Using the right tool for the job: principal component regression and partial least squares regression for analysis of multivariate ecological data.

Eric R. Scott1, Elizabeth Crone1

1Tufts University, Department of Biology

# Abstract

# Introduction

Ecologists often want to know which of many possible measured predictor variables are associated with some response. These kinds of analyses can be challenging when there are a large number of predictor variables, relative to observations, and when the predictor variables include correlated factors that have different ecological meanings. Multivariate data analysis 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 (Sokal and Rohlf 1995). However, in recent years, the scale of data available for ecological research has increased due to advancements in high-throughput metabolite sampling (Kallenbach et al. 2014; Kfoury et al. 2017), high-throughput phenotyping (Berger et al. 2010; Fahlgren et al. 2015), automated and remote data loggers (Cooke et al. 2004; Porter et al. 2005), remote sensing (Roughgarden et al. 1991; Aplin 2005), high-throughput sequencing technologies (Reuter et al. 2015), 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. Because multivariate data present problems for linear modeling methods (e.g. multiple regression) (Carrascal et al. 2009), multivariate inference is typically approached two ways: by reducing dimensionality with unsupervised approaches and conducting hypothesis testing on dimensionally reduced data, or through supervised multivariate analyses. Here, our purpose is to illustrate and clarify these two approaches.

A common unsupervised approach to multivariate inference in ecology is principal component analysis (PCA) followed by a statistical test on the principal component axes generated. PCA creates orthogonal latent variables called principal components that attempt to explain the variation in the predictor variables. Then, the data can be projected onto principal component axes (“scores”) and can be used to look for visual separation or perhaps as predictors in a statistical test for a relationship with some response variable (Figure 1A). Unsupervised techniques like PCA are agnostic to response variables and simply describe the variation in a set of predictor variables. PCA followed by a statistical test on principal components is sometimes used to look for multivariate relationships between many predictors and a response. This unsupervised approach answers the questions: “What are the main axes of variation in the data?” and “Do those axes have a relationship with the response variable?” In contrastmultivariate analysesesSess:Which variables in a multivariate data set best explain a response variable and “Does this set of variables have statistically significant predictive power?”(see Fibure 1B). difference in some cases supervised and unsupervised methods

In some cases, ecologists may be interested in asking questions that relate most naturally to unsupervised analyses. For example, the leaf economics spectrum (LES) is a highly repeatable principal component that explains variation in leaf traits from a slow to fast rate of return on investments in nutrients and leaf dry matter resulting from a trade-off between metabolic rates and investment to leaf structure (Wright et al. 2004). 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 (Gagliardi et al. 2015; Muir et al. 2017). Similarly, Smith et al. (Smith and Kriebel 2018) investigate evolution of floral shape by first using PCA to reduce the dimensionality of a dataset of floral morphology measurements (a “morphospace”). They then use this principal component, which captures the greatest sources of variation in the morphological traits measured, as a predictor to test hypotheses about evolution and pollination syndromes. Again, this analysis makes sense because the authors are interested specifically in the floral traits that vary the most among species to investigate the causes of that variation.

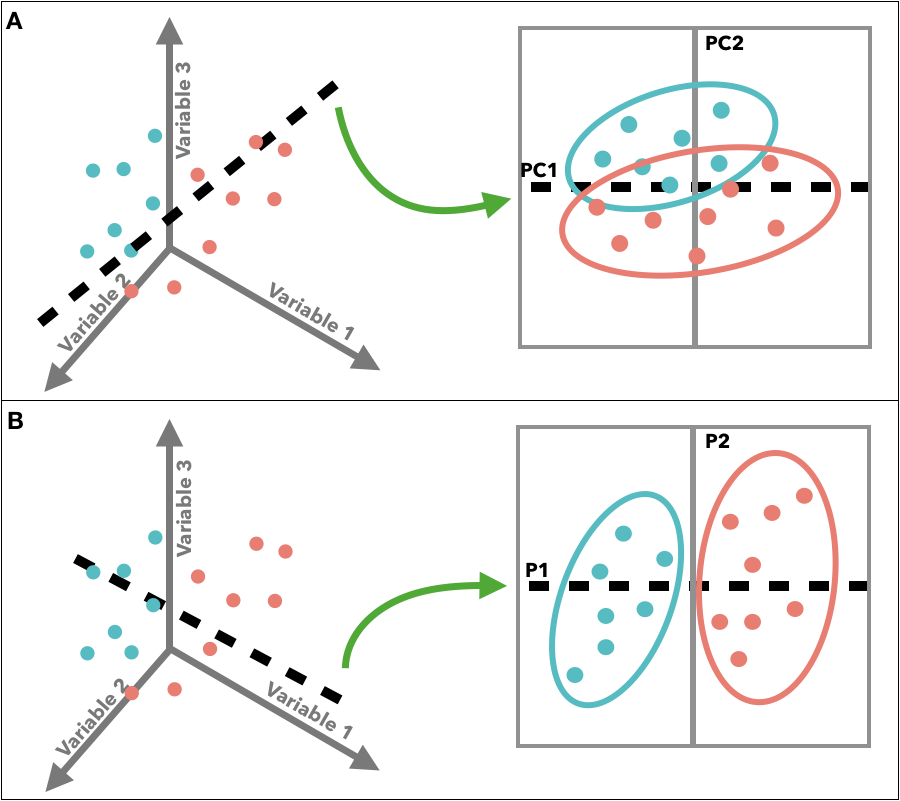


Figure . A hypothetical schematic of unsupervised (A) and supervised (B) dimensionality reduction. In the unsupervised case, a line is drawn through a three-dimensional cloud of points to capture the greatest amount of variation in the data, irrespective of point color. In the supervised case, the line is instead drawn to explain the greatest amount of co-variation with a response variable—in this case a categorical variable represented by point color.

However, one could take the same ecological problems and ask different questions using supervised analyses. Instead of asking if the position along the LES varies by elevation one could ask “which plant traits vary most with elevation?” and instead of asking if the floral morphological traits that vary the most are good predictors of pollinator identity, one could ask “which floral traits are the best predictors of pollinators?” The use of supervised multivariate analyses in ecology may not be as common as unsupervised approaches because of some of the challenges ecological data typically creates, including missing values, multicollinearity, small sample sizes, and more variables than observations (also called the “curse of dimensionality”). For example, multivariate analysis of variance (MANOVA), a fairly common supervised analysis in ecology, requires more independent observations than predictor variables. Other techniques, including partial least squares (PLS) regression and its discriminant analysis extension (PLS-DA), handle the above-mentioned challenges of ecological data especially well. In this paper, we contrast inference from PLS and PCA as representative examples of supervised and unsupervised multivariate analyses (respectively). We hope this comparison will help ecologists make informed decisions about which multivariate analyses best answer the questions researchers are interested in.

PLS was first developed in the late 1970’s in the field of econometrics (Wold 1975). and later adopted by analytical chemistry (Geladi and Kowalski 1986). In ecology, the use of PLS regression has been largely limited to the sub-field of chemical ecology (Hervé et al. 2018, but see Carrascal et al. 2009). The underlying assumption of PLS regression is that covariation between the predictor variable(s) and the response variable(s) can be explained by a small number of “latent” variables that may or may not be representative of some underlying causal variable. For example, a change in the concentrations of a large number of metabolites may be a result of a change in a single enzyme or hormone in a metabolic pathway (Lortzing and Steppuhn 2016; Padovan et al. 2017). In other cases, environmental conditions (e.g., climate and land use) may covary for reasons that are not specifically causal (Thogmartin et al. 2017, Crone et al. *in review*). PLS(-DA) has been widely adopted for the analysis of chemometric and metabolomic data and implemented into many metabolomics-specific statistical software (Barker and Rayens 2003; de Souza et al. 2017). However, the utility of PLS is not limited to metabolomic data. In fact, one of the strengths of PLS is that, unlike permutational MANOVA for example, it does not rely on distance or dissimilarity measures and therefore retains information about the relative importance of individual predictor variables (rather than the distance between them). PLS is a useful but underappreciated 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.

In this paper, we first present a a case study that originally used PCA regression to answer a slightly different question. We re-analyze this case study using PLS, and use these contrasting analyses to demonstrate some of the key differences in conclusions one might reach using a supervised vs. unsupervised approaches. We also use this case study to illustrate reporting of PLS model outputs. . We then use simulated data to explore some of the properties of PLS-DA and PCA discriminant analysis under different data scenarios. Finally, we provide some tips and best practices for implementing PLS models for ecological data.

# Case Study

Muir et al. (2017) investigated coordination between physiological and morphological leaf traits within a genus in order to understand evolutionary constraints on leaf trait trade-offs. In a common garden experiment, they collected data on leaf traits of 16 *Solanum* species grown from seed. Measured traits included leaf mass per area (LMA), leaf thickness, leaf dry matter content (LDMC), stomatal conductance (*gs*), mesophyll conductance (*gm*), assimilation rate (*A*N), water use efficiency (WUE), maximum rate of carboxylation (*V*cmax) and leaf respiration (*R*dark). In their analysis, Muir et al. performed PCA to recapitulate the leaf economics spectrum (LES). They then used the first principal component (the LES) to answer questions about how the position along the LES varies among habitats and phylogenetic relationships. Here, we reproduce their PCA-based analysis, then use PLS regression to answer a different question—which leaf traits vary with mean annual precipitation and temperature of species’ habitats?

### Methods

*Solanum* leaf trait data and coordinates of species habitats are available on Dryad (Muir et al. 2016). We used the habitat coordinates to download mean annual precipitation and temperature data from WorldClim (Fick and Hijmans 2017). Muir et al. (2017) identified one species, *S.* *juglandifolium*, as a potential outlier and performed their analyses both with and without that species. For the sake of brevity, we only include analyses with this species excluded.

To replicate the approach of the original paper, we performed PCA on leaf traits to recapitulate the LES. Then, we used the first principal component axis (i.e., the LES) as a predictor variable in two regressions with either mean annual temperature or precipitation as response variables. PCA was calculated using the *opls* package in R (CITE), and regressions were performed using the XX function in R.

We then used PLS regression to test the hypothesis that there is a relationship between climate variables and leaf traits. The two PLS analyses included all leaf traits as predictor variables and habitat mean temperature or precipitation as response variables. It is possible to include multiple response variables in PLS models to account for collinearity. However, in this dataset, temperature and precipitation were not correlated (Pearson’s correlation test, r = -0.10, df = 63, p = 0.421) so we chose to perform PLS for temperature and precipitation separately to improve interpretability, and for easier comparison with PCA results. PLS was performed using the *opls* function in the *ropls* package in R with default settings other than increasing the number of permutations to 1000 for PLS (Thévenot et al. 2015).

### Results and Discussion

*Interpretation of PCA regression*

The PCA of leaf traits resulted in two principal components with a cumulative R2 of 0.558. The first principal component axis (PC1), which explained 32% of the variation in leaf trait data, effectively recapitulated the LES with a strong positive correlation with LMA, LDMC, and leaf thickness and a strong negative correlation with *A*N, *g*m, *V*cmax and *g*s (Table 1). This principal component did not vary significantly with preciptiation (F(1,63) = 1.22, p = 0.273), but was associatied with temperature (F(1,63) = 9.07, p = 0.004) . However, the significant temperature association showed (Table 1)expected (cf. Wright et al, others XXX) Based on these results, Muir et al., concluded that there was limited evidence of association between leaf traits (i.e., position along LES position) with climate.

*Interpretation of PLS regression*

PLS showed significant associations between leaf traits and both temperature and precipitation, but the trait combinations that led to these relationships did not fall neatly along the LES. The PLS for precipitation produced a model with a single predictive component that explained about 20% of the total variation in leaf traits (Table 2, R2X(cum)) and about 42% of the variation in precipitation (Table 2, R2Y(cum)). Q2 is a statistic generated by internal cross-validation that can indicate overfitting or poor predictive power of a model (Eriksson et al. 2006). Q2 is always lower than R2Y, but when it is much lower, it indicates overfitting. In this case, Both R2Y and Q2 were relatively low, but similar, indicating moderate explanatory and predictive power, but a lack of overfitting. The significance of the model was determined by permutation and re-calculation of the R2Y and Q2 values. The p-values for R2Y and Q2 both indicate there was a highly significant, moderately strong, relationship between precipitation and leaf traits. PLS loadings showed that LMA, LDMC, and leaf thickness were positively correlated with precipitation (Table 1). Plants living in wet climates have thicker, denser leaves compared to those living in dry climates. The loadings for the precipitation PLS were similar to those for the first principal component, but without the strong tradeoffs with measures of photosynthetic capacity that are part of the LES. The lack of statistical significance in the PCA regression for precipitation was likely because photosynthetic traits were strongly loaded on PC1 and did not actually vary with precipitation.

The PLS for temperature also produced a highly significant single component model with similar explanatory (R2Y) and predictive (Q2) power to the precipitation PLS. VIP scores and predictive component loadings indicated a negative relationship between temperature and *V*cmax, *A*N, WUE, *R*dark and leaf thickness and a positive relationship with LDMC. Unlike the first principal component, physical leaf traits were not correlated on this predictive axis and there was a positive relationship between temperature and LDMC, but a slight negative relationship with thickness. Plants living in hotter climates therefore have more dense and thin leaves compared to plants from cooler climates.

*Comparison of PCA regression to PLS regression*

If our question of interest is “are there leaf trait–climate associations?”, then PLS regression is an appropriate approach that tells us there is a small, but highly significant relationship between leaf traits and both precipitation and temperature. The main axis of co-variation with temperature was quite different from that with precipitation and neither were very similar to the main axis of overall variation along the LES (i.e. PC1). These differences are not surprising given that PCA is agnostic to habitat variables (temperature, precipitation) while PLS regression is explicitly attempting to explain co-variation with habitat variables.

Table . Loadings from the first two principal components from PCA (PC1, PC2) and the first predictive component from PLS regressions with temperature and precipitation as response variables.



Table . Model diagnostics from PLS regression for both precipitation and temperature.



# Simulated data

### Methods

To further explore some of the properties of PLS and PCA, we applied both techniques to simulated 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 predictor 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 (Figure 2A). Here, there are no variables that distinguish the two groups.

2) “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) (Figure 2B). Here, the variables that distinguish the two groups fall along the main axis of covariance in the predictor variables.

3) “Needle in a haystack”: two groups of 10 variables with covariance of 0.5 and 5 variables with a difference in means of 2 (Figure 2C). Here, the variables that distinguish the two groups do not fall along the main axis of covariance in the predictor variables.

Multivariate data were simulated in R using the *holodeck* package (Scott 2019), 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 discriminant analysis (PCA-DA) was performed by first fitting a PCA using the *opls* function from the *ropls* package, which selects a number of principal components using an autofit criterion (Thévenot et al. 2015). Then a logistic regression was performed using all retained principal components as predictor variables and a p-value from a likelihood ratio test is reported (Aguilera et al. 2006). PLS-DA was conducted using the *opls* function from the *ropls* package with default settings except increasing the number of permutations to 500 to calculate p-values (Thévenot et al. 2015). We repeated PLS-DA analyses, twice, once using only statistically significant predictive axes (following standard best practices, CITE), and once forcing models to have the same number of predictive axes as were created by PCA (for comparison with PCA).For both PCA-DA and PLS-DA, root mean squared error of prediction (RMSEP) was calculated by external 7-fold cross-validation with the help of the *rsample* package (Kuhn and Wickham 2019). For PLS-DA, we used R2Y output by the *opls* package. For PCA-DA, we calculated R2 using XXX formula (Tjur 2009).

We were also interested in the ability of PCA-DA and PLS-DA to identify important discriminating variables in the “control” and “needle in a haystack” scenarios. To test this, we set criteria for both methods (PCA and PLS-DA) to identify important discriminating variables. For PCA-DA, a variable was considered important if had a significant Pearson correlation (p < 0.05) with any of the principal components that had a significant relationship with the response variable. Two separate criteria were used to determine important variables from PLS-DA models. First, variables were selected as important if they had significant correlations to the predictive axes retained in the PLS-DA models, similar to the criteria used for PCA-DA. In the second method, a variable was selected if it had a variable importance in projection (VIP) score greater than 1—a common criterion used for variable selection in PLS-DA (Chong and Jun 2005). VIP scores estimate the explanatory power of predictor variables over all significant predictive axes for PLS models. For each model, each variable was categorized as important or not. We then compared these values to known variable identities (generated as 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 Cohen’s kappa for each dataset, which describes the accuracy of the method for choosing discriminating variable (Cohen 1960). A kappa 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. Cohen’s kappa was calculated using the *psych* package for R (Revelle 2018).

See supplemental files for reproducible R scripts.

### Results and Discussion

Under the null scenario, both PCA-DA and PLS-DA are expected to find no separation between groups. As expected, both methods generally resulted in non-significant models. Due to a lack of convergence, 3 of the PCA-DA models failed. For the remaining 97 models, PCA explained ~ 55% of the variation in the predictor variables on average (Table 3, ) and the resulting PCs explain about 15% of the variation in the response variable (Table 3, ). The mean p-value from the PCA-DA models is high (*p* = 0.406), and there is a clear lack of separation in the score plot (Figure 2D). For PLS-DA, 75 out of 100 models failed because the first predictive component was not significant (Thévenot et al. 2015). The 25 PLS-DA models that were successfully built have low and values, which is an indicator of poor model performance (Table 3). When we forced PLS\_DA to produce a 1 component model (for comparison with PCA), the mean value was negative ( = -0.189), indicating very poor model performance. Even in such models with poor predictive power and non-significant p-values, some separation in the PLS-DA score plot can be observed (Figure 2G). This is a feature of PLS because some predictor variables will be associated with the response by chance, and this chance of spurious separation increases with the number of variables. Therefore, it is not generally recommended to show or interpret score plots for non-significant models (Kjeldahl and Bro 2010; Worley et al. 2013), and we have only included such a plot for heuristic purposes.

Under the control scenario, both PCA-DA and PLS-DA are expected to find significant separation between groups because the discriminating variables also fall along the main axis of variation in the predictor variables. As expected, both methods resulted in highly significant models that explained variation in group membership well. The PCA-DA for one of the datasets failed due to lack of convergence. For the remainng 99 data sets, PCA explained ~ 56% of the variation in the predictor variables () and the resulting PCs explained ~ 99% of the variation in the response (). The logistic regression portion of PCA-DA is highly significant under this scenario (p < 0.001). PLS-DA explained less variation in the predictor variables compared to PCA ( = 0.313), This difference is expected since PCA attempts to explain variation in the predictor variables and PLS-DA is only attempting to explain co-variation with the response variable. PLS-DA models explained variable under the control scenario values were calculated differently for PCA-DA and PLS-DA and therefore are not directly comparable between methods (TABLE) (Tjur 2009; Thévenot et al. 2015). RMSEP was calculated in the same way for both models, and indicated that PLS-DA had slightly higher predictive power (INSERT STATS HERE). For PLS-DA, mean was also high, and permutation testing wass highly significant ( = 0.002 ± 0.001), indicating a strong and highly significant relationship between the response and predictors. Overall, because the discriminating variables are also the variables that contribute the most to overall covariation in the dataset, PCA-DA and PLS-DA are nearly equivalent, despite answering slightly different questions.

For the needle in a haystack scenario, PCA-DA is expected to find principal components that do not explain differences between groups, since the variables that distinguish groups do not co-vary strongly. In contrast, PLS-DA is expected to identify the actual discriminating variables. In most ways, this contrast played out as expected. PCA-DA models were built for 97 models (3 failed due to convergence errors) and had a mean similar to the control scenario (Table 3, Figure 2F). However, was lower compared to the control scenario and the relationship between the PCs and the response variable was not usually statistically significant (mean *p* = 0.103, X% significant at the *p* < 0.05 level). RMSEP was generally high … (ADD STATS…?). PLS-DA, on the other hand, explained less variation in the predictor variables (), but nearly 91% of the variation in the response variable, even though only 5 out of 25 variables differed between in the two groups. was high, indicating good model performance, and both and are highly significant (Table 3, Figure 2I). The RMSEP for PLS-DA was noticeably lower than PCA-DA in this scenario. It was also similar to the RMSEP for PLS-DA in the control scenario, suggesting that PLS-DA does just as well when there are only 5 discriminating variables out of 25 that do not create the main axis of variation in the predictor variables, as when there are 10 out of 25 discriminating variables, along the main axis of covariance.

Table . Model statistics from simulation study. For each statistic a mean ± standard deviation is reported calculated from n models. For PCA-DA and PLS-DA the number of components was determined by the *opls* function, and in the case of the null scenario the best number of components was zero, in which case a model wasn’t built. For PLS-DA (ncomp forced), a one or two component model was forced.



A close up of a map

Description automatically generated

Figure . Multivariate analysis of representative datasets from three data scenarios: “null” (A, D, G), “control” (B, E, H) and “needle in a haystack” (C, F, I). The first row shows correlation heatmaps (A, B, C). 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. Score plots for PCA (D, E, F) and PLS-DA (G, H, I) plot samples as belonging to one of two groups (red circles with a solid ellipse or teal triangles with a dashed ellipse). 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. Ellipses represent 95% confidence bounds, parenthetical numbers on axis labels are the percent of total variation explained by the axis. Note that in G, the PLS-DA is clearly not a good model due to low Q2 value. We have only included this plot for heuristic reasons and do not recommend including a such PLS-DA score plot in a publication.

In terms of identifying specific causal variables, PLS-DA identified variables that were created with different means in the two levels of our factor better than PCA-DA (Figure 3). Under the control scenario, values of kappa were always greater than 0 for PCA-DA and both variable selection methods for PLS-DA. However, even in the control scenario where PCA-DA and PLS-DA performed similarly in finding separation between groups, PLS-DA outperformed PCA-DA at correctly identifying the variables most responsible for that separation, especially when using VIP scores (Figure 3A). Under the needle in a haystack scenario, the difference between PCA-DA and PLS-DA is more dramatic. PLS-DA using the VIP score method of identifying important variables performs the best with a mean Kappa of 0.802. Using correlation between axis scores and data produces more similar results between PCA-DA and PLS-DA, although there was a broader spread with more high scores (Kappa approaching 1) for PLS-DA.

We investigated the poor performance of correlation coefficients for PLS-DA by visually examining confusion matrices (not shown). Poor performance of this statistic reflected a large number of false positives, rather than false negatives. We hypothesize that the poor performance of correlation scores is partly due to the fact that, on average, the PLS-DA models for the needle in a haystack scenario included two predictive axes when there was actually only one axis separating groups in the data. False positives therefore came from two sources—non-discriminating variables which were correlated with group membership by chance, and significant correlations with the spurious second predictive axis. VIP scores do two things that the correlation method does not. VIP scores represent the *relative* importance of variables (the mean of squared VIP values always equals 1), and variable influence is summed over all axes, weighting the axes by their explanatory power (Chong and Jun 2005; Eriksson et al. 2006). This provides a more accurate prediction of important variables in this and many other cases, although it can fail when a large proportion of variables are actually causal (Chong and Jun 2005).

A close up of a map

Description automatically generated

Figure . Kernel density estimate of the distribution of Cohen’s kappa for the control (A) and needle in a haystack (B) scenarios for PCA-DA (red, solid line) and PLS-DA using two importance criteria: correlation with predictive axes (green, dashed line) and VIP > 1 (blue, dotted line). A kappa 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. Negative values of kappa indicate variable selection worse than expected by random chance.

# Conclusion

With multivariate data analysis, it is important to think about the question you are trying to answer. If you are primarily interested in predictor variables that vary among individuals or observations, then an unsupervised method like PCA regression might be most appropriate. However, if you are looking for the predictors that best explain a response variable, a supervised approach like PLS is more appropriate. From very early on in the adoption of PCA regression in ecology, it was suggested that this method might miss important relationships with a response variable, but only in the case when “nature is downright mean” (Jolliffe 1982). We have shown in a case study that nature can be downright mean—that is, reducing dimensionality of leaf trait data with PCA obscured statistically significant relationships between leaf traits and habitat climate in tomato species that were revealed with PLS regression. It is not unreasonable to expect situations in nature where the best predictors of a response variable are not the predictors that (co)vary the most among all observations. Advances in statistical techniques and high-power computing mean there is no need to rely on the benevolence of nature and hope that important variables aren’t buried in a principal component analysis. Instead, PLS and many other supervised approaches are available for testing multivariate hypotheses.

### Guidelines for using and reporting results from PLS(DA).

While our work in this paper uses only one of many software implementations of PLS regression, the same general guidelines can be generalized to the use of PLS regression and reporting of results using any statistical package. First, because PLS regression focuses on co-variation between predictors and response, it’s possible for a statistically significant relationship to be due to a small number of predictor variables (as in our needle in a haystack scenario). We recommend reporting R2X in addition to R2Y to give the reader a sense of how much variation in the predictors and the response, respectively, are explained by a PLS model. Second, because PLS is prone to overfitting, cross-validation of some sort is necessary to determine model predictive power (Eriksson et al. 2006; Kjeldahl and Bro 2010). When cross-validation indicates a weak model (a low Q2 value, for example), this should also be taken into account when interpreting results. Worley and Powers (2016) suggest using separation in a PCA score plot as validation method for PLS-DA. Their implicit assumption is that the variables which co-vary the most are also the ones contributing to group separation. We disagree with this method since, as we’ve shown in our needle in a haystack scenario, there can be a significant PLS-DA model with no separation of groups in a PCA score plot. Third, it is important to remember that score plots for PLS models differ from PCA because the predictive axes of PLS models are built to maximize co-variation with the response variable. This means that even statistically non-significant models may show some visual separation in score plots. Because Euclidean distance between points has little meaning in a PLS-DA score plot, we recommend that score plots be used sparingly for PLS-DA. However, in the case that there are more than two groups, or if a score plot is displayed alongside a loading plot, they may be useful visuals, especially when combined with visual aids like 95% confidence ellipses (Worley et al. 2013). We recommend being explicit about interpretation of model strength and significance using appropriate statistics (e.g. Q2 and p-values) rather than visual separation in score plot figure captions. The nature of PLS score plots also means that unlike PCA regression or PCA-DA, it is entirely inappropriate to base statistical significance of a PLS model on any statistical test of predictive axis scores (Weldegergis et al. 2018). Instead, permutation testing of R2Y or Q2 values (Thévenot et al. 2015), or CV-ANOVA (Eriksson et al. 2008) can be used to assess statistical significance of PLS models. Finally, variable selection can be used to identify which predictors best explain variation in the response. This can be done with VIP scores, which estimate the explanatory power of predictor variables across all significant predictive component axes. It is important to note that the cutoff of VIP > 1 is somewhat arbitrary and may require adjustment based on properties of particular datasets (Chong and Jun 2005). Additionally, if you expect many predictor variables to be unrelated to the response, a different approach using sparse PLS (sPLS) might be valuable. Sparse PLS-DA (sPLS-DA) performs variable selection and classification together rather than using VIP scores post-hoc to determine important predictors (Lê Cao et al. 2011).

PLS is, of course, only one of many supervised multivariate analysis techniques available. Other techniques may be more appropriate depending on the data and the questions. Redundancy analysis, for example, may be more appropriate for factorial experimental designs (Hervé et al. 2018). [Other examples, citations to reviews, etc.]

* Suggest other ecological questions where PLS might be helpful (cite examples)

# Works Cited

Aguilera AM, Escabias M, Valderrama MJ (2006) Using principal components for estimating logistic regression with high-dimensional multicollinear data. Comput Stat Data Anal 50:1905–1924. doi: 10.1016/J.CSDA.2005.03.011

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

Barker M, Rayens W (2003) Partial least squares for discrimination. J Chemom 17:166–173. doi: 10.1002/cem.785

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

Carrascal LM, Galván I, Gordo O (2009) Partial least squares regression as an alternative to current regression methods used in ecology. Oikos 118:681–690. doi: 10.1111/j.1600-0706.2008.16881.x

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

Cohen J (1960) A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas 20:37–46. doi: 10.1177/001316446002000104

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

de Souza LP, Naake T, Tohge T, Fernie AR (2017) From chromatogram to analyte to metabolite. How to pick horses for courses from the massive web resources for mass spectral plant metabolomics. Gigascience 6:1–20

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

Eriksson L, Johansson E, Kettaneh-Wold N, et al (2006) Multi- and Megavariate Data Analysis Part II Advanced Applications and Method Extensions. In: Multi- and Megavariate Data Analysis Part II Advanced Applications and Method Extensions

Eriksson L, Trygg J, Wold S (2008) CV-ANOVA for significance testing of PLS and OPLS® models. J Chemom 22:594–600. doi: 10.1002/cem.1187

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

Fick SE, Hijmans RJ (2017) WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int J Climatol. doi: 10.1002/joc.5086

Gagliardi S, Martin AR, Filho E de MV, et al (2015) Intraspecific leaf economic trait variation partially explains coffee performance across agroforestry management regimes. Agric Ecosyst Environ 200:151–160. doi: 10.1016/J.AGEE.2014.11.014

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

Jolliffe IT (1982) A Note on the Use of Principal Components in Regression. Appl Stat 31:300. doi: 10.2307/2348005

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

Kjeldahl K, Bro R (2010) Some common misunderstandings in chemometrics. J Chemom 24:558–564. doi: 10.1002/cem.1346

Kuhn M, Wickham H (2019) rsample: General Resampling Infrastructure

Lê Cao K-A, Boitard S, Besse P (2011) Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics 12:253. doi: 10.1186/1471-2105-12-253

Lortzing T, Steppuhn A (2016) Jasmonate signalling in plants shapes plant-insect interaction ecology. Curr Opin Insect Sci 14:32–39. doi: 10.1016/j.cois.2016.01.002

Muir CD, Conesa MÀ, Roldán EJ, et al (2017) Weak coordination between leaf structure and function among closely related tomato species. New Phytol 213:1642–1653. doi: 10.1111/nph.14285

Muir CD, Conesa MÀ, Roldán EJ, et al (2016) Data from: Weak coordination between leaf structure and function among closely related tomato species. Dryad Digit Repos. doi: 10.5061/dryad.1r8c2

Padovan A, Keszei A, Hassan Y, et al (2017) Four terpene synthases contribute to the generation of chemotypes in tea tree (Melaleuca alternifolia). BMC Plant Biol 17:160. doi: 10.1186/s12870-017-1107-2

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

Reuter JA, Spacek D V., Snyder MP (2015) High-Throughput Sequencing Technologies. Mol Cell 58:586–597. doi: 10.1016/J.MOLCEL.2015.05.004

Revelle W (2018) psych: Procedures for Psychological, Psychometric, and Personality Research

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

Scott ER (2019) holodeck: A Tidy Interface for Simulating Multivariate Data

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

Smith SD, Kriebel R (2018) Convergent evolution of floral shape tied to pollinator shifts in Iochrominae (Solanaceae)\*. Evolution (N Y) 72:688–697. doi: 10.1111/evo.13416

Sokal RR, Rohlf FJ (1995) Biometry: The Principles and Practices of Statistics in Biological Research. Biol Res

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

Tjur T (2009) Coefficients of Determination in Logistic Regression Models—A New Proposal: The Coefficient of Discrimination. Am Stat 63:366–372. doi: 10.1198/tast.2009.08210

Weldegergis BT, Zhu F, Poelman EH, Dicke M (2018) Correction to: Drought stress affects plant metabolites and herbivore preference but not host location by its parasitoids. Oecologia 187:873–874. doi: 10.1007/s00442-018-4149-8

Wold H (1975) Soft Modelling by Latent Variables: The Non-Linear Iterative Partial Least Squares (NIPALS) Approach. J Appl Probab 12:117–142. doi: 10.1017/S0021900200047604

Worley B, Halouska S, Powers R (2013) Utilities for quantifying separation in PCA/PLS-DA scores plots. Anal Biochem 433:102–104. doi: 10.1016/J.AB.2012.10.011

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