Analytical Biochemistry 433 (2013) 102–104

Contents lists available at SciVerse ScienceDirect

Analytical Biochemistry

journal homepage: www.elsevier.com/locate/yabio

Notes & Tips

Utilities for quantifying separation in PCA/PLS-DA scores plots

Bradley Worley, Steven Halouska, Robert Powers *

Department of Chemistry, University of Nebraska-Lincoln, NE 68588-0304, USA

ARTICLE INFO

Article history:
Received 21 August 2012
Received in revised form 6 October 2012
Accepted 6 October 2012
Available online 15 October 2012

Keywords:
PCA
PLS-DA
MVA
UPGMA
Metabolomics

ABSTRACT

Metabolic fingerprinting studies rely on interpretations drawn from low-dimensional representations of spectral data generated by methods of multivariate analysis such as principal components analysis and projection to latent structures discriminant analysis. The growth of metabolic fingerprinting and chemometric analyses involving these low-dimensional scores plots necessitates the use of quantitative statistical measures to describe significant differences between experimental groups. Our updated version of the PCAtoTree software provides methods to reliably visualize and quantify separations in scores plots through dendrograms employing both nonparametric and parametric hypothesis testing to assess node significance, as well as scores plots identifying 95% confidence ellipsoids for all experimental groups.

A trademark of metabolomics experiments—more specifically metabolic fingerprinting and nontargeted metabolic profiling studies—is the use of multivariate analysis techniques, most commonly principal components analysis (PCA) and projection to latent structures discriminant analysis (PLS-DA) [1,2]. While these techniques provide low-dimensional representations of complex datasets through visually interpretable scores plots, the task of inferring biologically relevant conclusions from scores plots has been largely based on subjective examinations by expert users. Correspondingly, the continued growth in metabolomics and the associated application of chemometric analysis have created a strong need for a quantitative means to justify conclusions drawn from these scores plots. Toward this goal, we recently described the application of our PCAtoTree software to generate metabolic tree diagrams from scores plots and the use of standard bootstrapping techniques to infer the statistical significance of each resulting tree node [3]. This note presents a new set of portable software tools that enhance and improve upon our original methodology. Our updated version of the PCAtoTree software provides quantification of scores–space separation using both nonparametric bootstrapping and multivariate Hotelling's T^2 hypothesis testing to generate easily interpretable dendrograms of differences between experimental groups. Notably, the new software is now stand-alone and no longer dependent on PHYLIP (http://www.phylip.com/) [4].

Scores plots generated from unsupervised PCA or supervised PLS-DA methods provide visualizable representations of information-rich spectral data by means of dimensionality reduction. In the case of PCA, orthogonal lines of maximum gross variation are found within the data, termed the "principal axes", onto which the input data are transformed [5]. This operation preserves as much original gross variation as possible in the first few transformed dimensions and reveals separations between experimental groups only when within-group variability is sufficiently less than between-group variability. Alternatively, PLS-DA is a supervised method that guides this transformation informed by between-group variability to better reveal group structure [6,7]. In any case, the resultant two- or three-dimensional scores plot is used to identify spectral features contributing to between-group variability based on separations observed between groups in the scores plot.

The importance placed on interpretation of PCA and PLS-DA scores plots necessitates the use of quantitative procedures to determine the significance of these group separations. However, no de facto protocol or metric exists to provide a means of reporting the degree or significance of cluster separation [3,8,9]. Anderson et al. used the J_2 criterion [10,11] to assess the quality of resulting scores clusters according to the average within-group and between-group scatters for all groups. However, the J_2 metric provides only an overall estimation of cluster separation without fine-grained information on each pair of groups [11]. A similar problem exists with the related Davies–Bouldin index [12], which chooses a worst-case estimate of cluster overlap as its figure of merit. Dixon et al. [13] also comprehensively reported the performances of four cluster separation indices based on modifications of metrics used to validate separation for unsupervised clustering algorithms. Alternatively, our PCAtoTree protocol constructs dendrograms from distance matrices based on PCA scores for the PHYLIP software suite using a bootstrapping routine to determine node significance [3,4]. However, it was recently shown that hypothesis testing using a Mahalanobis distance metric and the

* Corresponding author. Fax: +1 402 472 9402.
E-mail address: rpowers3@unl.edu (R. Powers).

0003-2697/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ab.2012.10.011

T^2 and F distributions can provide a statistical means to infer cluster similarity [8], suggesting the possibility of returning p values for full statistical quantitation of PCA group separations.

Methods

The methods described below were implemented in software using the C programming language with minimal external dependencies, so the programs may be compiled and executed on any modern GNU/Linux distribution.

Probability calculation

Under the assumption that each group in the scores space is distributed as a multivariate normal random variable, the distances between groups may be calculated using the squared Mahalanobis distance metric [14],

$$D_M^2 = (\boldsymbol{u}_j - \boldsymbol{u}_i)^T \boldsymbol{S}_p^{-1} (\boldsymbol{u}_j - \boldsymbol{u}_i).$$

Here, $\boldsymbol{u}_i$ and $\boldsymbol{u}_j$ are the p-variate sample means of groups i and j, respectively, and $\boldsymbol{S}_p$ is the pooled p-by-p variance–covariance matrix, a weighted average of the covariance matrices from groups i and j. The Mahalanobis distance may then be related to a Hotelling's T^2 statistic by the scaling [15].

$$T^2 = \left(\frac{n_i n_j}{n_i + n_j}\right) D_M^2,$$

where n_i and n_j are the number of data points in groups i and j, respectively. This T^2 statistic is an extension of the Student t statistic to hypothesis tests in multiple dimensions and can be related to an F distribution by a final scaling [15]:

$$x_F = \frac{n_i + n_j - p - 1}{p(n_i + n_j - 2)} T^2 \sim F(p, n_i + n_j - p - 1).$$

It can be seen from this final relation that evaluation of the complement of the cumulative F-distribution function at x_F yields the p value for accepting the null hypothesis: the points in groups i and j are in fact drawn from the same multivariate normal distribution.

Tree generation

The implementation of the tree-generation procedure is a classical UPGMA algorithm [16]. When p values are reported at each branch point, a single tree is generated based on the matrix of Mahalanobis distances between groups. In the case of bootstrapped trees, the groups are randomly resampled with replacement while preserving group size. The desired number of trees is then generated using Euclidean distances between group means. The final tree used to report bootstrap probabilities is built using a Euclidean distance matrix calculated from the original (non-resampled) dataset.

Confidence ellipse calculation

When viewing PCA and PLS-DA scores plots, it is common practice to apply hand-drawn ellipses to inform group membership or even to omit such ellipses entirely. This may lead to inconsistent or erroneous interpretation of experimental results. Instead, the fact that the Mahalanobis distances of a set of p-variate points from their sample mean follow a χ^2 distribution having p degrees of freedom [17] may be leveraged to estimate 95% confidence ellipsoids for scores in any number of dimensions. The sample mean $\boldsymbol{u}$ and covariance matrix $\boldsymbol{S}$ for each group must first be calculated from its scores space data. Then, the group covariance matrix is decomposed into its eigenvalues and eigenvectors,

$$\boldsymbol{S} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1},$$

where $\boldsymbol{Q}$ is a p-by-p matrix whose columns are the eigenvectors of $\boldsymbol{S}$, and $\boldsymbol{\Lambda}$ is a diagonal matrix of the corresponding eigenvalues of $\boldsymbol{S}$. For the case of two-dimensional scores data, the 95% confidence ellipse for the group follows,

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \boldsymbol{u} + \boldsymbol{Q}\sqrt{\boldsymbol{\Lambda} F_{0.95,2}^{-1}} \begin{bmatrix} \cos t \\ \sin t \end{bmatrix},$$

where $F_{0.95,2}^{-1}$ is the value of the inverse χ^2 cumulative distribution function at $\alpha = 0.05$ and 2 degrees of freedom, and the square root is taken element-wise over $\boldsymbol{\Lambda}$. Similarly, a three-dimensional (3D) confidence ellipsoid may be obtained from the parametric equation

$$\begin{bmatrix} x(u,v) \\ y(u,v) \\ z(u,v) \end{bmatrix} = \boldsymbol{u} + \boldsymbol{Q}\sqrt{\boldsymbol{\Lambda} F_{0.95,3}^{-1}} \begin{bmatrix} \cos u \cos v \\ \cos u \sin v \\ \sin v \end{bmatrix},$$

where the parameters t, u, and v are all evaluated on $(0,2\pi)$. These methods allow for the inclusion of confidence regions onto two- and three-dimensional scores plots that reflect the 95% membership boundaries for each group. The approach assumes normally distributed data. The approach assumes normally distributed data. Fig. 1(a) and Supplemenal Fig. S1 illustrate the inclusion of these group confidence regions in representative PCA and OPLS-DA scores plots [18,19]. The ellipses and ellipsoids clearly define statistically significant class separation and also provide an example in which multiple groups actually belong to the same biological classification.

Discussion

Our updated and enhanced PCAtoTree software package consists of a set of stand-alone C programs that generate dendrograms from PCA/PLS-DA scores, report p values and bootstrap numbers, and incorporate confidence ellipse/ellipsoids into scores plots. The p values reported for every pair of distinct groups in a PCA/PLS-DA scores plot provide a truly quantitative means to discuss group separations. We also included support for the generation of dendrograms that use these p values at each branch point to address the question of tree uniqueness. This eliminated the prior dependency on PHYLIP [4]. The reporting of p values is complementary to bootstrapping methods in cases of highly overlapped groups, in that it provides a more direct, interpretable quantitation of group separation.

The PCAtoTree software package now uses Mahalanobis distances because this metric is more appropriate for multivariate data. De Maesschalck et al. [20] provide an exceptional introduction to the use of Mahalanobis distances with PCA. Specifically, Mahalanobis distances account for different variances in each direction (PC1, PC2, PC3) and are scale-invariant. Moreover, the use of a Mahalanobis distance metric for dendrogram generation includes cluster shape and orientation in the analysis of group separation. Also, Mahalanobis distances calculated between groups in PCA scores space will closely approximate those calculated on the original data while avoiding possible collinearity of the original variables. This is not true of Mahalanobis distances in PLS-DA scores space, because of the underlying supervision of PLS. These features differ from the Euclidean metric, which is a special case of the Mahalanobis metric with the group covariance matrices equaling the identity. Fig. 1(b) illustrates the dendrogram structure based on the use of Mahalanobis distances determined a set of scores. Supplemental Fig. S2 shows the dendrogram structure based on Euclidean distances from the same scores.

It is important to note that our software is not a means of inferring the reliability of PCA or PLS-DA models, but only a toolset for quantifying the scores that those models produce. In the case of PCA scores, significance of the principal components used must be inferred based on the explained sum of squares or another cross-validation technique [21,22]. PLS-DA models require rigorous cross-validation to ensure model reliability, as they almost always yield perfect separations between the scores of different groups [23]. With that in mind, separations between groups not under

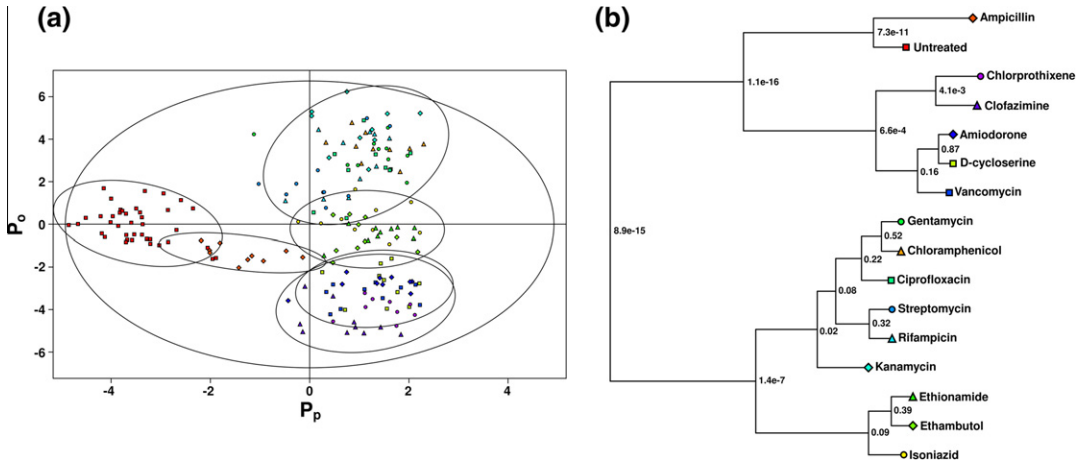

Fig.1. (a) 2D OPLS-DA scores plot illustrating 95% confidence ellipses for data having one predictive and one orthogonal PLS component. The symbol shape and color of each point correspond to the groups in (b). Discrimination in the first component is between wild-type and antibiotic-treated *Mycobacterium smegmatis*, and separations along the second component indicate metabolic differences between various antibiotic treatments. The antibiotics cluster together based on a shared biological target (cell wall synthesis, mycolic acid biosynthesis, or transcription, translation and DNA supercoiling). Three compounds of unknown *in vivo* activity were shown to clustered together with inhibitors of cell wall synthesis, implying a potential biological target. Interestingly, the *M. smegmatis* strain is resistant to ampicillin resulting in the ampicillin-treated cells clustering closer to untreated cells. The ellipses define the statistical significance of class separation and provide an illustration where two groups actually belong to the same biological classification. (b) Dendrogram generated from scores in (a) using Mahalanobis distances, with *p* values for the null hypothesis reported at each branch.

discrimination may be due to true experimental differences in PLS-DA scores plots, as opposed to the forced separations between discriminated groups. Thus, interpretation of the results of our PCAtoTree software must be done with the knowledge of the underlying algorithm's mathematical intent and only after the model has been validated. While we demonstrated our software using only 2D and 3D scores plots, our software places no restrictions on the number of components or on which components are used during dendrogram generation and *p* value calculation. Any dimensionality or choice of scores may be used with our PCAtoTree software provided all components are suitably validated.

Our updated and enhanced PCAtoTree software package provides a novel means of quantifying and visualizing separation significance in PCA and PLS-DA scores plots. Importantly, our new software enables single-step methodologies for generating informative scores plots and dendrograms of experimental groups in *any* study utilizing PCA or PLS-DA to elucidate group structure in complex datasets, including metabolic fingerprinting and nontargeted metabolic profiling. The tools are distributed under version 3.0 of the GNU General Public License and are freely available at http://bionmr.unl.edu/pca-utils.php.

Acknowledgments

The authors acknowledge Teklab Gebregiworgis, Bo Zhang, and Shulei Lei for their generous contributions of representative PCA and OPLS-DA scores plots used to develop and test the updated PCAtoTree software. This work was supported in part by funds from the National Institutes of Health (RO1 AI087668, R21 AI087561), the NIH National Center for Research Resources (P20 RR-17675), the America Heart Association (0860033Z), and the Nebraska Research Council. The research was performed in facilities renovated with support from the National Institutes of Health (RR015468-01).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ab.2012.10.011.

References

[1] T. Gebregiworgis, R. Powers, Application of NMR metabolomics to search for human disease biomarkers, Comb. Chem. High Throughput Screening 15 (2012) 595–610.
[2] B. Zhang, R. Powers, Using NMR-based metabolomics to study the regulation of biofilm formation, Future Med. Chem. 4 (2012) 1273–1306.
[3] M.T. Werth, S. Halouska, M.D. Shortridge, B. Zhang, R. Powers, Analysis of metabolomic PCA data using tree diagrams, Anal. Biochem. 399 (2010) 58–63.
[4] J.D. Retief, Phylogenetic analysis using PHYLIP, Methods Mol. Biol. 132 (2000) 243–258.
[5] I.T. Jolliffe, Principal Component Analysis, Springer, New York, 2002.
[6] M. Barker, W. Rayens, Partial least squares for discrimination, J. Chemom. 17 (2003) 166–173.
[7] S. Wold, M. Sjostrom, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab 58 (2001) 109–130.
[8] A.M. Goodpaster, M.A. Kennedy, Quantification and statistical significance analysis of group separation in NMR-based metabonomics studies, Chemom. Intell. Lab 109 (2011) 162–170.
[9] A.M. Goodpaster, L.E. Romick-Rosendale, M.A. Kennedy, Statistical significance analysis of nuclear magnetic resonance-based metabonomics data, Anal. Biochem. 401 (2010) 134–143.
[10] P.E. Anderson, N.V. Reo, N.J. DelRaso, T.E. Doom, M.L. Raymer, Gaussian binning: a new kernel-based method for processing NMR spectroscopic data for metabolomics, Metabolomics 4 (2008) 261–272.
[11] K. Koutroumbas, S. Theodoridis, Pattern Recognition, Elsevier/Academic Press, Amsterdam, 2006.
[12] D.L. Davies, D.W. Bouldin, A cluster separation measure, IEEE Trans. Pattern Anal. Mach. Intell. 1 (1979) 224–227.
[13] S.J. Dixon, N. Heinrich, M. Holmboe, M.L. Schaefer, R.R. Reed, J. Trevejo, R.G. Brereton, Use of cluster separation indices and the influence of outliers: application of two new separation indices, the modified silhouette index and the overlap coefficient to simulated data and mouse urine metabolomic profiles, J. Chemom. 23 (2009) 19–31.
[14] P.C. Mahalanobis, On the generalized distance in statistics, Proc. Natl. Inst. Sci. India 2 (1936) 7.
[15] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Academic Press, San Diego, 1979.
[16] C. Sokal, C. Michener, A statistical method for evaluating systematic relationships, Univ. Kansas Sci. Bull. 38 (1958) 30.
[17] H. Hotelling, The generalization of Student's ratio, Ann. Math. Stat. 2 (1931) 360–378.
[18] N.V. Chaika, T. Gebregiworgis, M.E. Lewallen, V. Purohit, P. Radhakrishnan, X. Liu, B. Zhang, K. Mehla, R.B. Brown, T. Caffrey, F. Yu, K.R. Johnson, R. Powers, M.A. Hollingsworth, P.K. Singh, MUC1 mucin stabilizes and activates hypoxia-inducible factor 1 alpha to regulate metabolism in pancreatic cancer, Proc. Natl. Acad. Sci. USA 109 (2012) 13787–13792.
[19] S. Halouska, R.J. Fenton, R.G. Barletta, R. Powers, Predicting the in vivo mechanism of action for drug leads using NMR metabolomics, ACS Chem. Biol. 7 (2012) 166–171.
[20] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, The Mahalanobis distance, Chemom. Intell. Lab. 50 (2000) 1–18.
[21] H.T. Eastment, W.J. Krzanowski, Cross-validatory choice of the number of components from a principal component analysis, Technometrics 24 (1982) 73–77.
[22] W.J. Krzanowski, Cross-validation in principal component analysis, Biometrics 43 (1987) 575–584.
[23] K. Kjeldahl, R. Bro, Some common misunderstandings in chemometrics, J. Chemom. 24 (2010) 558–564.