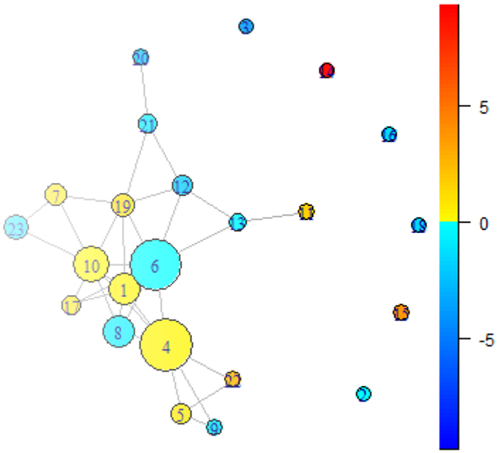
Regional Analysis with Topological Data Analysis Ball Mapper

Session 1: A Methodological Motivation and Introduction

Dr Simon Rudkin

University of Manchester

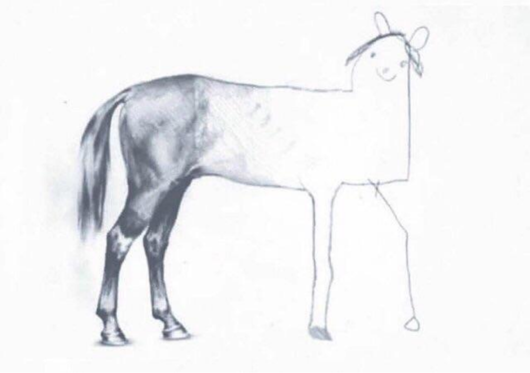


In this Session...

- Importance of Topology
- Artificial Data
- Scatterplots
- TDA Ball Mapper algorithms

This session serves as an introduction to Toplogical Data Analysis Ball Mapper (TDABM) as based upon the original working paper of Dłotko (2019).

Importance of Topology



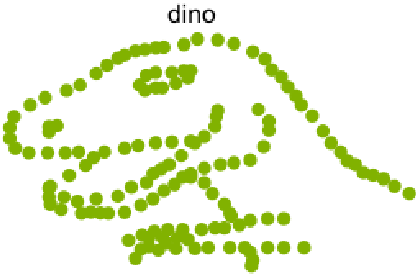
- "Unfinished Horse" reminds the mind can fill in many details from pictures
- Code the pixels according to their colour and machine can see too
- Being able to visualise and evaluate is critical

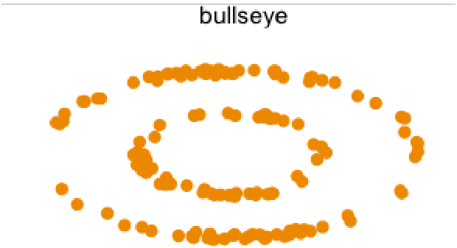
Importance of Topology 2

Sketch a scatterplot of two variables X and Y with the following information:

- The mean of X is 54 and mean of Y is 47.
- The standard deviation of X is 17 and the standard deviation of Y is 27.
- The correlation between X and Y is -0.06

Importance of Topology 3

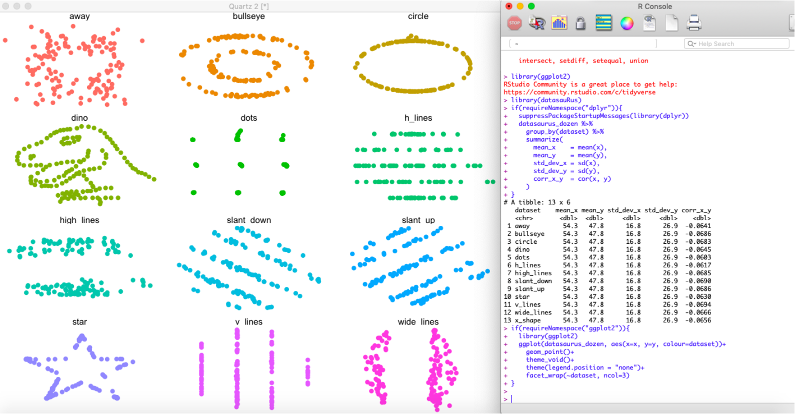




- Both datasets have correlation between horizontal and vertical of -0.06
- Neither are the plot you would expect
- Examples are from Matejka and Fitzmaurice (2017)

Importance of Topology 4

- See Matejka and Fitzmaurice (2017)
- Our eyes tell us these are not the same dataset even though the summary statistics are identical



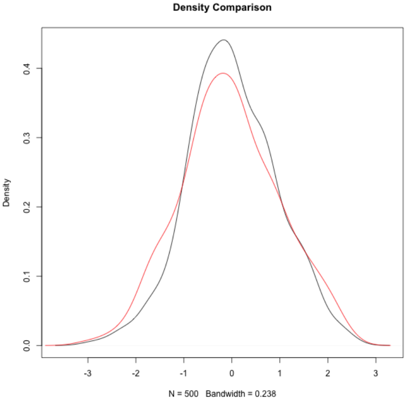
```
    intersect, setdiff, setequal, union

> library(ggplot2)
RStudio Community is a great place to get help:
https://community.rstudio.com/c/tidyverse
> library(datasauRus)
> if(requireNamespace("dplyr")){
+   suppressPackageStartupMessages(library(dplyr))
+   datasaurus_dozen %>%
+     group_by(dataset) %>%
+     summarize(
+       mean_x    = mean(x),
+       mean_y    = mean(y),
+       std_dev_x = sd(x),
+       std_dev_y = sd(y),
+       corr_x_y  = cor(x, y)
+     )
+ }
# A tibble: 13 x 6
   dataset    mean_x mean_y std_dev_x std_dev_y corr_x_y
   <chr>       <dbl>  <dbl>     <dbl>     <dbl>    <dbl>
 1 away         54.3   47.8      16.8      26.9  -0.0641
 2 bullseye     54.3   47.8      16.8      26.9  -0.0686
 3 circle       54.3   47.8      16.8      26.9  -0.0683
 4 dino         54.3   47.8      16.8      26.9  -0.0645
 5 dots         54.3   47.8      16.8      26.9  -0.0603
 6 h_lines      54.3   47.8      16.8      26.9  -0.0617
 7 high_lines   54.3   47.8      16.8      26.9  -0.0685
 8 slant_down   54.3   47.8      16.8      26.9  -0.0690
 9 slant_up     54.3   47.8      16.8      26.9  -0.0686
10 star         54.3   47.8      16.8      26.9  -0.0630
11 v_lines      54.3   47.8      16.8      26.9  -0.0694
12 wide_lines   54.3   47.8      16.8      26.9  -0.0666
13 x_shape      54.3   47.8      16.8      26.9  -0.0656
> if(requireNamespace("ggplot2")){
+   library(ggplot2)
+   ggplot(datasaurus_dozen, aes(x=x, y=y, colour=dataset))+
+     geom_point()+
+     theme_void()+
+     theme(legend.position = "none")+
+     facet_wrap(~dataset, ncol=3)
+ }
>
> |
```

Hidden in Plain Sight?

“So there are still many unknowns. And I think, because they’re so huge - so obvious - people haven’t really thought to study them in that much detail.” (Dr Alex Monro, from BBC Article)

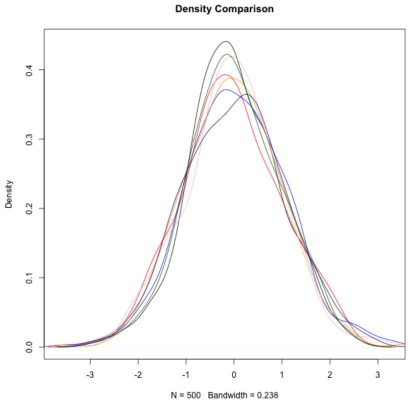
BBC article as at 4th July 2022

Artificial Variables



- Create variables to demonstrate properties
- Here draw from standard normal distribtuion of mean 0 variance 1
- For example a normally distributed variable X_1 and X_2 with 500 points
- Each random draw is slightly different
- Important to control for variation in other parameters

Artificial Variables 2



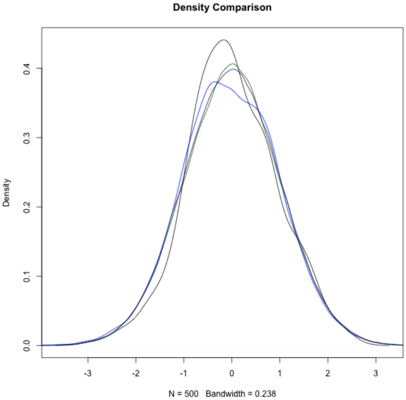
- Create variables to demonstrate properties
- For example a normally distributed variable X_1 with 500 points
- Each random draw is slightly different
- Here show 5 other sets of 500 points from $N \sim (0, 1)$

Artificial Variables 3

Variable	Mean	s.d	Min	Q25	Q75	Max
X_1	0.005	0.928	-2.968	-0.598	0.630	2.575
X_2	0.049	1.046	-3.466	-0.656	0.738	3.378
X_3	-0.023	0.995	-2.658	-0.639	0.610	3.297
X_4	-0.021	0.992	-3.532	-0.740	0.652	3.526
X_5	0.029	1.043	-2.860	-0.680	0.707	2.952
X_6	0.006	0.958	-2.987	-0.648	0.641	2.627

- Notable variation in estimates for the mean (true value 0) and standard deviation (true value 1) - All samples with 500 points
- Impact in tails is much larger

Artificial Variables 4



- Create variables to demonstrate properties
- For example a normally distributed variable X_1 with 500 points
- Each random draw is slightly different
- Increase number of points to 5000, 50000, 500000...
- Convergence to underlying normal distribution shape

Artificial Variables 5

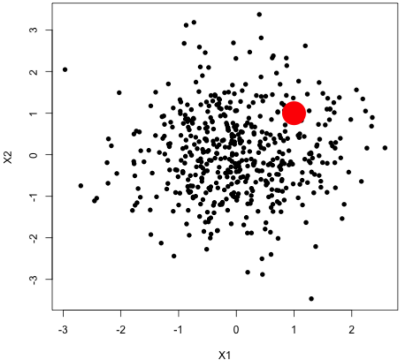
Points	Mean	s.d	Min	Q25	Q75	Max
500	0.005	0.928	-2.968	-0.598	0.630	2.575
5000	-0.031	1.000	-3.638	-0.710	0.657	3.292
50000	0.004	0.999	-4.472	-0.669	0.674	4.178
500000	-0.000	1.001	-5.051	0.677	0.678	4.485
5000000	0.000	1.000	-4.841	-0.674	0.675	5.162

- Increased numbers create closer value for mean and standard deviation
- Towards the law of large numbers

Summary of Part 1

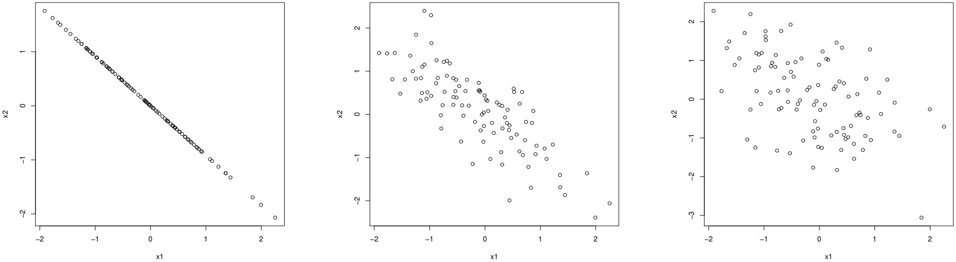
- Summary statistics do not tell the full story
- Important results are often hidden because we do not look
- All analyses are subject to a law of large numbers - we want more data
- Next we consider basic visualisations...

Scatterplots



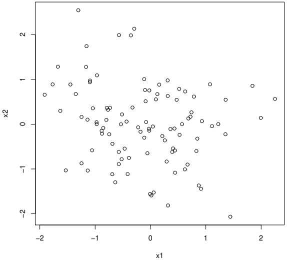
- Each point defined by value on horizontal and vertical axis
- Large red point here has $X_1 = 1$, $X_2 = 1$
- Other points are X_1 and X_2 from artifical set
- Total of 501 points

Scatterplots: Correlated Variables

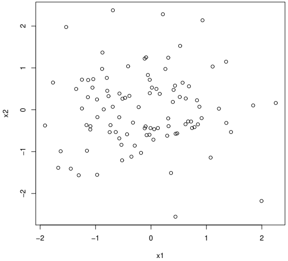


(a) $\rho = -1$ (b) $\rho = -0.8$ (c) $\rho = -0.5$

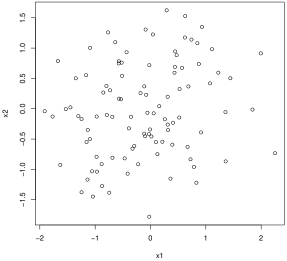
Scatterplots: Correlated Variables 2



(d) $\rho = -0.2$

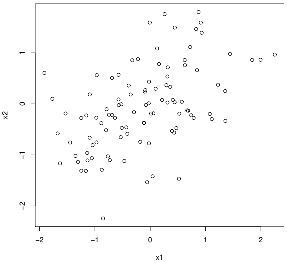


(e) $\rho = 0$

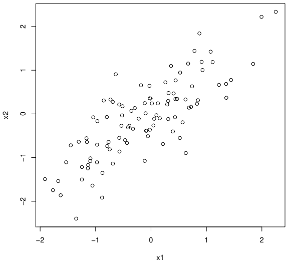


(f) $\rho = 0.2$

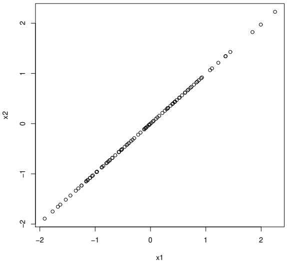
Scatterplots: Correlated Variables 3



(g) $\rho = 0.5$

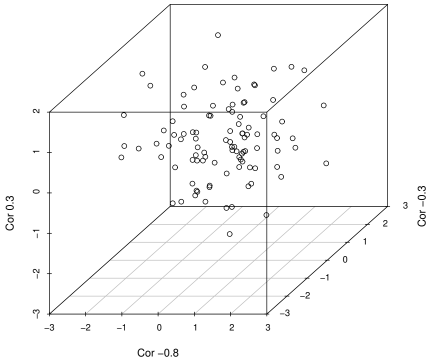


(h) $\rho = 0.8$



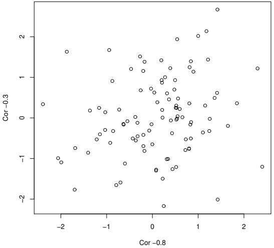
(i) $\rho = 1$

3d Scatterplots

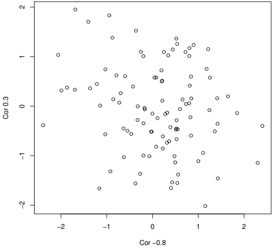


- Considers a third variable
- Can be viewed from any angle
- Difficult to fully understand the message that is being shown
- One option is to produce three pairwise plots

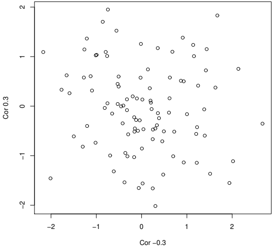
3d Scatterplots 2



(b) X_1 and X_2

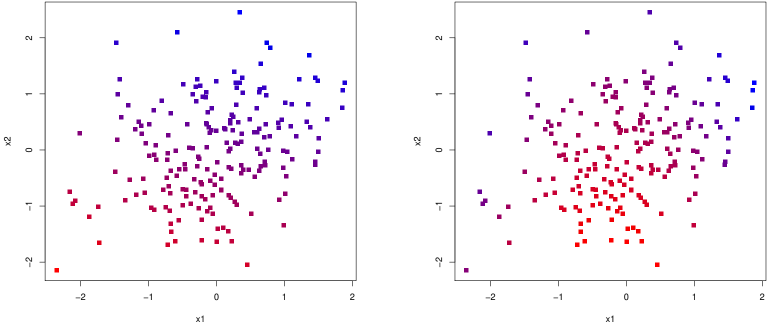


(c) X_1 and X_3



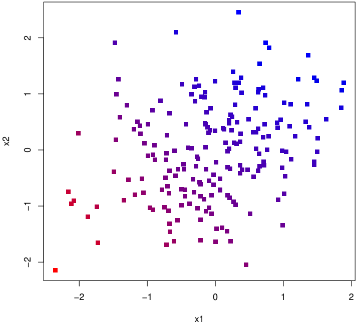
(d) X_2 and X_3

Scatters with Outcomes

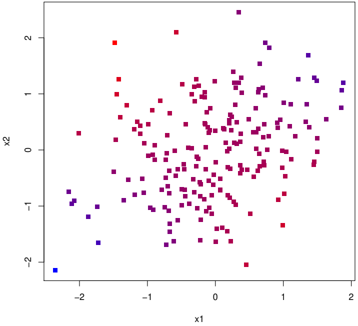


(a) $y_1 = 0.2x_1 + 0.7x_2$ (b) $y_2 = 0.2x_1 + 0.4x_1^2 + 0.5x_2$

Scatters with Outcomes 2



(c) $y_3 = 0.8x_1 + 0.2x_1^2 + 0.9x_2$

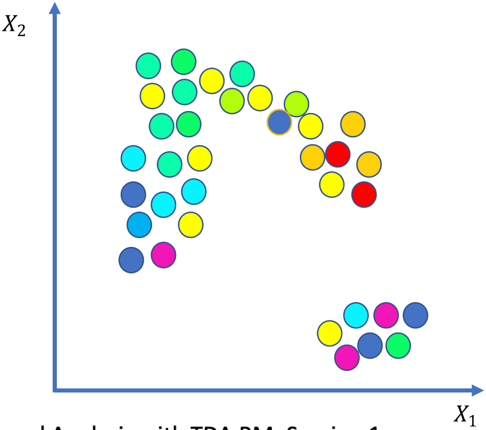


(d) $y_4 = 3x_1x_2$

Summary of Part 2

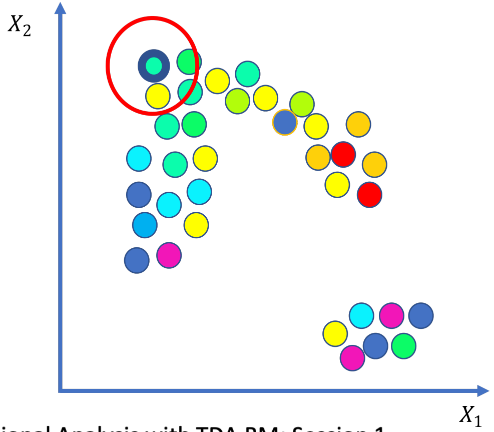
- Scatter plots provide a large amount of information considering simple plot
- Shape of scatter plots is influenced by correlation
- Moving beyond 2 variables typically means pair-wise plotting
- Want to be able to visualise in more dimensions...

Artificial Data



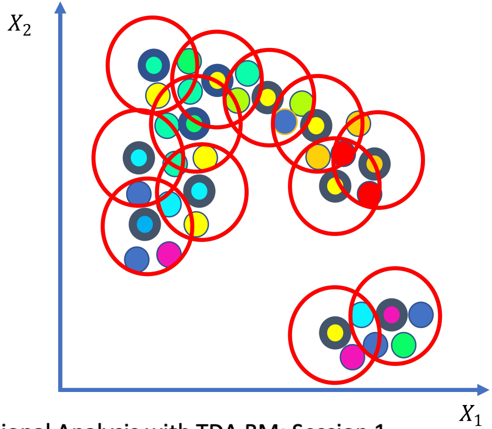
- Dataset coloured by outcome variable

Artificial Data



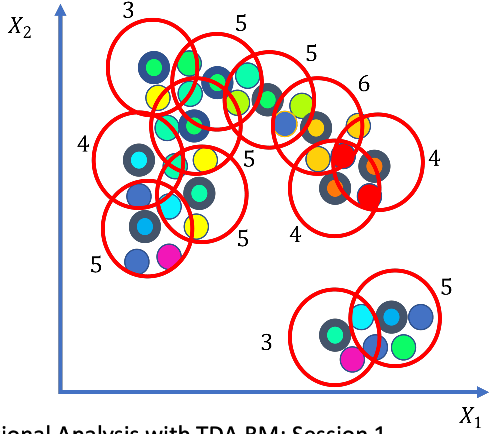
- Add a ball
- Radius of circle is the only parameter

Artificial Data



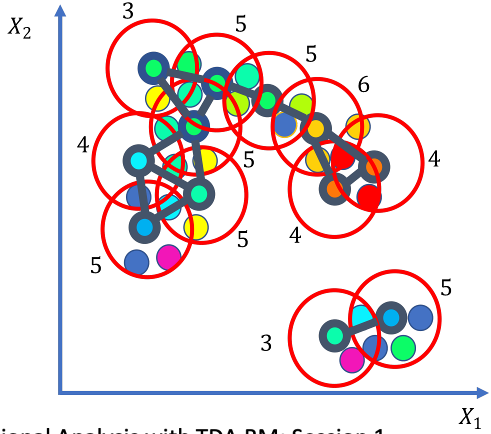
- Next centre must be a randomly chosen uncovered point
- Full coverage of the map

Artificial Data



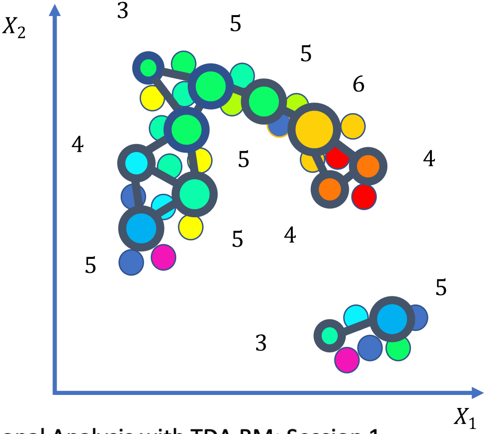
- Recolour to show average value
- Can use a different function
- Numbers to help us remember - computer would know

Artificial Data



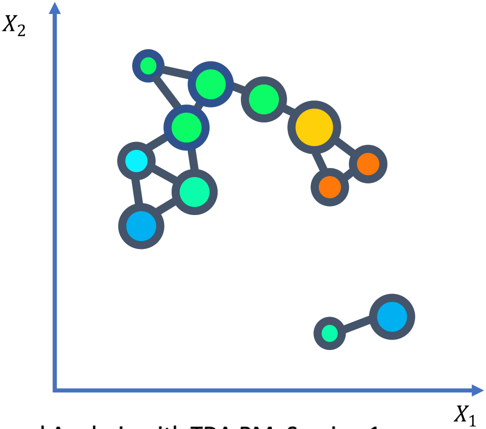
- Draw edges if points in intersection

Artificial Data



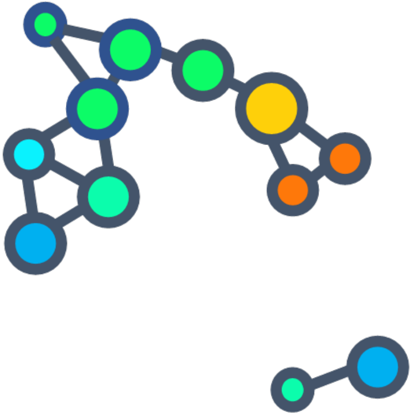
- Resize balls to show points in ball
- Indicative of density of space

Artificial Data



- Remove all information to produce essentially the map

Artificial Data

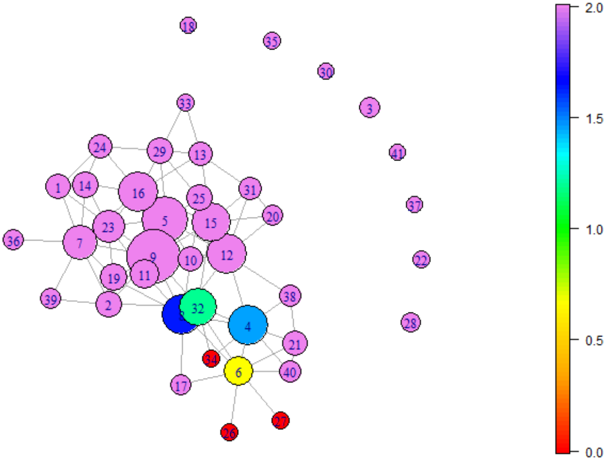


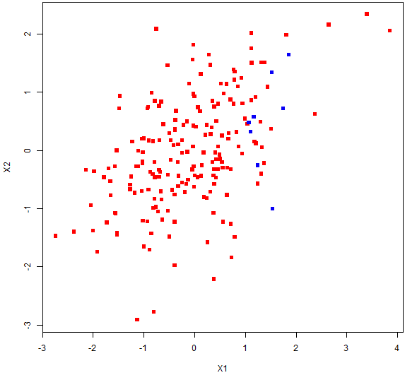
- Abstract with axes removed
- We could rotate this, flip it, or perform any transformation that keeps the links in tact
- Provided we have links we can understand the data - which direction corresponds to more X_1 can be set as the colouring variable

Summary of Part 3

- Ball Mapper begins with a point cloud of mutiple continuous variables
- Build a cover from balls until there is no data point not in at least one ball
- Connectivity in the BM graph shows points that are closeby in characteristic space
- Ball size shows the density of the joint distribution in that neighbourhood
- Colour of balls is a function on the values of members - usually simply the average of a sated variable
- Being able to see the data is just the start...

Understanding Outcomes



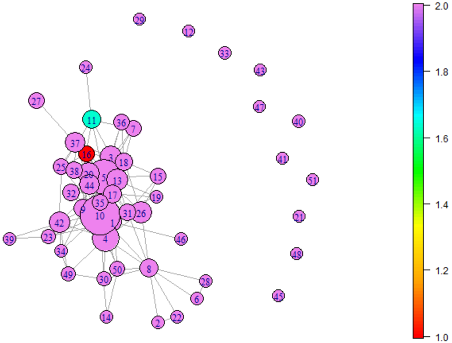


Regional Analysis with TDA BM: Session 1
Dr Simon Rudkin

Understanding Outcomes 2

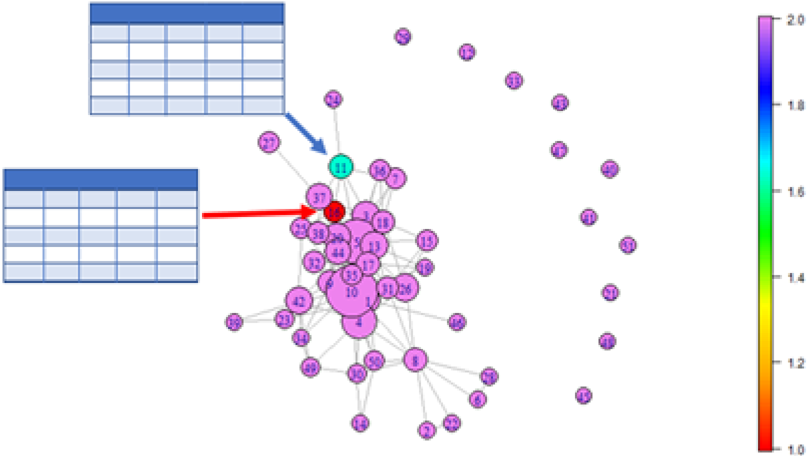
- The colouration function applied in panel (a) is based upon the values of X_1 and X_3.
- Scatter shows X_1 and X_2 - X_3 means blue squares are not clustered within the space.
- If we knew that X_3 was important then the next step would be a plot of X_1 and X_3
- Second case there are five variables, X_1 to X_5, and the condition for the negative outcome is that $1 < X_1 < 2$ and $1 < X_3 < 2$ and $X_5 > 1$.
- Correlations between X_1 and X_2 to X_5 are -0.1, 0.7, 0.5, and 0.1 respectively.

Understanding Outcomes 3



- Artificial examples can create outcome desired
- Colouration rule is chosen carefully
- However, real data may still produce this pattern - “when the stars align”

Understanding Outcomes 4



Data sits behind all of the pictures so we may query, gain insight, compare and evaluate

Elements of a BM Graph

Element	Brief Description
edges	List of edges providing the ball number that they run from and to. These edges are the connecting lines on the BM plot
edges _strength	This is a single column of numbers representing the number of points in the intersection of each pair defined in the edges list.
points_covered_by_landmarks	For each ball this gives a list of points that are within the ball. The list is separated by spaces and fits within one column
landmarks	Data points that are used as the centre of balls
coloring	Value used to colour the ball
coverage	For each point this gives a list of balls that contain the point. Again this list is in one column separated by spaces

Understanding Axes

```
colorByAllVariables(bm1,data,"bm1")
```

Inputs are:

- BallMapper object
- Data used to construct BM graph - one column per axis
- A prefix to use in the file names

Output will be of the form `bm11.png`, `bm12.png`, ...

Generating BM Graphs from R

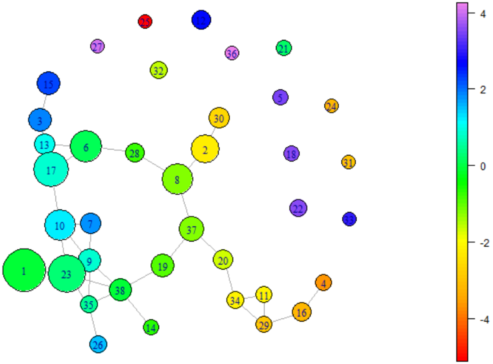
```
bm<-BallMapper(x,y,0.5)
```

Inputs are:

- A data frame with your characteristics , one column per axis
- A data frame which contains the outcome variable
- ϵ parameter to be used

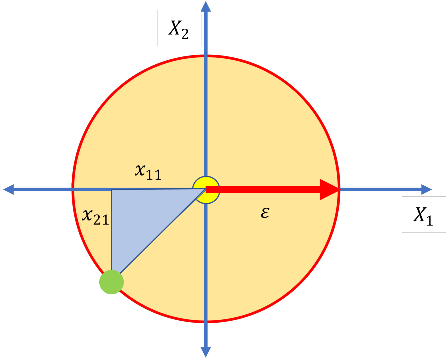
```
ColorIgraphPlot(bm,seed_for_plotting = 1)
```

First Example



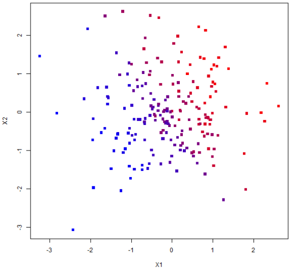
- X_1 200 draws from standard normal distribution
- X_2 also 200 draws from standard normal distribution but then transformed to give correlation with X_1 of 0.5
- $y_i = x_{1i} + x_{2i}$ is the colouring value for each point $i \in X$

Link with Epsilon

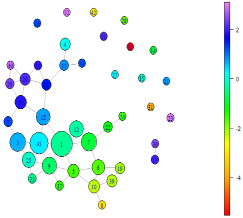


- Yellow point is landmark
- Aim to construct ball of radius ϵ
- Ball is formed
- If the other variable is identical then two points within the same ball can be 2ϵ apart
- Using Pythagoras we may compute the way that the differentiation would change moving round the ball

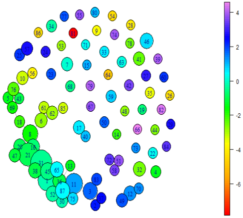
Effect of Adding Variables



(a) Scatter ($\rho = 0$)

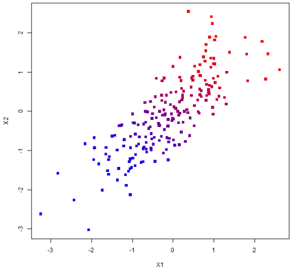


(b) Two Variables

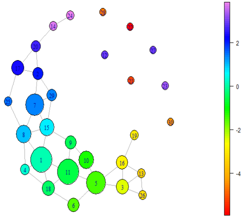


(c) Add Third Variable

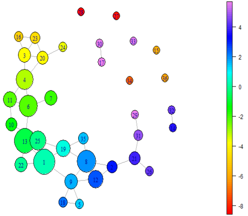
Effect of Adding Variables



(a) Scatter ($\rho = 0.8$)



(b) Two Variables

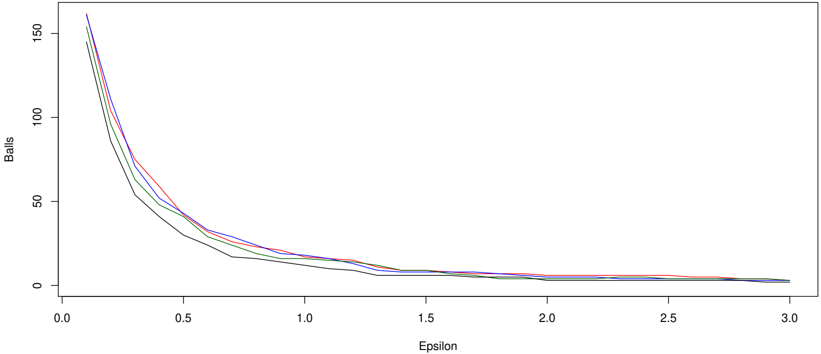


(c) Add Third Variable

Effect of Adding Variables 3

- Effect depends on correlation relative to that between X_1 and X_2
- In all cases the number of outliers increases because now the $\epsilon = 0.5$ must be split across three variables instead of two.
- Keeping ϵ in proportion to the number of variables is important.
- Unlike other algorithms the BM approach does not require the computation of a full distance matrix
- Instead a partial distance matrix to be constructed.
- More balls means more for computer to do calculating the relationship between the balls.
- Numbers of variables play an important part in that relationship as each dimension means more calculations and, for a given ball radius, more balls.

Effect of Radius

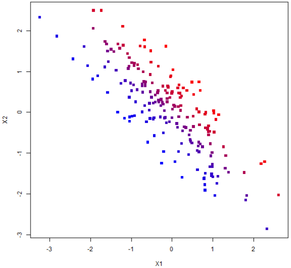


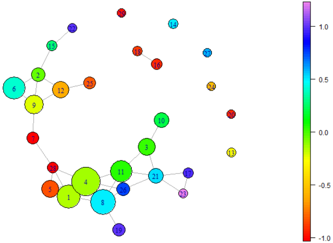
Here the red line denotes the cloud with $\rho = 0$, blue has $\rho = 0.2$, dark green has $\rho = 0.5$ and the black line has $\rho = 0.8$.

Effect of Radius 2

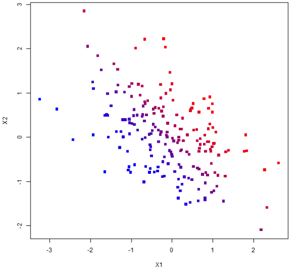
	Epsilon						
	0.1	0.5	1	1.5	2	2.5	3
$\rho = 0$	162	42	17	9	6	6	3
$\rho = 0.2$	161	43	18	8	5	4	3
$\rho = 0.5$	154	41	16	9	4	4	3
$\rho = 0.8$	145	30	12	6	3	3	2

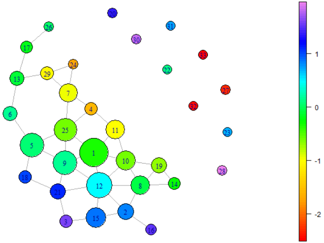
Correlation and BM Graphs $\rho = -0.8$



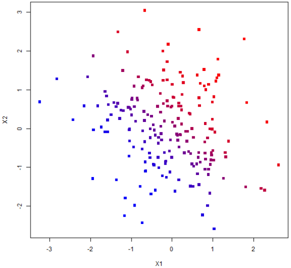


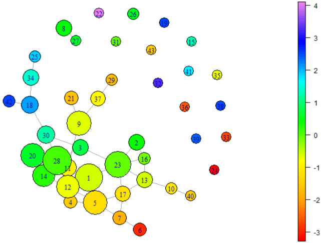
Correlation and BM Graphs $\rho = -0.5$



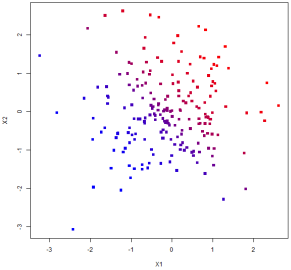


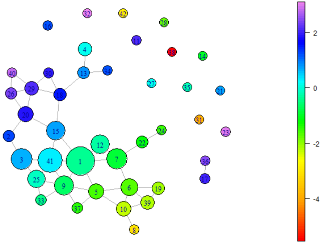
Correlation and BM Graphs $\rho = -0.2$





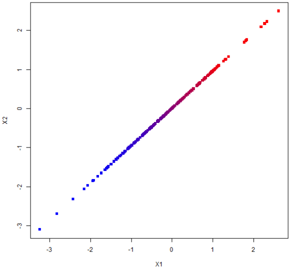
Correlation and BM Graphs $\rho = 0$

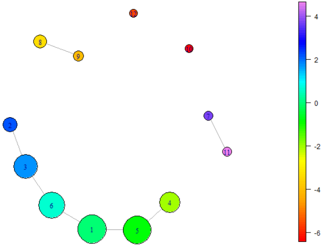




Regional Analysis with TDA BM: Session 1
Dr Simon Rudkin

Correlation and BM Graphs $\rho = 1$

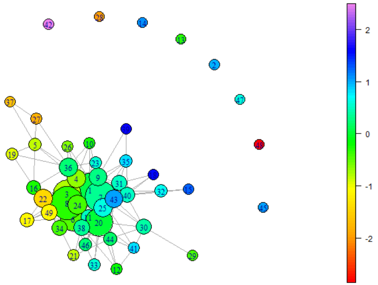




Correlation and BM Graphs 2

(a) Low Correlation

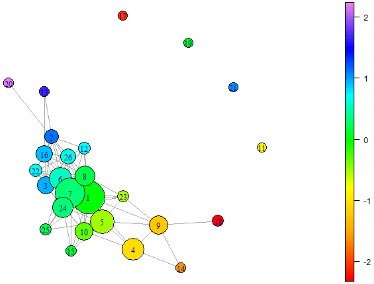
	X_1	X_2	X_3	X_4	X_5
X_1	1				
X_2	0.2	1			
X_3	0.1	-0.107	1		
X_4	0	0.091	0.095	1	
X_5	-0.1	-0.025	-0.073	-0.048	1
X_6	-0.2	-0.009	-0.057	0.094	0.019



Correlation and BM Graphs 3: High Correlation

(b) High Correlation

	X_1	X_2	X_3	X_4	X_5
X_1	1				
X_2	0.8	1			
X_3	0.7	0.517	1		
X_4	0.6	0.434	0.477	1	
X_5	-0.7	-0.558	-0.458	-0.429	1
X_6	-0.8	-0.613	-0.597	-0.477	0.610



Summary

- Simple plots of data help us to visualise
- Many of our inferences can be understood econometrically
- Moving into multiple dimensions complicates things - plot as pairs?
- Introduced the TDA Ball Mapper algroithm as a way to show data
- Session 3 will look at how to use the algorithm in R

Dłotko, P. (2019). Ball mapper: a shape summary for topological data analysis. *arXiv preprint arXiv:1901.07410*.