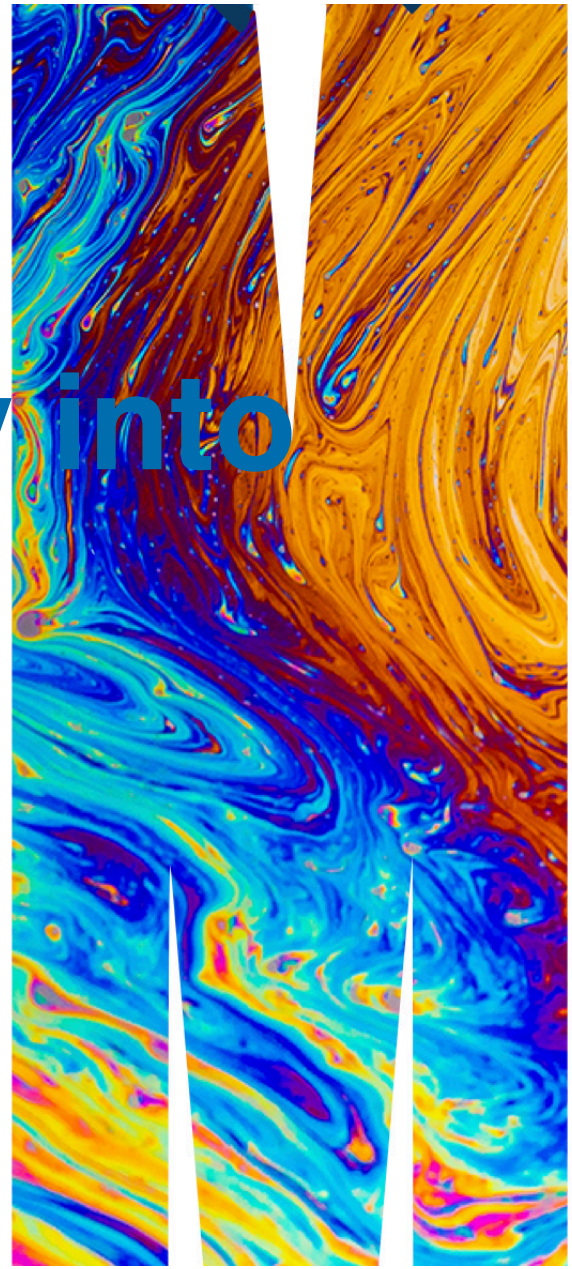
MONASH University

ETC5521: Diving Deeply into Data Exploration

Going beyond two variables, exploring high dimensions

Professor Di Cook

Department of Econometrics and Business Statistics



Outline

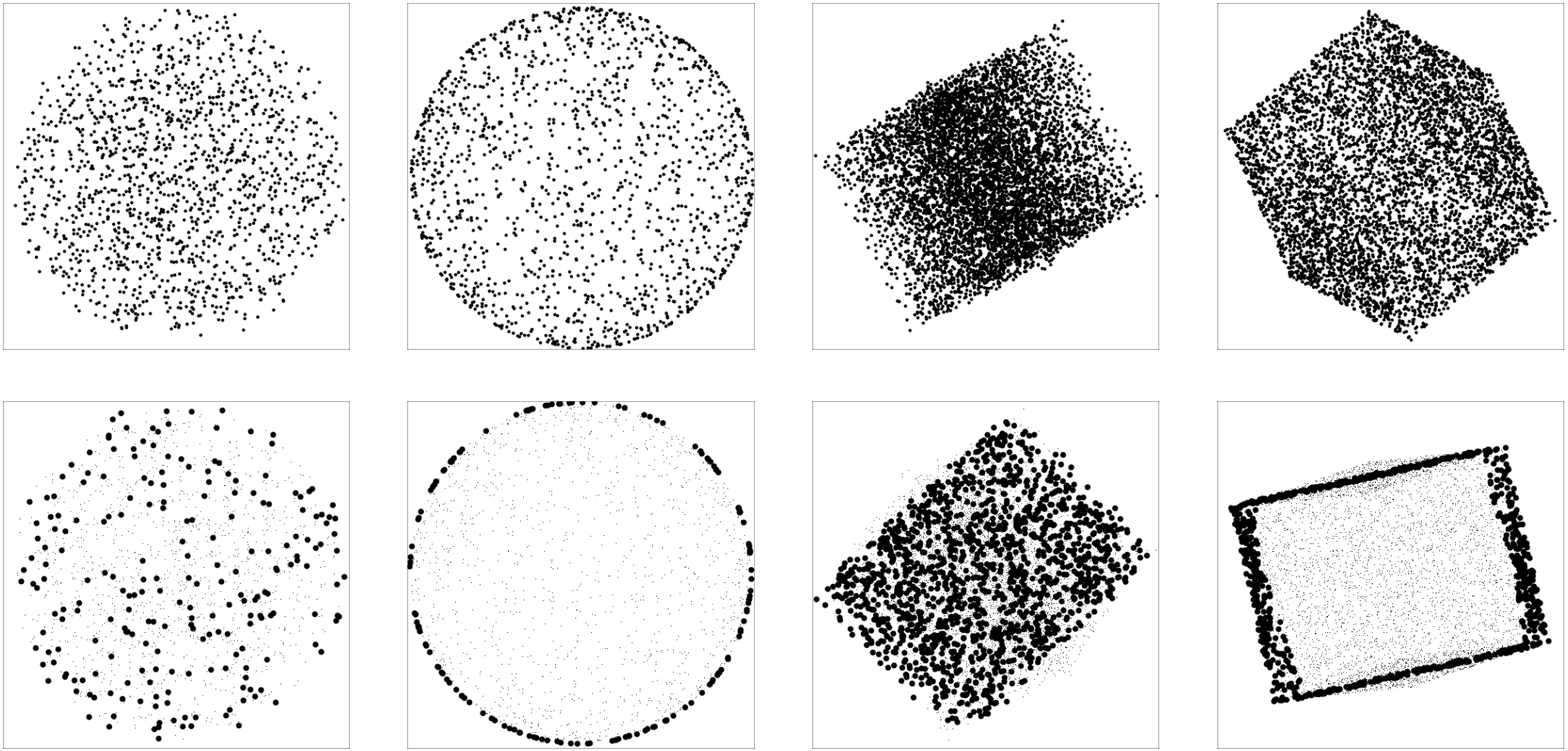
- What is high-dimensional data? (If all variables are quantitative)
- Exploring relationships between more than two variables
  - Tours - scatterplots of combinations of variables
  - Matrix of plots
  - Parallel coordinates
- What can be hidden
- Automating the search for pairwise relationships using scagnostics
- Linking elements of multiple plots
- Exploring multiple categorical variables

Flatland

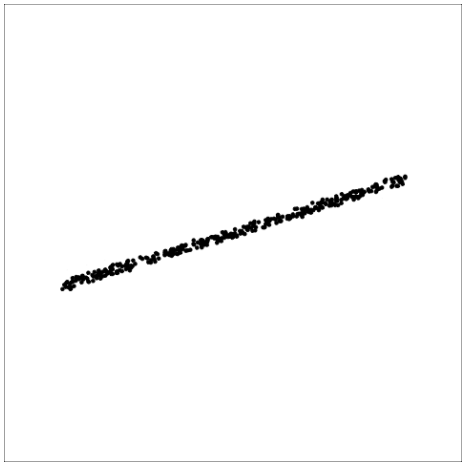
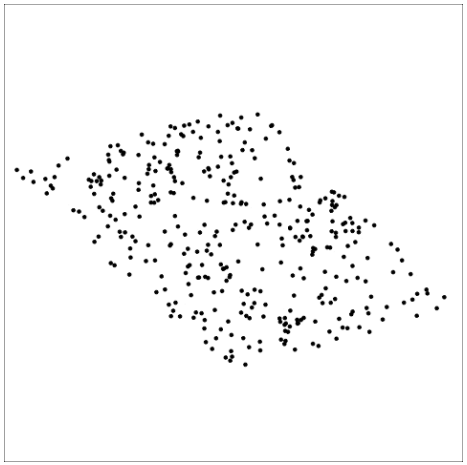
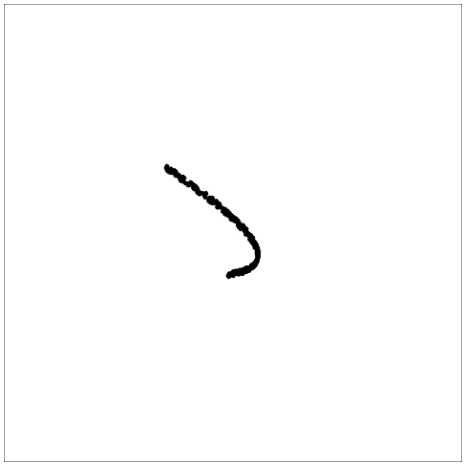
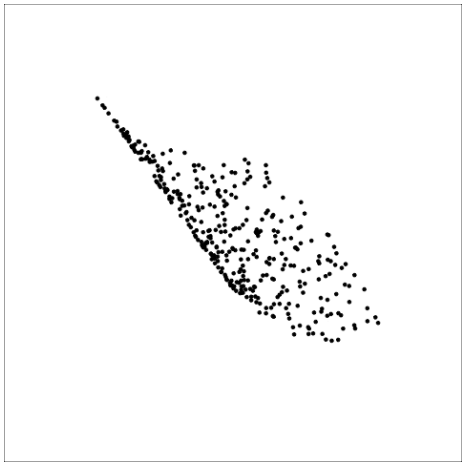
Click here to watch video

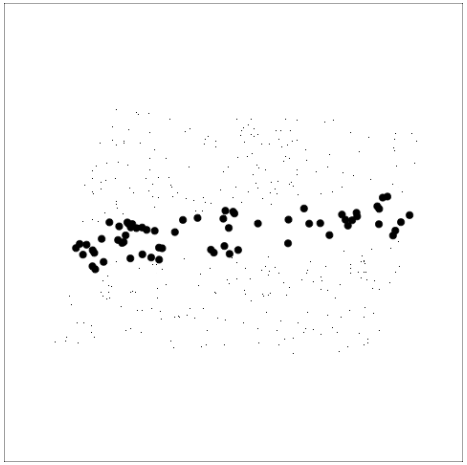
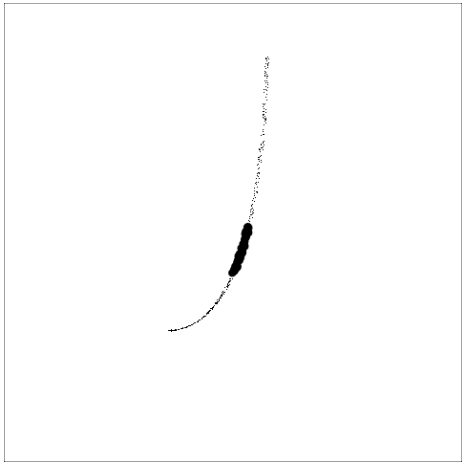
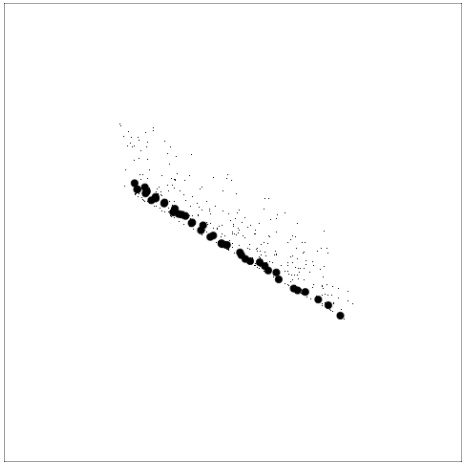
Trailer for “FLATLAND 2: SPHERELAND”. Original book, and movie information at wikipedia

High-dimensional shapes: shadows and slices

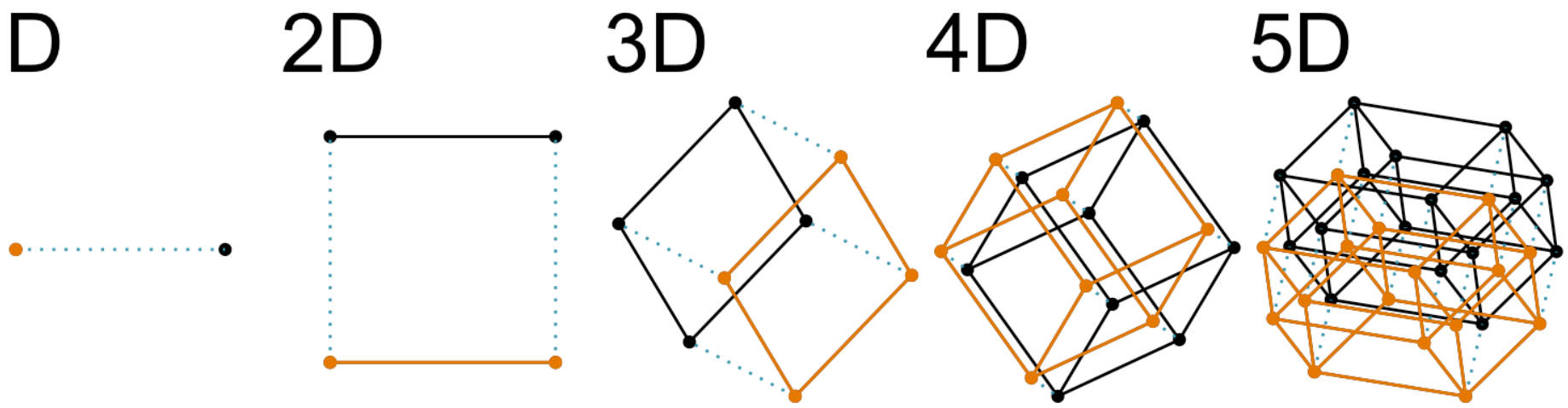


Low-dimensional shapes in high-dimensions

What is high-dimensions?



When all variables are quantitative, an extra variable adds an extra orthogonal axis. It has a name, Euclidean space which dates back to the ancient Greeks.

Features to find

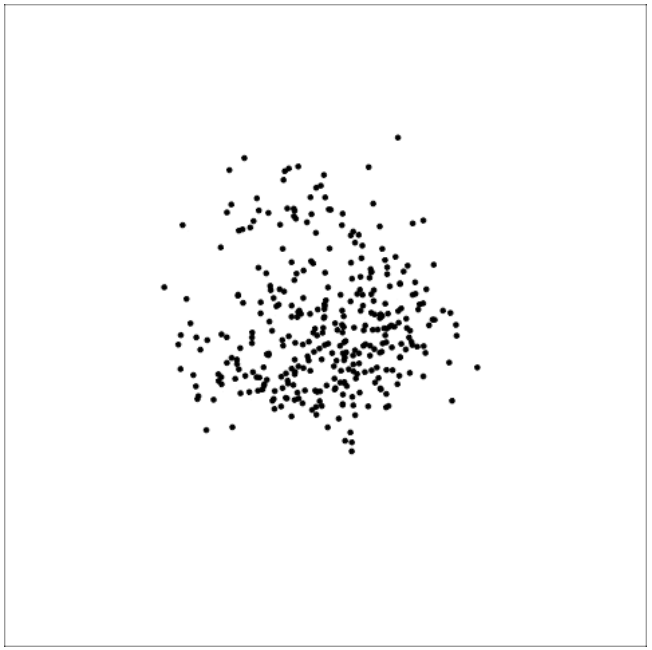
Feature	Example	Description
linear form		The shape is linear
nonlinear form		The shape is more of a curve
outliers		There are one or more points that do not fit the pattern on the others
clusters		The observations group into multiple clumps
gaps		There is a gap, or gaps, but its not clumped
barrier		There is combination of the variables which appears impossible
l-shape		When one variable changes the other is approximately constant
discreteness		Relationship between two variables is different from the overall, and observations are in a striped pattern

Any of the features from 2D are patterns to find in higher dimensions.

A movie of linear combinations: tour

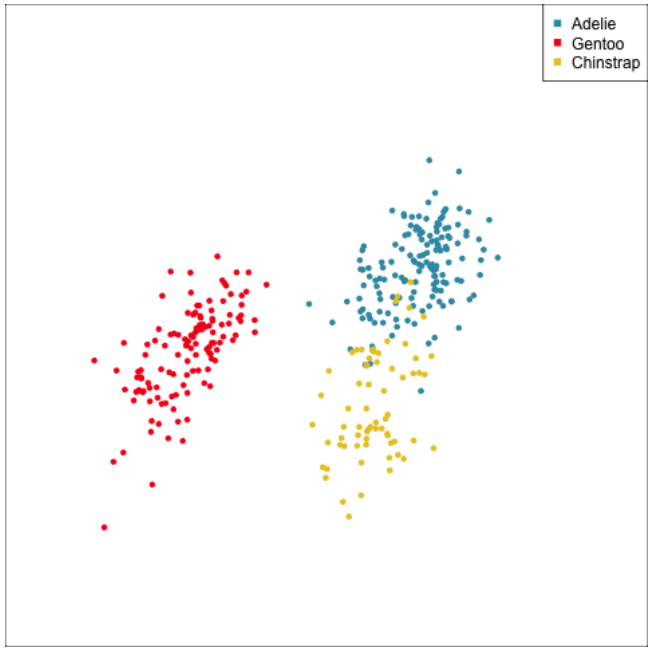
Grand tour

- data-processing
- Code



How many clusters?

- Code



The clusters correspond the three species.

What does linear combination of variables mean?

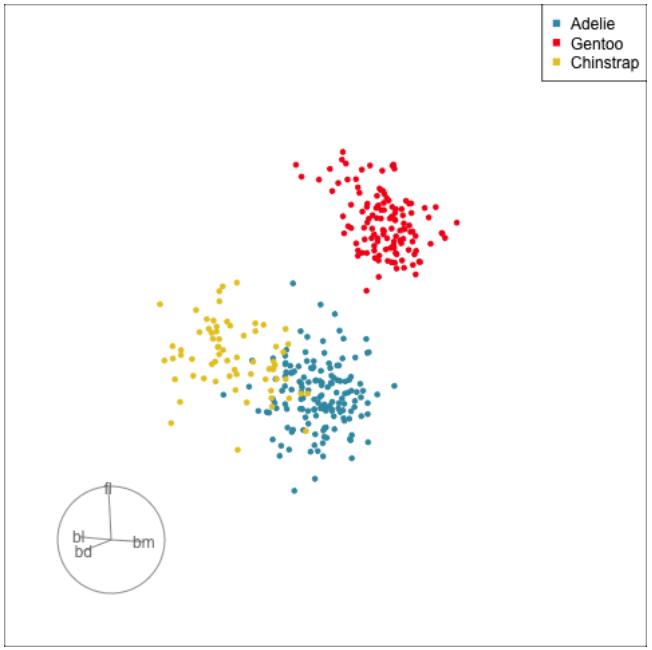
Click to see demo

If your data values are *-0.88, 0.78, -1.42, -0.56* and the coefficients are *0.23, -0.63, 0.67, -0.31* then the projected data value is *-1.47*. It's like a regression equation.

To make a scatterplot, two linear combinations are used. *With special care: (1) the sum of square of each equals 1, and (2) the sum of the product of the two linear combinations equals 0*.

Guided tour

- Code

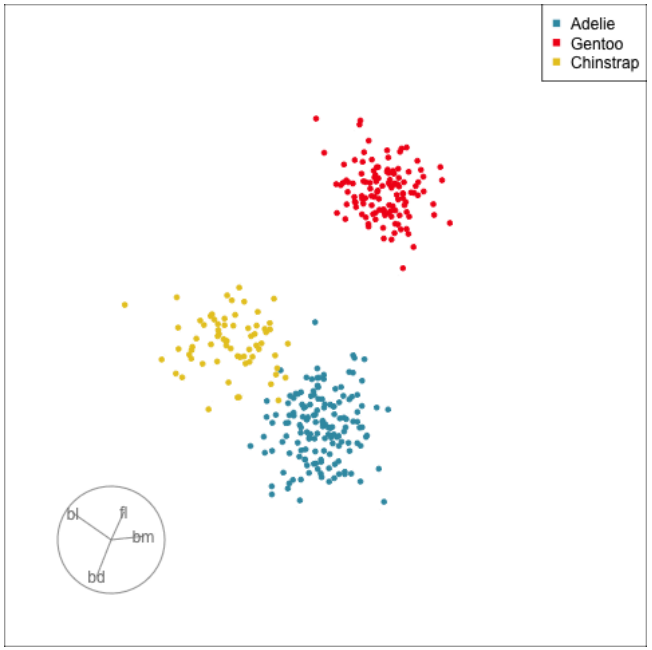


Define what structure is interesting, numerically, calculated by a function. Use an optimiser to choose linear combinations that maximise this function.

More on creating functions defining interesting structure soon!

Manual/radial tour

- Code



Remove a variable to see what the change to the pattern is.

Use this to assess whether a variable is important for a pattern, and hence the relationship between multiple variables.

Scale your data!

> The scale of the variables can affect how you see the relationships between multiple variables.

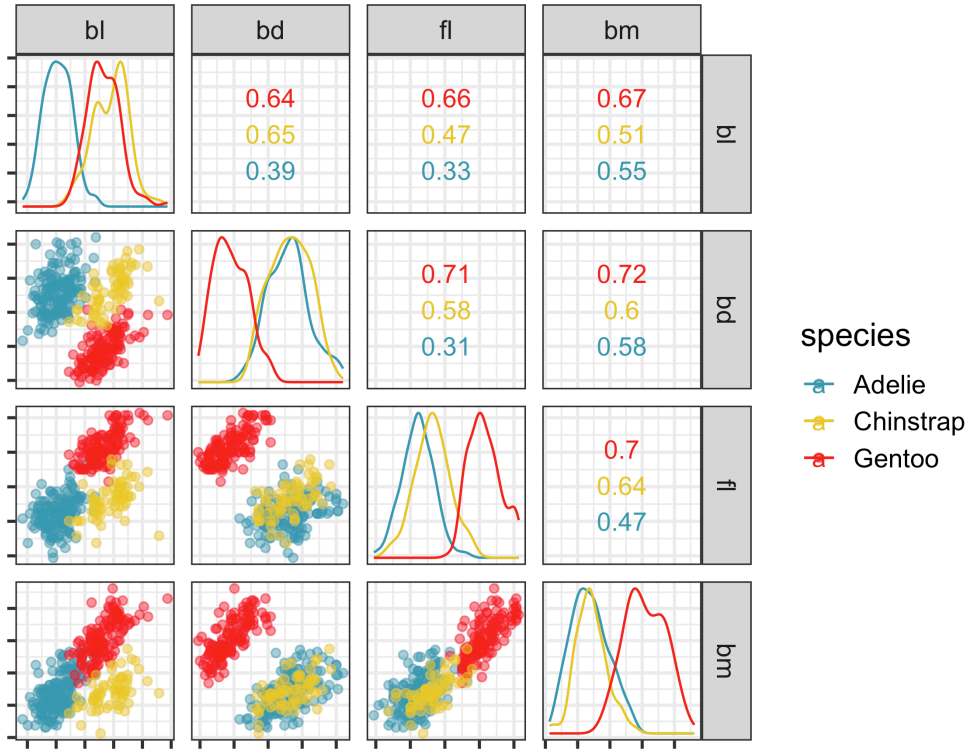
Generally, each variable should be scaled to have mean 0, standard deviation 1. (Or min -1 and max 1.)

If different scales on different variables is meaningful, and they are in the same units, you can scale with global mean and standard deviation (or minimum and maximum).

Static plots of multivariate data

Simpler: scatterplot matrix

▶ Code

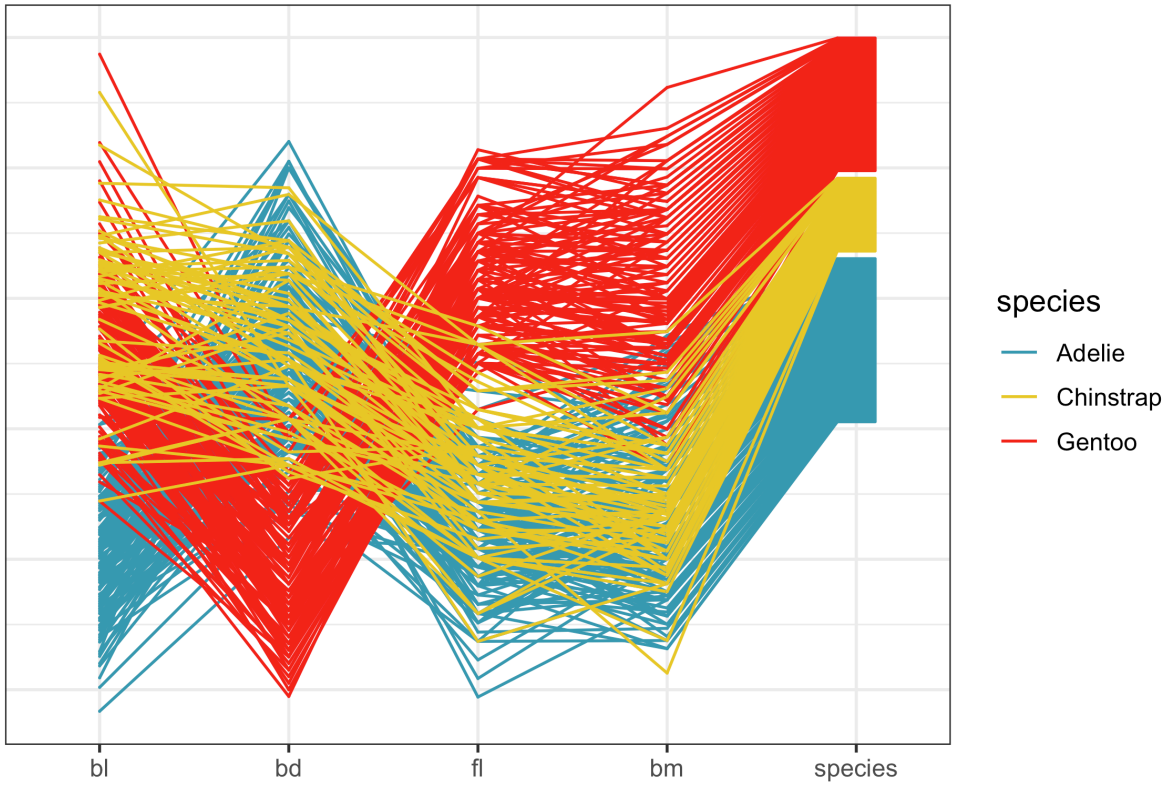


Plot

- all the pairs of variables.
- univariate distributions.
- maybe show correlations, too.

Parallel coordinate plot

▶ Code



Like side-by-side dot plots, where points are connected.

Look for patterns in the lines, such as

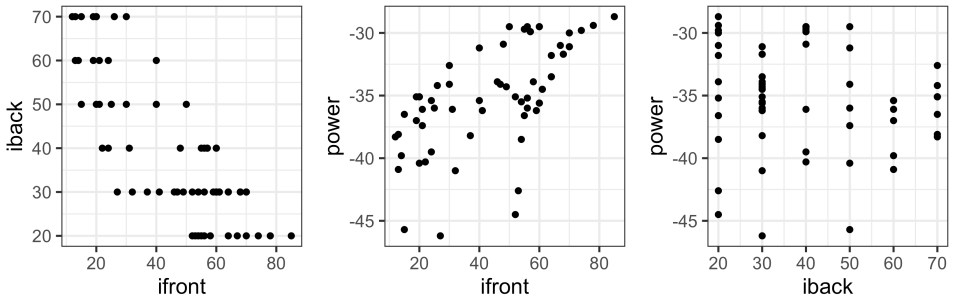
- grouped together indicating clustering.
- some lines going in different directions, indicating outliers.
- parallel or crossing lines, indicating linear positive and negative association.

What you might miss without a tour

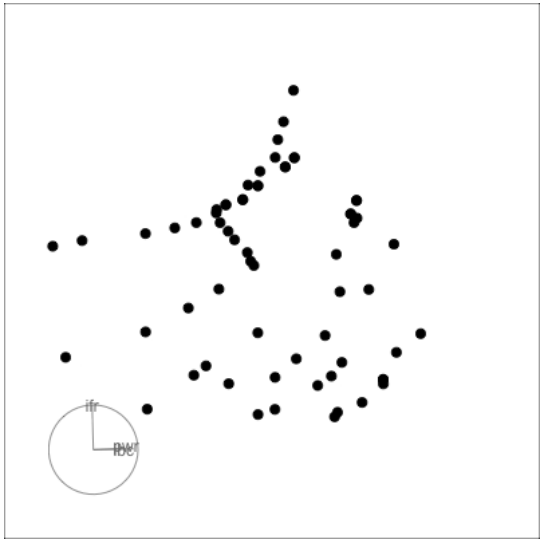
Hidden structure (1/3)

Many relationships can only be seen when multiple variables are combined:

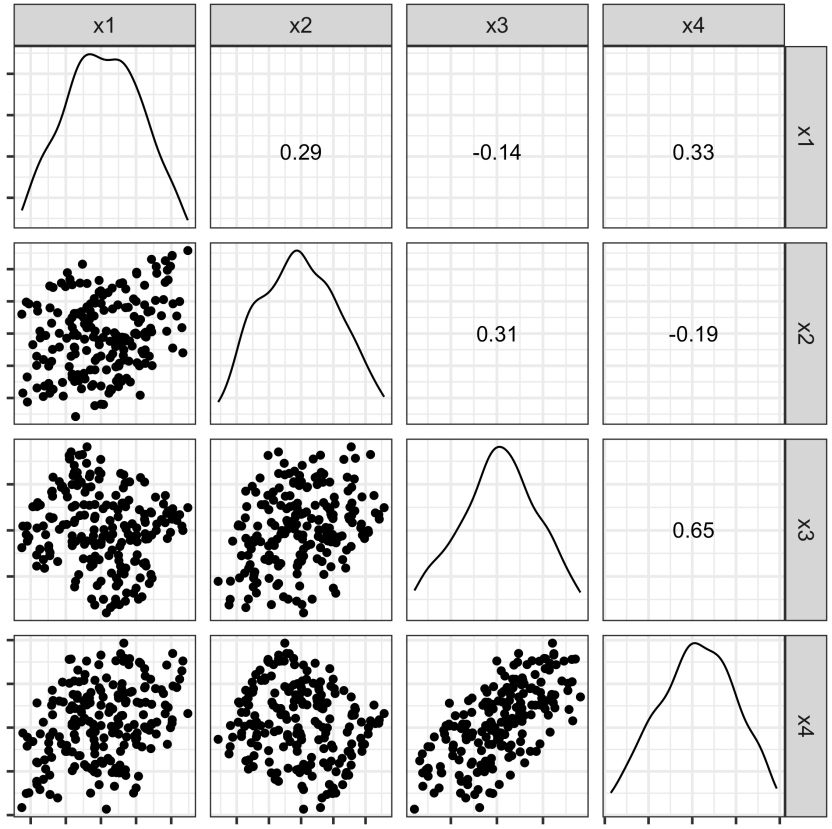
- outlier(s)
- clustering
- non-linearity

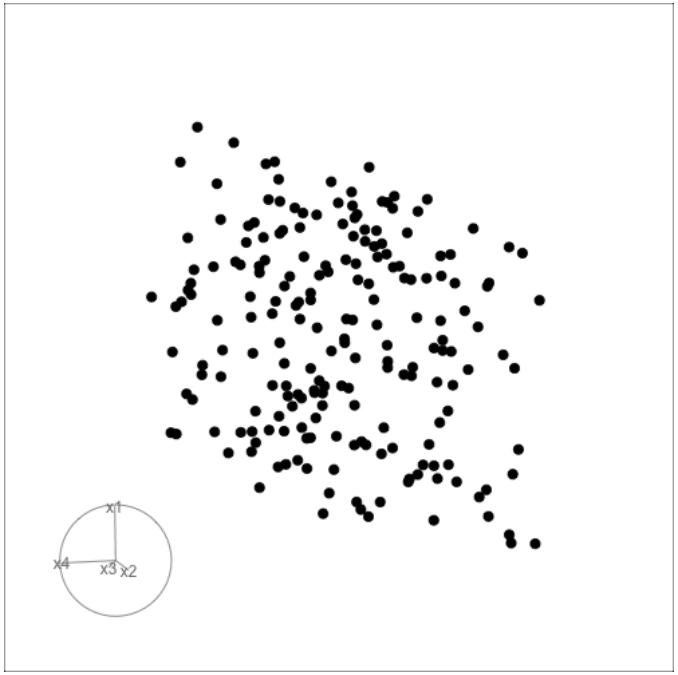


Example from my experience with early experiments in laser equipment construction.



Hidden structure (2/3)

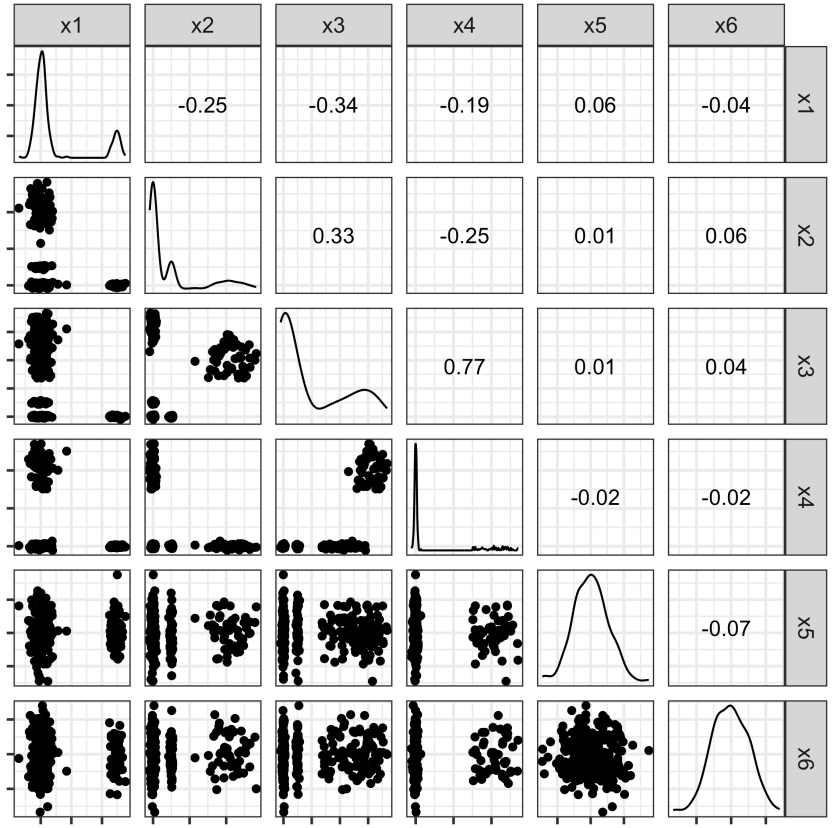


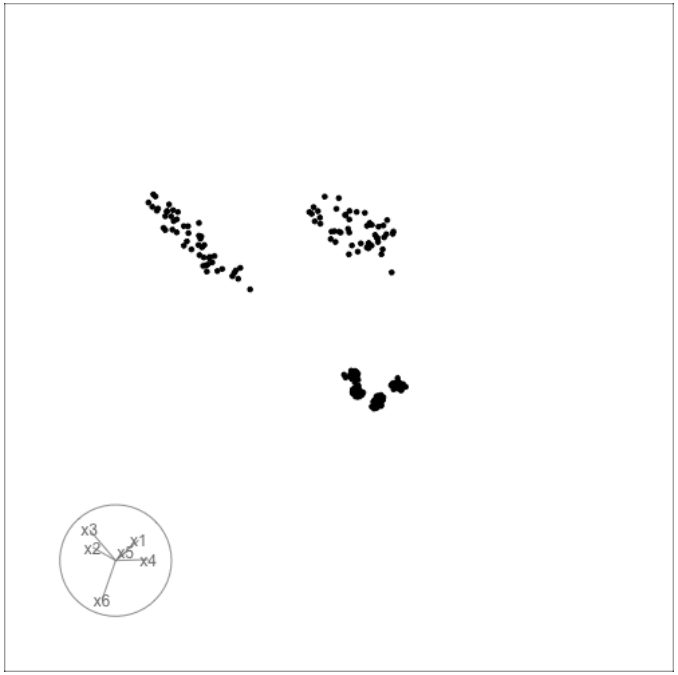


Can you see it now?

Can you see an anomaly?

Hidden structure (3/3)





How many can you see now?

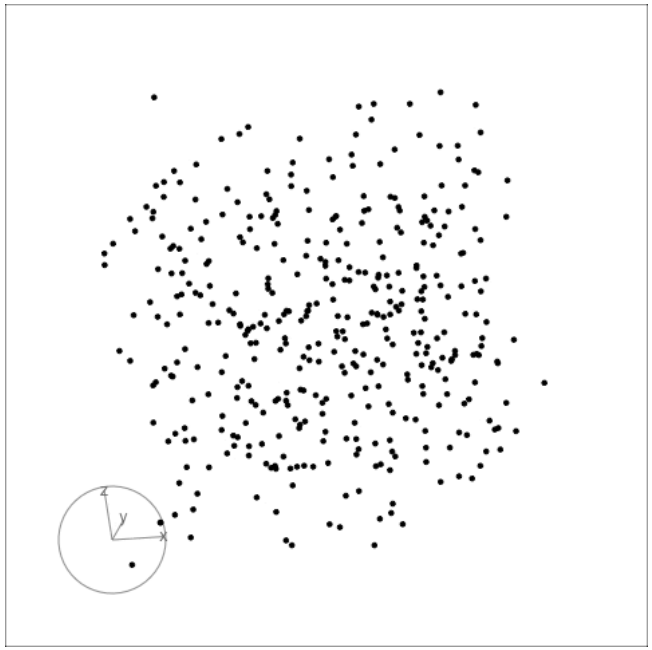
How many clusters?

Famous example: RANDU

RANDU[1] is a linear congruential pseudorandom number generator (LCG) used primarily in the 1960s and 1970s. Using RANDU for sampling a unit cube will only sample 15 parallel planes. As a result of the wide use of RANDU in the early 1970s, many results from that time are seen as suspicious.

Read more on wikipedia

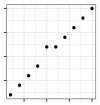
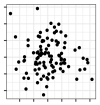
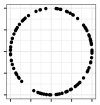
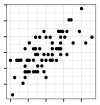
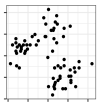
▶ Code



Automating the search with scagnostics

Scagnostics

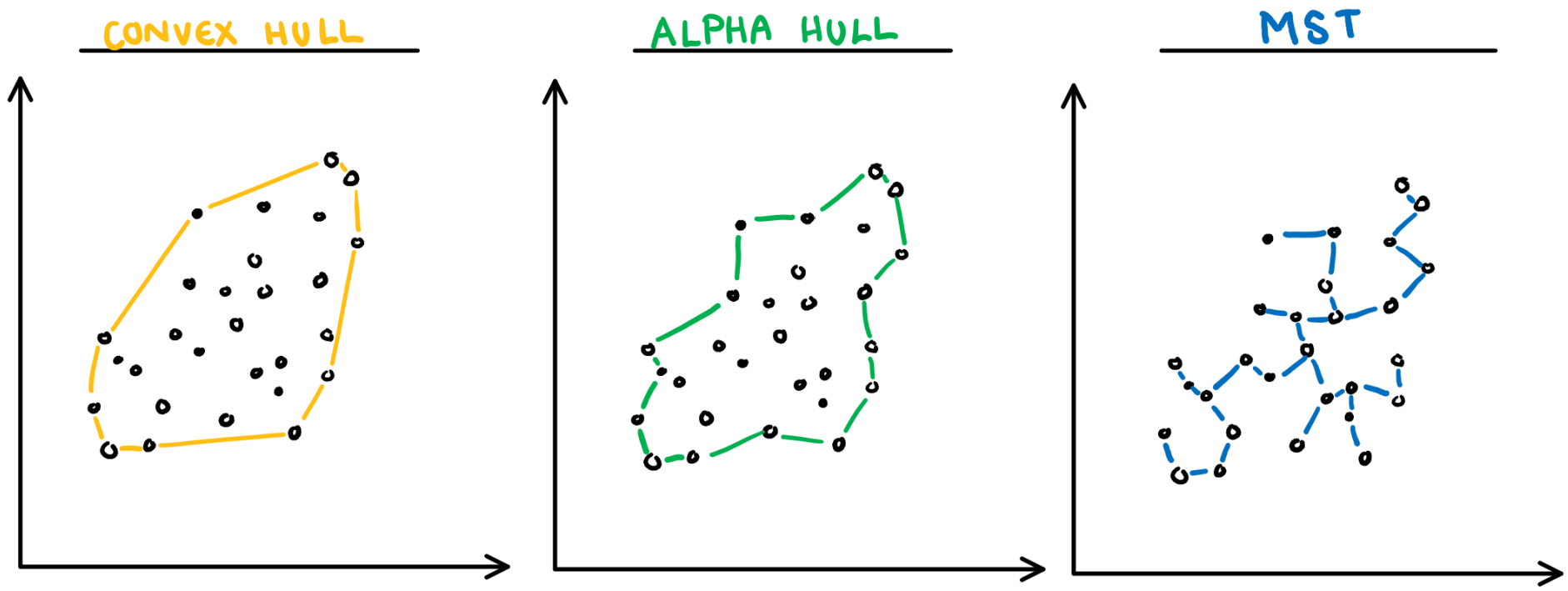
▶ Code

plot	set	outlying	stringy	striated	clumpy	sparse	monotonic	dcor
	line	0.000	1.00	0.60	0.37	0.157	0.997	0.99
	norm	0.190	0.79	0.33	0.60	0.095	0.013	0.16
	circle	0.000	1.00	0.98	0.97	0.065	0.009	0.25
	stripes	0.129	0.70	0.34	0.98	0.094	0.665	0.63
	clumps	0.038	0.61	0.23	0.99	0.107	0.375	0.50

- clumpy
- convex
- dcor
- monotonic
- outlying
- skewed
- skinny
- sparse
- splines
- striated
- stringy
- striped

How are scagnostics calculated? (1/3)

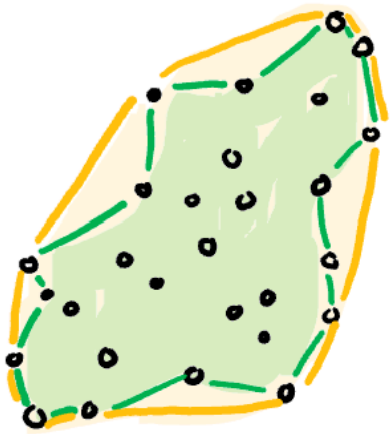
The building blocks are: convex hull, alpha hull, and minimal spanning tree



Sketches made by Harriet Mason

How are scagnostics calculated? (2/3)

Convex: Measure of how convex the shape of the data is. Computed as the ratio between the area of the alpha hull (A) and convex hull (C).



$$\frac{\text{ALPHA HULL AREA}}{\text{CONVEX HULL AREA}}$$

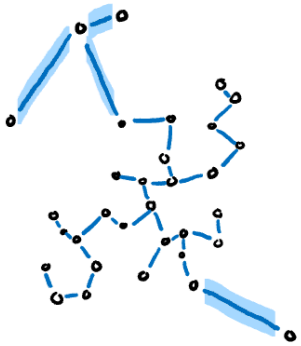
Skinny: A measure of how "thin" the shape of the data is. It is calculated as the ratio between the area and perimeter of the alpha hull (A) with some normalisation such that 0 correspond to a perfect circle and values close to 1 indicate a skinny polygon.

$$1 - \sqrt{\frac{4\pi \cdot \text{ALPHA HULL AREA}}{\text{ALPHA HULL PERIMETER}}}$$

Sketches made by Harriet Mason

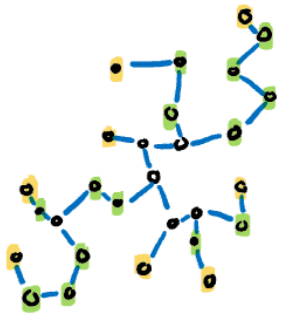
How are scagnostics calculated? (3/3)

Outlying: A measure of proportion and severity of outliers in dataset. Calculated by comparing the edge lengths of the outlying points in the MST with the length of the entire MST.



$$\frac{\text{EDGE LENGTHS OF OUTLYING POINTS}}{\text{EDGE LENGTHS OF ORIGINAL MST}}$$

Stringy: This measure identifies a "stringy" shape with no branches, such as a thin line of data. It is calculated by comparing the number of vertices of degree two ($V^{(2)}$) with the total number of vertices (V), dropping those of degree one ($V^{(1)}$).



$$\frac{\text{NUMBER OF } V^{(2)}}{\text{TOTAL NUMBER OF } V - \text{NUMBER OF } V^{(1)}}$$

Sketches made by Harriet Mason

Scagnostics from familiar measures

There are many more ways to numerically characterise association that can be used as scagnostics too:

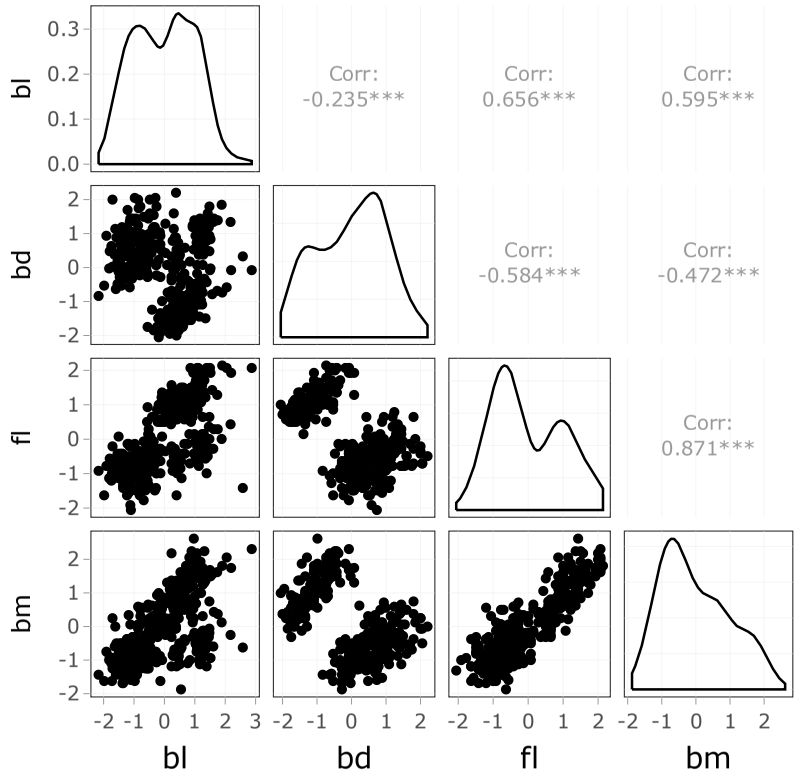
- We used those available in the cassowaryr R package
- Slope, intercept, error, R^2 from a simple linear model
- Also beyond scatterplots there are:
  - tignostics for time series (feasts R package)
  - longnostics for longitudinal data (brolgar R package)

Linking elements of multiple plots

Brushing in a scatterplot matrix

▶ Code

Selecting points using a square “brush”, using `plotly`, allows you to see where observations lie in the other plots (pairs of variables).



Linking between a tour and a scatterplot

▶ Demo

Linking between plots allows some queries to be made interactively. The penguins data has variables, `sex` and `island` which provide more information.

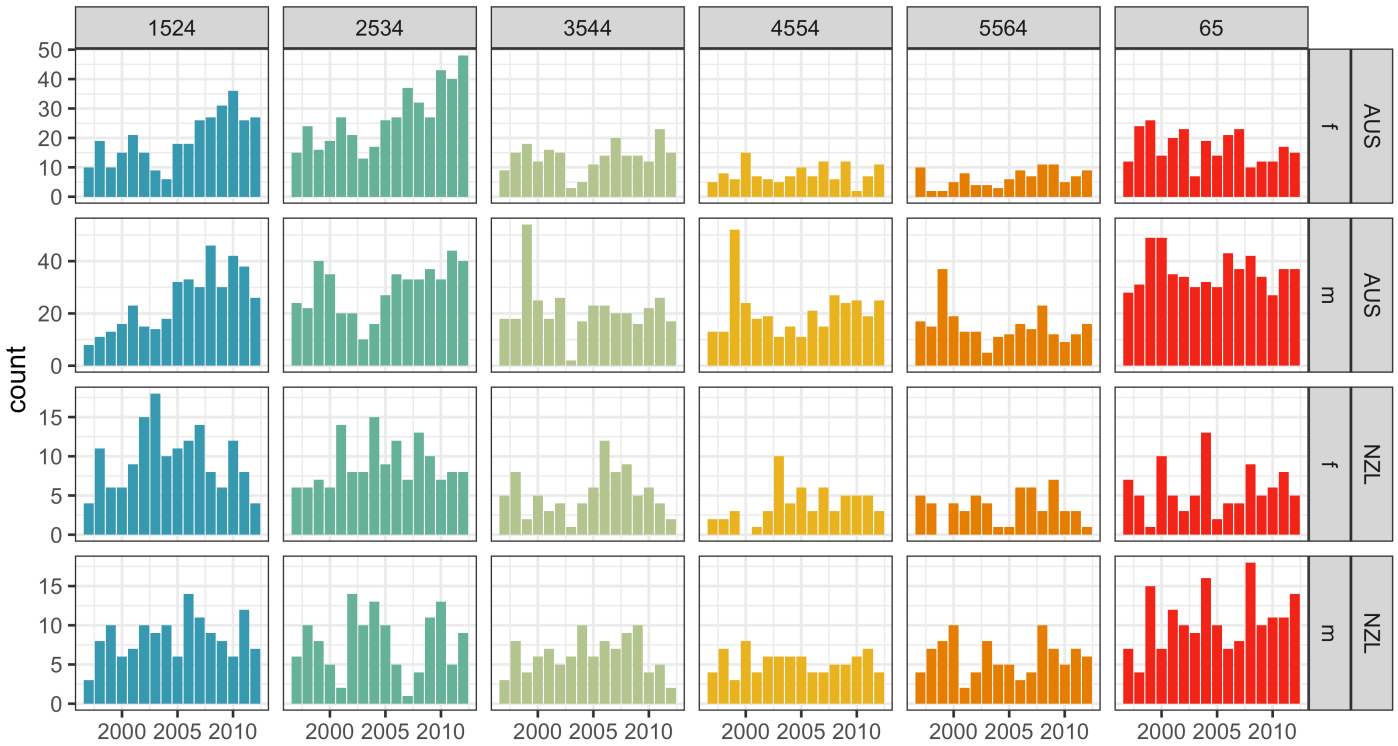
This demo code is setup to learn more the penguins: a (jittered) scatterplot shows island and sex, and a tour shows the four size measurements.

What can you learn in response to these questions?

1. Is the size of the penguins different between the sexes?
2. How does the size differ between penguins recorded at different locations?
3. Which island did the outlier come from?

Exploring multiple categorical variables

Facetted plots



Additional variables can be folded in to horizontal and vertical facets.

Ordering could be important to change, and scales on the axes need care.

Switching to proportions may allow direct comparison.

Tabulation

Overall counts

age	f	m
1524	314	395
2534	432	469
3544	216	345
4554	126	332
5564	103	244
65	269	575
Total	**1460**	**2360**

age	f	m
1524	0.44	0.56
2534	0.48	0.52
3544	0.39	0.61
4554	0.28	0.72
5564	0.30	0.70
65	0.32	0.68
Total	**0.38**	**0.62**

What percentage was taken? What is the comparison? What other percentages could be useful?

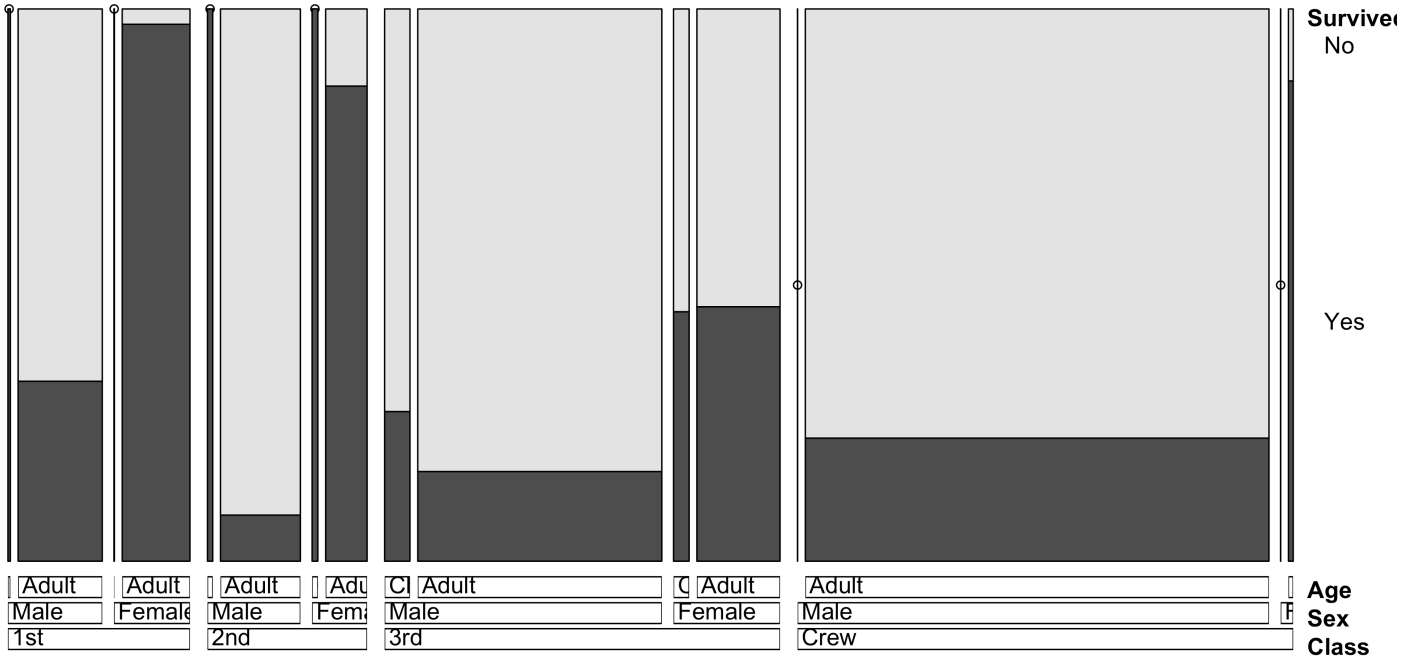
The way percentages are calculated corresponds to conditional distributions, e.g. if age is 15-24 what is the distribution of tuberculosis between the sexes.

Famous example: titanic

This data set provides information on the fate of passengers on the fatal maiden voyage of the ocean liner “Titanic”, summarized according to economic status (class), sex, age and survival.

Mosaic plots: conditional distributions

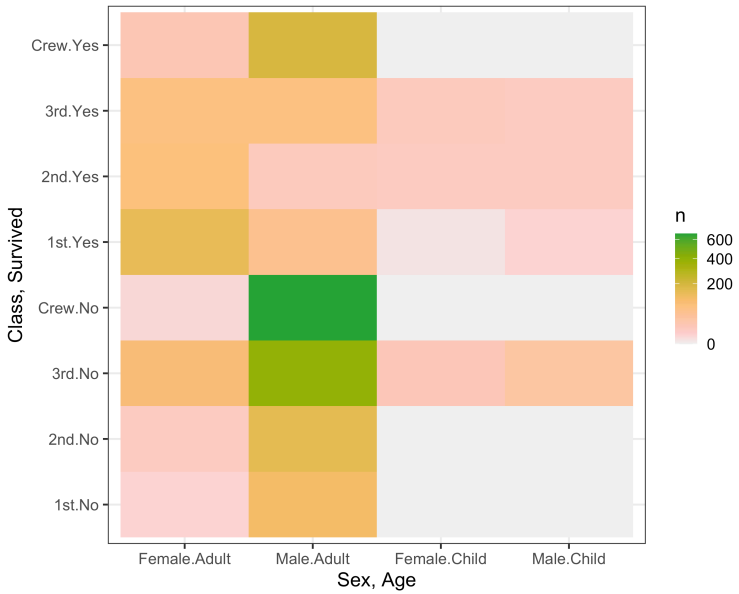
- Code



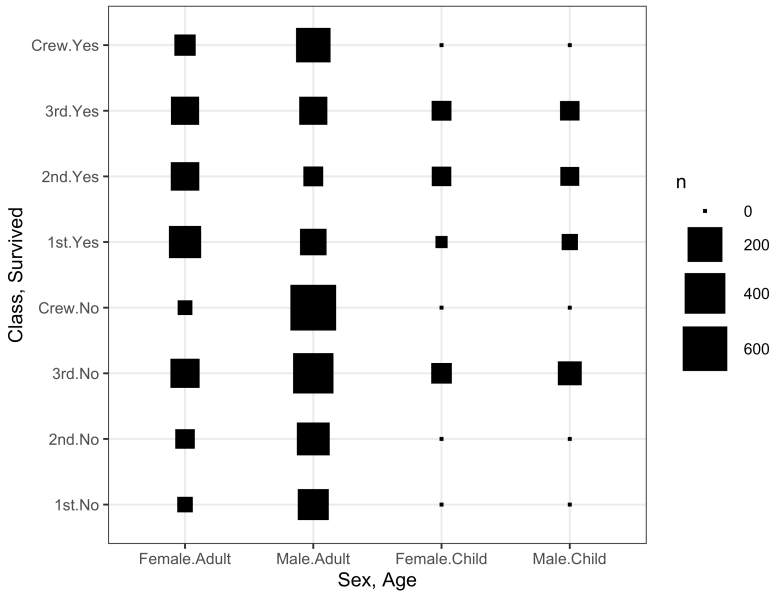
Order in which variables are entered can affect the conditioning.

Fluctuation diagrams: joint distributions

- Code



- Code



Overall counts mapped to tiling colour or size. Big counts completely dominate.

Resources

- Cook and Laa (2023) Interactively exploring high-dimensional data and models in R
- Wickham et al (2011). tourr: An R Package for Exploring Multivariate Data with Projections
- Sievert (2019) Interactive web-based data visualization with R, plotly, and shiny
- Mason, Lee, Laa, and Cook (2022). cassowaryr: Compute Scagnostics on Pairs of Numeric Variables in a Data Set