Analytical Letters

ISSN: 0003-2719 (Print) 1532-236X (Online) Journal homepage: www.tandfonline.com/journals/lanl20

Principal Component Analysis (PCA) Loading and Statistical Tests for Nuclear Magnetic Resonance (NMR) Metabolomics Involving Multiple Study Groups

Lin Jiang, Hunter Sullivan & Bo Wang

To cite this article: Lin Jiang, Hunter Sullivan & Bo Wang (2022) Principal Component Analysis (PCA) Loading and Statistical Tests for Nuclear Magnetic Resonance (NMR) Metabolomics Involving Multiple Study Groups, Analytical Letters, 55:10, 1648-1662, DOI: 10.1080/00032719.2021.2019758

To link to this article: https://doi.org/10.1080/00032719.2021.2019758

Published online: 18 Jan 2022.

ANALYTICAL LETTERS
2022, VOL. 55, NO. 10, 1648–1662
https://doi.org/10.1080/00032719.2021.2019758

Taylor & Francis
Taylor & Francis Group

BIOANALYTICAL

Principal Component Analysis (PCA) Loading and Statistical Tests for Nuclear Magnetic Resonance (NMR) Metabolomics Involving Multiple Study Groups

Lin Jiang[a], Hunter Sullivan[a], and Bo Wang[b]

[a]Division of Natural Sciences, New College of Florida, Sarasota, FL, USA; [b]Department of Chemistry, North Carolina A&T State University, Greensboro, NC, USA

ABSTRACT

Metabolomics is an interdisciplinary area that integrates knowledge of instrumentation, data science, and biochemistry. Metabolomics studies the changes in a large number of metabolites after various treatments using analytical platforms. However, the interpretation approaches have not been completely investigated. Principal component analysis (PCA) is an unsupervised method that describes high throughput metabolite data, which is different from supervised approaches such as partial least-squares discriminant analysis (PLS-DA) which frequently has overfitting problems. The interpretation of PCA loadings, especially for studies with multiple study groups, is not well developed for metabolomics. In this study, a new method was reported that integrates PCA loading values with the commonly used statistical *t*-test analysis to significantly improve the convenience and efficiency of interpretation. The method was demonstrated using practical studies of NMR metabolomics on the extracts from sea anemone that were treated with six atrazine concentrations. The results indicated that the approach is suitable for multiple groups of metabolomics for early-stage discoveries, such as low concentrations and potentially longitudinal studies. In summary, this methodology may be critical in studies such as environmental metabolomics with various stimuli factors where the data interpretation was previously incompletely developed.

ARTICLE HISTORY
Received 16 September 2021
Accepted 14 December 2021

KEYWORDS
Metabolomics; nuclear magnetic resonance (NMR); principal component analysis (PCA)

Introduction

Metabolomics is an approach that simultaneously studies the changes of a large number of metabolites in a biological specimen using analytical platforms (Atzori et al. 2012; Emwas et al. 2019). Both nuclear magnetic resonance (NMR) spectroscopy (Boroujerdi et al. 2009; Wang et al. 2015) and liquid chromatography-mass spectrometry (LC-MS) (Wilson et al. 2005; Sangster et al. 2006) have been widely used in metabolomics. NMR shows high reproducibility and limited sample preparation requirements compared to LC-MS, which is commonly used in health metabolomics (Markley et al. 2017). NMR is a nondestructive technique that has rare pollution problems and limited noise for

CONTACT Bo Wang bwang1@ncat.edu Department of Chemistry, North Carolina A&T State University, 1601 E. Market St, Greensboro, NC 27411, USA; Lin Jiang ljiang@ncf.edu Division of Natural Sciences, New College of Florida, 5800 Bay Shore Road, Sarasota, FL 34243, USA.

samples over a period of at least 6 months (Wang, Goodpaster, and Kennedy 2013), which is important for studies with multiple research groups, such as longitudinal studies. Due to the large number of metabolites using various instruments, statistical modelings such as principal component analysis (PCA) (Werth et al. 2010; Chawla 2011), partial least-squares discriminant analysis (PLS-DA) (Trygg, Holmes, and Lundstedt 2007; Gu et al. 2011), and other approaches such as random forest (Xi et al. 2014) and support vector machine (SVM) (Kumar et al. 2017) have been used for metabolomics. Factor analysis (FA) has been applied in solid-state NMR (Brus et al. 2011; Urbanova, Kobera, and Brus 2013) to investigate large size data sets but requires given common factors (Smilde et al. 2010). The quality control approaches (Kumar et al. 2020; Kumar et al. 2018) have also been developed to ensure the reliability of NMR metabolomics. As an unsupervised method, PCA is one of the most popular methods in metabolomics (Yata and Aoshima 2012) and is mainly used to describe the distribution of a large number of metabolites after dimensional reduction. PCA has been widely applied in metabolomics biomarker discovery studies in human diseases such as cancer (Li, Qiu, and Zhang 2016), diabetes (Choubey et al. 2020), and Alzheimer's disease (Ahmad and Dar 2018) and also in plant (Gadekallu et al. 2021; Yagmur and Gunes 2021) and environmental studies (Scheel et al. 2019). Although the limitation of PCA lies in the potential problems from in-group noise (Halouska and Powers 2006), it has been reported to be powerful in many studies including animal materials (Ahmadi et al. 2020) that have a relatively small number of samples. PLS-DA has shown to be powerful for classification and was reported to separate random groups with many features (Ruiz-Perez et al. 2020). However, there are issues such as overfitting (Westerhuis et al. 2008) or ill-performed cross-validation, which are difficult when the number of samples (observation) is small (Rodriguez-Perez, Fernandez, and Marco 2018). The interpretation of PCA is important, especially when the number of samples is small (<20) as in most animal studies (Mora-Ortiz et al. 2019a).

PCA loading has also been used in metabolomics (Hernandez-Bolio et al. 2021). However, the applications of PCA loadings are limited, and statistical significance studies such as the Welch's t-test (Wang, Goodpaster, and Kennedy 2013; Ni et al. 2019) are more common. However, t-test methods have limited information about the combinational effects of metabolites which also have family-wise error concerns (Wang, Goodpaster, and Kennedy 2013). The combination of PCA loading plot and combinational t-tests have been applied to NMR (Goodpaster, Romick-Rosendale, and Kennedy 2010), but application in metabolomics, especially when multiple study groups were involved, is still not well developed.

Various hypothesis-based approaches have been developed in multiple group studies which include analysis of variance (ANOVA) or a Kruskal-Wallis (KW) test (Elliott and Hynan 2011; Spicer, Salek, and Steinbeck 2017). ANOVA studies (Ametaj et al. 2010) are good at interpreting multiple groups of data, but the interpretation of the multidimensional information and the gradual metabolite changing is limited. While the significant difference of individual metabolites is important in metabolomics, early markers can only be observed as a pathway of metabolites, meaning modeling is important in early marker discoveries. Supervised machine learning methods are powerful in classifications, but methods such as PLS-DA have potential overfitting (Westerhuis et al.

2008), especially when the popular lab studies have limited sample numbers (Kumazoe et al. 2015; Mora-Ortiz et al. 2019b) for training and testing groups in machine learning models. In this study, the unsupervised PCA was used to analyze data and introduce a new way to represent the PCA loading plot, which may also be used in multiple group studies to discover the potential early metabolic biomarkers on sea anemones. Sea anemone such as *Exaiptasia diaphana (Rapp 1829)*, which is a relative of coral, reproduces rapidly and reliably and is recommended as a model organism for monitoring the marine environment and understanding healthy zooxanthellate cnidarian physiology (Trenfield et al. 2017). Therefore, this study used the metabolomic response to *E. diaphana* as an example (Jiang et al. 2021) to demonstrate the performance of the approach. This method provides an easy and visible approach for general end-users to process metabolomics data, especially for work with multiple study groups.

Methodology

Sample preparation

The study is demonstrated by the treatment of sea anemone species *Exaiptasia diaphana* samples with several doses of atrazine. The aqueous phase of the sea anemone extracts was analyzed using a Bruker Ascend 400 MHz high-resolution NMR with an Xpress autosampler (Jiang et al. 2021). The polar metabolites of *E. diaphana* were dissolved in deuterium water (D_2O) containing 0.1 M phosphate buffer and 0.5 mM trimethylsilylpropanoic acid (TSP). The NMR spectra were obtained in Amix 4.0 (Bruker BioSpin) and the 1D NOESY experiments (noesygppr1d) were used for all samples with 32 k increments, 64 scans, and a 4 s relaxation delay (d1). All NMR spectra were bucketed using a previously reported automated method (Wang, Maldonado-Devincci, and Jiang 2020). The processed data were normalized to the total peak intensity. Metabolite identification was carried out using Chenomx 8.6. The study includes a control and six treatment groups, with groups number Class 1 through Class 6 having increasing concentrations of atrazine. Each group contains 6 parallel samples.

Data analysis

Principal component analysis (PCA) was carried out in PLS-toolbox (Eigenvector research) including the PCA score plot, loading plot, and the combined plot. All data were mean-centered and Pareto scaled before PCA. The box plots were plotted in Matlab (R2020, Mathworks) and the Student's *t*-test (two tails) was calculated in Excel (Microsoft). Metaboanalyst was used for the ANOVA study. The applied approaches are briefly introduced.

PCA is a data visualization method that describes data with dimensional reduction. Both scores and loadings are generally used to describe the results (Wang, Goodpaster, and Kennedy 2013; Ruiz-Perez et al. 2020). The score plot shows the largest variance among the samples while the loadings represent the metabolites' contributions to the plot. Specifically, the score plot shows if the samples from different groups are separated, the loadings in the same direction of the score plot (Figure 1) showed a positive contribution to the group. Large loading numbers are more important to the score plot

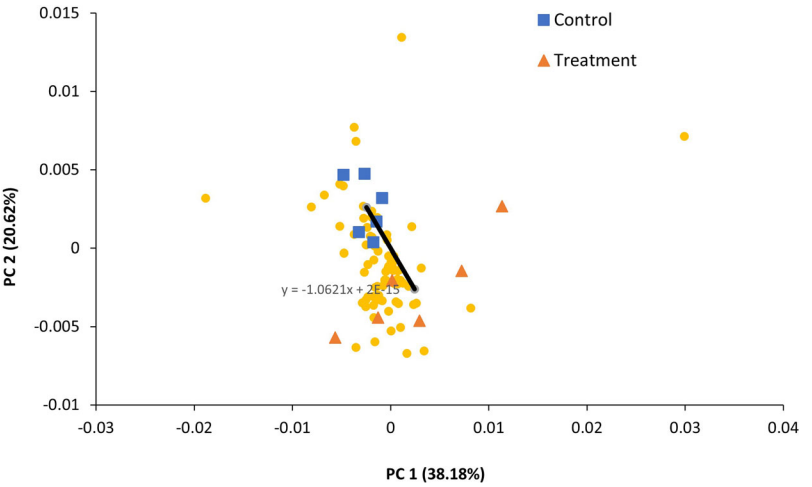

Figure 1. Example of a combined PCA score plot and loading plot. The blue squares and the orange triangles are the score plots, and each symbol represents one sample. The yellow dots are the loading plots and each dot represents one variable (metabolite). The black line represents the separation direction between the control and treatment groups. The loadings represent the correlation with the score plot when PCA was calculated using autoscaling. In this case, Pareto scaling was applied, so the correlation relationship was adjusted with the variance.

separation. Due to the noise and data reliability, Pareto scaling is usually recommended in NMR-based metabolomics. The Pareto approach used mean-centered data scaled to the square root of the standard deviation (van den Berg et al. 2006a).

Student's *t*-test uses the null hypothesis to compare the means of two groups (Zimmerman and Zumbo 1993). An assumption of no difference is made before calculation, and the probability of no difference between groups is represented by a *p*-value. A *p*-value smaller than 0.05 indicates the probability of the two groups being the same is 5% (95% different) (Wang, Goodpaster, and Kennedy 2013). Both Student's *t*-test and Welch's *t*-test are based on normal distributions, but the standard deviations of the two groups are considered to be the same in the former but different in the latter (Zimmerman and Zumbo 1993). In metabolomics, both Bonferroni corrections and false discovery rate (FDR) were used for multiple-comparison problems by using the critical *p* values smaller than 0.05 to reduce the false positive rate (Benjamini and Yekutieli 2001). The critical *p* values used in the Bonferroni correction were calculated by 0.05 divided by the number of variables. Critical *p* values were calculated for every variable in FDR after ranking *p* values from low to high. The critical *p*-value for one variable was calculated by 0.05 times the rank number and divided by the total variable number (Benjamini and Hochberg 1995).

Both analysis of variance (ANOVA) (St and Wold 1989) and the Kruskal-Wallis (KW) test (Acar and Sun 2013) were designed for studies with more than two groups. Similar to Student's *t* test, they compare the means between the groups and report if they are statistically significantly different. ANOVA has a normal distribution assumption for the data, but the Kruskal-Wallis test is a rank-based method without a distribution assumption (Odiase and Ogbonmwan 2005).

When a regression line was fit to scatter points, R^2 is generally calculated by the ratio of the variable variation explained by the model and the total variation (Howarth 2017). Higher R^2 values are expected in better models.

Method description

The PCA score plot and loading plot were first prepared in a combined figure. Due to the scales of the score values and loading values being different, the combined score and loading were plotted to fit both data together using the PLS toolbox (Eigenvector Research Inc). To convert the PCA score plot to the scale of the PCA loading plot, each PCA score value was divided by the maximum score absolute value to obtain a percentage. The percentage of each score value was multiplied by the maximum length of the loadings to provide a similar scale for the loading plot. The score plot distribution was scaled but undistorted in the converting process. An example of a PCA study with a combined PCA score plot and loading plot is illustrated in Figure 1. The samples include all metabolites in the control and Class 7 which is labeled as treatment in this example.

For two groups of PCA studies, first the separation direction was determined using the center of each group in the PCA score plot. The center was calculated using the average of each group's first two principal components (PC) for which an example is shown in Figure 1. The two ends of the black line in Figure 1 are the centers of each group, and the black line is the separation direction for the two groups. Second, the PCA loading points were projected onto the line determined by the separation of the two groups. The projection of each score plot was calculated using the distance of the point (one metabolite) to the line in an x-y axis. Examples of loading values before and after projections are shown in Figure 2B. The projection of loading points to the separation line of the score plot was calculated using Matlab (R2020, Mathworks) with a laboratory-written script. Third, the projection of loading plots was used as the contribution of the corresponding metabolites to the separation, and the distance between the loading to the center of the axis (the 0) was calculated using Excel (Microsoft) based on the distance calculation method (Figure 2B). The positive or negative loading values in the x-axis direction were used to determine the sign of the distance, which is important for the determination of the correlations to the score groups. For example, the loading ($x2$, $y2$) in Figure 2B has a positive value, meaning the loading has a positive relationship with the treatment group, which has a center with a positive x value. While the loading ($x1$, $y1$) has a negative value, there is a positive relationship with the control group (negative relationship with the treatment group). The distances are called loading factors in the remainder of this article. Finally, a plot between the loading factors and the p values between the two groups was prepared to interpret the results for easy visualization by end-users (Figure 3).

For studies with multiple groups, the linear fit of the centers from all study groups was applied using this methodology, and the loading plots were projected to the linear fit with the similar methods used in the two-group study. Loading factors and p values were plotted using the p values between the two groups.

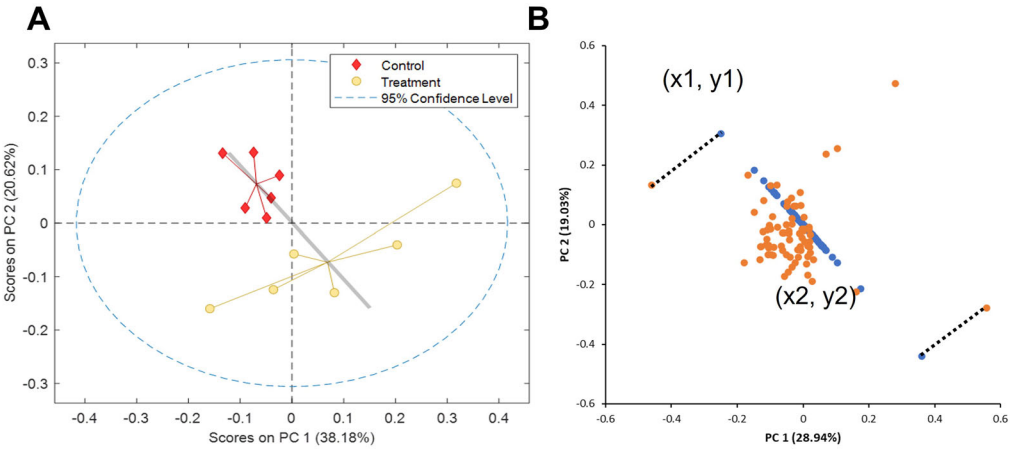

Figure 2. (A) PCA score plot of two study groups. The two groups showed a distinct separation after the PCA study and were connected to the center of each group. Red diamonds are the controls and yellow dots are the treated samples. (B) Corresponding PCA loading plot after projection. The blue dots are the original PCA loadings generated by PCA. The orange dots are the loadings after projection on the separation direction of the two study groups. The black dash lines showed two loadings before projection (blue dots) and after projection (orange dots). The distance to the center for ($x1$, $y1$) is $-\sqrt{x1^2 + y1^2}$ where the sign of $x1$ is negative, and the distance for ($x2$, $y2$) is $\sqrt{x2^2 + y2^2}$ where the sign of $x2$ is positive. PCA loadings are the metabolites' contributions to the PCA score plot (Fig. 2A).

Results

Demonstration with two groups

The methodology was first tested in a two-group study of the controls and one high atrazine concentration treatment in the sea anemone samples. The PCA score plot showed a distinct difference between the groups (control and treatment) using the first two principal components (PCs) (Figure 1, squares and triangles). The combined PCA score and loading plot (Figure 1) showed the separation between the two (black line). The loadings (Figure 1, dots) show the metabolite correlations with the score values. The loadings (metabolites) in the same direction of the treatment group are positively related, and those in the opposite direction negatively correlated. PCA is designed to describe data without supervision (Karhunen and Joutsensalo 1994), which indicated the metabolic difference as the whole system. The PCA loading values showed the correlation with the score plot when autoscaling was applied (Yamamoto et al. 2014). The Pareto scaling, which modifies the data to the square root of the standard deviation, suppresses the smaller peaks. Small peaks have more noise (Wang, Goodpaster, and Kennedy 2013), so Pareto scaling can reduce its effect on the modeling. This method has been applied in NMR metabolomics in several studies (van den Berg et al. 2006b; Lindon, Nicholson, and Holmes 2007). The loadings showed the potential metabolites that highly contributed to the PCA model (Figure 1). However, the interpretation of the PCA loading plot is difficult. Although it is possible to label the statistical significance with p values between study groups using a color map as previously reported (Goodpaster, Romick-Rosendale, and Kennedy 2010), the data analysis is still difficult due to the crowded data points and the absence of a loading value interpretation.

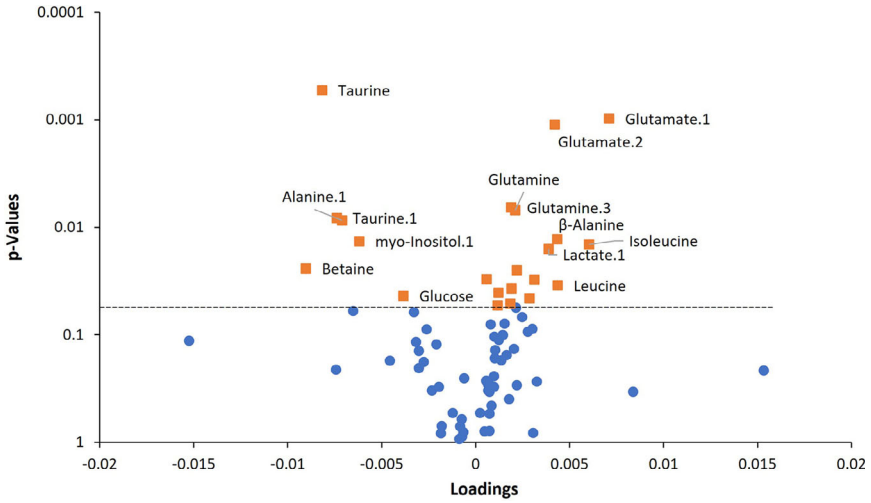

Figure 3. PCA loading factors versus the *p* values of the metabolites between two groups. The metabolites with larger loading factors and small *p* values are highlighted in orange squares. The metabolites with larger *p* values are in blue and were not further studied. The same metabolite may have more than one peak, so the metabolites were labeled with numbers after a dot. The dashed line is the cutoff of the *p* values (0.05).

When the loading plots were projected in a separate direction from the PCA score plot, the distance between the loading points to the axis became the loading factors (Figure 2B). The contribution of the metabolites that have large loading values but not in the direction of the separation is suppressed by the projection process, while the large loading values in the separation direction are not reduced. A plot using Student's *t*-test *p* values between log scale versus the loading factor shows the relationships of the loadings and *p* values. The plot provides a powerful tool to characterize the contribution of each metabolite to the model and the confidence level of a single metabolite (Figure 3). Figure 3 shows a clear plot of the important metabolites with both modeling and *p*-value information. In this case, glutamine and glutamate showed the importance in the separation and *p* values which indicated the potential perturbation in the glutamine pathways.

Applications in multiple study groups

Studies with multiple groups are usually too complex to be analyzed. However, PCA has been shown to be able to initially analyze the general trend of metabolite changes as a system model. The study with seven groups including a control and six atrazine concentration treatments is used as an example. This study shows that there is a general direction to the changing patterns after increasing the dose of atrazine. However, the data interpretation is difficult due to the slight difference of the in-group variance (Figure 4A). The center of the data shows a general trend with a relatively high linear fit R^2 value, and the changing patterns give evidence to the potential gradual changes with increased atrazine concentration (Figure 4B). The early metabolic markers are usually weak, so partial separation is normal in metabolomics which increases the difficulty in data interpretation. The combination of the loading factors of all groups versus the *p*

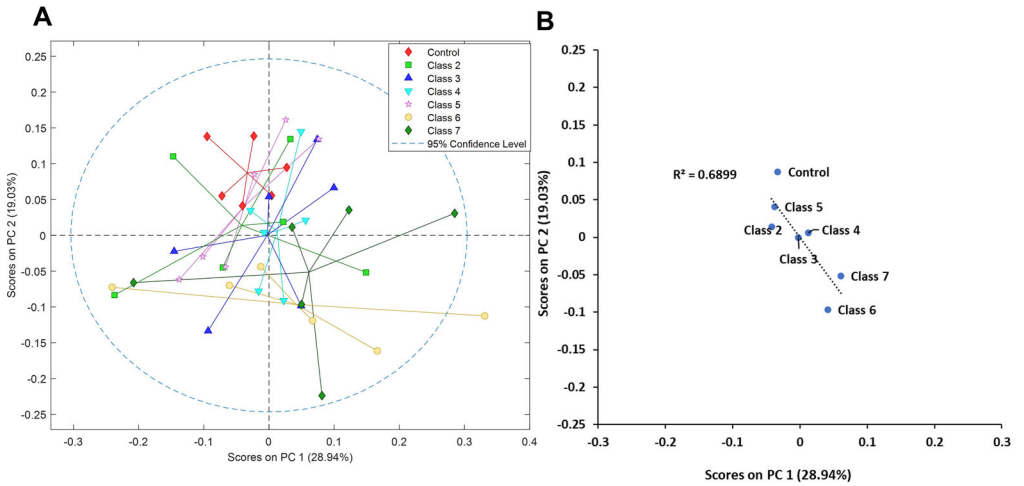

Figure 4. (A) PCA score plot of 7 study groups. (B) Linear fit showing the direction of the changes with the atrazine concentration. Class 5 is a potential outlier but has a limited effect on the data analysis which also indicated the performance of the method when outliers are present.

values of the comparison between a low concentration group and the controls provides higher confidence to analyze the early metabolic biomarkers.

In this example, the score plot shows that Class 2 has partial separation from the control groups (Figure 4A). Student's *t*-test showed that several metabolites were significantly different after treatment ($p < 0.05$, Figure 5). However, significant single metabolite changes may have potential family-wise error problems (Benjamini and Yekutieli 2001). The general metabolic profiling changes and the potential pathway models are more powerful to discover the influence of environmental stimuli upon the growth of sea anemones. For example, the results showed that the phenylalanine peak has a low *p*-value ($p < 0.05$, Figure 5) which may be considered important for single metabolite studies. However, the loading factor is also small which means the metabolite did not show a high contribution to the PCA models. In contrast, the glutamate showed low *p* values as well as high loading values, which were consistently observed in all high concentration groups (Figure 6). Therefore, glutamate may be recognized as the potential key biomarker in the sea anemones for the atrazine exposure. The one-way analysis of variance (ANOVA) also showed significant *p* values for metabolites such as glutamate when analyzing all groups (Table 1) which is consistent with this method.

Discussion

Applications of the method in PCA loading analysis for two groups

PCA loading plots are critical in interpreting PCA results, especially when a large number of metabolites were applied. Metabolomics aims to study the combinational changes of metabolite responses to environmental stimuli, and the PCA loading values may contribute to important metabolites in the model. The loading values of PCA provide important information that fills the disadvantages of Student's *t*-test or Welch's *t*-test which only consider a single metabolite (Wang, Goodpaster, and Kennedy 2013). These

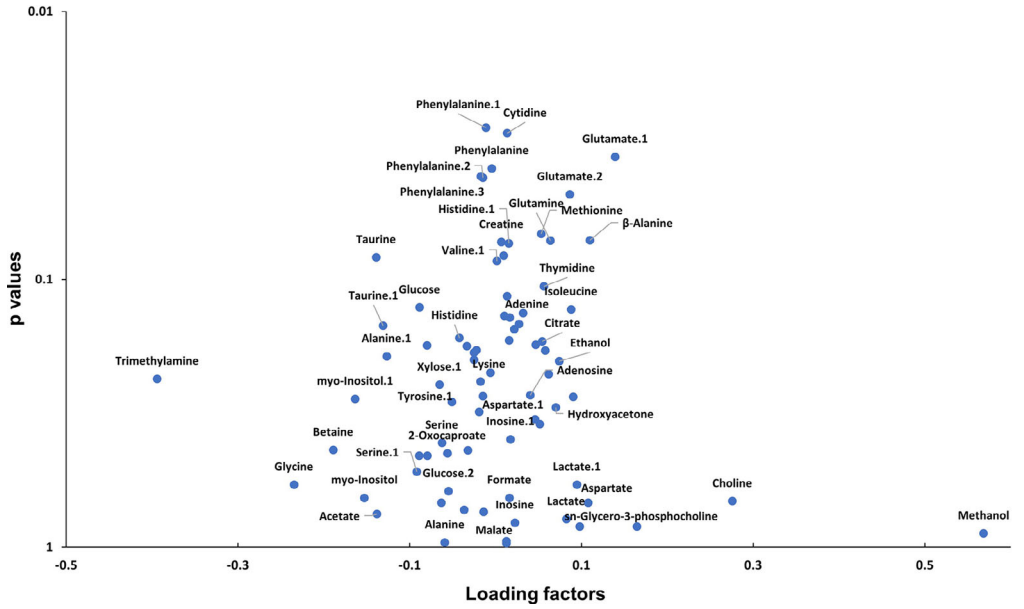

Figure 5. Example plot of *p* values versus loading factors using the latter from a multiple group PCA study. The *p* values are controls versus class 2 which is the lowest atrazine concentration of the treatments.

results show that when the loading factors are applied together with the *p* values of a *t*-test, an easy tool is provided to interpret the PCA loading plot. For example, metabolites with higher loading factors and low *p* values were highlighted in Figure 3 (orange dots) and metabolites such as glutamate and glutamine were significantly different between the two study groups but were critical metabolites which led to the separation of the groups. The interpretation of PCA loadings is easier with higher confidence. For example, the *p* values of glucose and lactate are lower than 0.05 but are higher than 0.01, which may be easily excluded when a p-value approach such as Bonferroni correction (Benjamini and Hochberg 1995) is applied (Figure 3). However, when relatively large loadings factors were observed, the important roles of the two metabolites in the PCA model are observed and the conversion of glucose to lactate in glycolysis is consistent with their opposite signs (glucose is negative and lactate is positive) in the PCA model. Therefore, the method also provides an approach to connect the loading analysis with metabolic pathway analysis. The average of the absolute value of all loading factors is suggested to be the cutoff for the loading factors in this study. Although the statistical significance of the metabolites has been studied with several methods, including the Welch's test and Mann-Whitney U test (Goodpaster, Romick-Rosendale, and Kennedy 2010; McKnight and Najab 2010), this study specifies the most suitable statistical methods for two groups depending on the metabolite distributions. This study applied loading factors in analyzing the PCA loading numerically and is suitable for study groups that include seven groups as in the demonstration data.

Applications of the method in multiple group studies

PCA is an important approach to study for multiple groups and considers the combined effect (Gogos et al. 2000) of each variable. The loading factor is important to show the

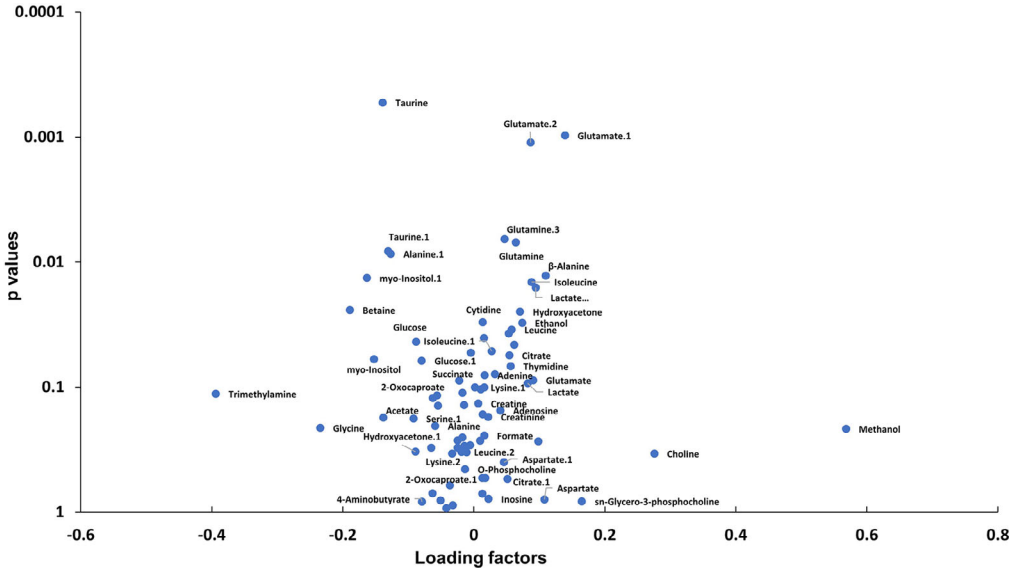

Figure 6. Example plot of *p* values versus loading factors using the latter from a multiple group PCA study. The *p* values are controls versus class 7 which is the highest atrazine concentration of the treatments.

Table 1. One-way analysis of variance (ANOVA) for metabolites with potential significant *p* values for the studies with seven groups.

Metabolites	*f* value	*p* value	False discovery rate
Glucose	4.84	1.08×10^{-3}	4.45×10^{-2}
myo-Inositol.1	4.37	2.15×10^{-3}	4.45×10^{-2}
Glutamate.2	4.28	2.46×10^{-3}	4.45×10^{-2}
myo-Inositol	4.26	2.55×10^{-3}	4.45×10^{-2}
Glutamate.1	4.18	2.85×10^{-3}	4.45×10^{-2}

The same metabolite may have more than one peak, so the metabolites were labeled with numbers after a dot.

metabolites as a group, not just one metabolite. This method is excellent to interpret gradual changes when a linear fit to the center of the PCA score plot was applied. When *p* values were plotted versus the loading factors calculated using the whole PCA study with all groups, the early marker was easily observed from a large number of metabolites. In metabolomics, the loading values show a similar position in the PCA loading plots from the same or related metabolic pathways. For example, the glutamate and glutamine in Figure 5 are from the glutamate pathway. The results also showed consistency with the one-way ANOVA study and the loading factor versus p-value plot showed clear information of the metabolite contributions to the PCA model. Most importantly, the potential early biomarkers or pathways are observed using this methodology. On the other hand, metabolites with significant *p* values in one group treatment but are unrelated to the general trend may be excluded. For example, phenylalanine showed a significant difference after the low concentration treatment (Class 2). However, the small loading factors (Figure 5) indicate the low contribution of the metabolites to change patterns of the principal metabolites and should not be considered to be an important early marker. The

results were further confirmed when phenylalanine did not show consistency in the higher atrazine concentration group (Figure 6).

Conclusion

Metabolomics studies are popular in human health and environmental studies; however, combination metabolite analysis is still difficult due to the large size of the datasets. Many studies focused on the significance level (p values) of a single metabolite instead of the combinational effects of multiple species. Although PCA showed excellent performance in metabolomics, the interpretation of PCA loadings has not been well studied. NMR has high reproducibility in metabolomics and is suitable for studies that need a long data acquisition process such as multiple group studies. In this work, an excellent methodology was examined to apply the PCA loading plot information in NMR-based metabolomics using sea anemone extracts following atrazine treatment. The methodology not only shows an efficient approach to analyze the PCA loading plots with potential combination information, but also indicates powerful applications in multiple-group studies with better visualization. The different concentrations of atrazine on sea anemones showed a powerful way to investigate the gradual changes of responses to environmental influences. For the multiple group studies, this work is suitable for multiple groups with gradual changes such as early-stage discoveries with multiple concentrations, or variable time sample collection after drug treatment. The approaches for studies with completely different groups will be further developed in our future work. The method may have limitations when outliers exist in studies that distort the PCA score plot and therefore lead to less powerful loading factors. Future investigation will characterize optimized approaches to minimize the potential outliers for PCA. In summary, the method is an efficient approach to discover potential early metabolic biomarkers in metabolomics and improves the applications in environmental health.

Acknowledgments

BW is thankful to the Department of Chemistry of North Carolina A&T State University for the startup funding and the support for the NMR facility. The authors thank Kennedi Jackson, Janece Sewell, and Mya Nicholas at North Carolina A&T State University for the English revision. This material is based upon work supported in part by the National Science Foundation under Grant No. CHE- 2137575.

Disclosure statement

There are no relevant financial or non-financial competing interests to report.

Funding

BW is thankful to the Department of Chemistry of North Carolina A&T State University for the startup funding. LJ would like to acknowledge the financial support of the New College of Florida.

References

Acar, E. F., and L. Sun. 2013. A generalized Kruskal-Wallis test incorporating group uncertainty with application to genetic association studies. *Biometrics* 69 (2):427–35. doi:10.1111/biom.12006.

Ahmad, F., and W. M. Dar. 2018. Classification of Alzheimer's disease stages: An approach using PCA-based algorithm. *American Journal of Alzheimer's Disease and Other Dementias* 33 (7): 433–9. doi:10.1177/1533317518790038.

Ahmadi, S., A. Razazan, R. Nagpal, S. Jain, B. Wang, S. P. Mishra, S. Wang, J. Justice, J. Ding, D. A. McClain, et al. 2020. Metformin reduces aging-related leaky gut and improves cognitive function by beneficially modulating gut microbiome/goblet cell/mucin axis. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences* 75 (7):e9–e21. doi:10.1093/gerona/glaa056.

Ametaj, B., Q. Zebeli, F. Saleem, N. Psychogios, M. Lewis, S. Dunn, J. Xia, and D. Wishart. 2010. Metabolomics reveals unhealthy alterations in rumen metabolism with increased proportion of cereal grain in the diet of dairy cows. *Metabolomics* 6 (4):583–94. doi:10.1007/s11306-010-0227-6.

Atzori, L., J. L. Griffin, A. Noto, and V. Fanos. 2012. Review metabolomics: A new approach to drug delivery in perinatology. *Current Medicinal Chemistry* 19 (27):4654–61. doi:10.2174/092986712803306448.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1):289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.

Benjamini, Y., and D. Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 29:1165–88.

Boroujerdi, A. F. B., M. I. Vizcaino, A. Meyers, E. C. Pollock, S. L. Huynh, T. B. Schock, P. J. Morris, and D. W. Bearden. 2009. NMR-based microbial metabolomics and the temperature-dependent coral pathogen vibrio coralliilyticus. *Environmental Science & Technology* 43 (20): 7658–64. doi:10.1021/es901675w.

Brus, J., M. Urbanova, I. Sedenkova, and H. Brusova. 2011. New perspectives of 19f mas NMR in the characterization of amorphous forms of atorvastatin in dosage formulations. *International Journal of Pharmaceutics* 409 (1–2):62–74. doi:10.1016/j.ijpharm.2011.02.030.

Chawla, M. P. S. 2011. PCA and ICA processing methods for removal of artifacts and noise in electrocardiograms: A survey and comparison. *Applied Soft Computing* 11 (2):2216–26. doi:10.1016/j.asoc.2010.08.001.

Choubey, D. K. M., Kumar, V. Shukla, S. Tripathi, and V. K. Dhandhania. 2020. Comparative analysis of classification methods with PCA and LDA for diabetes. *Current Diabetes Reviews* 16 (8):833–50. doi:10.2174/1573399816666200123124008.

Elliott, A. C., and L. S. Hynan. 2011. A SAS(®) macro implementation of a multiple comparison post hoc test for a Kruskal-Wallis analysis. *Computer Methods and Programs in Biomedicine* 102 (1):75–80. doi:10.1016/j.cmpb.2010.11.002.

Emwas, A.-H., R. Roy, R. T. McKay, L. Tenori, E. Saccenti, G. A. N. Gowda, D. Raftery, F. Alahmari, L. Jaremko, M. Jaremko, et al. 2019. NMR spectroscopy for metabolomics research. *Metabolites* 9 (7):123. doi:10.3390/metabo9070123.

Gadekallu, T. R., D. S. Rajput, M. Reddy, K. Lakshmanna, S. Bhattacharya, S. Singh, A. Jolfaei, and M. Alazab. 2021. A novel PCA–whale optimization-based deep neural network model for classification of tomato plant diseases using GPU. *Journal of Real-Time Image Processing* 18 (4):1383–96. doi:10.1007/s11554-020-00987-8.

Gogos, A., D. Jantz, S. Senturker, D. Richardson, M. Dizdaroglu, and N. D. Clarke. 2000. Assignment of enzyme substrate specificity by principal component analysis of aligned protein sequences: An experimental test using DNA glycosylase homologs. *Proteins: Structure, Function, and Genetics* 40 (1):98–105. doi:10.1002/(SICI)1097-0134(20000701)40:1<98::AID-PROT110>3.0.CO;2-S.

Goodpaster, A., L. Romick-Rosendale, and M. Kennedy. 2010. Statistical significance analysis of nuclear magnetic resonance-based metabonomics data. *Analytical Biochemistry* 401 (1):134–43. doi:10.1016/j.ab.2010.02.005.

Gu, H., Z. Pan, B. Xi, V. Asiago, B. Musselman, and D. Raftery. 2011. Principal component directed partial least squares analysis for combining nuclear magnetic resonance and mass spectrometry data in metabolomics: Application to the detection of breast cancer. *Analytica Chimica Acta* 686 (1–2):57–63. doi:10.1016/j.aca.2010.11.040.

Halouska, S., and R. Powers. 2006. Negative impact of noise on the principal component analysis of NMR data. *Journal of Magnetic Resonance (San Diego, Calif.: 1997)* 178 (1):88–95. doi:10.1016/j.jmr.2005.08.016.

Hernandez-Bolio, G. I., A. Fagundo-Mollineda, E. E. Caamal-Fuentes, D. Robledo, Y. Freile-Pelegrin, and E. Hernandez-Nunez. 2021. NMR metabolic profiling of sargassum species under different stabilization/extraction processes. *Journal of Phycology* 57 (2):655–63. doi:10.1111/jpy.13117.

Howarth, R. 2017. *Dictionary of mathematical geosciences*. Cham: Springer.

Jiang, L., H. Sullivan, C. Seligman, S. Gilchrist, and B. Wang. 2021. An NMR-based metabolomics study on sea anemones Exaiptasia diaphana (Rapp, 1829) with atrazine exposure. *Molecular Omics* 17 (6):1012–20. doi:10.1039/d1mo00223f.

Karhunen, J., and J. Joutsensalo. 1994. Representation and separation of signals using nonlinear PCA type learning. *Neural Networks* 7 (1):113–27. doi:10.1016/0893-6080(94)90060-4.

Kumar, A., S. Kumar, A. K. Maurya, and V. K. Agnihotri. 2020. NMR based metabolic profiling of Saussurea lappa roots and aerial parts from western Himalaya. *Analytical Chemistry Letters* 10 (4):428–41. doi:10.1080/22297928.2020.1796783.

Kumar, A., A. K. Maurya, G. Chand, and V. K. Agnihotri. 2018. Comparative metabolic profiling of costus speciosus leaves and rhizomes using NMR, GC-MS and UPLC/ESI-MS/MS. *Natural Product Research* 32 (7):826–33. doi:10.1080/14786419.2017.1365069.

Kumar, N., M. Shahjaman, M. N. H. Mollah, S. M. S. Islam, and M. A. Hoque. 2017. Serum and plasma metabolomic biomarkers for lung cancer. *Bioinformation* 13 (06):202–8. doi:10.6026/97320630013202.

Kumazoe, M., Y. Fujimura, S. Hidaka, Y. Kim, K. Murayama, M. Takai, Y. Huang, S. Yamashita, M. Murata, D. Miura, et al. 2015. Metabolic profiling-based data-mining for an effective chemical combination to induce apoptosis of cancer cells. *Scientific Reports* 5 (1):9474. doi:10.1038/srep09474.

Li, Y. F., S. Qiu, and A. H. Zhang. 2016. High-throughput metabolomics to identify metabolites to serve as diagnostic biomarkers of prostate cancer. *Analytical Methods* 8 (16):3284–90. doi:10.1039/C6AY00127K.

Lindon, J. C., J. K. Nicholson, and E. Holmes. 2007. *The handbook of metabonomics*. Elsevier. Amsterdam, the Netherlands.

Markley, J. L., R. Bruschweiler, A. S. Edison, H. R. Eghbalnia, R. Powers, D. Raftery, and D. S. Wishart. 2017. The future of NMR-based metabolomics. *Current Opinion in Biotechnology* 43:34–40. doi:10.1016/j.copbio.2016.08.001.

McKnight, P. E., and J. Najab. 2010. Mann-Whitney U test. In *The Corsini encyclopedia of psychology*. John Wiley & Sons, Inc. Englewood Cliffs, NJ, US

Mora-Ortiz, M., P. Nuñez Ramos, A. Oregioni, and S. P. Claus. 2019a. NMR metabolomics identifies over 60 biomarkers associated with type ii diabetes impairment in db/db mice. *Metabolomics: Official Journal of the Metabolomic Society* 15 (6):89. doi:10.1007/s11306-019-1548-8.

Mora-Ortiz, M., P. Nuñez Ramos, A. Oregioni, and S. P. Claus. 2019b. NMR metabolomics identifies over 60 biomarkers associated with type ii diabetes impairment in db/db mice. *Metabolomics* 15 (6):16. doi:10.1007/s11306-019-1548-8.

Ni, J. J., L. Xu, W. Li, C. M. Zheng, and L. J. Wu. 2019. Targeted metabolomics for serum amino acids and acylcarnitines in patients with lung cancer. *Experimental and Therapeutic Medicine* 18 (1):188–98. doi:10.3892/etm.2019.7533.

Odiase, J. I., and S. M. Ogbonmwan. 2005. Jmasm20: Exact permutation critical values for the Kruskal-Wallis one-way ANOVA. *Journal of Modern Applied Statistical Methods* 4 (2):609–20. doi:10.22237/jmasm/1130804820.

Rodriguez-Perez, R., L. Fernandez, and S. Marco. 2018. Overoptimism in cross-validation when using partial least squares-discriminant analysis for omics data: A systematic study. *Analytical and Bioanalytical Chemistry* 410 (23):5981–92. doi:10.1007/s00216-018-1217-1.

Ruiz-Perez, D., H. Guan, P. Madhivanan, K. Mathee, and G. Narasimhan. 2020. So you think you can PLS-DA? *BMC Bioinformatics* 21 (Suppl 1):2–10. doi:10.1186/s12859-019-3310-7.

Sangster, T., H. Major, R. Plumb, A. J. Wilson, and I. D. Wilson. 2006. A pragmatic and readily implemented quality control strategy for HPLC-MS and GC-MS-based metabonomic analysis. *The Analyst* 131 (10):1075–8. doi:10.1039/b604498k.

Scheel, G. L., E. D. Pauli, M. Rakocevic, R. E. Bruns, and I. S. Scarminio. 2019. Environmental stress evaluation of coffea arabica l. Leaves from spectrophotometric fingerprints by PCA and OSC-PLS-DA. *Arabian Journal of Chemistry* 12 (8):4251–7. doi:10.1016/j.arabjc.2016.05.014.

Smilde, A., J. Westerhuis, H. Hoefsloot, S. Bijlsma, C. Rubingh, D. Vis, R. Jellema, H. Pijl, F. Roelfsema, and J. Van Der Greef. 2010. Dynamic metabolomic data analysis: A tutorial review. *Metabolomics: Official Journal of the Metabolomic Society* 6 (1):3–17. doi:10.1007/s11306-009-0191-1.

Spicer, R. A., R. Salek, and C. Steinbeck. 2017. Analysis: Compliance with minimum information guidelines in public metabolomics repositories. *Scientific Data* 4:170137. doi:10.1038/sdata.2017.137.

St, L., and S. Wold. 1989. Analysis of variance (ANOVA). *Chemometrics and Intelligent Laboratory Systems* 6 (4):259–72. doi:10.1016/0169-7439(89)80095-4.

Trenfield, M. A., J. W. van Dam, A. J. Harford, D. Parry, C. Streten, K. Gibb, and R. A. van Dam. 2017. Assessing the chronic toxicity of copper and aluminium to the tropical sea anemone Exaiptasia pallida. *Ecotoxicology and Environmental Safety* 139:408–15. doi:10.1016/j.ecoenv.2017.02.007.

Trygg, J., E. Holmes, and T. Lundstedt. 2007. Chemometrics in metabonomics. *Journal of Proteome Research* 6 (2):469–79. doi:10.1021/pr060594q.

Urbanova, M., L. Kobera, and J. Brus. 2013. Factor analysis of 27Al mas NMR spectra for identifying nanocrystalline phases in amorphous geopolymers. *Magnetic Resonance in Chemistry: MRC* 51 (11):734–42. doi:10.1002/mrc.4009.

van den Berg, R. A., H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf. 2006b. Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics* 7 (1):1–15. doi:10.1186/1471-2164-7-142.

van den Berg, R. H., Hoefsloot, J. Westerhuis, A. Smilde, and M. van der Werf. 2006a. Centering, scaling, and transformations: Improving the biological information content of metabolomics data. *BMC Genomics* 7:142.

Wang, B., A. M. Goodpaster, and M. A. Kennedy. 2013. Coefficient of variation, signal-to-noise ratio, and effects of normalization in validation of biomarkers from NMR-based metabonomics studies. *Chemometrics and Intelligent Laboratory Systems* 128:9–16. doi:10.1016/j.chemolab.2013.07.007.

Wang, B., A. M. Maldonado-Devincci, and L. Jiang. 2020. Evaluating line-broadening factors on a reference spectrum as a bucketing method for NMR based metabolomics. *Analytical Biochemistry* 606:113872. doi:10.1016/j.ab.2020.113872.

Wang, B. S., Sheriff, A. Balasubramaniam, and M. A. Kennedy. 2015. NMR based metabolomics study of y2 receptor activation by neuropeptide y in the sk-n-be2 human neuroblastoma cell line. *Metabolomics* 11 (5):1243–52. doi:10.1007/s11306-015-0782-y.

Wang, B., Z. Shi, G. Weber, and M. Kennedy. 2013. Introduction of a new critical p value correction method for statistical significance analysis of metabonomics data. *Analytical and Bioanalytical Chemistry* 405 (26):8419–29. doi:10.1007/s00216-013-7284-4.

Werth, M., S. Halouska, M. Shortridge, B. Zhang, and R. Powers. 2010. Analysis of metabolomic pca data using tree diagrams. *Analytical Biochemistry* 399 (1):58–63. doi:10.1016/j.ab.2009.12.022.

Westerhuis, J., H. Hoefsloot, S. Smit, D. Vis, A. Smilde, E. van Velzen, J. van Duijnhoven, and F. van Dorsten. 2008. Assessment of plsda cross validation. *Metabolomics* 4 (1):81–9. doi:10.1007/s11306-007-0099-6.

Wilson, I., R. Plumb, J. Granger, H. Major, R. Williams, and E. Lenz. 2005. HPLC-MS-based methods for the study of metabonomics. *Journal of Chromatography. B, Analytical Technologies in the Biomedical and Life Sciences* 817 (1):67–76. doi:10.1016/j.jchromb.2004.07.045.

Xi, B., H. Gu, H. Baniasadi, and D. Raftery. 2014. Statistical analysis and modeling of mass spectrometry-based metabolomics data. *Methods in Molecular Biology* 1198:333–53.

Yagmur, B., and A. Gunes. 2021. Evaluation of the effects of plant growth promoting rhizobacteria (pgpr) on yield and quality parameters of tomato plants in organic agriculture by principal component analysis (PCA). *Gesunde Pflanzen* 73 (2):219–28. doi:10.1007/s10343-021-00543-9.

Yamamoto, H., T. Fujimori, H. Sato, G. Ishikawa, K. Kami, and Y. Ohashi. 2014. Statistical hypothesis testing of factor loading in principal component analysis and its application to metabolite set enrichment analysis. *BMC Bioinformatics* 15:51. doi:10.1186/1471-2105-15-51.

Yata, K., and M. Aoshima. 2012. Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *Journal of Multivariate Analysis* 105 (1): 193–215. doi:10.1016/j.jmva.2011.09.002.

Zimmerman, D. W., and B. D. Zumbo. 1993. Rank transformations and the power of the student t test and Welch t'test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale* 47 (3):523–39. doi: 10.1037/h0078850.