

## Supplementary Information 1

**Article title:** Development of a classification and ranking method for the identification of possible biomarkers in two-dimensional gel-electrophoresis based on Principal Component Analysis and variable selection procedures

**Authors:** Elisa Robotti<sup>1</sup>, Marco Demartini<sup>1</sup>, Fabio Gosetti<sup>1</sup>, Giorgio Calabrese<sup>2</sup>, Emilio Marengo<sup>1\*</sup>

**Affiliation:** <sup>1</sup> Department of Environmental and Life Sciences – University of Eastern Piedmont – Viale T. Michel 11 – 15121 Alessandria – Italy; <sup>2</sup> Università Cattolica del Sacro Cuore - Via Emilia Parmense, 84 - 29122 Piacenza - Italy

**Supplementary Information 1:** Pseudocode adopted in Matlab environment for Ranking-PCA in cross-validation with autoscaling

```
% FUNCTION MAIN
```

```
1 Define itermax (maximum number of iterations to achieve convergence in PCs calculation)
```

```
2 Load the data
```

```
3 Initalise ind1old=0
```

```
% SELECTION OF THE FIRST VARIABLE
```

```
4 for j=1: nvar (number of variables)
```

```
5     calculation of the first PC in cross-validation
```

```
6     calculation of ind1, ind2, distance between the centroids of the two classes
```

```
7     if ind1>ind1old,
8         ind1old=ind1; ind2old=ind2 (results of the best variable identified so far are
           recorded)
9     end if
10 exit cycle on j
11 ind1, ind2 and the distance are calculated in calibration with the best variable selected
12 results in both calibration and validation are recorded

% SELECTION OF SUBSEQUENT VARIABLES

13 for p=nvar-1:-1:1 (cycle on the selection cycles to be carried out)
14     initialise ind2old=0;
15     define the maximum number of PCs to be calculated as the minimum among: (sam1-1-
           VALID), (sam2-1-VALID), nmaxPC, v (sam1=number of samples of the first class; sam2 =
           number of samples of the second class; VALID = number of object left out during cross-
           validation; nmaxPC= maximum number of PCs to be calculated; v= number of variables
           added at cycle p)
16     for j=1:size(data,2) (cycle on the number of variables to be evaluated for selection)
17         calculation of the PCs in cross-validation by NIPALS with the j-th variable included
18         initialise ind1old=0
19         for I=1:modelnum (cycle on the number of models in PCs to be evaluated)
20             calculation of ind1, ind2, distance between the centroids
21             if convergence on PCs is achieved
22                 if ind1>ind1old,
23                     if ind2>ind2old,
```

24 results are recorded for the best variable evaluated so far:  
ind1old=ind1, ind2old=ind2

25 end all if cycles

26 exit cycle on I

27 calculation of the results in calibration with the variable selected

28 exit cycle on j

29 save results

30 exit cycle on selection cycles p