Paisley Caves *Equus* phalanx analysis

BK McHorse, EB Davis, E Scott, and D Jenkins

February 2016

This analysis is for identifying isolated equid phalanges found coeval with human coprolites in the Paisley Caves of Oregon. We use linear discriminant analysis to determine the likely species of the Paisley Caves fossil horses, using a training set of identified phalanges from contemporaneous faunas. For questions, please contact Brianna McHorse (bmchorse@fas.harvard.edu).

1. Setup and data cleaning
2. Assumption testing
3. Discriminant analysis for Paisley Caves phalanges
4. Stout vs. stilt setup, cleaning, and assumptions
5. Stout vs. stilt logistic regression

All data analysis and figure generation was performed using R *v*3.1.1 (R Core Team, 2014). Data formatting used packages plyr and reshape2 (Wickham, 2007, 2011); discriminant analysis used MASS (Venables and Ripley, 2002); QDA visualization used dr (Weisberg, 2002); plots were created using ggplot2 and Ggally (Wickham, 2009; Schloerke et al., 2011). The code output report was created using knitr (Xie, 2015). Data and annotated code can be found in the online supplemental material.

Section 1: Setup and Data Cleaning

First, set up the workspace and call relevant libraries.

```
rm(list=ls())
setwd("filepath-to-data")
library(MASS)
library(plyr)
library(ggplot2)
library(GGally)
library(reshape2)
library(knitr)
library(dr)
library(cowplot)
```

```
# Bring in the data and check it out.
rawtoes <- read.csv("supplementary_data_2.csv", header = T)
kable(head(rawtoes))
```

Catalogue.Prefix	Specimen.Number	Locality	Stout.Stilt	Genus	Species	Species.simplified	Side	Age	GL	Bp	BFp	Dp	SD	Bd	MinLA	MinLB
F:AM or AMNH	Chan. 48-1085	Dalhart Quarry	Stilt	Equus (Hemionus)	NW Stilt-legged	NW Stilt-legged	?	AD	40.25	35.89	33.40	25.65	30.62	32.77	30.94	37.01
F:AM or AMNH	Chan. 48-1085 D	Dalhart Quarry	Stilt	Equus (Hemionus)	NW Stilt-legged	NW Stilt-legged	?	AD	41.57	37.32	33.76	26.16	34.14	36.71	30.98	39.19
F:AM or AMNH	Chan. 48-1085 G	Dalhart Quarry	Stilt	Equus (Hemionus)	NW Stilt-legged	NW Stilt-legged	?	AD	40.96	37.49	35.72	24.92	34.15	36.47	29.43	37.80
F:AM or AMNH	Chan. 48-1085 E	Dalhart Quarry	Stilt	Equus (Hemionus)	NW Stilt-legged	NW Stilt-legged	?	AD	40.50	40.88	37.21	25.92	37.09	38.53	30.74	37.13
F:AM or AMNH	Chan. 48-1085 C	Dalhart Quarry	Stilt	Equus (Hemionus)	NW Stilt-legged	NW Stilt-legged	?	AD	41.86	41.57	37.55	27.23	37.49	37.48	30.42	39.03
F:AM or AMNH	Chan. 48-1085 A	Dalhart Quarry	Stilt	Equus (Hemionus)	NW Stilt-legged	NW Stilt-legged	?	AD	41.25	40.97	36.82	25.97	36.96	39.14	31.14	37.80

```
str(rawtoes)
```

```
## 'data.frame':    463 obs. of  19 variables:
##  $ Catalogue.Prefix  : Factor w/ 8 levels "CMNFV","F:AM or AMNH",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Specimen.Number   : Factor w/ 407 levels "---","1294-PC-2/3B-24",..: 258 262 265 263 261 259 258 264 260 258 ...
##  $ Locality          : Factor w/ 17 levels "","Bonanza Crk. #73",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Stout.Stilt       : Factor w/ 3 levels "","Stilt","Stout": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Genus             : Factor w/ 2 levels "Equus","Equus (Hemionus)": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Species           : Factor w/ 9 levels "","\"occidentalis\"",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ Species.simplified: Factor w/ 7 levels "","\"occidentalis\"",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ Side              : Factor w/ 6 levels "","?","?L","?R",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Age               : Factor w/ 3 levels "","AD","SA": 2 2 2 2 2 2 2 2 2 2 ...
##  $ GL                : num  40.2 41.6 41 40.5 41.9 ...
##  $ Bp                : num  35.9 37.3 37.5 40.9 41.6 ...
##  $ BFp               : num  33.4 33.8 35.7 37.2 37.5 ...
##  $ Dp                : num  25.6 26.2 24.9 25.9 27.2 ...
##  $ SD                : num  30.6 34.1 34.1 37.1 37.5 ...
##  $ Bd                : num  32.8 36.7 36.5 38.5 37.5 ...
##  $ MinLA             : num  30.9 31 29.4 30.7 30.4 ...
##  $ MinLB             : num  37 39.2 37.8 37.1 39 ...
##  $ Repository        : Factor w/ 7 levels "AMNH","CMN","GCPM",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Notes             : Factor w/ 126 levels "","Abraded dorsally; some measurements not possible.  Drawer VF-239.",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
# Do you want to save high-resolution figures for publication? Set as "yes" if so.
plotflag <- "no"
```

Now, let's create a few levels of cleaned datasets to work with.

```
# Remove specimens with NA in any of the measurements
toes <- rawtoes[complete.cases(rawtoes[,10:17]),]
```

```r
# Remove non-identified specimens
toescleaned <- toes[toes$Species.simplified != "sp.",]
toescleaned$Species.simplified <- droplevels(toescleaned$Species.simplified)

PChorses <- toescleaned[toescleaned$Species.simplified == "",] # Make a data frame just for the Paisley Caves horses
squeakytoes <- toescleaned[toescleaned$Species.simplified != "",] # Remove the Paisley Caves horses from the cleaned data frame.
squeakytoes$Species.simplified <- droplevels(squeakytoes$Species.simplified) # Drop the empty levels.
```

We now have four options for data: **rawtoes**, which is the original dataset; **toes**, which has had incomplete specimens removed; **toescleaned**, which has removed specimens without a species identification; and **squeakytoes**, which is the cleaned dataset but also with the Paisley Caves horses removed.

To summarize our fully cleaned dataset:

```r
# Summaries of locality and species sampling
kable(count(squeakytoes, "Locality"))
```

Locality	freq
Bonanza Crk. #73	2
Dalhart Quarry	34
Eldorado Crk. #45	1
Fossil Lake	18
Hunker Crk. #10	3
Hunker Crk. #51	1
Old Crow R. #152	2
Old Crow R. #163	2
Old Crow R. #74-78	2
Rancho La Brea	112
San Josecito Cave	54
Silver Lake	37

```r
kable(count(squeakytoes, "Species.simplified"))
```

Species.simplified	freq
“occidentalis”	112
conversidens	54
lambei	13
NW Stilt-legged	34
scotti	55

```r
length(squeakytoes[,1]) - 2 # Total number of specimens sampled, excluding two PC specimens
```

```
## [1] 266
```

Section 2: Assumption testing

The main assumptions of discriminant analysis include normally distributed data, equivalent covariance matrices, and independence of data points. The data are independently sampled by default except for the possibility of occasionally including two phalanges from the same animal, but we will test the other assumptions quantitatively.

We begin by testing normality.

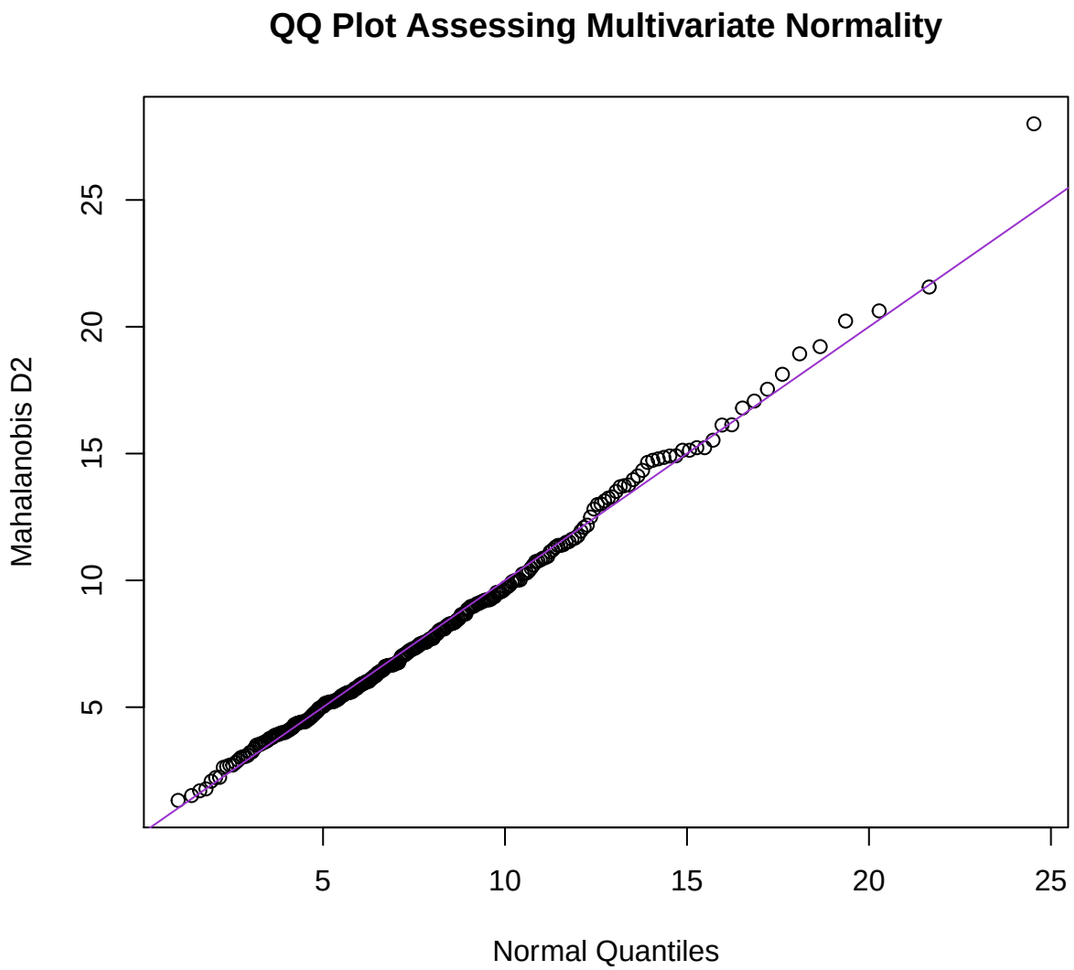
```r
# Within each species, is each measurement normally distributed? Use the Shapiro test, then print the p-value, species, and column.
# p < 0.05 suggests non-normality.
# Using 'squeakytoes' data because singletons cannot be Shapiro tested.

toes.shap <- ddply(squeakytoes, "Species.simplified", summarize,
                   glp = shapiro.test(GL)$p.value,
                   bpp = shapiro.test(Bp)$p.value,
                   bfpp = shapiro.test(BFp)$p.value,
                   dpp = shapiro.test(Dp)$p.value,
                   sdp = shapiro.test(SD)$p.value,
                   bdp = shapiro.test(Bd)$p.value,
                   mlap = shapiro.test(MinLA)$p.value,
                   mlbp = shapiro.test(MinLB)$p.value)

# Round the output to three digits
toes.shap[1:nrow(toes.shap),2:ncol(toes.shap)] <- round(toes.shap[1:nrow(toes.shap),2:ncol(toes.shap)], digits=3)

# Show the table.
kable(toes.shap)
```

Species.simplified	glp	bpp	bfpp	dpp	sdp	bdp	mlap	mlbp
“occidentalis”	0.000	0.008	0.280	0.001	0.245	0.422	0.001	0.000
conversidens	0.680	0.708	0.072	0.955	0.661	0.172	0.418	0.216
lambei	0.279	0.070	0.106	0.115	0.108	0.097	0.217	0.287
NW Stilt-legged	0.343	0.002	0.000	0.565	0.125	0.043	0.995	0.620
scotti	0.285	0.594	0.803	0.421	0.320	0.457	0.088	0.037

The measurements are generally normally distributed.

Now let's test multivariate normality.

```
# Are all measurements normally distributed in multivariate space?
# QQ plot - code from http://www.statmethods.net/stats/anovaAssumptions.html
  m.toes <- as.matrix(log(squeakytoes[,10:17])) # n x p numeric matrix
  center <- colMeans(m.toes) # centroid
  n <- nrow(m.toes); p <- ncol(m.toes); cov <- cov(m.toes);
  d <- mahalanobis(m.toes, center, cov) # distances
  qqplot(qchisq(ppoints(n),df=p),d,
         main="QQ Plot Assessing Multivariate Normality",
         ylab="Mahalanobis D2", xlab="Normal Quantiles")
  abline(a=0,b=1, col="darkorchid")
```



The log-transformed data appear to be fairly multivariately normally distributed, with a few outliers; so, let's log-transform our datasheets.

```
rawtoes[,10:17] <- log(rawtoes[,10:17])
toes[,10:17] <- log(toes[,10:17])
toescleaned[,10:17] <- log(toescleaned[,10:17])
```

```r
squeakytoes[,10:17] <- log(squeakytoes[,10:17])
PChorses[,10:17] <- log(PChorses[,10:17])
```

If you'd like, you can go back and perform the univariate Shapiro tests again to make sure that log-transforming the data has not somehow made any of them significantly deviate from normal. (It has not, in this case.)

Next, we test equivalence of variance.

```r
# Bartlett Test of Homogeneity of Variances
bartlett.toes <- c(glb = bartlett.test(GL ~ Species.simplified, data = squeakytoes)$p.value,
                   bpb = bartlett.test(Bp ~ Species.simplified, data = squeakytoes)$p.value,
                   bfpb = bartlett.test(BFp ~ Species.simplified, data = squeakytoes)$p.value,
                   dpb = bartlett.test(Dp ~ Species.simplified, data = squeakytoes)$p.value,
                   sdb = bartlett.test(SD ~ Species.simplified, data = squeakytoes)$p.value,
                   bdb = bartlett.test(Bd ~ Species.simplified, data = squeakytoes)$p.value,
                   mlab = bartlett.test(MinLA ~ Species.simplified, data = squeakytoes)$p.value,
                   mlbb = bartlett.test(MinLB ~ Species.simplified, data = squeakytoes)$p.value)
print(round(bartlett.toes, digits=3))
```

```
##   glb   bpb  bfpb   dpb   sdb   bdb  mlab  mlbb
## 0.059 0.055 0.000 0.110 0.000 0.003 0.829 0.026
```

For some variables, the hypothesis of equal variance is rejected by the Bartlett test (where $p < 0.05$). We will therefore use quadratic discriminant analysis (QDA) rather than linear discriminant analysis (LDA), as the former does not make the assumption of equal variance/covariance.

Section 3: Discriminant analysis

We will perform a jackknifed QDA, which uses leave-one-out cross-validation to report a more accurate identification accuracy. We then use the discriminant function to predict the identity of the Paisley Caves horses and return the prior probabilities of those predictions.

QDA, cross-validated, even priors, on **squeakytoes**:

```r
# Does not predict new specimens.
qCVfit <- qda(Species.simplified ~ GL + Bp + BFp + Dp + SD + Bd + MinLA + MinLB,
           data=squeakytoes,
           prior = seq(from=1, to=1, length.out=length(levels(squeakytoes$Species.simplified)))/length(levels(squeakytoes$Species.simplified)),
           CV = TRUE)
qCVtab <- table(squeakytoes$Species.simplified, qCVfit$class)
kable(qCVtab) # Matrix of actual vs. predicted IDs
```

	“occidentalis”	conversidens	lambei	NW Stilt-legged	scotti
“occidentalis”	103	1	0	1	7
conversidens	0	50	0	4	0
lambei	0	4	7	0	2
NW Stilt-legged	1	3	0	30	0
scotti	9	0	0	0	46

```
print(diag(round(prop.table(qCVtab, 1), digits=3))) # Proportion correct for each species
```

```
##  "occidentalis"    conversidens          lambei NW Stilt-legged
##           0.920           0.926           0.538           0.882
##          scotti
##           0.836
```

```
print(sum(round((diag(prop.table(qCVtab))), digits=3))) # Total proportion correct
```

```
## [1] 0.881
```

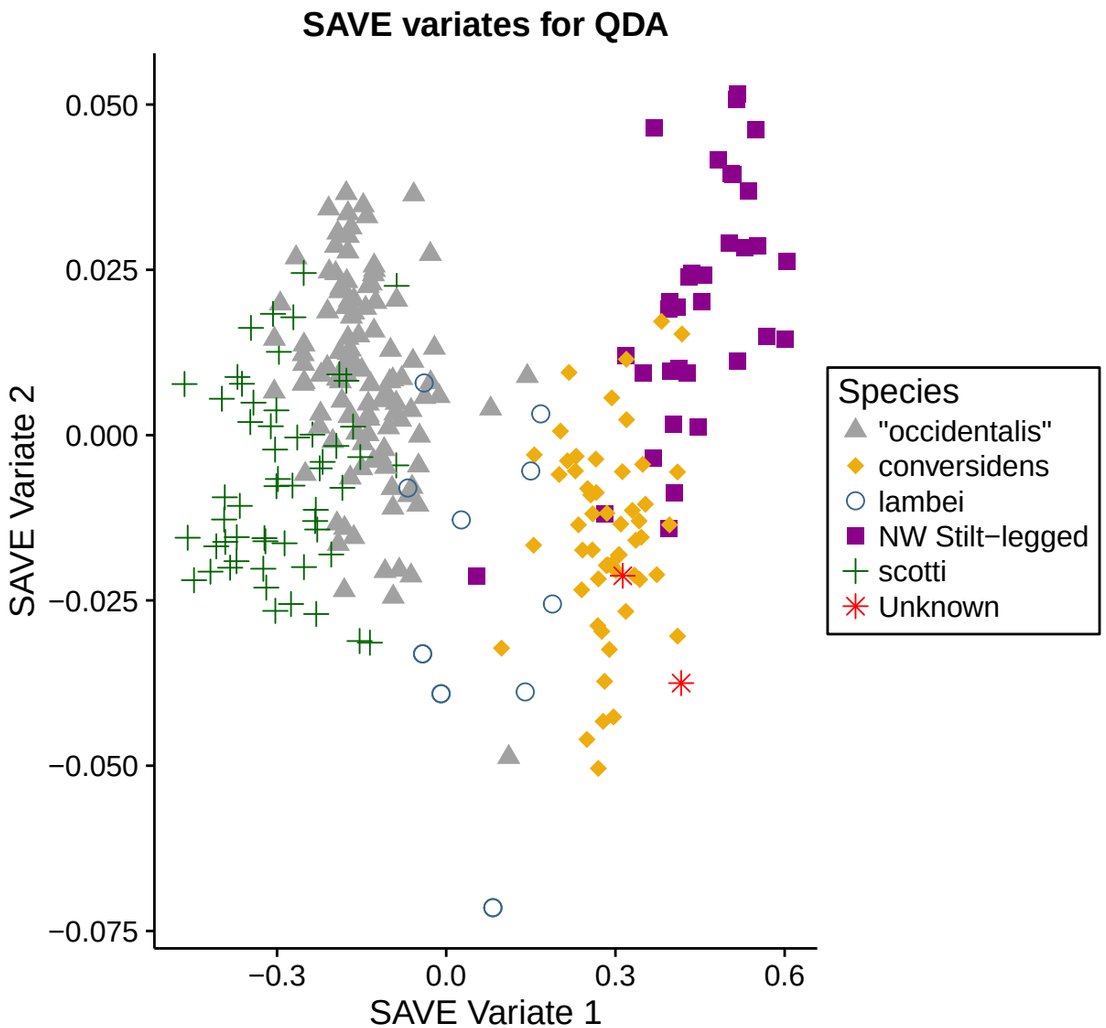
QDA, not cross-validated, even priors, to predict identity of Paisley Caves phalanges:

```
# Now we use non-jackknifed LDA to predict the identity of the PC horses.
qfit <- qda(Species.simplified ~ GL + Bp + BFp + Dp + SD + Bd + MinLA + MinLB,
           data=squeakytoes,
           prior = seq(from=1, to=1, length.out=length(levels(squeakytoes$Species.simplified)))/length(levels(squeakytoes$Species.simplified)),
           CV = FALSE)
qpnew <- predict(qfit, PChorses) # Predict the Paisley Caves horses...
qpostprob <- qpnew$posterior # Assign the posterior probabilities to an object
kable(round(qpostprob, digits = 4)) # Round for easier viewing.
```

	“occidentalis”	conversidens	lambei	NW Stilt-legged	scotti
58	0	0.9931	0	0.0069	0
62	0	1.0000	0	0.0000	0

Finally, a plot to visualize the discriminant space. We’ll be using SAVE variates, which offer visualization for QDA analogous to plotting the first two canonical axes of LDA.

```r
qdaplot <- rbind(squeakytoes, PChorses) # Add the PC horses back in, so we can see them on the plot.

qda.save <- dr(Species.simplified ~ GL + Bp + BFp + Dp + SD + Bd + MinLA + MinLB,
               data=qdaplot,
               na.action=na.omit,
               method = "save")

# Save the first and second SAVE variates and labels. Label the PC horses as Unknown.
variates <- dr.direction(qda.save, which = 1:2)
qy <- dr.y(qda.save)
qy.labels <- as.character(qy)
qy.labels[qy.labels == ""] <- "Unknown"

plotdata <- as.data.frame(cbind(qy, variates))

# Note that for the paper, colors but no positions have been changed.
plotpalette <- c("gray63", "darkgoldenrod2", "steelblue4", "darkmagenta", "darkgreen", "red")

ggplot(plotdata, aes(Dir1,Dir2)) +
  geom_point(aes(color=qy.labels, shape=qy.labels), size=3) +
  scale_shape_manual(name = "Species", values = c(17, 18, 1, 15, 3, 8)) +
  scale_color_manual(name = "Species", values=plotpalette) +
  theme(legend.position = "right",
        legend.background = element_rect(fill="white", size=0.5, linetype="solid", colour ="black")) +
  labs(title = "SAVE variates for QDA", x = "SAVE Variate 1", y = "SAVE Variate 2", color = "Species")
```



Section 4: Stout vs. stilt setup, cleaning, and assumptions

We will create a **SStoes** dataset, which has non-identifications removed. We'll make a second object, **SStoesPC**, that includes the Paisley Caves specimens; finally, we'll create separate **stout** and **stilt** subsets.

```r
# Set up data
SStoes <- toescleaned[toescleaned$Stout.Stilt != "",] # Remove specimens not identified as stout- or stilt-legged
SStoesPC <- rbind(SStoes, PChorses) # Add the Paisley Caves specimens

stout <- SStoes[SStoes$Stout.Stilt == "Stout",]
stilt <- SStoes[SStoes$Stout.Stilt == "Stilt",]


# Within each group, is each measurement normally distributed? Use the Shapiro test, then print the p-value, species, and column.
# p < 0.05 suggests non-normality.

SS.shap <- ddply(SStoes, "Stout.Stilt", summarize,
                      glp = shapiro.test(GL)$p.value,
                      bpp = shapiro.test(Bp)$p.value,
                      bfpp = shapiro.test(BFp)$p.value,
                      dpp = shapiro.test(Dp)$p.value,
                      sdp = shapiro.test(SD)$p.value,
                      bdp = shapiro.test(Bd)$p.value,
                      mlap = shapiro.test(MinLA)$p.value,
                      mlbp = shapiro.test(MinLB)$p.value)

kable(SS.shap)
```

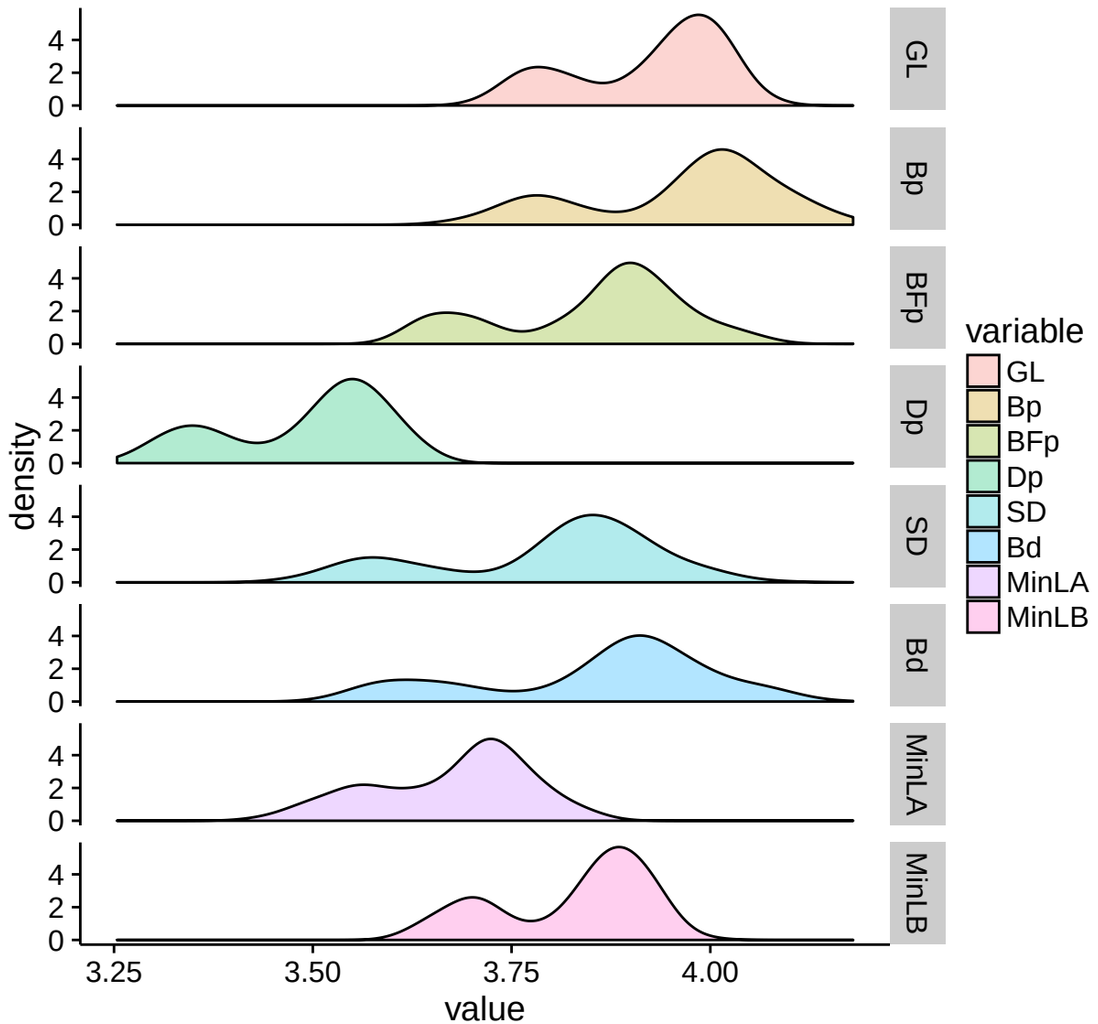
Stout.Stilt	glp	bpp	bfpp	dpp	sdp	bdp	mlap	mlbp
Stilt	0.3558584	0.0135265	0.0021442	0.7341139	0.2226622	0.127868	0.9798398	0.5716486
Stout	0.0000000	0.0000000	0.0000000	0.0000000	0.0000000	0.000000	0.0000017	0.0000000

The stilt-legged horses are normally distributed, but the stout-legged horses are not. We can look at the distribution of measurements to get an idea why. This section formats the data and then plots the distribution of each measurement.

```r
# Make a new frame of stout measurements only
stoutmeasures <- stout[,10:17]
stoutplot <- cbind(stout$Species.simplified, stoutmeasures)
stoutplot <- melt(stoutplot)
```

```
## Using stout$Species.simplified as id variables
```

```
# Plot the stouts
ggplot(stoutplot, aes(x=value, fill=variable)) + geom_density(alpha=.3) +
  facet_grid(variable ~ .) +
  theme(panel.background = element_blank(), panel.grid.minor = element_blank())
```



It appears that most of the variables are bimodally distributed for the stout-legged horses, and this is why the normality assumption fails. While DFA is generally suggested to be robust to violations of normality, we will use a logistic regression instead. Logistic regression performs a similar function but makes fewer assumptions.

Section 5: Stout vs. stilt

We will now perform a logistic regression using **toescleaned**, using measurements to determine stout- vs. stilt-leggedness. First, though, look at the correlations of variables:

```
kable(cor(SStoes[,10:17])) # Look at pairwise correlations among variables
```

	GL	Bp	BFp	Dp	SD	Bd	MinLA	MinLB
GL	1.0000000	0.9300824	0.9137295	0.9616560	0.8788865	0.8782730	0.9708705	0.9823300
Bp	0.9300824	1.0000000	0.9768153	0.9417756	0.9597156	0.9491106	0.9215259	0.9076904
BFp	0.9137295	0.9768153	1.0000000	0.9145659	0.9695930	0.9670727	0.9037038	0.9042725
Dp	0.9616560	0.9417756	0.9145659	1.0000000	0.8852221	0.8734981	0.9378894	0.9404338
SD	0.8788865	0.9597156	0.9695930	0.8852221	1.0000000	0.9842142	0.8536062	0.8834488
Bd	0.8782730	0.9491106	0.9670727	0.8734981	0.9842142	1.0000000	0.8564275	0.8840194
MinLA	0.9708705	0.9215259	0.9037038	0.9378894	0.8536062	0.8564275	1.0000000	0.9519595
MinLB	0.9823300	0.9076904	0.9042725	0.9404338	0.8834488	0.8840194	0.9519595	1.0000000

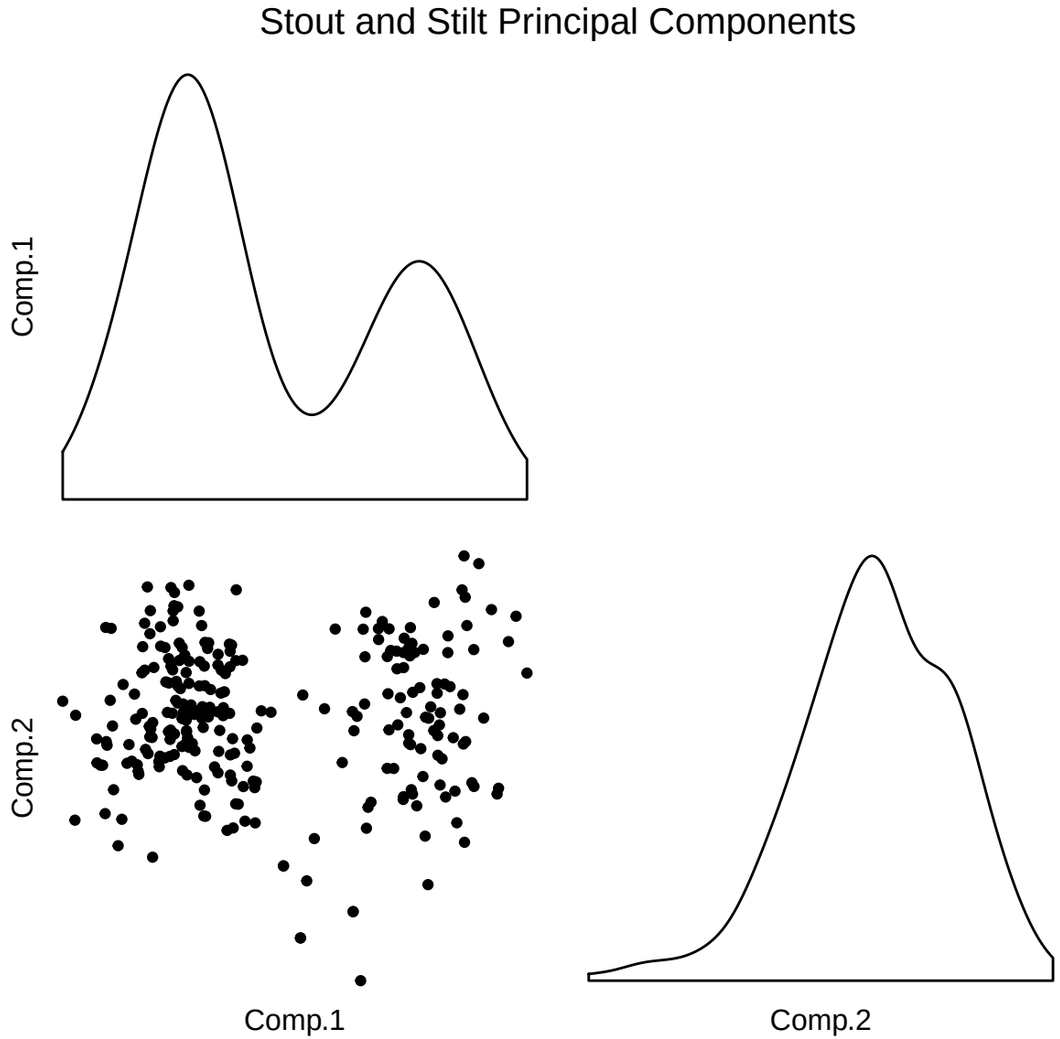
The variables are all quite tightly correlated, violating assumptions of logistic regression. To solve this problem, we will use principal components to collapse the variation into orthogonal axes and then perform logistic regression on the first two PC axes; in this way, we combine the information from all 8 dimensions without violating assumptions.

```
SSpca <- princomp(SStoesPC[,10:17], cor=TRUE)
summary(SSpca) # How much variance included in each principal component?
```

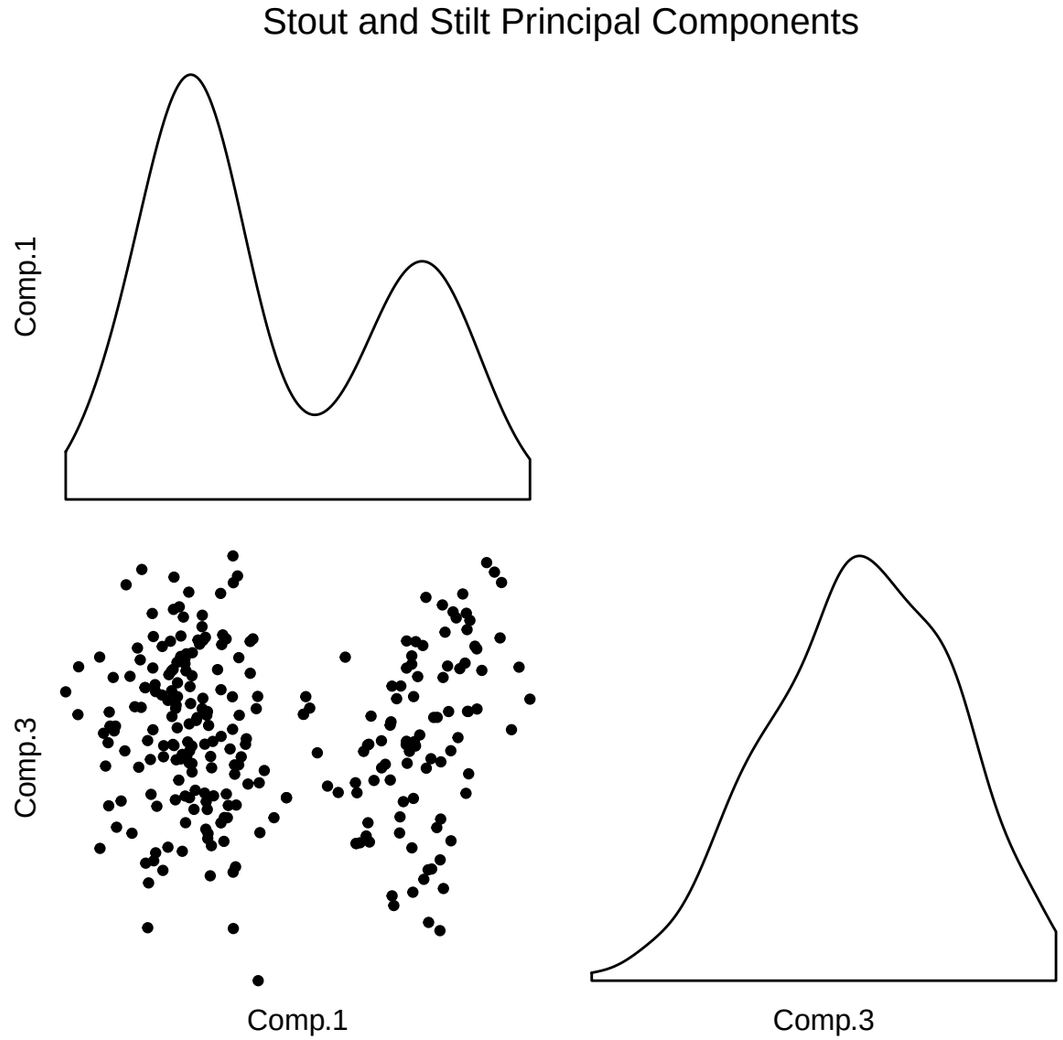
```
## Importance of components:
##                           Comp.1     Comp.2      Comp.3      Comp.4
## Standard deviation     2.7341766 0.56686515 0.280515299 0.242262863
## Proportion of Variance 0.9344652 0.04016701 0.009836104 0.007336412
## Cumulative Proportion  0.9344652 0.97463225 0.984468354 0.991804766
##                            Comp.5      Comp.6     Comp.7      Comp.8
## Standard deviation     0.156614516 0.136014031 0.11628792 0.094926659
## Proportion of Variance 0.003066013 0.002312477 0.00169036 0.001126384
## Cumulative Proportion  0.994870779 0.997183256 0.99887362 1.000000000
```

```
PCs <- SSpca$scores # Save scores
SStoesPC <- cbind(SStoesPC, PCs) # Add the principal component scores to the data frame
```

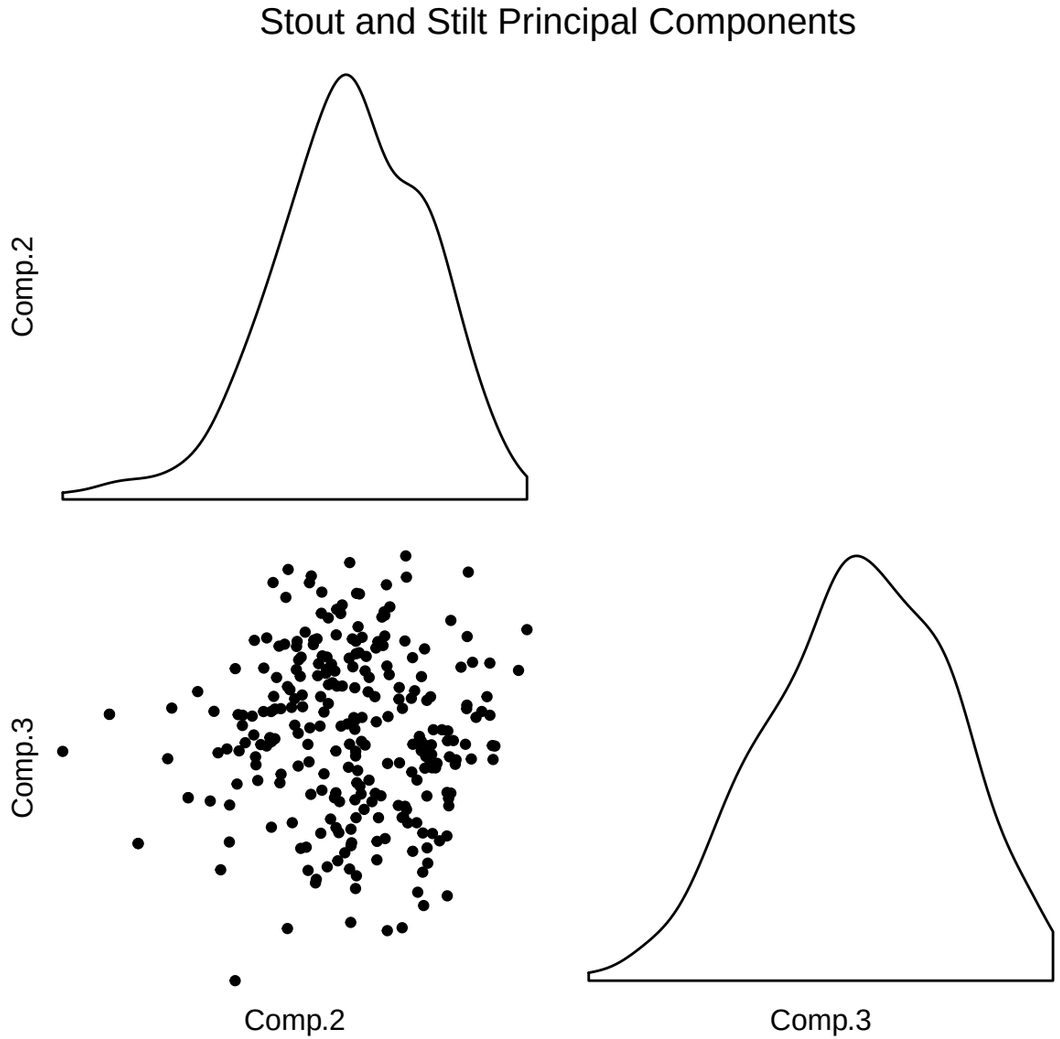
```r
# Visualize the first two principal components
ggpairs(SStoesPC[SStoesPC$Stout.Stilt != "",], # Remove the PC horses so they don't throw off the colors
        columns=c("Comp.1", "Comp.2"),
        colour='Stout.Stilt',
        title="Stout and Stilt Principal Components",
        lower=list(continuous='points'),
        axisLabels='none',
        upper=list(continuous='blank'),
        legends=T)
```


Stout and Stilt Principal Components
Comp.1
Comp.2
Comp.1
Comp.2

```r
# Visualize the first and third principal components
ggpairs(SStoesPC[SStoesPC$Stout.Stilt != "",], # Remove the PC horses so they don't throw off the colors
        columns=c("Comp.1", "Comp.3"),
        colour='Stout.Stilt',
        title="Stout and Stilt Principal Components",
        lower=list(continuous='points'),
        axisLabels='none',
        upper=list(continuous='blank'),
        legends=T)
```


Stout and Stilt Principal Components
Comp.1
Comp.3
Comp.1
Comp.3

```r
# Visualize the second and third principal components
ggpairs(SStoesPC[SStoesPC$Stout.Stilt != "",], # Remove the PC horses so they don't throw off the colors
        columns=c("Comp.2", "Comp.3"),
        colour='Stout.Stilt',
        title="Stout and Stilt Principal Components",
        lower=list(continuous='points'),
        axisLabels='none',
        upper=list(continuous='blank'),
        legends=T)
```


Stout and Stilt Principal Components
Comp.2
Comp.3
Comp.2
Comp.3

We now perform the logistic regression.

```
# First we split apart the Paisley Caves unknowns and the training set again.
SSnoPC <- SStoesPC[SStoesPC$Repository != "OSMA",]
PCpc <- SStoesPC[SStoesPC$Repository == "OSMA",]

# Now we fit the logistic, which is a specific form of the generalized linear model command.
SSfit <- glm(Stout.Stilt ~ Comp.1 + Comp.2 + Comp.3 + Comp.4 + Comp.5 + Comp.6 +
               Comp.7 + Comp.8, data = SSnoPC, family = binomial())

summary(SSfit) # Summarize the logistic model
```

```
##
## Call:
## glm(formula = Stout.Stilt ~ Comp.1 + Comp.2 + Comp.3 + Comp.4 +
##     Comp.5 + Comp.6 + Comp.7 + Comp.8, family = binomial(), data = SSnoPC)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4658   0.0017   0.0097   0.0484   1.3704
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   7.4836     1.8888   3.962 7.43e-05 ***
## Comp.1       -1.6978     0.4647  -3.654 0.000258 ***
## Comp.2        1.3724     1.0683   1.285 0.198924
## Comp.3       -9.8047     3.7018  -2.649 0.008083 **
## Comp.4       -8.8996     3.1053  -2.866 0.004157 **
## Comp.5       -4.5017     3.4790  -1.294 0.195681
## Comp.6       -3.5381     4.5436  -0.779 0.436157
## Comp.7        2.9700     4.7454   0.626 0.531394
## Comp.8       -5.3299     8.2486  -0.646 0.518176
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 203.886  on 267  degrees of freedom
## Residual deviance:  29.382  on 259  degrees of freedom
## AIC: 47.382
##
## Number of Fisher Scoring iterations: 9
```

```r
SSnoPC$IsStout <- SSnoPC$Stout.Stilt == "Stout"
SSpredict <- predict(SSfit, SSnoPC, type = "response") # Use the logistic model to predict on the training set.
SSresults <- table(actual.stout = SSnoPC$IsStout, predicted.stout = SSpredict > 0.5)
SSresults # Confusion matrix. "False" vs. "True" indicates whether or not the specimen is (or is predicted to be) stout-legged.
```

```
##             predicted.stout
## actual.stout FALSE TRUE
##        FALSE    31    3
##        TRUE      1  233
```

```r
sum(diag(prop.table(SSresults))) # Shows total percent correct.
```

```
## [1] 0.9850746
```

```r
# Next, we predict the identity of the Paisley Caves phalanges.
PCpredict <- predict(SSfit, PCpc, type = "response")
PCpredicts <- PCpredict > 0.5 # Where TRUE predicts stout-legged and FALSE suggests stilt-legged.
PCpredicts
```

```
##   58   62
## TRUE TRUE
```

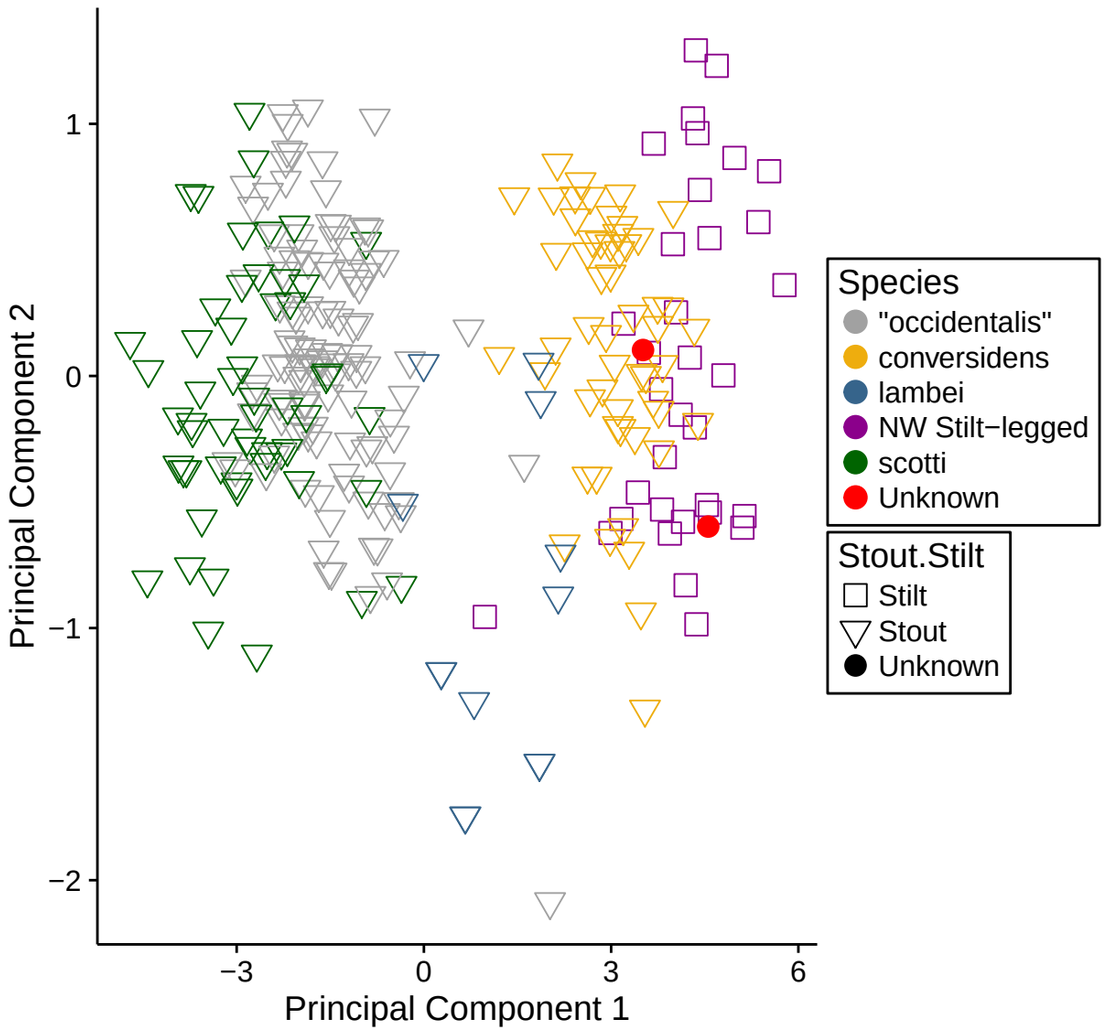
```r
PCpredict <- round(PCpredict, digits = 3) # Round strength of predictions
PCpredict # Gives probability of "TRUE" for each specimen, i.e., probability of being stout-legged.
```

```
##    58    62
## 0.967 1.000
```

The stout vs. stilt logistic predicts both Paisley Caves specimens as stout-legged. See paper for discussion; here we plot the principal components of each species, as well as the unknown Paisley Caves horses (red dots), to explore why the logistic makes these predictions.

```r
# We'll add the PC horses back in and label their species and stout/stilt status as Unknown.
SStoesPC$Species.simplified <- as.character(SStoesPC$Species.simplified)
SStoesPC[SStoesPC$Species.simplified == "",7] <- "Unknown"
SStoesPC$Species.simplified <- as.factor(SStoesPC$Species.simplified)

SStoesPC$Stout.Stilt <- as.character(SStoesPC$Stout.Stilt)
SStoesPC[SStoesPC$Stout.Stilt == "",4] <- "Unknown"
```

```r
SStoesPC$Stout.Stilt <- as.factor(SStoesPC$Stout.Stilt)

# Visualize the first two principal components

SSplot <- cbind(SStoesPC[,20:27])
SSplot$Stout.Stilt <- SStoesPC$Stout.Stilt
SSplot$Species <- SStoesPC$Species.simplified

ggplot(SSplot, aes(Comp.1, Comp.2)) +
  geom_point(aes(color=Species, shape=Stout.Stilt), size=4) +
  scale_shape_manual(values = c(0, 6, 16)) +
  scale_color_manual(name = "Species", values=plotpalette) +
  theme(legend.position = "right",
        legend.background = element_rect(fill="white", size=0.5, linetype="solid", colour ="black")) +
  labs(x = "Principal Component 1", y = "Principal Component 2")
```


Principal Component 2
1
0
−1
−2
−3
0
3
6
Principal Component 1
Species
"occidentalis"
conversidens
lambei
NW Stilt−legged
scotti
Unknown
Stout.Stilt
Stilt
Stout
Unknown

References

R Core Team. 2014. R: A Language and Environment for Statistical Computing (version 3.1.0). Vienna: R Foundation for Statistical Computing.

Schloerke, B., J. Crowley, D. Cook, H. Hofmann, H. Wickham, F. Briatte, and M. Marbach. 2011. Ggally: Extension to ggplot2.

Venables, W. N., and B. D. Ripley. 2002. Modern Applied Statistics with S. Springer-Verlag.

Weisberg, S. 2002. Dimension reduction regression in R. Journal of Statistical Software 7:1-22.

Wickham, H. 2007. Reshaping data with the reshape package. Journal of Statistical Software 21:1-20.

Wickham, H. 2009. ggplot2: Elegant Graphics for Data Analysis. Springer, 221 pp.

Wickham, H. 2011. The split-apply-combine strategy for data analysis. Journal of Statistical Software 40:1-29.

Xie, M. Y. 2015. Knitr: A General-Purpose Package for Dynamic Report Generation in R.