# Introduction

In recent years, the scale of data available for ecological research has increased due to advancements in high-throughput sampling technology (Kfoury et al. 2017) [other citations], image processing [e.g., morphometrics, herbivory of leaves], automated data logging [ibuttons], remote sensing[citations], and citizen science[citation]. Resulting large, 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 data is of course not a new phenomenon in ecological research. [something about community ecology/ species richness/ vegan package].

Generally, there are two types of questions one might ask using multivariate datasets that require two different statistical approaches—unsupervised and supervised analyses. Unsupervised analyses are appropriate if the main goal of the analysis is to describe the covariation in the data. For example, the leaf economics spectrum (LES) is a spectrum varying from “acquisitive” to “????” growth types in plants. It is not measured directly, but instead represents covariation in a group of correlated morphological and physiological plant traits. Unsupervised analyses, like PCA, are often used to describe the covariation in a multivariate dataset in fewer axes than there are variables (dimensionality reduction) These techniques create new, latent variables that explain the greatest amount of variation through multivariate space. Ideally, this type of analysis results in an axis that makes ecological sense (e.g. the LES), however, interpretation is more difficult when a PC axis can’t be described as a trade-off or continuum that makes biological sense.

Supervised analyses are appropriate if the goal is to describe a multivariate response to some predictor variable, or to describe multivariate differences between treatment groups (discriminant analysis). These types of analyses are also not new to Ecology [something about discriminant function analysis and other techinques/ analagous to bi-variate regression]. However, these techniques cannot be used when the number of variables exceeds the number of samples [because…].

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.

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. PLS was first described in [YEAR, citation] and has since gained popularity in metabolomics[citation], a field that regularly deals with datasets with many more variables (metabolites) than samples. 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.

# Methods (briefly)

## Simulated data methods

I generated a multivariate dataset with a random covariance structure. I am not sure how to “customize” this yet. Here’s the code:

library(MASS)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':  
##   
## select

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

vars = 50 #how many variables?  
vars.diff = 5 #how many of the vars are going to contribute to differences between groups?  
N = 20 #how many samples?  
seed = 100  
set.seed(seed)  
A <- matrix(runif((vars)^2)\*2-1, ncol=vars)  
Sigma <- t(A) %\*% A  
data3 <- mvrnorm(n = N, mu = rep(0, vars), Sigma = Sigma) %>% as.data.frame()

I then assigned group membership for each row.

data4 <- data3 %>%  
 mutate(group = c(rep("a", nrow(data3)/2), rep("b", nrow(data3)/2))) %>% #adds a column, "group", with a's and b's  
 select(group, everything())

Then I created a second data set that added 5 new variables that discriminated between groups

set.seed(seed)  
# add x variables that are based on existing ones but with differences between groups  
x = 5  
# strength of difference (passed to rnorm())  
mu = 5  
data5 <- data4 %>%  
 mutate\_at(vars(num\_range("V", 1:x)),  
 # if its in group a, add a random number, if group b, subtract a random number  
 funs(D = ifelse(group == "a", . + rnorm(1, mu, 1), . - rnorm(1, mu, 1))))

Then I did PCA and PLS-DA on both datasets to produce the figure below. I tried this with several random seeds to purposefully cherry-pick an example where PCA reveals no separation, but PLS-DA is significant. I then did univariate t-tests for each variable to check that there were no differences just by chance in my totally random dataset.

## 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

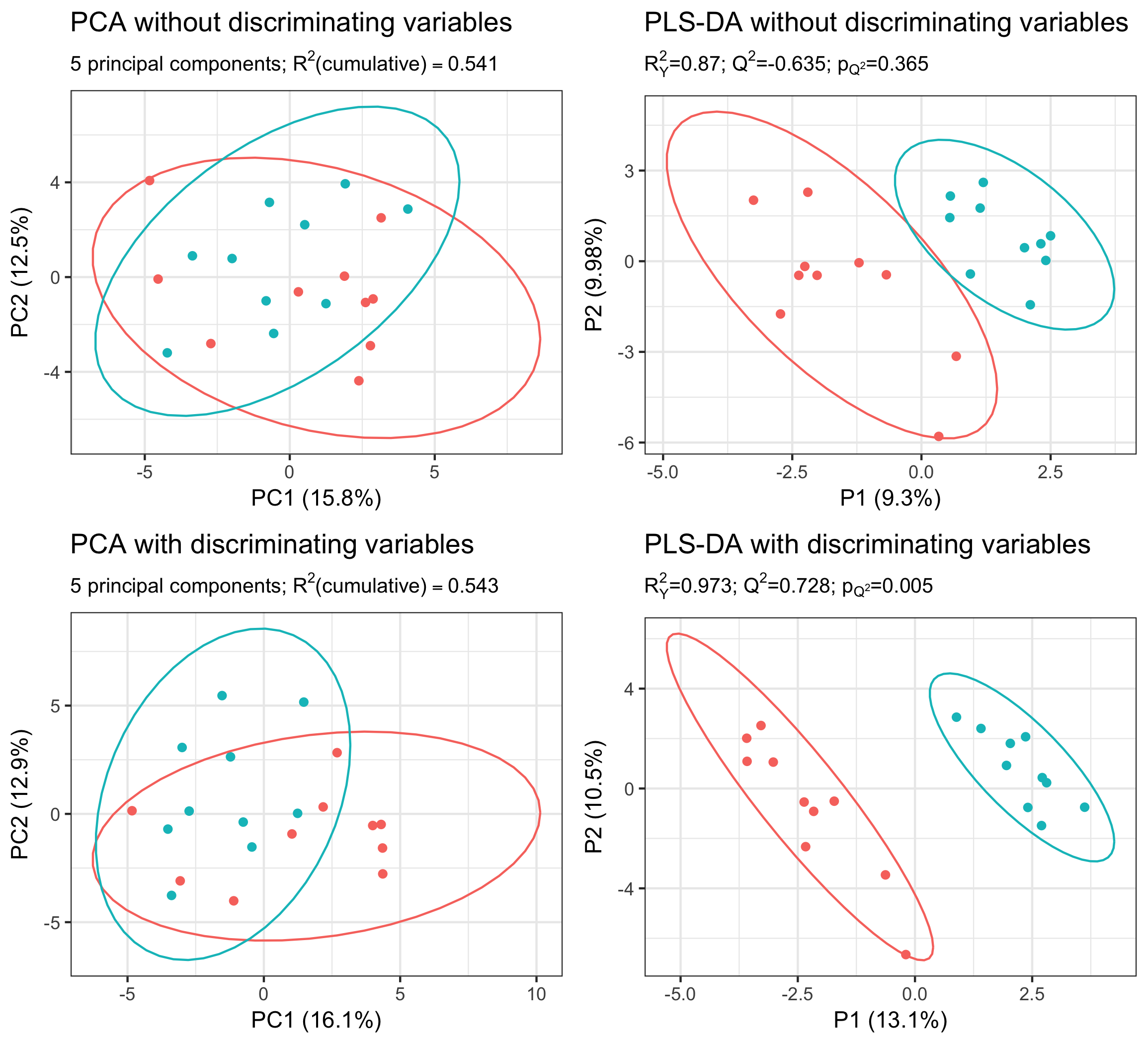


Figure 1: Multivariate analysis of simulated data with random group assignment (A, B) and with 5 aditional variables generated to discriminate between groups (C, D). For PCA plots (A, C), 5 prinicpal components were retained, and the first two principal components are plotted. For PLS-DA plots, the first two predictive axes are plotted, Q2 values are calculated using 7-fold cross validation, and pQ2 is calculated with 200 permutations. Ellipses represent 95% confidence bounds, parenthetical numbers on axis labesl are the percent of total variation explained by the axis. Note, in figure B, 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 55 variables were created to distinguish the groups, the first predictive axis of the PLS-DA on the full data set describes 13.1% of the total variation in the data. In addition, the VIP scores of 5 non-discriminating varibles were over 1. Maybe VIP scores alone shouldn’t be used to determine importance of variables. False-discovery-rate adjusted t-tests more accurately identify distinguishing variables in this case.

Table 1: Table of variables with variable importance in projection (VIP) scores greater than 1. Means for each group are reported for each variable, and p-values from t-tests are reported as raw (p) and false discovery rate adjusted (FDR adjusted). Variables ending in “D” were those generated to discriminate between groups.



## Cupcakes vs. Muffins

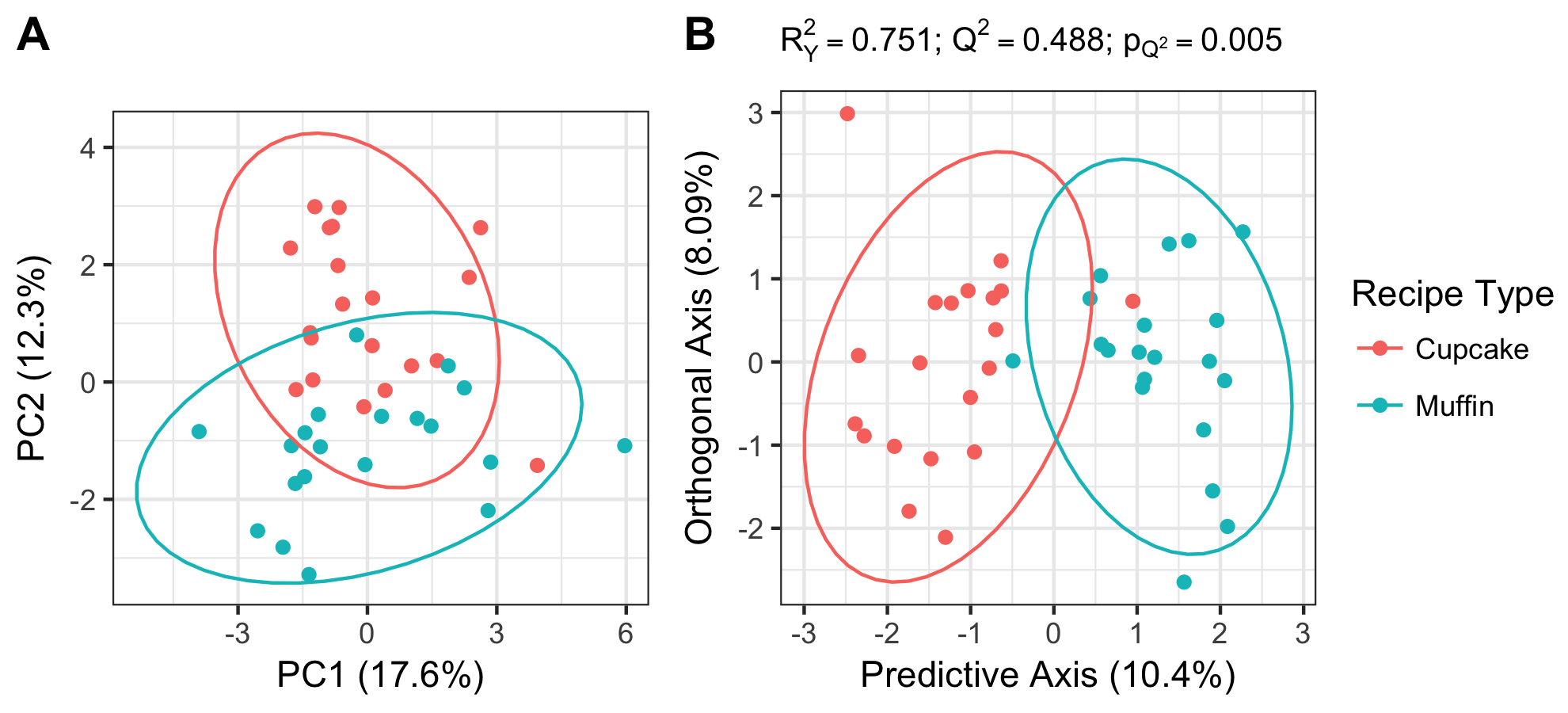


Figure 2: PCA (A) and OPLS-DA (B) on 20 cupcake and 20 muffin recipes.

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

* 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)

* What, if any, are the assumptions of PLS?
* How robust is it to departures from these assumptions?
* 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?
* How do I alter the covariation of a randomly generated dataset? (what’s a covariance matrix?)

# Works Cited

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

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