ETC3250/5250: Introduction to Machine Learning

Visualisation of multivariate data

Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 5a

Why is visualisation important?

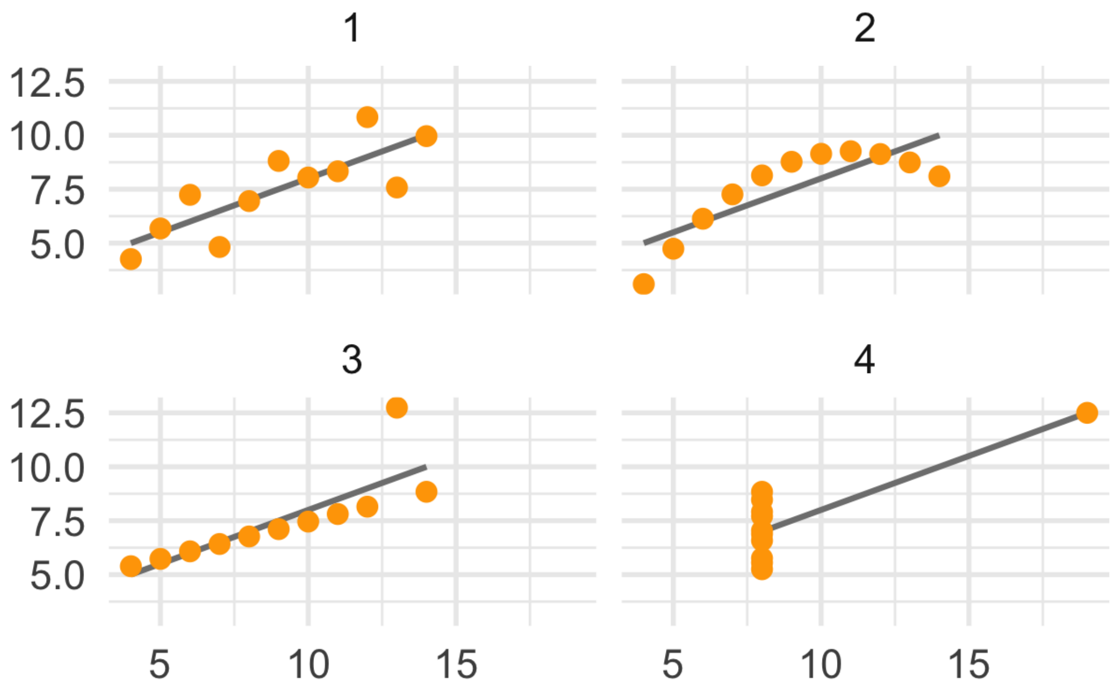
Consider the following datasets, known as Anscombe's quartet, which all have the same numerical statistical summaries.

set	mx	my	sx	sy	r
1	9	7.5	3.32	2.03	0.82
2	9	7.5	3.32	2.03	0.82
3	9	7.5	3.32	2.03	0.82
4	9	7.5	3.32	2.03	0.82

Why is visualisation important?

Consider the following datasets, known as Anscombe's quartet, which all have the same numerical statistical summaries.

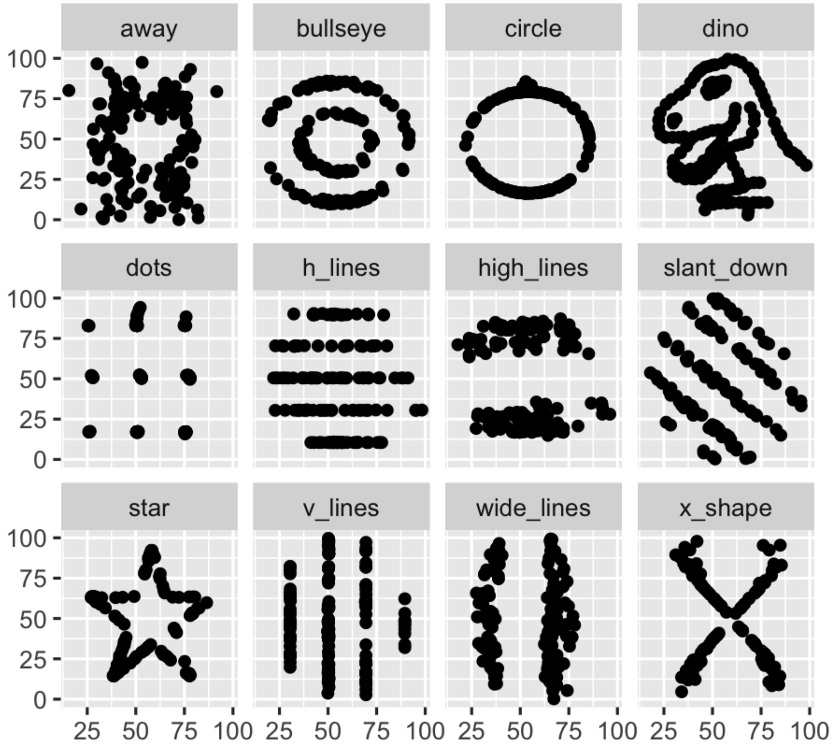
Plots provide a more detailed statistical summary.



Why is visualisation important?

Very different data can have the same numerical summaries.

Datasaurus dozen: all have the same means, standard deviations and correlation, also.



Why is visualisation important?

In machine learning visualisation is used for:

- Initial data analysis: to examine whether the data
  - satisfies assumptions required for the method
  - has unexpected complications like outliers or nonlinearity
- Assess the model fit:
  - predicted vs observed
  - residuals
  - boundaries between classes

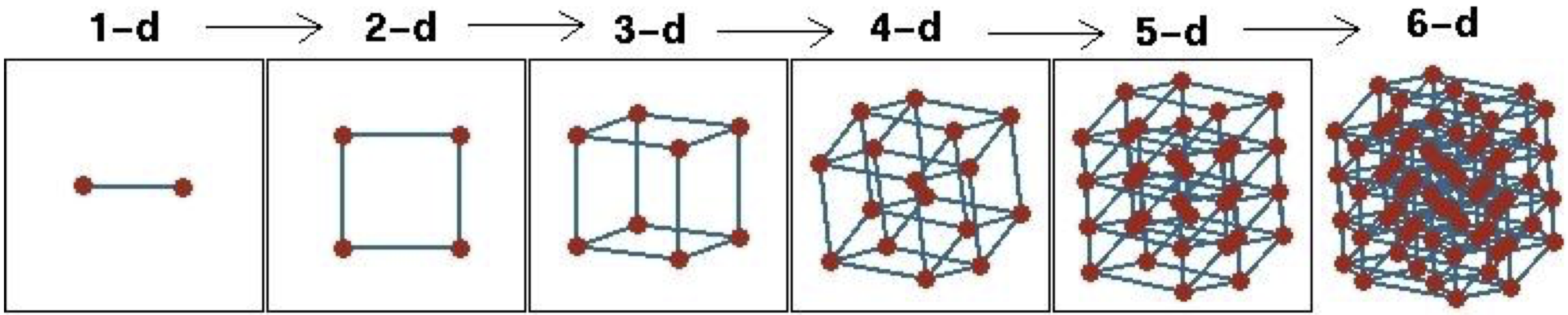
High dimensional visualisation

Common methods for visualising high-dimensions

- Tour
- Parallel coordinate plot
- Scatterplot matrix
- Mosaic plot

Most of what you find when you google "visualising high-dimensions" is awful, e.g. use colour and symbol after 3D to show 5D; PCA, MDS, tSNE, are visualisation methods; "you can't see beyond 3D".... Rubbish!

Dimensionality



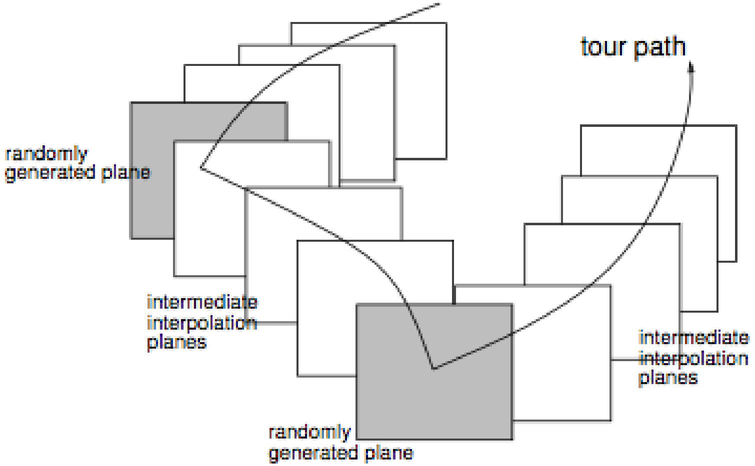
- When you add another variable, you implicitly add another orthogonal axis.
- The space is effectively a p-dimensional cube
- The data might not fill the cube, and then dimension reduction might make it a $k(< p)$-dimensional cube

Grand tours

A grand tour is by definition a movie of low-dimensional projections constructed in such a way that it comes arbitrarily close to showing all possible low-dimensional projections; in other words, a grand tour is a space-filling curve in the manifold of low-dimensional projections of high-dimensional data spaces.

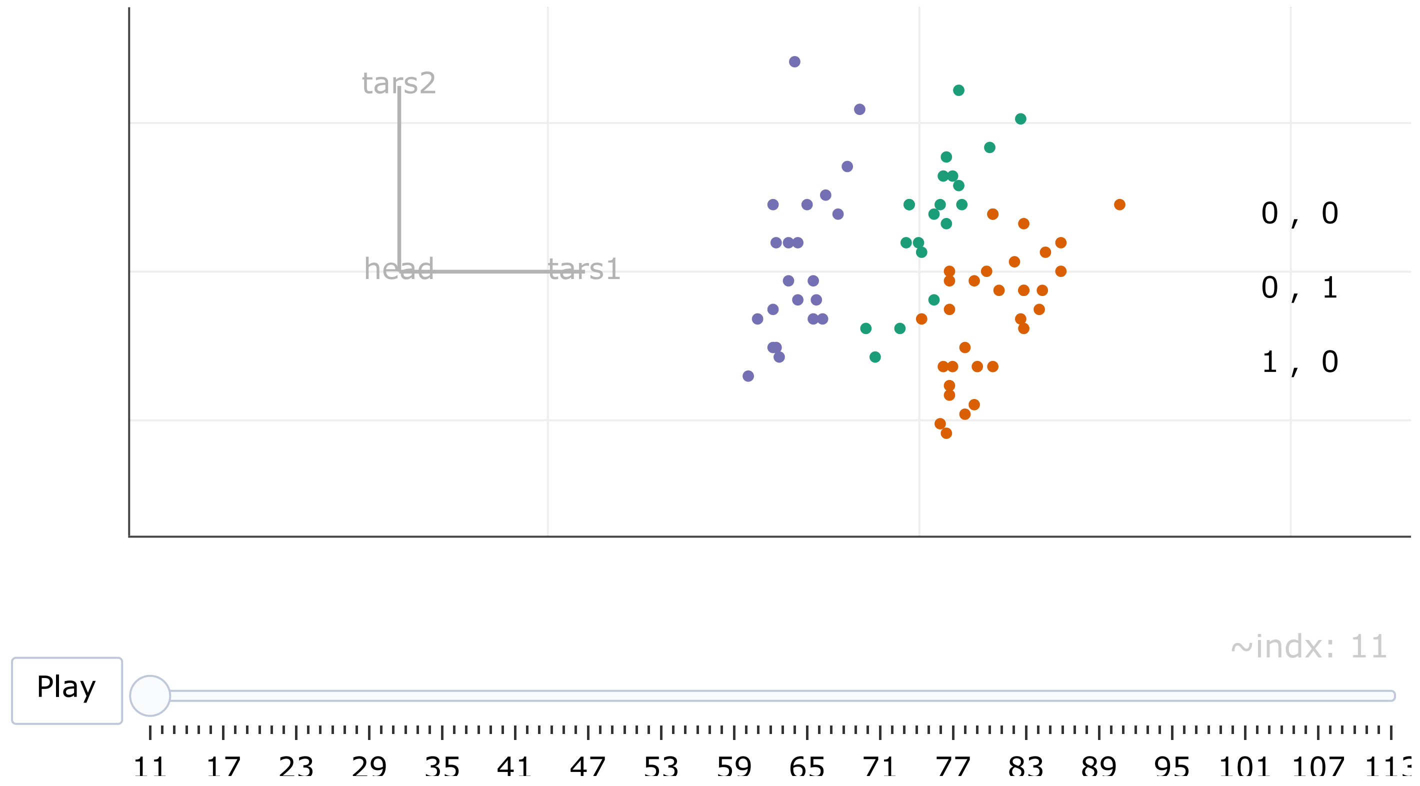
Notation

- $\mathbf{x}_i \in \mathcal{R}^p$, i^{th} data vector
- d projection dimension
- F is a $p \times d$ orthonormal basis, $F'F = I_d$
- The projection of $\mathbf{x}$ onto F is $\mathbf{y}_i = F'\mathbf{x}_i$.
- Its a movie, so the tour is indexed by time, $F(t)$, where $t \in [a, z]$. Starting and target frame denoted as $F_a = F(a), F_z = F(t)$.
- The animation of the projected data is given by a path $\mathbf{y}_i(t) = F'(t)\mathbf{x}_i$.



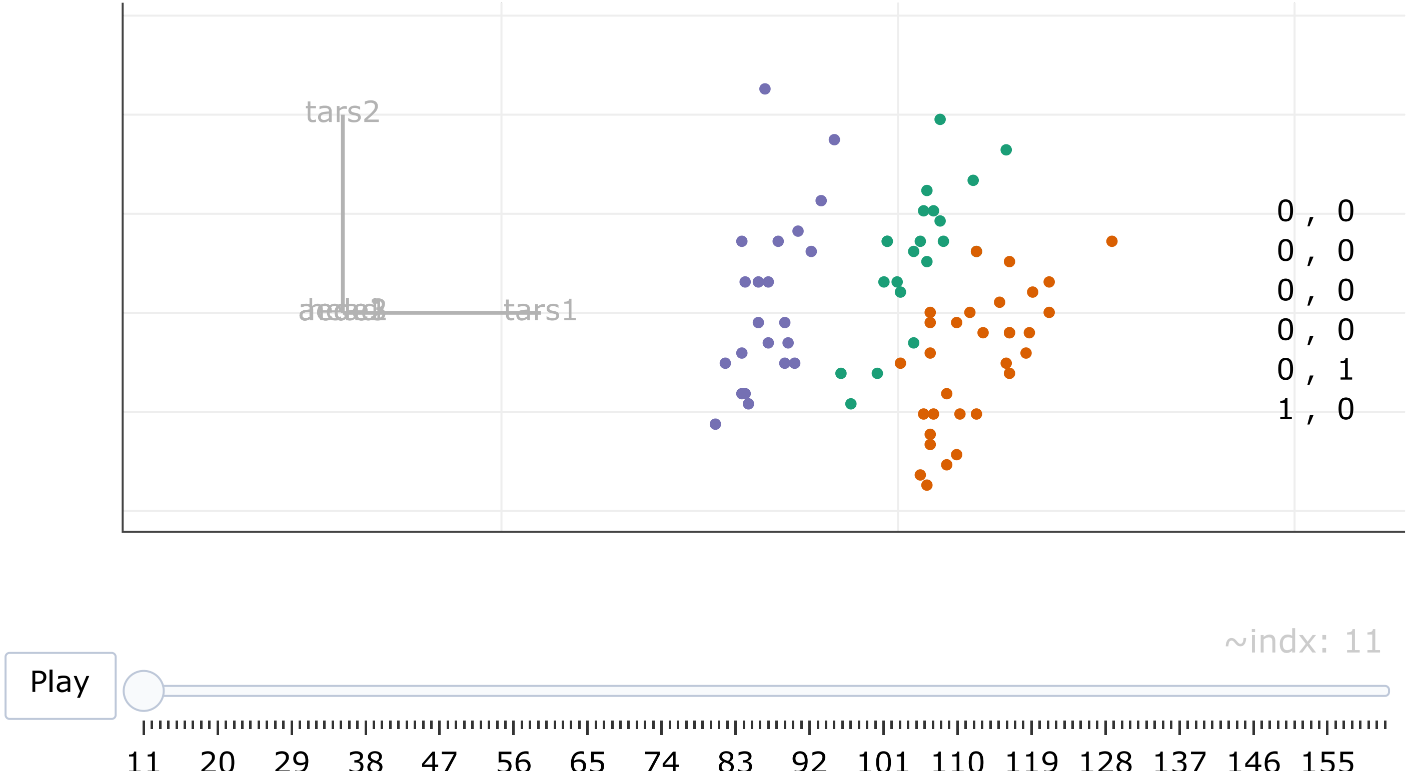
Examples

$p = 3, d = 2$



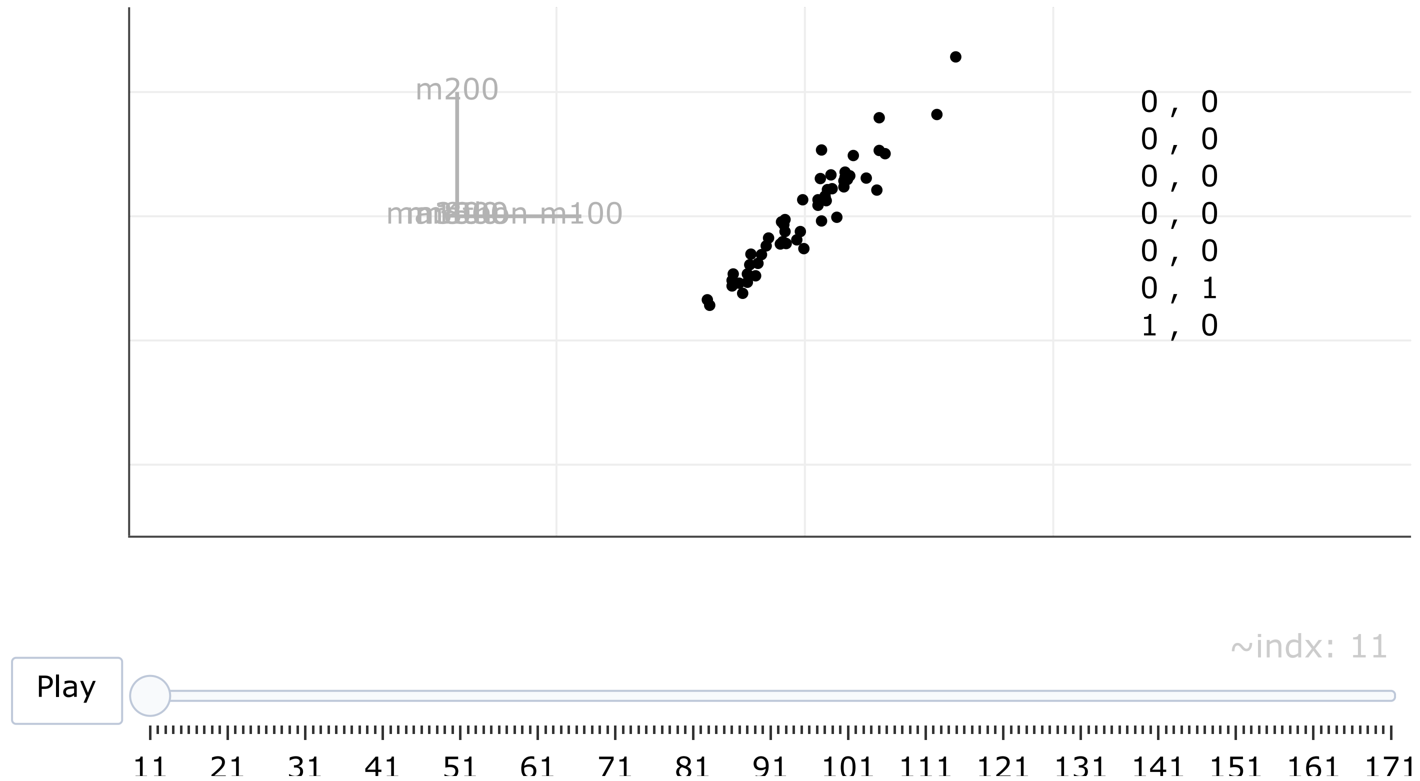
Examples

$p = 6, d = 2$



Examples

$p = 7, d = 2$



Examples

With the grand tour, you can get a good overall sense of the distribution (shape) of the data in its p-dimensional space:

- In the first data set, the primary shape are three well separated clusters
- In the track data, the primary shape is that it lives in essentially a 1-D subspace, with a small amount of variation in other directions. It is also possible to see several outliers.

Guided tours

Remember: projection pursuit

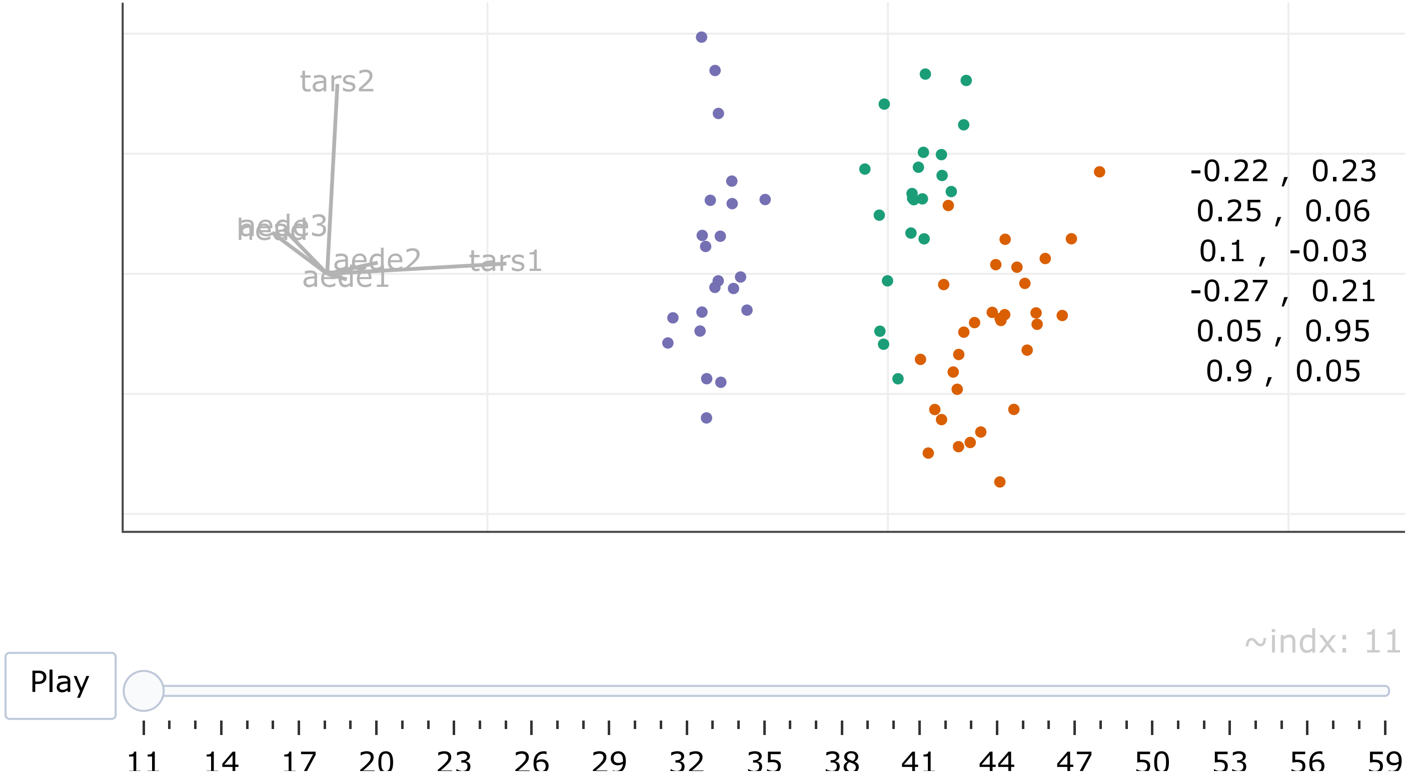
$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} f\left(\sum_{j=1}^{p} \phi_{j1} x_{ij}\right) \text{ subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1$$

The guided tour chooses new target projections by optimising a function of interest:

- `holes`: This is an inverse Gaussian filter, which is optimised wheren there is not much data in the center of the projection, i.e. a "hole" or donut shape in 2D.
- `central mass`: The opposite of holes, high density in the centre of the projection, and often "outliers" on the edges.
- `LDA`: An index based on the linear discrimination dimension reduction, optimised by projections where the named classes are most separated.

Guided tours

$p = 6, d = 2$, guidance using the LDA index



Usage - in R

To run a tour live:

```r
library(tourr)
# On a Mac use this graphics device
# quartz()
animate_xy(flea[, 1:6])
animate(flea[, 1:6], tour_path=grand_tour(),
  display=display_xy(axes = "bottomleft"))
library(RColorBrewer)
pal <- brewer.pal(3, "Dark2")
col <- pal[as.numeric(flea$species)]
animate_xy(flea[,-7], col=col)
```

Others

- Little: Interpolates between all possible pairs of variables. Like the scatterplot matrix, but animated between them.
- Local: Rocks back and forth from a given projection, so shows all possible projections within a radius.
- Frozen: Fixes some of the values of the orthonormal projection matrix and allows the others to vary freely according to any of the other tour methods.
- Manual: Control the contribution of a single variable, and move along this axis. This is really useful to examine the sensitivity of structure (e.g. clustering) to the contribution of a variable. Maybe the variable can be "zero'd out" and the structure would not be affected, thus simplifying the "model". This is available in the spinifex package.

Rendering

The projection dimension d can be 1, 2, 3, ... It is just a projection of the data, and then you need to decide how to *render* the data.

- *d=1* : The projected data can be displayed as a dotplot, density, histogram, boxplot, ...
- *d>2* : Use stereo (for *d=3*) or a scatterplot matrix (or parallel coordinate plot)

Lecturer: *Professor Di Cook*

Department of Econometrics and Business Statistics

ETC3250.Clayton-x@monash.edu

Week 5a