ICIC Express Letters
Volume 6, Number 5, May 2012

ICIC International ©2012 ISSN 1881-803X
pp. 1335–1339

MULTIVARIATE ANALYSIS OF GENE EXPRESSION DATA AND MISSING VALUE IMPUTATION BASED ON LLSIMPUTE ALGORITHM

KOHBALAN MOORTHY[1], MOHD SABERI MOHAMAD[1], SAFAAI DERIS[1] AND ZUWAIRIE IBRAHIM[2]

[1]*Artificial Intelligence and Bioinformatics Research Group*
Faculty of Computer Science and Information Systems
[2]Department of Mechatronics and Robotics
Center of Artificial Intelligence and Robotics
Faculty of Electrical Engineering
Universiti Teknologi Malaysia
81310 UTM Skudai, Johor, Malaysia
kohbalan@gmail.com; {saberi; safaai}@utm.my; zuwairiee@fke.utm.my

Received July 2011; accepted October 2011

ABSTRACT. *Microarray ade4 or known as MADE4 is a multivariate software analysis package for microarray gene expression data. This software package is capable of accepting wide variety of gene expression data formats such as Bioconductor AffyBatch and exprSet. This MADE4 R package extends the advantages of ade4 package in multivariate statistical and graphical functions for the use in the microarray data application. Moreover, MADE4 provides new graphical and visualization tools that assist in the interpretation of multivariate analysis of microarray data. Besides that, LLSimpute algorithm has been incorporated to assist in handling of datasets with missing values and this has eased the application for the users to analysis on gene expression data that contain missing values.*

Keywords: Multivariate analysis, Correspondence analysis, Between group analysis, Co-inertia analysis, Microarray data, Gene expression data, LLSimpute algorithm

1. **Introduction.** MADE4, or known as microarray ade4, is a software package based on R language that was developed to facilitate multivariate analysis of microarray gene-expression data. Furthermore, MADE4 accepts a wide variety of gene-expression data formats and takes advantage of the extensive multivariate statistical and graphical functions in the R package ade4, extending these for application to microarray data. In addition, MADE4 provides new graphical and visualization tools that assist in interpretation of multivariate analysis of microarray data. The aim of this development of microarray ade4 (MADE4) is to provide a simple-to-use tool for multivariate analysis of microarray data [1]. Multivariate analysis encompasses much more methods than these examples of linear modelling implied by [2].

Besides that, the input of various types of datasets has been a key factor for the usability of made4 in gene expression analysis, as wide variety of gene expression data input formats such as Bioconductor AffyBatch, exprSet, marrayRaw, and standard R matrix formats (data.frame or matrix) are available from all different sources for multivariate analysis. This multivariate analysis software (made4) has limitation for dataset pre-processing, whereby the dataset that contains missing values cannot be used to perform multivariate analysis as the missing values estimation is a crucial step in the datasets pre-processing [3]. Therefore, a function for reading and inserting the missing values to the dataset based on local least squares imputation method (LLSimpute) has been proposed. This LLSimpute algorithm is taken from [3].

2. **LLSimpute Algorithm.** The LLSimpute algorithm is used to estimate the missing values in target genes as the linear combination of their most k-similar neighbors chosen by the first k smallest Euclidean distance. For example, assuming that the target gene g_1 contains a missing value in the first position of its total $n = 5$ experiment measures, k similar genes are chosen, which consist of complete measurements before imputing the missing value in target gene, then matrix A is constructed, vectors b and w, and the missing value as follows:

$$\begin{pmatrix} \alpha & w^T \\ b & A \end{pmatrix} = \begin{pmatrix} \alpha & w_1 & w_2 & w_3 & w_4 \\ b_1 & A_{1,1} & A_{1,2} & A_{1,3} & A_{1,4} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ b_k & A_{k,1} & A_{k,2} & A_{k,3} & A_{k,4} \end{pmatrix} \tag{1}$$

where α is the missing value in g_1, $w^T \in R^{1x(n-1)}$ contains $n-1$ elements of g_1 whose first missing item is deleted, the elements of $b \in R^{kx1}$ are the first components of the k-nearest genes, and the rows of the matrix A contain k-nearest neighbor genes with their first values deleted. With the above definition, the least squares problem based on L_2-norm can be formulated as,

$$\min_x \left\| A^T x - w \right\|_2 \tag{2}$$

Then, the missing value α is estimated as linear combination of the vector b:

$$\alpha = b^T x = b^T (A^T)^\dagger w \tag{3}$$

where $(A^T)^\dagger$ is the pseudoinverse of A^T. This procedure will be implemented in the LLSimpute function.

3. **Datasets.** The made4 software package includes two microarray gene expression dataset, which are *khan* and NCI60. *Khan* is a microarray gene expression dataset from [4] which contains SRBCT gene expression data. NCI60 is the microarray gene expression profiles of the NCI 60 cell lines. Both of the datasets included are incomplete due to the limitation of the R-package size. Therefore, the complete dataset of *khan* and NCI60 has been used to replace the included dataset from the made4 package.

Besides that, eight more datasets have been added to this made4 software for multivariate analysis, making it to ten dataset in total. The datasets are Adenocarcinoma, Brain, Breast, Colon, Leukemia, Lymphoma, and Prostate. For the Breast cancer dataset, there are two separate datasets, which contain class 2 and class 3. The detailed information for each dataset is listed in Table 1.

TABLE 1. Main characteristics of the microarray datasets used

Dataset Name	Genes	Patients	Classes	Reference
Adenocarcinoma	9868	76	2	[5]
Brain	5597	42	5	[6]
Breast2	4869	77	2	[7]
Breast3	4869	95	3	[7]
Colon	2000	62	2	[8]
Leukemia	3051	38	2	[9]
Lymphoma	4026	62	3	[10]
NCI60	5244	61	8	[11]
Prostate	6033	102	2	[12]
SRBCT	2308	63	4	[4]

4. **Results and Discussion.** The development and integration of LLSimpute algorithm for the missing values imputation has solved the usability of the datasets for the multivariate analysis of the gene expression data. Without the missing values imputation, the datasets could have been rendered useless for the multivariate analysis since incomplete datasets cannot be analyzed.

Due to the huge amount of visualization output generated for each datasets, all of the results generated through these multivariate analyses have been recorded and presented in the supplementary page, which can be downloaded at http://www.utm.my/aibig/people/mohd-saberi-mohamad/research/supplementary-information.html. An example output for each analysis is presented in the sections below to further facilitate the understanding of the functions of the multivariate analysis.

4.1. **Overview of dataset.** The overview function is a very simple wrapper function that draws a boxplot, histogram, and hierarchical tree of expression data. The hierarchical plot is produced using average linkage cluster analysis with Pearson's correlation metric. An example using Brain dataset is shown in Figure 1.

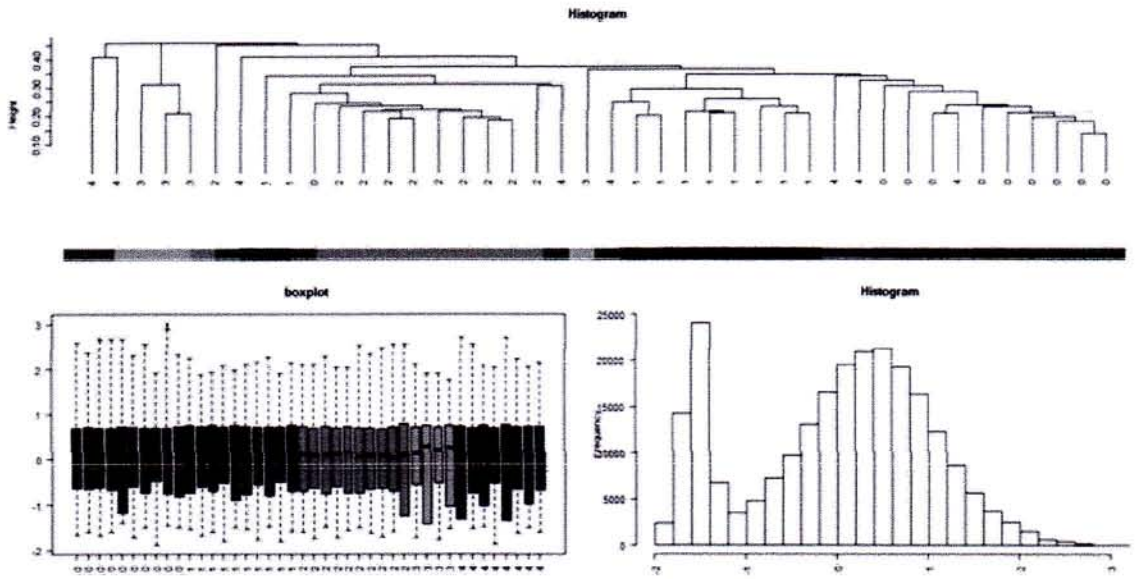

FIGURE 1. Overview of the brain dataset

4.2. **Correspondence analysis (COA).** The application of correspondence analysis is to study the association between microarray samples and genes in a reduced dimensional space. It is more like principal component analysis, where it displays a low-dimensional projection of the data, e.g., into a plane. This is done for two variables simultaneously, thus revealing associations between them.

Once the correspondence analysis is done, a plot is produced with four separate visual outputs, where the first view on the top left is a plot of the eigenvalues, followed by the top right view for the projection of microarray samples from patient with tumor types, whereby each tumor is labeled in a separate color. In the bottom left view, we can see the projection of genes (gray filled circles) is shown and finally the bottom right view contains the biplot showing both genes and samples. Samples and genes with a strong associated are projected in the same direction from the origin. The greater distance from the origin produces the stronger the association. An example using Brain dataset is shown in Figure 2.

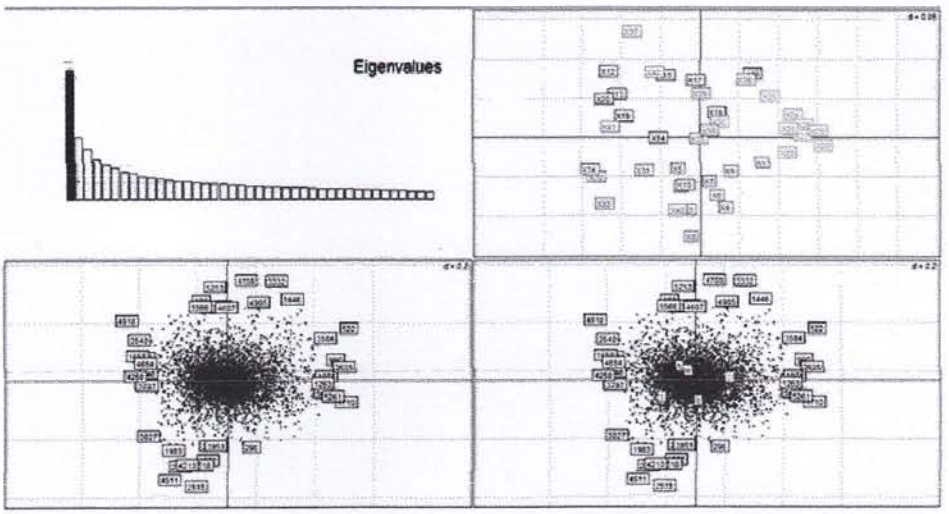

FIGURE 2. Correspondence analysis of the brain dataset

4.3. **Between groups analysis (BGA).** Between Group Analysis (BGA) is a supervised classification method. Therefore, the classification and class prediction is done using Between Group Analysis. The basis of BGA is to ordinate the groups rather than the individual samples. The plots done below are single dimension plot for the BGA analysis. An example using Brain dataset is shown in Figure 3.

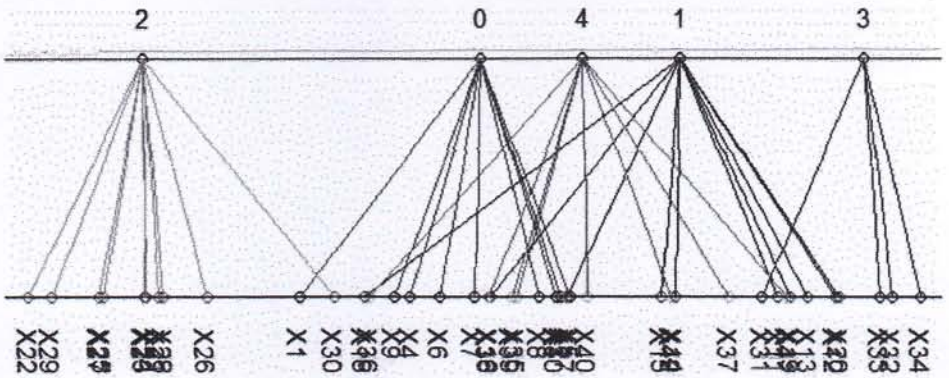

FIGURE 3. Between groups analysis of the brain dataset

5. **Conclusion.** There are many functions in MADE4 to visualize the results. The MADE4 package can accept a wide variety of gene expression data input formats had made it a simple to use tool for multivariate analysis. The integration of LLSimpute algorithm for the missing values imputation has been crucial in pre-processing of the datasets since incomplete datasets cannot be used for multivariate analysis. Moreover, with this added function, most of the datasets with missing values can be solved. For the visualization of the analysis, a simplest way to view results is to use plot functions. Apart from that, there are functions for drawing 1D and 3D plots for higher dimension analysis. Hence, it can be said that MADE4 has been the simplest tool for multivariate analysis of gene expression data.

Acknowledgments. This work is financed by Institutional Scholarship MyPhD provided by the Ministry of Higher Education of Malaysia. We also would like to thank Universiti Teknologi Malaysia for supporting this research by UTM GUP research grants (vot number: Q.J130000.7123.00H67 and Q.J130000.7107.01H29).

REFERENCES

[1] C. A. Culhane, J. Thioulouse and D. G. Higgins, MADE4: An R package for multivariate analysis of gene expression data, *Gene Expression*, vol.21, no.11, pp.2789-2790, 2005.

[2] R. Brereton, Chemometrics: Data analysis for the laboratory and chemical plant, in *Chichester*, 1st Edition, John Wiley & Sons Ltd, 2003.

[3] H. Kim, G. H. Golub and H. Park, Missing value estimation for DNA microarray gene expression data: Local least squares imputation, *Bioinformatics*, vol.21, no.2, pp.187-198, 2005.

[4] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson and P. S. Meltzer, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, *Nat. Med.*, vol.6, pp.673-679, 2001.

[5] S. Ramaswamy, K. N. Ross, E. S. Lander and T. R. Golub, A molecular signature of metastasis in primary solid tumors, *Nature Genctics*, vol.33, pp.49-54, 2003.

[6] S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander and T. R. Golub, Prediction of central nervous system embryonal tumour outcome based on gene expression, *Nature*, vol.415, pp.436-442, 2002.

[7] L. J. van't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards and S. H. Friend, Gene expression profiling predicts clinical outcome of breast cancer, *Nature*, vol.415, pp.530-536, 2002.

[8] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack and A. J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. of Natl. Acad. Sci.*, vol.96, no.12, pp.6745-6750, 1999.

[9] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander, Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, vol.286, no.5439, pp.531-537, 1999.

[10] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Losses, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Jr. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown and L. M. Staudt, Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, vol.403, no.6769, pp.503-511, 2000.

[11] D. T. Ross, U. Scherf, M. B. Eisen, C. M. Perou, C. Rees, P. Spellman, V. Iyer, S. S. Jeffrey, M. V. de Rijn, M. Waltham, A. Pergamenschikov, J. C. Lee, D. Lashkari, D. Shalon, T. G. Myers, J. N. Weinstein, D. Botstein and P. O. Brown, Systematic variation in gene expression patterns in human cancer cell lines, *Nature Genetics*, vol.24, no.3, pp.227-235, 2000.

[12] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub and W. R. Sellers, Gene expression correlates of clinical prostate cancer behaviour, *Cancer Cell*, vol.1, pp.203-209, 2002.