Mahalanobis-Distance 1.0
(MD)

NANOBRIDGES
-ACollaborative Project

UNIWERSYTET GDAŃSKI

Marie Curie Actions
EUROPEAN UNION

THE AUTHORS ARE GRATEFUL FOR THE FINANCIAL SUPPORT FROM THE EUROPEAN COMMISSION THROUGH THE MARIE CURIE **IRSES** PROGRAM, NANOBRIDGES PROJECT (FP7-PEOPLE-2011-IRSES, GRANT AGREEMENT NUMBER 295128).

Last updated: 22.05.2015

Mahalanobis-Distance 1.0

Background:

The multivariate chemometrical analysis involves the measurement of distances between objects and variables. The two widely and most commonly used distance measures are Euclidean distance (ED) and the Mahalanobis distance (MD). The ED is easy to compute and interpret but the calculation of MD takes into account the correlation in the data, since it is calculated using the inverse of the variance-covariance matrix of the dataset. When the data are measured over the large number of variables, due to the redundant or correlated information leads to singular or nearly singular variance-covariance matrix that cannot be inverted. Subsequently, for the calculation of variance-covariance matrix, the number of objects in dataset should be larger than the number of variables. So, feature reduction i.e selecting a small number of meaningful variables is required prior to the finalizing the input file for MD calculation [1].

The MD is useful for the many purposes such as detection of outliers, the selection of calibration samples from a larger set of measurement and for investing the representativity between the two datasets. In pattern recognition, the MD is applied in clustering techniques such as k-Nearest Neighbour method (kNN), in discrimination techniques such as linear, quadratic and regularised discriminating analysis (LDA, QDA and RDA) [2] and in class modelling techniques such as UNEQ (multivariate normal class model assuming an individual dispersion of each class) [3], EQ (multivariate normal class model assuming equal dispersion of each class) [4] and modifications of Soft Independent Modelling of Class Analogy (SIMCA).

About the algorithm: The algorithm involved in the calculation of MD is as follows:

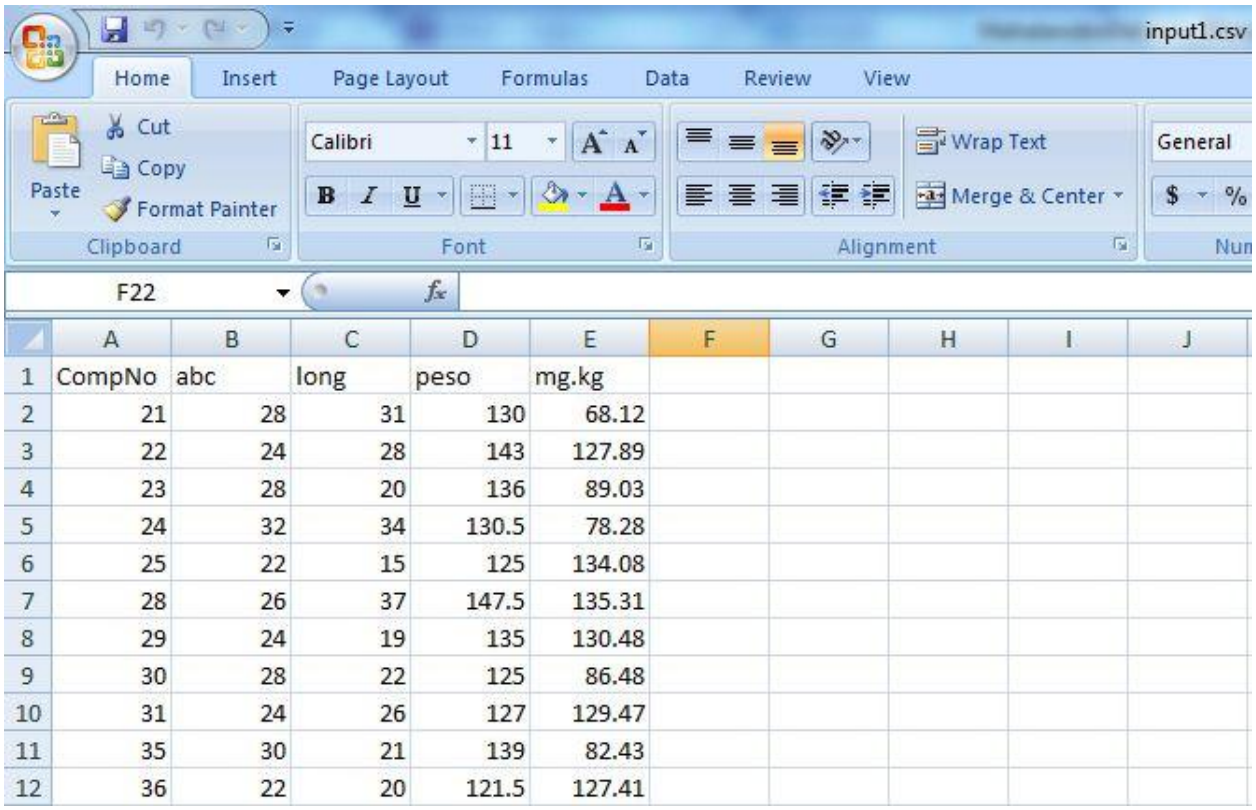
$$C_x = \frac{1}{(n-1)}(X_c)^T(X_c) \qquad \text{(i)}$$

$$MD_i = \sqrt{(x_i - x)C_x^{-1}(x_i - \bar{x})^T} \qquad \text{(ii)}$$

Where, C_x is the variance-covariance matrix; X is the data matrix containing n objects. X_c is the column-centered data matrix ($X - \overline{X}$)

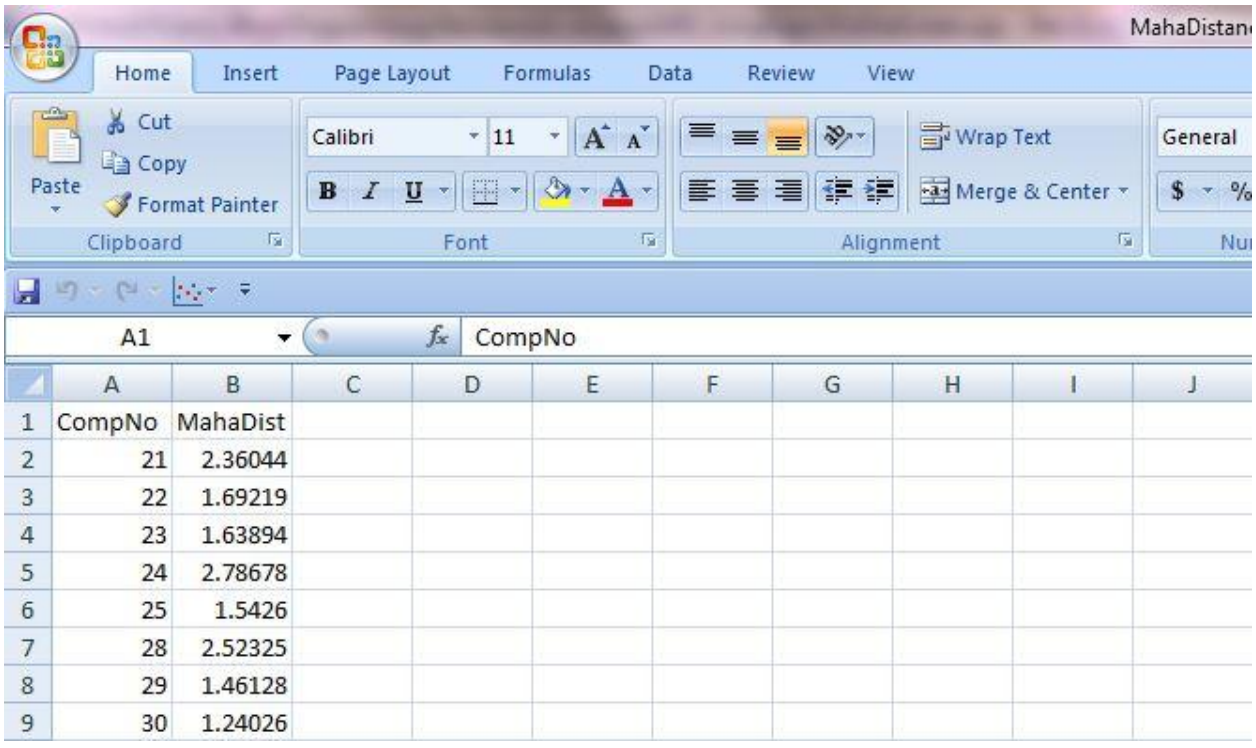
The eigen library was used for the calculation of matrix determinant and inverse calculation[5].

Input file format: This program takes a input file in CSV format.



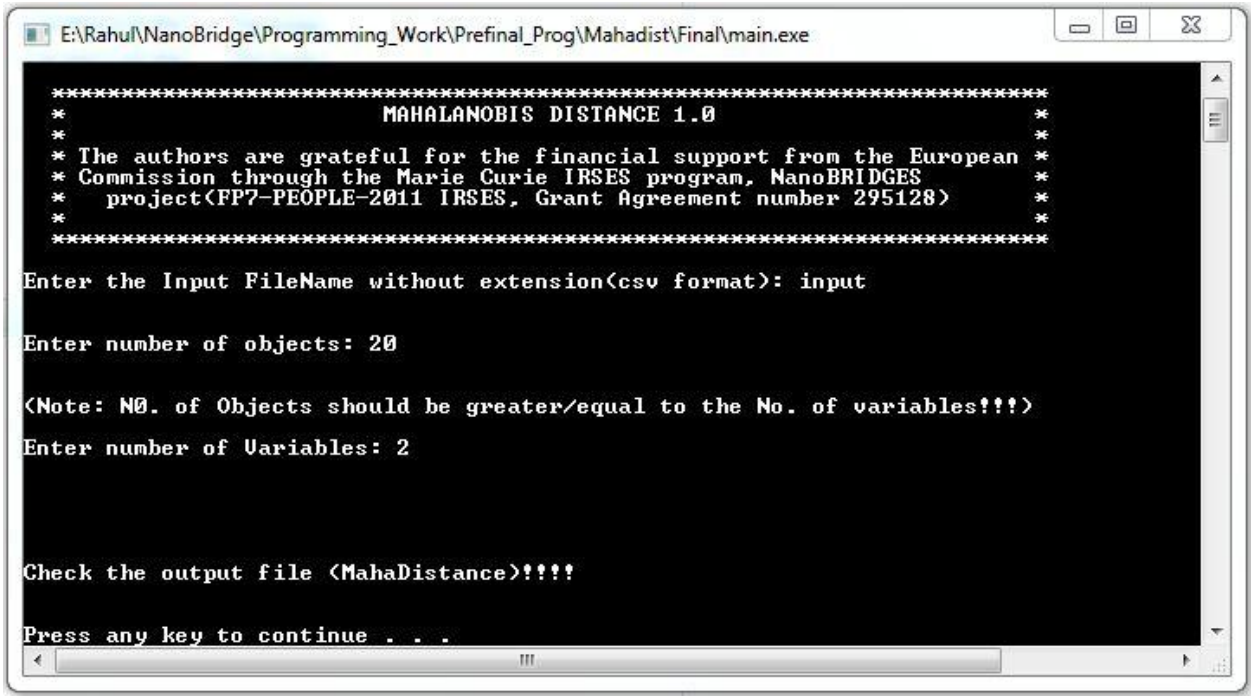
	A	B	C	D	E	F	G	H	I	J
1	CompNo	abc	long	peso	mg.kg					
2	21	28	31	130	68.12					
3	22	24	28	143	127.89					
4	23	28	20	136	89.03					
5	24	32	34	130.5	78.28					
6	25	22	15	125	134.08					
7	28	26	37	147.5	135.31					
8	29	24	19	135	130.48					
9	30	28	22	125	86.48					
10	31	24	26	127	129.47					
11	35	30	21	139	82.43					
12	36	22	20	121.5	127.41					

Output file format:



	A	B	C	D	E	F	G	H	I	J
1	CompNo	MahaDist								
2	21	2.36044								
3	22	1.69219								
4	23	1.63894								
5	24	2.78678								
6	25	1.5426								
7	28	2.52325								
8	29	1.46128								
9	30	1.24026								

How to run a program:



Disclaimer

For academic purpose only.

The program Mahalanobis-Distance 1.0 has been developed in C++ language and is validated on the known data sets. This program is compatible with both 32- and 64-bit Windows operating system. Please report for discrepancy of result for any other dataset. Contact us at any of the following addresses:

Dr. Tomasz Puzyn,
NanaBRIDGESProject
Coordinator,
Faculty of Chemistry,
University of Gdansk,
Gdansk,
Poland 80-952
Email Id: puzi@qsar.eu.org

Dr. Kunal Roy,
Drug Theoretics and Cheminformatics Lab.,
Dept. of Pharmaceutical Technology,
Jadavpur University,
Kolkata, West Bengal,
INDIA-700032
Email Id: kunalroy_in@yahoo.com

Rahul BalasahebAher,
Phd Research Scholar,
Drug Theoretics and Cheminformatics Lab.,
Dept. of Pharmaceutical Technology,
Jadavpur University,
Kolkata, West Bengal,
INDIA-700032
E-mail Id: rahulba26@gmail.com(*for any queries regarding the program)

References:

1. De Maesschalck, Roy, DelphineJouan-Rimbaud, and Désiré L. Massart. "The mahalanobis distance." *Chemometrics and intelligent laboratory systems* 50, no. 1 (2000): 1-18.
2. Wu, W., Y. Mallet, B. Walczak, W. Penninckx, D. L. Massart, S. Heuerding, and F. Erni. "Comparison of regularized discriminant analysis linear discriminant analysis and quadratic discriminant analysis applied to NIR data."*AnalyticaChimicaActa* 329, no. 3 (1996): 257-265.
3. Derde, M. P., and D. L. Massart. "UNEQ: a disjoint modelling technique for pattern recognition based on normal distribution." *AnalyticaChimicaActa* 184 (1986): 33-51.
4. Coomans, D., I. Broeckaert, M. P. Derde, A. Tassin, D. L. Massart, and S. Wold. "Use of a microcomputer for the definition of multivariate confidence regions in medical diagnosis based on clinical laboratory profiles." *Computers and biomedical research* 17, no. 1 (1984): 1-14.
5. http://eigen.tuxfamily.org/index.php?title=Main_Page