A Practical Use of Principal Component Analysis

Kian Kermani
University of California, Irvine

November 2023

Table of Contents

Dimensionality Reduction: Why?

- Large data sets can be hard to analyze.
- Not all of the information is useful.

Dimensionality Reduction: Why?

- Large data sets can be hard to analyze.
- Not all of the information is useful.

Team	W	D	L	G	GA	GD
Liverpool	32	3	3	85	33	52
Manchester City	26	3	9	102	35	67
Manchester United	18	12	8	66	36	30
Chelsea	20	6	12	69	54	15
Leicester City	18	8	12	67	41	26
Tottenham Hotspur	16	11	11	61	47	14
Wolverhampton	15	14	9	51	40	11
Arsenal	14	14	10	56	48	8
Sheffield United	14	12	12	39	39	0
Burnley	15	9	14	43	50	-7

Dimensionality Reduction: Why?

- With PCA, we can simplify our data down to its most useful components.

Dimensionality Reduction: Why?

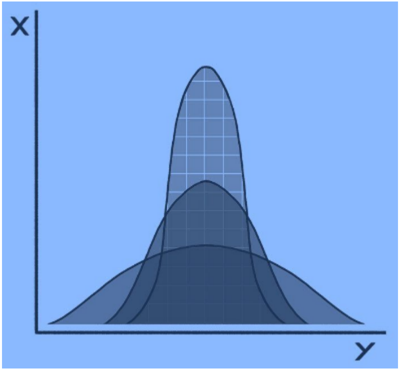
- With PCA, we can simplify our data down to its most useful components.
- Variables with high variance will be of most interest.

Dimensionality Reduction: Why?

- With PCA, we can simplify our data down to its most useful components.
- Variables with high variance will be of most interest.

Variance

The measure of the spread of data within a data set.



Covariance Matrix

We want to measure how much two variables *vary* with respect to each other (i.e. the covariance), and we want to do this across n dimensions.

Covariance Matrix

We want to measure how much two variables *vary* with respect to each other (i.e. the covariance), and we want to do this across n dimensions.

$$C = \begin{bmatrix} cov(x_1, x_1) & \dots & cov(x_1, x_n) \\ \vdots & \ddots & \vdots \\ cov(x_n, x_1) & \dots & cov(x_n, x_n) \end{bmatrix}$$

- Calculating the eigenvectors and eigenvalues of the covariance matrix gives us the **principal components** of our data set.

The Covariance Matrix

- Calculating the eigenvectors and eigenvalues of the covariance matrix gives us the **principal components** of our data set.

Let $v_1, ..., v_n$ be eigenvectors in $\mathbb{R}^n$ and $\lambda_1, ..., \lambda_n$ be the corresponding eigenvalues.

The Covariance Matrix

- Calculating the eigenvectors and eigenvalues of the covariance matrix gives us the **principal components** of our data set.

Let $v_1, ..., v_n$ be eigenvectors in $\mathbb{R}^n$ and $\lambda_1, ..., \lambda_n$ be the corresponding eigenvalues.

$$D = \text{diag}(\lambda_1, ..., \lambda_n)$$
$$V = [v_1, ..., v_n]$$

- Calculating the eigenvectors and eigenvalues of the covariance matrix gives us the **principal components** of our data set.

Let $v_1, ..., v_n$ be eigenvectors in $\mathbb{R}^n$ and $\lambda_1, ..., \lambda_n$ be the corresponding eigenvalues.

$$D = \text{diag}(\lambda_1, ..., \lambda_n)$$
$$V = [v_1, ..., v_n]$$

Note: The values are listed such that $\lambda_1 \geq ... \geq \lambda_n$

Interpreting the Principal Components

Principal Component 1 to n

The first principal component (corresponding to λ_1) explains the *highest* proportion of the variability in the data set. The nth component accounts for the *lowest* proportion of variability.

Principal Component 1 to n

The first principal component (corresponding to λ_1) explains the *highest* proportion of the variability in the data set. The nth component accounts for the *lowest* proportion of variability.

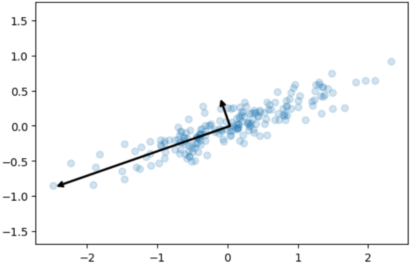


Figure: Eigenvectors (Principal Components) point in direction of highest variance

Interpreting the Principal Components

Principal Component 1 to n

The first principal component (corresponding to λ_1) explains the *highest* proportion of the variability in the data set. The nth component accounts for the *lowest* proportion of variability.

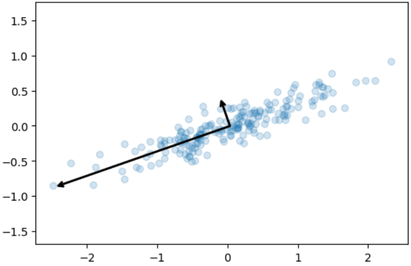


Figure: Eigenvectors (Principal Components) point in direction of highest variance

We can choose to keep some p (with $p \leq n$) number of PC's to reduce our data's dimension!

A Quick Demonstration

Team	W	D	L	G	GA	GD
Liverpool	32	3	3	85	33	52
Manchester City	26	3	9	102	35	67
Manchester United	18	12	8	66	36	30
Chelsea	20	6	12	69	54	15
Leicester City	18	8	12	67	41	26
Tottenham Hotspur	16	11	11	61	47	14
Wolverhampton	15	14	9	51	40	11
Arsenal	14	14	10	56	48	8
Sheffield United	14	12	12	39	39	0
Burnley	15	9	14	43	50	-7

Figure: 2019-2020 Premier League data

A Quick Demonstration

- After creating the covariance matrix, we compute the eigenvalues.

- After creating the covariance matrix, we compute the eigenvalues.

$$D = \text{diag}(1300,\ 71.9,\ 8.05,\ 4.62,\ -2.65e-14,\ -3.73e-14)$$

- Two of our eigenvalues come out very close to zero!

- We can see the proportion of variability explained by each eigenvalue if we take

$$P_i = \frac{\lambda_i}{\lambda_1 + \lambda_2 + ... + \lambda_n}$$

A Quick Demonstration

- We can see the proportion of variability explained by each eigenvalue if we take

$$P_i = \frac{\lambda_i}{\lambda_1 + \lambda_2 + ... + \lambda_n}$$

$$D = \mathsf{diag}(1300,\ 71.9,\ 8.05,\ 4.62,\ -2.65e-14,\ -3.73e-14)$$

A Quick Demonstration

- We can see the proportion of variability explained by each eigenvalue if we take

$$P_i = \frac{\lambda_i}{\lambda_1 + \lambda_2 + ... + \lambda_n}$$

$$D = \mathsf{diag}(1300,\ 71.9,\ 8.05,\ 4.62,\ -2.65e-14,\ -3.73e-14)$$

$$P = (0.939,\ 0.052,\ 0.00583,\ 0.00334,\ -192e-17,\ -2.7e-17)$$

A Quick Demonstration

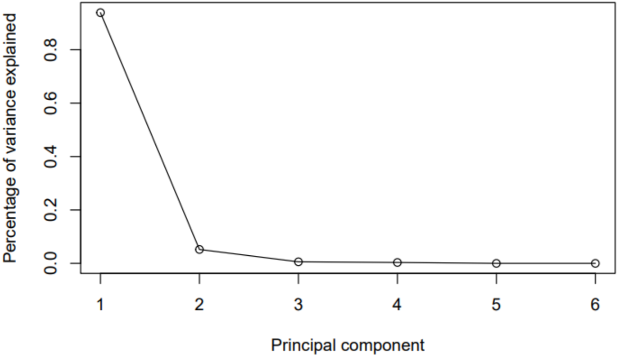


Figure: Scree Plot of our PCs

A Quick Demonstration

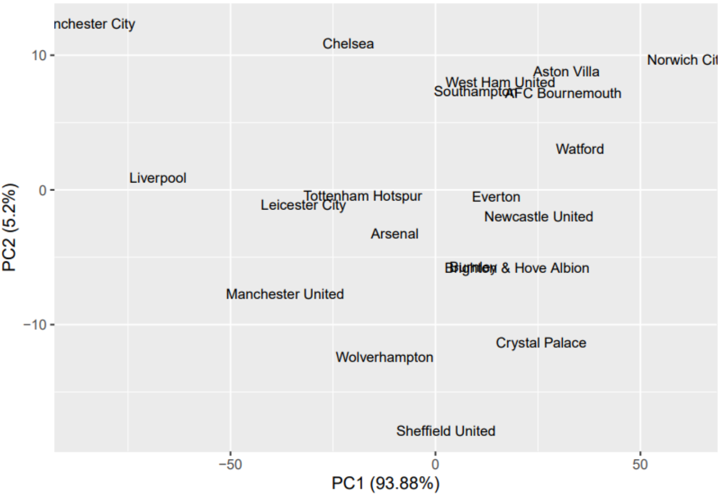


Figure: Visualization of differences between teams

Thank you for listening!