Dimension Reduction (PCA)

07-10-2023

What is the goal of dimension reduction?

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

- **REFINED QUESTION:** can we "project" data to a lower dimension but keep maximal information?
- **SHARPER QUESTION:** is there another "basis", which better expresses the information in our original data set?

Linear Algebra Interlude

- 2D vectors - magnitude ('norm') and direction ('dot product')
- Projection: length of the shadow of the given vector over another vector

$$Proj_w(v) = (v^T w^*)w^*$$

where:

$$w^* = \frac{w}{||w||}$$

and

$v^T w^*$ = degree of information preserved about v after projecting onto w

MATRICES can be thought of as:

- data
- functions (linear transformations)

EIGENVALUES AND EIGENVALUES:

$$Au = \lambda u$$

λ = eigenvalue

u = eigenvector

Importance:

- Eigenvectors basically stay invariant to rotation after beign acted on by A – "holding ground after being acted on by A"

PCA

GOAL: can we find p new directions that preserves:

- linearity
- maximizes variance explained
- are orthogonal

Let u be the vector that preserves the most information from the data:

$$max \sum_{i=1}^{p} (x_i^T u)^2$$

s.t. $u^T u = 1$, or equivalently: $u^T u - 1 = 0$

Then to find the other principal components:

$$max \sum_{i=1}^{p} (x_i^T u)^2$$

s.t. $u_2^T u_2 = 1$ AND $u_1 \perp u2$

- PCA explores the covariance between variables and combines variables into a smaller set of uncorrelated variables called principal components (PCs)
- The first principal component is the linear combination of the p variables that has the **largest variance**. The amount of variability captured goes in descending order.

Singular Value Decomposition (SVD)

X is the convariance matrix

$$X = UDV^T$$

* Matrices U and V contain the left and right sinular 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

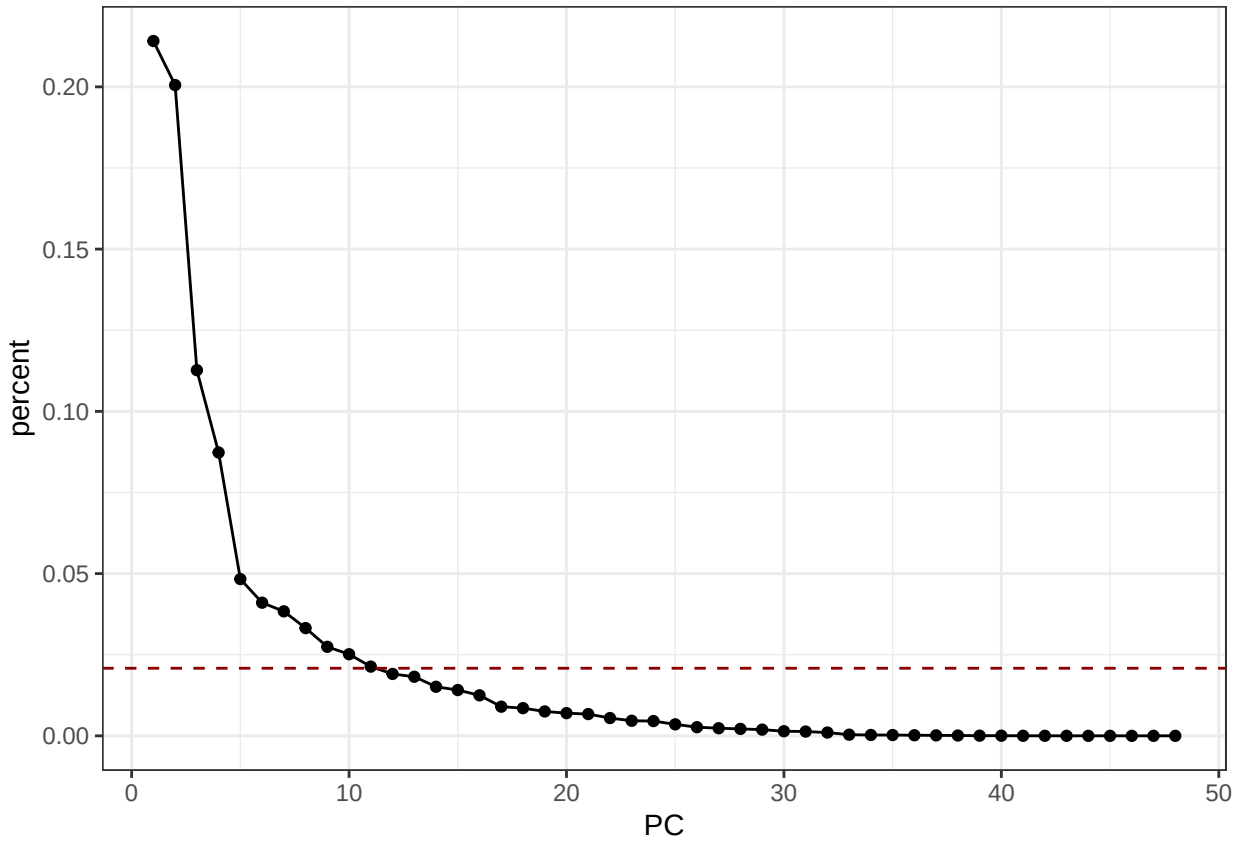
Z = XV is the PC matrix

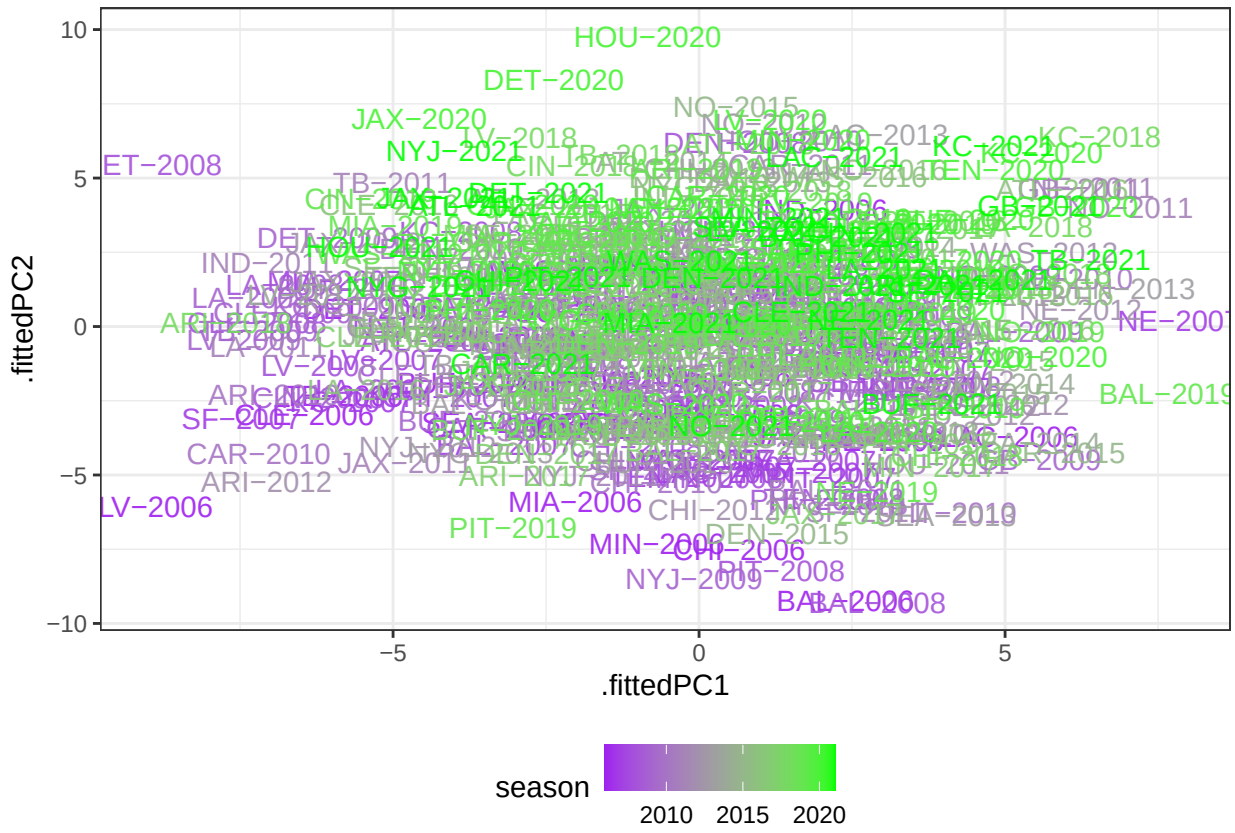
Eigenvalue Decomposition

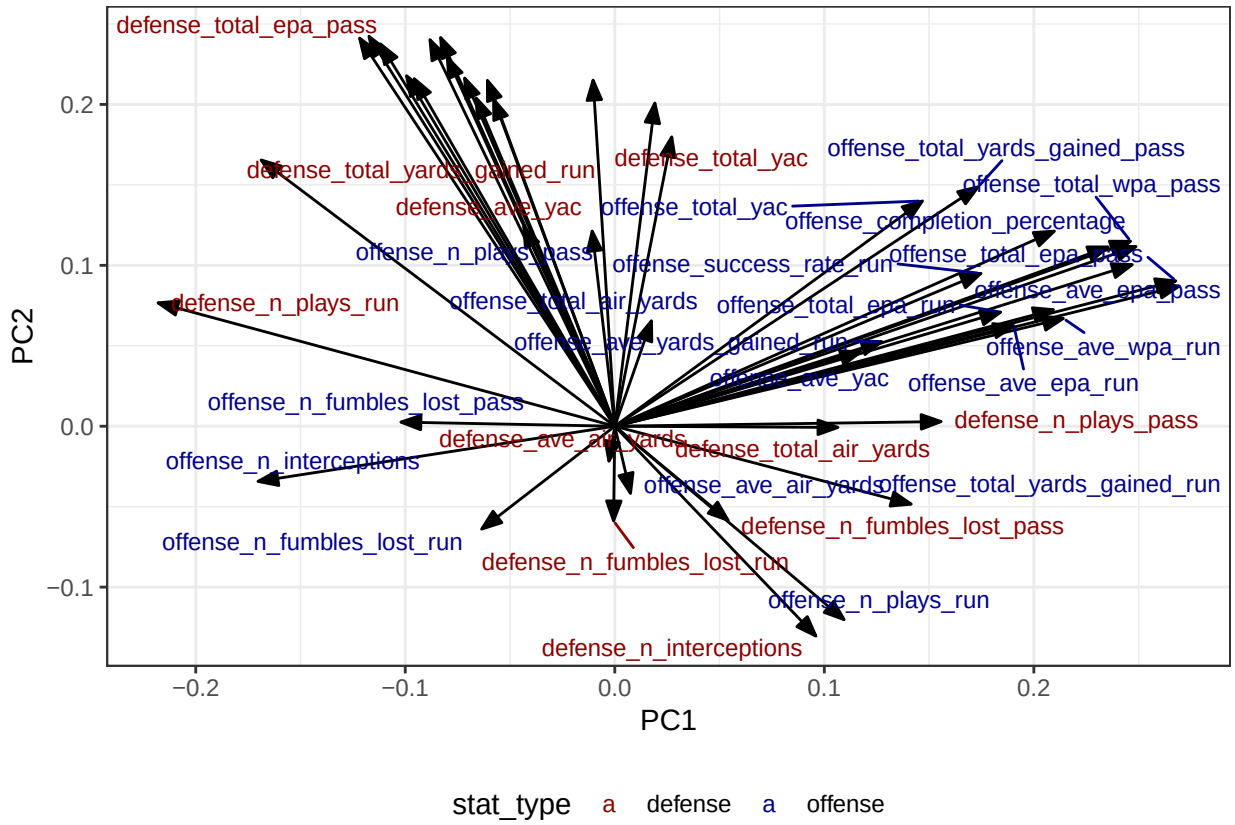
- V are **eigenvectors** of $X^T X$
- U are the **eigenvectors** of XX^T
- The singular values (diagonal of D) are square roots of teh **eigenvalues** of$X^T X$ or XX^T
- Meaning that Z = UD

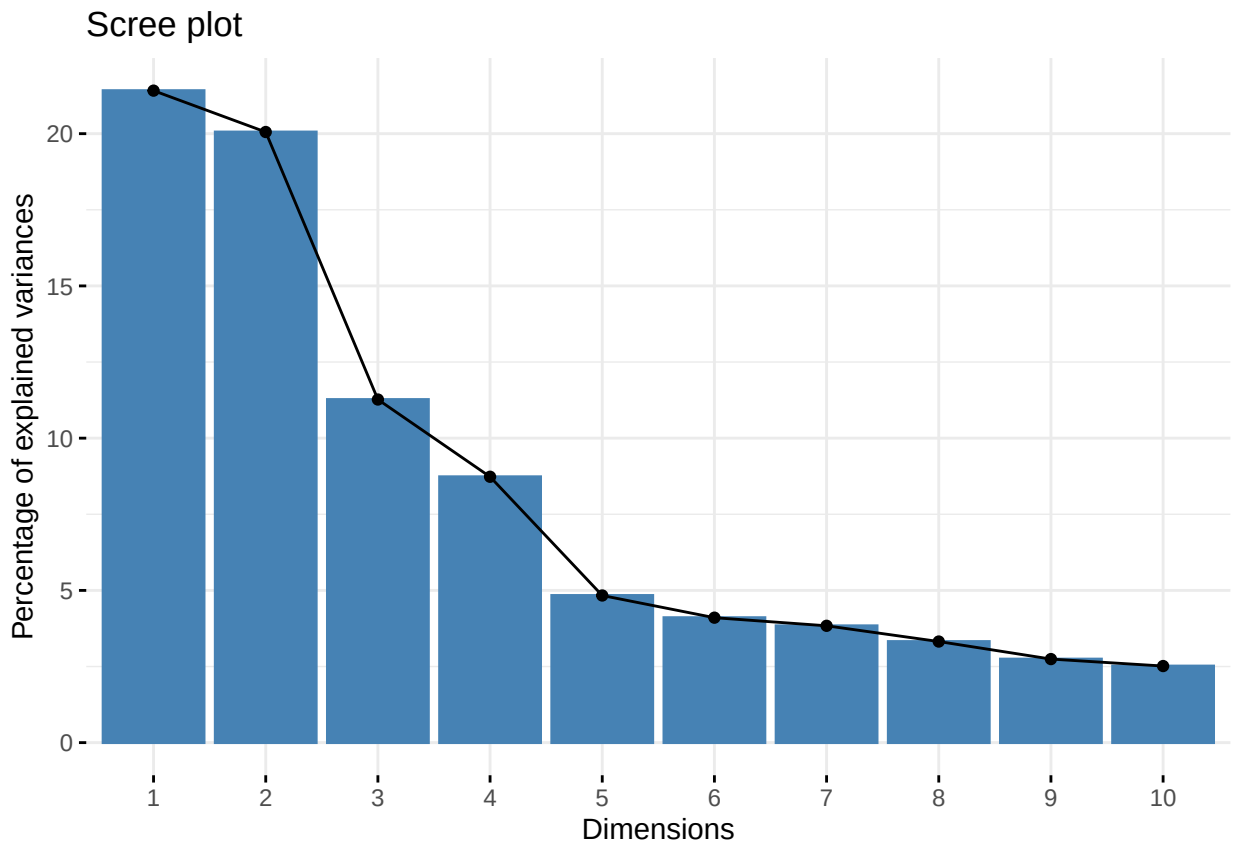
Example

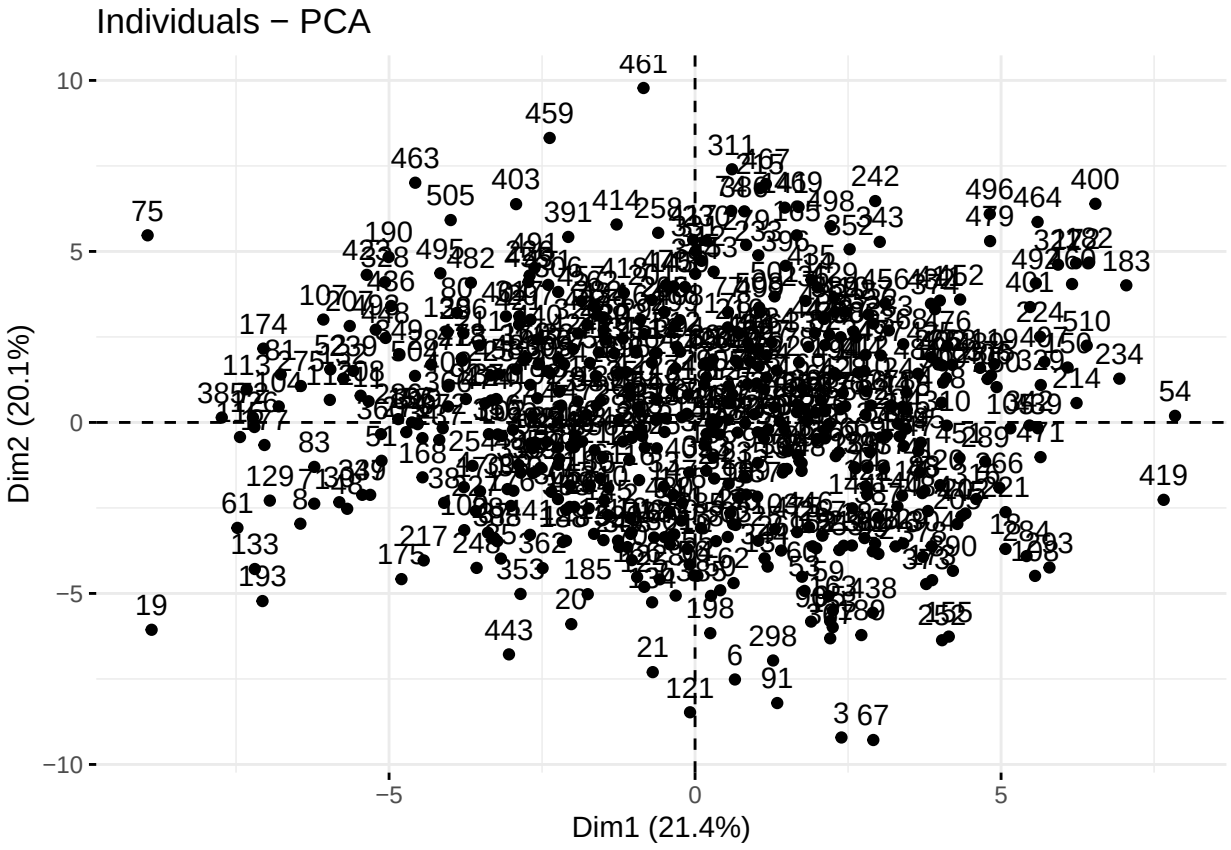
```
## Importance of components:
##                            PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     3.2060 3.1026 2.3257 2.04728 1.52301 1.40350 1.35714
## Proportion of Variance 0.2141 0.2006 0.1127 0.08732 0.04832 0.04104 0.03837
## Cumulative Proportion  0.2141 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
##                           PC29    PC30    PC31    PC32    PC33    PC34    PC35
## Standard deviation     0.30357 0.26161 0.25289 0.22149 0.13146 0.11459 0.10964
## Proportion of Variance 0.00192 0.00143 0.00133 0.00102 0.00036 0.00027 0.00025
## Cumulative Proportion  0.99468 0.99610 0.99744 0.99846 0.99882 0.99909 0.99934
##                           PC36    PC37    PC38    PC39    PC40    PC41    PC42
## Standard deviation     0.09672 0.08397 0.07385 0.05223 0.04814 0.03391 0.02901
## Proportion of Variance 0.00019 0.00015 0.00011 0.00006 0.00005 0.00002 0.00002
## Cumulative Proportion  0.99954 0.99968 0.99980 0.99985 0.99990 0.99993 0.99994
##                           PC43    PC44    PC45    PC46    PC47    PC48
## Standard deviation     0.02562 0.02290 0.02213 0.02139 0.01718 0.01670
## Proportion of Variance 0.00001 0.00001 0.00001 0.00001 0.00001 0.00001
## Cumulative Proportion  0.99996 0.99997 0.99998 0.99999 0.99999 1.00000
```

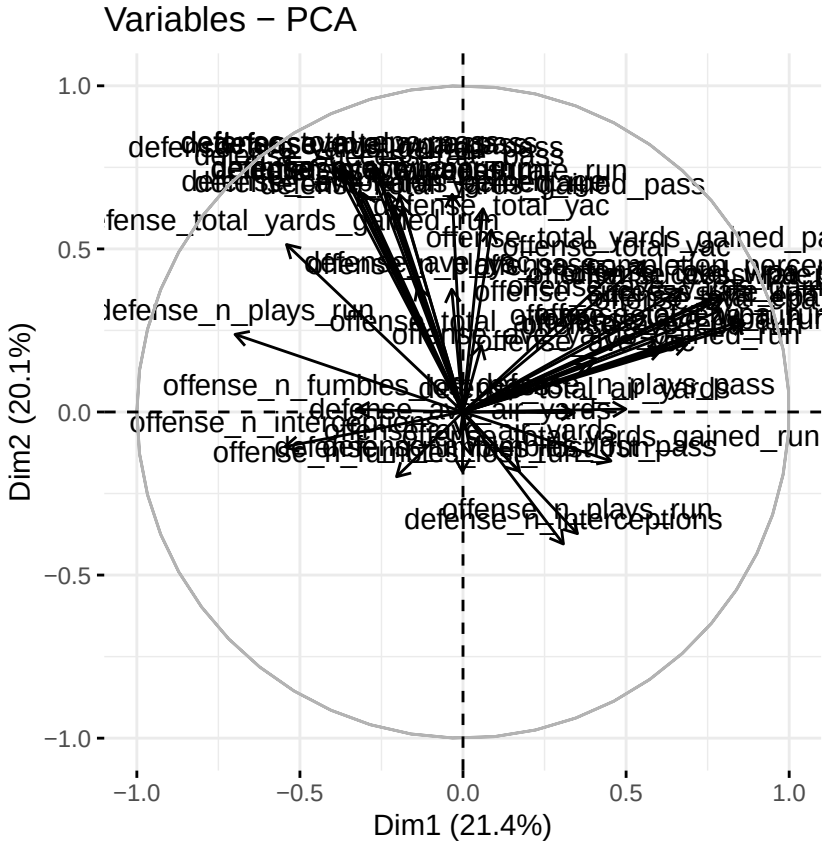

percent
0.20
0.15
0.10
0.05
0.00
0
10
20
30
40
50
PC


.fittedPC2
10
5
0
−5
−10
−5
0
5
.fittedPC1
HOU−2020
DET−2020
JAX−2020
NYJ−2021
ET−2008
TB−2011
IND−2011
SF−2006
CAR−2010
ARI−2012
LV−2006
MIA−2006
PIT−2019
MIN−2006
NYJ−2009
KC−2021
KC−2018
TB−2021
NE−200
BAL−2019
season
2010
2015
2020


defense_total_epa_pass
defense_total_yards_gained_run
defense_total_yac
offense_total_yards_gained_pass
offense_total_wpa_pass
defense_ave_yac
offense_total_yac
offense_completion_percentage
offense_n_plays_pass
offense_success_rate_run
offense_total_epa_pass
defense_n_plays_run
offense_total_air_yards
offense_total_epa_run
offense_ave_epa_pass
offense_ave_yards_gained_run
offense_ave_wpa_run
offense_ave_yac
offense_ave_epa_run
offense_n_fumbles_lost_pass
defense_n_plays_pass
defense_ave_air_yards
defense_total_air_yards
offense_n_interceptions
offense_ave_air_yards
offense_total_yards_gained_run
defense_n_fumbles_lost_pass
offense_n_fumbles_lost_run
defense_n_fumbles_lost_run
offense_n_plays_run
defense_n_interceptions
PC2
PC1
0.2
0.1
0.0
−0.1
−0.2
−0.1
0.0
0.1
0.2
stat_type
a defense
a offense


Scree plot
Percentage of explained variances
Dimensions
20
15
10
5
0
1
2
3
4
5
6
7
8
9
10


Individuals − PCA
Dim2 (20.1%)
Dim1 (21.4%)
10
5
0
−5
−10
−5
0
5


Variables − PCA
Dim2 (20.1%)
Dim1 (21.4%)
1.0
0.5
0.0
−0.5
−1.0
offense_n_plays_run
defense_n_interceptions

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