Principal Components Analysis (part I)

Predictive Modeling & Statistical Learning

Gaston Sanchez

Introduction

NBA Team Stats

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

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

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

Data Perspectives

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

Main Interests

At its simplest we are interested in 2 fundamental purposes:

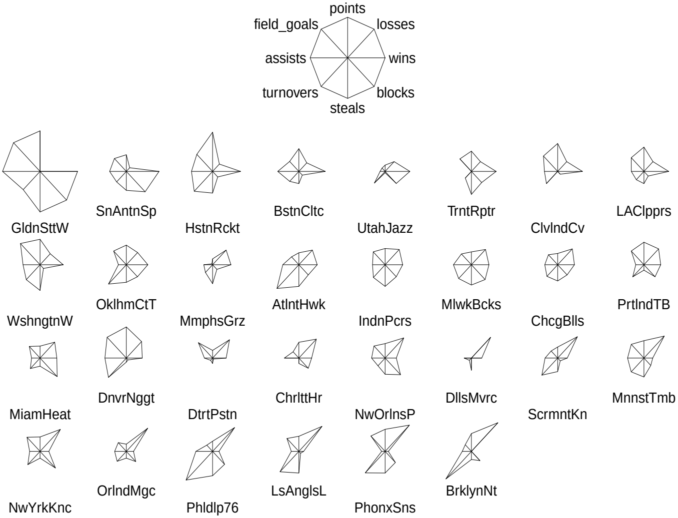
- Study resemblance among individuals
  (resemblance among NBA teams)
- Study relationship among variables
  (relationship among team statistics)

Exploration

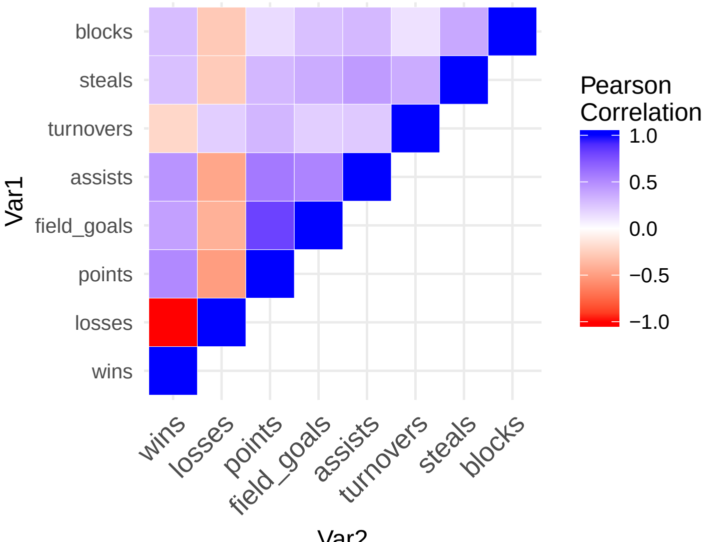
Likewise, we can explore variables at different stages:

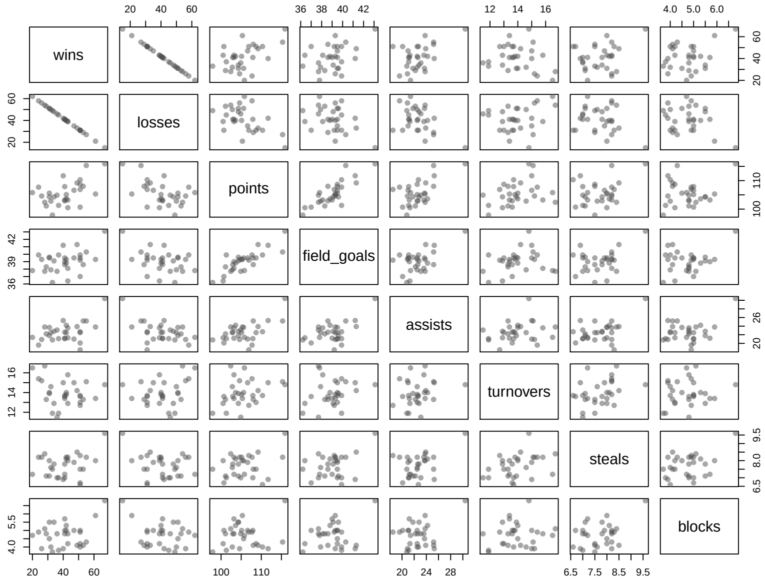
- Univariate: one variable at a time
- Bivariate: two variables simultaneously
- Multivariate: multiple variables

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


points
losses
wins
blocks
steals
turnovers
assists
field_goals
GldnSttW
SnAntnSp
HstnRckt
BstnCltc
UtahJazz
TrntRptr
ClvlndCv
LAClpprs
WshngtnW
OklhmCtT
MmphsGrz
AtlntHwk
IndnPcrs
MlwkBcks
ChcgBlls
PrtlndTB
MiamHeat
DnvrNggt
DtrtPstn
ChrlttHr
NwOrlnsP
DllsMvrc
ScrmntKn
MnnstTmb
NwYrkKnc
OrlndMgc
Phldlp76
LsAnglsL
PhonxSns
BrklynNt

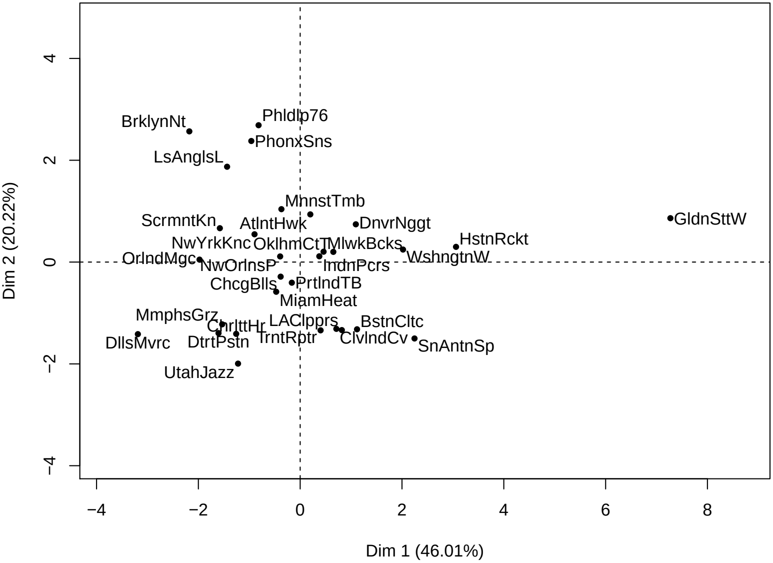
Correlation heatmap



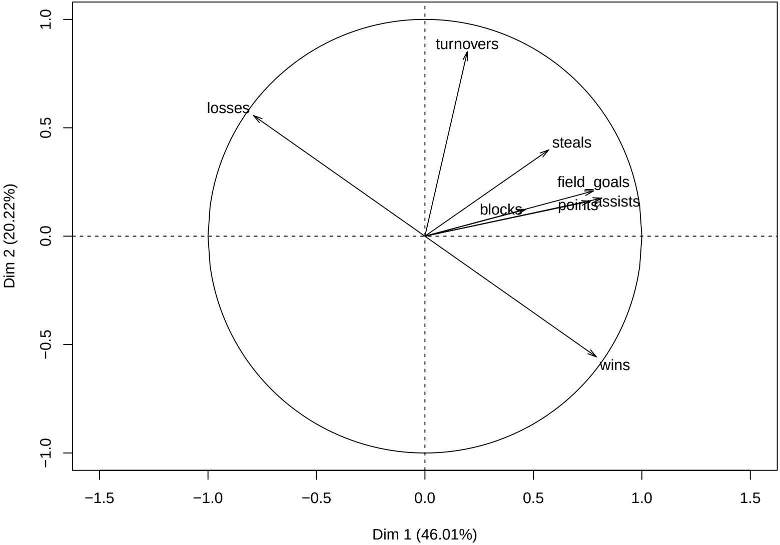


wins
losses
points
field_goals
assists
turnovers
steals
blocks

What if we could get a better low-dimensional summary of the data?



BrklynNt
Phldlp76
PhonxSns
LsAnglsL
MnnstTmb
ScrmntKn
AtlntHwk
DnvrNggt
GldnSttW
HstnRckt
NwYrkKnc
OklhmCtT
MlwkBcks
WshngtnW
OrlndMgc
NwOrlnsP
IndnPcrs
ChcgBlls
PrtlndTB
MiamHeat
MmphsGrz
ChrlttHr
LAClpprs
BstnCltc
DllsMvrc
DtrtPstn
TrntRptr
ClvlndCv
SnAntnSp
UtahJazz
Dim 2 (20.22%)
Dim 1 (46.01%)
−4
−2
0
2
4
6
8


turnovers
losses
steals
field_goals
blocks
points
assists
wins
1.0
0.5
0.0
−0.5
−1.0
Dim 2 (20.22%)
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Dim 1 (46.01%)

About PCA

Data Structure

Principal Components Analysis (PCA) is a multivariate method that allows us to study and explore a set of quantitative variables measured on some objects.

Landmarks

- PCA was first introduced by Karl Pearson (1904)
  On lines and planes of closest fit to systems of points in space
- Further developed by Harold Hotelling (1933)
  Analysis of a complex of statistical variables into principal components
- Singular Value Decomposition (SVD) theorem by Eckart-Young (1936)
  The approximation of a matrix by another of a lower rank
- Computationally implemented in the 1960s

Core Idea

With PCA we seek to **reduce the dimensionality** (condense information in variables) of a data set while retaining as much as possible of the variation present in the data

PCA: Overall Goals

- Summarize a data set with the help of a small number of synthetic variables.
- Visualize the position (ressemblance) of individuals (among each other).
- Visualize how variables are correlated.
- Interpret the synthetic variables.

Applications

PCA can be used for

1. Dimension Reduction
2. Visualization
3. Feature Extraction
4. Data Compression
5. Smoothing of Data
6. Detection of Outliers
7. Preliminary process for further analyses

About PCA

The most common approaches:

PCA can be presented using various—different but equivalent—approaches. Each approach corresponds to a unique perspective and a way of thinking about data.

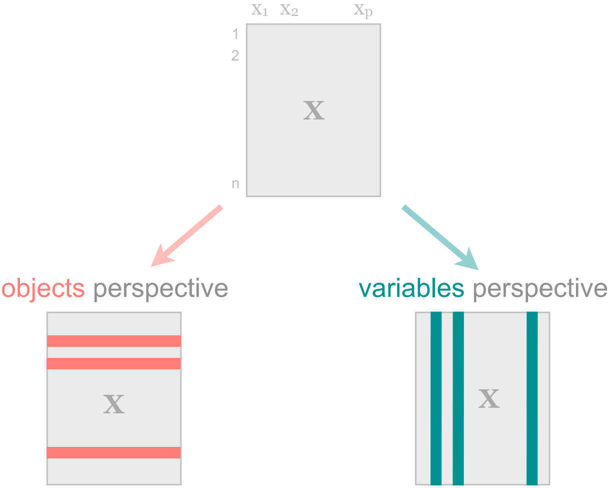
- Data in terms of variation (spread/dispersion)
- Data as points (i.e. vectors) in a multidimensional space
- Data that follows a decomposition model

I will present PCA by mixing and connecting all of these approaches.

Data Matrix Duality Recap

Data Perspectives

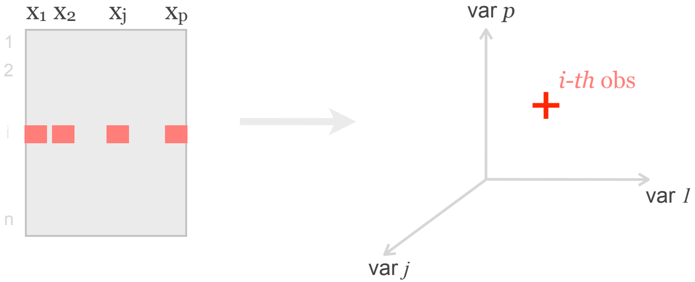
looking at a data matrix from two perspectives



Objects in Multidimensional Space

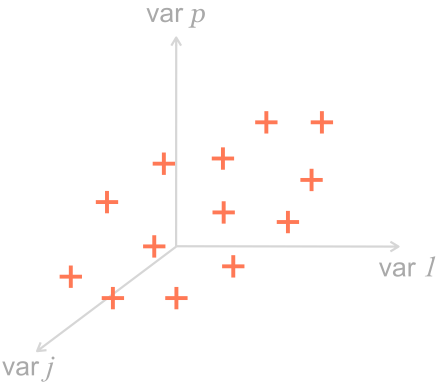
each object described
by *p* variables

Associated
p-dimensional space

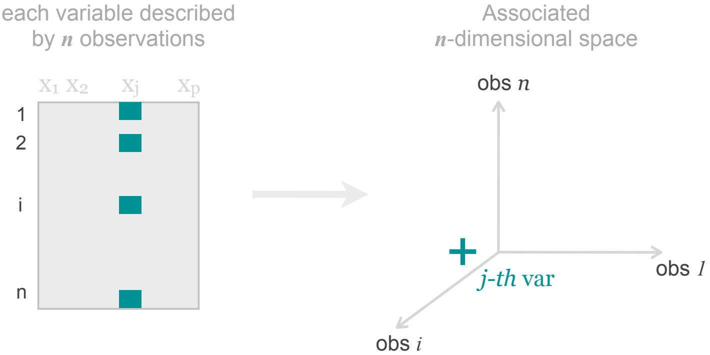


Cloud of objects

Objects as points in a p-dimensional space

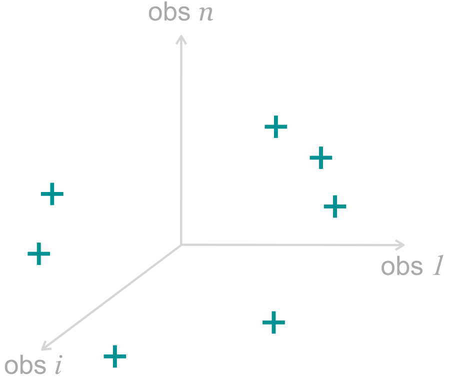


Variables in Multidimensional Space



Cloud of variables

Variables as points in a n-dimensional space



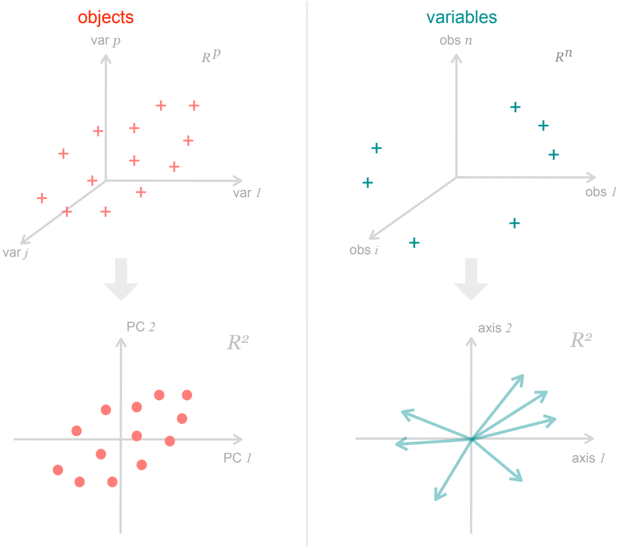
Overall Goal

PCA Visualization

One way to present PCA is based on a data visualization approach.

We look for the "best" graphical representation that allows us to visualize the data in a low dimensional space (usually 2-dimensions).

Best representation in low dimensional space

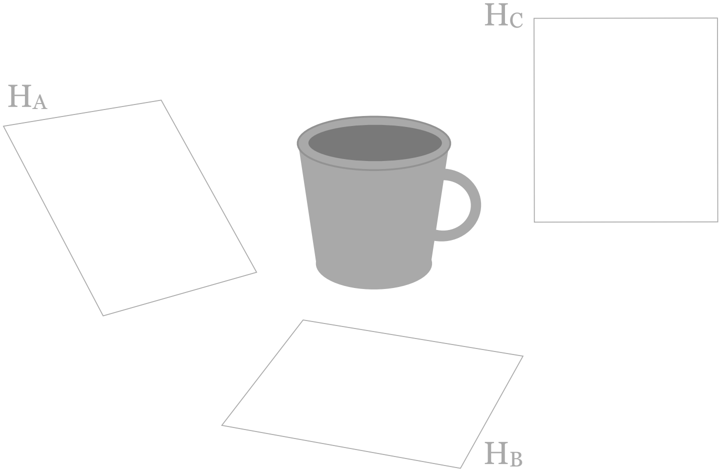


Geometric mindset

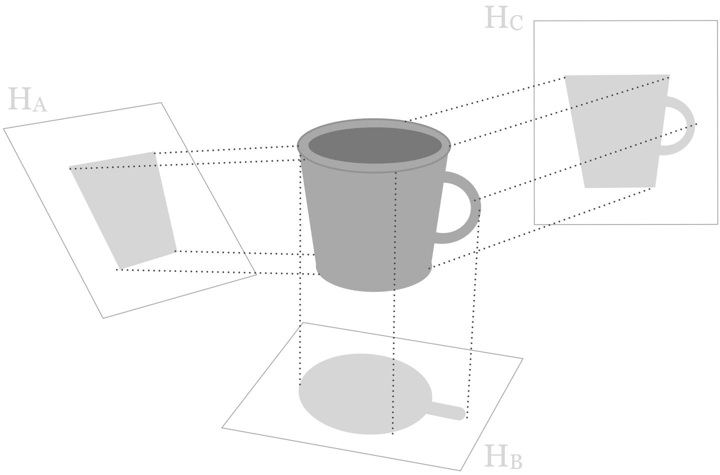
To help you understand the main idea of PCA from a geometric standpoint, I'd like to begin showing you my *mug-data* example.

Imagine we have some data in a "high-dimensional space"

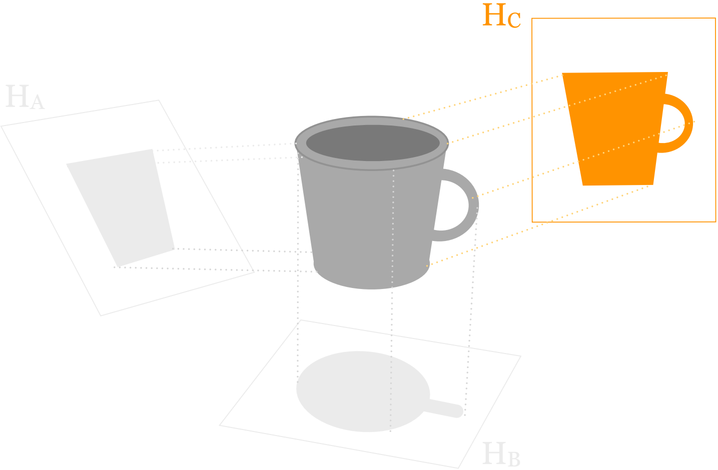
We are looking for Candidate Subspaces



with the best low-dimensional representation



Best low-dimensional projection



Projections!!!

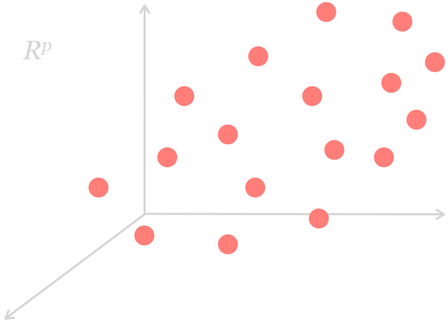
Projection

We want to find a subspace that provides us the best **projection** of the data

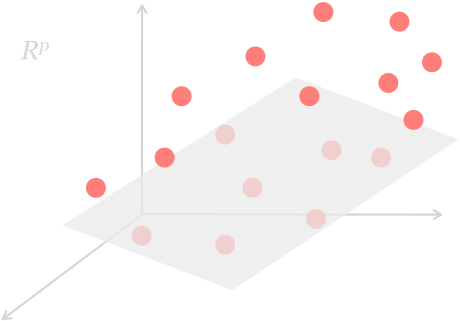
Key Message

PCA involves projecting the data onto a low-dimensional space that best captures the original dispersion in the data.

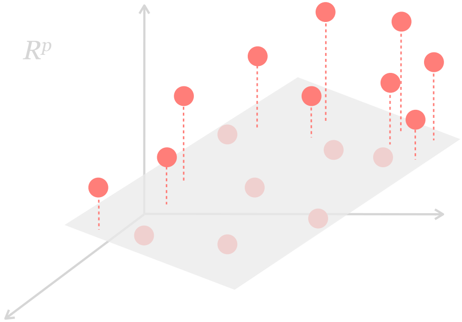
Objects in a high-dimensional space



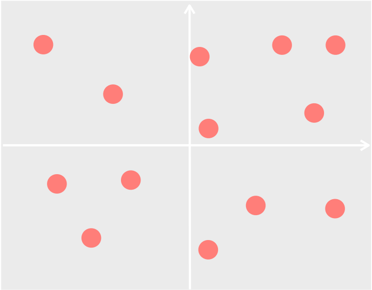
We look for a subspace such that



the projection of points on it



is the best low-dimensional representation



How do you find the associated axes?

Main Idea

In order to find the “best” low dimensional representation, we need to be able to measure the **amount of spread** (i.e. dispersion).

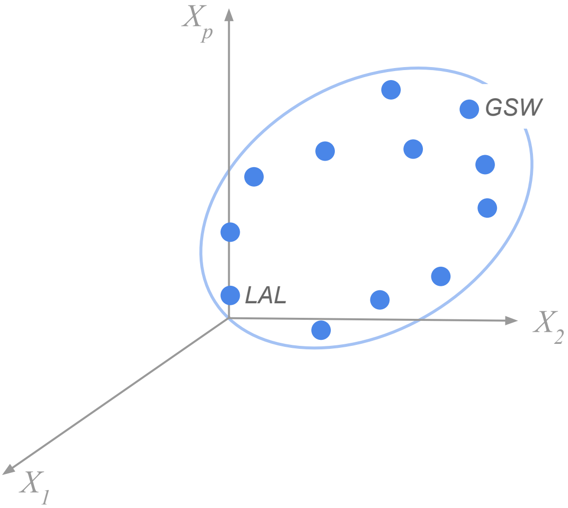
How to measure dispersion?

Inertia

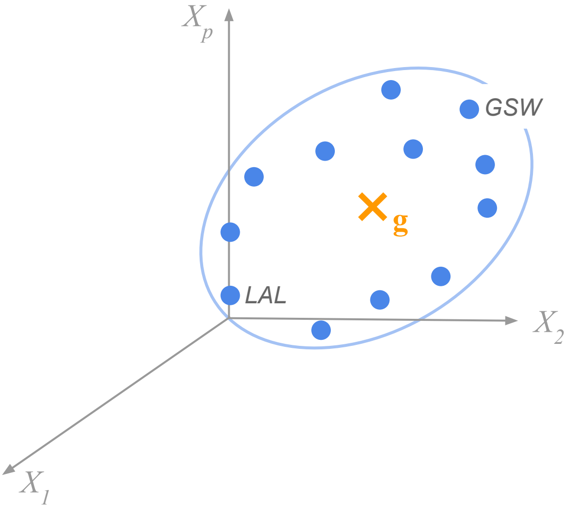
One way to take into account the dispersion of the data is with the concept of **Inertia**.

- Inertia is a term borrowed from the *moment of inertia* in mechanics.
- 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)

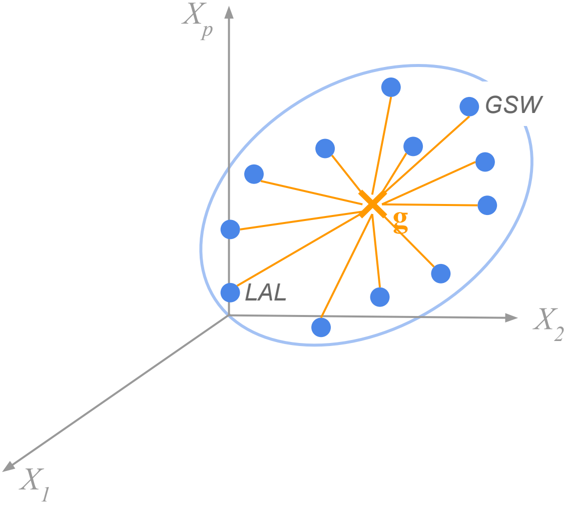


Formula of Total Inertia

The Total Inertia, I, is a weighted sum of square 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)



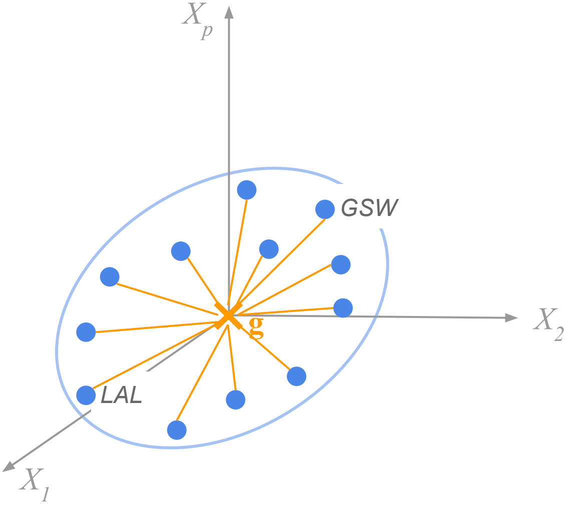
Formula of Total Inertia

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 square distances around the centroid g

Centered data: centroid is the origin

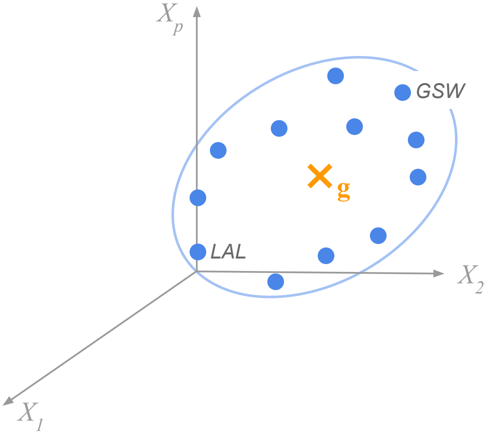


Computing Inertia

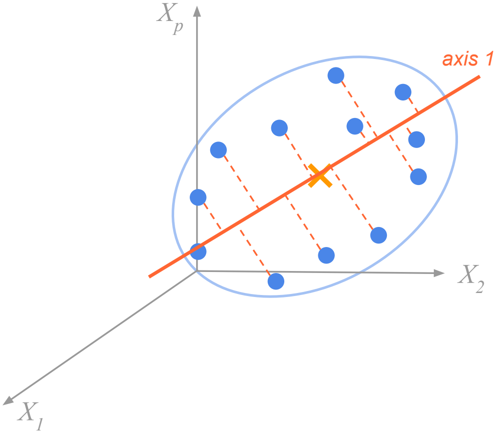
$$
\begin{aligned}
Inertia &= \sum_{i=1}^{n} m_i d^2(\mathbf{x_i}, \mathbf{g}) \\
&= \sum_{i=1}^{n} \frac{1}{n}(\mathbf{x_i} - \mathbf{g})^\mathsf{T}(\mathbf{x_i} - \mathbf{g}) \\
&= \frac{1}{n} tr(\mathbf{X}^\mathsf{T}\mathbf{X}) \\
&= \frac{1}{n} tr(\mathbf{X}\mathbf{X}^\mathsf{T})
\end{aligned}
$$

Principal Components

Looking for an axis 1



1st axis

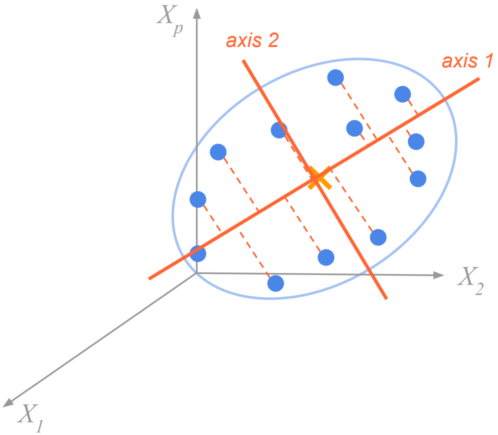


We want a 1st axis that retains most of the projected inertia

First Axis and Principal Component

- The axis Δ_1 passes through the centroid $\mathbf{g}$ (with centered data, $\mathbf{g}$ is the origin)
- The axis Δ_1 is created by the unit-norm vector $\mathbf{v_1}$, eigenvector of $\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X}$, associated to the largest eigenvalue λ_1
- The explained inertia by the axis Δ_1 is equal to λ_1
- With standardized data, the proportion of explained inertia by Δ_1 is λ_1/p

2nd axis



We want a 2nd axis, orthogonal to Δ_1, that retains most of the remaining projected inertia

Second Axis and Principal Component

- The axis Δ_2 passes through the centroid $\mathbf{g}$ (with centered data, $\mathbf{g}$ is the origin)
- The axis Δ_2 is created by the unit-norm vector $\mathbf{v_2}$, eigenvector of $\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X}$, associated to the second largest eigenvalue λ_2
- The explained inertia by the axis Δ_2 is equal to λ_2
- With standardized data, the proportion of explained inertia by Δ_2 is λ_2/p

Computational note

In practice, most software routines for PCA don't really work with the *population covariance* matrix $\frac{1}{n}\mathbf{X}^\mathsf{T}\mathbf{X}$.

Instead, most programs work with the sample covariance matrix: $\frac{1}{n-1}\mathbf{X}^\mathsf{T}\mathbf{X}$

Notice that with standardized data, $\frac{1}{n-1}\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{R}$, is the correlation matrix.

PCA of NBA Team Stats

Eigenvalues

	eigenvalue	percentage	cumulative perc
comp 1	3.6806	46.007	46.01
comp 2	1.6177	20.221	66.23
comp 3	1.0185	12.732	78.96
comp 4	0.6214	7.768	86.73
comp 5	0.4720	5.900	92.63
comp 6	0.4619	5.774	98.40
comp 7	0.1279	1.598	100.00
comp 8	0.0000	0.000	100.00

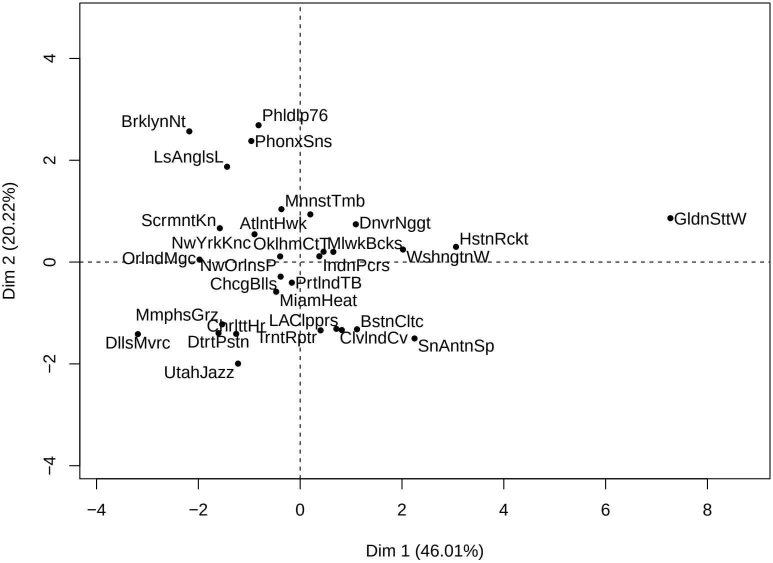
What's going on with eigenvalue of PC8?

Eigenvectors

	v1	v2	v3	v4	v5	v6	v7
wins	0.412	-0.437	0.054	-0.187	-0.138	-0.255	-0.129
losses	-0.412	0.437	-0.054	0.187	0.138	0.255	0.129
points	0.425	0.138	-0.449	0.160	-0.163	-0.048	0.738
field_goals	0.405	0.164	-0.330	0.412	-0.203	0.400	-0.573
assists	0.398	0.127	-0.030	-0.127	0.897	0.047	-0.042
turnovers	0.102	0.669	-0.049	-0.191	-0.146	-0.649	-0.246
steals	0.297	0.313	0.418	-0.544	-0.260	0.512	0.118
blocks	0.243	0.097	0.711	0.622	0.005	-0.149	0.132

Principal Components

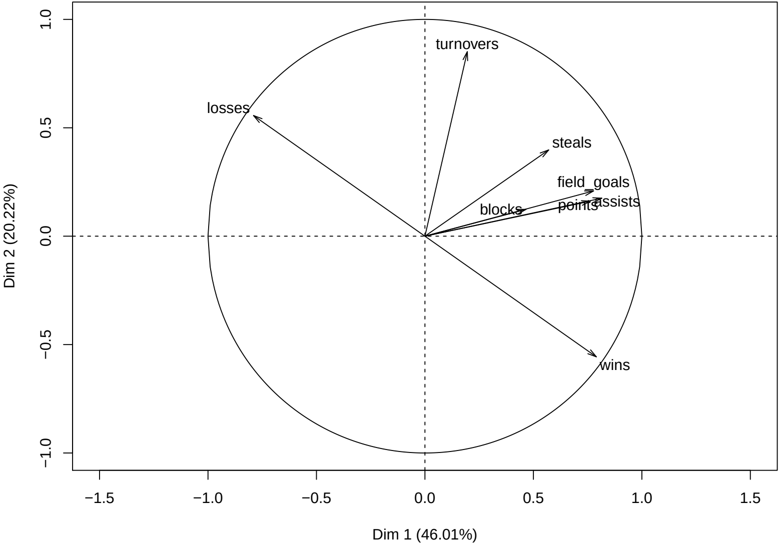
	PC1	PC2	PC3	PC4	PC5	PC6	PC7
GldnSttW	7.150	0.848	1.324	0.369	0.687	0.606	0.024
SnAntnSp	2.208	-1.475	1.521	0.186	-0.086	-0.546	-0.261
HstnRckt	3.010	0.294	-1.418	-0.842	-0.194	-0.454	0.646
BstnCltc	1.098	-1.298	-0.827	-0.875	0.869	-0.340	0.257
UtahJazz	-1.200	-1.961	0.770	0.147	-0.341	-1.686	-0.295
TrntRptr	0.394	-1.318	0.560	-0.162	-2.078	0.553	0.401
ClvlndCv	0.699	-1.290	-2.052	0.398	-0.059	-0.848	-0.018
LAClpprs	0.805	-1.313	-0.982	-0.232	-0.295	0.071	0.195
WshngtnW	1.986	0.242	-1.002	-0.802	-0.491	0.878	-0.492
OklhmCtT	0.640	0.197	0.208	-0.023	-1.104	-0.631	-0.227



BrklynNt
Phldlp76
PhonxSns
LsAnglsL
MnnstTmb
ScrmntKn
AtlntHwk
DnvrNggt
GldnSttW
HstnRckt
NwYrkKnc
OklhmCtT
MlwkBcks
OrlndMgc
NwOrlnsP
IndnPcrs
WshngtnW
ChcgBlls
PrtlndTB
MiamHeat
MmphsGrz
ChrlttHr
LAClpprs
BstnCltc
DllsMvrc
DtrtPstn
TrntRptr
ClvlndCv
SnAntnSp
UtahJazz
Dim 2 (20.22%)
Dim 1 (46.01%)
−4
−2
0
2
4
6
8

Correlations between variables and PCs

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


turnovers
losses
steals
field_goals
blocks
points
assists
wins
Dim 1 (46.01%)
Dim 2 (20.22%)
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

Principal Components?

Meaning of *Principal*

The term **Principal**, as used in PCA, has to do with the notion of **principal axis** from geometry and linear algebra

Principal Axis

A *principal axis* is a certain line in a Euclidean space associated to an ellipsoid or hyperboloid, generalizing the major and minor axes of an ellipse