MONASH University

ETC5521: Exploratory Data Analysis

Going beyond two variables, exploring high dimensions

Lecturer: *Di Cook*

ETC5521.Clayton-x@monash.edu

Week 8 - Session 2

This lecture

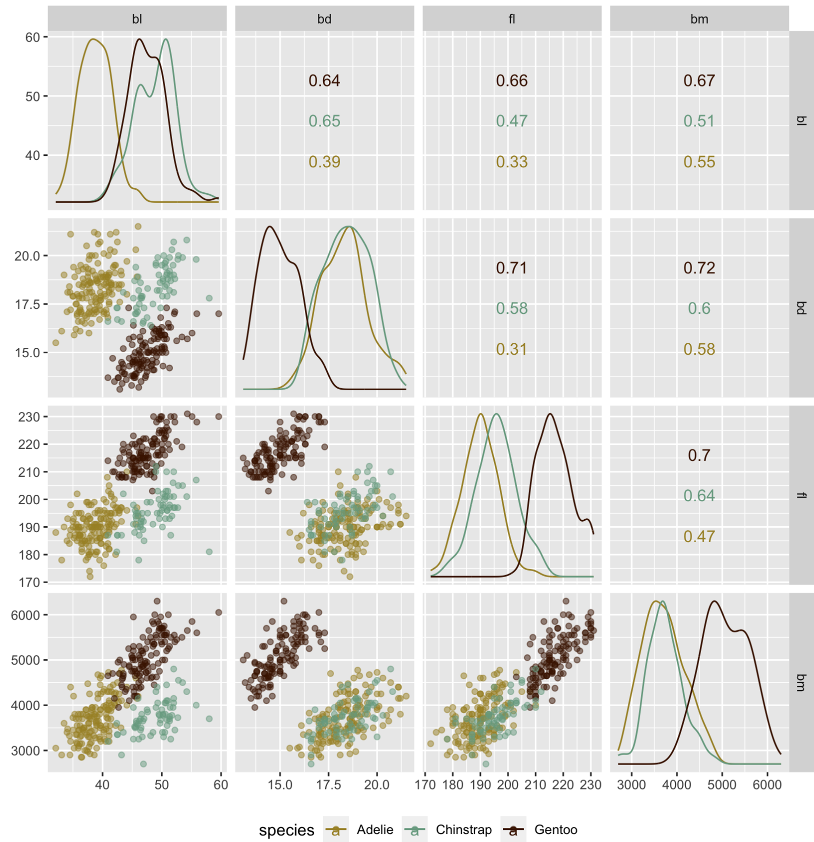
linked brushing between plots
parallel rather than orthogonal axes
tours (rotations) through high-dimensions

Case study 4 Penguins

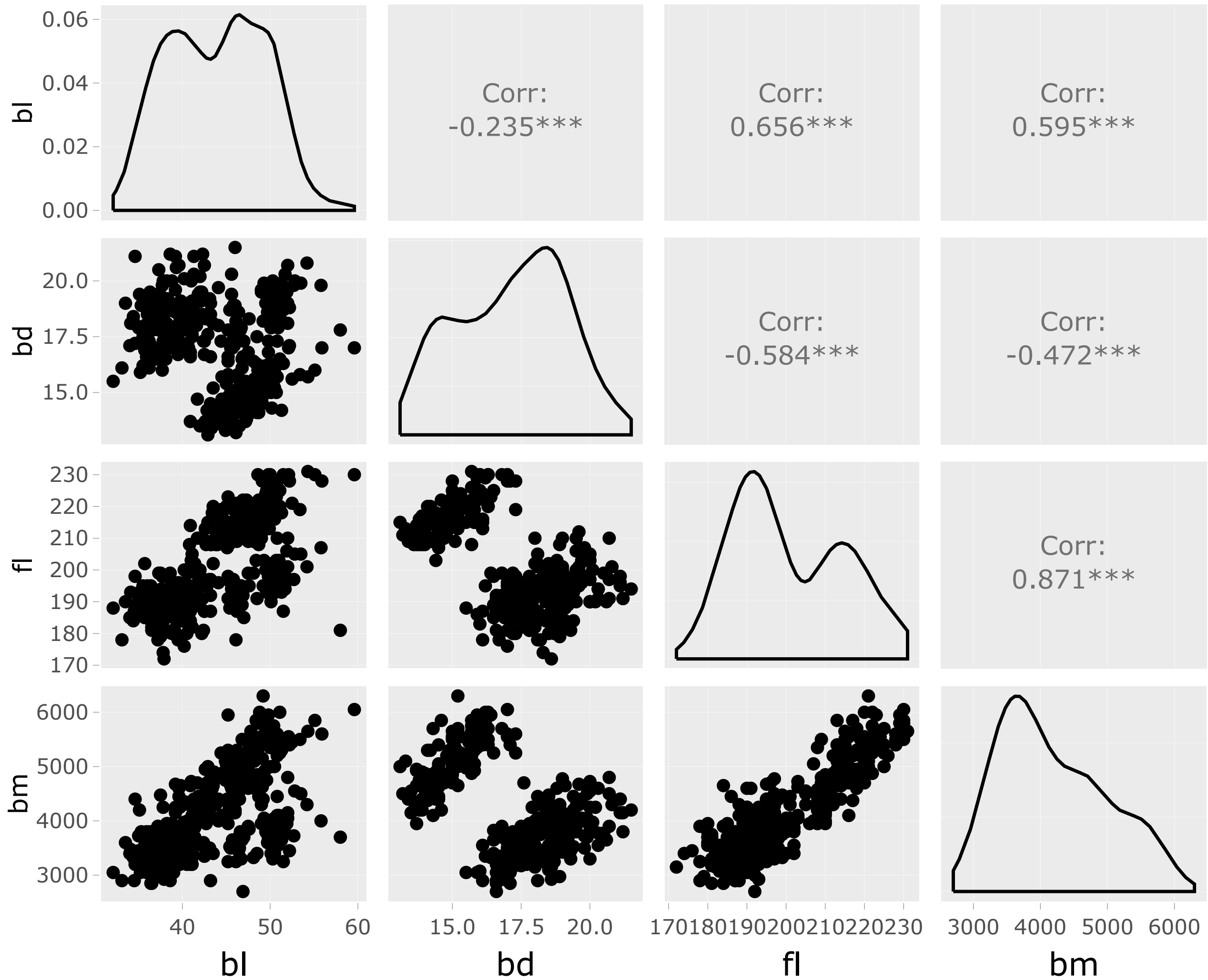
R

Size measurements for adult foraging penguins near Palmer Station, Antarctica

- species: a factor denoting penguin species (Adélie, Chinstrap and Gentoo)
- bill_length_mm: a number denoting bill length (millimeters)
- bill_depth_mm: a number denoting bill depth (millimeters)
- flipper_length_mm: an integer denoting flipper length (millimeters)
- body_mass_g:an integer denoting body mass (grams)



Linking between plots



If you have interactive plots you can investigate whether

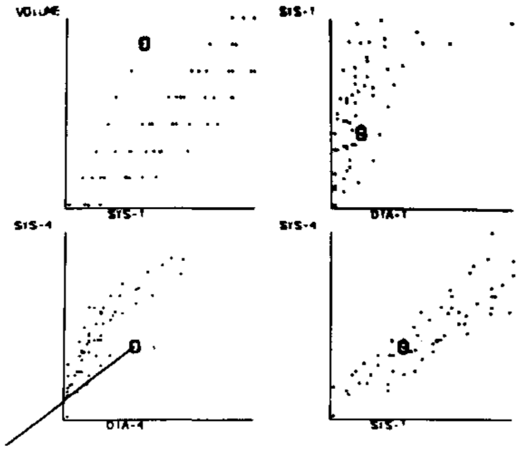
- an outlier in one or two variables is one of the extremes in other variables ⟶ then it is a multivariate outlier
- clusters of observations in one plot are concentrated in a cluster in other variables ⟶ the data is multimodal in multivariate space, and there are likely sub-populations.

Your turn, cut and paste the code into your R console, and select regions of the resulting plot to examine where these points lie in other plots.

```
# Load the penguins data with code from previous slide
library(tidyverse)
library(tourr)
library(plotly)
highlight_key(penguins) %>%
  GGally::ggpairs(aes(color = species),
                  columns = 3:6) %>%
  ggplotly() %>%
  highlight("plotly_selected")
```

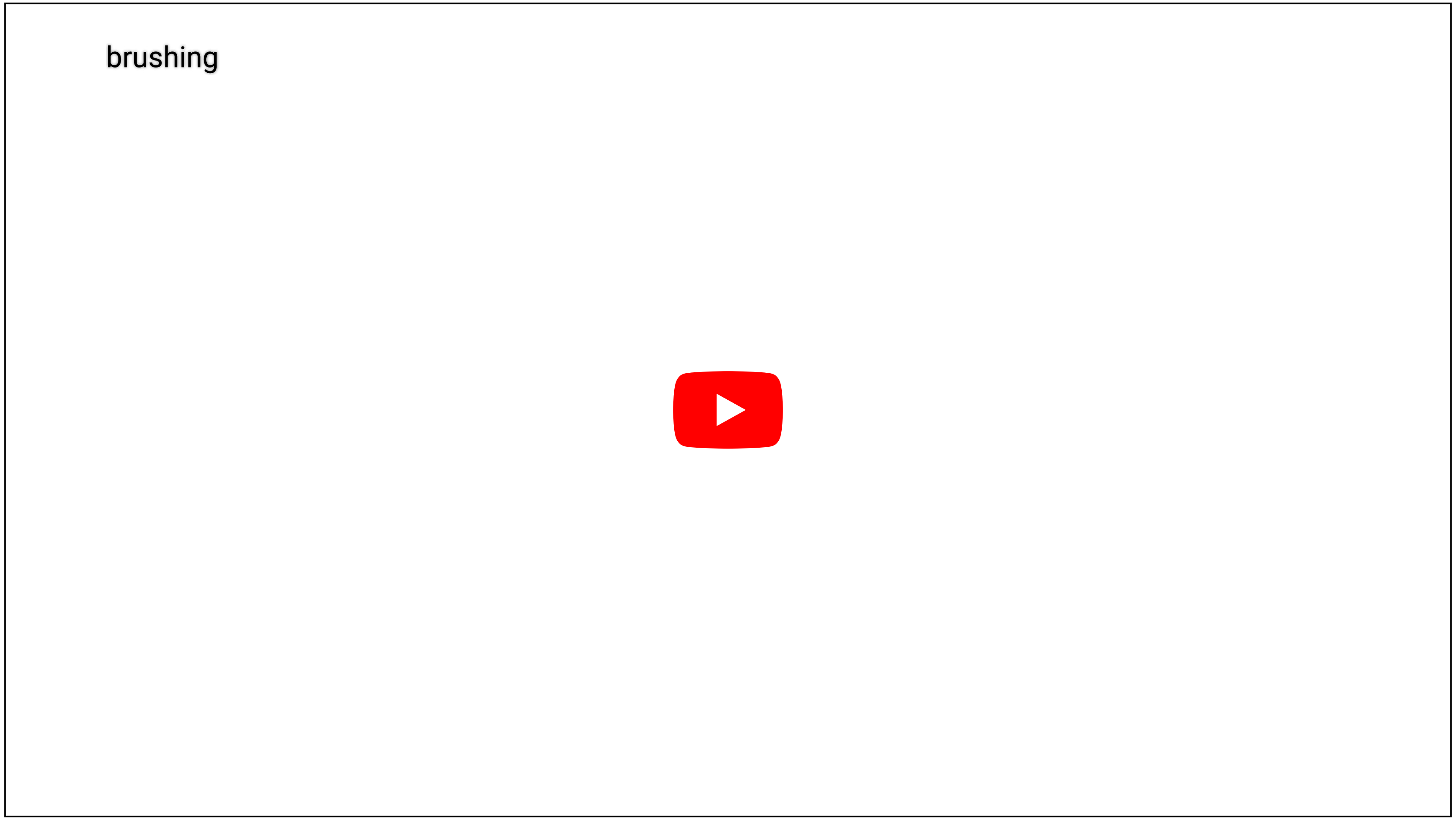
05:00

History



<u>Figure 2</u>: The user has designated a point on one graph by means of the light-pen. It and corresponding points on the other graphs then are identified by a circle, and the values of all variables for that case are as follows:

First linked brushing across a set of scatterplots done by Carol Newton (1978) "Graphics from Alpha to Omega in Data Analysis" in Graphical Representation of Multivariate Data.

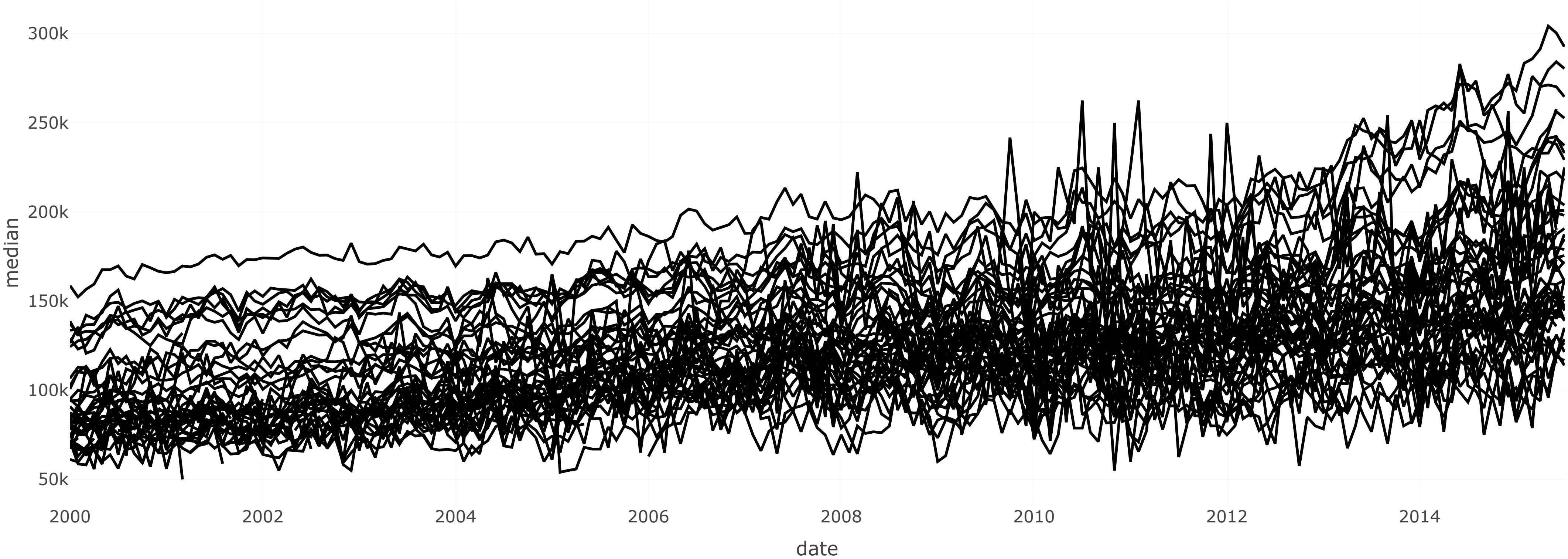


Brushing on a scatterplot matrix, Rick Becker (Becker and Cleveland, 1987 Brushing Scatterplots, Technometrics)

Parallel coordinate plots

Another way to display multivariate display

You can think about parallel coordinate plots like time series plots. Each variable is treated as a parallel axis. Observations are drawn as a line across the axes.



Median house price by date for a number of cities in Texas.

learn R

What can you learn?

- lines that are parallel indicate positive linear association, possibly across many variables
- lines that cross indicate negative linear association
- lines that go up and down differently from any other lines indicate multivariate outliers
- lines that go up and down together, but differently from other lines indicate multivariate clustering

You need to have an interactive parallel coordinate plot for them to be effective for exploring data

Your turn, cut and paste the code into your R console, and click in the resulting plot to examine the line for an observation.

```
# Using the same penguins data subset as earlier in the slides
library(shiny)
library(plotly)
library(tidyverse)
ui <- fluidPage(
  plotlyOutput("parcoords"),
  verbatimTextOutput("data"))

server <- function(input, output, session) {
  penguins_numeric <- penguins[,3:6] %>%
    na.omit()

  output$parcoords <- renderPlotly({
    dims <- Map(function(x, y) {
      list(values = x, range = range(x), label = y)
```

05:00

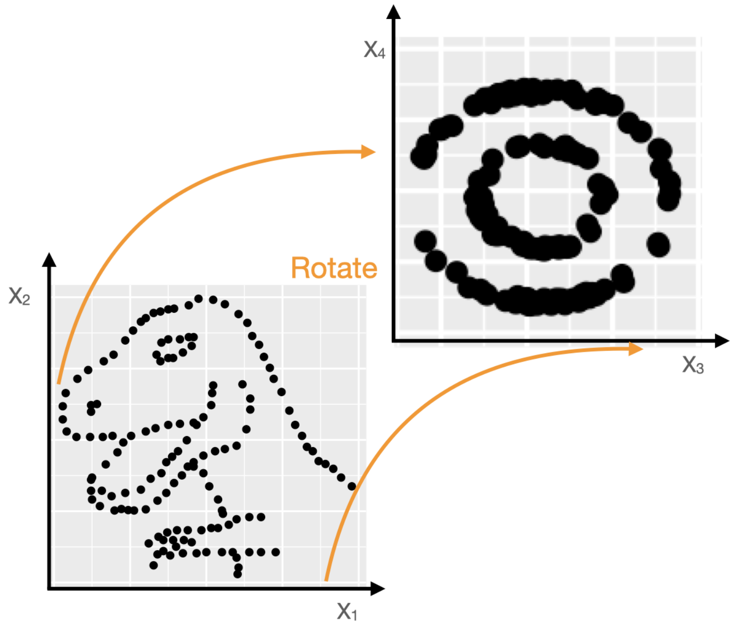
Dynamic graphics: tours

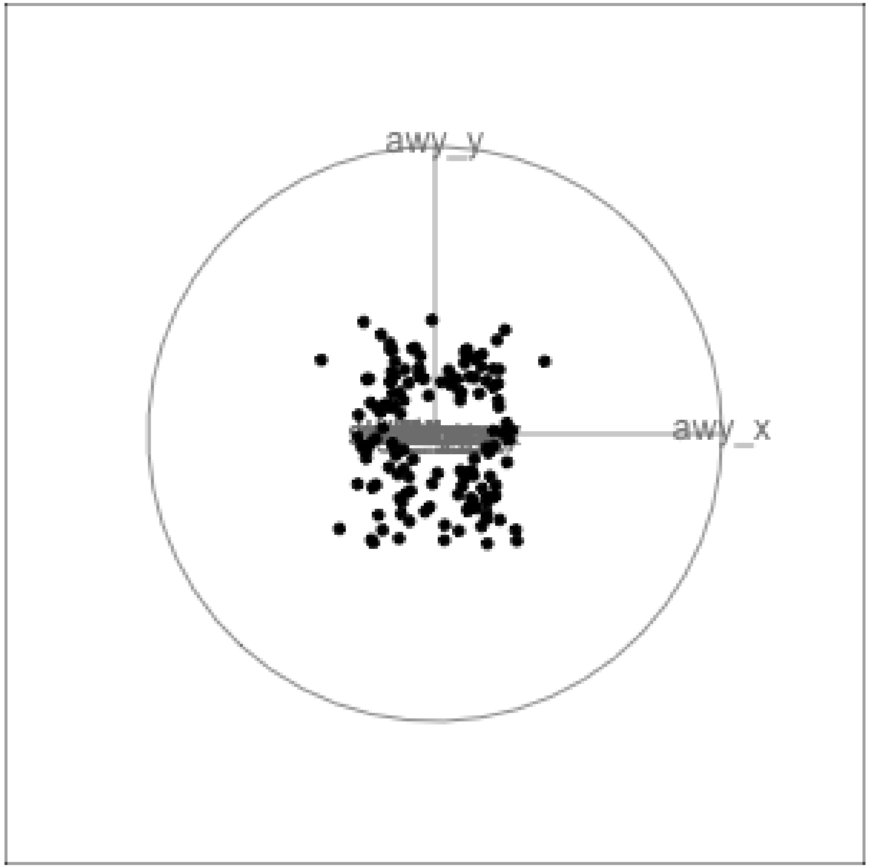
Remember Tukey's PRIM-9?? This is the evolution of that, and allows us to look at all possible linear combination of variables.

Touring between pairs of variables

Dinosaurus dozen in wide form as multivariate data.

Rotate one pair of variables into the other - special type of interpolation/animation called a *little tour*





Tours of multivariate data

We look at combinations of variables, as well as the individual variables. The grand tour does this in an elegant way so that there's a chance of seeing all possible combinations of the variables, and it glues these together so that there is a smooth change from one to another. With an important note: that always when scatterplots (or higher dimensional combinations are shown) that the axes are always orthogonal.

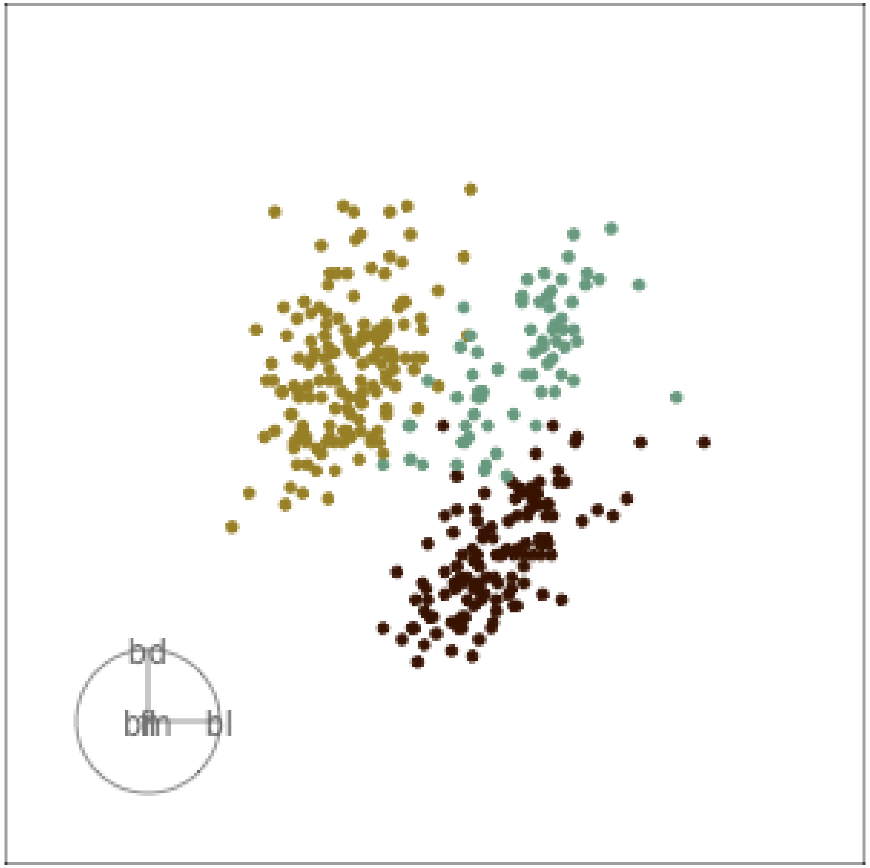
Using combinations of variables, it can be possible to see separations between groups, outliers, linear and non-linear dependencies that were not visible in the single variable plots.

3D plots aren't enough. You need tours to find unusual multiple variable relationsips.

Our first grand tour

Here's the code

```
clrs <- ochre_pal(
  palette="nolan_ned")(3)
col <- clrs[
  as.numeric(
    penguins$species)]
animate_xy(penguins[,3:6],
           col=col,
           axes="off")
```



Case study 4 Penguins

What do you learn about this data?

- clusters ✅
- outliers ✅
- linear dependence ✅
- elliptical clusters with slightly different shapes ✅
- separated elliptical clusters with slightly different shapes ✅

What is a tour?

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.

$\mathbf{x}_i \in \mathrm{R}^p$, i^{th} data vector

F is a $p \times d$ orthonormal matrix, $F'F = I_d$, where d is the projection dimension.

The projection of $\mathbf{x}_i$ onto F is $\mathbf{y}_i = F'\mathbf{x}_i$.

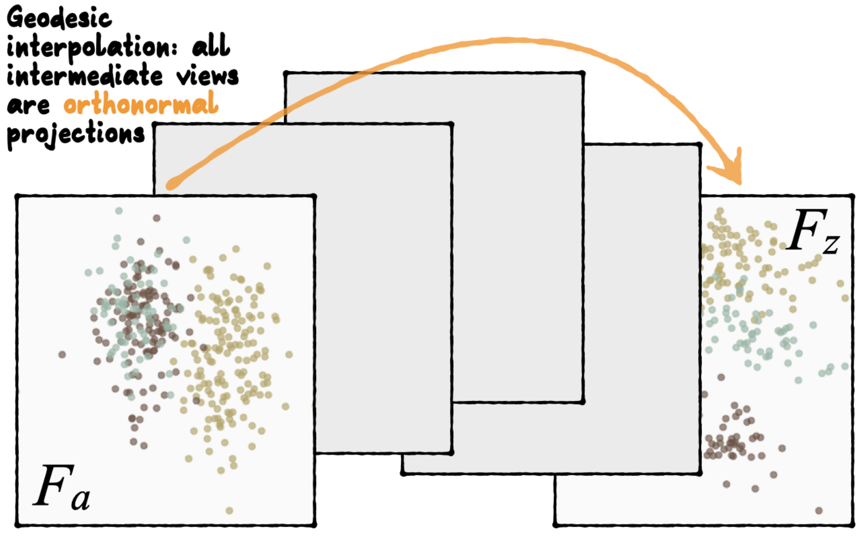
Tour is indexed by time, $F(t)$, where $t \in [a, z]$. Starting and target frame denoted as $F_a = F(a), F_z = F(t)$.

The animation of the projected data is given by a path $\mathbf{y}_i(t) = F'(t)\mathbf{x}_i$.

Geodesic interpolation between planes

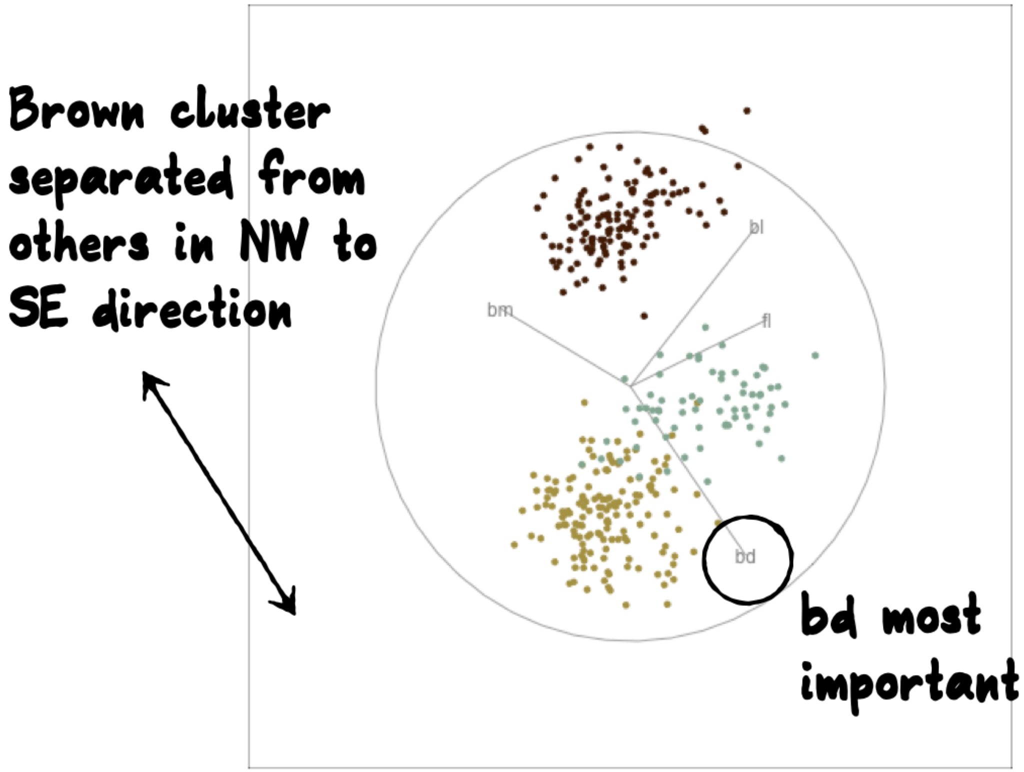
Tour is indexed by time, $F(t)$, where $t \in [a, z]$.
Starting and target frame denoted as
$F_a = F(a), F_z = F(t)$.

The animation of the projected data is given by a path
$\mathbf{y}_i(t) = F'(t)\mathbf{x}_i$.

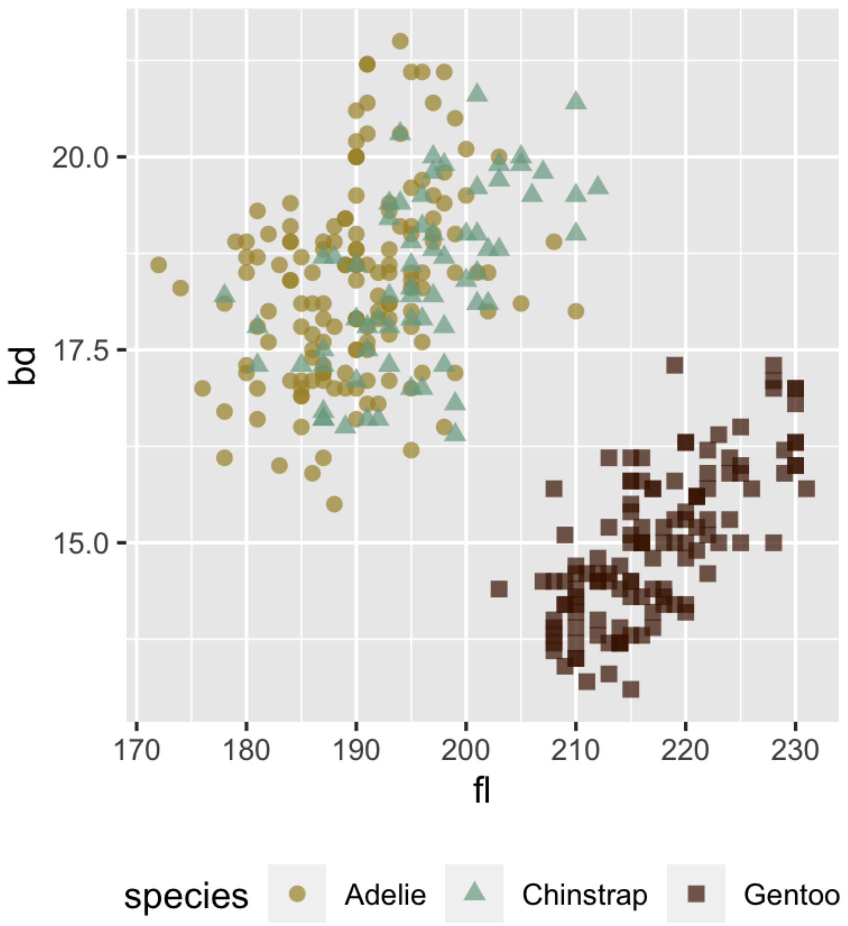


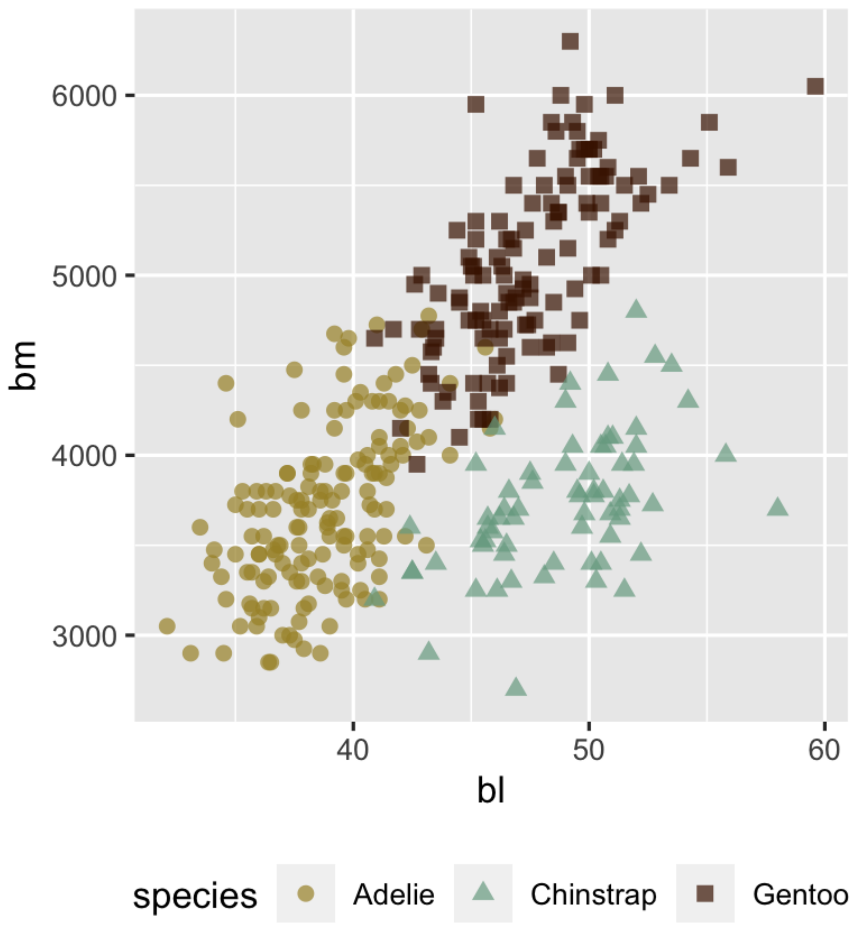
Reading axes - interpretation

Length and direction of axes relative to the pattern of interest



Case study 4 Penguins





Gentoo from others in contrast of fl, bd

Chinstrap from others in contrast of bl, bm

There may be multiple and different combinations of variables that reveal similar structure. ☹

The tour can help to discover these, too. 😂

Other tour types

- guided: follows the optimisation path for a projection pursuit index.
- little: interpolates between all variables.
- local: rocks back and forth from a given projection, so shows all possible projections within a radius.
- dependence: two independent 1D tours
- frozen: fixes some variable coefficients, others vary freely.
- manual: control coefficient of one variable, to examine the sensitivity of structure this variable. (In the spinifex package)
- slice: use a section instead of a projection.

Guided tour

New target bases are chosen using a projection pursuit index function

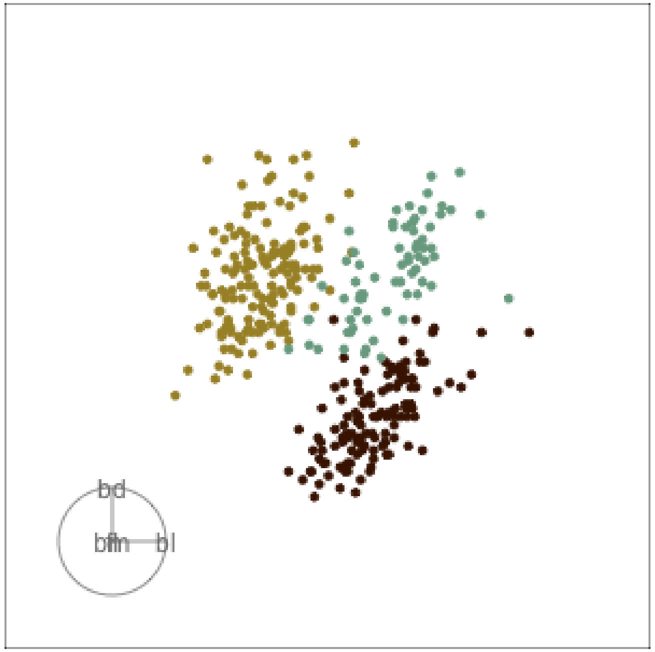
$$\underset{F}{\text{maximize}}\, g(F'x) \quad \text{subject to } F \text{ being orthonormal}$$

- `holes`: This is an inverse Gaussian filter, which is optimised when 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`/`PDA`: An index based on the linear discriminant dimension reduction (and penalised), optimised by projections where the named classes are most separated.

Case study 4 Penguins

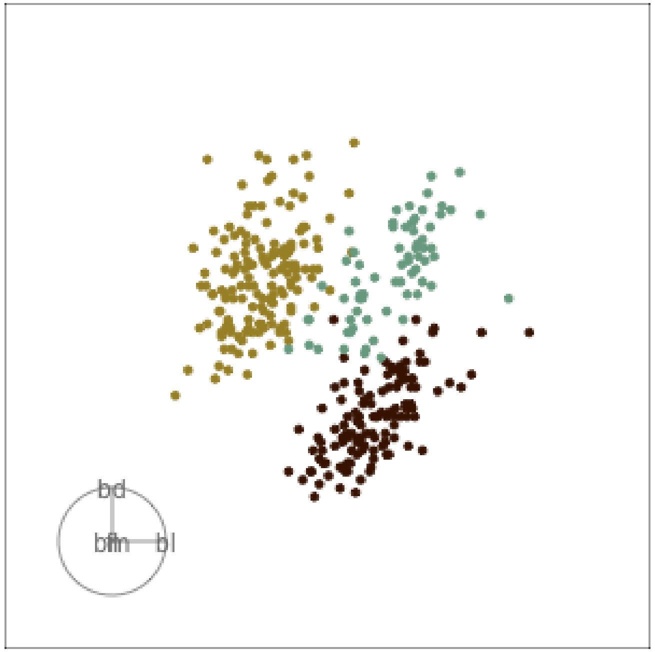
R

Grand



Might accidentally see best separation

Guided, using LDA index



Moves to the best separation

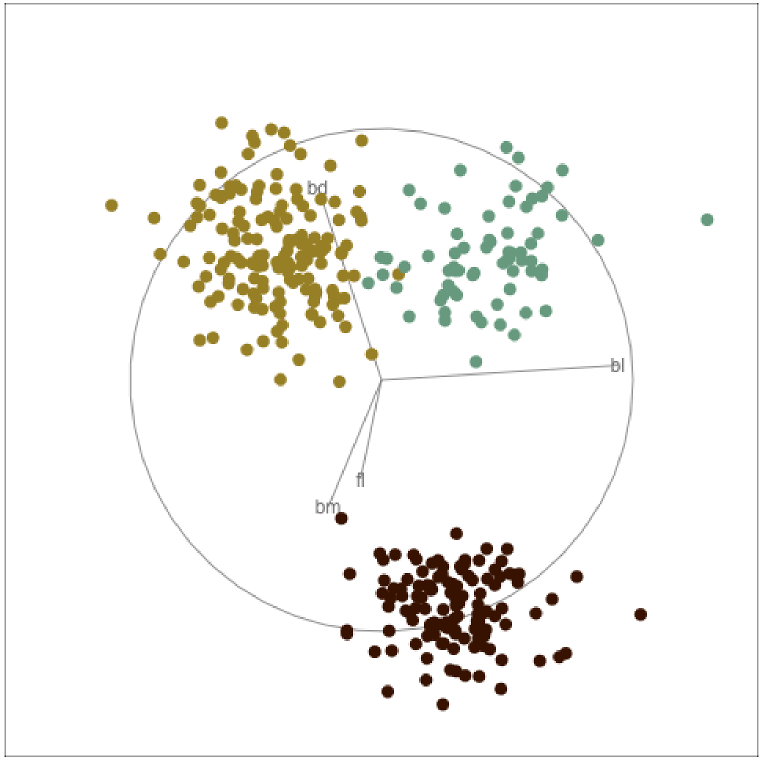
Manual tour

Control the coefficient of one variable, reduce it to zero, then increase it to 1, maintaining orthonormality

Case study 4 Penguins

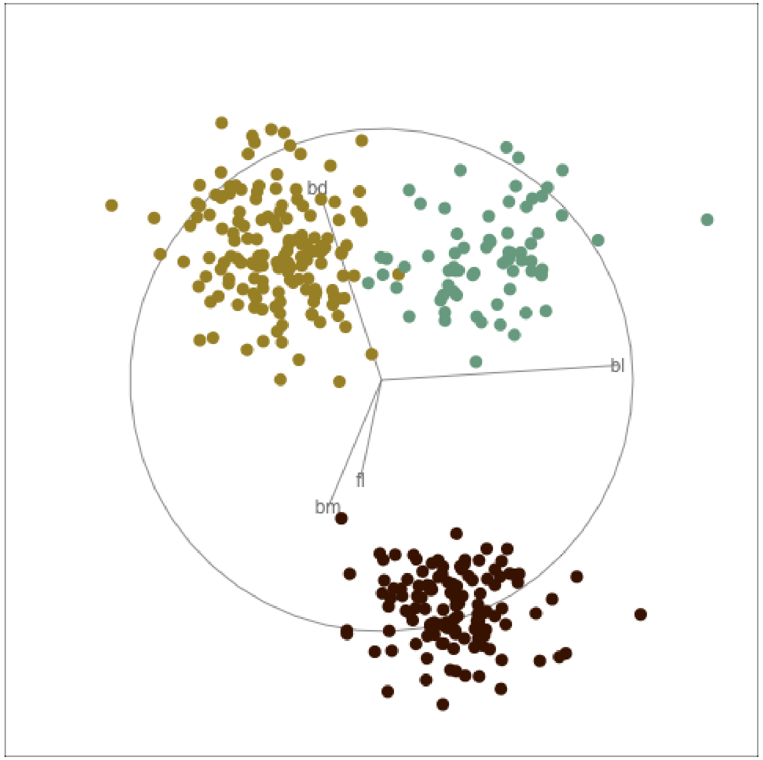
🖼️ R

- start from best projection, given by projection pursuit
- `bl` contribution controlled
- if `bl` is removed from projection, Adelie and chinstrap are mixed
- `bl` is important for Adelie



Case study 4 Penguins

- start from best projection, given by projection pursuit
- `fl` contribution controlled
- cluster less separated when fl is fully contributing
- `fl` is important, in small amounts, for Gentoo



Resources

- Wickham et al (2011). tourr: An R Package for Exploring Multivariate Data with Projections. http://www.jstatsoft.org/v40/i02/.
- Cook and Laa (2023) Interactively exploring high-dimensional data and models in R, https://dicook.github.io/mulgar_book/
- Sievert (2019) Interactive web-based data visualization with R, plotly, and shiny, https://plotly-r.com
- Horst et al (2020 https://allisonhorst.github.io/palmerpenguins/
- Gorman KB, Williams TD, Fraser WR (2014) Ecological Sexual Dimorphism and Environmental Variability within a Community of Antarctic Penguins (Genus Pygoscelis). PLoS ONE 9(3): e90081. doi:10.1371/journal.pone.0090081

Lecturer: *Di Cook*

ETC5521.Clayton-x@monash.edu

Week 8 - Session 2