# Introduction

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). The resulting multivariate datasets have the potential to increase our understanding of ecological phenomena, given that adequate statistical techniques are used to separate signal from potentially increasing noise.

Multivariate analysis techniques can be divided into two types—unsupervised and supervised. Unsupervised techniques are agnostic to experimental design. That is, they do not take information about dependent variables or group membership as input. Unsupervised multivariate analyses such as principal component analysis (PCA), principal coordinate analysis (PCoA), and non-metric multidimensional scaling (NMDS) are often employed to simplify interpretation of multidimensional data by collapsing correlated measured variables into fewer dimensions or latent variables. Unsupervised analyses are powerful tools for describing the variation in a multivariate data set. However, they are not designed with hypothesis testing in mind. Yet, many ecologists use unsupervised analyses in the context of testing relationships with dependent variables either by visually inspecting score plots that result from unsupervised techniques or by using the latent variables they create as independent variables in statistical tests.

Supervised multivariate analyses, on the other hand, take dependent variables into account and test how response variables co-vary with one or more dependent variables. Using a supervised technique answers a fundamentally different question than carrying out hypothesis tests on latent variables created by unsupervised techniques. When using a supervised analysis, you are answering the question “Does [dependent variable] explain the co-variation in the response variables?” while the latter approach—for example, a PCA followed by an ANOVA using principal component scores as an independent variable—answers the question “Is there a main axis of variation that describes the response variables, and does [dependent variable] explain variation in that main axis of variation?” While the second approach may be appropriate in some situations, it is important to understand the difference. As we will show in this paper, significant multivariate relationships are often missed by the unsupervised approach, for example when the variables that explain the dependent variable are not the same ones that account for a large amount of variation in the dataset.

In the past, ecologists may have avoided supervised techniques because they were subject to the so-called “curse of dimensionality”. That is, some supervised multivariate techniques cannot be used when the number of variables exceeds the number of observations, when there is a high degree of multicollinearity, or when data sets contain missing values—all common occurrences in ecological data sets. However, several supervised multivariate analysis methods have been created that do not have these limitations. Here, we focus on one such method, partial least squares regression. Unlike unsupervised techniques like PCA which create latent variables that explain the most variation in the data, PLS creates latent variables (axes) that explain the most **co-variation** with 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

Partial least squares regression (PLS) and its discriminant analysis extension (PLS-DA) are supervised multivariate analysis techniques created to be robust to small sample sizes (relative to the number of variables measured), missing values, and multicollinearity [citation to creator, to paper saying its robust]. It 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 PLS is not limited to metabolomic data, or even to data all of one type. In fact, one of the strengths of PLS is that it does not rely on distance or dissimilarity measures and therefore retains information about the relative importance of individual variables which can be 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 large 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.

PLS does show up in ecological literature, but often methods and results are reported poorly or incorrectly, perhaps because researchers are accustomed to presenting and seeing results of unsupervised methods like PCA. Therefore, in this paper we offer advice and best practices on when to use PLS, how to use PLS, and how to report results.

# Methods (briefly)

## Simulated data methods

Multivariate data were simulated in R using [my package name?], which provides wrapper functions for mvrnorm() from the MASS package. The sim\_covar() function adds columns to a data frame that co-vary with a user supplied variance and covariance. The sim\_discr() function adds multivariate normal data that differ in means between levels of a categorical variable in the data frame. The sim\_missing() function simply randomly replaces a user supplied fraction of the data observations with missing values (NA).

Using these functions, we created data sets with 2 levels of a factor, 10 observations per level (N = 20) and 25 variables under [three?] different scenarios:

1 ) “Null”, where 5 variables had zero covariance, and two groups of 10 variables co-varied with covariance 0.5.

2) “Needle in a haystack” where two groups of 10 variables co-varied with covariance of 0.5 and 5 variables discriminated between groups (difference in means = 2).

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

We then created 100 randomly generated datasets using the same parameters under each of these three scenarios and subjected them to PCA and PLS-DA. For each PCA we calculated the distance between centroids of the two groups and for PLS-DA we calculated an R2, Q2, and pQ2 value. Additionally, for each pair of PCA and PLS-DA results we extracted PC1 loading values and calculated VIP scores for PLS-DA models. For VIP scores, we used a cutoff of VIP > 1 to select important variables and for PCA we chose a cutoff of loading > 0.15 to select important variables. We then 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. {DESCRIBE WHAT KAPPA COEF MEANS}

## Cupcakes vs. Muffins methods

I took a random subsample of 40 recipes (20 muffins and 20 cupcakes) where each variable is an ingredient (in cups per serving) and applied PCA and PLS-DA on it. I tried this with several random seeds to purposefully cherry-pick an example where PCA revealed some separation to compare the results of PCA vs PLS-DA.

I should also do PLS regression on calories per serving and include that instead or in addition to the cupcakes vs. muffins plots.

# Results

## Simulated data set

What I can say for almost sure:

1. Separation in PCA space is terrible at finding a needle in a haystack. That is, if your discriminating variables are a small proportion (or even only half) of your total variables, it does an exceptionally terrible job at detecting differences between groups.
2. Other statistical methods that use F-tests (PERMANOVA, RDA, PC-ANOVA) are not going to work well if there is a lot of covariation in the data set that is NOT related to the dependent variable. PLS-DA seems relatively un-phased by this situation and can still find the variables that contribute to group differences
3. Even if a PLS-DA model is **terrible** there can be visible separation in a plot of the first two axes. DO NOT publish PLS-DA plots if the model is not significant. In fact, don’t publish these plots on their own at all. They are useful as bi-plots or side-by-side with loading plots, but if all you’re trying to show is visual separation, **don’t do it**. It is misleading to readers. *Maybe* they’re useful if you have more than 2 groups, so you can see that some groups are more similar to each other than others.
4. It is irresponsible to get VIP scores from a poor PLS-DA model. (Chong and Jun 2005) shows that no matter what, some VIP scores will be >1, because that’s how VIP works (I think the mean VIP has to be 1 or something??). Therefore, the VIP > 1 cutoff is somewhat arbitrary. If it’s a weaker model, use a higher cutoff.

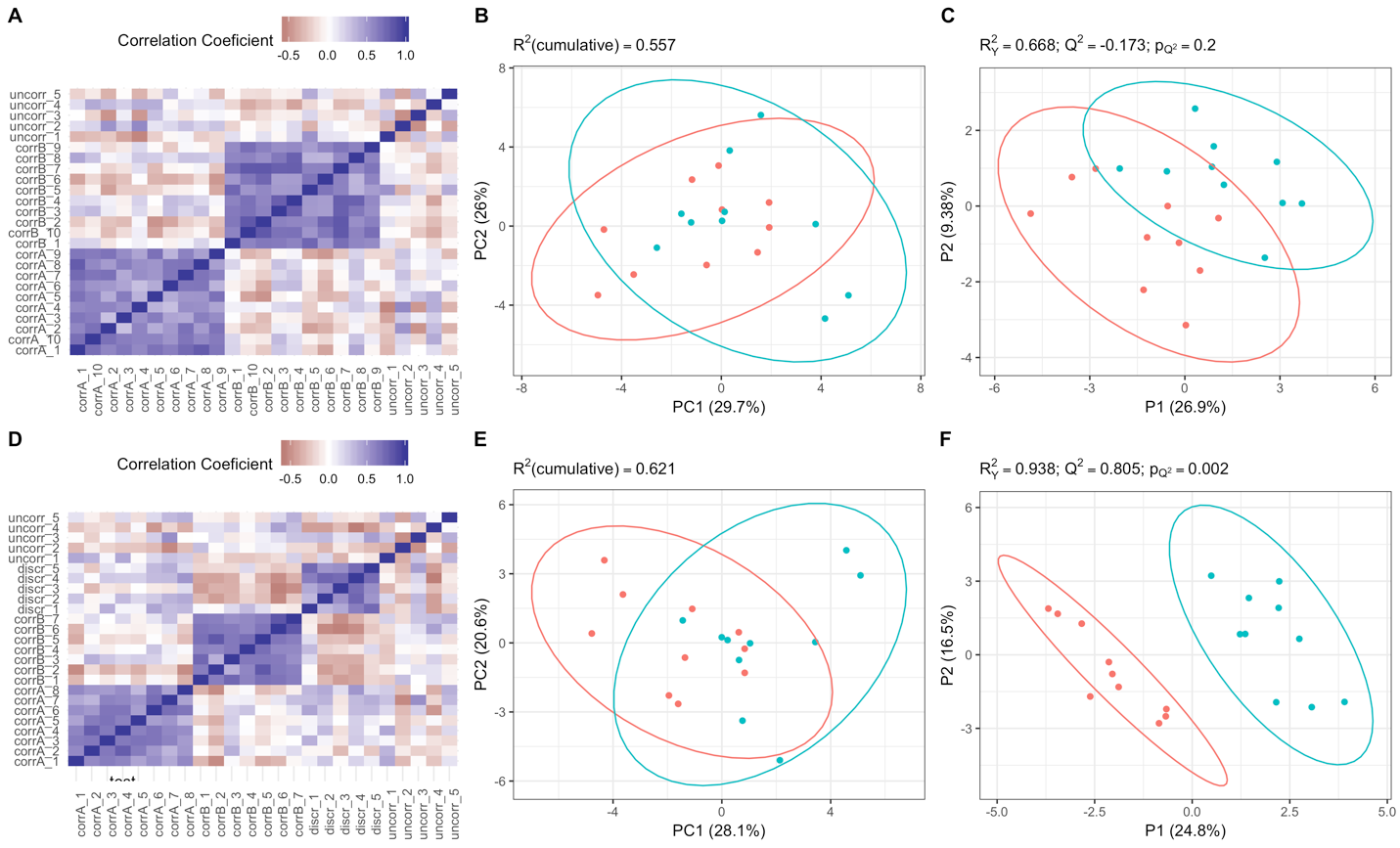


Figure 1: Multivariate analysis of simulated data with random group assignment (A, B, C) and with 5 variables generated to discriminate between groups (D, E, F). A and B show correlation heatmaps of randomly generated datasets. B and E are PCA score plots of the first two PC axes. For PLS-DA plots (C, F), 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 C, the PLS-DA is clearly not a good model due to low Q2 and a high p-value. We recommend not including a PLS-DA plot for non-significant results in a publication.

The addition of 5 discriminating variables has a negligible effect on the PCA. There is still essentially no separation between the two groups along either PC1 or PC2. However, the effect of these discriminating variables on the PLS-DA is apparent both in the visual separation between groups as well as the and values. The PLS-DA on completely random data also demonstrates the tendency of PLS to overfit. Without any cross-validation, one might conclude that the two groups were different, however the extremely low and high p value from this model indicates that this separation is due to chance. Without reporting these cross-validation measures, the PLS-DA plot alone would be extremely misleading. We therefore recommend that plots of non-significant PLS models not be included in publications. It’s also worth noting that even though only 5 of 25 variables were created to differ between groups, the first predictive axis of the PLS-DA on the full data set describes 24.8% of the total variation in the data.

Table 1: Variable importance in projection (VIP) scores from the PLS-DA model, and loadings of the first two PCA axes resulting from analysis of dataset 2, with 5 discriminating variables. Variables beginning with “corr” were created with a covariance of 0.55, variables beginning with “uncorr” had a covariance of 0, and variables beginning with “discr” had a mean difference of 2 between groups with a covariance of 0. VIP scores greater than 1 are typically considered significant contributors.



## Cupcakes vs. Muffins

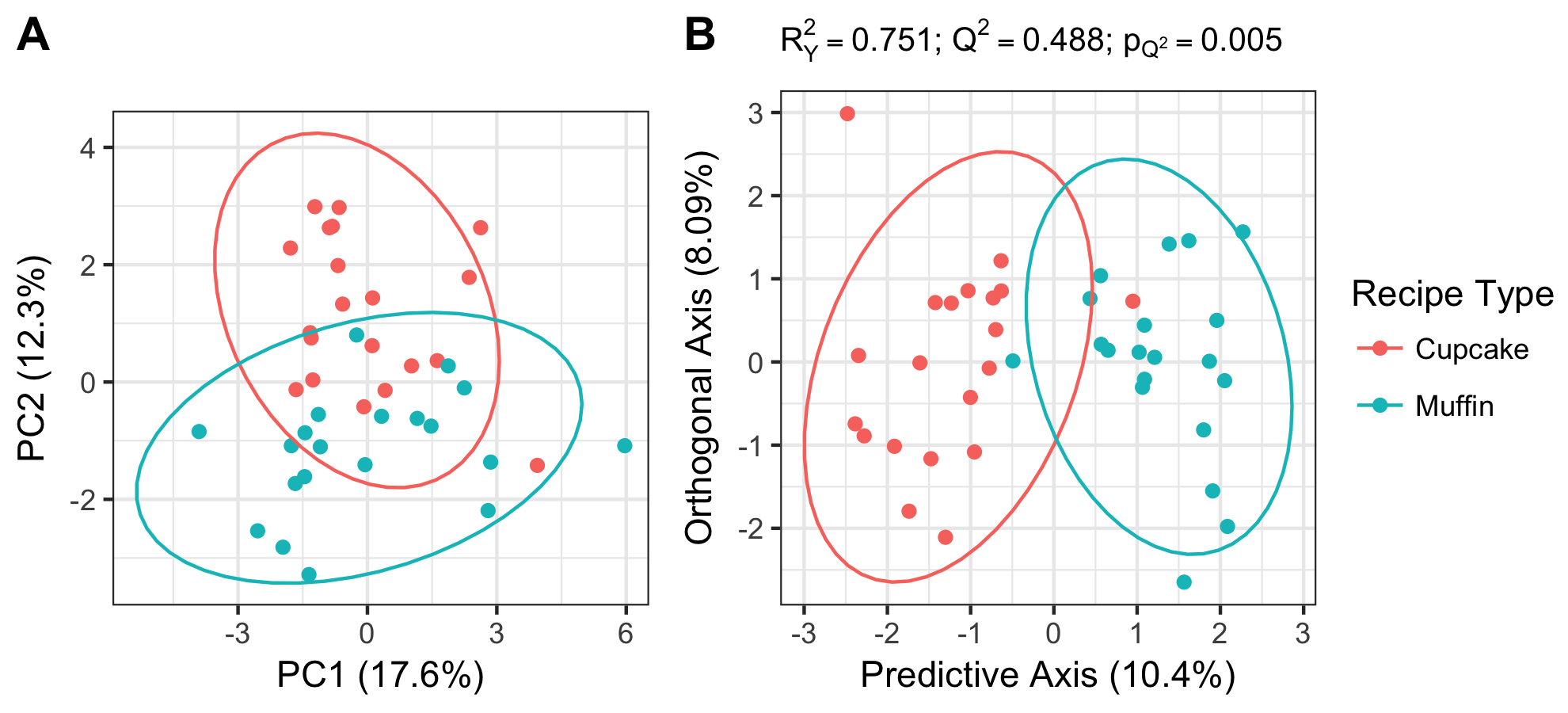


Figure 2: PCA (A) and OPLS-DA (B) on 20 cupcake and 20 muffin recipes.

Table 2: Variable importance in separating muffins from cupcakes. For each variable we report a variable importance in projection (VIP) score, the loading/correlation with the OPLS-DA predictive axis, and the loadings/correlations with the first two PC axes. VIP scores > 1 are generally considered important to separation among groups.



PCA and OPLS-DA both show some separation between cupcakes and muffins, but there are substantial differences in the two methods. First, although PC1 explains the greatest amount of covariation in the ingredients, it does not show any separation between cupcakes and muffins. PC2 shows some weak separation between cupcakes and muffins. This indicates that the variables with the greatest (co)variation in the dataset are not good predictors of the type of baked good. Even when comparing PC2, which separates cupcakes and muffins, with the OPLS-DA predictive axis, there are substantial differences. For example, “unitless” ingredients (e.g. “one sweet potato”, “25 blueberries”) are strongly negatively correlated with PC2 (toward muffins), however, it is not a good predictor of muffins vs. cupcakes as evidenced by its low VIP score and its weak correlation with the OPLS-DA predictive axis. Conversely, “spice” is a good predictor of muffins vs. cupcakes (muffins have more spices than cupcakes) as evidenced by a VIP greater than 1 and a stronger correlation with the OPLS-DA predictive axis, but have a weak correlation with PC2.

# Discussion

Results of simulated data.

Worley and Powers (2016) also compare separation in PCA score space with that in OPLS-DA score space. While the authors show that for OPLS-DA, as model validity decreases, separation of groups in score space is basically unaffected, their suggestion that PCA is a reliable indicator for OPLS-DA reliability is not necessarily true under all situations. When discriminating variables are a small portion of the total number of variables, PCA is especially unreliable at detecting separation. That is, the variables responsible for the greatest amount of variation (explained by PC axis 1) are not necessarily the same variables that reliably distinguish groups. In this situation, poor separation in PCA space is not an indicator that real differences between samples don't exist. In this case, (O)PLS-DA may still be able to detect differences between groups and identify the distinguishing variables with high VIP scores reliably.

“﻿Indeed, discriminant analyses can be prone to ‘overfitting the data’. Overfitting means that the underlying statistical model is too adjusted to the data and becomes ungeneralizable.”(Hervé et al. 2018)

* Discuss (mention?) variations on PLS and their advantages and disadvantages
  + Orthogonal PLS (OPLS, OPLS-DA)
  + Sparse PLS (sPLS), implemented in mixOmics. Does variable selection at the same time as PLS
  + Multi-level PLS, implemented in mixOmics. Allows for nested and repeated-measures designs.
* What to report/not report for PLS
  + MUST report some measure of cross-validation, ideally multiple measures
  + PCA is a good companion, but just because PCA shows no separation, doesn’t invalidate PLS-DA (contrary to recommendations of Worley and Powers (2016)).
  + What % total variation explained by predictive axes? If it’s small, that’s an indicator that discriminating variables don’t vary much, but might still be important
  + DON’T report plots of non-significant PLS-DA. Visual separation in the plot is misleading!
  + NEVER do univariate tests on PLS axis loadings. Use cross-validation to determine if model is significant and explanatory.
* Who should use PLS/PLS-DA? When to use?

# Questions I have (that I should find answers for)

* How exactly is Q2 calculated (from cross-validation, but what does it mean)? What is RMSE?
* What are the alternatives to permutation testing to get p-values implemented in other packages. CV-ANOVA?
* What exactly is VIP? How is it related to the s-plot in SIMCA?

# 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

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

# Writing graveyard:

Unsupervised analyses are powerful tools for describing the variation in a multivariate data set. However, they are not designed with hypothesis testing in mind. Yet, many ecologists use unsupervised analyses in the context of testing relationships with dependent variables either by visually inspecting score plots that result from unsupervised techniques or by using the latent variables they create as independent variables in statistical tests. This method may be sufficient in some situations, especially when latent variables created by unsupervised analyses (e.g. principal component axes) are interpretable. When latent variables aren’t interpretable (e.g. a mix of seemingly biologically unrelated variables) or when the best explanatory variables do not contribute strongly to overall variation in the data, important results will be missed by this strategy of dimensionality reduction followed by visual inspection of score plots and univariate tests on latent variables.

Many multivariate analysis methods exist to describe the variation in multivariate data. For example, PCA (and related PCoA) is an unsupervised technique that describes variation in the data in fewer axes or latent variables than the number of measured variables. This is useful for creating plots of the data in “PCA space” to visualize patterns of variation. Unsupervised analyses are well suited to describing patterns of co-variation. For example the leaf economics spectrum (LES) is a principal component axis that describes [\_\_\_%] of the variation in a set of plant traits and can be interpreted as a continuum from a resource acquisitive strategy to a resource conservative strategy (Wright et al. 2004). However, unsupervised techniques such as PCA are agnostic to response variables such as treatment groups, and the aim of many ecological studies is to describe the response of individuals to a manipulation or along an environmental gradient. Unsupervised techniques like PCA are not well suited for answering these types of questions (despite this, separation of points in PCA space is often used as an indicator of treatment effect in ecological literature).

Supervised analysis techniques ……. What do they do?

However, when variables may be highly correlated (multicollinearity), when sample sizes are small relative to number of variables measured, or when missing values are present, many supervised multivariate analysis techniques don’t perform well.

PCA is an extremely common multivariate analysis technique for dealing with large dimensional ecological data. It aims to describe the (co)variation in the data with fewer axes than there are variables (dimensionality reduction) and is often used to visualize patterns of variation. However, PCA is an unsupervised technique, that is, it is agnostic to any grouping or treatment. Unsupervised analyses are well suited to describing patterns of co-variation. For example the leaf economics spectrum (LES) is a principal component axis that describes [\_\_\_%] of the variation in a set of plant traits and can be interpreted as a continuum from a resource acquisitive strategy to a resource conservative strategy (Wright et al. 2004). However, the aim of many ecological studies is to describe the response of individuals to a manipulation or along an environmental gradient and supervised analyses are required.

Many ecological studies aim to do more than just describe variation in these

Partial least squares regression (PLS) and its discriminant analysis extension (PLS-DA) are supervised multivariate analysis techniques created to be robust to small sample sizes (relative to the number of variables measured), missing values, and multicollinearity [citation to creator, to paper saying its robust]. 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 by chemical ecologists measuring large numbers of metabolites (eco-metabolomics) which may co-vary due to changes in some underlying metabolic pathway. However, the usefulness of this technique is not limited to use on variables of only one type.

Multivariate data can be analyzed in two fundamentally different ways—unsupervised, and supervised analyses. Unsupervised analyses describe patterns in the data and are often used in more descriptive studies. For example, principal component analysis is a widely used unsupervised technique that describes the covariation among variables by creating new axes that represent linear combinations of variables. These axes can sometimes be thought of as “latent variables”. For example the leaf economics spectrum (LES) is a principal component axis that describes covariation in a set of plant traits and ranges from a resource acquisitive strategy to a resource conservative strategy (Wright et al. 2004).

Supervised analyses, on the other hand, are appropriate if the goal is to describe a multivariate response to some predictor variable, or to describe multivariate differences between *a priori* chosen groups such as treatment groups in a manipulative experiment (i.e., discriminant analysis).

Although unsupervised analyses can’t be used to discriminate groups or find relationships in and of themselves, the latent variables they produce are sometimes used by ecologists in further analyses with the intent of discriminating groups. For example, one might test if two plant species differ in their location along the LES axis with a t-test on principal component axis scores. This type of analysis may be justifiable when a principal component axis clearly represents a spectrum of biological variation, such as the LES. However, this strategy of using unsupervised analyses to reduce dimensionality followed by statistical tests is usually not appropriate, as we will demonstrate below.

For a variety of reasons, ecology researchers often use an unsupervised analysis, such as PCA, to reduce dimensionality, then look for visual patterns in labeled datapoints in a two-dimensional plot of new, latent variable axes. To derive some p-value to report, researchers may perform some kind of univariate hypothesis testing on the values of the datapoints along these new latent variables (e.g. PC axis scores). This can result in complicated interpretation of ecologicaly meaningful results because a significant effect of a PC axis does not necessarily convey meaning unless the axis iteslf makes good biological sense. Not only does this obscure the interpretation of results, but it can also lead to incorrect conclusions because unsupervised and supervised analyses are fundamentally different. For example, unsupervised and supervised analyses are likely to lead to different conclusions when the response variables that most strongly influence the independent variable don’t contribute much to overal covariation in the dataset, as we will demonstrate below. Additionally, determining which variables best predict the independent variable (whether it be categorical or continuous) is complicated in unsupervised analyses like PCA since the axes that best explain the independent variable are likely not the linear combination of variables that best explain the independent variable. Therefore, researchers often use post-hoc univariate tests to determine which variables are driving the relationship after seeing separation in PCA space, unnecessarily inflating type I error.

Supervised analyses in [WHO USES CCA AND LDA AND RA?] have been used for a long time. For example, redundancy analysis was devolped in 1968 (Stewart and Lowe) and does some stuff…… However, redundancy analysis performs poorly on datasets with missing values, with small sample sizes (relative to number of variables) or a high degree of multicollinearity (Liu 2011). [EXPLAIN MULTICOLLINEARITY]. PERMANOVA, which is (maybe???) an extension of redundancy analysis, is able to [do some things well], but it requires a distance matrix as an input, and therefore you lose all information about individual variables. It also does not deal with multicollinearity well, as we will demonstrate. PLS doesn’t have these problems.

Partial least squares regression (PLS, also called “projection to latent structures”) and its discriminant analysis extension (PLS-DA), are supervised statistical techniques that work on datasets where the number of variables is greater than the number of samples. The underlying assumption of PLS is that the measured variables are influenced by some process which is driven by a much smaller number of variables. PLS was designed to work with datasets with a high degree of multicollinearity, missing values, and small sample sizes relative to number of variables (Liu 2011). It has therefore gained popularity in metabolomics[citation], a field that regularly deals with datasets of many metabolites that are generated by underlying metabolic pathways. Several statistical software packages have been developed around this technique, specifically for analyzing metabolomic data [SIMCA and metaboanalyst.com]. PLS and its extentions have been adopted by many chemical ecologists[citations], but the usefulness of these techniques is not limited to metabolomic data and is an appropriate approach for answering many ecological questions.

In this primer we intend to demonstrate advantages of PLS over dimensionality reduction followed by univariate hypothesis testing, discuss model validation and hypothesis testing with PLS including important caveats, and demonstrate the use, reporting, and interpretation of PLS results on an ecological dataset.