

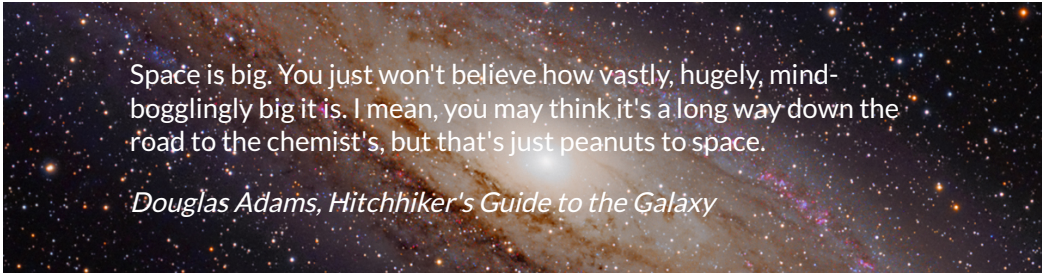
# ETC3250: Dimension reduction

Semester 1, 2019

Professor Di Cook

Econometrics and Business Statistics  
Monash University

Week 4 (a)



Space is big. You just won't believe how vastly, hugely, mind-bogglingly big it is. I mean, you may think it's a long way down the road to the chemist's, but that's just peanuts to space.

*Douglas Adams, Hitchhiker's Guide to the Galaxy*

## Outline

- High dimensions
- Definition

Remember, our data can be denoted as:

$$\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N, \text{ where } x_i = (x_{i1}, \dots, x_{ip})^T$$

then

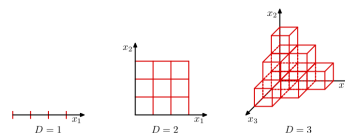
Dimension of the data is  $p$ , the number of variables.

3 / 27

## Outline

- High dimensions
- Definition
- Cubes and spheres

Space expands exponentially with dimension:



As dimension increases the volume of a sphere of same radius as cube side length becomes much smaller than the volume of the cube:



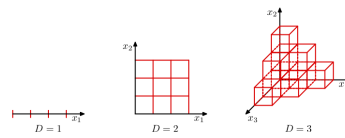
4 / 27

## Outline

### High dimensions

- Definition
- Cubes and spheres

Space expands exponentially with dimension:



As dimension increases the volume of a sphere of same radius as cube side length becomes much smaller than the volume of the cube:



4 / 27

## Outline

### High dimensions

- Definition
- Cubes and spheres
- Sub-spaces

Data will often be confined to a region of the space having lower intrinsic dimensionality. The data lives in a low-dimensional subspace.

SO, reduce dimensionality, to the subspace containing the data.

5 / 27

## Outline

### High dimensions

- Definition
- Cubes and spheres
- Sub-spaces

Data will often be confined to a region of the space having lower **intrinsic dimensionality**. The data lives in a low-dimensional subspace.

SO, **reduce dimensionality**, to the subspace containing the data.

5 / 27

## Outline

### High dimensions

#### PCA

- Definition

Principal component analysis (PCA) produces a low-dimensional representation of a dataset. It finds a sequence of linear combinations of the variables that have **maximal variance**, and are **mutually uncorrelated**. It is an unsupervised learning method.

Why?

- We may have too many predictors for a regression. Instead, we can use the first few principal components.
- Understanding relationships between variables.
- Data visualization. We can plot a small number of variables more easily than a large number of variables.

6 / 27

## Outline

- High dimensions
- PCA
- Definition

Principal component analysis (PCA) produces a low-dimensional representation of a dataset. It finds a sequence of linear combinations of the variables that have **maximal variance**, and are **mutually uncorrelated**. It is an unsupervised learning method.

Why?

- We may have too many predictors for a regression. Instead, we can use the first few principal components.
- Understanding relationships between variables.
- Data visualization. We can plot a small number of variables more easily than a large number of variables.

6 / 27

## Outline

- High dimensions
- PCA
- Definition

### First principal component

The first principal component of a set of variables  $x_1, x_2, \dots, x_p$  is the linear combination

$$z_1 = \phi_{11}x_1 + \phi_{21}x_2 + \dots + \phi_{p1}x_p$$

that has the largest variance such that

$$\sum_{j=1}^p \phi_{j1}^2 = 1$$

The elements  $\phi_{11}, \dots, \phi_{p1}$  are the **loadings** of the first principal component.

7 / 27

## Outline

- High dimensions
- PCA
- Definition

### First principal component

The first principal component of a set of variables  $x_1, x_2, \dots, x_p$  is the linear combination

$$z_1 = \phi_{11}x_1 + \phi_{21}x_2 + \dots + \phi_{p1}x_p$$

that has the largest variance such that

$$\sum_{j=1}^p \phi_{j1}^2 = 1$$

The elements  $\phi_{11}, \dots, \phi_{p1}$  are the **loadings** of the first principal component.

7 / 27

## Outline

- High dimensions
- PCA
- Definition
- Geometry

The loading vector  $\phi_1 = [\phi_{11}, \dots, \phi_{p1}]'$  defines direction in feature space along which data vary most.

If we project the  $n$  data points  $x_1, \dots, x_n$  onto this direction, the projected values are the principal component scores

$$z_{11}, \dots, z_{n1}$$

The second principal component is the linear combination  $z_{12} = \phi_{12}x_{11} + \phi_{22}x_{12} + \dots + \phi_{p2}x_{1p}$  that has maximal variance among all linear combinations that are *uncorrelated* with  $z_{11}$ .

Equivalent to constraining  $\phi_2$  to be orthogonal (perpendicular) to  $\phi_1$ . And so on.

There are at most  $\min(n-1, p)$  PCs.

8 / 27

## Outline

- High dimensions
- PCA
  - Definition
  - Geometry

The loading vector  $\phi_1 = [\phi_{11}, \dots, \phi_{p1}]'$  defines direction in feature space along which data vary most.

If we project the  $n$  data points  $x_1, \dots, x_n$  onto this direction, the projected values are the principal component scores

$$z_{11}, \dots, z_{n1}$$

The second principal component is the linear combination  $z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \dots + \phi_{p2}x_{ip}$  that has maximal variance among all linear combinations that are *uncorrelated* with  $z_{i1}$ .

Equivalent to constraining  $\phi_2$  to be orthogonal (perpendicular) to  $\phi_1$ . And so on.

There are at most  $\min(n-1, p)$  PCs.

8 / 27

## Outline

- High dimensions
- PCA
  - Definition
  - Geometry

The loading vector  $\phi_1 = [\phi_{11}, \dots, \phi_{p1}]'$  defines direction in feature space along which data vary most.

If we project the  $n$  data points  $x_1, \dots, x_n$  onto this direction, the projected values are the principal component scores

$$z_{11}, \dots, z_{n1}$$

The second principal component is the linear combination  $z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \dots + \phi_{p2}x_{ip}$  that has maximal variance among all linear combinations that are *uncorrelated* with  $z_{i1}$ .

Equivalent to constraining  $\phi_2$  to be orthogonal (perpendicular) to  $\phi_1$ . And so on.

There are at most  $\min(n-1, p)$  PCs.

8 / 27

## Outline

- High dimensions
- PCA
  - Definition
  - Geometry

The loading vector  $\phi_1 = [\phi_{11}, \dots, \phi_{p1}]'$  defines direction in feature space along which data vary most.

If we project the  $n$  data points  $x_1, \dots, x_n$  onto this direction, the projected values are the principal component scores

$$z_{11}, \dots, z_{n1}$$

The second principal component is the linear combination  $z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \dots + \phi_{p2}x_{ip}$  that has maximal variance among all linear combinations that are *uncorrelated* with  $z_{i1}$ .

Equivalent to constraining  $\phi_2$  to be orthogonal (perpendicular) to  $\phi_1$ . And so on.

There are at most  $\min(n-1, p)$  PCs.

8 / 27

## Outline

- High dimensions
- PCA
  - Definition
  - Geometry

The loading vector  $\phi_1 = [\phi_{11}, \dots, \phi_{p1}]'$  defines direction in feature space along which data vary most.

If we project the  $n$  data points  $x_1, \dots, x_n$  onto this direction, the projected values are the principal component scores

$$z_{11}, \dots, z_{n1}$$

The second principal component is the linear combination  $z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \dots + \phi_{p2}x_{ip}$  that has maximal variance among all linear combinations that are *uncorrelated* with  $z_{i1}$ .

Equivalent to constraining  $\phi_2$  to be orthogonal (perpendicular) to  $\phi_1$ . And so on.

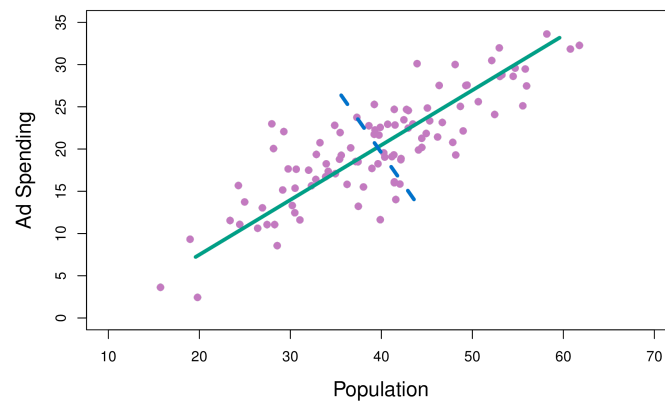
There are at most  $\min(n-1, p)$  PCs.

8 / 27



## Outline

- High dimensions
- PCA
  - Definition
  - Geometry
  - Example

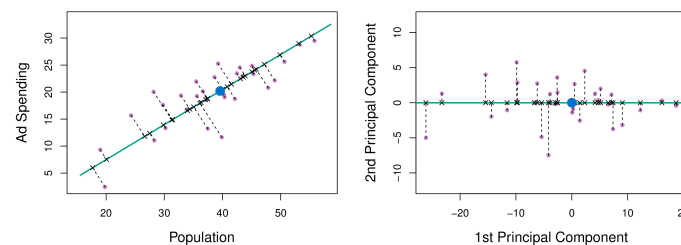


First PC; second PC

9 / 27

## Outline

- High dimensions
- PCA
  - Definition
  - Geometry
  - Example



If you think of the first few PCs like a linear model fit, and the others as the error, it is like regression, except that errors are orthogonal to model.

(Chapter6/6.15.pdf)

10 / 27

## Outline

- High dimensions
- PCA
  - Definition
  - Geometry
  - Example
  - Computation

PCA can be thought of as fitting an  $n$ -dimensional ellipsoid to the data, where each axis of the ellipsoid represents a principal component. The new variables produced by principal components correspond to **rotating** and **scaling** the ellipse **into a circle**.

11 / 27

## Outline

- High dimensions
- PCA
  - Definition
  - Geometry
  - Example
  - Computation

Suppose we have a  $n \times p$  data set  $X = [x_{ij}]$ .

Centre each of the variables to have mean zero (i.e., the column means of  $X$  are zero).

$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \dots + \phi_{p1}x_{ip}$

Sample variance of  $z_{i1}$  is  $\frac{1}{n} \sum_{i=1}^n z_{i1}^2$ .

$$\underset{\phi_{11}, \dots, \phi_{p1}}{\text{maximize}} \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^p \phi_{j1} x_{ij} \right)^2 \text{ subject to } \sum_{j=1}^p \phi_{j1}^2 = 1$$

12 / 27

## Outline

High dimensions

PCA

- Definition
- Geometry
- Example
- Computation

1. Compute the covariance matrix (after scaling the columns of  $X$ )

$$C = X'X$$

2. Find eigenvalues and eigenvectors:

$$C = VDV'$$

where columns of  $V$  are orthonormal (i.e.,  $V'V = I$ )

3. Compute PCs:  $\Phi = V$ ,  $Z = X\Phi$ .

13 / 27

## Outline

High dimensions

PCA

- Definition
- Geometry
- Example
- Computation

### Singular Value Decomposition

$$X = U\Lambda V'$$

$X$  is an  $n \times p$  matrix

$U$  is  $n \times r$  matrix with orthonormal columns ( $U'U = I$ )

$\Lambda$  is  $r \times r$  diagonal matrix with non-negative elements.

$V$  is  $p \times r$  matrix with orthonormal columns ( $V'V = I$ ).

It is always possible to uniquely decompose a matrix in this way.

14 / 27

## Outline

High dimensions

PCA

Definition

Geometry

Example

Computation

1. Compute SVD:  $X = U\Lambda V'$ .

2. Compute PCs:  $\Phi = V$ .  $Z = X\Phi$ .

Relationship with covariance:

$$C = X'X = V\Lambda U'U\Lambda V' = V\Lambda^2 V' = VDV'$$

Eigenvalues of  $C$  are squares of singular values of  $X$ .

Eigenvectors of  $C$  are right singular vectors of  $X$ .

The PC directions  $\phi_1, \phi_2, \phi_3, \dots$  are the right singular vectors of the matrix  $X$ .

15 / 27

## Outline

High dimensions

PCA

Definition

Geometry

Example

Computation

Total variance

Total variance in data (assuming variables centered at 0):

$$TV = \sum_{j=1}^p \text{Var}(x_j) = \sum_{j=1}^p \frac{1}{n} \sum_{i=1}^n x_{ij}^2$$

If variables are standardised, TV=number of variables!

Variance explained by  $m$ 'th PC:

$$V_m = \text{Var}(z_m) = \frac{1}{n} \sum_{i=1}^n z_{im}^2$$

$$TV = \sum_{m=1}^M V_m \text{ where } M = \min(n-1, p).$$

16 / 27

## Outline

High dimensions

PCA

- Definition
- Geometry
- Example
- Computation
- Total variance

Total variance in data (assuming variables centered at 0):

$$TV = \sum_{j=1}^p \text{Var}(x_j) = \sum_{j=1}^p \frac{1}{n} \sum_{i=1}^n x_{ij}^2$$

If variables are standardised, TV=number of variables!

Variance explained by  $m$ 'th PC:

$$V_m = \text{Var}(z_m) = \frac{1}{n} \sum_{i=1}^n z_{im}^2$$

$$TV = \sum_{m=1}^M V_m \text{ where } M = \min(n-1, p).$$

16 / 27

## Outline

High dimensions

PCA

- Definition
- Geometry
- Example
- Computation
- Total variance

Total variance in data (assuming variables centered at 0):

$$TV = \sum_{j=1}^p \text{Var}(x_j) = \sum_{j=1}^p \frac{1}{n} \sum_{i=1}^n x_{ij}^2$$

If variables are standardised, TV=number of variables!

Variance explained by  $m$ 'th PC:

$$V_m = \text{Var}(z_m) = \frac{1}{n} \sum_{i=1}^n z_{im}^2$$

$$TV = \sum_{m=1}^M V_m \text{ where } M = \min(n-1, p).$$

16 / 27

## Outline

High dimensions

PCA

- Definition
- Geometry
- Example
- Computation
- Total variance
- Choosing  $k$

### Proportion of variance explained:

$$\text{PVE}_m = \frac{V_m}{TV}$$

Choosing the number of PCs that adequately summarises the variation in  $X$ , is achieved by examining the cumulative proportion of variance explained.

Cumulative proportion of variance explained:

$$\text{CPVE}_k = \sum_{m=1}^k \frac{V_m}{TV}$$

and also by a scree plot.

17 / 27

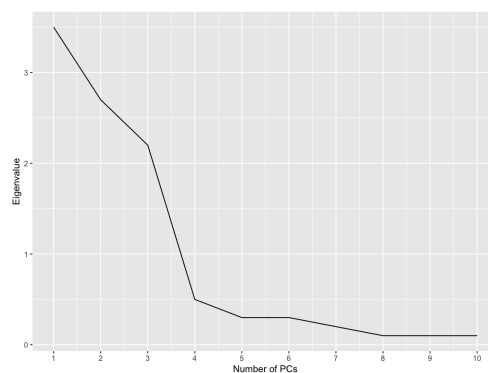
## Outline

High dimensions

PCA

- Definition
- Geometry
- Example
- Computation
- Total variance
- Choosing  $k$

**Scree plot:** Plot of variance explained by each component vs number of component.

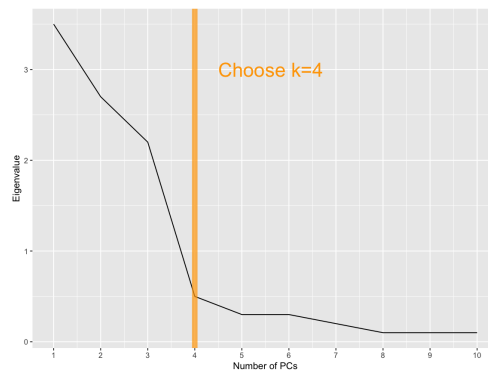


18 / 27

## Outline

- High dimensions
- PCA
  - Definition
  - Geometry
  - Example
  - Computation
  - Total variance
  - Choosing  $k$

**Scree plot:** Plot of variance explained by each component vs number of component.



18 / 27

## Outline

- High dimensions
- PCA
- Example
  - Data

The data on national track records for women (as at 1984).

```
## Observations: 55
## Variables: 8
## $ m100    <dbl> 11.61, 11.20, 11.43, 11.41, 11.46, 11.31, 12.14, 1.
## $ m200    <dbl> 22.94, 22.35, 23.09, 23.04, 23.05, 23.17, 24.47, 2.
## $ m400    <dbl> 54.50, 51.08, 50.62, 52.00, 53.30, 52.80, 55.00, 5.
## $ m800    <dbl> 2.15, 1.98, 1.99, 2.00, 2.16, 2.10, 2.18, 2.00, 2.
## $ m1500   <dbl> 4.43, 4.13, 4.22, 4.14, 4.58, 4.49, 4.45, 4.06, 4.
## $ m3000   <dbl> 9.79, 9.08, 9.34, 8.88, 9.81, 9.77, 9.51, 8.81, 9.
## $ marathon <dbl> 178.52, 152.37, 159.37, 157.85, 169.98, 168.75, 19.
## $ country <chr> "argentin", "australi", "austria", "belgium", "berl"
```

Source: Johnson and Wichern, Applied multivariate analysis

19 / 27

## Outline

- High dimensions
- PCA
- Example
- Data
- Explore

m100 m200 m400 m800 m1500 m3000 marathon 20 / 27

## Outline

- High dimensions
- PCA
- Example
- Data
- Explore
- Compute

```
track_pca <- prcomp(track[,1:7], center=TRUE, scale=TRUE)
track_pca
```

```
## Standard deviations (1, .., p=7):
## [1] 2.41 0.81 0.55 0.35 0.23 0.20 0.15
##
## Rotation (n x k) = (7 x 7):
##      PC1  PC2  PC3  PC4  PC5  PC6  PC7
## m100  0.37  0.49 -0.286 0.319 0.231 0.6198 0.052
## m200  0.37  0.54 -0.230 -0.083 0.041 -0.7108 -0.109
## m400  0.38  0.25  0.515 -0.347 -0.572 0.1909 0.208
## m800  0.38 -0.16  0.585 -0.042 0.620 -0.0191 -0.315
## m1500 0.39 -0.36  0.013 0.430 0.030 -0.2312 0.693
## m3000 0.39 -0.35 -0.153 0.363 -0.463 0.0093 -0.598
## marathon 0.37 -0.37 -0.484 -0.672 0.131 0.1423 0.070
```

21 / 27



## Outline

- High dimensions
- PCA
- Example
  - Data
  - Explore
  - Compute
  - Assess

Summary of the principal components:

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Variance	5.81	0.65	0.30	0.13	0.05	0.04	0.02
Proportion	0.83	0.09	0.04	0.02	0.01	0.01	0.00
Cum. prop	0.83	0.92	0.97	0.98	0.99	1.00	1.00

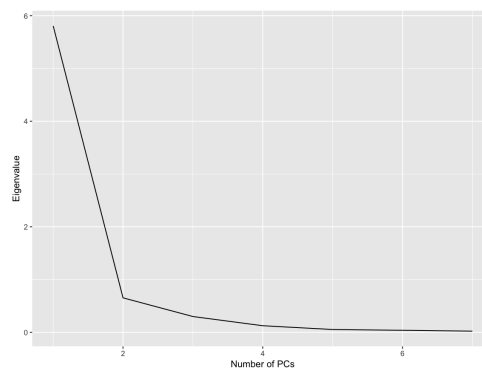
Increase in variance explained large until  $k = 3$  PCs, and then tapers off. A choice of 3 PCs would explain 97% of the total variance.

22 / 27

## Outline

- High dimensions
- PCA
- Example
  - Data
  - Explore
  - Compute
  - Assess

Scree plot: Where is the elbow?



At  $k = 2$ , thus the scree plot suggests 2 PCs would be

23 / 27

## Outline

- High dimensions
- PCA
- Example
  - Data
  - Explore
  - Compute
  - Assess

**Visualise model using a biplot:** Plot the principal component scores, and also the contribution of the original variables to the principal component.



24 / 27

## Outline

- High dimensions
- PCA
- Example
  - Data
  - Explore
  - Compute
  - Assess
  - Interpret

**Explain and interpret:** using the coefficients of the principal components.

PC1 measures overall magnitude, the strength of the athletics program. High positive values indicate **poor** programs with generally slow times across events.

PC2 measures the **contrast** in the program between **short and long distance** events. Some countries have relatively stronger long distance athletes, while others have relatively stronger short distance athletes.

There are several **outliers** visible in this plot, **wsamoa**, **cookis**, **dpkorea**. PCA, because it is computed using the

variance in the data, can be affected by outliers. It may be

25 / 27

## Outline

- High dimensions
- PCA
- Example
  - Data
  - Explore
  - Compute
  - Assess
  - Interpret

**Explain and interpret:** using the coefficients of the principal components.

PC1 measures overall magnitude, the strength of the athletics program. High positive values indicate **poor** programs with generally slow times across events.

PC2 measures the **contrast** in the program between **short and long distance** events. Some countries have relatively stronger long distance athletes, while others have relatively stronger short distance athletes.

There are several **outliers** visible in this plot, **wsamoa**, **cookis**, **dpkorea**. PCA, because it is computed using the

variance in the data, can be affected by outliers. It may be

25 / 27

## Outline

- High dimensions
- PCA
- Example
  - Data
  - Explore
  - Compute
  - Assess
  - Interpret

**Explain and interpret:** using the coefficients of the principal components.

PC1 measures overall magnitude, the strength of the athletics program. High positive values indicate **poor** programs with generally slow times across events.

PC2 measures the **contrast** in the program between **short and long distance** events. Some countries have relatively stronger long distance athletes, while others have relatively stronger short distance athletes.

There are several **outliers** visible in this plot, **wsamoa**, **cookis**, **dpkorea**. PCA, because it is computed using the

variance in the data, can be affected by outliers. It may be

25 / 27

## Outline

High dimensions

PCA

Example

• Data

• Explore

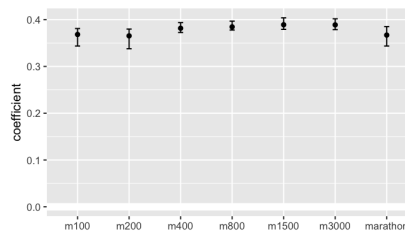
• Compute

• Assess

• Interpret

• Accuracy

Bootstrap can be used to assess whether the coefficients of a PC are significantly different from 0. The 95% bootstrap confidence intervals are:



All of the coefficients on PC1 are significantly different from 0, and positive, approximately equal, **not significantly different from each other**.

26 / 27

 Made by a human with a computer

Slides at <https://monba.dicook.org>.

Code and data at

[https://github.com/dicook/Business\\_Analytics](https://github.com/dicook/Business_Analytics).

Created using R Markdown with flair by **xaringan**, and **kunoichi** (female ninja) style.



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

27 / 27

