BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 17 2012, pages 2288–2290
doi:10.1093/bioinformatics/bts439

Data and text mining

Advance Access publication June 18, 2012

imDEV: a graphical user interface to R multivariate analysis tools in Microsoft Excel

Dmitry Grapov[1,2] and John W. Newman[1,2,*]

[1]Department of Nutrition, University of California and [2]Obesity and Metabolism Research Unit, USDA, ARS, Western Human Nutrition Research Center, Davis, CA 95616, USA

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Interactive modules for Data Exploration and Visualization (imDEV) is a Microsoft Excel spreadsheet embedded application providing an integrated environment for the analysis of omics data through a user-friendly interface. Individual modules enables interactive and dynamic analyses of large data by interfacing R's multivariate statistics and highly customizable visualizations with the spreadsheet environment, aiding robust inferences and generating information-rich data visualizations. This tool provides access to multiple comparisons with false discovery correction, hierarchical clustering, principal and independent component analyses, partial least squares regression and discriminant analysis, through an intuitive interface for creating high-quality two- and a three-dimensional visualizations including scatter plot matrices, distribution plots, dendrograms, heat maps, biplots, trellis biplots and correlation networks.

Availability and implementation: Freely available for download at http://sourceforge.net/projects/imdev/. Implemented in R and VBA and supported by Microsoft Excel (2003, 2007 and 2010).

Contact: John.Newman@ars.usda.gov

Supplementary Information: Installation instructions, tutorials and users manual are available at http://sourceforge.net/projects/imdev/.

Received on March 27, 2012; revised on July 21, 2012; accepted on July 5, 2012

1 INTRODUCTION

Omics experiments generate complex high-dimensional data requiring multivariate analyses. Although basic spreadsheets are widely used for data storage and low-level statistical analyses, these currently lack tools for multivariate analyses and visualization. In contrast, the R Project for Statistical Computing is a freely available software environment (R Development Core Team, 2011) that provides a variety of multivariate data analysis and visualization methods. Despite its power, the command-line interface of R is a barrier to its broad use. Through the application of RExcel (Baier and Neuwirth, 2007), Interactive modules for Data Exploration and Visualization (imDEV) unites the Microsoft Excel (MS Excel) spreadsheet and R, to provide a user-friendly graphical interface to multivariate analysis and visualization. imDEV also facilitates multivariate data interpretation by integrating dynamic data visualizations with univariate statistics, dimensional reduction methods, predictive modeling and network analyses tools.

* To whom correspondence should be addressed.

2 METHODS

2.1 Software

imDEV (http://sourceforge.net/projects/imdev/) is implemented in Visual Basic for Applications (VBA) and R programming languages and depends on the open source applications statconnDCOM (v3.1-2B7) and RExcel (v3.2.2) (http://rcom.univie.ac.at/) for integration as an add-in into MS Excel versions 2003, 2007 and 2010.

Pull down menus in MS Excel provide a modular interface organized by analysis type. Background processes translate user inputs from VBA to R, and R source scripts execute calculations and generate visualizations. Numerical outputs from R are returned to user-defined ranges or worksheets in MS Excel. Visualizations can be exported directly from the R plotting interface in a variety of file formats. User-defined inputs are stored as named ranges in MS Excel, enabling dynamic loading of R objects between analyses sessions.

2.2 Features

Interactive modules allow the user to rapidly progress from routine data analysis tasks to the generation of complex network representations of multivariate classification or prediction models. Data pre-treatment tools include variable-specific normality transformations, missing values imputation, centering and a variety of scaling methods. Hypothesis testing is supported by parametric and non-parametric comparisons of class means with multiple comparison adjustments, q-value calculations and tabular result outputs.

Analysis and visualization of variable correlations depend on the significance of correlation coefficients from user-selected parametric or rank-order tests (i.e. Pearson, Spearman and Kendall). Analysis of variable or sample relatedness is facilitated by hierarchal cluster analyses supported by a variety of distance and agglomeration methods, and aided by multivariate, group-specific visualizations of variable distributions, linear correlations, correlation matrix heat maps and dendrograms.

Exploratory data analyses are supported through dimensional reduction and projection pursuits. These include outlier-insensitive methods such as probabilistic and Bayesian principal component analyses (PCA) (Stacklies *et al.*, 2007) and independent components analysis, useful for separating analytical noise from biologically relevant information (Marchini *et al.*, 2010). Interpretation of data projections is aided by customizable, single to multiple-component, two- or three-dimensional biplots for model eigenvalues, scores and loadings.

Multivariate predictive modeling can be used to reduce data complexity through the segregation of informative from non-informative variables. Multivariate regression and classification are supported through an interface for fitting, validating and optimizing partial least squares (PLS) regression and classification models (Mevik *et al.*, 2011). Models can be fit using various algorithms including the Non-Linear Iterative Partial Least Squares (NIPALS) method (Wold, 1966), the Statistically Inspired Modification of the PLS (SIMPLS) method (De Jong, 1993) and

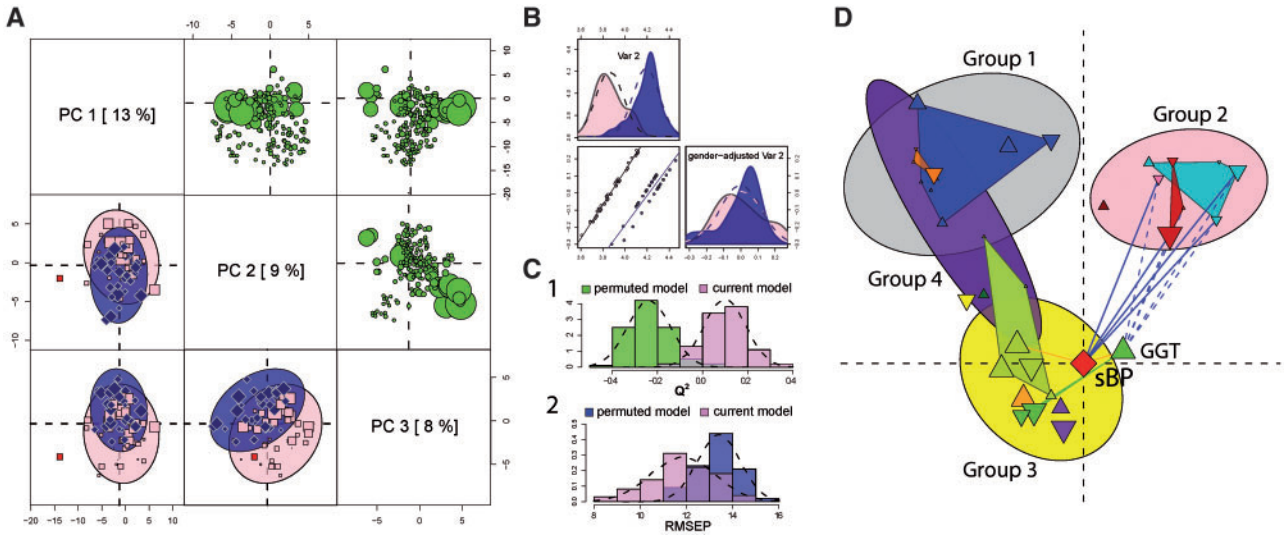

Fig. 1. Analysis of the relationship between circulating metabolite levels and systolic blood pressure (sBP). Exploratory PCA was used to identify gender-specific differences in metabolite concentrations and sBP. A PLS model was developed for the prediction of sBP given gender-adjusted metabolite concentrations. Network analysis of the PLS model parameters was used to highlight a previously known relationship between sBP and GGT, and identified a novel group of related metabolites (Group 2) that are negatively correlated with both sBP and GGT. (**A**) PCA scores and loadings trellis-plots. Scores (bottom left) where size indicates sBP, while color and shape indicate gender (pink squares, female; blue diamonds, male), with an outlier highlighted in red. Loading (top right) sizes indicate *P*-values from Mann–Whitney U-test for gender-specific differences in metabolite concentrations. (**B**) Variable distribution and scatter plot matrix displaying the effect of covariate adjustment for gender on a representative variable. (**C**) Comparison of gender-adjusted sBP PLS model Q^2 (**C1**) and RMSEP (**C2**) statistic distributions to their respective permuted null distributions. (**D**) A multi-dimensionally scaled PLS model correlation network visualizing correlations (Spearman's rho, $P < 0.05$) displayed by colored edges (orange, positive; blue, negative) between sBP (red diamond) and model parameters (triangles). Triangle (i.e. network vertex) characteristics encode PLS coefficient magnitude (size) and sign (upward, positive; inverted, negative). Major groups of correlated variables, defined by a hierarchal cluster analysis, are displayed by ellipses, and biologically related classes of metabolites are shown using similar vertex colors and polygons

various kernal methods. To avoid over-fitting, tools to compare results with permuted null models and training and test set-based validations are provided. Filter (Romanski, 2009), wrapper (Markowetz and Spang, 2005) and hybrid variable selection methods are implemented to optimize model performance, which is internally and externally cross-validated using user-defined or PCA optimized (Kennard and Stone, 1969) training and/or test splits of the original data. Tools for penalized feature selection methods (Goeman *et al.*, 2012; Kraemer and Boulesteix, 2011), e.g. L1 (lasso) and L2 (ridge), are under development.

Correlation networks provide an intuitive method to integrate and analyze relationships among data and metadata. User-defined, projection-based or multidimensional scaled graph layouts are used to generate undirected two- or three-dimensional weighted graphs, which through a dynamic edge drawing and vertex annotating interface can be used to explore, highlight or identify linear dependencies among variables. To limit spurious edges and visual complexity of correlation-driven networks, future expansions will implement methods for undirected Gaussian graphical Markov model generation (Castelo and Roverato, 2006).

3 RESULTS AND DISCUSSION

imDEV provides a visualization centric integrated environment for multivariate analyses based on projection pursuits, multivariate predictive modeling and network generation. To highlight the versatility of imDEV, a previously reported metabolomic dataset comprising serum free fatty acids, oxylipins and endocannabinoids (Psychogios *et al.*, 2011) was combined with associated clinical data (NIH Project Number 5R01HL076238-03, Oxidized Linoleic Acid, Aldosterone and Obesity) and analyzed to identify relationships between circulating metabolites and systolic blood pressure (sBP).

The relationships between metabolomic variables ($n = 174$) and sBP in men ($n = 30$) and women ($n = 38$) are explored (Fig. 1). Visual inspection of PCA scores and loadings (Fig. 1A) for the second and third components highlights gender-specific differences in metabolite and sBP measurements. Since females displayed lower sBP than their male counterparts (Mann–Whitney U test, $P < 0.05$), gender-independent predictors for sBP were explored. Variable-specific transformations (Box and Cox, 1964) were used to achieve normal distributions (D'Agostino and Stephens, 1986). Missing values were imputed using a 10 component probabilistic PCA model (Stacklies *et al.*, 2007). Pre-treated values were linearly adjusted for gender (Fig. 1B), centered, scaled to unit variance and used to fit a PLS predictive model for sBP. Hybrid, filter/subset, feature selection was used to reduced model complexity by 81%, compared with the full parameter model, while maintaining goodness-of-fit (Q^2) and root mean squared error of prediction (RMSEP) model statistics. The selected feature set/model was externally validated by comparing the distributions for its Q^2 (Fig. 1 C1) and RMSEP (Fig. 1 C2) statistics, based on 100 permutations of model training (2/3 of the samples) and testing (1/3 hold out set) procedures, to their respective null distributions, generated from 100 randomly permuted sBP models. The out of sample error rate defined by the RMSEP for the

developed model, containing 6 clinical and 27 metabolite parameters, was significantly lower than that expected by random chance (Fig. 1 C2). A correlation network (Fig. 1D) was used to interpret the model and study the interactions between its components. A hierarchal cluster analysis was used to define four major groups of correlated parameters, which are highlighted by ellipses in the multidimensional scaled defined network space (Fig. 1D). In this graph, vertex positions are defined by similarities in rank-order correlations and vertex shape and size by the sign and magnitude of the PLS model coefficients. Vertex and polygon colors are the same for biologically related species. Correlations between model parameters and sBP are shown by connecting edges. Systolic blood pressure is shown to be positively correlated with gamma glutamyl transferase (GGT) and negatively correlated with metabolite Group 2. GGT is a known independent risk factor for cardiovascular disease (CVD) and metabolic syndrome (Lee *et al*., 2007). Interestingly, metabolite Group 2, consisting of two classes of biologically related molecules, is negatively correlated with both sBP and GGT. Further investigation of theses species may provide sensitive markers for CVD risk and aid in the development of a mechanistic relationship between sBP, GGT and these circulating metabolites.

In summary, imDEV consists of a VBA-based graphical user interface that accesses R-based visualizations and statistical algorithms, which operate on data stored within and returned to named ranges or worksheets within Excel. This developed frame work for imDEV ensures that new analytical tools and methods can be incorporated into future interactive modules. The developed software interfacing exploratory, predictive and networking tasks allows for a rapid and dynamic multivariate analysis workflow that focuses on understanding the data at a systems level, aiding its interpretation within a biological context.

ACKNOWLEDGEMENTS

We thank Dr. Tobias Kind for editorial support for this manuscript and acknowledge the exceptional work of the R Development Core Team, authors of the R community contributed packages used in imDEV, the RExcel and statconn developers and imDEV early adopters, who have made this project possible.

Funding: This work was funded in part by NIGMS-NIH T32-GM008799, NIH-NIDDK R01DK078328-01, and USDA-ARS intramural Project 5306-51530-016-19D. The USDA is an equal opportunity provider and employer.

Conflict of Interest: none declared.

REFERENCES

Baier,T. and Neuwirth,E. (2007) Excel :: COM :: R. *Comput. Stat.*, **22**, 91–108.

Box,G.E.P. and Cox,D.R. (1964) An analysis of transformations. *J. Roy. Stat. Soc.*, **26**, 211–252.

Castelo,R. and Roverato,A. (2006) A robust procedure for Gaussian graphical model search from microarray data with *p* larger than *n*. *J. Mach. Learn. Res.*, **7**, 2621–2650.

D'Agostino,R.B. and Stephens,M.A. (eds.) (1986) *Goodness-of-Fit Techniques*. Marcel Dekker, New York.

De Jong,S. (1993) SIMPLS: an alternative approach to partial least squares regression. *Chemometr. Intell. Lab.*, **18**, 251–263.

Goeman,J.J. *et al.* (2012) *Penalized: L1 (lasso and fused lasso) and L2 (ridge) Penalized Estimation in GLMs and in the Cox Model.*. R Foundation for Statistical Computing, Vienna.

Kennard,R.W. and Stone,L.A. (1969) Computer aided design of experiments. *Technometrics*, **11**, 137–148.

Kraemer,N. and Boulesteix,A.-L. (2011) *ppls: Penalized Partial Least Squares*. R Foundation for Statistical Computing, Vienna.

Lee,D.S. *et al.* (2007) Gamma glutamyl transferase and metabolic syndrome, cardiovascular disease, and mortality risk: the Framingham Heart Study. *Arterioscler. Thrombo. Vasc. Biol.*, **27**, 127–133.

Marchini,J.L. *et al.* (2010) *fastICA: FastICA Algorithms to Perform ICA and Projection Pursuit*.. R Foundation for Statistical Computing, Vienna.

Markowetz,F. and Spang,R. (2005) Molecular diagnosis. Classification, model selection and performance evaluation. *Methods Inf. Med.*, **44**, 438–443.

Mevik,B.-H. *et al.* (2011) *pls: Partial Least Squares and Principal Component regression*.. R Foundation for Statistical Computing, Vienna.

Psychogios,N. *et al.* (2011) The human serum metabolome. *PLoS One*, **6**, e16957.

R Development Core Team*R: A Language and Environment for Statistical Computing*.. R Foundation for Statistical Computing, Vienna.

Romanski,P. (2009) *FSelector: Selecting Attributes*.. R Foundation for Statistical Computing, Vienna.

Stacklies,W. *et al.* (2007) pcaMethods—a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, **23**, 1164–1167.

Wold,H. *et al.* (2007) Estimation of principal components and related models by iterative least squares. In Krishnaiah,P.R. (ed.) *Multivariate Analysis*. Academic Press, NY, pp. 1164–1167.