a SCA-based model built only on the cells from the control individuals (group label 0), cells from the diseased/response (group label 1) are projected in this obtained space. This is useful when we to focus the analysis on how a response deviate from a 'normal' condition.

The first function in this section gives two plots. On the left, you can see the PRESS (Predicted Residuals Sum of Squares), the lower the value the better. On the right, you can see the Variance explained per PC. Based on these Figures you can decide which PCs you want to include. In the example of LPS, you might want to include PC1 and PC2 (as the error with the other PCs goes high). Then type: [1 2]



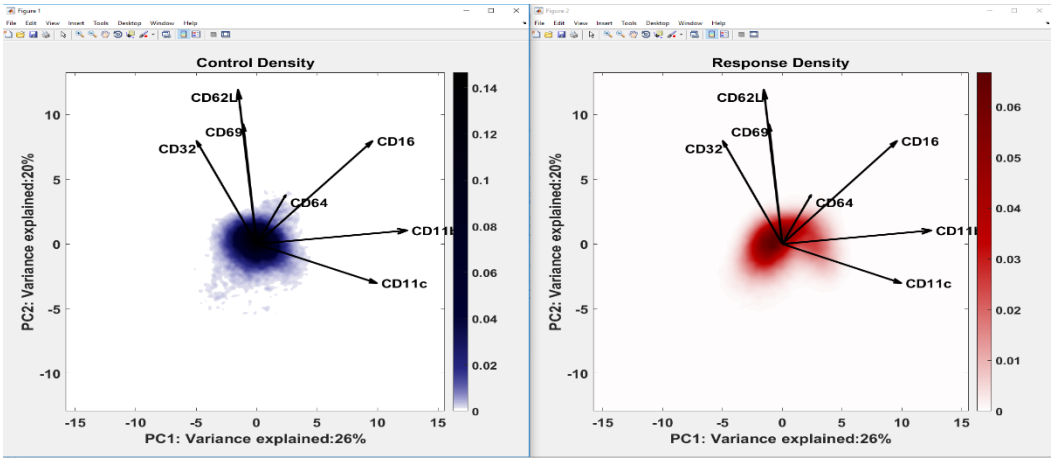
- SECTION E – Compute KDE probability densities of cell distribution in the Control Space

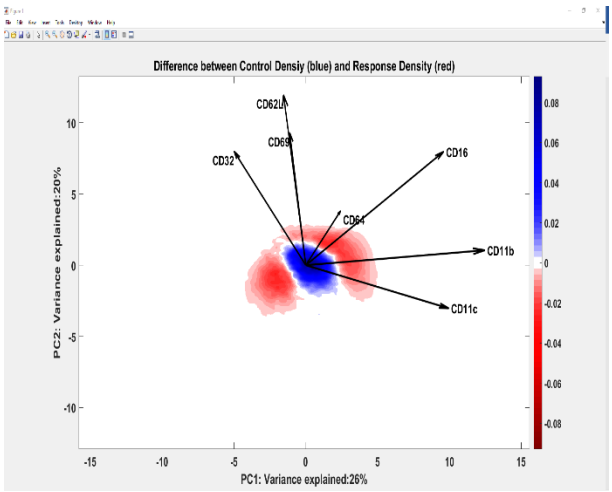
This section will estimate the probability densities estimation of the cells scores in the Control Model. The estimation of these densities will make comparable the cell distribution of diseased/response again the control individuals. Visualization of the KDEs is possible through the next section.

- SECTION Fa/Fb – Visualization of the estimated probability densities

In the case of section Fa, visualization of:

The cumulative Control Density (of all the cells from control individuals), the cumulative Response Density (of all the cells from response individuals), and the difference between them will be plotted, as in the example below for the LPS data:





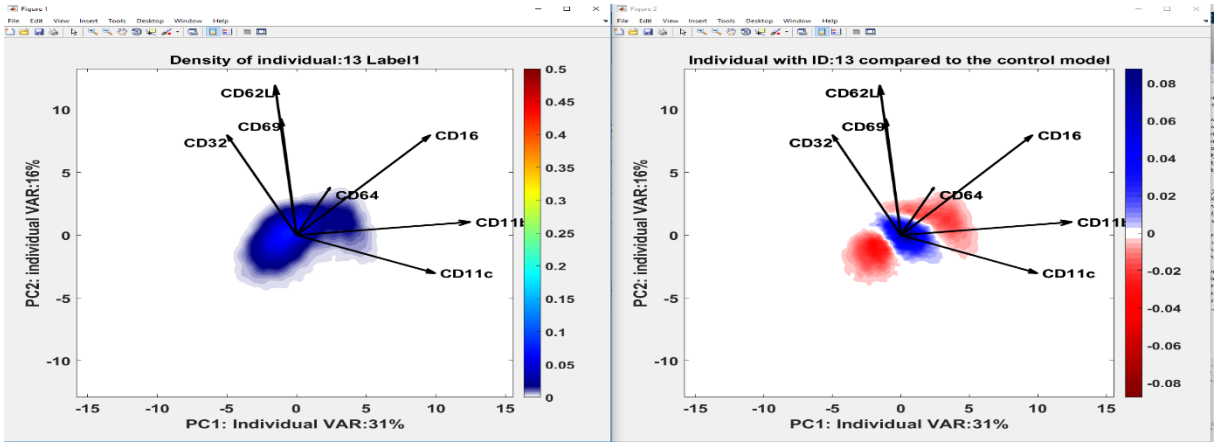
The difference between densities graph will show the results of Control Density – Response Density, shown on the first panels. The red region intensity (negative) indicates the location where more likely are present cells over-produced in the responder; the blue region (positive) indicates the location where more likely are present healthy cells. The white area corresponds to a value of Control Density – Response Density=0, which can indicate bins with no cells or equal intensity of control and responder estimates.

After inspecting the plots, you might want to close the figures. In this case type:

```
close all
```

In the case of section Gb, visualization of:

Probability density of single individuals, after specifying the ID(s) of the sample(s). The difference between densities plot will show the results of Control Density – Individual Density.



- SECTION H – Indices to extract 'abnormal' cells

This section will build indices to extract cells which exceed the Sum Prediction Error limit (so they are not described by the Control Model) and cells found by the Difference between Densities to be more present in the diseased/response individuals.

- SECTION J – Elimination Step

In the elimination step, abnormal cells will be identified (so the normal will be eliminated from the subsequent modeling). The new matrix S_in contains the orginal raw data

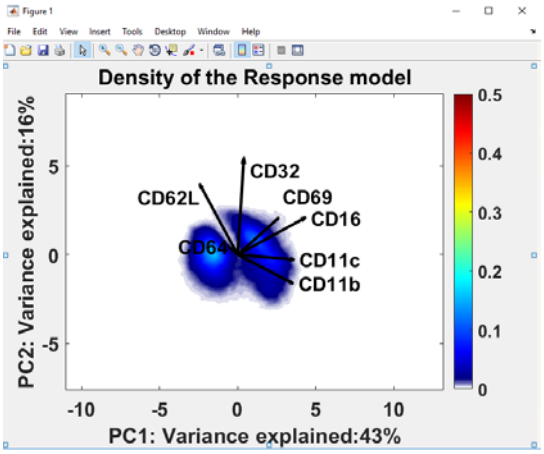
- SECTION K – Select again the variables for S_in (after removal of normal cells) + preprocess the data

The subsequent analysis, ECLIPSE model, will be built on the S_in matrix, which contain raw data of only the abnormal cells. Variables selection and pre-processing is needed again.

- SECTION L – ECLIPSE Model

Two different options are possible:

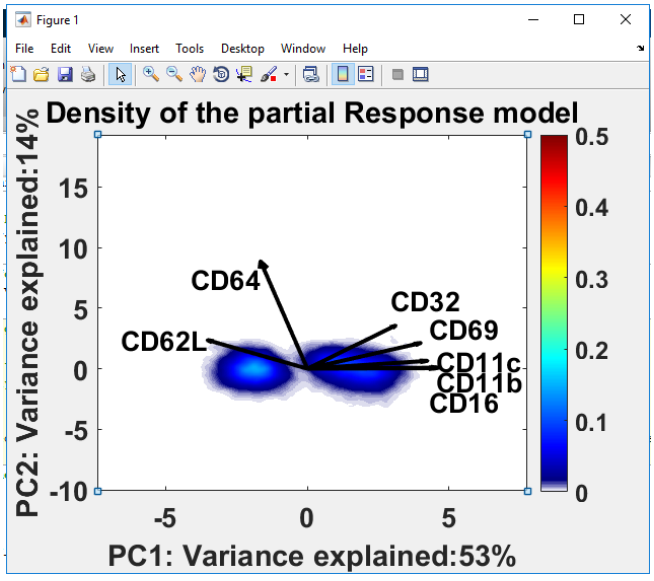
General model: build a PCA-based model on all the responder individuals.



If you wish to adjust the colormap range type caxis([0 0.5])

Partial model: select only a subset of responder individual and build a PCA-based model on those.

The following IDs belong to Responders:
1: ID:9
2: ID:10
3: ID:11
4: ID:12
5: ID:13
6: ID:14
7: ID:15
8: ID:16
Enter the IDs of the individuals you want to base the model on. ([1 3 5], 1:5, etc.) : [15 16]



- `SECTION L - Build individual ECLIPSE mode`

Build a PCA-model on the responding cells of single individual

```
%Note: Drawback is that the personal models cannot be compared
between individuals
```