Principal Components Analysis (part II)

Predictive Modeling & Statistical Learning

Gaston Sanchez

PCA of NBA Team Stats

NBA Team Stats

- NBA Team Stats: regular season (2016-17)
- Github file: `data/nba-teams-2017.csv`
- Source: **stats.nba.com**
- `http://stats.nba.com/teams/traditional/#!?sort=GP&dir=-1`

SEASON: 2016-17 | SEASON TYPE: Regular Season | PER MODE: Per Game | SEASON SEGMENT: All Games | Advanced Filters

RECENT FILTERS | GLOSSARY | SHARE

	TEAM	GP	W	L	WIN%	MIN	PTS	FGM	FGA	FG%	3PM	3PA	3P%	FTM	FTA	FT%	OREB	DREB	REB	AST	TOV	STL	BLK	BLKA	PF	PFD	+/-
1	Miami Heat	82	41	41	.500	48.2	103.2	39.0	85.8	45.5	9.9	27.0	36.5	15.2	21.6	70.6	10.6	33.0	43.6	21.2	13.4	7.2	5.7	4.9	20.5	18.7	1.1
1	Atlanta Hawks	82	43	39	.524	48.5	103.2	38.1	84.4	45.1	8.9	26.1	34.1	18.1	24.9	72.8	10.3	34.1	44.3	23.6	15.8	8.2	4.8	5.2	18.2	21.6	-0.9
1	Brooklyn Nets	82	20	62	.244	48.2	105.8	37.8	85.2	44.4	10.7	31.6	33.8	19.4	24.6	78.8	8.8	35.1	43.9	21.4	16.5	7.2	4.7	5.6	21.0	20.4	-6.7
1	Charlotte Hornets	82	36	46	.439	48.4	104.9	37.7	85.4	44.2	10.0	28.6	35.1	19.4	23.8	81.5	8.8	34.8	43.6	23.1	11.5	7.0	4.8	5.5	16.6	19.9	0.2
1	Chicago Bulls	82	41	41	.500	48.2	102.9	38.6	87.1	44.4	7.6	22.3	34.0	18.0	22.5	79.8	12.2	34.1	46.3	22.6	13.6	7.8	4.8	4.6	17.7	18.8	0.4
1	Cleveland Cavaliers	82	51	31	.622	48.5	110.3	39.9	84.9	47.0	13.0	33.9	38.4	17.5	23.3	74.8	9.3	34.4	43.7	22.7	13.7	6.6	4.0	4.3	18.1	20.6	3.2
1	Dallas Mavericks	82	33	49	.402	48.2	97.9	36.2	82.3	44.0	10.7	30.2	35.5	14.8	18.5	80.1	7.9	30.7	38.6	20.8	11.9	7.5	3.7	3.4	19.1	19.4	-2.9
1	Denver Nuggets	82	40	42	.488	48.2	111.7	41.2	87.7	46.9	10.6	28.8	36.8	18.7	24.2	77.4	11.8	34.6	46.4	25.3	15.0	6.9	3.9	4.9	19.1	20.2	0.5
1	Detroit Pistons	82	37	45	.451	48.3	101.3	39.9	88.8	44.9	7.7	23.4	33.0	13.9	19.3	71.9	11.1	34.6	45.7	21.1	11.9	7.0	3.8	4.1	17.9	17.5	-1.1
1	Golden State Warriors	82	67	15	.817	48.2	115.9	43.1	87.1	49.5	12.0	31.2	38.3	17.8	22.6	78.8	9.4	35.0	44.4	30.4	14.8	9.6	6.8	3.8	19.3	19.4	11.6

```r
# variables
dat <- read.csv('data/nba-teams-2017.csv')

names(dat)
```

```
 [1] "team"                "games_played"        "wins"
 [4] "losses"              "win_prop"            "minutes"
 [7] "points"              "field_goals"         "field_goals_attempted"
[10] "field_goals_prop"    "points3"             "points3_attempted"
[13] "points3_prop"        "free_throws"         "free_throws_att"
[16] "free_throws_prop"    "off_rebounds"        "def_rebounds"
[19] "rebounds"            "assists"             "turnovers"
[22] "steals"              "blocks"              "block_fga"
[25] "personal_fouls"      "personal_fouls_drawn" "plus_minus"
```

Active and Supplementary Elements

Which Variables?

Active Variables

We are going to focus the analysis on the following **active** variables:

- `wins`
- `losses`
- `points`
- `field_goals`
- `assists`
- `turnovers`
- `steals`
- `blocks`

"Active" means these are the variables used to compute PCs.

Which Variables?

Supplementary Variables

Among the rest of the variables, we are going to consider three **supplementary** variables:

- `points3`
- `rebounds`
- `personal_fouls`

"Supplementary" means these variables are NOT used to compute PCs, but we will take them into account during the interpretation phase.

Which Individuals?

Active Individuals

All of the teams in season 2016-2017

Supplementary Individuals

Warriors and Cavaliers 2015-2016

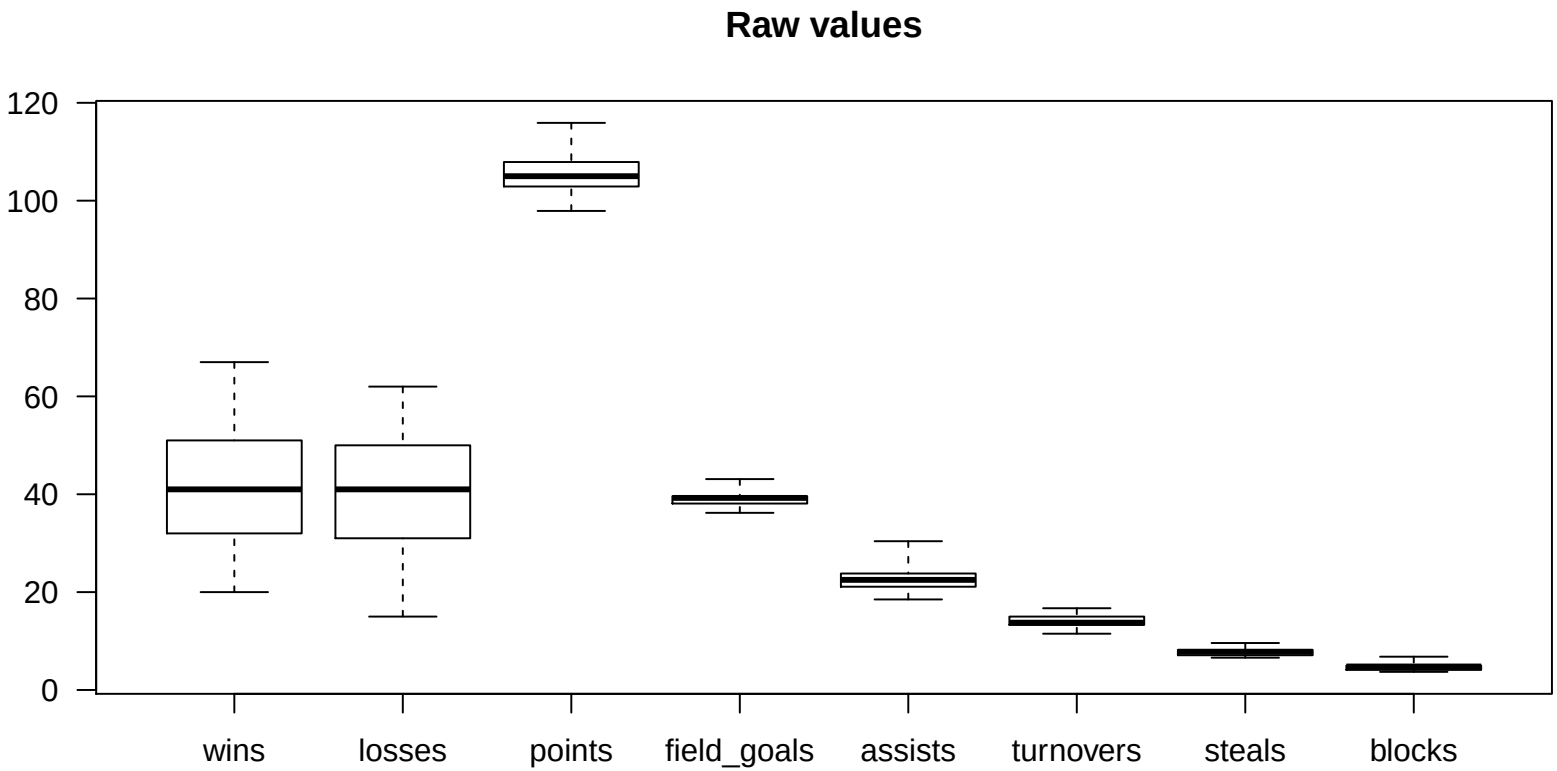
Scale of Variables

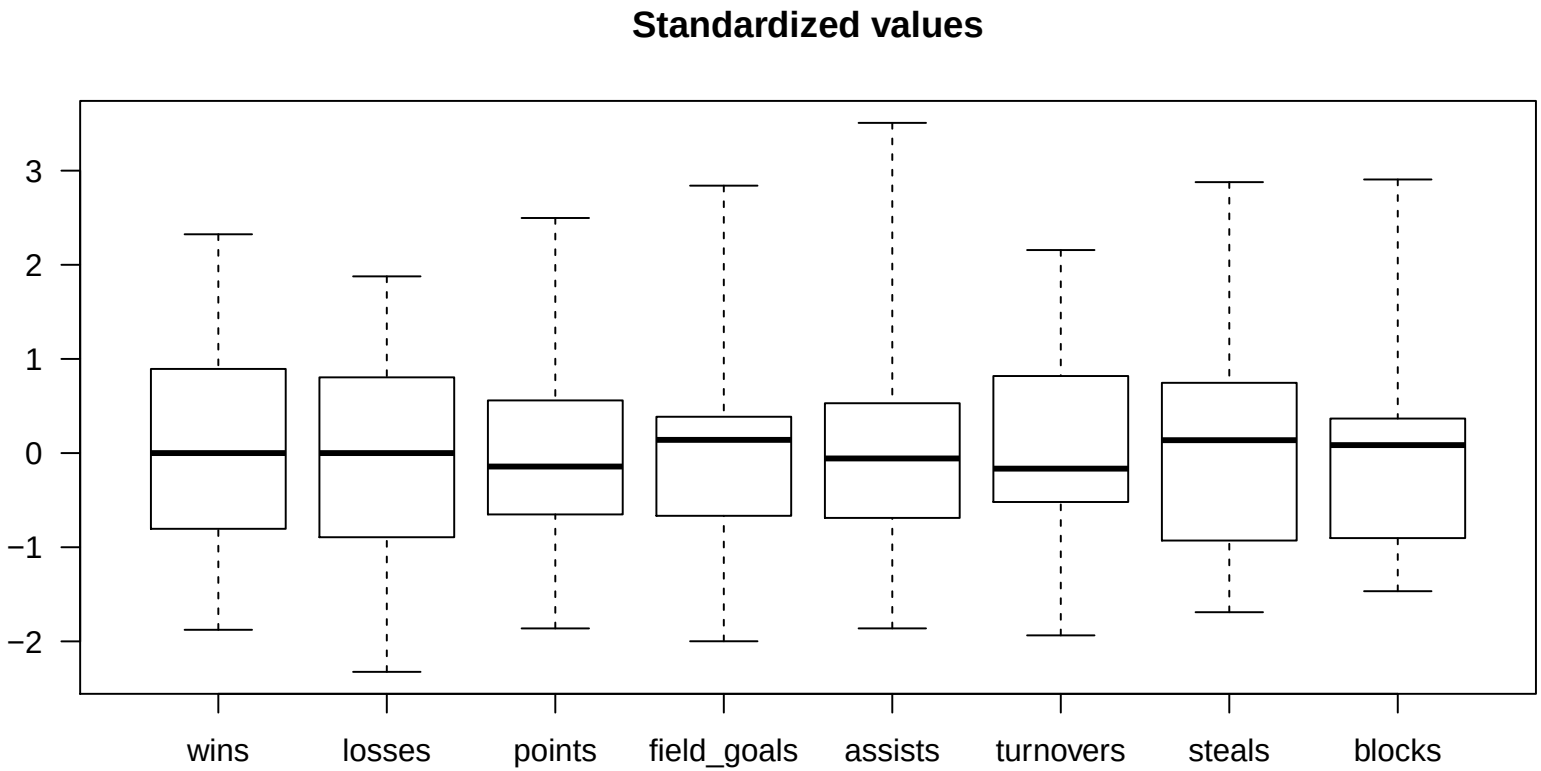
Importance of Variables

To standardize or not?

- A key issue has to do with the scale of the variables
- If variables have different units of measurement, then we should standardize them
- If variables have the same units:
  - you could leave them unstandardized
  - or you could standardize them (strongly suggested)

Regardless of the scaling decision, we operate on centered data.


Raw values
120
100
80
60
40
20
0
wins
losses
points
field_goals
assists
turnovers
steals
blocks


Standardized values
3
2
1
0
−1
−2
wins
losses
points
field_goals
assists
turnovers
steals
blocks

PCA

PCA via EVD

Let's work on the standardized variables.

PCA involves computing the eigenvalue decomposition of the (sample) correlation matrix $\mathbf{R} = \frac{1}{n-1}\mathbf{X}^\mathsf{T}\mathbf{X}$

$$\mathbf{R} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\mathsf{T}$$

where:

- $\mathbf{V}$ is the matrix of eigenvectors
- $\mathbf{\Lambda}$ is the matrix of eigenvalues

PCA via EVD

Principal Components (aka Scores)

$$\mathbf{Z} = \mathbf{X}\mathbf{V}$$

Loadings: the eigenvectors in $\mathbf{V}$ are referred to as loadings

The eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ capture the projected inertia.

How many PCs to retain?

Various criteria

- Screeplot (see if there's an "elbow")
- Predetermined amount of variation
- Kaiser rule
- Jollife rule

Table of Eigenvalues

	eigenvalues	proportion	cum_prop
comp1	3.68	46.01	46.01
comp2	1.62	20.22	66.23
comp3	1.02	12.73	78.96
comp4	0.62	7.77	86.73
comp5	0.47	5.90	92.63
comp6	0.46	5.77	98.40
comp7	0.13	1.60	100.00
comp8	0.00	0.00	100.00

What's going on with eigenvalue of PC8?

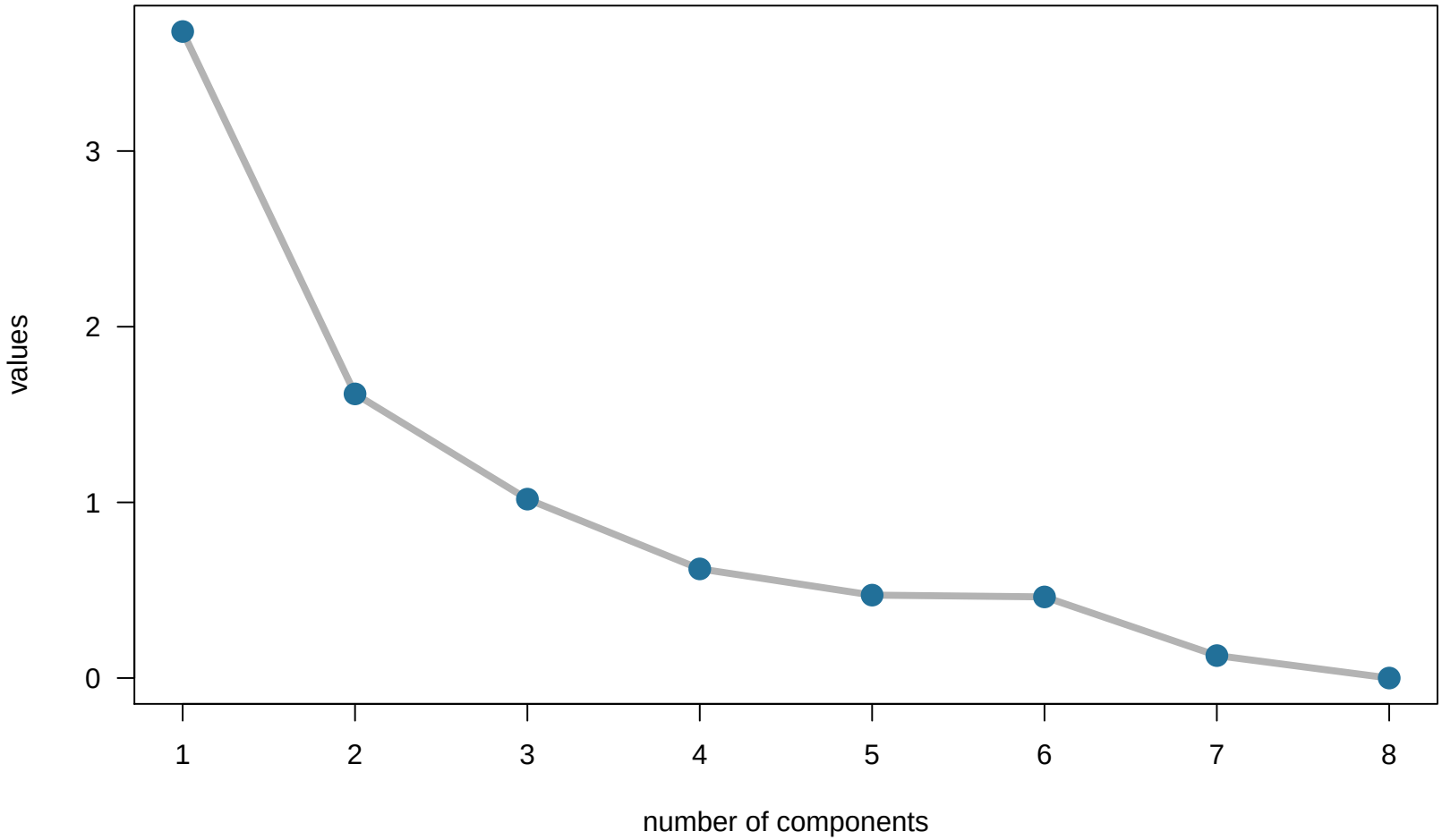
Table of Eigenvalues

When working with standardized data, the total amount of variation (i.e. total inertia) is equal to the number of Variables: ($p = 8$) in this case.

Likewise, the sum of the eigenvalues equals this total amount of variation: 8

Therefore, we calculate the portion of variation captured by each PC.

Screeplot of eigenvalues



Look for an "elbow"

Predetermined amount of variation

One option to decide how many PCs to retain, consists of predefining a specified portion of variation: e.g. 60% or 70%

```
      eigenvalues proportion cum_prop
comp1      3.6806    46.0071  46.0071
comp2      1.6177    20.2214  66.2285
```

Kaiser's Rule

Another criterion to decide how many PCs to keep, is the so-called Kaiser's rule, which consists of retaining those PCs with eigenvalues $\lambda_k > 1$

```
      eigenvalues proportion cum_prop
comp1    3.680569   46.00711 46.00711
comp2    1.617713   20.22142 66.22853
comp3    1.018539   12.73174 78.96027
```

Jollife's Rule

An alternative to Kaiser's rule is the less known Jollife's rule, in which we retain those PCs with eigenvalues $\lambda_k > 0.7$

```
      eigenvalues proportion cum_prop
comp1    3.680569   46.00711 46.00711
comp2    1.617713   20.22142 66.22853
comp3    1.018539   12.73174 78.96027
```

Studying the Individuals

Quality of Representation

When studying the individuals, we typically pay attention to:

- Scatterplots of PCs
- Quality of representation
- Contributions

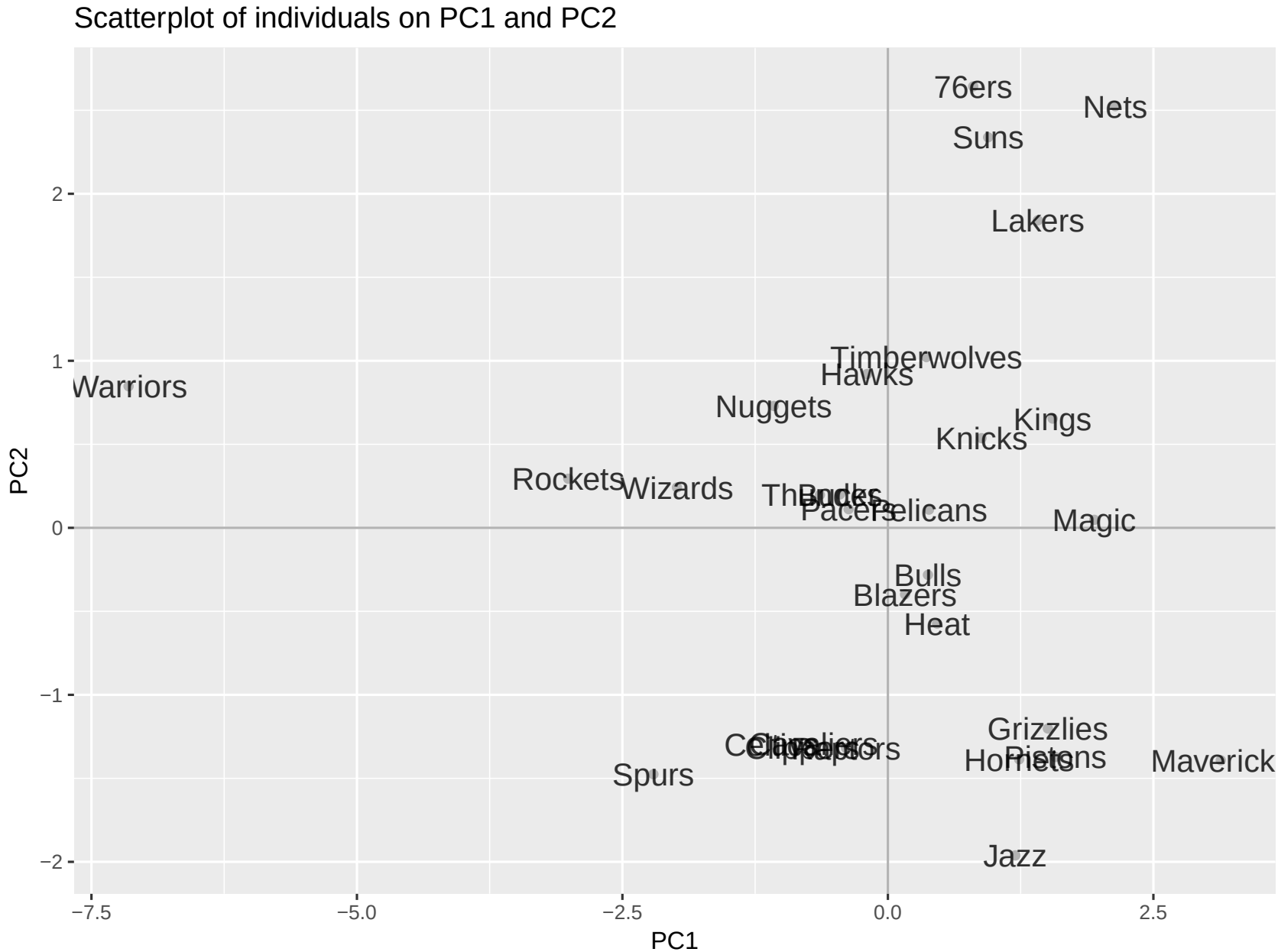
Principal Components

PCs are typically given by: $\mathbf{Z} = \mathbf{XV}$, although it is possible to rescale them (e.g. variance $= 1$ or unit-norm)

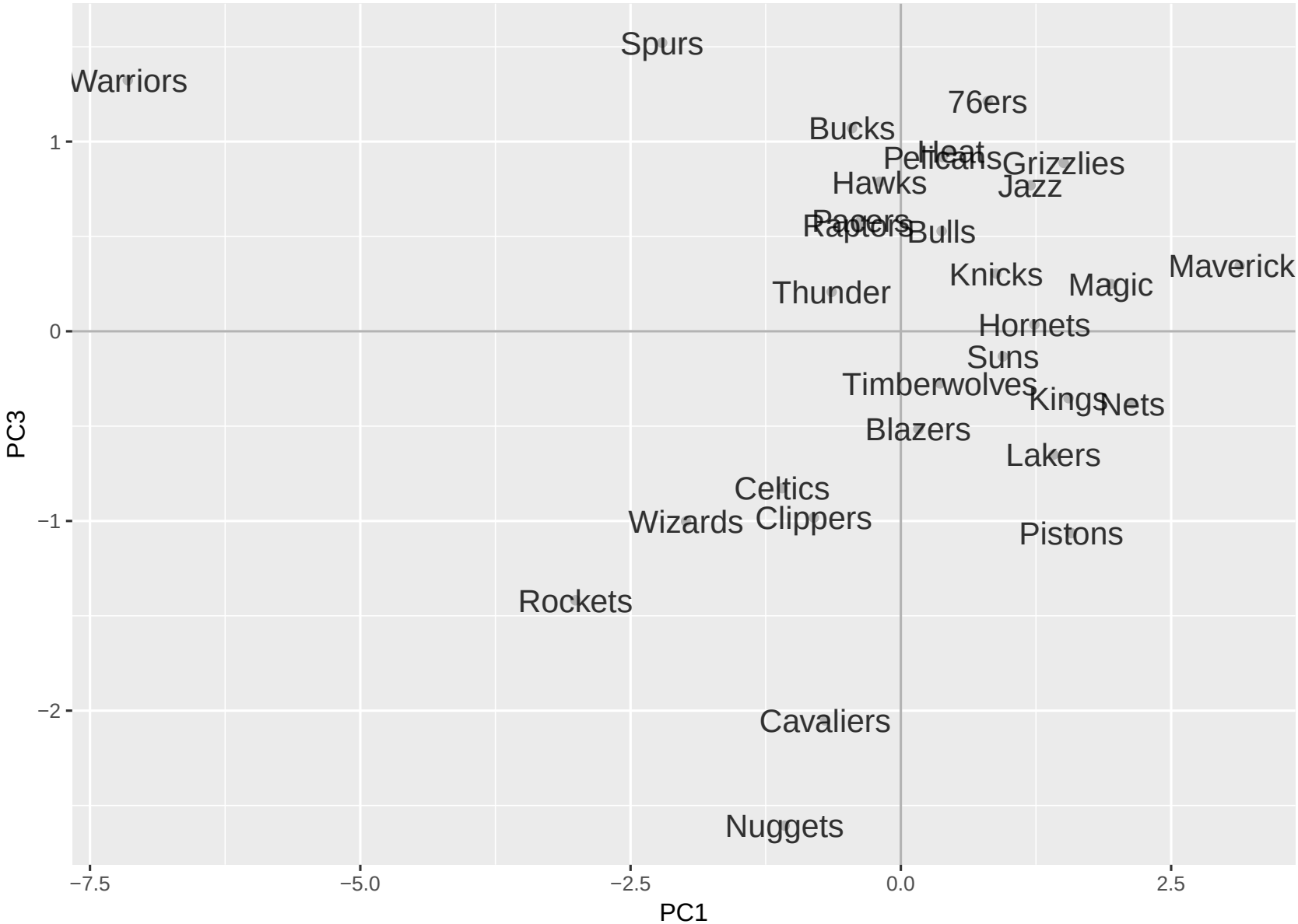
The first 10 rows of each PC are given below:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Warriors	-7.150	0.848	1.324	0.369	-0.687	-0.606	-0.024	0
Spurs	-2.208	-1.475	1.521	0.186	0.086	0.546	0.261	0
Rockets	-3.010	0.294	-1.418	-0.842	0.194	0.454	-0.646	0
Celtics	-1.098	-1.298	-0.827	-0.875	-0.869	0.340	-0.257	0
Jazz	1.200	-1.961	0.770	0.147	0.341	1.686	0.295	0
Raptors	-0.394	-1.318	0.560	-0.162	2.078	-0.553	-0.401	0
Cavaliers	-0.699	-1.290	-2.052	0.398	0.059	0.848	0.018	0
Clippers	-0.805	-1.313	-0.982	-0.232	0.295	-0.071	-0.195	0
Wizards	-1.986	0.242	-1.002	-0.802	0.491	-0.878	0.492	0
Thunder	-0.640	0.197	0.208	-0.023	1.104	0.631	0.227	0

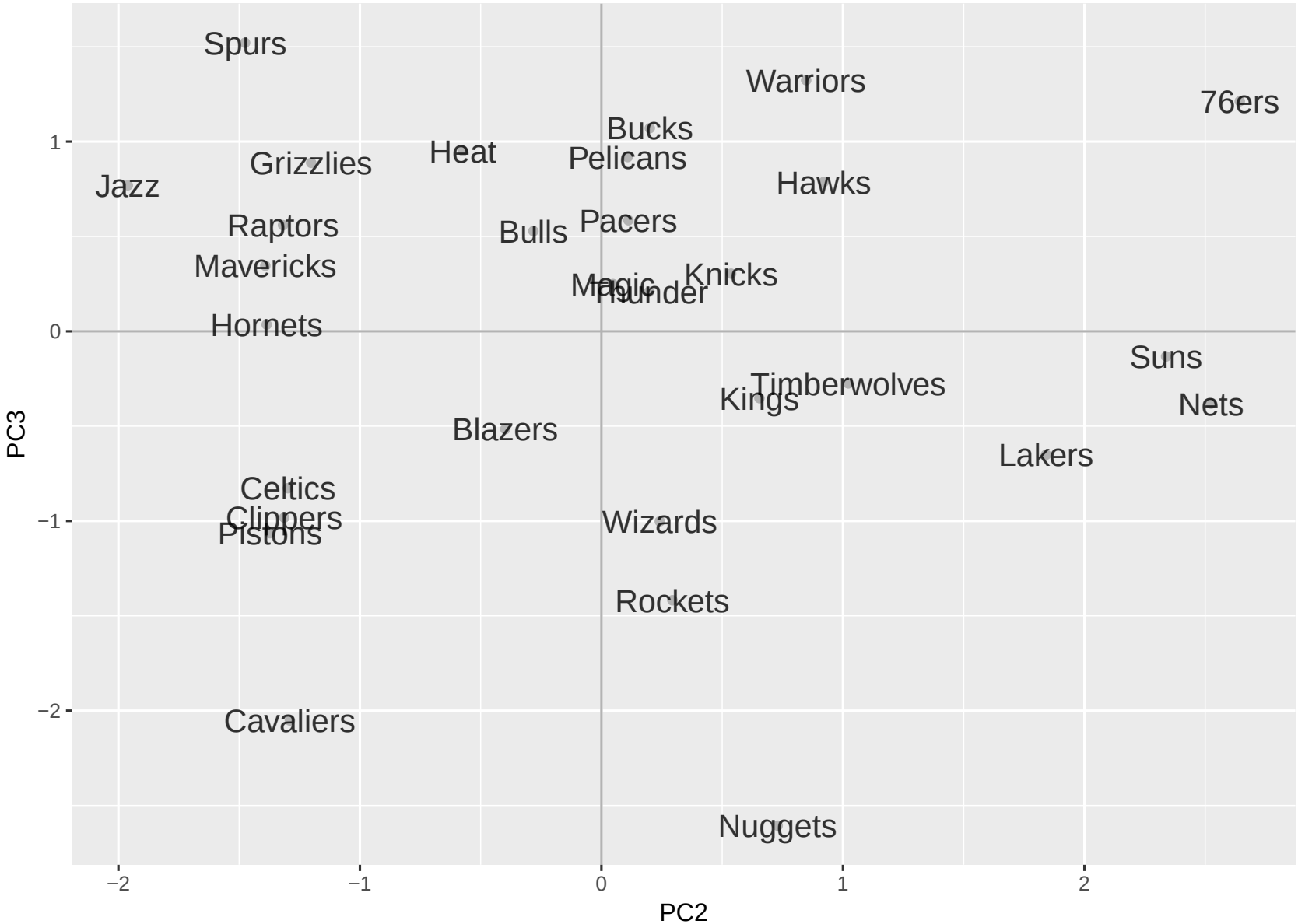
notice what happens with the last PC


Scatterplot of individuals on PC1 and PC2
76ers
Nets
Suns
Lakers
Timberwolves
Hawks
Warriors
Nuggets
Kings
Knicks
Rockets
Wizards
Pelicans
Magic
Bulls
Blazers
Heat
Grizzlies
Spurs
Hornets
Maverick
Jazz
PC2
PC1
2
1
0
−1
−2
−7.5
−5.0
−2.5
0.0
2.5

Scatterplot of individuals on PC1 and PC3



Scatterplot of individuals on PC2 and PC3



Quality of Representation

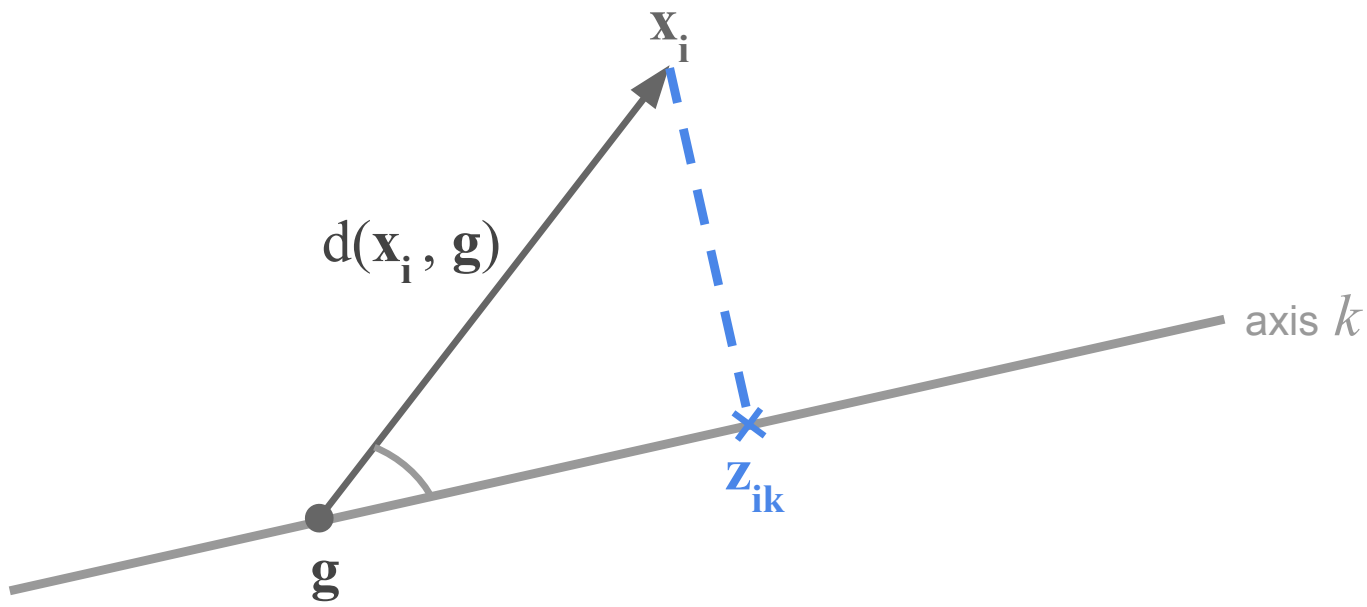
$$cos^2(i,k) = \frac{z_{ik}^2}{d^2(i,g)}$$

where:

- z_{ik} is the value of k-th PC for individual i
- $d^2(\mathbf{x_i}, \mathbf{g})$ is the squared distance of individual i to the centroig $\mathbf{g}$
- recall that with centered data, $\mathbf{g}$ is the origin

Cosine equal to 1 indicates that individual i is exactly on axis Δ_k (angle zero). Conversely, a cosine equal to 0 indicates that the individual i is on an orthogonal direction of axis Δ_k.

Quality of Representation



$$cos^2(i,k) = \frac{z_{ik}^2}{d^2(\mathbf{z_i}, \mathbf{g})}$$

Quality of Representation

Adding the squared cosines over all principal axes for a given individual, we get:

$$\sum_{k=1}^{p} = cos^2(i,k) = 1$$

This sum provides, in percentages, the "quality" of the representation of an individual on the subspace defined by the principal axes.

Quality of Representation

First 6 rows of $cos^2(i,k)$ for $k = 1, 2, 3, 4$

	PC1	PC2	PC3	PC4
Warriors	0.93682873	0.013170110	0.03212084	0.002495688
Spurs	0.49881739	0.222735807	0.23665382	0.003536342
Rockets	0.72317632	0.006904138	0.16053335	0.056596224
Celtics	0.22852854	0.319227201	0.12961118	0.144959032
Jazz	0.16098822	0.429843158	0.06618774	0.002415628
Raptors	0.02216717	0.247608673	0.04463893	0.003743516

Note that Warriors has a value close to 1 on PC1. On the other hand, Raptors has a value close to zero on PC1.

Quality of Representation

The squared cosine is used to evaluate the quality of the representation On a given PC, some distances between individuals will be well represented, while other distances will be highly distorted.

You can add the squared cosines of an individual over different axes, resulting in a "quality" measure of how well that individual is represented in that subspace.

Contributions

$$ctr(i,k) = \frac{m_i z_{ik}^2}{\lambda_k} \times 100$$

where:

- m_i is the mass or weight of individual i, in thise case: $(\frac{1}{n-1})$
- z_{ik} is the value of k-th PC for individual i
- λ_k is the eigenvalue associated to k-th PC

Contributions

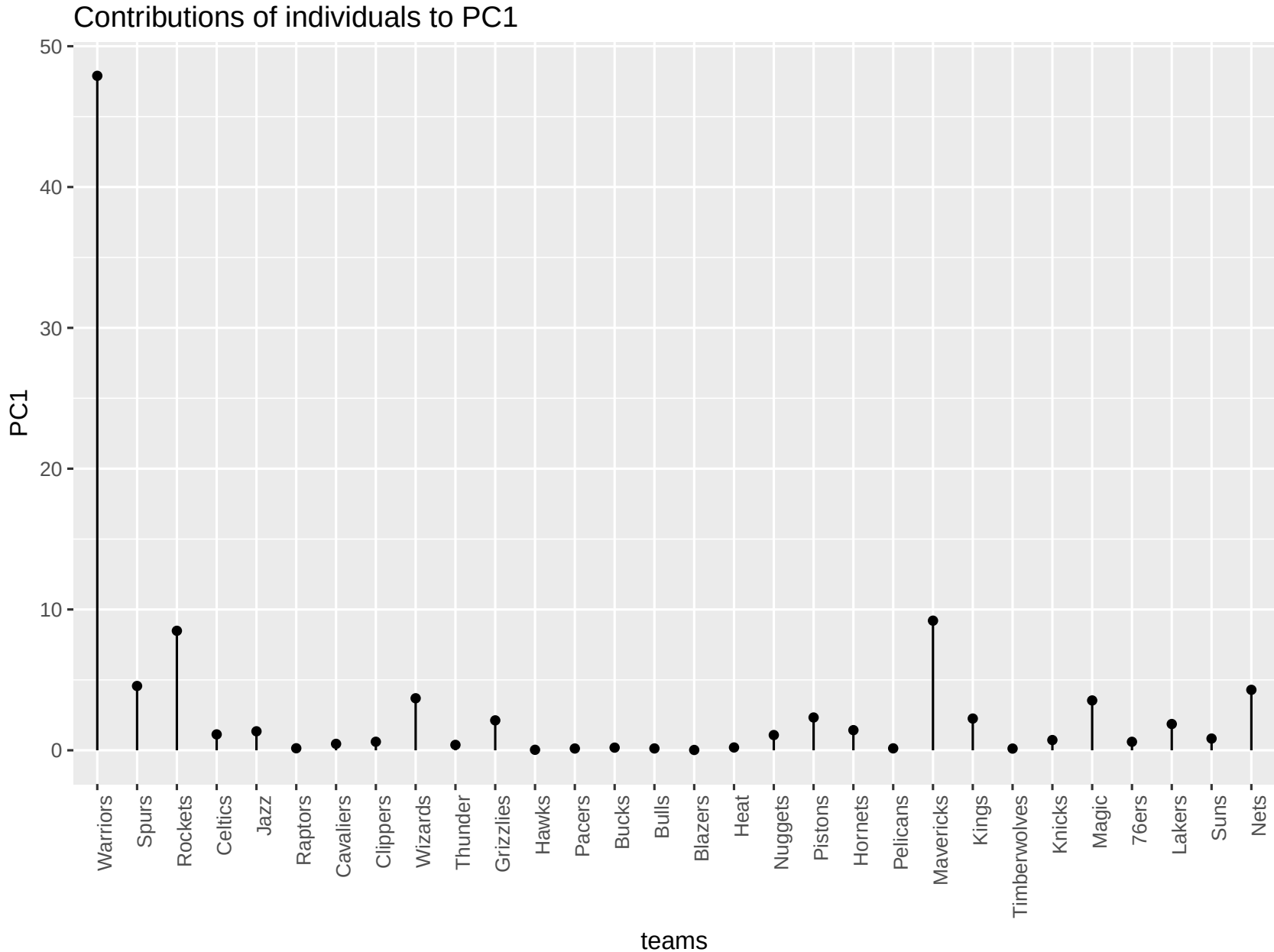
- $ctr(i, k)$ is the contribution of an individual to the construction of component k.
- Note that contributions are expressed in percentages.
- The contribution is calculated as the inertia explained by the individual i on the component k
- Outliers have an influence, and it is interesting to know to what extent their influence affects the construction of the components.
- Detecting those individuals that contribute to a given PC helps to assess the component's stability.

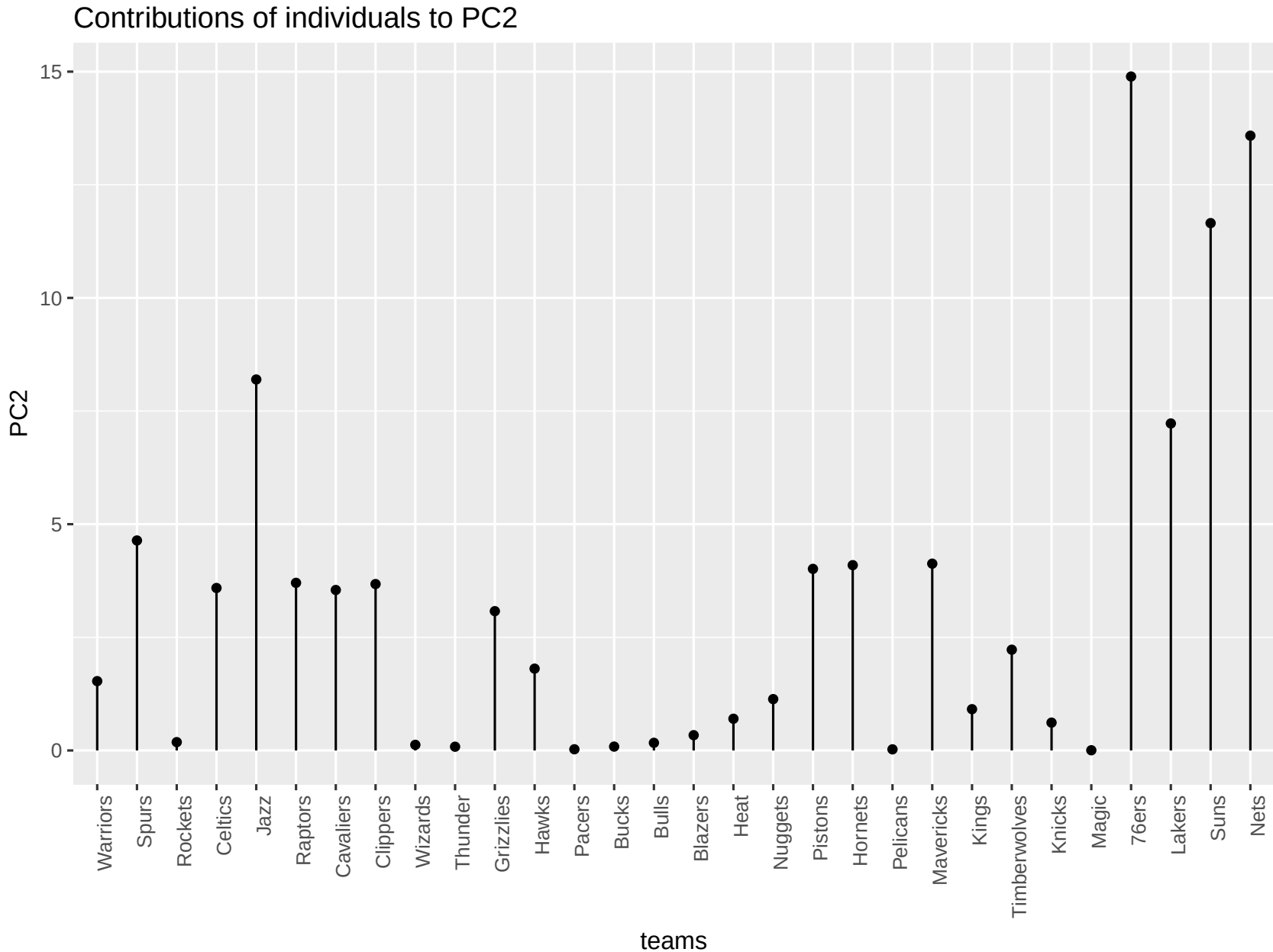
Contributions

First 6 rows of $ctr(i, k)$ for $k = 1, 2, 3, 4$

	PC1	PC2	PC3	PC4
Warriors	47.8979178	1.5320017	5.934452	0.7557724
Spurs	4.5678923	4.6406383	7.831142	0.1918107
Rockets	8.4869097	0.1843437	6.807819	3.9340253
Celtics	1.1297516	3.5905074	2.315380	4.2445545
Jazz	1.3495386	8.1981272	2.004962	0.1199404
Raptors	0.1457559	3.7042069	1.060638	0.1457941

Note that Warriors has a large contribution to PC1. On the other hand, Raptors has a value close to zero on PC1.


Contributions of individuals to PC1
PC1
50
40
30
20
10
0
Warriors
Spurs
Rockets
Celtics
Jazz
Raptors
Cavaliers
Clippers
Wizards
Thunder
Grizzlies
Hawks
Pacers
Bucks
Bulls
Blazers
Heat
Nuggets
Pistons
Hornets
Pelicans
Mavericks
Kings
Timberwolves
Knicks
Magic
76ers
Lakers
Suns
Nets
teams


Contributions of individuals to PC2
15
10
5
0
PC2
Warriors
Spurs
Rockets
Celtics
Jazz
Raptors
Cavaliers
Clippers
Wizards
Thunder
Grizzlies
Hawks
Pacers
Bucks
Bulls
Blazers
Heat
Nuggets
Pistons
Hornets
Pelicans
Mavericks
Kings
Timberwolves
Knicks
Magic
76ers
Lakers
Suns
Nets
teams

More about Contributions

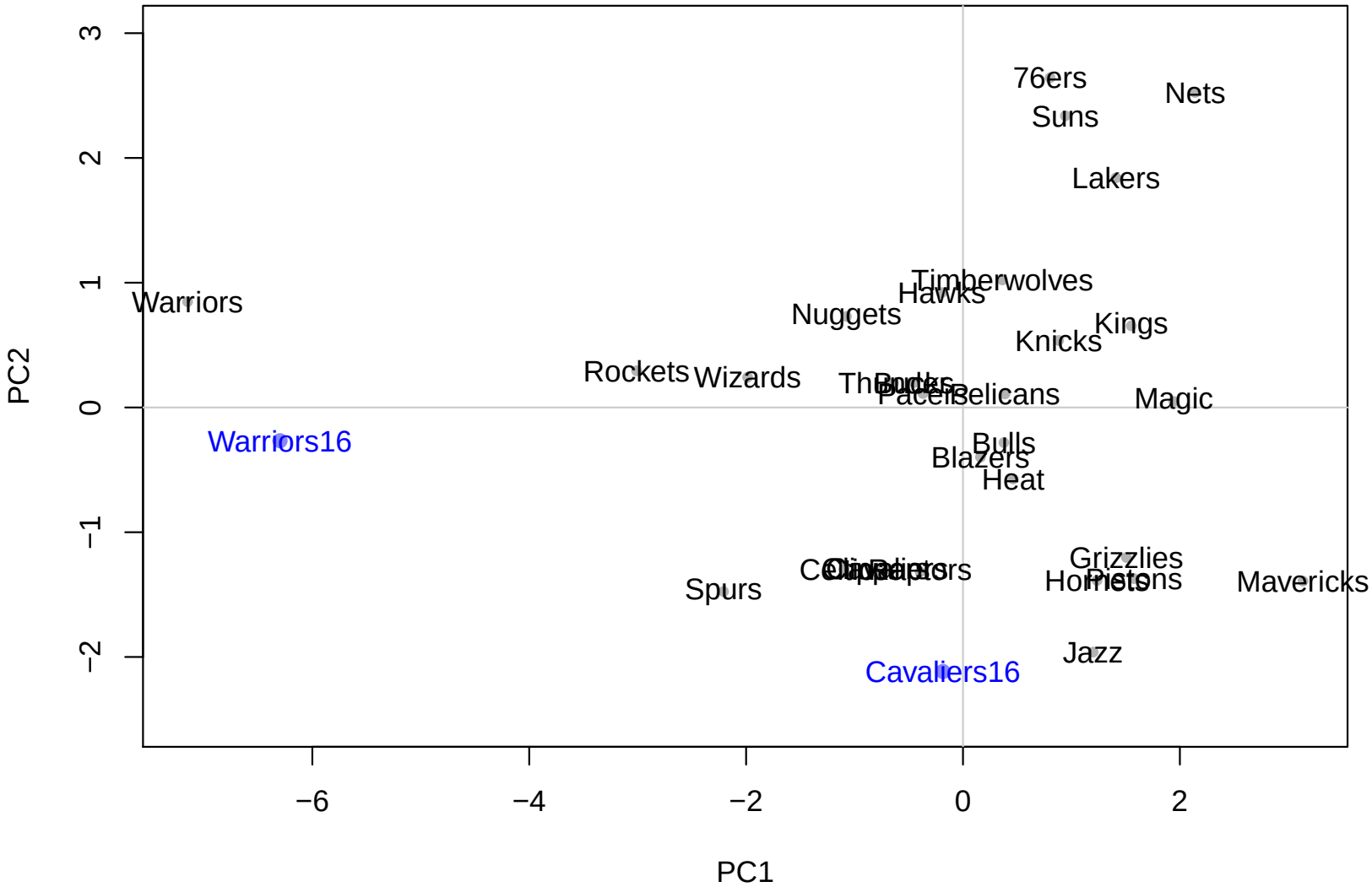
The variance of $\mathbf{z_k}$ is equal to:

$$var(\mathbf{z_k}) = \sum_{i=1}^{n} m_i z_{ik}^2$$

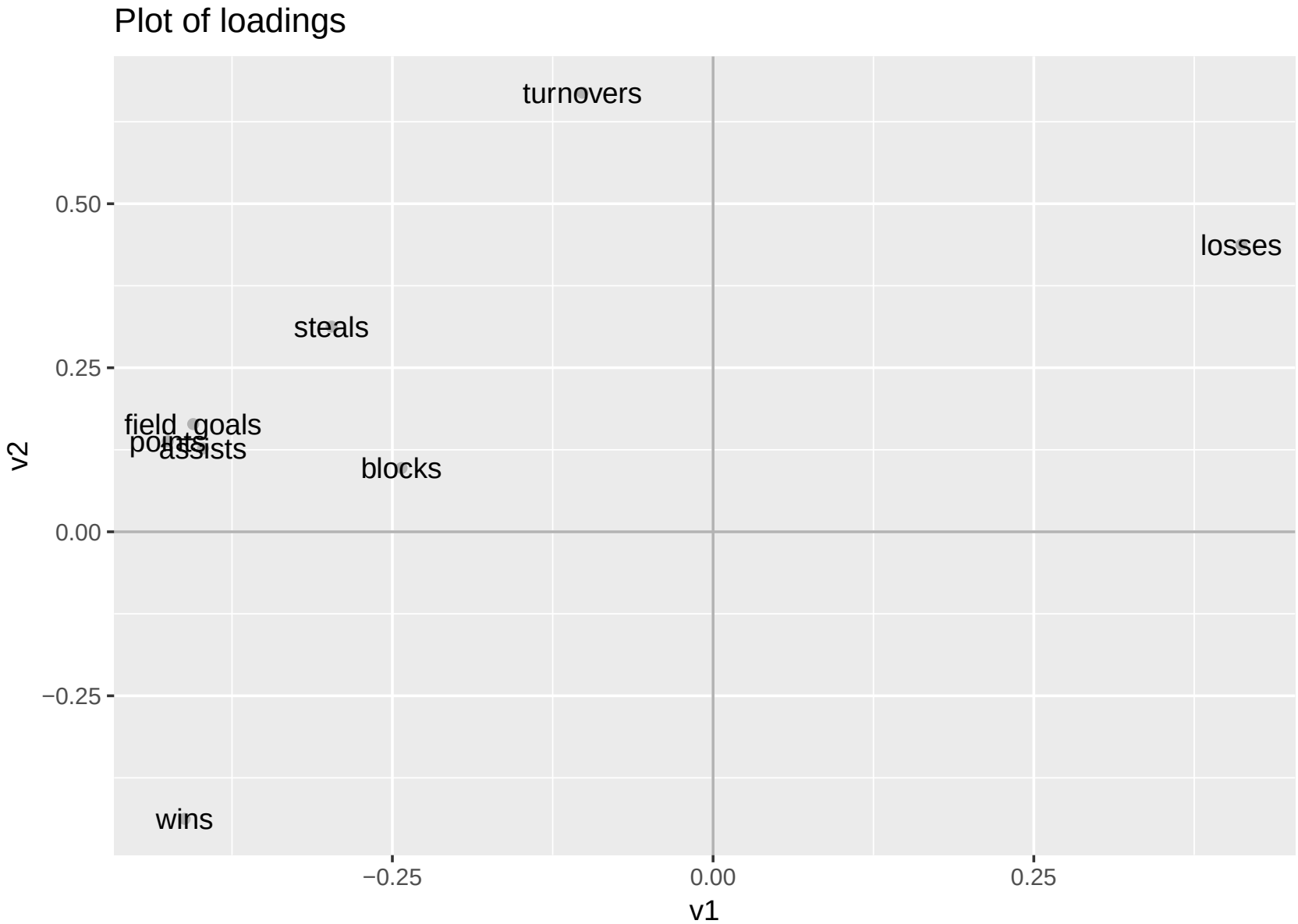
which is equal to λ_k

If all individuals had a uniform contribution to a given PC, then they would have a contribution $ctr(i,k) = \frac{1}{n-1}$.

Representing Supplementary Individuals

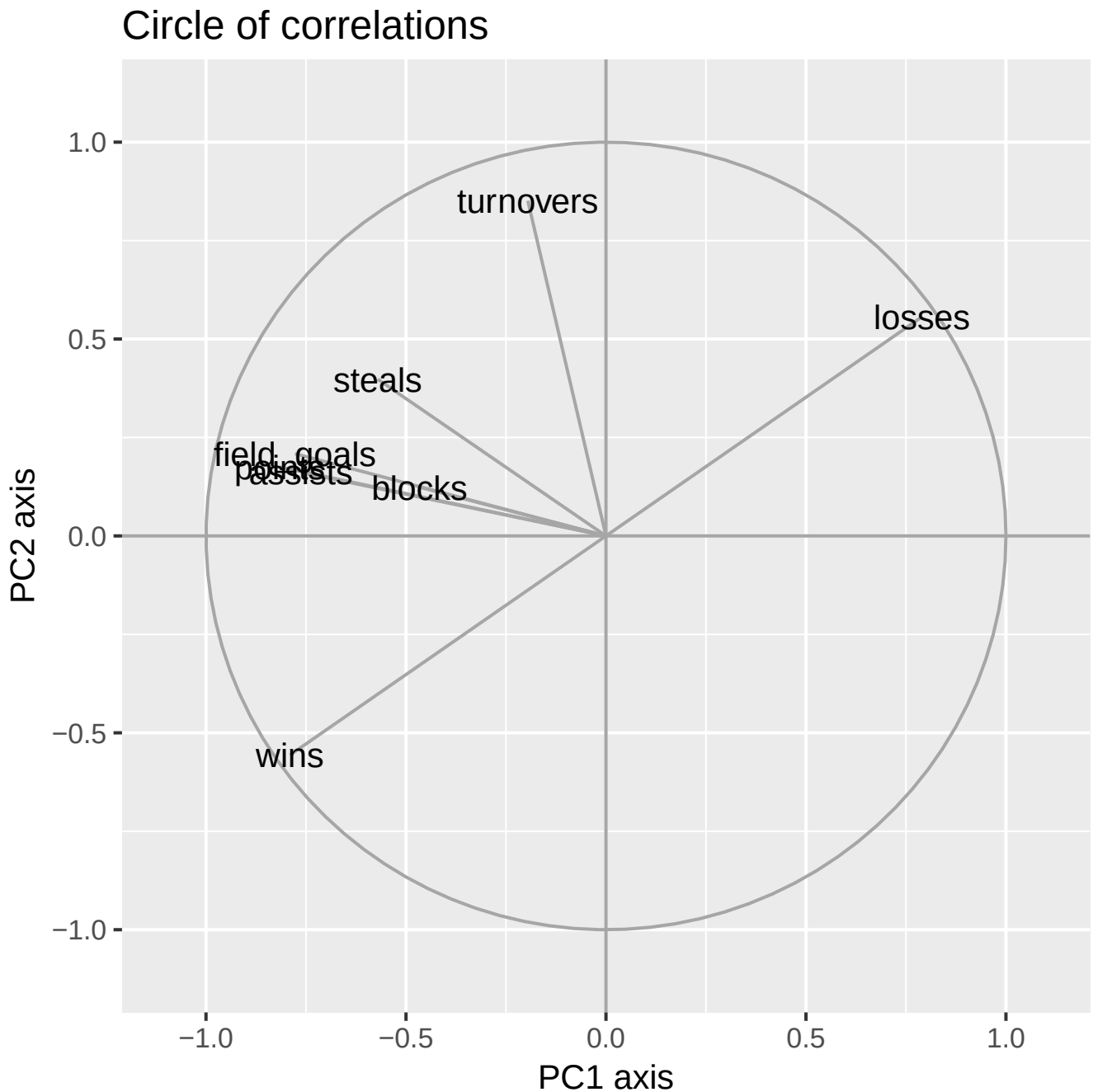

76ers
Nets
Suns
Lakers
Timberwolves
Hawks
Warriors
Nuggets
Kings
Knicks
Rockets
Wizards
Pelicans
Magic
Warriors16
Bulls
Blazers
Heat
Grizzlies
Spurs
Hornets
Pistons
Mavericks
Jazz
Cavaliers16
PC1
PC2
−6
−4
−2
0
2
3
2
1
0
−1
−2

Study cloud of Variables


Plot of loadings
turnovers
losses
steals
field_goals
points
assists
blocks
wins
0.50
0.25
0.00
−0.25
v2
−0.25
0.00
0.25
v1

Correlations between Variables and PCs

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
wins	-0.790	-0.556	0.055	-0.148	0.095	0.174	0.046	-0.793
losses	0.790	0.556	-0.055	0.148	-0.095	-0.174	-0.046	0.793
points	-0.815	0.175	-0.453	0.126	0.112	0.032	-0.264	0.005
field_goals	-0.777	0.209	-0.333	0.325	0.140	-0.272	0.205	0.185
assists	-0.763	0.162	-0.030	-0.100	-0.616	-0.032	0.015	-0.155
turnovers	-0.195	0.851	-0.049	-0.150	0.101	0.441	0.088	0.531
steals	-0.571	0.398	0.422	-0.428	0.179	-0.348	-0.042	-0.047
blocks	-0.466	0.124	0.718	0.490	-0.003	0.101	-0.047	-0.126


Circle of correlations
turnovers
losses
steals
field_goals
blocks
wins
PC2 axis
PC1 axis
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0

Interpreting the PC Dimensions

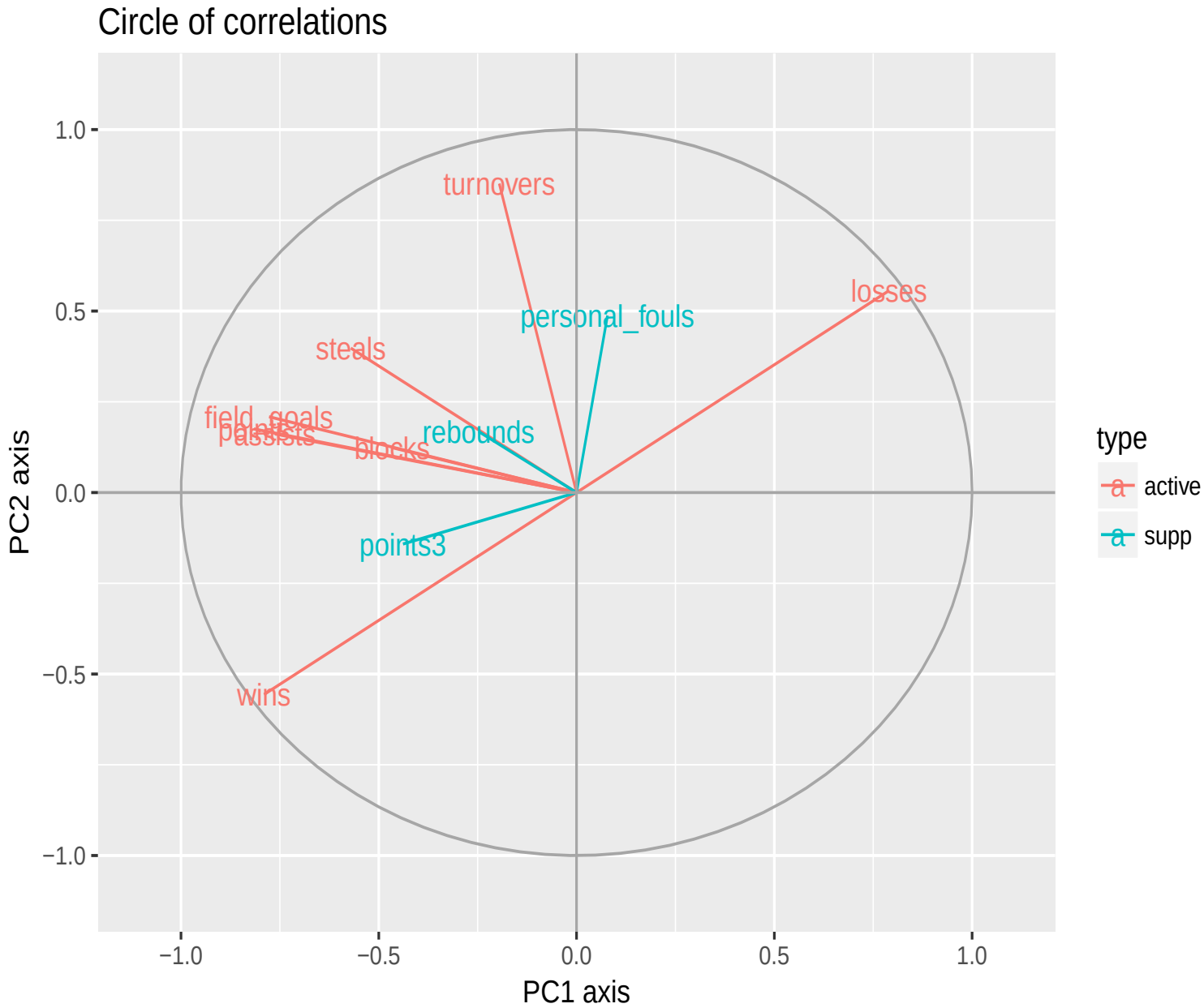
Interpreting PCs

- Being linear combinations of the studied variables, PCs can sometimes being interpreted
- Analysts try to label them in some meaningful way
- This is useful, but not always possible
- You can look at the magnitude of the loadings
- You can also look at the correlations between variables and PCs

Representing Supplementary Variables

Correlations between all Variables and PCs

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
wins	-0.790	-0.556	0.055	-0.148	0.095	0.174	0.046	-0.793
losses	0.790	0.556	-0.055	0.148	-0.095	-0.174	-0.046	0.793
points	-0.815	0.175	-0.453	0.126	0.112	0.032	-0.264	0.005
field_goals	-0.777	0.209	-0.333	0.325	0.140	-0.272	0.205	0.185
assists	-0.763	0.162	-0.030	-0.100	-0.616	-0.032	0.015	-0.155
turnovers	-0.195	0.851	-0.049	-0.150	0.101	0.441	0.088	0.531
steals	-0.571	0.398	0.422	-0.428	0.179	-0.348	-0.042	-0.047
blocks	-0.466	0.124	0.718	0.490	-0.003	0.101	-0.047	-0.126
points3	-0.439	-0.142	-0.353	-0.123	-0.233	0.369	-0.425	-0.324
rebounds	-0.247	0.168	-0.219	0.423	0.235	0.148	0.206	0.248
personal_fouls	0.078	0.488	0.021	-0.148	0.474	-0.002	-0.153	0.363


Circle of correlations
1.0
0.5
0.0
−0.5
−1.0
PC2 axis
turnovers
losses
personal_fouls
steals
field_goals
rebounds
blocks
points3
wins
−1.0
−0.5
0.0
0.5
1.0
PC1 axis
type
active
supp

References

- **Exploratory Multivariate Analysis by Example Using R** by Husson, Le and Pages (2010). *Chapter 1: Principal Component Analysis (PCA)*. CRC Press.
- **An R and S-Plus Companion to Multivariate Analysis** by Brian Everitt (2004). *Chapter 3: Principal Components Analysis*. Springer.
- **Principal Component Analysis** by Ian jolliffe (2002). Springer.
- **Data Mining and Statistics for Decision Making** by Stephane Tuffery (2011). *Chapter 7: Factor Analysis*. Editions Technip, Paris.

References (French Literature)

- **Statistique Exploratoire Multidimensionnelle** by Lebart et al (2004). *Chapter 3, section 3: Analyse factorielle discriminante*. Dunod, Paris.
- **Probabilites, analyse des donnees et statistique** by Gilbert Saporta (2011). *Chapter 6: Analyse en Composantes Principaux*. Editions Technip, Paris.
- **Statistique: Methodes pour decrire, expliquer et prevoir** by Michel Tenenhaus (2008). *Chapter 10: L'analyse discriminante*. Dunod, Paris.
- **Analyses factorielles simples et multiples** by Brigitte Escofier et Jerome Pages (2016, 5th edition). *Chapter 2: L'analyse discriminante*. Dunod, Paris.