Common Principal Component Analysis: Palmer Penguins

Benjamin Draves

10/19/2020

```
#Plot original dataset
scale <- 5

hist_top <- ggplot(penguins, aes(flipper_length_mm)) +
  geom_histogram(color = 'black', alpha = .8, fill = 'orange') +
  labs(x = '', y = '',title = 'Penguin Size') + theme_bw()

hist_right <- ggplot(penguins, aes(body_mass_g)) +
  geom_histogram(color = 'black', alpha = .8, fill = 'orange') +
  labs(x = '', y = '') + theme_bw()+ coord_flip()

scatter <- ggplot(penguins, aes(flipper_length_mm, body_mass_g))+
  geom_point(col = 'orange', alpha = .8)+  theme_bw()+
  labs(x = 'Flipper Length (mm)', y = 'Body Mass (g)')

empty <- ggplot()+geom_point(aes(1,1), colour="white")+
         theme(axis.ticks=element_blank(),
               panel.background=element_blank(),
               axis.text.x=element_blank(), axis.text.y=element_blank(),
               axis.title.x=element_blank(), axis.title.y=element_blank())

pdf(paste0(fig_path, 'penguin_size.pdf'), width = 7, height = 7)
grid.arrange(hist_top, empty, scatter, hist_right, ncol=2, nrow=2, widths=c(4, 1), heights=c(1, 4))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 2 rows containing non-finite values (stat_bin).
## Warning: Removed 2 rows containing missing values (geom_point).
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

```
dev.off()
```

```
## pdf
##   2
```

```
png(paste0(fig_path, 'penguin_size.png'))
grid.arrange(hist_top, empty, scatter, hist_right, ncol=2, nrow=2, widths=c(4, 1), heights=c(1, 4))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Warning: Removed 2 rows containing non-finite values (stat_bin).
## Warning: Removed 2 rows containing missing values (geom_point).
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

```
dev.off()
```

```
## pdf
##   2
```

```
#Plot centered dataset
hist_top <- ggplot(penguins, aes((flipper_length_mm - mean(flipper_length_mm, na.rm = TRUE))/sd(flipper
  geom_histogram(color = 'black', alpha = .8, fill = 'turquoise') +
  labs(x = '', y = '', title = 'Center/Scaled Penguin Size') + theme_bw()

hist_right <- ggplot(penguins, aes((body_mass_g- mean(body_mass_g, na.rm = TRUE))/sd(body_mass_g, na.rm
  geom_histogram(color = 'black', alpha = .8, fill = 'turquoise') +
  labs(x = '', y = '') + theme_bw()+ coord_flip()+theme(plot.margin=unit(c(scale, scale, scale, -2*scal

scatter <- ggplot(penguins, aes((flipper_length_mm - mean(flipper_length_mm, na.rm = TRUE))/sd(flipper_
                                (body_mass_g - mean(body_mass_g, na.rm = TRUE))/sd(body_mass_g, na.rm =
  geom_point(col = 'turquoise', alpha = .8)+  theme_bw()+
  labs(x = 'Center/Scaled Flipper Length (mm)', y = 'Center/Scaled Body Mass (g)')

pdf(paste0(fig_path, 'penguin_size_centered.pdf'), width = 7, height = 7)
grid.arrange(hist_top, empty, scatter, hist_right, ncol=2, nrow=2, widths=c(4, 1), heights=c(1, 4))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

```
dev.off()
```

```
## pdf
##   2
```

```
png(paste0(fig_path, 'penguin_size_centered.png'))
grid.arrange(hist_top, empty, scatter, hist_right, ncol=2, nrow=2, widths=c(4, 1), heights=c(1, 4))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

```
dev.off()
```

```
## pdf
##   2
```

```
#carry out dim reduction
X <- cbind(penguins$flipper_length_mm, penguins$body_mass_g)
X <- X[!is.na(X[,1]),]
cmean <- apply(X, 2, mean)
```

```r
Xc <- apply(X, 2, function(x) (x - mean(x))/sd(x))
Xc.svd <- svd(Xc)
PC <- Xc %*% Xc.svd$v
pc_dat <- data.frame(PC1 = PC[,1], PC2 = PC[,2])

#plot principal directions over original plot
v1 <- Xc.svd$v[,1]; v2 <- Xc.svd$v[,2]
s1  <- .1 * sqrt(Xc.svd$d[1])
s2  <- -.1 * sqrt(Xc.svd$d[2])

scatter2 <- scatter +
  geom_segment(aes(x = 0, y = 0, xend = s1 * v1[1], yend = s1 * v1[2]),
                arrow = arrow(length = unit(0.2, "cm")),
                color = 'blueviolet')+
  geom_segment(aes(x = 0, y = 0, xend = s2 * v2[1], yend = s2 * v2[2]),
                arrow = arrow(length = unit(0.2, "cm")),
                color = 'blueviolet')

pdf(paste0(fig_path, 'penguin_size_pc.pdf'), width = 7, height = 7)
grid.arrange(hist_top, empty, scatter2, hist_right, ncol=2, nrow=2, widths=c(4, 1), heights=c(1, 4))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

```r
dev.off()
```

```
## pdf
##   2
```

```r
png(paste0(fig_path, 'penguin_size_pc.png'))
grid.arrange(hist_top, empty, scatter2, hist_right, ncol=2, nrow=2, widths=c(4, 1), heights=c(1, 4))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing non-finite values (stat_bin).
```

```r
dev.off()
```

```
## pdf
##   2
```

```r
#Plot principal components
hist_top <- ggplot(pc_dat, aes(PC1)) +
  geom_histogram(color = 'black', alpha = .8, fill = 'purple') +
  xlim(min(PC[,1]) - .2, max(PC[,1]) + .2) +
  labs(x = '', y = '', title = 'Principal Components') + theme_bw()
```

```r
hist_right <- ggplot(pc_dat, aes(PC2)) +
  geom_histogram(color = 'black', alpha = .8, fill = 'purple') +
  xlim(min(PC[,1]) - .2, max(PC[,1]) + .2) +
  labs(x = '', y = '') + theme_bw() + coord_flip()

scatter <- ggplot(pc_dat, aes(PC1, PC2))+
  geom_point(col = 'purple', alpha = .8)+  theme_bw()+
  ylim(min(PC[,1]) - .2, max(PC[,1]) + .2) +
  xlim(min(PC[,1]) - .2, max(PC[,1]) + .2) +
  labs(x = 'Principal Component 1', y = 'Principal Component 2')

pdf(paste0(fig_path, 'penguin_size_pc_plot.pdf'), width = 7, height = 7)
grid.arrange(hist_top, empty, scatter, hist_right, ncol=2, nrow=2, widths=c(4, 1), heights=c(1, 4))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

```r
dev.off()
```

```
## pdf
##   2
```

```r
png(paste0(fig_path, 'penguin_size_pc_plot.png'))
grid.arrange(hist_top, empty, scatter, hist_right, ncol=2, nrow=2, widths=c(4, 1), heights=c(1, 4))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2 rows containing missing values (geom_bar).
```

```r
dev.off()
```

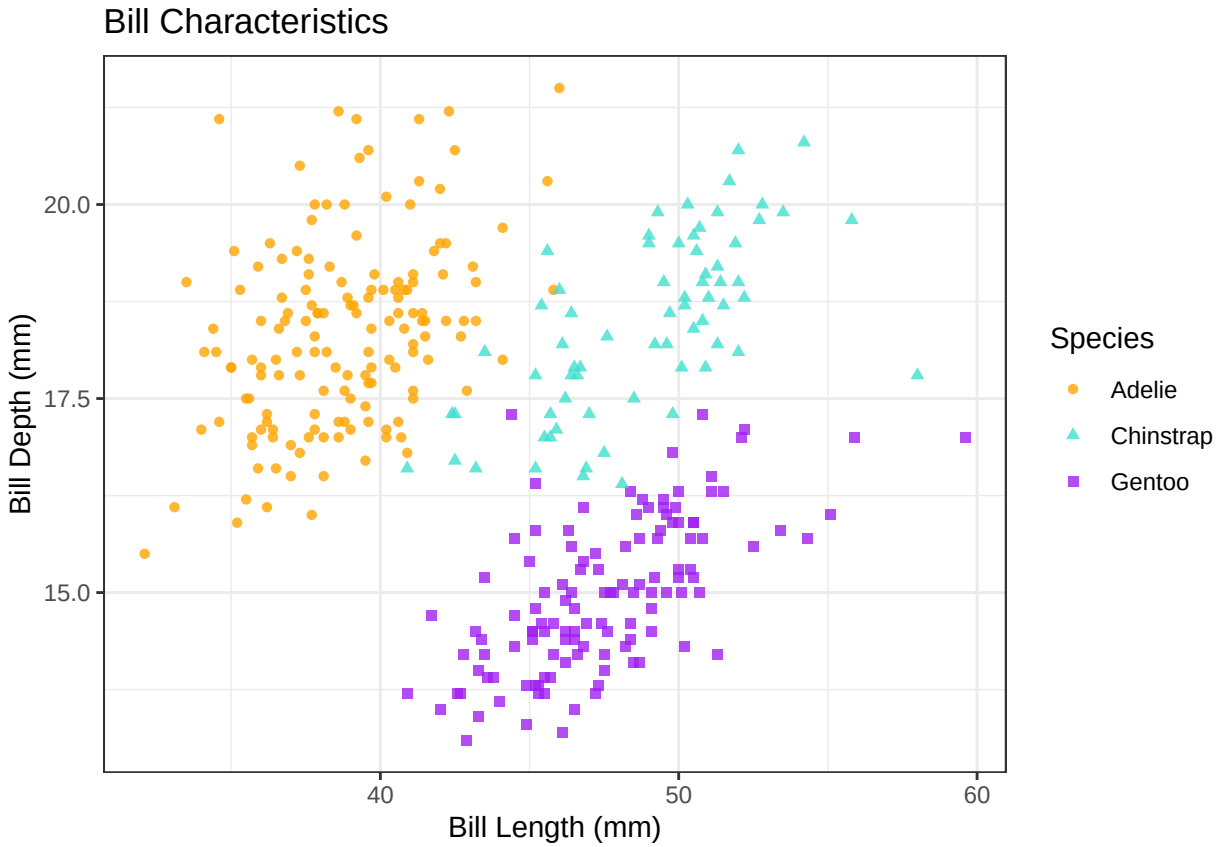
```
## pdf
##   2
```

```r
fig_path <- '/Users/benjamindraves/Desktop/CPCA/figures/cpca/'

#center and scale by group
scaled_penguins <- penguins %>%
  group_by(species) %>%
  mutate(flipper_length_scaled = (flipper_length_mm - mean(flipper_length_mm, na.rm = TRUE))/sd(flipper
         bill_length_scaled = (bill_length_mm - mean(bill_length_mm, na.rm = TRUE))/sd(bill_length_mm,
         bill_depth_scaled = (bill_depth_mm - mean(bill_depth_mm, na.rm = TRUE))/sd(bill_depth_mm, na.r
         )

#Bill Depth vs Bill Length
ggplot(penguins, aes(bill_length_mm, bill_depth_mm, color = species, shape = species))+
  geom_point(alpha = .8) +
  theme_bw()+ scale_color_manual(values=c('orange', 'turquoise', 'purple'))+
  labs(x = 'Bill Length (mm)', y = 'Bill Depth (mm)', title = 'Bill Characteristics',
       color = 'Species', shape ='Species')
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



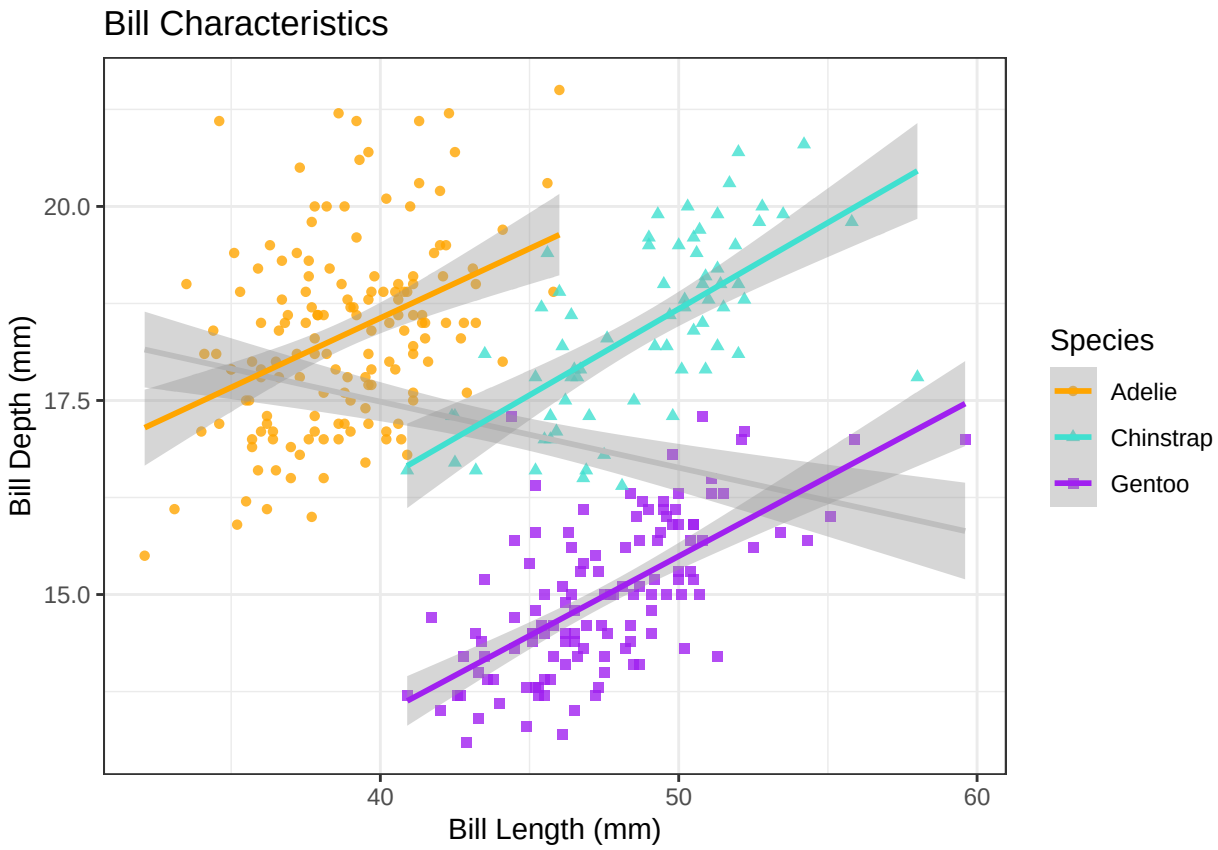

```
ggsave('bill_length_v_depth.pdf', path = fig_path, width = 6, height = 5)
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

```
ggplot(penguins, aes(bill_length_mm, bill_depth_mm))+
  geom_point(aes(color = species, shape = species), alpha = .8) +
  geom_smooth(method = 'lm', col = 'grey')+
  geom_smooth(aes(color = species), method = 'lm') +
  theme_bw()+ scale_color_manual(values=c('orange', 'turquoise', 'purple'))+
  labs(x = 'Bill Length (mm)', y = 'Bill Depth (mm)',  title = 'Bill Characteristics',
       color = 'Species', shape ='Species')
```

```
## `geom_smooth()` using formula 'y ~ x'
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
## `geom_smooth()` using formula 'y ~ x'
## Warning: Removed 2 rows containing non-finite values (stat_smooth).

## Warning: Removed 2 rows containing missing values (geom_point).
```




```
ggsave('simpsons_paradox.pdf', path = fig_path, width = 6, height = 5)
```

```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

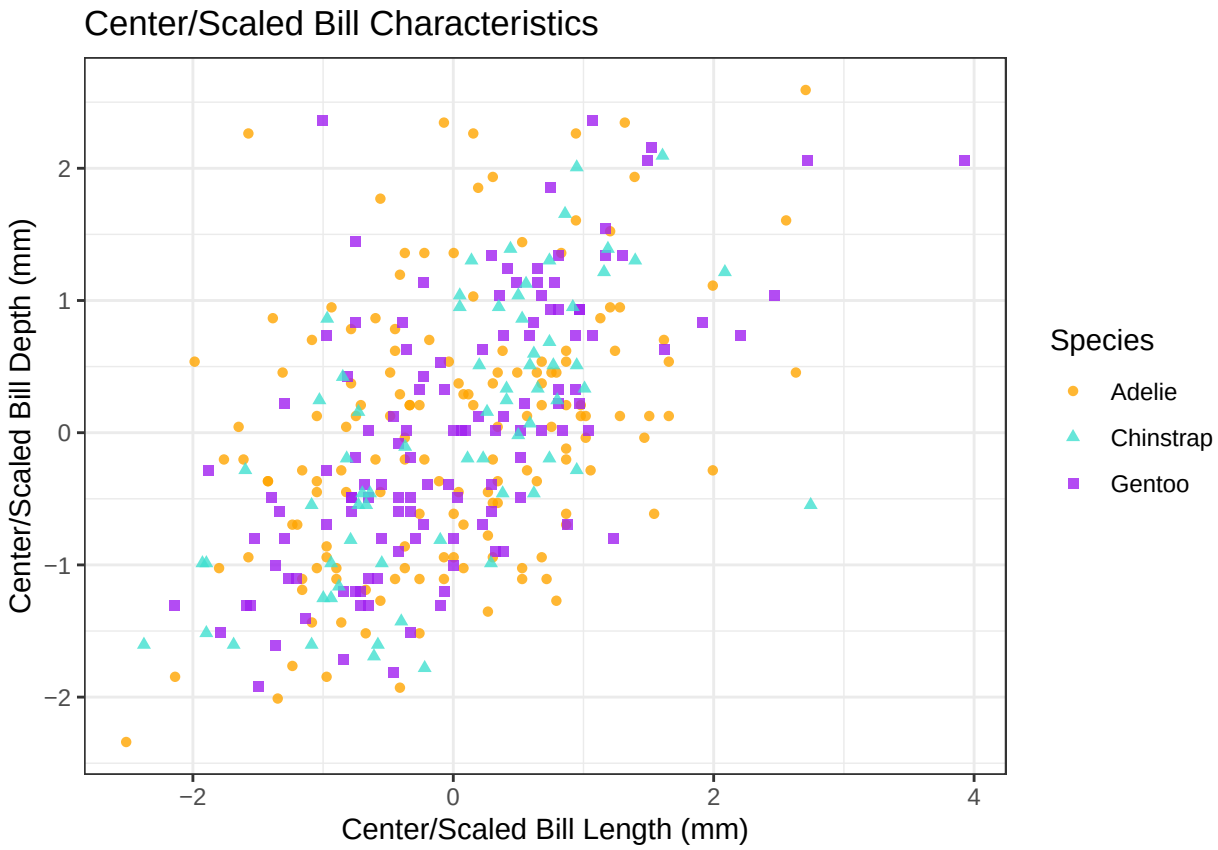
```
## `geom_smooth()` using formula 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

```
ggplot(scaled_penguins, aes(bill_length_scaled, bill_depth_scaled, color = species))+
  geom_point(aes(shape = species), alpha = .8) +
  theme_bw()+ scale_color_manual(values=c('orange', 'turquoise', 'purple'))+
  labs(x = 'Center/Scaled Bill Length (mm)', y = 'Center/Scaled Bill Depth (mm)', title = 'Center/Scale
       color = 'Species', shape ='Species')
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```




```
ggsave('bill_length_v_depth_scaled.pdf', path = fig_path, width = 6, height = 5)
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```

```
# Extract each X
X.Adelie <- scaled_penguins %>% select(bill_length_scaled, bill_depth_scaled) %>%
  filter(species == 'Adelie') %>% ungroup() %>%
  select(-'species') %>% na.omit() %>% as.matrix
```

```
## Adding missing grouping variables: `species`
```

```
X.Chinstrap <- scaled_penguins %>% select(bill_length_scaled, bill_depth_scaled) %>%
  filter(species == 'Chinstrap') %>% ungroup() %>%
  select(-'species') %>% na.omit() %>% as.matrix
```

```
## Adding missing grouping variables: `species`
```

```
X.Gentoo <- scaled_penguins %>% select(bill_length_scaled, bill_depth_scaled) %>%
  filter(species == 'Gentoo') %>% ungroup() %>%
  select(-'species') %>% na.omit() %>% as.matrix
```

```
## Adding missing grouping variables: `species`
```

```
# Estimate principal directions
X.ext <- rbind(X.Adelie, X.Chinstrap, X.Gentoo)
V <- svd(X.ext)$v

# Variance by population
Lambda <- list()
Lambda[[1]] <- diag(crossprod(X.Adelie %*% V) / (nrow(X.Adelie) - 1))
```

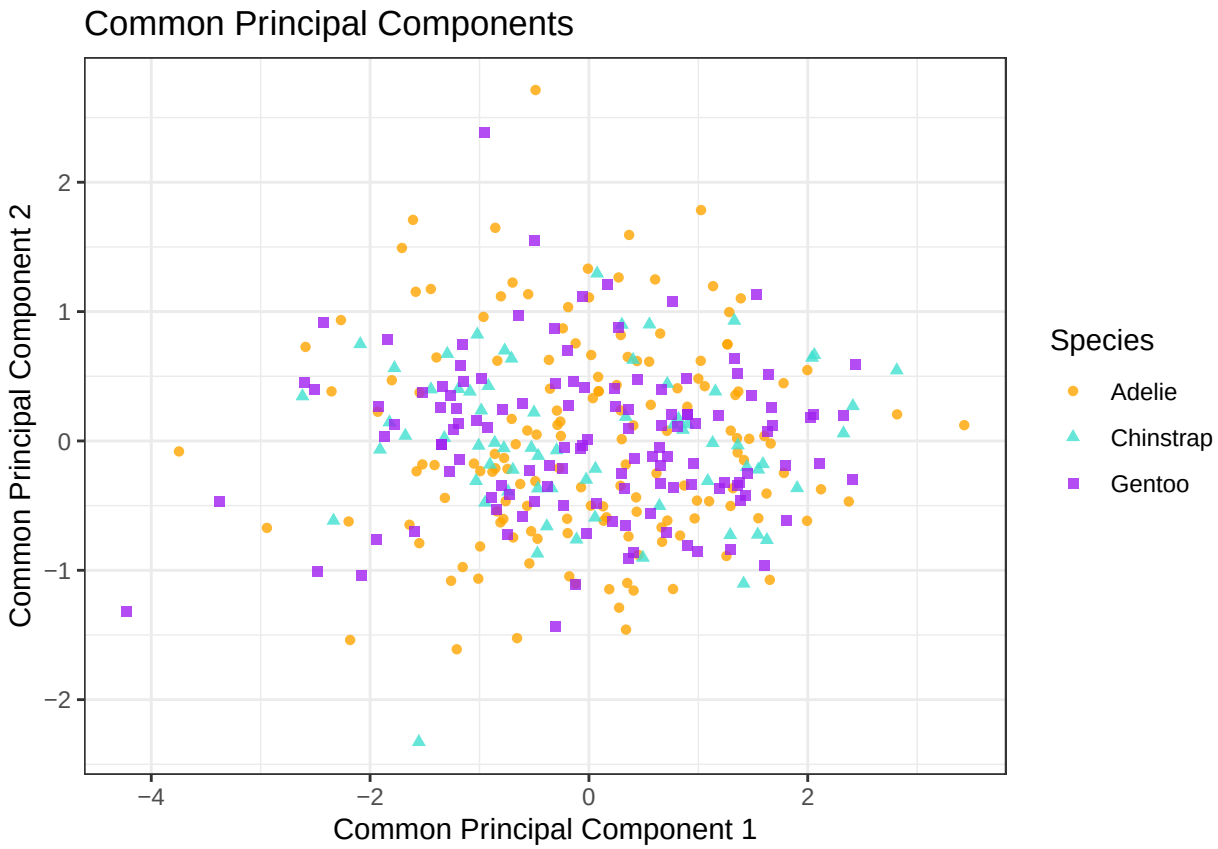
```r
Lambda[[2]] <- diag(crossprod(X.Chinstrap %*% V) / (nrow(X.Chinstrap) - 1))
Lambda[[3]] <- diag(crossprod(X.Gentoo %*% V) / (nrow(X.Gentoo) - 1))

#Get percentage explained
names(Lambda) <- c('Adelie', 'Chinstrap', 'Gentoo')
lapply(Lambda, function(x) x / sum(x))
```

```
## $Adelie
## [1] 0.6957458 0.3042542
##
## $Chinstrap
## [1] 0.8267681 0.1732319
##
## $Gentoo
## [1] 0.821692 0.178308
```

```r
#Get Common Principal Components
PC <- X.ext %*% V
pca_dat <- data.frame(PC1 = PC[,1], PC2 = PC[,2],
                      species = c(rep('Adelie', nrow(X.Adelie)),
                                  rep('Chinstrap', nrow(X.Chinstrap)),
                                  rep('Gentoo', nrow(X.Gentoo))))

#Plot CPCs
ggplot(pca_dat, aes(PC1, PC2, color = species))+
  geom_point(aes(shape = species), alpha = .8) +
  theme_bw()+ scale_color_manual(values=c('orange', 'turquoise', 'purple'))+
  labs(x = 'Common Principal Component 1', y = 'Common Principal Component 2', title = 'Common Principa
       color = 'Species', shape ='Species')
```




```
ggsave('common_principal_components.pdf', path = fig_path, width = 6, height = 5)
```