Inertia: Multivariate Dispersion

Matrix Algebra 4 Statistical Learning

Gaston Sanchez

Introduction

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	SEASON TYPE	PER MODE	SEASON SEGMENT	
2016-17	**Regular Season**	**Per Game**	**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')
```

```r
dim(dat)
```

```
[1] 30 27
```

```r
names(dat)
```

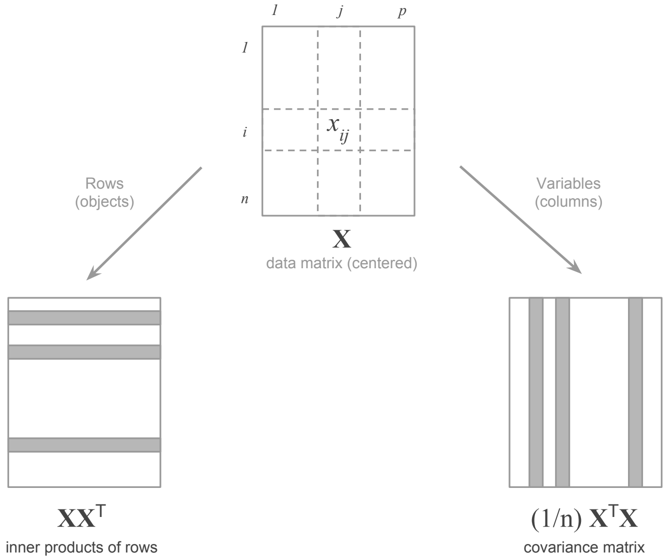
```
 [1] "team"                "games_played"          "wins"
 [4] "losses"              "win_prop"              "minutes"
 [7] "points"              "field_goals"           "field_goals_attempted"
[10] "field_goals_prop"    "points3"               "points3_attempted"
[13] "points3_prop"        "free_throws"           "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"
```

Exploratory Data Analysis

For illustration purposes, let's focus on the following variables:

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

EDA: Objects and Variables Perspectives



EDA: Objects and Variables Perspectives

Data Perspectives

We are interested in analyzing a data set from both perspectives: **objects** and **variables**

At its simplest we are interested in 2 fundamental purposes:

- Study resemblance among individuals
  (resemblance among NBA teams)
- Study relationship among variables
  (relationship among team statistics)

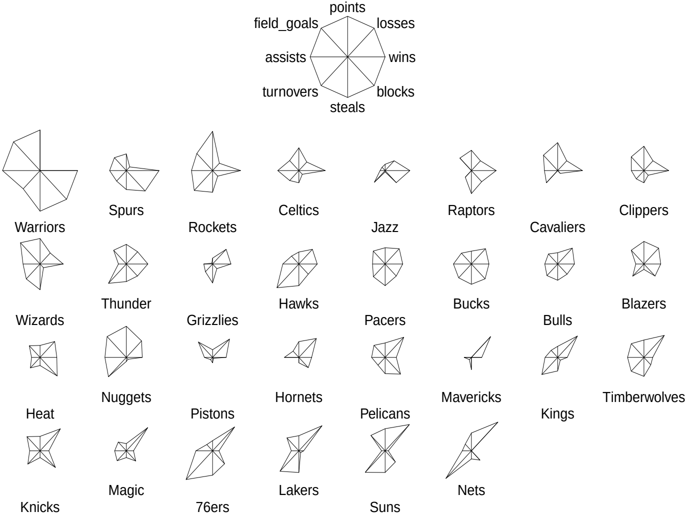
EDA

Exploration

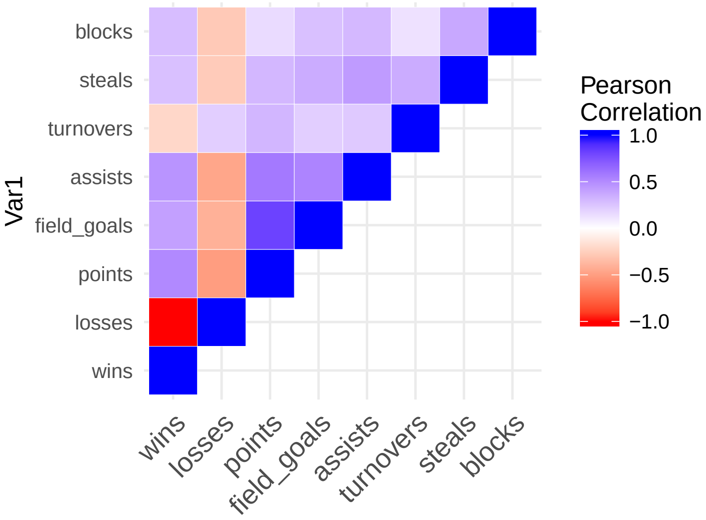
Likewise, we can explore variables at different stages:

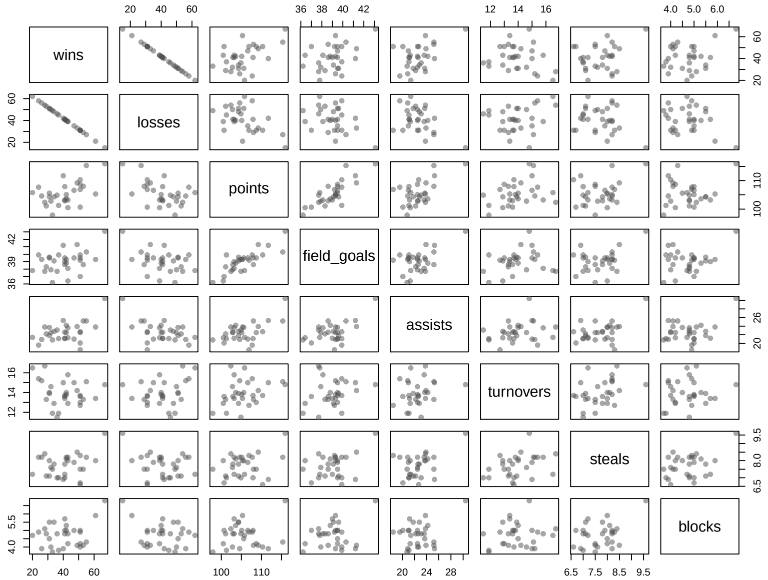
- Univariate: one variable at a time
- Bivariate: two variables simultaneously
- Multivariate: multiple variables

Let's see a shiny-app demo (see apps/ folder in github repo)


points
losses
wins
blocks
steals
turnovers
assists
field_goals
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

Correlation heatmap




wins
losses
points
field_goals
assists
turnovers
steals
blocks

Can we get a measure of multivariate dispersion?

How to measure dispersion? The concept of Inertia

Sum of Squared Distances

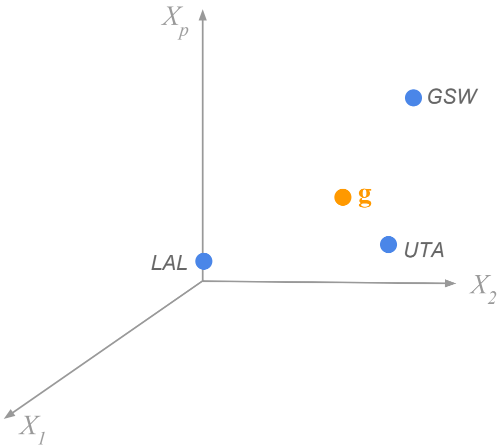
Pair-wise Squared distances

One way to consider the dispersion of data (in a mathematical form) is by adding the squared distances among all pairs of points.

Squared distances from centroid

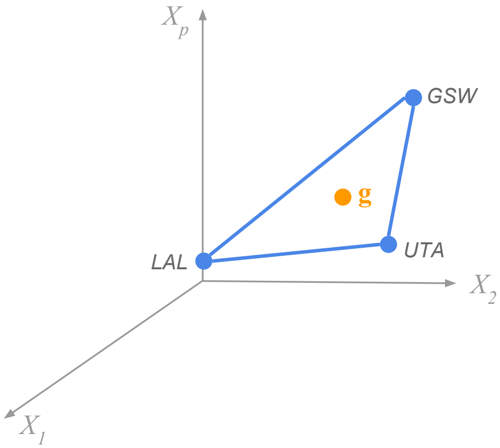
Another way to measure the dispersion of data is by considering the squared distances of all points around the center of gravity (i.e. centroid)

Imagine 3 points and its centroid



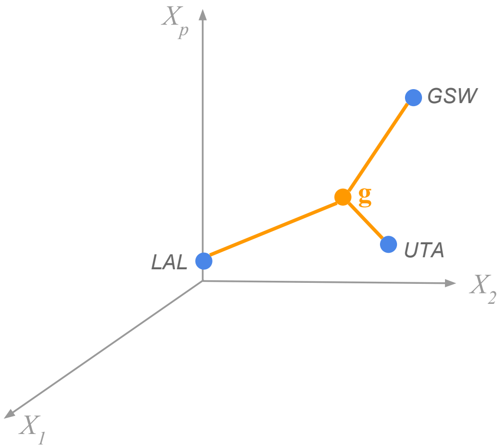
Centroid $\mathbf{g}$ is the "average" team.

Dispersion: Sum of all squared dists



$$\text{SSD} = 2d^2(\text{LAL}, \text{GSW}) + 2d^2(\text{LAL}, \text{UTA}) + 2d^2(\text{GSW}, \text{UTA})$$

2n × (sum of squared dists w.r.t. centroid)



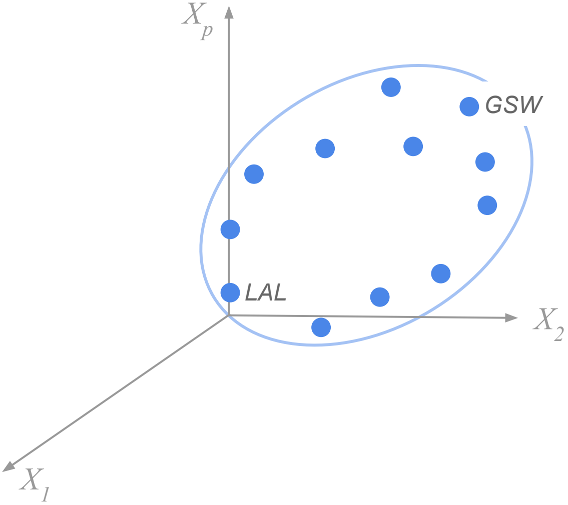
$$\mathsf{SSD} = (2 \times 3) \times \{d^2(\mathsf{LAL}, \mathbf{g}) + d^2(\mathsf{GSW}, \mathbf{g}) + d^2(\mathsf{UTA}, \mathbf{g})\}$$

Inertia

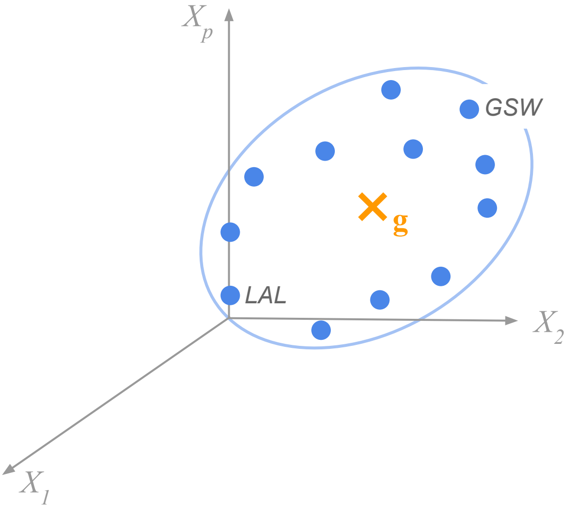
One way to take into account the dispersion of the data is with the concept of **Inertia**.

- Inertia is a term borrowed from the *moment of inertia* in mechanics (physics).
- This involves thinking about data as a rigid body (i.e. particles).
- We use the term Inertia to convey the idea of dispersion in the data.
- In multivariate methods, the term **Inertia generalizes the notion of variance**.
- Think of Inertia as a "multidimensional variance"

Cloud of teams in p-dimensional space



Centroid (i.e. the average team)

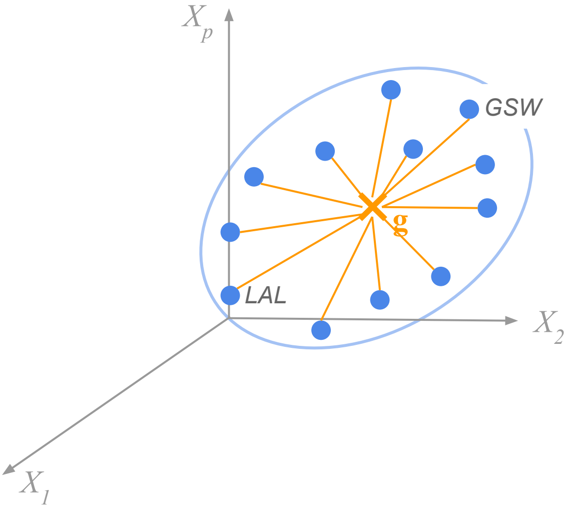


Formula of Total Inertia

The Total Inertia, I, is a weighted sum of squared distances among all pairs of objects:

$$I = \frac{1}{2n^2} \sum_{i=1}^{n} \sum_{h=1}^{n} d^2(i, h)$$

Overall variation/spread (around centroid)



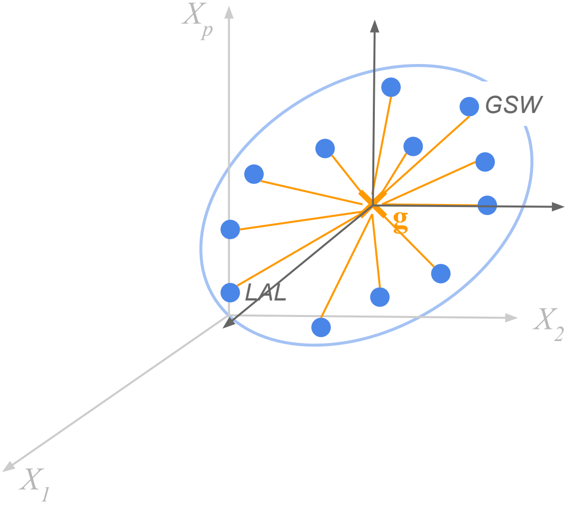
Formula of Total Inertia

Equivalently, the Total Inertia can be calculated in terms of the centoid $\mathbf{g}$:

$$I = \frac{1}{n} \sum_{i=1}^{n} d^2(\mathbf{x_i}, \mathbf{g})$$

The Inertia is an average sum of squared distances around the centroid $\mathbf{g}$

Centered data: centroid is the origin



Computing Inertia

$$
\begin{aligned}
Inertia &= \sum_{i=1}^{n} m_i d^2(\mathbf{x_i}, \mathbf{g}) \\
&= \sum_{i=1}^{n} \frac{1}{n} (\mathbf{x_i} - \mathbf{g})^\mathsf{T} (\mathbf{x_i} - \mathbf{g}) \\
&= \frac{1}{n} tr(\mathbf{X}^\mathsf{T} \mathbf{X}) \\
&= \frac{1}{n} tr(\mathbf{X} \mathbf{X}^\mathsf{T})
\end{aligned}
$$

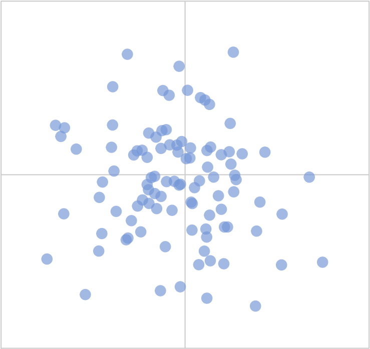
where m_i is the mass (i.e. weight) of individual i, usually $1/n$

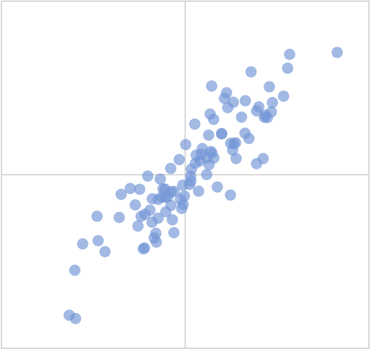
Inertia? What for?

What's Important?

Two data sets can have the same inertia. The amount of dispersion is important, but it is also important the shape-form of that dispersion.

Two data sets with similar inertia but different shape


Inertia = 2.02


Inertia = 2