# Introduction

Data collected on many variables for the same individuals presents some unique challenges for data analysis. Multivariate data is not new to the field of ecology, and ecologists have been using techniques like principal component analysis to reduce dimensionality and simplify the interpretation of multivariate data for a long time (CITATION TO SOME OLD PCA PAPER). However, in recent years, the scale of data available for ecological research has increased due to advancements in high-throughput sampling technology (Kallenbach et al. 2014; Kfoury et al. 2017), image processing (Berger et al. 2010; Fahlgren et al. 2015), automated and remote data logging (Cooke et al. 2004; Porter et al. 2005), remote sensing (Roughgarden et al. 1991; Aplin 2005), high-throughput sequencing technologies (Soininen et al. 2009), and citizen science (Silvertown 2009; Bonney et al. 2009; Dickinson et al. 2012). Simultaneously, there is perhaps and increasing interest in going beyond simply describing multivariate data to inferring processes from patterns.

One typical approach to multivariate statistical inference in ecology is to first reduce dimensionality through an unsupervised technique like PCA, and then to look for visual separation in a score plot or perhaps to use the derived latent variables (principle components in the case of PCA) as predictors in a statistical test for a relationship with some response variable. However, this may not answer the question the researcher is actually asking. Unsupervised techniques like PCA are agnostic to response variables and can only describe the variation in data. PCA followed by a statistical test on principle components answers the slightly different question “What are the main axes of variation in the data? Do those axes have a relationship with the response variable?” This approach can be justified when one is truly interested in the axis that describes variation. For example, the leaf economics spectrum (LES) is a highly repeatable principle component that explains variation in leaf traits from “slow” to “fast” lifestyles. Because this multivariate trade-off exists across all plants and within groups of plants, it is reasonable to ask questions about how position along the LES varies among habitats, along environmental gradients, or among clades. However, a failure to find a relationship with the LES does not mean that measured leaf traits do not vary among habitats, along environmental gradients, or among clades.

Supervised approaches, on the other hand, take response variables into account and test how response variables co-vary with one or more predictor variables. Using a supervised technique answers a fundamentally different question than the unsupervised approach described above. When using a supervised analysis, you are answering the question “Is there a multivariate relationship between the predictors and the response variable?” This may seem a minor distinction, but can lead to completely different conclusions, and can even miss statistically significant relationships, as we will demonstrate.

The use of supervised multivariate analyses in ecology may not be as common because of some of the challenges ecological data typically creates, including missing values, multicollinearity, small sample sizes, and more variables than observations (AKA the “curse of dimensionality”). Previously, these challenges limited the use of supervised multivariate statistics, but this is no longer the case. In fact, several techniques, including partial least squares regression (PLSR) and its discriminant analysis extension (PLS-DA), handle the above-mentioned challenges of ecological data especially well. Unlike PCA, which creates axes that explain the most variation in the data, PLS creates axes that explain the most co-variationwith a dependent variable. This is an important distinction because it is not safe to assume that the best explanatory variables will also show the most overall variation among samples.

PLSR was first developed in the late seventies in the field of econometrics (Wold…) and later adopted by analytical chemistry (Geladi and Kowalski 1986), but has only been widely adopted in ecology by the sub-field of chemical ecology (Citation). The underlying assumption of PLS is that covariation between the predictor variable(s) and the response variable(s) is due to a small number of “latent” variables. For this reason, the technique has been readily adopted for the analysis of metabolomic data because a change in a large number of metabolites may be a result of a change in a single enzyme or metabolic pathway. In fact, PLS has been implemented into many metabolomics-specific statistical software (SIMCA, metaboanalyst). However, the utility of PLSR is not limited to metabolomic data, or even to data all of one type. In fact, one of the strengths of PLSR is that, unlike permutational MANOVA, it does not rely on distance or dissimilarity measures and therefore retains information about the relative importance of individual variables which can be used to make loading plots and summarized with a variable importance in projection (VIP) score. VIP scores have been shown to be very robust to determining which predictor variables are responsible for variation in the response variable (Chong and Jun 2005). This makes PLS an ideal technique for highly multivariate ecological datasets where multicollinearity may be present, and the researcher would not only like to address a hypothesis about the effects of some independent variable on a multivariate dataset, but also to determine which measured variables are most responsible for the multivariate relationship.

To demonstrate the use of PLSR and compare with a PCA approach, we employ a case study as well as simulated data…

# Case Study

## Methods

## Results and Discussion

It is important to perform cross-validation on PLS models to avoid over-fitting. Cross-validation involves leaving parts of the data out of the model development, then predicting it with the model and comparing with the actual values. Cross-validation is often used as a criterion to select a number of components for PLS models. In the case that zero components are selected, the model is considered non-significant and one can conclude that there is no significant multivariate relationship with the response variable. The explanatory power of a PLS model can be evaluated with an R2Y statistic which is analogous to a regression R2 and is the amount of variation in the Y variable explained by the model. Additionally, some software packages (e.g. *ropls* R package and SIMCA) provide a Q2 value, which is generated through cross-validation and can be used to evaluate the predictive power of a PLS model. A Q2 value that is much lower than the R2Y value indicates that the R2Y is not particularly robust. [some sentence about what to do if Q2 is really low. Ask Elizabeth about this]. Worley and Powers (2016) suggest using PCA as a second validation tool. However, we demonstrate below that it is entirely possible to have a highly significant PLS model with no visible separation in a PCA score plot. Statistical significance of the PLS model can be derived through permutation testing. By randomly permuting data labels and re-fitting the PLS, a p-value can be generated for both R2Y and Q2 values to determine how likely those values are to be as high or higher simply by chance.

Plotting

If a PLS model doesn’t pass the checks of cross-validation, we strongly recommend against plotting the resulting score plot. As we show below, visible separation may occur in PLS score plots even when cross-validation indicates the model is not significant. If the PLS is significant, a score plot still could be misleading, especially given that even randomly permuted data can produce PLS score plots with visible separation between groups or ordering along a continuous response variable (Worley and Powers, 2016). Score plots are useful, however, when plotted alongside loading plots or when discriminating between more than two groups. Loading plots allow readers to interpret predictive axes. For a PLS-DA with multiple levels of a categorical response variable, readers can use a score plot to assess relative similarity of groups along predictive axes. Finally, because cross-validation is so essential when conducting PLS, we recommend always including cross-validation results of some kind in the figure legend or in the figure itself (fig…).

# Simulated data

## Methods

To demonstrate some of the properties of PLS and PCA, we used randomly generated multivariate data created with different covariance structures. All multivariate datasets had 20 observations, one factor with two levels (10 observations per level), and 25 continuous variables. All 25 variables had a variance of 1 and a mean of 0 when they were not discriminating between factor levels. Covariance and the difference in means between factor levels was adjusted depending on the scenario:

1 ) “Null”: 5 variables with covariance of 0, and two groups of 10 variables with a covariance of 0.5 (Fig 1A).

2) “Needle in a haystack”: two groups of 10 variables with covariance of 0.5 and 5 variables with a difference in means of 2 (Fig 1D).

3) “Red Herring” where 10 variables covaried moderately with a covariance of 0.5 but distinguished groups poorly (difference in means = 1); 10 variables with zero covariance, but that distinguished groups more strongly (difference in means = 2); 5 variables did not covary or distinguish groups (i.e. noise).

4) “Control” where two sets of 5 variables covary moderately with covariance = 0.5 and discriminate between groups (difference in means = 2); 5 variables with covariance = 0.5 and no difference in means; and 10 variables that do not covary or distinguish groups (i.e. noise) (Fig 1G).

Multivariate data were simulated in R using the *tidymvsim* package (citation), which allows simple generation of multivariate data frames with varying correlation structures. We created 100 randomly generated datasets using the same parameters under each of these scenarios. PCA and PLS-DA were conducted using the *opls* function from the *ropls* package with default settings (Thévenot et al. 2015). For permutation testing to calculate p-values for PLS-DA models, 500 permutations were used. In the case that a PLS model couldn’t be created for a dataset (for example, because the first component was not significant), those datasets were removed after recording the number of failed models.

To test accuracy of identification of important discriminating variables, we used the “control” and “needle in a haystack” scenarios where variables were either discriminating or not. We set criteria for both methods (PCA and PLS-DA) to identify important discriminating variables. For PLS-DA, a variable was considered identified as discriminating if it had a variable importance in projection (VIP) score greater than 1 (Chong and Jun 2005). For PCA, a variable was considered discriminating if its distance from 0 in a correlation plot of the first two principal components was greater than 0.38, which is equivalent to the threshold Pearson correlation coefficient that would be significant at alpha = 0.05. We then compared these to known variable identities (discriminating or not) and created a confusion matrix for each dataset with the number of discriminating variables correctly identified as important being a true positive. From this we calculated a kappa coefficient for each dataset, which describes the accuracy of the method for choosing discriminating variable. A Kappa coefficient of 1 indicates complete accuracy while a kappa of 0 indicates important variables are selected no better than by chance. A negative kappa indicates that selection of important variables is worse than chance. See supplemental files for reproducible R scripts.

## Results and Discussion

*PLS-DA is prone to overfitting.* PLS models should be cross-validated to determine model predictive power and statistical significance. In the *ropls* package this is done through 7-fold cross validation where some of the data is left out for model generation, then predicted by the model and compared to true values. Cross validation is used to select a number of components. In the case that zero components are selected as optimum, the model should be considered non-significant. If at least one component is appropriate, an R2X, R2Y, and Q2 value will be calculated and p-values generated by permutation testing. R2X is the variation in the predictor variables explained by the model (analogous to R2 for PCA), R2Y is the variation in the response variable explained by the model (analogous to R2 for a regression), and Q2 describes the predictive power of the model and is generated through cross-validation (Eriksson, 2006). A Q2 close to 1 means the model has strong predictive power and a Q2 value close to or below zero means the model has weak predictive power. If Q2 is much lower than R2Y or close to or below zero, this is an indication that the model has very low predictive power and there is probably not a significant relationship between the response and predictors (Thenevot). A p-value is also reported by the *ropls* package by randomly permuting data labels and re-calculating R2Y and Q2 and measuring the proportion of permuted models with R2Y and Q2 values higher than the model created with the true data.

Under the null scenario, both PCA and PLS-DA are expected to find no separation between groups. For PLS-DA, 75 out of 100 models failed because the first predictive component was not significant (Thevenot et al 2015). A lack of effect of the treatment is evident for the PCA due to the lack of separation of groups in the score plot (Fig 1. B). For the PLS-DA models that were successfully built however, there is visible separation between groups in the score plot, and the permutation tests are marginally significant on average (pQ2 = 0.080 ± 0.054). It is important to note that even under a scenario of completely random data, PLS-DA score plots will often show visual separation between groups, which is why cross validation is necessary. The PLS-DA models that were built have low R2Y and Q2 values (mean R2Y = 0.577 ± 0.171, mean Q2 = 0.219 ± 0.148) which is an indicator of poor model performance. In the case that a model has poor predictive power (low Q2) it is recommended that a score plot not be shown, and no p-value reported, as even under this null scenario visible separation is shown, and this is misleading.

A close up of a map

Description automatically generated

Figure 1: Multivariate analysis of representative datasets from three data scenarios: “null” (A, B, C), “control” (D, E, F) and “needle in a haystack” (G, H, I). The first column shows correlation heatmaps (A, D, G). Variable names on the axes that begin with “C” were generated with a covariance of 0. 5, those that begin with “N” were generated with a covariance of 0, and those that begin with “D” were generated to have different means in the two groups. The second column shows PCA score plots (B, E, H). The third column shows PLS-DA plots (C, F, I). For PLS-DA plots, the first two predictive axes are plotted, Q2 values are calculated using 7-fold cross validation, and pQ2 is calculated with 500 permutations (C, F, I). Ellipses represent 95% confidence bounds, parenthetical numbers on axis labels are the percent of total variation explained by the axis. Note that in C, the PLS-DA is clearly not a good model due to low Q2 value. We recommend not including a such PLS-DA score plot for non-significant results in a publication.

*PLS-DA finds obvious differences between groups just as well as PCA*. Under the control scenario, the PCA for one of the datasets failed due to a convergence error and PLS-DA models were built for all 100 datasets. Both PCA and PLS-DA show significant separation between groups. In PCA, there is separation between groups along PC1 (Fig. 1D). For PLS-DA, both mean R2Y and mean Q2 are high (0.835 ± 0.061 and 0.752 ± 0.072, respectively), and permutation testing is highly significant (pQ2 = 0.002 ± 0.001). Because the discriminating variables are also the variables that contribute the most to overall covariation in the dataset, PCA and PLS-DA are nearly equivalent, despite answering slightly different questions. Here, PCA is answering the question “Is there a main axis of variation in the data?” and then we are able to evaluate separation along that axis visually through a score plot or statistically by doing a t-test on principal component axis scores. PLS-DA, on the other hand, is answering the question “What variables (if any) explain the difference between groups?” and it answers this directly.

*PLS-DA outperforms PCA when discriminating variables aren’t responsible for the majority of covariation.* For the needle in a haystack scenario, there is poor separation between groups in the PCA score plot (Fig 1H) while PLS-DA shows strong separation (Fig 1I) and has a very high mean R2 and Q2 value (0.906 ± 0.046 and 0.709 ± 0.108, respectively) and a highly significant permutation test (p = 0.002 ± 0.001).

In the needle in a haystack scenario, PCA performs poorly at finding group separation, while PLS-DA is able to find strong separation between groups. This is because the variables that contribute to differences between the groups are not contributing greatly to the overall variation in the data. Again, this is by design because these methods answer different questions. PCA is finding the main axis of variation in the data, while PLS-DA is finding variables that co-vary with group membership. This is further demonstrated through calculation of the kappa coefficients for these analyses.

*PLS-DA identifies discriminating variables while PCA identifies variables with high variation.* Kappa coefficients for the two methods (Fig 2. table?) show that PLS identifies variables that were created with different means in the two levels of our factor consistently better than PCA. Even in the control scenario where PCA and PLS performed similarly in finding separation between groups, PLS far outperforms PCA at correctly identifying the variables most responsible for that separation (Fig 2A). This is by design since PCA is agnostic to the response variable and just attempting to explain variation in the data while PLS-DA is specifically identifying which variables (if any) contribute to differences between the groups.

A close up of text on a white background

Description automatically generated

Figure 2. Distribution of Cohen’s Kappa for the control (A) and needle in a haystack (B) scenarios. A value of 1 means perfect identification of discriminating variables by the model while a value of zero indicates the model performed no better than random chance.

# Works Cited

Aplin P (2005) Remote sensing: ecology. Prog Phys Geogr 29:104–113. doi: 10.1191/030913305pp437pr

Berger B, Parent B, Tester M (2010) High-throughput shoot imaging to study drought responses. J Exp Bot 61:3519–3528. doi: 10.1093/jxb/erq201

Bonney R, Cooper CB, Dickinson J, et al (2009) Citizen Science: A Developing Tool for Expanding Science Knowledge and Scientific Literacy. Bioscience 59:977–984. doi: 10.1525/bio.2009.59.11.9

Chong I-G, Jun C-H (2005) Performance of some variable selection methods when multicollinearity is present. Chemom Intell Lab Syst 78:103–112. doi: 10.1016/J.CHEMOLAB.2004.12.011

Cooke SJ, Hinch SG, Wikelski M, et al (2004) Biotelemetry: a mechanistic approach to ecology. Trends Ecol Evol 19:334–343. doi: 10.1016/J.TREE.2004.04.003

Dickinson JL, Shirk J, Bonter D, et al (2012) The current state of citizen science as a tool for ecological research and public engagement. Front Ecol Environ 10:291–297. doi: 10.1890/110236

Fahlgren N, Gehan MA, Baxter I (2015) Lights, camera, action: high-throughput plant phenotyping is ready for a close-up. Curr Opin Plant Biol 24:93–99. doi: 10.1016/J.PBI.2015.02.006

Geladi P, Kowalski BR (1986) Partial least-squares regression: a tutorial. Anal Chim Acta 185:1–17. doi: 10.1016/0003-2670(86)80028-9

Hervé MR, Nicolè F, Lê Cao K-A (2018) Multivariate Analysis of Multiple Datasets: a Practical Guide for Chemical Ecology. J Chem Ecol 44:215–234. doi: 10.1007/s10886-018-0932-6

Kallenbach M, Oh Y, Eilers EJ, et al (2014) A robust, simple, high-throughput technique for time-resolved plant volatile analysis in field experiments. Plant J 78:1060–1072. doi: 10.1111/tpj.12523

Kfoury N, Scott E, Orians C, Robbat A (2017) Direct Contact Sorptive Extraction: A Robust Method for Sampling Plant Volatiles in the Field. J Agric Food Chem 65:8501–8509. doi: 10.1021/acs.jafc.7b02847

Liu J (2011) Overview of redundancy analysis and partial linear squares and their extension to the frequency domain. Dalhousie University

Porter J, Arzberger P, Braun H-W, et al (2005) Wireless Sensor Networks for Ecology. Bioscience 55:561–572. doi: 10.1641/0006-3568(2005)055[0561:WSNFE]2.0.CO;2

Roughgarden J, Running SW, Matson PA (1991) What Does Remote Sensing Do For Ecology? Ecology 72:1918–1922. doi: 10.2307/1941546

Silvertown J (2009) A new dawn for citizen science. Trends Ecol Evol 24:467–471. doi: 10.1016/J.TREE.2009.03.017

Soininen EM, Valentini A, Coissac E, et al (2009) Analysing diet of small herbivores: the efficiency of DNA barcoding coupled with high-throughput pyrosequencing for deciphering the composition of complex plant mixtures. Front Zool 6:16. doi: 10.1186/1742-9994-6-16

Thévenot EA, Roux A, Xu Y, et al (2015) Analysis of the Human Adult Urinary Metabolome Variations with Age, Body Mass Index, and Gender by Implementing a Comprehensive Workflow for Univariate and OPLS Statistical Analyses. J Proteome Res 14:3322–3335. doi: 10.1021/acs.jproteome.5b00354

Worley B, Powers R (2016) PCA as a Practical Indicator of OPLS-DA Model Reliability. Curr Metabolomics 4:97–103. doi: 10.2174/2213235X04666160613122429

Wright IJ, Reich PB, Westoby M, et al (2004) The worldwide leaf economics spectrum. Nature 428:821–827