Dimension Reduction

Principal components analysis (PCA)

July 10th, 2023

What is the goal of dimension reduction?

We have p variables (columns) for n observations (rows) **BUT** which variables are **interesting**?

Can we find a smaller number of dimensions that captures the **interesting** structure in the data?

- Could examine all pairwise scatterplots of each variable - tedious, manual process

- Tuesday: clustered variables based on correlation

- Can we find a combination of the original p variables?

Dimension reduction:

- Focus on reducing the dimensionality of the feature space (i.e., number of columns),

- While **retaining** most of the information / **variability** in a lower dimensional space (i.e., reducing the number of columns)

Principal components analysis (PCA)

Step 1: Linear relationship that maximally summarizes variability of features
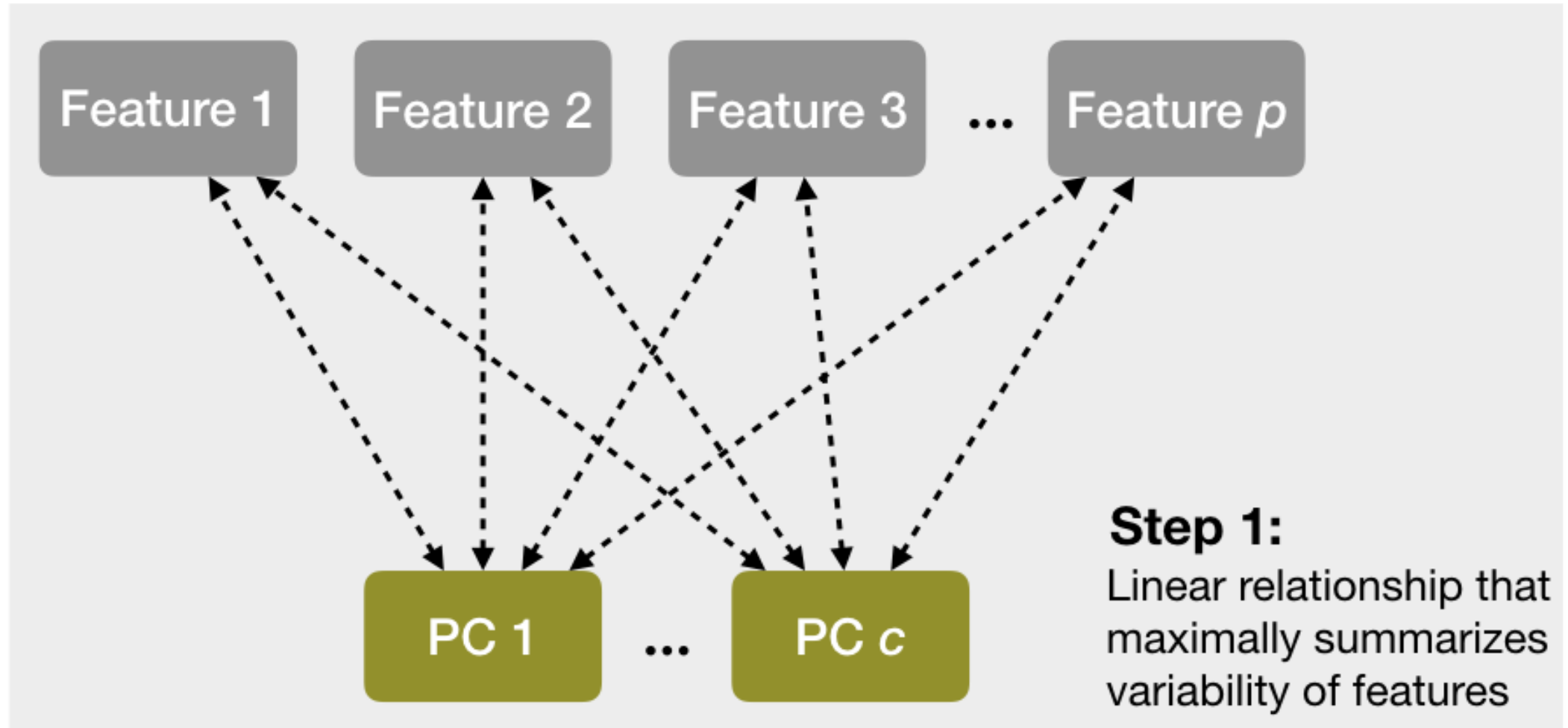
Principal components analysis (PCA)

- PCA explores the **covariance** between variables, and combines variables into a smaller set of **uncorrelated** variables called **principal components (PCs)**

- PCs are **weighted**, linear combinations of the original variables

  - Weights reveal how different variables are *loaded* into the PCs

- We want a **small number of PCs** to explain most of the information / variance in the data

First principal component:

$$Z_1 = \phi_{11} X_1 + \phi_{21} X_2 + \cdots + \phi_{p1} X_p$$

- ϕ_{j1} are the weights indicating the contributions of each variable $j \in 1, \ldots, p$

- Weights are normalized $\sum_{j=1}^{p} \phi_{j1}^2 = 1$

- $\phi_1 = (\phi_{11}, \phi_{21}, \ldots, \phi_{p1})$ is the **loading vector** for PC1

- Z_1 is a linear combination of the p variables that has the **largest variance**

Principal components analysis (PCA)

Second principal component:

$$Z_2 = \phi_{12}X_1 + \phi_{22}X_2 + \cdots + \phi_{p2}X_p$$
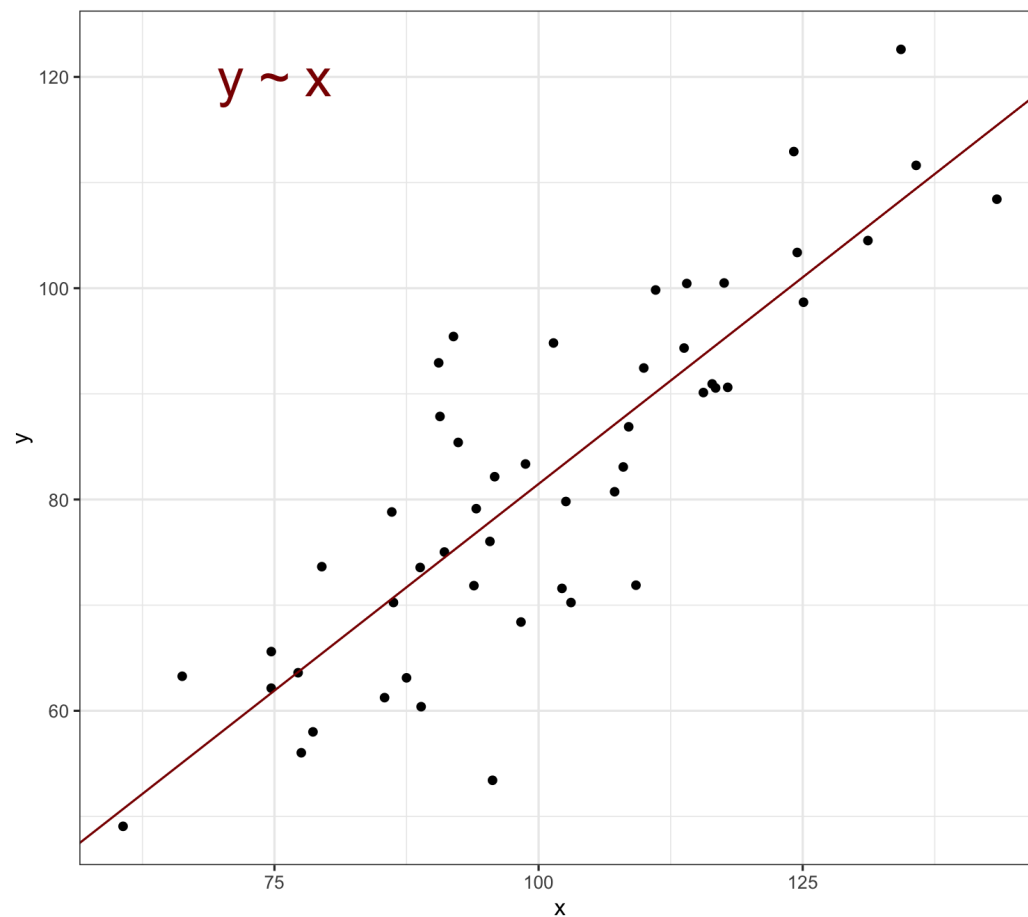
- ϕ_{j2} are the weights indicating the contributions of each variable $j \in 1, \ldots, p$

- Weights are normalized $\sum_{j=1}^{p} \phi_{j1}^2 = 1$

- $\phi_2 = (\phi_{12}, \phi_{22}, \ldots, \phi_{p2})$ is the **loading vector** for PC2

- Z_2 is a linear combination of the p variables that has the **largest variance**

  - **Subject to constraint it is uncorrelated with** Z_1

We repeat this process to create p principal components
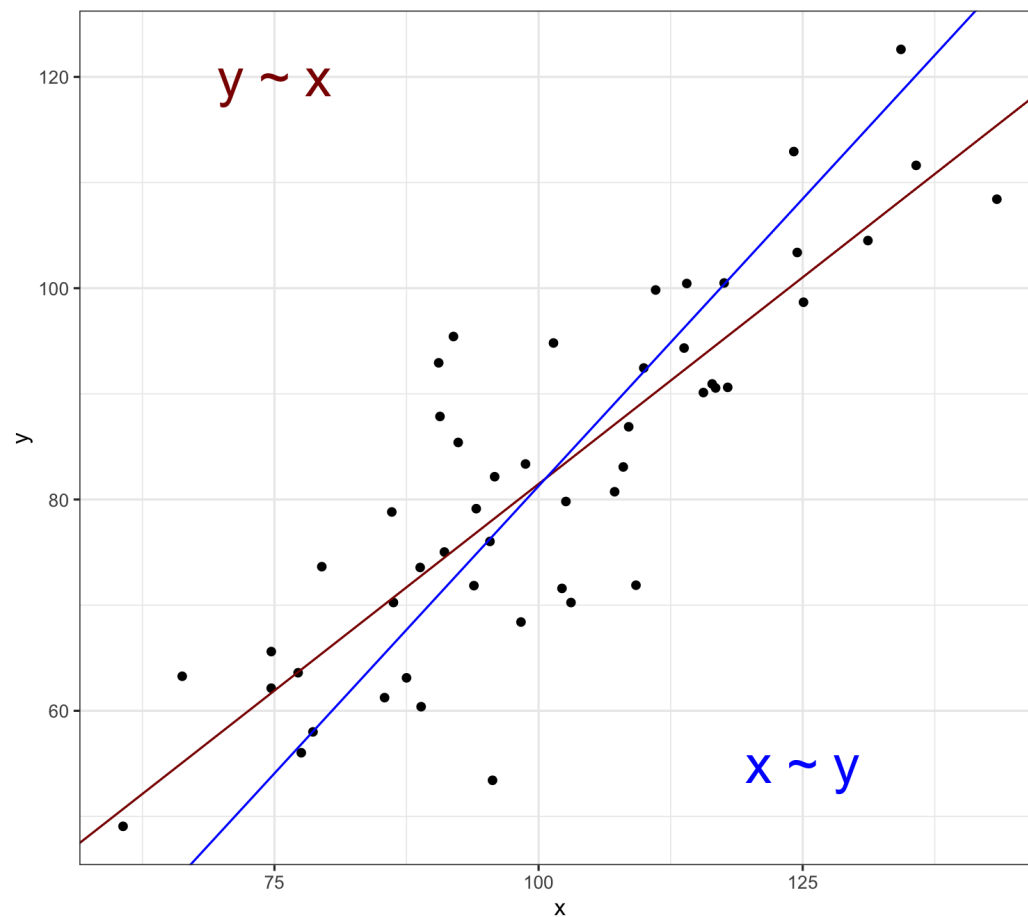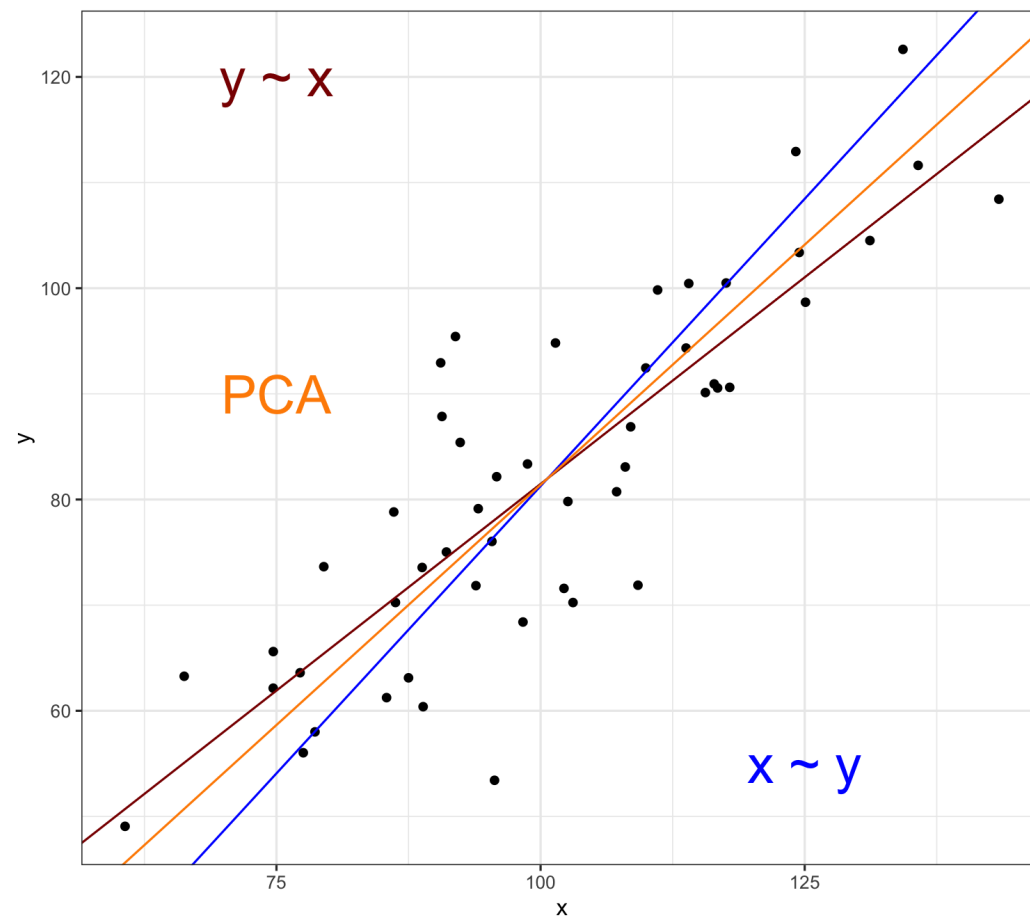
Visualizing PCA in two dimensions
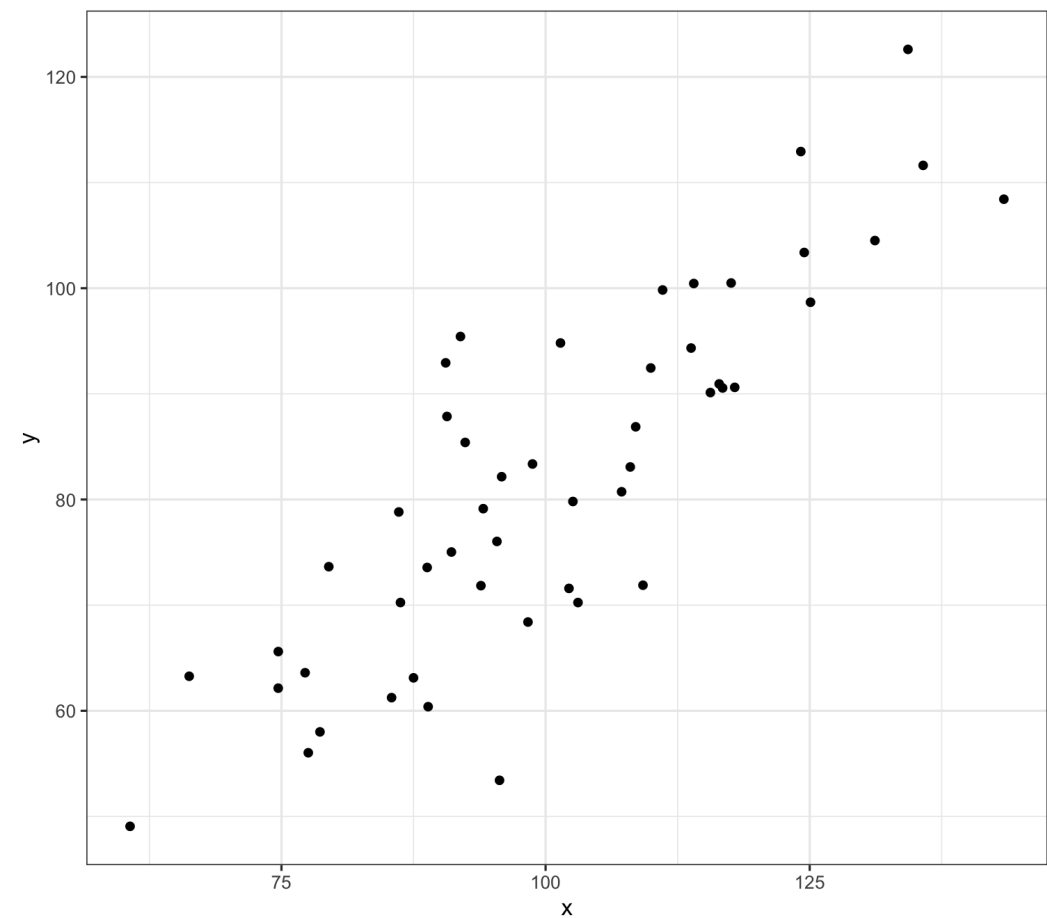
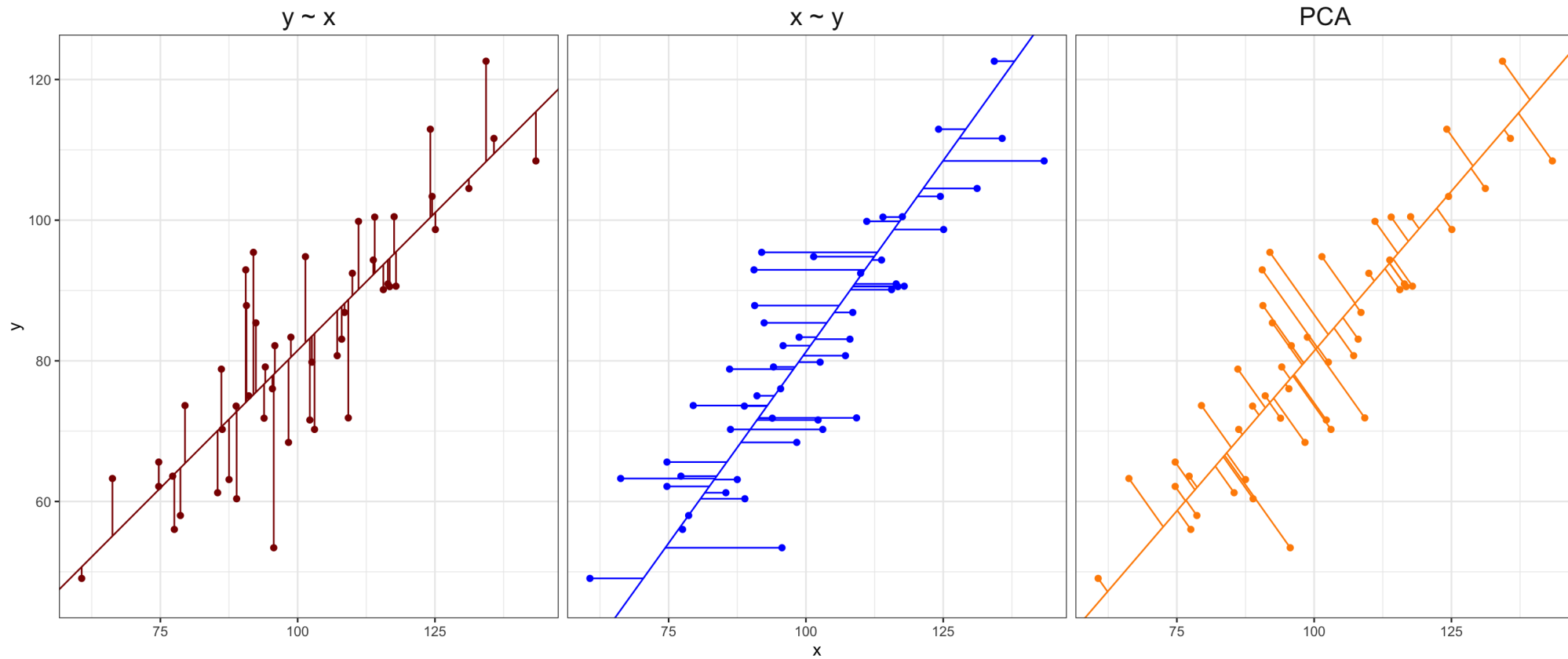# Visualizing PCA in two dimensions

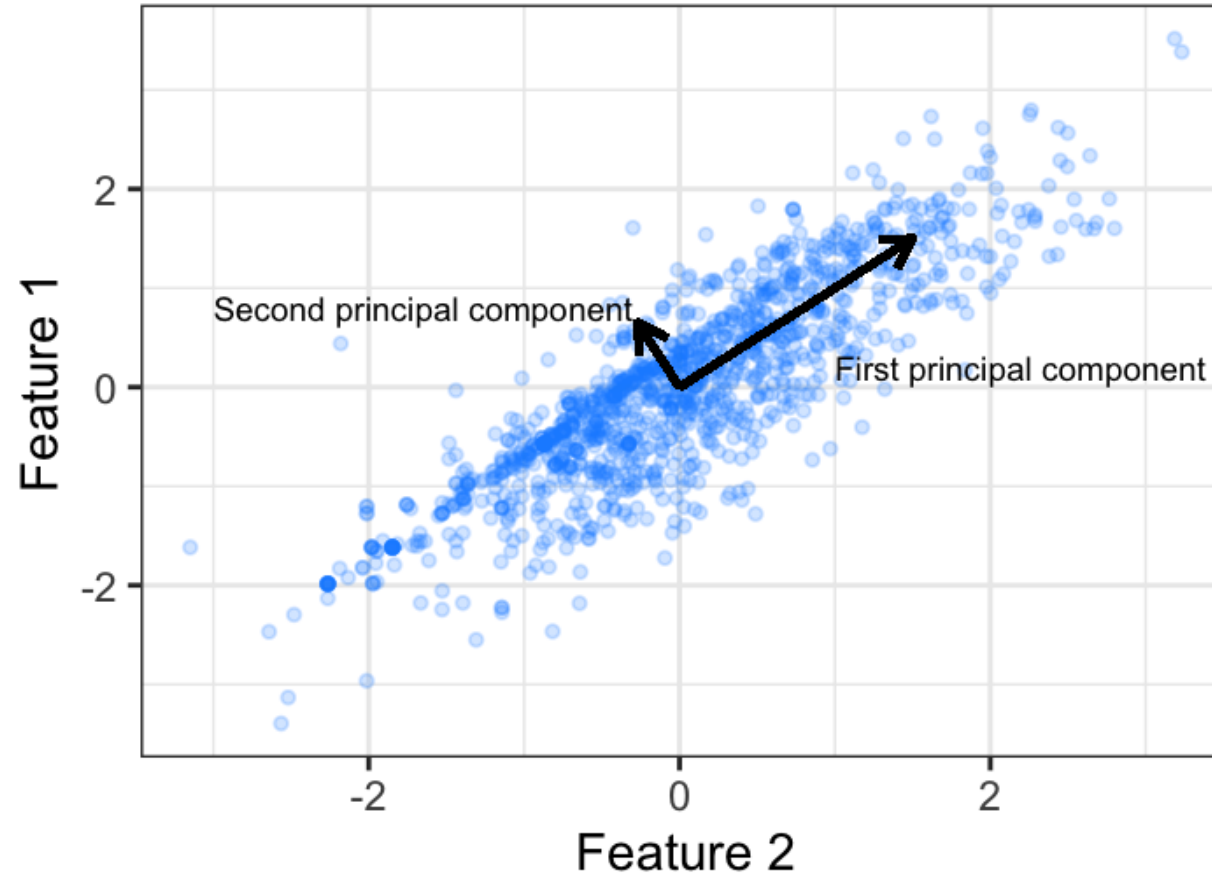# Visualizing PCA in two dimensions

Visualizing PCA in two dimensions

Visualizing PCA in two dimensions

Searching for variance in orthogonal directions

PCA: singular value decomposition (SVD)

$$X = UDV^T$$

- Matrices U and V contain the left and right (respectively) **singular vectors of scaled matrix** X

- D is the diagonal matrix of the **singular values**

- SVD simplifies matrix-vector multiplication as **rotate, scale, and rotate again**

V is called the **loading matrix** for X with ϕ_j as columns,

- $Z = XV$ is the PC matrix

BONUS **eigenvalue decomposition** (aka spectral decomposition)

- V are the **eigenvectors** of $X^T X$ (covariance matrix, ^{T} means *transpose*)

- U are the **eigenvectors** of XX^T

- The singular values (diagonal of D) are square roots of the **eigenvalues** of $X^T X$ or XX^T

- Meaning that $Z = UD$

Eigenvalues solve time travel?

Probably not... but they guide dimension reduction

We want to choose $p^* < p$ such that we are explaining variation in the data

Eigenvalues λ_j for $j \in 1, \ldots, p$ indicate **the variance explained by each component**

- $\sum_j^p \lambda_j = p$, meaning $\lambda_j \geq 1$ indicates $\mathrm{PC}j$ contains at least one variable's worth in variability

- λ_j/p equals proportion of variance explained by $\mathrm{PC}j$

- Arranged in descending order so that λ_1 is largest eigenvalue and corresponds to PC1

- Can compute the cumulative proportion of variance explained (CVE) with p^* components:

$$\mathrm{CVE}_{p^*} = \frac{\sum_j^{p*} \lambda_j}{p}$$

Can use **scree plot** to plot eigenvalues and guide choice for $p^* < p$ by looking for "elbow" (rapid to slow change)

Example data: NFL teams summary

Created dataset using `nflfastR` summarizing NFL team performances from 1999 to 2021

```r
library(tidyverse)
nfl_teams_data <- read_csv("https://shorturl.at/cfmpW")
nfl_model_data <- nfl_teams_data %>%
  mutate(score_diff = points_scored - points_allowed) %>%
  # Only use rows with air yards
  filter(season >= 2006) %>%
  dplyr::select(-wins, -losses, -ties, -points_scored, -points_allowed, -season, -team)
```

NFL PCA example

Use the `prcomp` function (uses SVD) for PCA on **centered** and **scaled** data

```
model_x <- as.matrix(dplyr::select(nfl_model_data, -score_diff))
pca_nfl <- prcomp(model_x, center = TRUE, scale = TRUE)
summary(pca_nfl)
```

```
## Importance of components:
##                             PC1     PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation       3.2060  3.1026  2.3257 2.04728 1.52301 1.40350 1.35714
## Proportion of Variance   0.2141  0.2006  0.1127 0.08732 0.04832 0.04104 0.03837
## Cumulative Proportion    0.2141  0.4147  0.5274 0.61468 0.66301 0.70405 0.74242
##                             PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation       1.26250 1.14773 1.09881 1.01200 0.95689 0.93513 0.85233
## Proportion of Variance   0.03321 0.02744 0.02515 0.02134 0.01908 0.01822 0.01513
## Cumulative Proportion    0.77562 0.80307 0.82822 0.84956 0.86863 0.88685 0.90199
##                            PC15    PC16    PC17    PC18    PC19    PC20    PC21
## Standard deviation       0.82315 0.77434 0.65692 0.64016 0.60076 0.5796 0.56756
## Proportion of Variance   0.01412 0.01249 0.00899 0.00854 0.00752 0.0070 0.00671
## Cumulative Proportion    0.91610 0.92859 0.93758 0.94612 0.95364 0.9606 0.96735
##                            PC22    PC23    PC24    PC25    PC26    PC27    PC28
## Standard deviation       0.51349 0.47233 0.46768 0.41284 0.35810 0.33597 0.32018
## Proportion of Variance   0.00549 0.00465 0.00456 0.00355 0.00267 0.00235 0.00214
## Cumulative Proportion    0.97284 0.97749 0.98205 0.98560 0.98827 0.99062 0.99276
```

Proportion of variance explained

`prcomp$sdev` corresponds to the singular values, i.e., $\sqrt{\lambda_j}$, what is `pca_nfl$sdev^2 / ncol(model_x)`?

Can use the `broom` package easily tidy `prcomp` summary for plotting

```r
library(broom)
pca_nfl %>%
  tidy(matrix = "eigenvalues") %>%
  ggplot(aes(x = PC, y = percent)) +
  geom_line() + geom_point() +
  geom_hline(yintercept = 1 / ncol(model_x),
             color = "darkred",
             linetype = "dashed") +
  theme_bw()
```

- Add reference line at $1/p$, *why*?
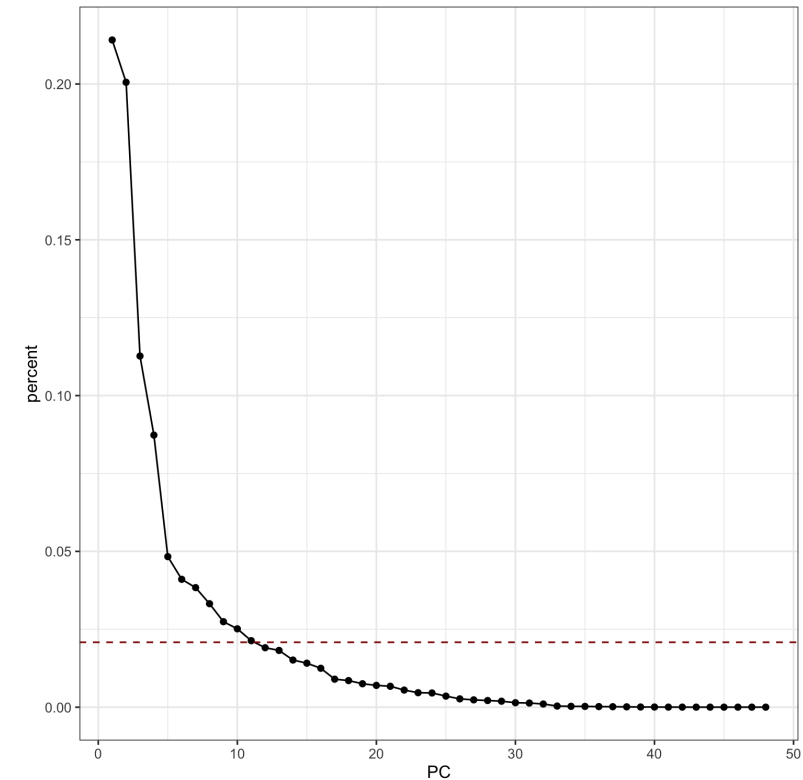
Display data in lower dimensions

`prcomp$x` corresponds to the matrix of **principal component scores**, i.e., $Z = XV$

Can `augment` dataset with PC scores for plotting

- Add team and season for context

```
pca_nfl %>%
  augment(nfl_model_data) %>%
  bind_cols({
    nfl_teams_data %>%
      filter(season >= 2006) %>%
      dplyr::select(season, team)
  }) %>%
  unite("team_id", team:season, sep = "-",
        remove = FALSE) %>%
  ggplot(aes(x = .fittedPC1, y = .fittedPC2,
             color = season)) +
  geom_text(aes(label = team_id), alpha = 0.9
  scale_color_gradient(low = "purple", high =
  theme_bw() + theme(legend.position = "botto
```

What are the loadings of these dimensions?

`prcomp$rotation` corresponds to the **loading matrix**, i.e., V

```r
arrow_style <- arrow(
  angle = 20, ends = "first", type = "closed"
  length = grid::unit(8, "pt")
)
library(ggrepel)
pca_nfl %>%
  tidy(matrix = "rotation") %>%
  pivot_wider(names_from = "PC", names_prefix
              values_from = "value") %>%
  mutate(stat_type = ifelse(str_detect(column
                          "offense", "defen
  ggplot(aes(PC1, PC2)) +
  geom_segment(xend = 0, yend = 0, arrow = ar
  geom_text_repel(aes(label = column, color =
                  size = 3) +
  scale_color_manual(values = c("darkred", "d
  theme_bw() +
  theme(legend.position = "bottom")
```
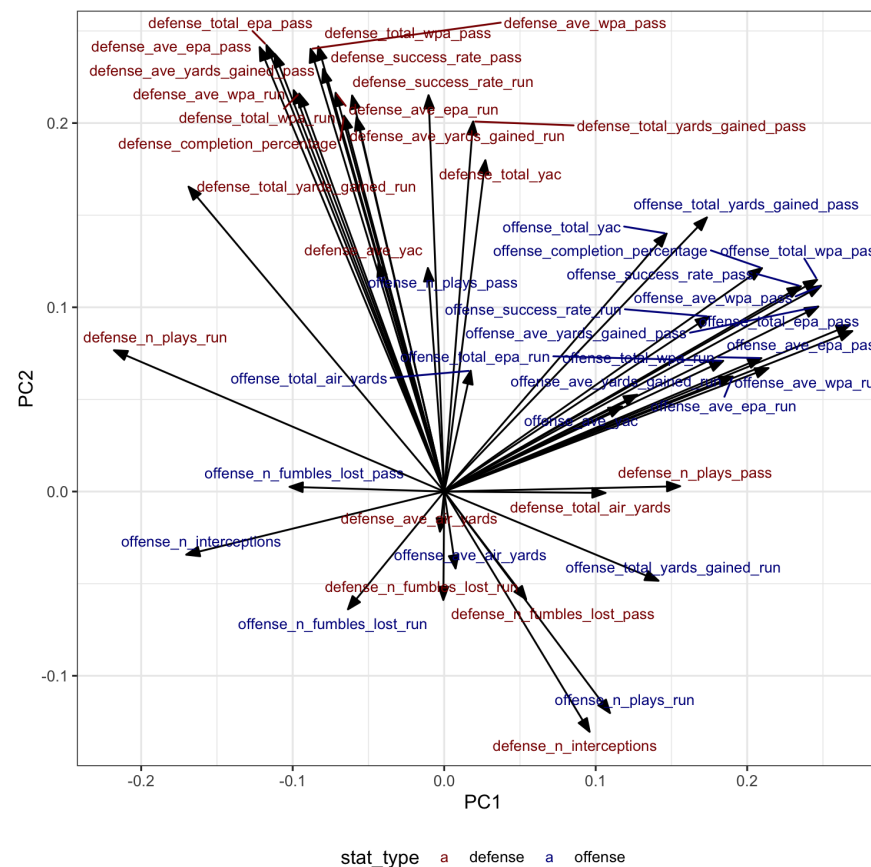
PCA analysis with `factoextra`

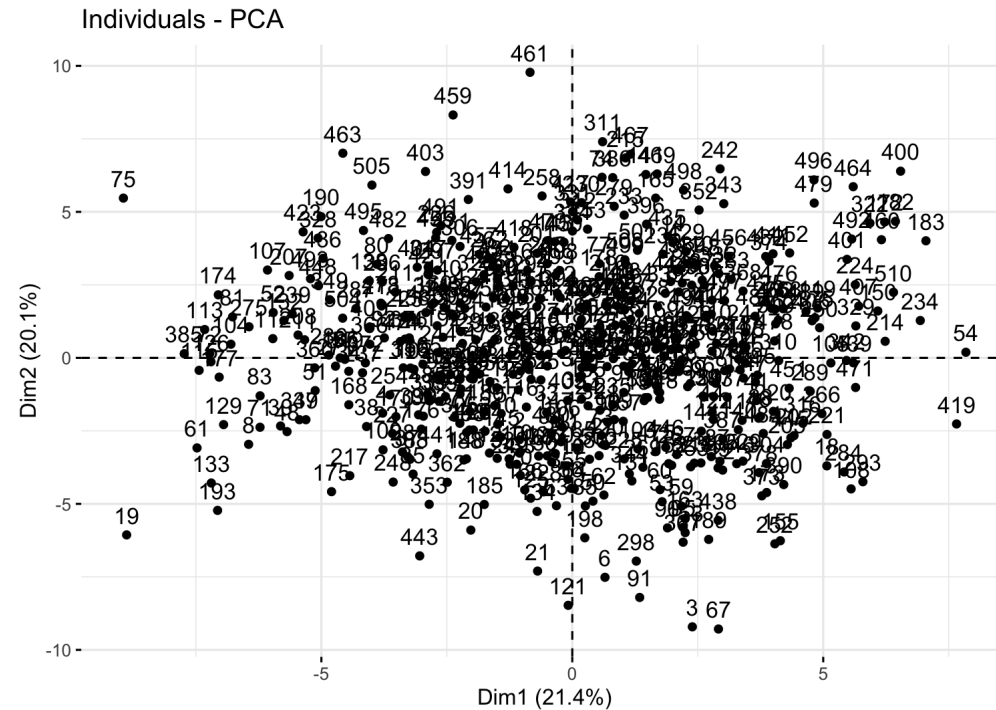Visualize the proportion of variance explained by each PC with `factoextra`

```
library(factoextra)
fviz_eig(pca_nfl)
```
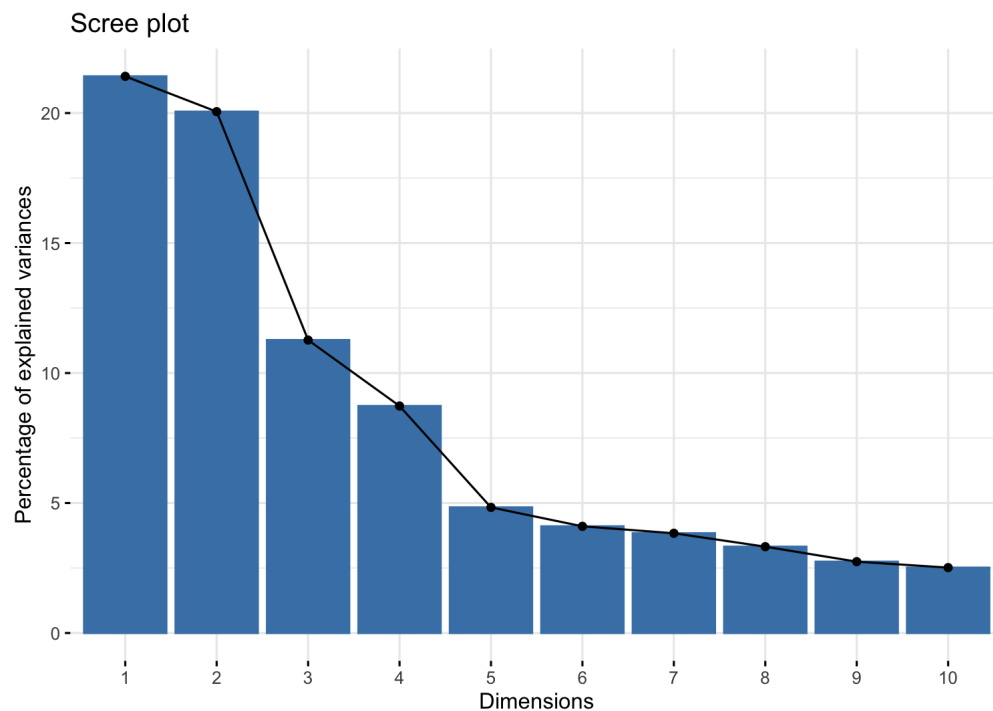
PCA analysis with `factoextra`

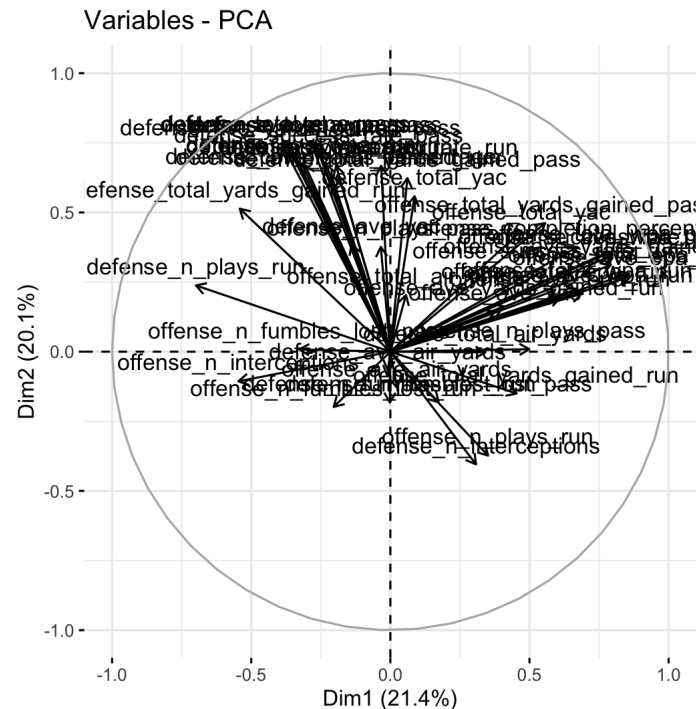Display observations with first two PC

```
fviz_pca_ind(pca_nfl)
```

PCA analysis with `factoextra`

Projection of variables - angles are interpreted as correlations, where negative correlated values point to opposite sides of graph

```
fviz_pca_var(pca_nfl)
```

PCA analysis with `factoextra`

Biplot displays both the space of observations and the space of variables

- Arrows represent the directions of the original variables

```
fviz_pca_biplot(pca_nfl)
```