Published in final edited form as:

Chemometr Intell Lab Syst. 2011 December 15; 109(2): 162–170. doi:10.1016/j.chemolab.2011.08.009.

Quantification and statistical significance analysis of group separation in NMR-based metabonomics studies

Aaron M. Goodpaster and Michael A. Kennedy*

Department of Chemistry and Biochemistry, Miami University, Oxford, OH 45056, USA

Abstract

Currently, no standard metrics are used to quantify cluster separation in PCA or PLS-DA scores plots for metabonomics studies or to determine if cluster separation is statistically significant. Lack of such measures makes it virtually impossible to compare independent or inter-laboratory studies and can lead to confusion in the metabonomics literature when authors putatively identify metabolites distinguishing classes of samples based on visual and qualitative inspection of scores plots that exhibit marginal separation. While previous papers have addressed quantification of cluster separation in PCA scores plots, none have advocated routine use of a quantitative measure of separation that is supported by a standard and rigorous assessment of whether or not the cluster separation is statistically significant. Here quantification and statistical significance of separation of group centroids in PCA and PLS-DA scores plots are considered. The Mahalanobis distance is used to quantify the distance between group centroids, and the two-sample Hotelling's T^2 test is computed for the data, related to an F-statistic, and then an F-test is applied to determine if the cluster separation is statistically significant. We demonstrate the value of this approach using four datasets containing various degrees of separation, ranging from groups that had no apparent visual cluster separation to groups that had no visual cluster overlap. Widespread adoption of such concrete metrics to quantify and evaluate the statistical significance of PCA and PLS-DA cluster separation would help standardize reporting of metabonomics data.

Keywords

PCA; PLS-DA; Scores plot; Metabonomics; Cluster separation; Statistical significance

1. Introduction

In general, metabonomics studies [1] rely on multivariate data analysis techniques to evaluate massive amounts of data. The two most widely used techniques in the literature are principal component analysis (PCA) [2] and partial least squares — discriminant analysis (PLS-DA) [3]. PCA is an *unsupervised* method that assesses variance across all observations in the raw data whereas in a *supervised* method like PLS-DA, a class discriminator, e.g. healthy sample versus diseased sample, is specified and used to maximize group separation according to class belonging. While PLS-DA tends to improve the separation between

*Corresponding author at: 106 Hughes Laboratories, Department of Chemistry and Biochemistry, Miami University, Oxford, OH 45056, USA. Tel.: +1 513 529 8267. kennedm4@muohio.edu (M.A. Kennedy).

groups compared to PCA, there is some risk that increased apparent separation can be an artifact of the PLS-DA algorithm and not reflect variances that truly distinguish between the groups [4].

While statistical validation of metabolite changes between groups identified by either PCA or PLS-DA is essential [5], examples exist in the metabonomics literature where metabolites are identified as changing between groups based on PCA or PLS-DA group separation in scores plots even though the visual separation between groups is questionable.

Unfortunately no standard metric has been introduced or widely adopted to quantify cluster separation and to assess the statistical significance of cluster separation in PCA and PLS-DA scores plots. If such standard protocols were widely adopted, it would standardize the reporting of data in the metabonomics literature and make the data easier to interpret.

Quantitative separation of group clusters in PCA and PLS-DA scores plots has been discussed infrequently in the literature; however, a few papers have attempted to address the issue. Fuzzy K-means clustering has been explored as a means of optimizing cluster separations to better classify class belonging of samples based on major phenotypical differences and minor phenotype subgroups observed in two different NMR datasets [6]. Another paper developed a novel method using what was defined as a PCA to Tree analysis that utilized bootstrapping techniques to improve the quantitative analysis of PCA clustering [7]. The PCA to Tree approach uses a phylogenetic algorithm to assess distance matrices resulting from various metabolic states which are organized into a phylogenetic-like tree format and a bootstrap algorithm is used to identify statistically relevant branch separations [7]. Anderson et al. used the J_2 criterion to determine the quality of the clusters which closely relates to the Davis–Bouldin index [8]. Dixon et al. provided the most extensive paper written on the subject of determining the separation in PCA scores plots [9]. In this paper they used simulated data to evaluate separation indices. The four indices investigated were the Davis–Bouldin index (DBI), silhouette width, modified silhouette width index and overlap coefficient.

While all the reports mentioned above explored valid ways to quantify group cluster separation in PCA and PLS-DA scores plots, none advocated reporting of a quantitative statistic to characterize cluster separation or reporting of group whether or not the cluster separations were statistically significant. Here we demonstrate the value of quantifying cluster separation in PCA and PLS-DA scores plots based on computation of the Mahalanobis distance between the centroids of the two cluster groups, and then the statistical significance of the cluster separation is assessed by calculating the Hotelling's T^2 two-sample statistic, relating this statistic to an F-value, and then applying an F-test. The approach is demonstrated using four experimental data sets that range from exhibiting no visual cluster separation to having complete visual cluster separation.

2. Material and methods

2.1. Datasets

Four datasets from previous metabonomics investigations performed in our lab were chosen for this study. These four datasets contained different amounts of apparent separation in the

PCA scores plot based on qualitative visual inspection. The four datasets were initially qualitatively classified as having *total separation*, *partial separation*, or *no separation*. The term *total separation* means that, based on visual inspection of the PCA scores plot, none of the points from one group overlapped with the points of the second group, at the 95% confidence interval. The term *partial separation* means that, while based on visual inspection of the PCA scores plot there were two distinct clusters of points for the two different groups being compared, some points from the first group mixed with the points from the second group, and some points from the second group mixed with the points from the first group, at the 95% confidence interval. Finally, the term *no separation* means that, based on visual inspection of the PCA scores plot, there was no distinct clustering of the points for the two groups and the points from the two groups appeared to be evenly intermixed. One dataset fell into the *total separation* category (17 controls and 24 treated) and one dataset fell into the *no separation* category (24 controls and 23 treated) while two datasets fell into the *partial separation* category (set #1: 14 controls and 19 treated; set #2: 14 controls and 19 treated). The *total separation* case was taken from the fecal dataset from the paper by Romick-Rosendale et al. [10]. The other three datasets were from unpublished studies.

2.2. Data collection

All samples were collected and frozen until NMR data collection. All samples were thawed on ice, buffered using a phosphate buffer, and centrifuged for 5 min at 10,000×g before 600 µL of the sample was placed into a 5 mm NMR tube. All NMR experiments were carried out on a Bruker Avance™III spectrometer operating at 850.10 MHz ¹H frequency and equipped with a room temperature 5 mm TXI triple resonance probe with inverse detection and controlled by the TopSpin 2.1.4 console software package (Bruker, Germany). All experiments were conducted at 293 K. All data was collected using a spectral width of 20.0 ppm. A standard one-dimensional (1D) presaturation experiment (zgpr) was run on all samples to assure that water presaturation (used to increase the dynamic range for signal detection) and shimming (required to minimize signal overlap) met certain specifications to collect reliable data. The shimming was considered to have met our specification when the full width at half height of the internal standard, trimethylsilyl propionate, was <1.0 Hz. For the *total separation* and one *partial separation* datasets, a 1D first increment of a NOESY (noesygppr1d) experiment was collected and for the *no separation* and the other *partial separation* dataset a CPMG (cpmgpr1d) experiment was collected. All experiments included on-resonance presaturation of the water peak achieved by irradiation during a recycle delay of 4.0 s with a pulse power of 54.89 dB and 51.04 dB for the NOESY and CPMG experiments, respectively. The 90° pulse width for every sample was determined using the automatic pulse calculation feature in TopSpin. The pulse widths varied between 9.5 and 13.0 µs depending on the different types of samples used for the different datasets. All experiments were run with 4 dummy scans to ensure a steady-state of recovered magnetization and 65K data points per spectrum. The number of transients and acquisition time varied between the four datasets depending on the type of sample that was being run. These parameters were optimized using standard procedures and would not have any influence on the cluster analysis algorithms introduced and demonstrated below.

2.3. Principal component analysis

All data were phase corrected and baseline corrected in TopSpin using the AU, apk0.noe macro. This assured that every spectrum was processed in exactly the same way to eliminate processing differences. Each dataset was then subjected to PCA using AMIX 3.9.7 (Bruker Biospin, GmbH). NMR spectra were binned into 0.03 ppm-wide buckets, except for the dataset that showed no separation, which was binned into 0.01 ppm wide buckets. All datasets were binned over the region δ 10.0 to 0.15 ppm. The region around the water peak for each dataset was excluded. This excluded region, which varied slightly for each dataset but in general was in the region of δ 4.5–5.0, was excluded to eliminate the effects of imperfect water suppression.

2.4. Quantification of PCA scores plot separation and statistical significance analysis

Quantification of the separation of two clusters in PCA or PLS-DA scores plots reduced to a problem of measuring the distance between cluster centroids for groups or populations on the basis of two discriminator variables, the PC1 and PC2 scores for each observation, that corresponded to the coordinates for the projection of each observation onto the first two, or in general any two, principal components. Quantification of the separation of groups based on a single discriminant variable, e.g. x , reduces to a discussion of the difference in the mean values of x for the two groups being compared. In the more general case where one is comparing two groups characterized by two or more discriminant variables, it is necessary to compare the centroids for the groups instead of just the mean values of a single discriminator variable. The solution to this problem, which was introduced long ago, is to compute the Mahalanobis distance [11] to quantify the magnitude of the separation of the clusters in the PCA and PLS-DA scores plots. The Mahalanobis distance is defined as:

$$D_M(PC1, PC2) = \sqrt{d' C_w^{-1} d}$$

where d is the 1×2 Euclidian difference vector between the centroids for the two groups computed as $d = [\overline{PC1}(2) - \overline{PC1}(1), \overline{PC2}(2) - \overline{PC2}(1)]$ and C_w^{-1} is the inverse of the pooled variance–covariance matrix between the two groups.

In order to determine whether or not the cluster separation in the PCA or PLS-DA scores plot was statistically significant, a statistical test like the Students t -test, which is used to assess the statistical significance of the difference between the means of two groups characterized by a single discriminant variable, was needed to evaluate the separation of the two centroids of the two groups being compared. The Hotelling's two-sample T^2 statistic is such a test, which can be related to an F-value and subjected to a F-test to determine if the cluster separation observed in the PCA scores plot for the two groups was statistically significant. Lattin et al. outline in chapter 12 how the Hotellings T^2 statistic relates to a Student's t -test when testing for statistically significant differences when more than one discriminator variable is present [12]. In order to calculate the Hotelling's two-sample T^2 statistic, the PC1 and PC2 coordinates used to project each spectrum onto the scores plot were exported from AMIX into an Excel spreadsheet. The centroid, i.e. the average PC1 and PC2 coordinates, of each group was calculated in Excel. All other calculations were

performed using MATLAB. The Hotelling's two-sample T^2 statistic was calculated by using the following equation:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} d' C_W^{-1} d$$

where n is the number of samples in each group, and C_W and d are defined as above. The T^2 statistic increases with increasing distance between the two group centroids in the PCA scores plot and with decreasing *within group* variance. Once the T^2 statistic was computed, it was converted to an F-value, which was then assessed using an F-test. Application of the F-test required computation of a F-statistic, which is the ratio of the *between group* variance to that of the *within group* variance, a function that follows a Fisher's F-distribution, and then the F-test is executed by comparing the F-value to the critical F-value, i.e. the value of F at a specified confidence level $(1-\alpha)$ on the F-distribution function. If the F-value is greater than the critical F-value, then the null hypothesis, which assumes that there is no separation between the groups, can be rejected, and it can be concluded that there is statistically significant separation between the groups. The relation between the Hotellings T^2 statistic and the F-value, or F-statistic, is defined as:

$$\frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} T^2 = F(p, n_1 + n_2 - p - 1)$$

where p is the number of discriminator variables, e.g. in the case of a two-dimensional PCA scores plot the PC1 and PC2 scores represent two discriminant variables, or $p = 2$. The F-value was then compared to a table of F-critical values to determine if the cluster separation in the scores plot was statistically significant. The critical F-value was determined using the website <http://www.danielsoper.com/statcalc/calc04.aspx>. The server requires input of the numerator degrees of freedom, indicated by the first index of the F-value function in the equation above, and the denominator degrees of freedom, indicated by the second index in the F-value equation above, which refers to the total number of subjects minus the number of discriminant variables minus 1. The F-statistic was evaluated at a specified probability, α , and the server automatically calculated the critical F-value. All critical F-values were calculated using $\alpha = 0.05$.

2.5. One-dimensional plots of Euclidian distances between cluster centroids

To provide a simplified visualization of centroid separation and point scatter in the individual clusters, we generated plots depicting the distance between centroids compared to the distances between individual points within a cluster relative to the cluster centroids. In these plots, the distance between the two horizontal lines indicates the Euclidian distance of separation between the cluster centroids and each vertical line represents the Euclidian distance between of that point from its own group centroid, plotted centered on its own horizontal centroid line.

3. Results and discussion

Four different experimental data sets were investigated to demonstrate the value of using quantitative metrics to evaluate the magnitude and statistical significance of cluster separations in PCA and PLS-DA scores plots. The results were analyzed in the context of the statistical significance of the PCA loadings that drive cluster separation in the PCA scores plot according to an approach developed previously in our lab [5]. PCA and PLS-DA scores plot separations were assessed by comparing “no scaling” and Pareto scaling data pretreatment prior to PCA and PLS-DA. Pareto data scaling (see discussion in [13]) emphasizes variances of smaller features in data sets, in our case weaker peaks in the NMR spectra, by dividing the dataset, i.e. the bucket intensities, by the square root of the standard deviation of the bucket intensity, thus leading to a heavier weighting of the variances of smaller features in the dataset in the variance–covariance matrix and, consequently, in the eigenvectors, i.e. principal components, and the corresponding PCA loadings. In principle, because Pareto-scaling produces a different variance–covariance matrix, the principal components will change, resulting in a different cluster separation profile in the PCA scores plot, and a different distribution of loadings in the PCA loadings plot. Therefore, Pareto scaling has the potential to make it easier to visually detect potentially important biomarkers from the PCA loadings corresponding to metabolites at low concentrations in the samples. The four different datasets used to demonstrate our approach were classified as *total separation*, *partial separation #1*, *partial separation #2*, and *no separation*. The four datasets were evaluated as outlined in Section 2.4 to quantify the magnitude and statistical significance of cluster separation and correlated with the distribution of statistically significant PCA loadings.

Figs. 1 through 4 show PCA and PLS-DA scores and loadings plots for the four datasets analyzed in this study. Fig. 1 shows the results for the *total separation* case. It is obvious from visual inspection of the PCA scores plots that the clusters for these two groups completely separate with either no bucket scaling (Fig. 1A) or Pareto scaling (Fig. 1C). However, it is impossible to determine by visual inspection if the magnitude of cluster separation changes with Pareto scaling compared to no bucket scaling. Computation of the Mahalanobis distance (D_M) provided a convenient metric to quantitatively compare the magnitude of cluster separation. The D_M for the cluster separation in the PCA scores plot in Fig. 1A is 7.65 compared to a D_M of 7.66 for the PCA scores plot in Fig. 1C (Table 1). The fact that the Mahalanobis distances do not depend on the nature of the pretreatment of the data in this example is not surprising since there is no predictable dependence of the principal components on the relative size of the elements of the variance–covariance matrix after scaling the data by the square root of the standard deviation of each variable, i.e. bucket, prior to computing the variance–covariance matrix. Remember that without Pareto scaling, the diagonal elements of the variance–covariance are equal to the square of the standard deviation of the variable, and the off-diagonal elements of the matrix are equal to the product of the standard deviations of the two different variables that produce the off-diagonal element. PLS-DA, on the other hand, is intended to maximize cluster separation between groups. Therefore, application of PLS-DA to the same dataset, with no bucket scaling, was expected to increase the magnitude of cluster separation in the PLS-DA scores

plot. As expected, the D_M for the cluster separation following PLS-DA in this data set (Fig. 1E) increased by about 6% from 7.65 to 8.15 (Table 1).

The same type of analysis was applied to the other three datasets. In Fig. 2, the PCA and PLS-DA scores plots are shown for the “*partial separation #1*” dataset. The D_M for the cluster separation in this dataset (Fig. 2A) was 0.93 (Table 1), which was substantially smaller than that measured for the cluster separation in Fig. 1A. Pareto scaling of this dataset (Fig. 2C) resulted in no significant change in the magnitude of the cluster separation and a corresponding D_M of 0.91 (Table 1). PLS-DA of this dataset (Fig. 2E) resulted in a 71% increase in the D_M from 0.93 to 1.59 (Table 1). A similar trend was observed for the “*partial separation #2*” dataset. The D_M between PCA clusters calculated with no bucket scaling (Fig. 3A) was 1.38 (Table 1), and Pareto scaling of this dataset resulted in no substantial change in the magnitude of the cluster separation with a D_M of 1.42 (Table 1). PLS-DA of this dataset resulted in a 34% increase in the magnitude of cluster separation with the D_M increasing from 1.38 to 1.85 (Table 1). Interestingly, while it would be difficult to characterize the relative magnitude of cluster separations based on a qualitative visual assessment of the PCA scores plots calculated without bucket scaling for the two *partial separation* cases, the D_M provided a convenient quantitative metric to characterize the relative magnitude of cluster separations, indicating a 48% increase in the D_M for *partial separation #2* dataset ($D_M = 1.38$) compared to a $D_M = 0.93$ for the *partial separation #1* dataset. Finally, analysis of the *no separation* dataset (Fig. 4) resulted in a relatively small D_M of 0.21 (Table 1) for the PCA scores plot with no scaling (Fig. 4A), and a small change in this distance after Pareto scaling ($D_M = 0.40$) (Table 1), however a 666% increase in the magnitude of the cluster separation was observed after PLS-DA ($D_M = 1.40$) (Table 1). Interestingly, and perhaps not surprisingly, PLSDA had the strongest effect on the magnitude of cluster separation in a case where there was the weakest evidence of cluster separation using raw PCA.

Next we demonstrate use of a simple metric to determine whether or not the cluster separations were statistically significant. Consider the PCA scores plot shown in Fig. 1A. To assess whether or not the observed cluster separation was statistically significant, we first calculated a T^2 statistic for the cluster separation, which was 582.21 (Table 1), converted this T^2 statistic into an F-score, which was 283.64 (Table 1), and finally applied an F-test, using a critical F-value of 3.96. Comparison of the F-value for this data set (283.64) (Table 1) to the critical F-value (3.96) (Table 1) showed that the F-value was substantially larger than the critical F-value, indicating that the null hypothesis, i.e. that there was no separation between the clusters, could be rejected, and that there was less than a 5% probability that the observed cluster separation was due to a random occurrence.

To provide an alternative visualization of the cluster separation quality for each dataset that was not scaled prior to PCA, we generated a plot in which the distance between the horizontal lines represented the Euclidian distance between the two group centroids and the vertical lines centered on each horizontal line represented the Euclidian distance between each point and its own centroid (Fig. 5). One can see that in the *no separation* case there is virtually no overlap between the individual points of each group and the centroid of the other group (Fig. 5A).

Assessment of the statistical significance of the cluster separation of the *total separation* dataset with Pareto scaling pretreatment yielded a similar result compared to the analysis of the data with no scaling pretreatment, with the T^2 statistic for the cluster separation of 584.49 (Table 1) and an F-score of 284.75 (Table 1), which when subjected to the F-test using a critical F-value of 3.96, indicated statistically significant separation between the clusters. Application of the statistical test to the *total separation* dataset analyzed using PLS-DA resulted in a 14% increase in the T^2 statistic for the cluster separation (661.56) (Table 1) compared to simple PCA with a corresponding F-score of 322.30 (Table 1), which when subjected to the F-test using a critical F-value of 3.96, indicated statistically significant separation between the clusters. Note that the statistical significance of the variances in the bucket intensities [5] do not change as a result of analyzing the data using PLS-DA compared to PCA, although the magnitude of the cluster separation and the statistical significance of the cluster separation increases; this point will be discussed further below.

The procedure for the statistical significance analysis of the cluster separations was then applied to the other three datasets. Perhaps the most interesting cases are those for which the cluster separation is marginal, as is the case for the *partial separation #1* and *partial separation #2* datasets. Analysis of the *partial separation #1* dataset indicated statistically significant separation between the clusters (Table 1). The one-dimensional Euclidian distance plot for this data (Fig. 5B) indicated substantial overlap between the individual distances from one group to its own centroid with the centroid of the other group. Analysis of the same dataset using Pareto scaling indicated that the cluster separation was not statistically significant (Table 1). Finally, application of PLS-DA to this dataset resulted in a 602% increase in the T^2 statistic for the cluster separation compared to the simple PCA, and indicated a statistically significant separation between the clusters (Table 1). Application of the same analyses to the *partial separation #2* dataset indicated statistically significant separation between the clusters (Table 1). Consistent with the statistical analysis, the one-dimensional Euclidian distance plot for this data (Fig. 5C) showed less overlap between the individual distances from one group to its own centroid with the centroid of the other group compared to the *partial separation #1* dataset. Pareto scaling for this dataset resulted in a statistically significant separation between the clusters (Table 1). Analysis of the PLS-DA for this dataset resulted in an 80% increase in the T^2 statistic for the cluster separation, and indicated statistically significant separation between the clusters (Table 1). Finally, application of this analysis to the *no separation* dataset indicated that the separation between clusters was not statistically significant (Table 1). The one-dimensional Euclidian distance plot for this data (Fig. 5D) showed virtually complete overlap between the individual distances from one group to its own centroid with respect to the centroid of the other group. Analysis of the Pareto scaled data for this dataset indicated that the separation between clusters was not statistically significant (Table 1). PLS-DA of this dataset resulted in a 4600% increase in the T^2 statistic for the cluster separation indicating that the separation between clusters was statistically significant (Table 1). Clearly, the variances of the individual bucket intensities, which ultimately drive separation in the PCA scores plots and the corresponding statistical significance of the differences between the bucket means [5], were unaffected by the PLS-DA procedure, and this relationship will be elaborated upon below.

Finally, we analyze the quantification of the magnitude and the statistical significance of the cluster separation in the context of the statistical significance of the differences in the means of the bucket intensities that ultimately drive the separation of the clusters in the PCA scores plots. Here, we utilized a hybrid data representation scheme that combines traditional PCA loadings plots with heat-map color-coding of the loadings plot points according to p-score magnitude, and ultimately statistical significance [5]. This hybrid data representation scheme enables rapid visual assessment of the distribution of statistically significant loadings in the PCA loadings plots. Since this analysis provides information with regard to statistically significant points in the loadings plot, we applied the algorithm only to the *total separation* dataset, which contained the largest number of statistically significant buckets. The loadings plot points corresponding to the traditional PCA scores plot in Fig. 1A were distributed generally as one would expect (Fig. 1B) with the statistically significant loadings occurring away from the origin in the loadings plot. In order to use a quantitative metric to characterize the distribution of statistically significant loadings from the origin of the loadings plot, we computed the average Euclidian distance of each statistically significant loading from the origin and grouped the measurements into whether the loadings corresponded to strong, medium or weak features (peaks) in the data sets. The results for each calculation are summarized in Table 2. For the *total separation* PCA with no bucket scaling, the average distances were 0.2084, 0.1047, and 0.0484 for the strong, medium, and weak categories, respectively. Applying the same analysis to the Pareto scaling of this dataset resulted in a substantial redistribution of statistically significant loadings in the loadings plot that was easily detected upon visual inspection (Fig. 1D). This qualitative interpretation of the pattern redistribution was confirmed by the quantitative calculation of the average Euclidian distances from the origin for each group that indicated about an 11% decrease in the average distance from the origin of the loadings (0.1862) (Table 2) corresponding to the strongest features in the dataset, an 18% increase in the average distance of the loadings of the medium intensity features from the origin (0.1235) (Table 2), and a 176% increase in the distance of the statistically significant loadings of the weakest bucket features from the origin (0.0852) (Table 2). Finally, as expected, evaluation of the PLS-DA indicated that the distribution of the statistically significant loadings was not substantially different compared to the simple PCA, based on the average distance of the statistically significant loadings from the origin in each of the three categories (Table 2).

In closing, we have developed tools to quantify the magnitude of cluster separations in PCA scores plots, and to determine if the observed cluster separations are statistically significant. It is worthwhile, then, to briefly discuss, in a metabolic profiling context, what it means if two groups experience only *partial separation* or *no separation* in a PCA scores plot. In the case where *no separation* is evident in the PCA scores plot, this indicates that the principal components that account for the directions of greatest variance across the two data sets do not correspond to variables, i.e. buckets or ultimately metabolites, which differ consistently between the two groups or populations. Instead, the differences in bucket intensities that dominate the eigenvectors vary strongly in both datasets, rather than being strong in one dataset and weak in the other dataset. For example, imagine comparing a healthy group and a group with cancer. If there is no distinct difference in the metabolic profiles due to the presence of the cancer, the direction of greatest variance, i.e. the principal component or

eigenvector, may be driven by variations in diet or some other factor instead of the presence of cancer, and in this case, there would be no distinct clustering of the groups, e.g. healthy versus cancer. In this case, the principal components obtained from the PCA might reflect variances due to diet, age, gender, or some other factor, but not due to the presence or absence of cancer; so in this example, there would be no clustering according to group membership. Alternatively, if one observed only *partial separation* in the PCA scores plot, this could indicate that there are distinct patterns in at least some metabolite levels that distinguish between the two groups, however the patterns may be weak, or confounded by strong variations in metabolites due to other factors associated with individual variabilities.

4. Conclusions

Here, we have demonstrated the utility of applying simple metrics for quantification of cluster separations, and for assessment of the statistical significance of cluster separations, in PCA and PLS-DA scores plots. The methods invoke computation of a Mahalanobis distance to characterize the distance between cluster centroids in two-dimensional PCA and PLS-DA scores plots, and rely on calculation of a Hotelling's T^2 statistic, an associated F-value, and application of an F-test to determine the statistical significance of the cluster separation. We demonstrated the utility of applying these metrics using four datasets that contained varying degrees of cluster separation, based on qualitative visual inspection, ranging from datasets with *total separation* to datasets with *no separation*. For the case of *total separation*, the quantitative statistical evaluation confirmed that the cluster separation was strongly statistically significant, whereas in the case of *no separation*, the cluster separation was not statistically significant, as expected. The techniques demonstrated more value in cases where the cluster separation was difficult to judge based on visual inspection alone. This was nicely illustrated using the two datasets that had *partial separation*, where the statistical significance analysis indicated that one dataset exhibited statistically significant cluster separation while the cluster separation in the other dataset was demonstrated to be not statistically significant. It was also demonstrated that the Mahalanobis distance provided a convenient quantitative metric to describe the relative magnitude of the cluster separation in these more ambiguous cases compared to visual inspection alone. The technique was also used to quantify the increased magnitude and statistical significance of the cluster separations in PLS-DA scores plots, even though the quantity and distribution of statistically significant buckets did not change, also as expected. In another application of these techniques, our approach enabled a quantitative assessment of the redistribution of statistically significant loadings as a consequence of pretreating the data by Pareto scaling, which demonstrated how statistically significant buckets from weak features in the dataset moved farther from the origin compared to the case of no data pretreatment.

Currently, there are no widely adopted practices used to quantify and report cluster separation in PCA or PLS-DA scores plots, or to assess whether or not the cluster separation is statistically significant. The lack of such metrics makes it virtually impossible to compare cluster separations measured in independent studies, and/or to compare studies conducted in different laboratories. Moreover, lack of standard use of such metrics can cause confusion in the metabonomics literature when conclusions are drawn based on marginal or questionable separation in scores plots and potentially important metabolites are identified and associated

with ill-characterized cluster separations. Adoption of, and standard reporting using, the simple metrics introduced in this paper should facilitate comparison metabonomics studies conducted at different times and by independent laboratories, and should provide a more quantitative framework in which to discuss metabonomics data analysis in the literature in the future.

Acknowledgements

The data collection was conducted at the Ohio Biomedicine Center of Excellence in Structural Biology and Metabonomics at Miami University. The authors acknowledge Lindsey Romick-Rosendale for providing some of the raw NMR data used for the demonstration exercises. The work was funded by the National Institutes of Health National Cancer Institutes; Grant number: 1R15CA152985.

Abbreviations

PCA	principal component analysis
PLS-DA	partial least squares — discriminant analysis
NMR	nuclear magnetic resonance
NOESY	nuclear Overhauser spectroscopy
CPMG	Carr–Purcell–Meiboom–Gill.

References

1. Lindon J, Nicholson J, Holmes E, Everett J. Metabonomics: metabolic processes studied by NMR spectroscopy of biofluids. *Concepts Magn. Res.* 2000; 12:289–320.
2. Wold S, Esbensen K, Geladi P. Principal component analysis. *Chemometr. Intell. Lab.* 1987; 2:37–52.
3. Barker M, Rayens W. Partial least squares for discrimination. *J. Chemometr.* 2003; 17:166–173.
4. Westerhuis J, Hoefsloot H, Smit S, Vis D, Smilde A, van Velzen E, van Duijnhoven J, van Dorsten F. Assessment of PLS-DA cross validation. *Metabolomics.* 2008; 4:81–89.
5. Goodpaster AM, Romick-Rosendale LE, Kennedy MA. Statistical significance analysis of nuclear magnetic resonance-based metabonomics data. *Anal. Biochem.* 2010; 401:134–143. [PubMed: 20159006]
6. Cuperlovi -Culf M, Belacel N, Culf AS, Chute IC, Ouellette RJ, Burton IW, Karakach TK, Walter JA. NMR metabolic analysis of samples using fuzzy K-means clustering. *Magn. Reson. Chem.* 2009; 47:S96–S104. [PubMed: 19731396]
7. Werth MT, Halouska S, Shortridge MD, Zhang B, Powers R. Analysis of metabolomic PCA data using tree diagrams. *Anal. Biochem.* 2010; 399:58–63. [PubMed: 20026297]
8. Anderson P, Reo N, DelRaso N, Doom T, Raymer M. Gaussian binning: a new kernel-based method for processing NMR spectroscopic data for metabolomics. *Metabolomics.* 2008; 3:261–272.
9. Dixon S, Heinrich N, Holmboe M, Schaefer M, Reed R, Trevejo J, Brereton R. 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. Chemometr.* 2009; 23:19–31.
10. Romick-Rosendale LE, Goodpaster AM, Hanwright PJ, Patel NB, Wheeler ET, Chona DL, Kennedy MA. NMR-based metabonomics analysis of mouse urine and fecal extracts following oral treatment with the broad-spectrum antibiotic enrofloxacin (Baytril). *Magn. Reson. Chem.* 2009; 47:S36–S46. [PubMed: 19768747]
11. Mahalanobis PC. On the generalised distance in statistics. *Proc. Natl. Inst. Sci. India.* 1936; 2:49–55.

12. Lattin, JM.; Carroll, JD.; Green, PE. Analyzing Multivariate Data. Pacific Grove, CA: Thomson Brooks/Cole; 2003.
13. Noda I. Scaling techniques to enhance two-dimensional correlation spectra. J. Mol. Struct. 2008; 883–884:216–227.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 1.

Scores and loadings plots for the *total separation* dataset. (A) PC1 versus PC2 PCA scores plot calculated with no bucket-scaling pretreatment. The solid line is drawn between the centroids for each cluster. (B) PCA loadings plot corresponding to the PCA scores plot in A. The loadings are color-coded according to p-score as described in [5]. (C) PC1 versus PC2 PCA scores plot calculated with Pareto-scaling pretreatment. The solid line is drawn between the centroids for each cluster. (D) PCA loadings plot corresponding to the PCA scores plot in C. (E) t[1] versus t[2] PLS-DA X-scores plot calculated with no bucket-scaling pretreatment. (F) PLS-DA loadings plot corresponding to the PCA scores plot in E.

Fig. 2.

Scores and loadings plots for the *partial separation #1* dataset. (A) PC1 versus PC2 PCA scores plot calculated with no bucket-scaling pretreatment. The solid line is drawn between the centroids for each cluster. (B) PCA loadings plot corresponding to the PCA scores plot in A. The loadings are color-coded according to p-score as described in [5]. (C) PC1 versus PC2 PCA scores plot calculated with Pareto-scaling pretreatment. The solid line is drawn between the centroids for each cluster. (D) PCA loadings plot corresponding to the PCA scores plot in C. (E) $t[1]$ versus $t[2]$ PLS-DA X-scores plot calculated with no bucket-scaling pretreatment. (F) PLS-DA loadings plot corresponding to the PCA scores plot in E.

Fig. 3.

Scores and loadings plots for the *partial separation #2* dataset. (A) PC1 versus PC2 PCA scores plot calculated with no bucket-scaling pretreatment. The solid line is drawn between the centroids for each cluster. (B) PCA loadings plot corresponding to the PCA scores plot in A. The loadings are color-coded according to p-score as described in [5]. (C) PC1 versus PC2 PCA scores plot calculated with Pareto-scaling pretreatment. The solid line is drawn between the centroids for each cluster. (D) PCA loadings plot corresponding to the PCA scores plot in C. (E) $t[1]$ versus $t[2]$ PLS-DA X-scores plot calculated with no bucket-scaling pretreatment. (F) PLS-DA loadings plot corresponding to the PCA scores plot in E.

Fig. 4.

Scores and loadings plots for the *no separation* dataset. (A) PC1 versus PC2 PCA scores plot calculated with no bucket-scaling pretreatment. The solid line is drawn between the centroids for each cluster. (B) PCA loadings plot corresponding to the PCA scores plot in A. The loadings are color-coded according to p-score as described in [5]. (C) PC1 versus PC2 PCA scores plot calculated with Pareto-scaling pretreatment. The solid line is drawn between the centroids for each cluster. (D) PCA loadings plot corresponding to the PCA scores plot in C. (E) t[1] versus t[2] PLS-DA X-scores plot calculated with no bucket-scaling pretreatment. (F) PLS-DA loadings plot corresponding to the PCA scores plot in E.

Fig. 5.

One-dimensional plots of Euclidian distances between cluster centroids. In each plot, the centroid of one dataset was drawn as a horizontal line value with a y-value of zero and the centroid for second cluster was indicated as a horizontal line with the y-value indicating Euclidian distance of the separation between the two cluster centroids. The distance of each observation from the centroid of its own group is depicted as a vertical line centered on the centroid line. These one-dimensional plots illustrate the magnitude of separation between the two centroids of the two groups in relation to the scatter of observation to centroid distances in each group. All examples shown had no scaling data pretreatment. The plots are constructed from the: (A) *total separation* PCA scores plot in Fig. 1A, (B) *partial separation #1* PCA scores plot in Fig. 2A, (C) *partial separation #2* PCA scores plot in Fig. 3A and, (D) *no separation* PCA scores plot in Fig. 4A.

Table 1

Summary of Mahalanobis distances for cluster separations and Hotellings T^2 and F-test statistics for various datasets and pretreatment conditions.

	Mahalanobis distance	Two-sample T^2 statistic	F-value	Critical F-value	Significant?
<i>No scaling</i>					
Total separation	7.65	582.21	283.64	3.24	Yes
Partial separation #1	0.93	6.97	3.37	3.32	Yes
Partial separation #2	1.38	15.57	7.53	3.32	Yes
No separation	0.21	0.50	0.24	3.21	No
<i>Pareto scaling</i>					
Total separation	7.66	584.49	284.75	3.24	Yes
Partial separation #1	0.91	6.63	3.21	3.32	No
Partial separation #2	1.42	16.40	7.93	3.32	Yes
No separation	0.40	1.86	0.91	3.21	No
<i>PLS-DA</i>					
Total separation	8.15	661.56	322.30	3.24	Yes
Partial separation #1	1.59	20.30	9.82	3.32	Yes
Partial separation #2	1.85	27.96	13.53	3.32	Yes
No separation	1.40	22.99	11.24	3.21	Yes

Table 2

Summary of average Euclidian distance of loadings to the origin of the loadings plots grouped into categories of strong (>0.012 normalized intensity), medium (0.005 – 0.012 normalized intensity), and weak (<0.005 normalized intensity) features in the total *separation* NMR dataset. Normalized intensities are calculated by dividing bucket intensities by the total integrated spectrum intensity.

	Average distance from origin		
	No scaling	Pareto scaling	PLS-DA
Strong	0.2084	0.1862	0.1900
Medium	0.1047	0.1235	0.0916
Weak	0.0484	0.0852	0.0431