PCA Demo

Caitlin Mothes

2023-03-01

Set up by loading libraries and data sets

install.packages("plotly")

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4  
## ✔ tibble 3.1.8 ✔ dplyr 1.0.9  
## ✔ tidyr 1.2.0 ✔ stringr 1.4.0  
## ✔ readr 2.1.2 ✔ forcats 0.5.2  
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(lterdatasampler)  
library(plotly)

##   
## Attaching package: 'plotly'  
##   
## The following object is masked from 'package:ggplot2':  
##   
## last\_plot  
##   
## The following object is masked from 'package:stats':  
##   
## filter  
##   
## The following object is masked from 'package:graphics':  
##   
## layout

library(ggfortify)  
  
  
# load crab data  
data("pie\_crab")

inspect the data

pie\_crab

## # A tibble: 392 × 9  
## date latitude site size air\_temp air\_temp\_sd water\_temp water…¹ name   
## <date> <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>  
## 1 2016-07-24 30 GTM 12.4 21.8 6.39 24.5 6.12 Guan…  
## 2 2016-07-24 30 GTM 14.2 21.8 6.39 24.5 6.12 Guan…  
## 3 2016-07-24 30 GTM 14.5 21.8 6.39 24.5 6.12 Guan…  
## 4 2016-07-24 30 GTM 12.9 21.8 6.39 24.5 6.12 Guan…  
## 5 2016-07-24 30 GTM 12.4 21.8 6.39 24.5 6.12 Guan…  
## 6 2016-07-24 30 GTM 13.0 21.8 6.39 24.5 6.12 Guan…  
## 7 2016-07-24 30 GTM 10.3 21.8 6.39 24.5 6.12 Guan…  
## 8 2016-07-24 30 GTM 11.2 21.8 6.39 24.5 6.12 Guan…  
## 9 2016-07-24 30 GTM 12.7 21.8 6.39 24.5 6.12 Guan…  
## 10 2016-07-24 30 GTM 14.6 21.8 6.39 24.5 6.12 Guan…  
## # … with 382 more rows, and abbreviated variable name ¹​water\_temp\_sd  
## # ℹ Use `print(n = ...)` to see more rows

Look at variable correlations

pie\_crab %>%   
 select(latitude, air\_temp:water\_temp\_sd) %>%   
 cor()

## latitude air\_temp air\_temp\_sd water\_temp water\_temp\_sd  
## latitude 1.00000000 -0.99497146 0.7932130 -0.9571738 0.04188273  
## air\_temp -0.99497146 1.00000000 -0.7780397 0.9632287 -0.04532179  
## air\_temp\_sd 0.79321301 -0.77803972 1.0000000 -0.7879147 0.40970338  
## water\_temp -0.95717376 0.96322867 -0.7879147 1.0000000 -0.11480905  
## water\_temp\_sd 0.04188273 -0.04532179 0.4097034 -0.1148090 1.00000000

Conduct the PCA test

crab\_pca <- pie\_crab %>%   
 select(latitude, air\_temp:water\_temp\_sd) %>%  
 prcomp()

Look at variance contributions

summary(crab\_pca)

## Importance of components:  
## PC1 PC2 PC3 PC4 PC5  
## Standard deviation 6.6533 1.21857 0.78276 0.45905 0.25065  
## Proportion of Variance 0.9492 0.03184 0.01314 0.00452 0.00135  
## Cumulative Proportion 0.9492 0.98100 0.99413 0.99865 1.00000

Look at variable loadings

crab\_pca$rotation

## PC1 PC2 PC3 PC4 PC5  
## latitude 0.63779536 -0.1456839 0.4171761 -0.06536301 0.62744330  
## air\_temp -0.56802027 0.1189354 -0.2222182 -0.30713175 0.72076107  
## air\_temp\_sd 0.11790786 0.3206932 0.2023075 -0.87678588 -0.27124121  
## water\_temp -0.50648072 -0.2207416 0.8270560 0.06654101 -0.07937944  
## water\_temp\_sd 0.01204459 0.9016982 0.2272295 0.35807342 0.08333967

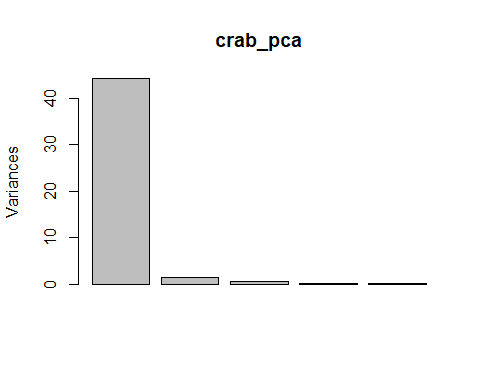
Look at site/observation scores

cor(crab\_pca$x)

## PC1 PC2 PC3 PC4 PC5  
## PC1 1.000000e+00 1.282922e-15 -5.203544e-15 -2.758252e-14 1.143021e-13  
## PC2 1.282922e-15 1.000000e+00 4.326717e-16 -9.482459e-16 7.313997e-16  
## PC3 -5.203544e-15 4.326717e-16 1.000000e+00 -1.202820e-15 5.418838e-15  
## PC4 -2.758252e-14 -9.482459e-16 -1.202820e-15 1.000000e+00 -8.592028e-15  
## PC5 1.143021e-13 7.313997e-16 5.418838e-15 -8.592028e-15 1.000000e+00

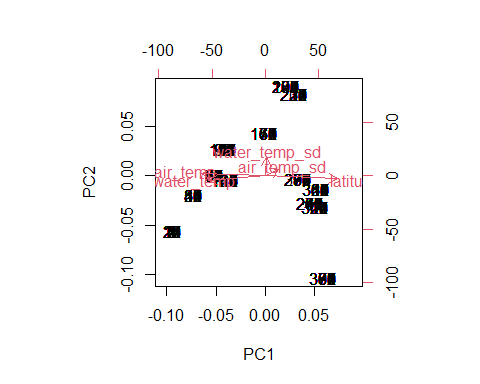
Make screeplot of variable contributions

screeplot(crab\_pca)



Make a biplot

biplot(crab\_pca)



Make a ggplot biplot (and make it interactive)

ggplotly(autoplot(crab\_pca, loadings = TRUE, loadings.label = TRUE) +  
 theme\_minimal())

Add PC axes to pie\_crab data

pie\_crab <- bind\_cols(pie\_crab, crab\_pca$x)

Conduct the multiple linear regression

model <- lm(size ~ PC1 + PC2 + PC3 + PC4 + PC5, data = pie\_crab)

summary(model)

##   
## Call:  
## lm(formula = size ~ PC1 + PC2 + PC3 + PC4 + PC5, data = pie\_crab)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.4215 -1.7999 -0.0104 1.7884 7.8223   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 14.65801 0.13517 108.443 < 2e-16 \*\*\*  
## PC1 0.30758 0.02034 15.120 < 2e-16 \*\*\*  
## PC2 -0.15108 0.11106 -1.360 0.17453   
## PC3 0.82621 0.17290 4.778 2.51e-06 \*\*\*  
## PC4 0.83790 0.29483 2.842 0.00472 \*\*   
## PC5 -2.56991 0.53996 -4.759 2.75e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.676 on 386 degrees of freedom  
## Multiple R-squared: 0.4239, Adjusted R-squared: 0.4165   
## F-statistic: 56.81 on 5 and 386 DF, p-value: < 2.2e-16