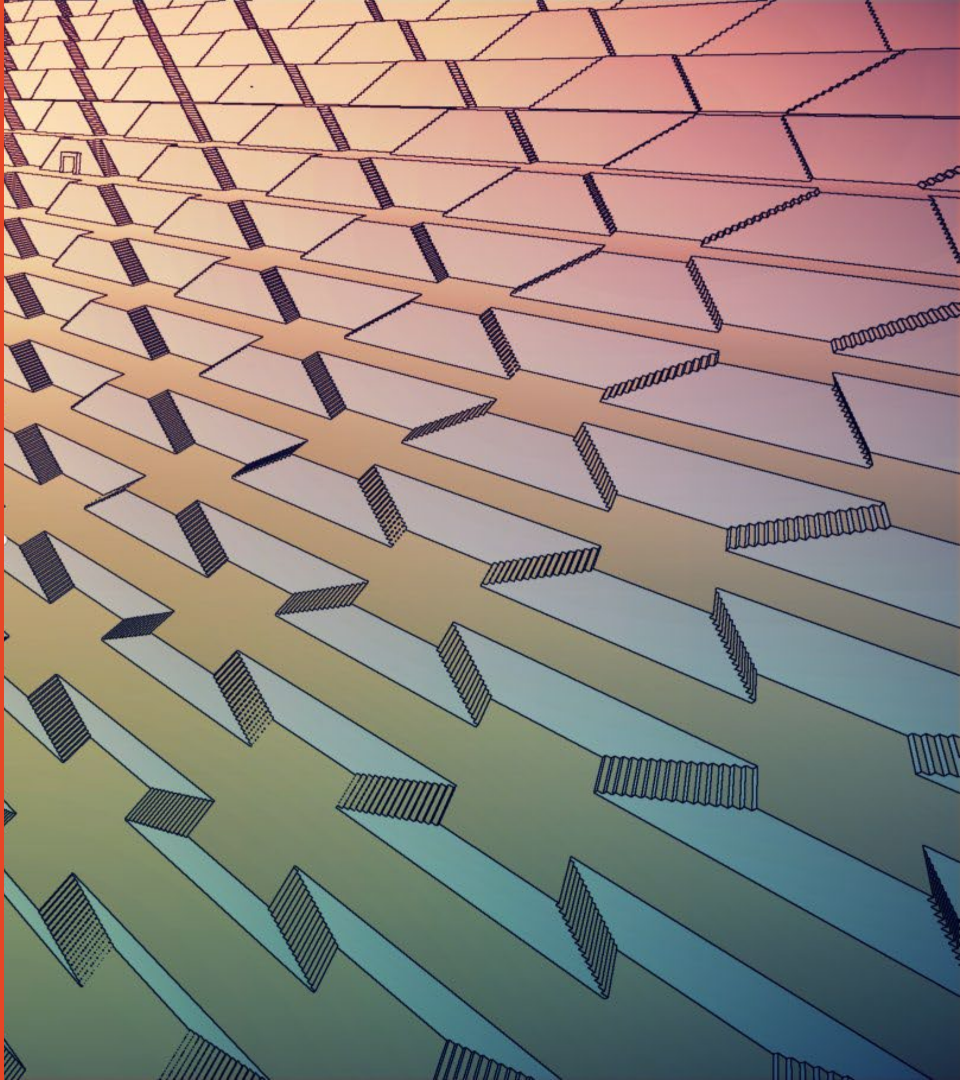
Multivariate Statistical Analysis 1: Dimension Reduction Methods

Presented by
Alex Shaw
Sydney Informatics Hub
Core Research Facilities
The University of Sydney



Acknowledging SIH

All University of Sydney resources are available to Sydney researchers **free of charge**. The use of the SIH services including the Artemis HPC and associated support and training warrants acknowledgement in any publications, conference proceedings or posters describing work facilitated by these services.

The continued acknowledgment of the use of SIH facilities ensures the sustainability of our services.

Suggested wording for use of workshops and workflows:

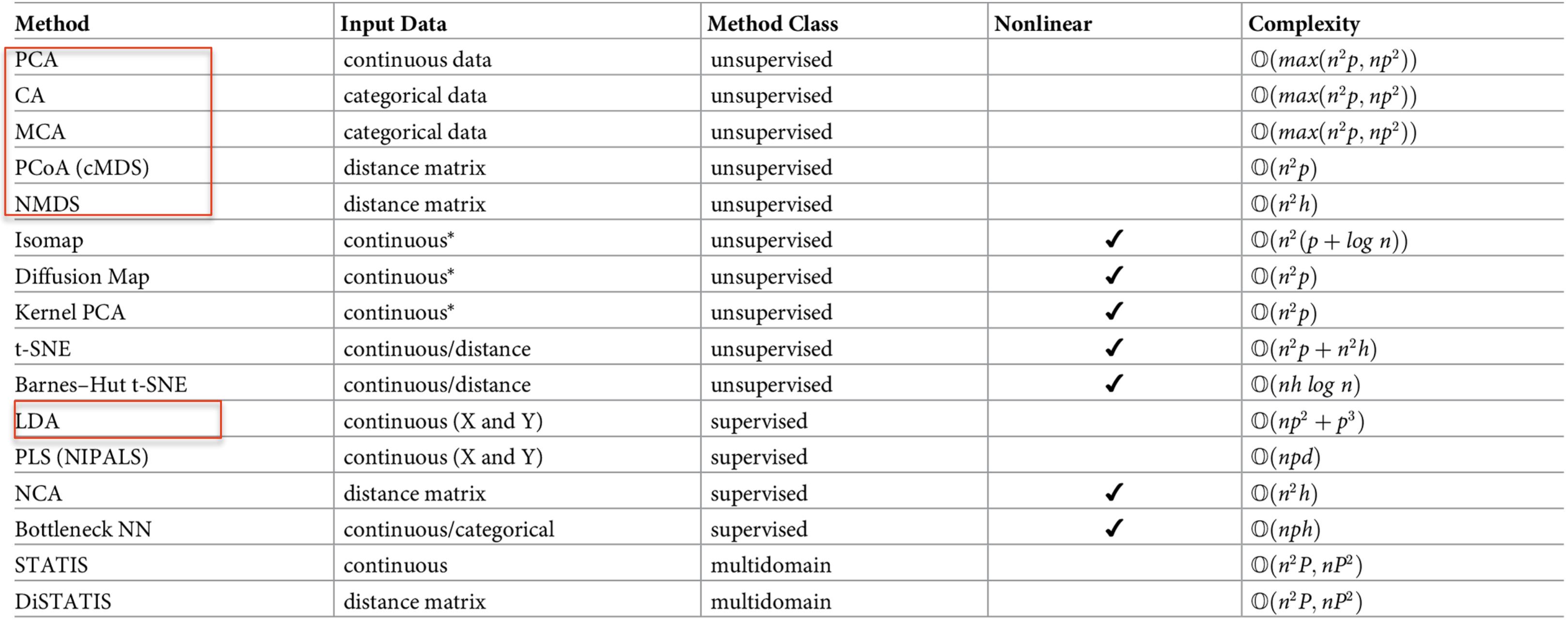
"The authors acknowledge the Statistical workshops and workflows provided by the Sydney Informatics Hub, a Core Research Facility of the University of Sydney."

How to use these slides

- These slides are additional material to help you understand the use of the software workflows

- Eventually all the material contained in these slides will be incorporated into the software workflows

Dimension reduction methods

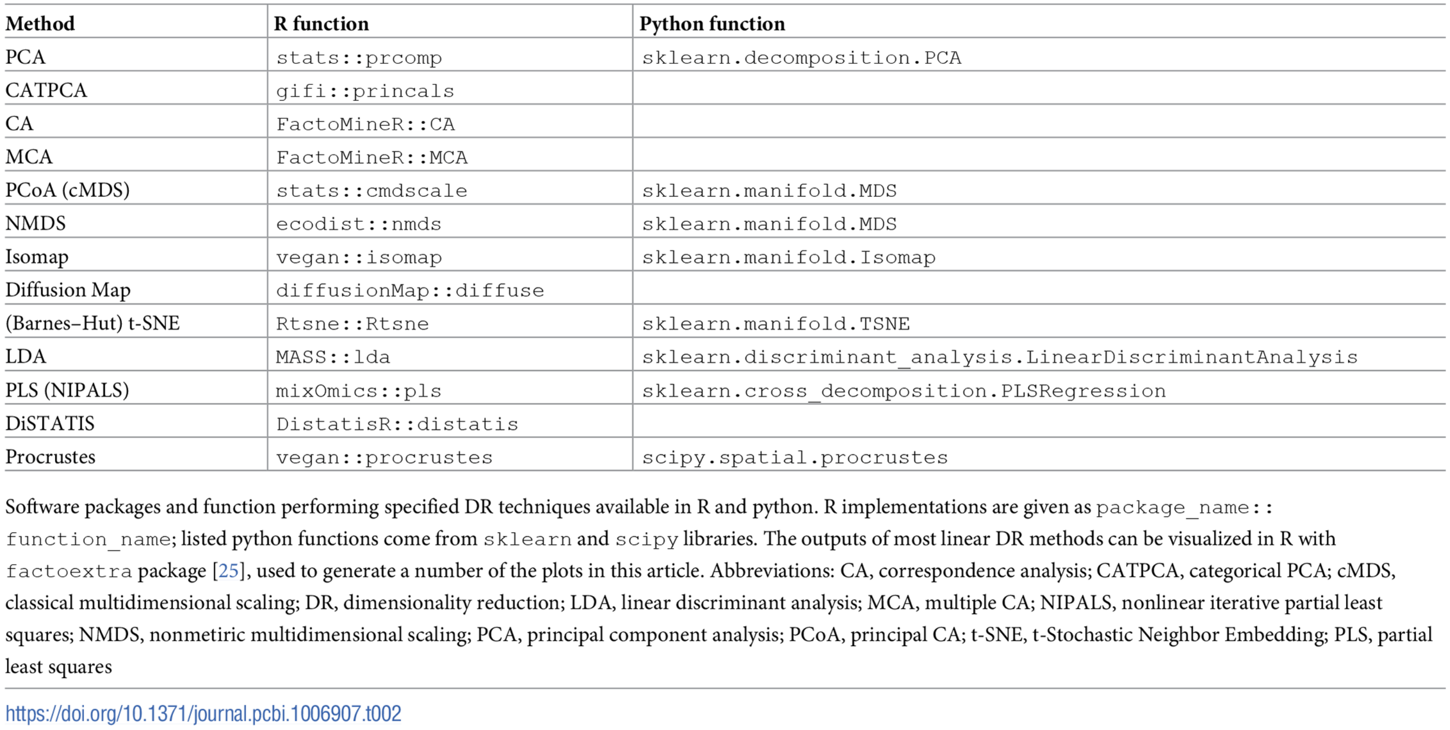


Method	Input Data	Method Class	Nonlinear	Complexity
PCA	continuous data	unsupervised		$\mathbb{O}(max(n^2p, np^2))$
CA	categorical data	unsupervised		$\mathbb{O}(max(n^2p, np^2))$
MCA	categorical data	unsupervised		$\mathbb{O}(max(n^2p, np^2))$
PCoA (cMDS)	distance matrix	unsupervised		$\mathbb{O}(n^2p)$
NMDS	distance matrix	unsupervised		$\mathbb{O}(n^2h)$
Isomap	continuous*	unsupervised	✓	$\mathbb{O}(n^2(p + log\ n))$
Diffusion Map	continuous*	unsupervised	✓	$\mathbb{O}(n^2p)$
Kernel PCA	continuous*	unsupervised	✓	$\mathbb{O}(n^2p)$
t-SNE	continuous/distance	unsupervised	✓	$\mathbb{O}(n^2p + n^2h)$
Barnes–Hut t-SNE	continuous/distance	unsupervised	✓	$\mathbb{O}(nh\ log\ n)$
LDA	continuous (X and Y)	supervised		$\mathbb{O}(np^2 + p^3)$
PLS (NIPALS)	continuous (X and Y)	supervised		$\mathbb{O}(npd)$
NCA	distance matrix	supervised	✓	$\mathbb{O}(n^2h)$
Bottleneck NN	continuous/categorical	supervised	✓	$\mathbb{O}(nph)$
STATIS	continuous	multidomain		$\mathbb{O}(n^2P, nP^2)$
DiSTATIS	distance matrix	multidomain		$\mathbb{O}(n^2P, nP^2)$

Mentioned in this workshop

Nguyen LH, Holmes S (2019) Ten quick tips for effective dimensionality reduction. PLOS Computational Biology 15(6): e1006907. https://doi.org/10.1371/journal.pcbi.1006907
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006907

Table 2. Example implementations.



Method	R function	Python function
PCA	`stats::prcomp`	`sklearn.decomposition.PCA`
CATPCA	`gifi::princals`	
CA	`FactoMineR::CA`	
MCA	`FactoMineR::MCA`	
PCoA (cMDS)	`stats::cmdscale`	`sklearn.manifold.MDS`
NMDS	`ecodist::nmds`	`sklearn.manifold.MDS`
Isomap	`vegan::isomap`	`sklearn.manifold.Isomap`
Diffusion Map	`diffusionMap::diffuse`	
(Barnes–Hut) t-SNE	`Rtsne::Rtsne`	`sklearn.manifold.TSNE`
LDA	`MASS::lda`	`sklearn.discriminant_analysis.LinearDiscriminantAnalysis`
PLS (NIPALS)	`mixOmics::pls`	`sklearn.cross_decomposition.PLSRegression`
DiSTATIS	`DistatisR::distatis`	
Procrustes	`vegan::procrustes`	`scipy.spatial.procrustes`

Software packages and function performing specified DR techniques available in R and python. R implementations are given as `package_name::function_name`; listed python functions come from `sklearn` and `scipy` libraries. The outputs of most linear DR methods can be visualized in R with `factoextra` package [25], used to generate a number of the plots in this article. Abbreviations: CA, correspondence analysis; CATPCA, categorical PCA; cMDS, classical multidimensional scaling; DR, dimensionality reduction; LDA, linear discriminant analysis; MCA, multiple CA; NIPALS, nonlinear iterative partial least squares; NMDS, nonmetiric multidimensional scaling; PCA, principal component analysis; PCoA, principal CA; t-SNE, t-Stochastic Neighbor Embedding; PLS, partial least squares

https://doi.org/10.1371/journal.pcbi.1006907.t002

Nguyen LH, Holmes S (2019) Ten quick tips for effective dimensionality reduction. PLOS Computational Biology 15(6): e1006907.
https://doi.org/10.1371/journal.pcbi.1006907
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006907

Dimensionality Reduction Techniques: PCA-like

Principal Component Analysis (PCA)

An example dataset: Decathlon

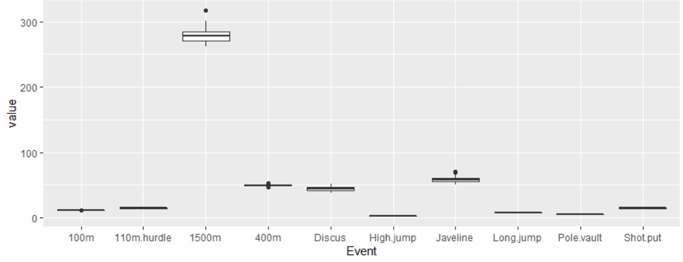
- Some of the concepts are best explained using a concrete example
- Let’s use the **decathlon** data set, which contains the performance of athletes in the 10 events of the decathlon
  - Track events: seconds to complete distance
  - Field events: distance travelled in metres
- So we have data from 41 performances across 10 variables
- Unless stated, all subsequent outputs are from this example

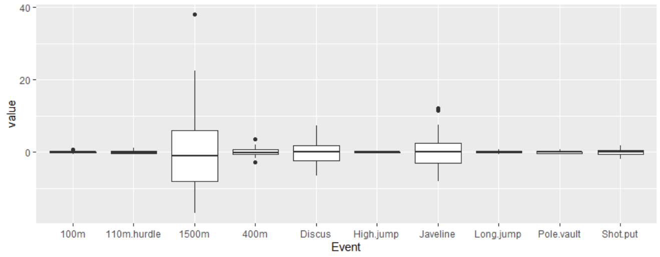
```
library(FactoMineR)
library(factoextra)
data(decathlon)
```

A workflow for dimension reduction

0. **Identify your variable types and perform appropriate Exploratory Data Analysis**
1. Run dimension reduction analysis
2. Examine the relationships between variables
3. Examine the relationships between subjects
4. Further summarising/interpretation. Choose how many dimensions to keep/examine.
5. Downstream analysis

Decathlon Dataset. Step 0: EDA





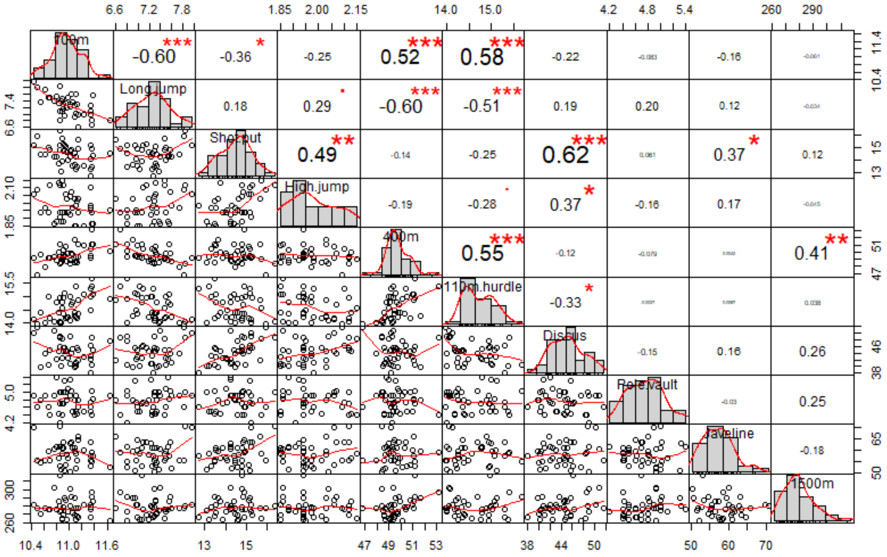
- We always mean centre the data (subtract the mean value from each variable) before PCA. Remember that we are using correlation, so the mean must be subtracted otherwise biases are introduced.

- It is also clear in this example that the different variables are on a different scale of measurement (different length track events with time in s, field events are distances in m)

- Beyond this, the variances for our variables are quite different to each other. We need to ‘rescale’ this data to get useful results (divide by the sample standard deviation).

```
data_long %>% ggplot(aes(x=Event,y=value)) + geom_boxplot()
data_long_centred %>% ggplot(aes(x=Event,y=value)) + geom_boxplot()
```

Decathlon Dataset. Step 0: EDA



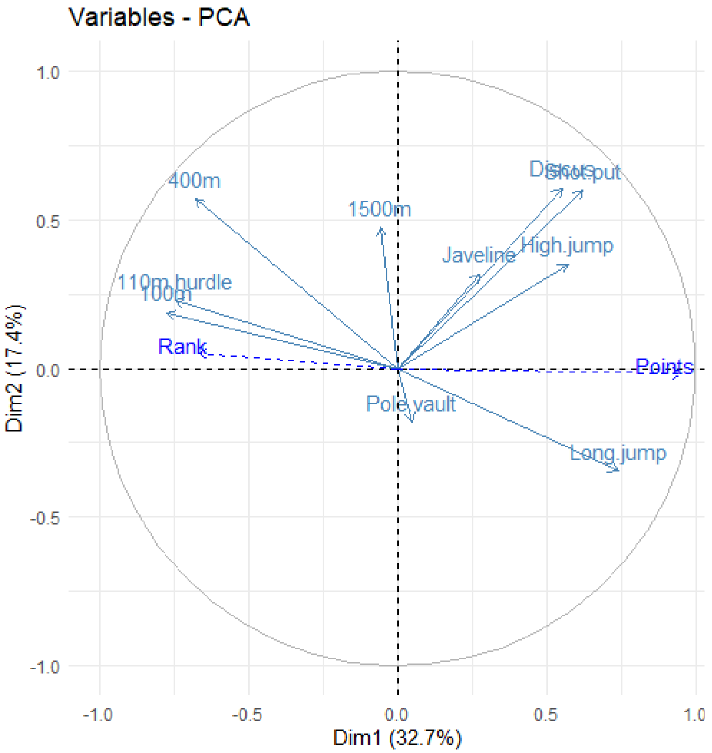
`chart.Correlation(data_wide %>% select(-Athlete))`

- Clear correlations between the similar field events, and similar track events
- Negative correlations reveal the track events have a different polarity to the field events
  - A bigger number is a good thing in long jump (you jumped further)
  - A bigger number is a bad thing in 100m (you took longer)
  - We decide to multiply all track times by -1 so all events have the same polarity

A workflow for dimension reduction

0. Identify your variable types and perform appropriate Exploratory Data Analysis
1. Run dimension reduction analysis
2. **Examine the relationships between variables**
3. **Examine the relationships between subjects**
4. Further summarising/interpretation. Choose how many dimensions to keep/examine.
5. Downstream analysis

Decathlon Dataset. PCA step 2: Examine the variable plot



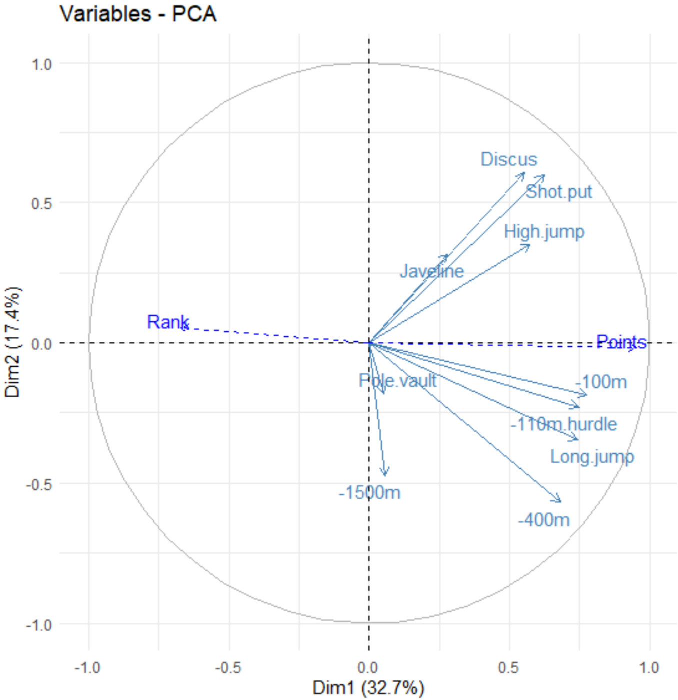
- Let’s run the PCA and look at the loading plot
- This loading plot shows what happens if we didn’t decide to multiply all track times by -1 (something we picked up in EDA). The ‘polarity’ problem separates track events and field events on this plot, which makes this plot more difficult to interpret
- Subsequent plots show the solution with the negative track times, and plots relabelled with track events having a minus sign in front

PCA Step 2: Examine the variable plot

- The loadings plot is where we see the dimension reduction in action. All of the original variables have been projected onto the 2D plane of the two PCs examined in each loading plot (PC1 ‘Dim1’ and PC2 ‘Dim2’ here)

- The correlation of the original variables with each PC is shown by the position of the arrowhead. Drop a perpendicular line from the arrowhead to each axis to see the correlation with that PC (unit circle shows the [-1,1] limits of correlation).

- Those original variables with shorter vectors (e.g. Pole Vault) are less loaded on to the PCs. Recall that the principal directions chosen in PCA preserve as much information as possible, so the variables poorly represented are less important for explaining variability overall in your data



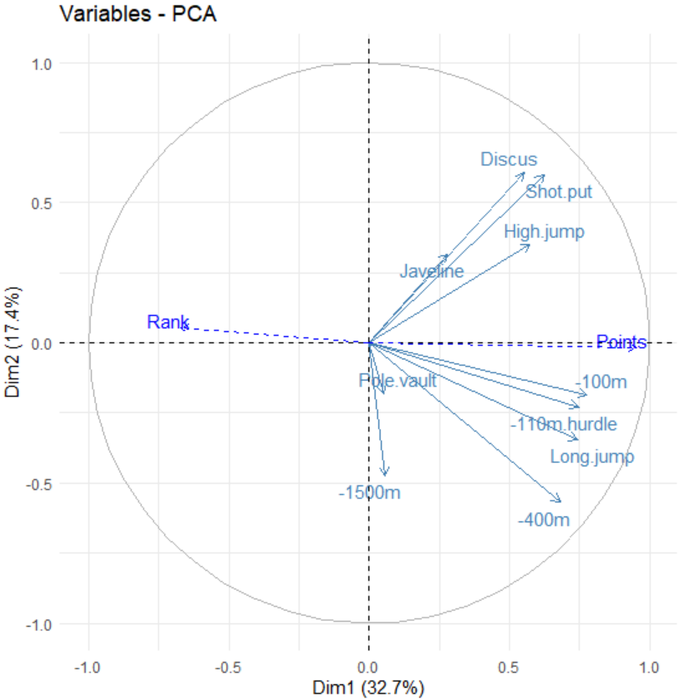
PC1 = 0.55(Discus) + 0.62(Shotput) + 0.28(Javeline) + …

PC2 = 0.61(Discus) + 0.60(Shotput) + 0.20(Javeline) + …

PCA Step 2: Examine the variable plot

- Sometimes you may have supplementary variables available that are not used in PCA analysis but potentially correlate with PCs. In this case we can use the ‘Points’ and ‘Rank’ which for each decathlon give the number of points achieved and the rank of each athlete respectively. These can be added to the loading plot by calculating their Pearson’s correlation with each PC.

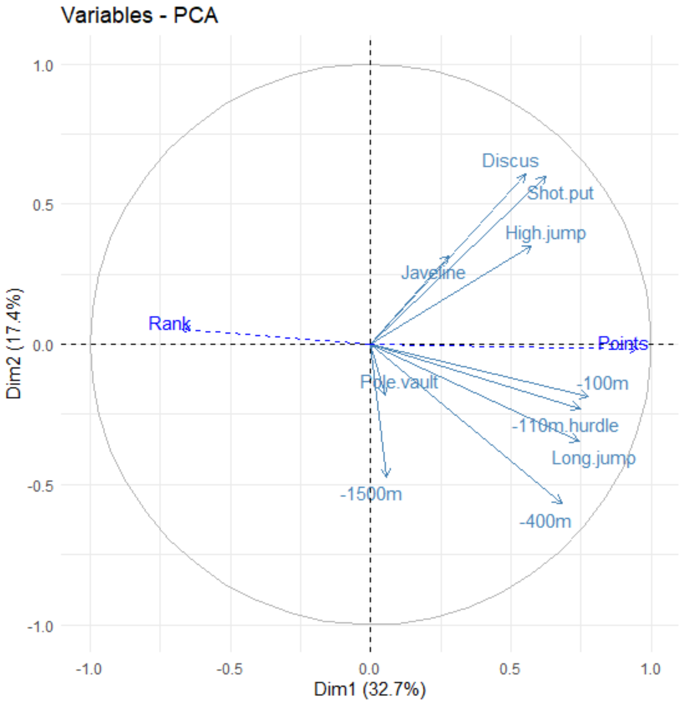
- Not surprisingly, these are correlated in different directions with respect to PC1. A smaller rank means a better performance, for which more points are awarded.



PCA Step 2: Examine the variable plot

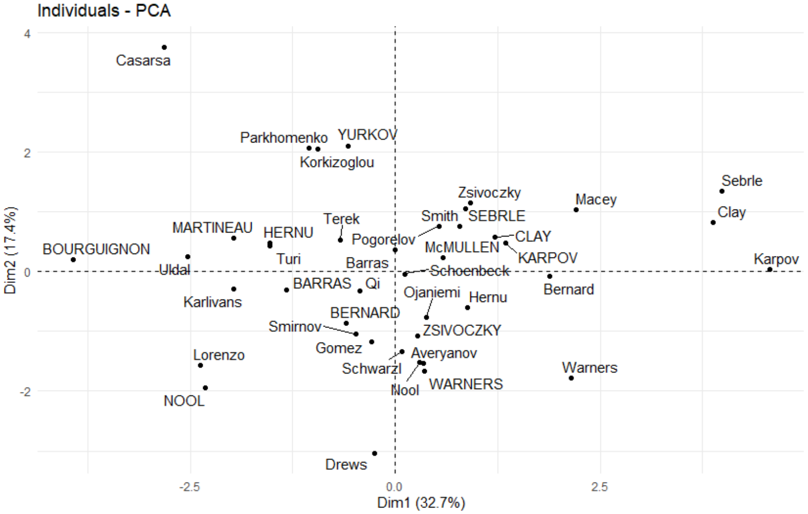
- Interpreting the PCs as latent variables:
  - Clearly PC1 represents performance of the athlete.
    - **A weighted sum** of all events – all variables are positively correlated.
    - Those with higher PC1 scores, have higher points and lower rank.
    - Despite explaining the most variability (33%), this is not particularly interesting *as a latent variable*.

  - PC2 appears to represent specialisation of athletes.
    - It is a **contrast** between events that involve sprinting, and those that don't, which have opposite signed correlations with PC2*.
    - Sprint specialists achieve by running really fast. Non-sprint specialists achieve most of their points by being good at other field events.
    - Despite explaining about half as much of the variability as PC1 (17%), PC2 has a more interesting interpretation as a latent variable.
    - If specialisation is present in decathletes, this PC will appear reliably in multiple samples.


Variables - PCA

*A technical note: the sign of the PC itself is arbitrary! The sign of the correlation between the variable and the PC is not arbitrary.

PCA Step 3: Examine the subject plot



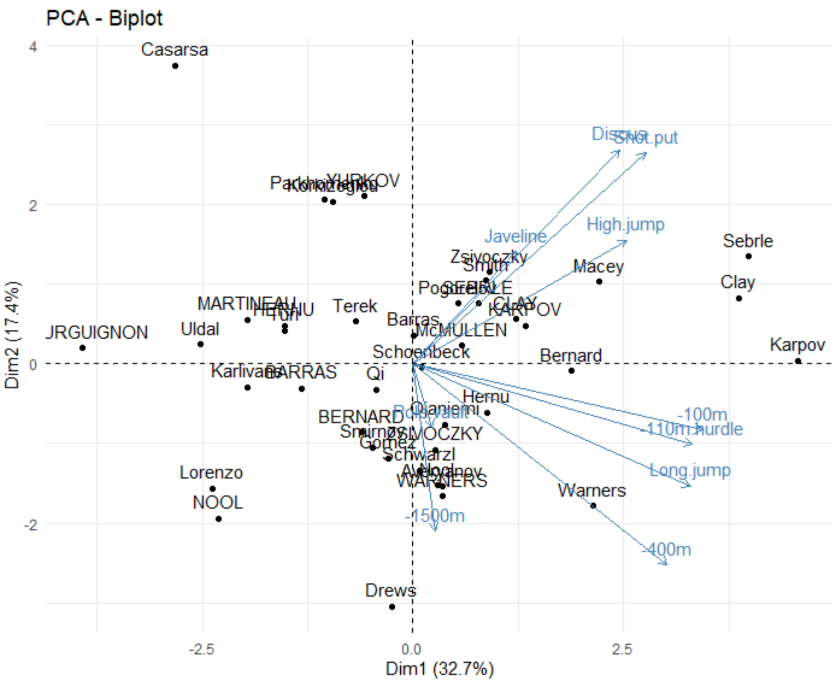
- We can plot the individual observations plotted in PC space, often called the score plot. Their PC scores for two dimensions at a time
- This is again a projection into PC space, but not of the variables, but instead the individual subjects
- Observations close to each other have similar values for the depicted PCs (but not necessarily similar overall). Recall latent variable interpretation for each PC.
- In some examples (not this one) there may be distinct clusters, and you can perform various types of clustering on observations in the PC space.

PCA Step 2 & 3: Examine the variable and subject plot

- The subjects and the variables can be examined on the same plot

- Note that the axes are the same as in the subject plot, and reflect the PC scores

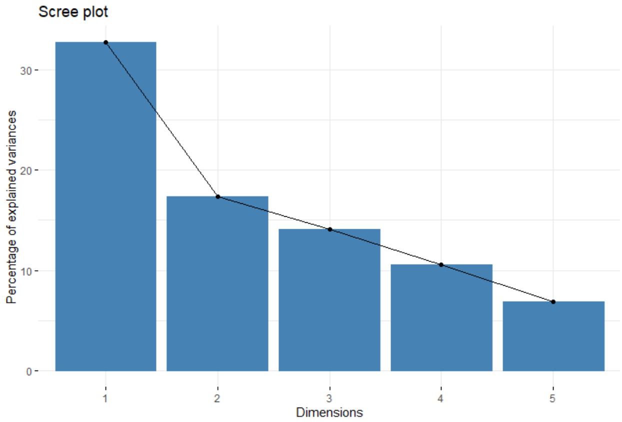
- The original variables can be thought of as additional axes showing the approximate* + relative position of each of the subjects on the original variable



---

* Approximate because the PC score is only correlated with, not proportional to any original variable

How many PCs should we consider?

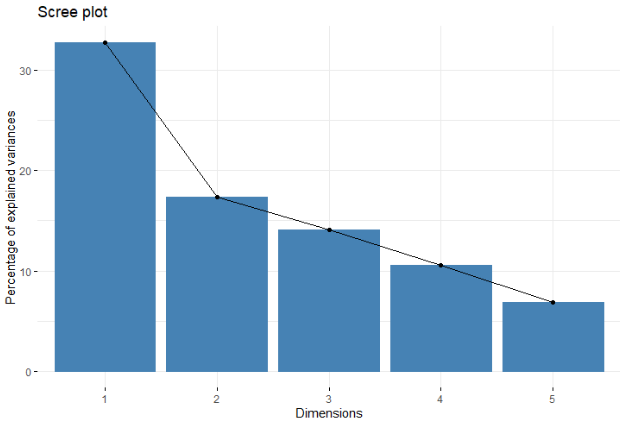


- This ‘Scree plot’ shows the % total variance of each PC. Each PC captures some amount of the total variance (across all original variables = across all dimensions), and this decreases for each subsequent PC

- We have seen that the first 2 PCs capture about 50% of the total variance and we have plausible latent variable interpretation for each. We should definitely consider these two.

- There are as many PCs as there were variables from the output, but we are only looking at the first 5.

How do you choose how many PCs?



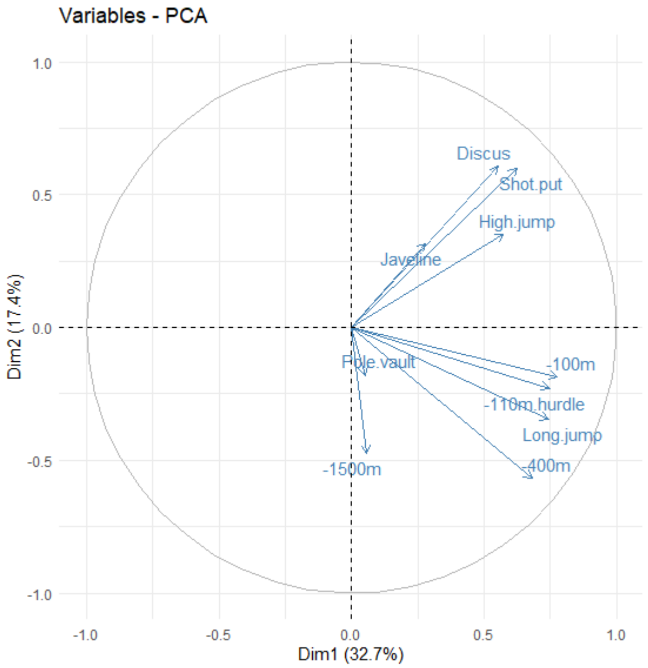
- The steeper the drop between bars, the less information is being captured in the kth PCs compared to the first (k – 1) PCs. We can consider dropping the kth and higher PCs

- Rules of thumb:
  - Smallest # of PCs that together hold 80-90% of variance
  - Keep components with an eigenvalue greater than the average of eigenvalues/Keep components with an eigenvalue > 1 when working with standardised variables as per this example
  - Look for the ‘elbow’ in the line plot

- If we collected many samples, we would probably find that the first dimensions were more stable in their (relative) directions between samples whereas the last dimensions would keep changing. This would show us that the first dimensions contain more ‘signal’, and the last dimensions contain more ‘noise’.

Challenge question – correlation between original variable and PC

What is the correlation between -100m (i.e. 100m time with a minus sign) and PC1 score?

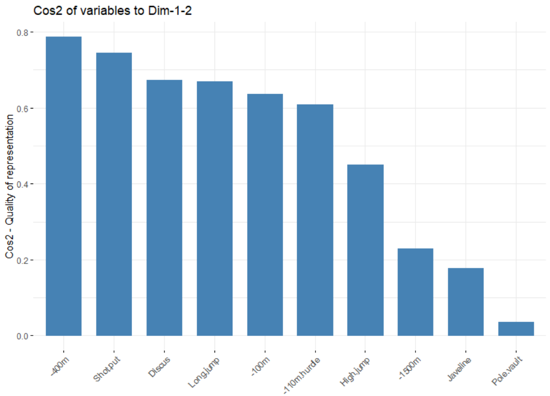


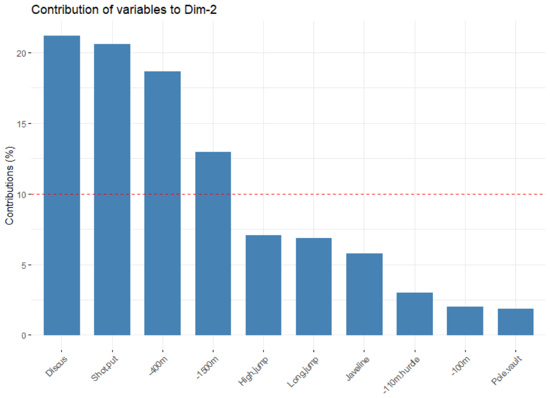
Challenge question – correlation between PCs

What is the correlation between PC1 and PC2 scores?

PCA Step 4: Further summarisation

- **Other interpretation aids:**
  - Quality of representation for variables and individuals on a given map (think about the relative length of the vectors on the loading plot)
  - Contribution of variables and individuals to each PC (again the relative length of the vectors, and how important each subject was to the PC output)





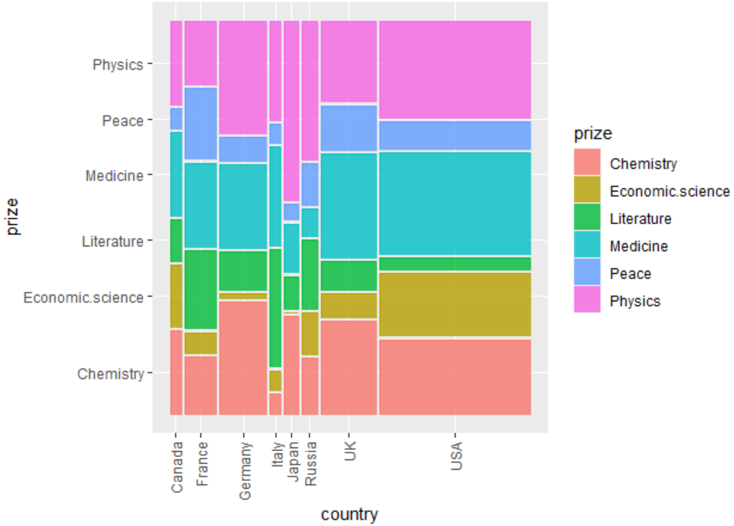
Correspondence Analysis (CA)

Correspondence Analysis (CA)

- Very commonly used in surveys with categorical responses.
- Often used as categorical/qualitative analogue of PCA
- Input for PCA and factor analysis is continuous observations
- Input for Correspondence Analysis is categorical observations on two variables: contingency table

Example: Nobel Prize Data

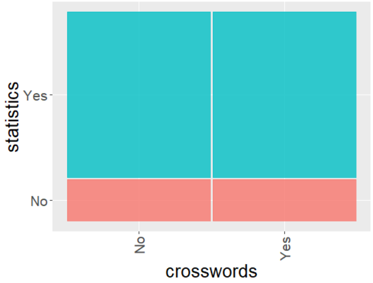
- Nobel prize winners by country from 1901 to 2015. Just G8 countries. Excluding mathematics.
- We could use a contingency table to show frequencies, or a mosaic plot as shown below:
  - The width of the columns represents the column proportions (country)
  - The area of the blocks represents the proportion of all prizes (prize x country)



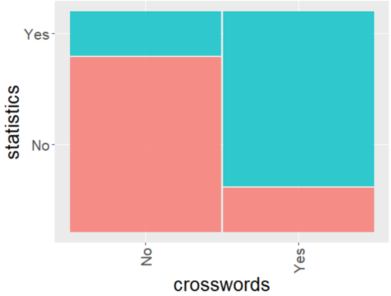
https://husson.github.io/MOOC.html

Association

- In an underlying population, there are two possible relationships between the two variables: independence, or association.



INDEPENDENCE
/no association



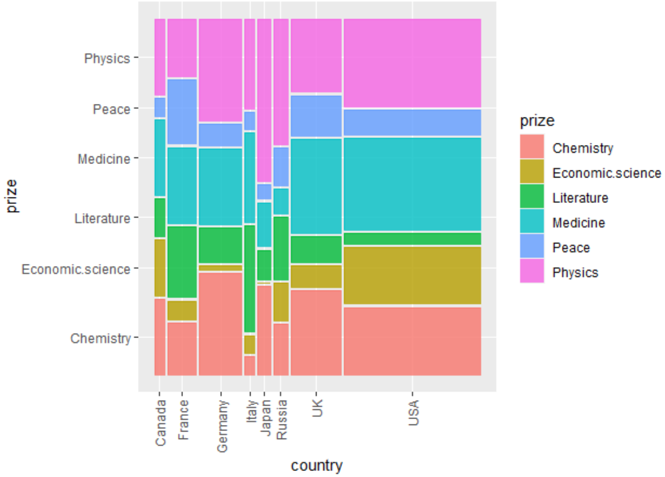
DEPENDENCE
/association

- We use a chi-squared test (or similar) to examine evidence against the null hypothesis of independence for a given sample. This provides a good summary of association for two variables with two categories in each.

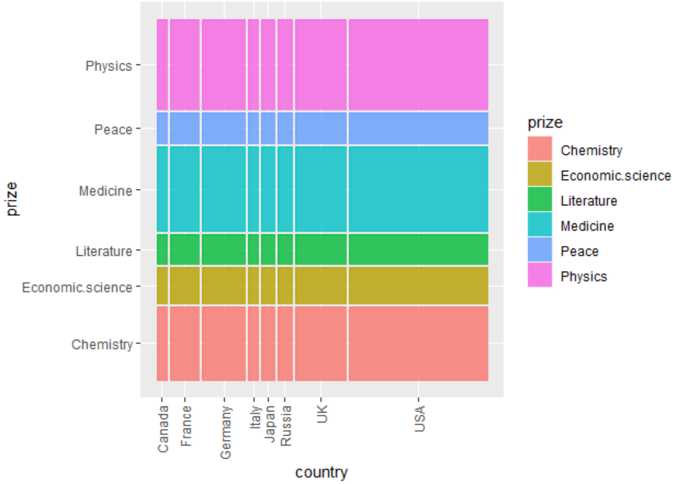
- But what about if there are more than two categories in either or both variables?

Comparing to the independence model

- Is there an association between country and type of Nobel prize won?
- We can compare the observed areas to what is expected under the independence model, where the frequency of each combination of country and prize follows the marginal (overall) frequencies of countries and of prizes



Observed data

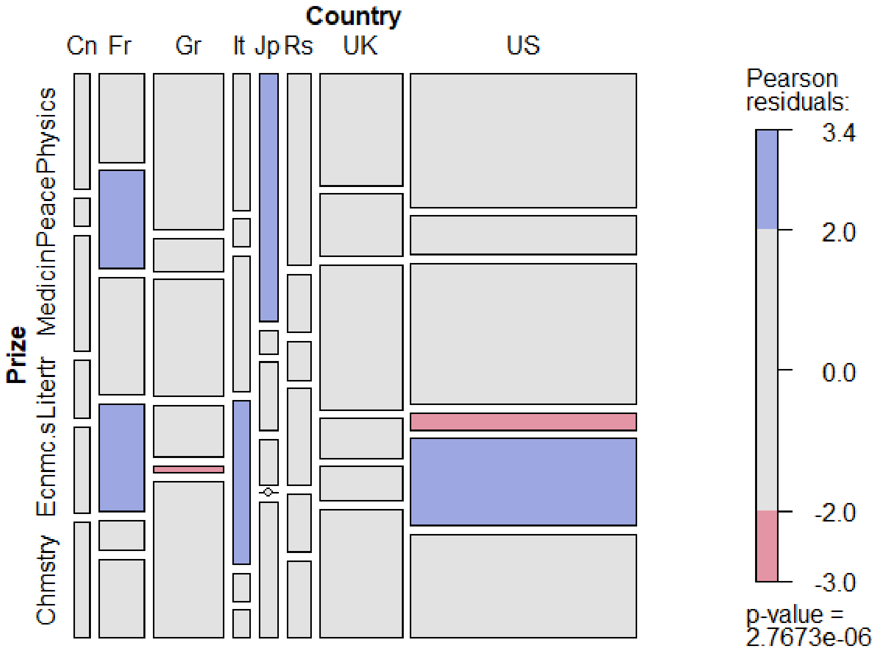


Expected under null hypothesis INDEPENDENCE

Nobel Prize data Step 0: EDA

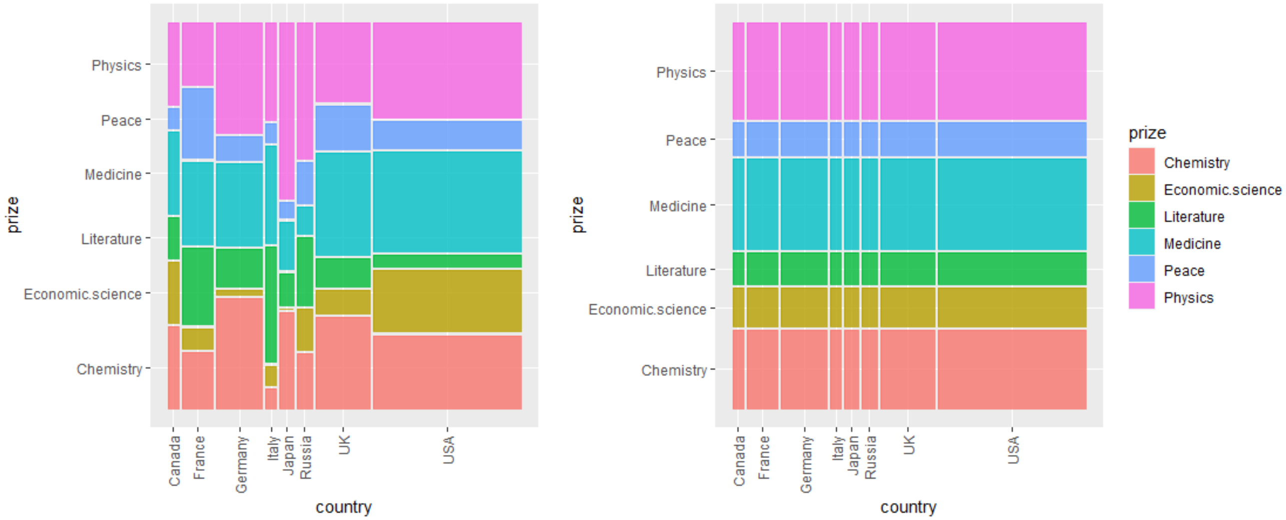
- We can perform a Chi-squared test for independence of the two variables (very strong evidence of association for these data)

- We can also look at the Pearson residuals to see which combination of categories have observed values furthest away from expectation under the independence model



How does correspondence analysis work?

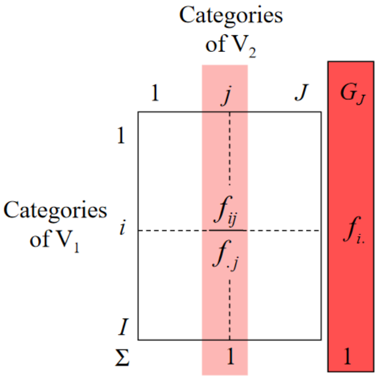
- *Can these frequencies be treated as continuous data?*
- *For each country (column), what is its profile? (Relative frequency of each prize)*

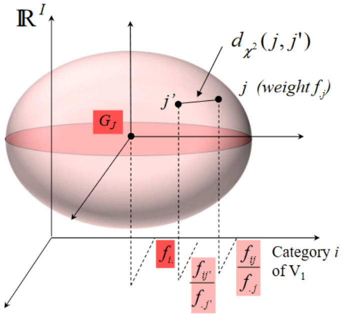


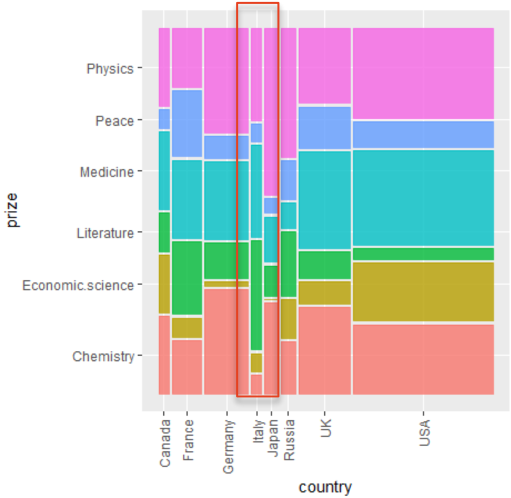
```
nobel_sums_long %>% mutate(Country=fct_relevel(Country,"Total",after = Inf)) %>% ggplot(aes(x=Country,fill=name,y=value)) +
geom_bar(position="fill",stat="identity") + scale_fill_brewer(palette = "Set2")
```

How does correspondence analysis work?

- Each column profile (country) is represented as a point in J-dimensional space, where J is the number of rows (prizes). The location depends on the relative frequency of each row.
- The mean column profile is represented as another point (G_i) that serves as a reference to observed column profiles
- You can calculate a Chi-squared distance between any two column profiles
- The total inertia is made up of the inertia (weighted distance) of each column profile from the mean column profile (i.e. the deviation of that column from independence)



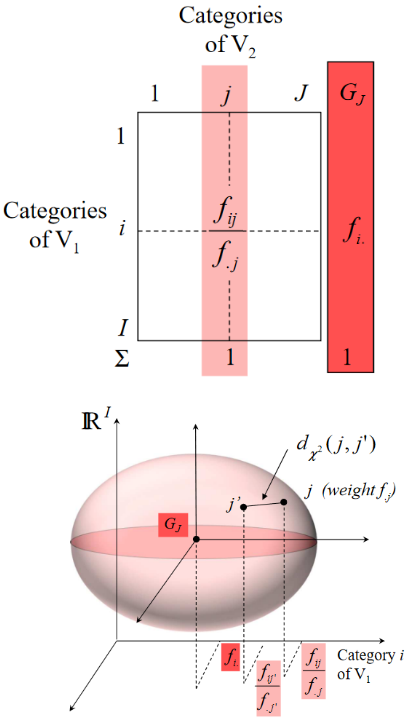


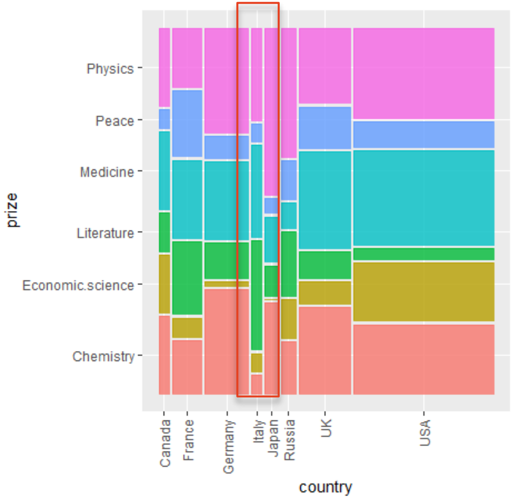


https://husson.github.io/MOOC.html

How does correspondence analysis work?

- The set up is analogous to PCA, except that the relative frequencies must add to 1 (every subject must be in one of the categories)
- Extraction of components proceeds as per PCA, choosing principal directions that maximise the inertia of the first dimension, and subsequent dimensions being orthogonal to previous dimensions
- The same analysis can be performed from the row profile point-of-view. The total inertia is the same. Columns and rows are symmetric in CA. Think association between two variables as not having a direction.



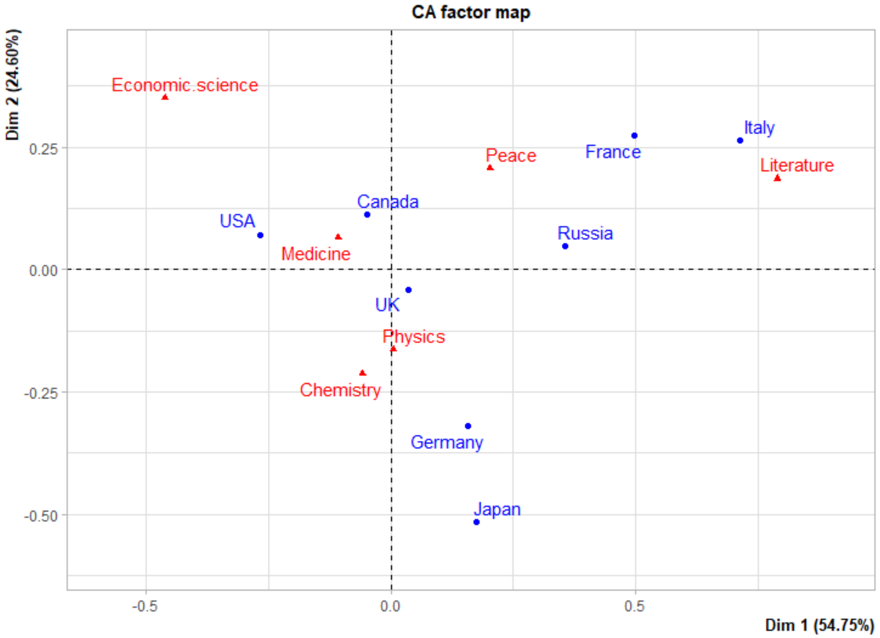


CA Steps 2&3: Correspondence Map

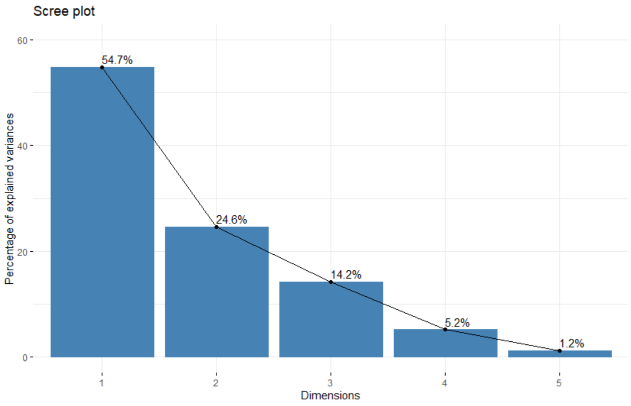
- This is the correspondence map for the Nobel prize data

- Rows (countries) appear in blue
- Columns (prize type) appear in red

- This is an analogue of the biplot for PCA:
  - Rows and columns play symmetric roles in CA
  - We don't usually include vectors and points, as the choice of which is 'variables' and which is 'subjects' is arbitrary.
  - For discussion we will consider countries as subjects and prizes as variables.



CA Steps 2&3: Scree plot



```
dim 1 dim 2 dim 3 dim 4 dim 5
 0.08  0.04  0.02  0.01  0.00
```

- Scree plot shows the percentage of inertia captured by each dimension
- The first two dimensions capture ~80% of total inertia

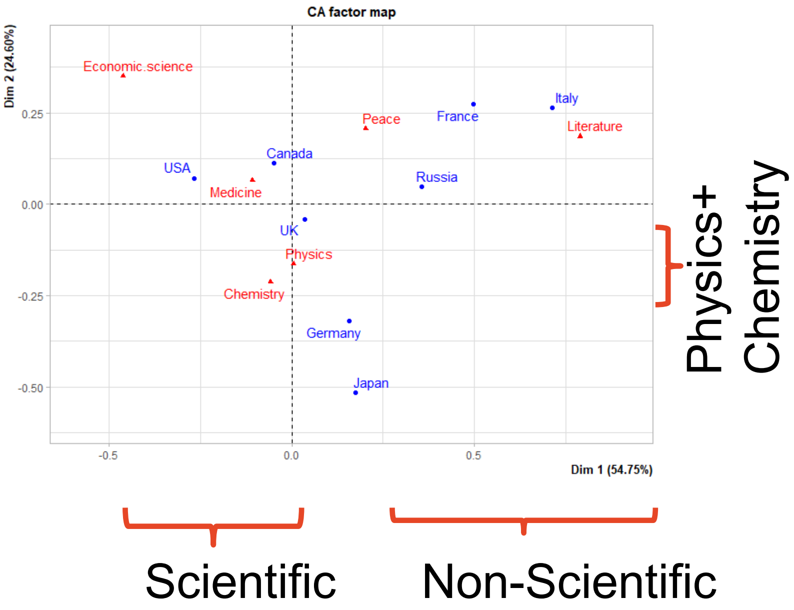
- The size of the eigenvalue itself also tells us something, an eigenvalue of 1 means perfect association between 1 column and 1 row (i.e. 1 country had all of its prizes in one category only, and no other countries had a prize in this category)

```
fviz_screeplot(res.ca, addlabels = TRUE, ylim = c(0, 60))
round(res.ca$eig,2)[,1]
```

CA Step 2: Examine the ‘variable’ plot

- Dimension 1 appears to contrast the scientific prizes from non-scientific

- Dimension 2 appears to contrast ‘hard sciences’ Physics and Chemistry from the others



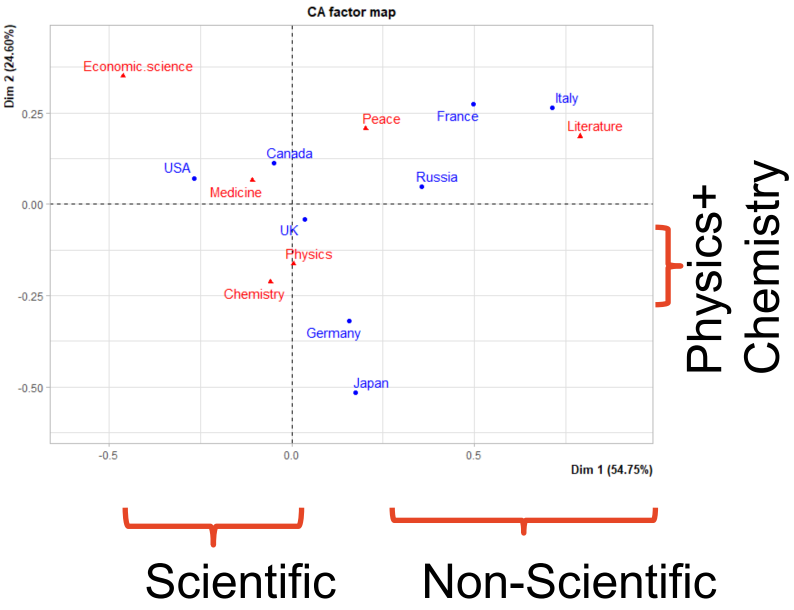
Physics+ Chemistry

Scientific

Non-Scientific

CA Step 3: Examine the ‘subject’ plot

- Same map, but lets concentrate on the rows now (countries, blue points)
- The origin indicates the mean row profile
- Can think of the country being pulled from the origin toward (and potentially beyond) the prizes with which it is most associated with
- The distance between blue and red points on this plot does not have a direct meaning, but rows and columns that are associated tend to be closer together. The plot can be scaled so that the distances between red points, or between blue points more accurately reflect similar profiles



Method Extension: Multiple Correspondence Analysis (MCA)

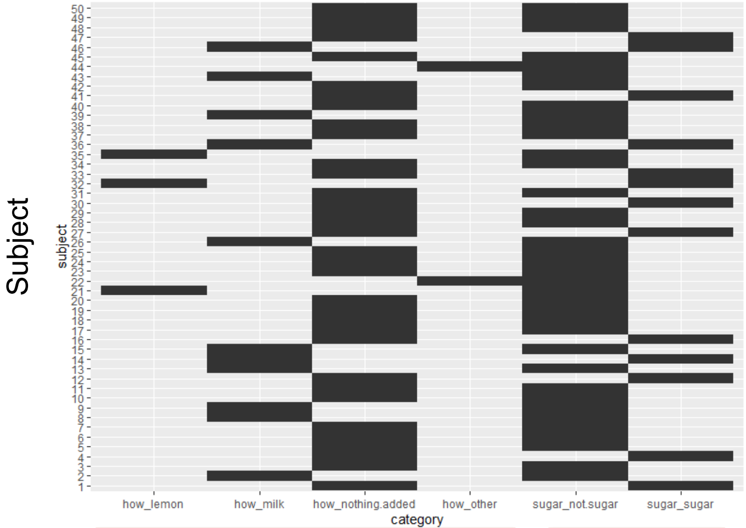
- In multiple correspondence analysis we consider more than two categorical variables
- We no longer start with a contingency table of two categorical variable frequencies, but instead with an ‘indicator matrix*’ of subjects vs. categories
- Correspondence map can show relation between subjects, variables and categories

How do you take your tea:

Milk	Lemon	Other	Straight up
○	○	○	○

Do you have sugar in your tea?

Yes	No
○	○



How do you take your tea:

…sugar in your tea?

*Matrix of indicator variables where e.g. (1 = Yes, 0 = No) for the relevant subject x category combination

Dimensionality Reduction Techniques: Distance based

THE UNIVERSITY OF SYDNEY

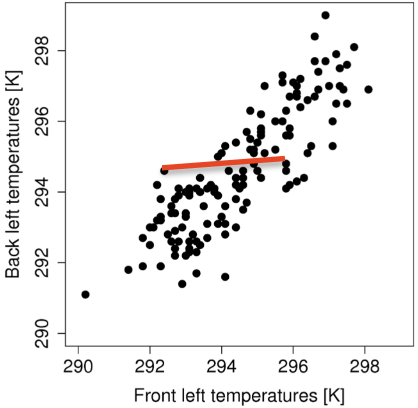
Distance Based Methods

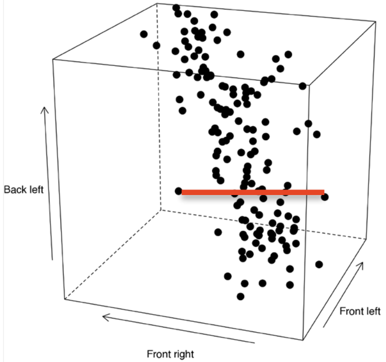
- The starting point of methods so far have been measurements on **variables**, either continuous (PCA, FA) or categorical (CA)

- Multidimensional Scaling (MDS) and similar methods use the distances between **subjects** as their starting point.

- The goal is to produce a low-dimensional map (often called an ordination or embedding for these methods) that most accurately visualises the distances between the subjects, i.e. preserves as much as possible the distance information from higher dimensions

- Better at accurately capturing local structure (distances between neighbouring subjects), than PCA-like methods, which are better at capturing global structure (distances between all subjects)





What are distances between subjects?

- We're (very) familiar with geographic distances, but what is the distance between subjects?
- The type of data will dictate the appropriate distance measurement. MDS is very flexible in that it can accommodate different kinds of distances.
    - The Euclidian distance is the straight-line distance between two points in your original variable space
    - Chi-squared distance, is the Euclidian distance between relative frequencies (see the CA example)
    - In ecology, the abundance of species at different sites is measured. Bray-Curtis distance is used, ranges between 0 (identical) and 1 (no similarity)

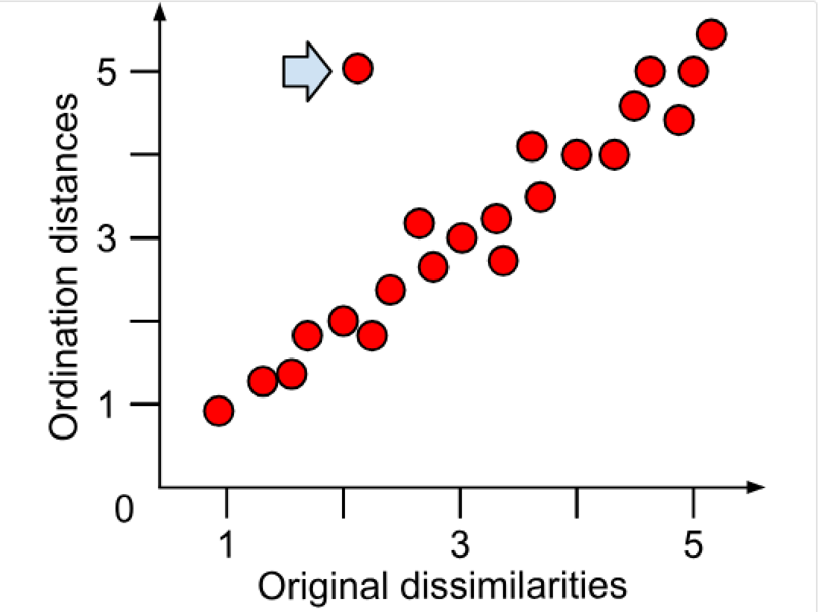
https://2kmfromhome.com/

Multidimensional Scaling (MDS)

MDS Step 1: Run the MDS

- MDS can be run as either metric or non-metric (nMDS). Metric is also called Principle Coordinates Analysis (PCoA)

- Unlike PCA, it is an iterative algorithm, starting with randomly positioned subjects on the lower dimensional map and stepping towards an optimal solution where the ordination (lower dimensional) distances are as close as possible to the original distances (higher dimensional)

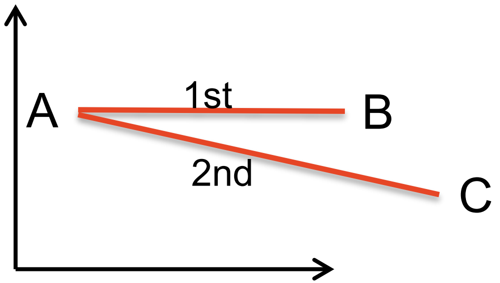


https://sites.google.com/site/mb3gustame/dissimilarity-based-methods/non-metric-multidimensional-scaling

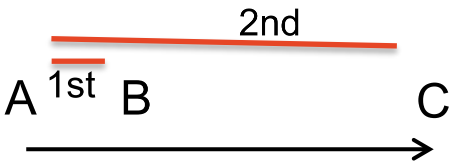
Metric (MDS) or Non-metric? (nMDS)

- Metric is appropriate when you expect there can be a linear relationship between the data distances (original variables) and the ordination distances (ordination map). The method and outputs are very similar to PCA. Metric with Euclidian distance *is* PCA.

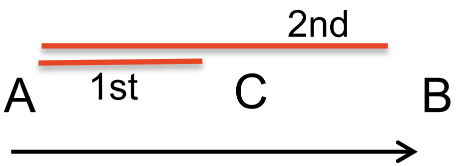
- In non-metric MDS (nMDS) the success of the ordination depends on preserving the ranks of dissimilarity between subject



ACTUAL DISTANCE
In higher dimensions



SAME DISSIMILARITY RANK
In lower dimension



DIFFERENT DISSIMILARITY RANK
In lower dimension

Differences of nMDS from PCA

- The output dimensions are not orthogonal to each other as in PCA
- Ordination map is optimised for a pre-specified number of dimensions
  - The algorithm for producing an optimal ordination works by iteration, not calculation of a single solution as PCA does
  - The dimensions are not ordered by variance explained as they are in PCA
- nMDS is able to handle and effectively summarise non-linear relationships

Example: Mac Nally bird abundance

- Let’s take an ecology example for MDS: Mac Nally bird abundance data
  - Ralph Mac Nally (1989)
  - Maximum abundance for 102 bird species
  - 37 sites, 5 different forest types (Gippsland manna gum, montane forest, woodland, box-ironbark and river redgum and mixed forest)

- Research question: Do the bird assemblages differ between forest types?

https://www.flutterbys.com.au/stats/tut/tut15.1.html

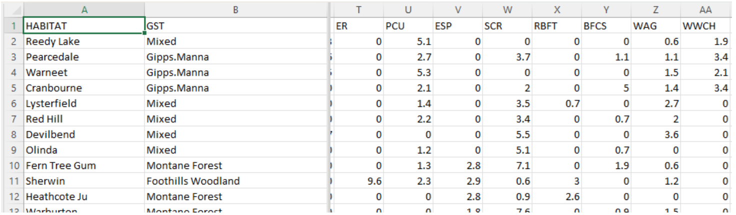
Bird abundance Step 0: EDA

- Very skewed data typical of abundances. A lot of variation in total abundance between sites (x-axis) and between relative species abundance (colour area)

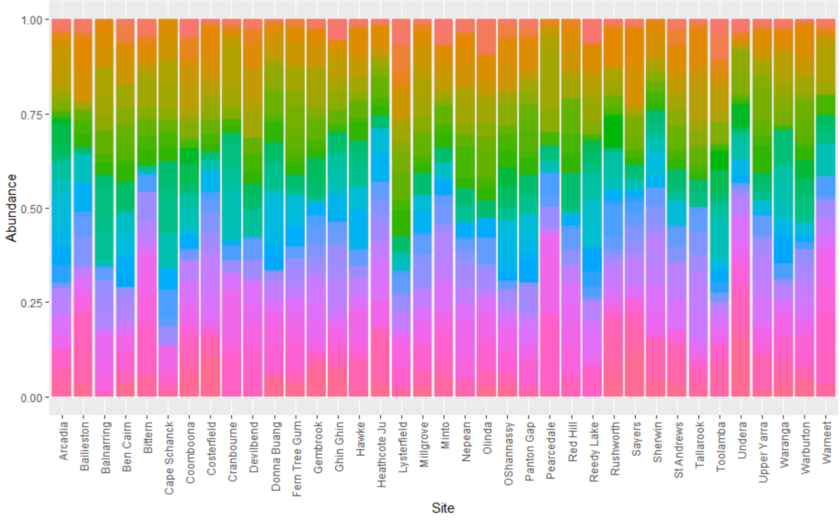
- Lots of 0 counts

- We need to standardise these abundances to meaningfully compare the composition of the sites:
  - Apply Wisconsin double standardisation (divide abundance by column maximum and row total), which reduces the influence of highly abundant species and equalises the relative importance of sites and species

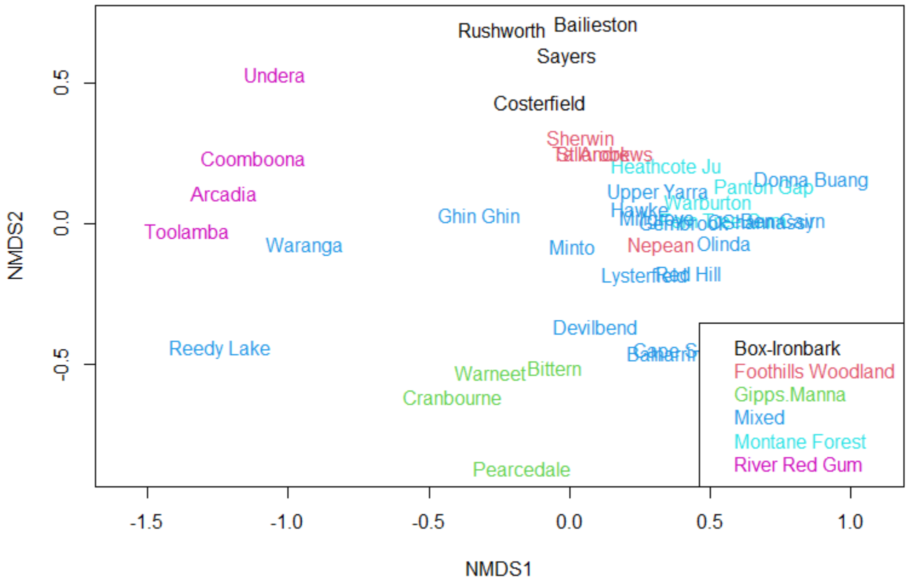
  - Then calculate the Bray-Curtis distance between sites, which ignores “shared absences” (when both sites have a 0 count for that species)



	A	B	T	U	V	W	X	Y	Z	AA
1	HABITAT	GST	ER	PCU	ESP	SCR	RBFT	BFCS	WAG	WWCH
2	Reedy Lake	Mixed	0	5.1	0	0	0	0	0.6	1.9
3	Pearcedale	Gipps.Manna	0	2.7	0	3.7	0	1.1	1.1	3.4
4	Warneet	Gipps.Manna	0	5.3	0	0	0	0	1.5	2.1
5	Cranbourne	Gipps.Manna	0	2.1	0	2	0	5	1.4	3.4
6	Lysterfield	Mixed	0	1.4	0	3.5	0.7	0	2.7	0
7	Red Hill	Mixed	0	2.2	0	3.4	0	0.7	2	0
8	Devilbend	Mixed	0	0	0	5.5	0	0	3.6	0
9	Olinda	Mixed	0	1.2	0	5.1	0	0.7	0	0
10	Fern Tree Gum	Montane Forest	0	1.3	2.8	7.1	0	1.9	0.6	0
11	Sherwin	Foothills Woodland	9.6	2.3	2.9	0.6	3	0	1.2	0
12	Heathcote Ju	Montane Forest	0	0	2.8	0.9	2.6	0	0	0
13	Warburton	Montane Forest	0	0	1.8	7.6	0	0.9	1.5	0



MDS Step 3: Examine the subject (site) plot

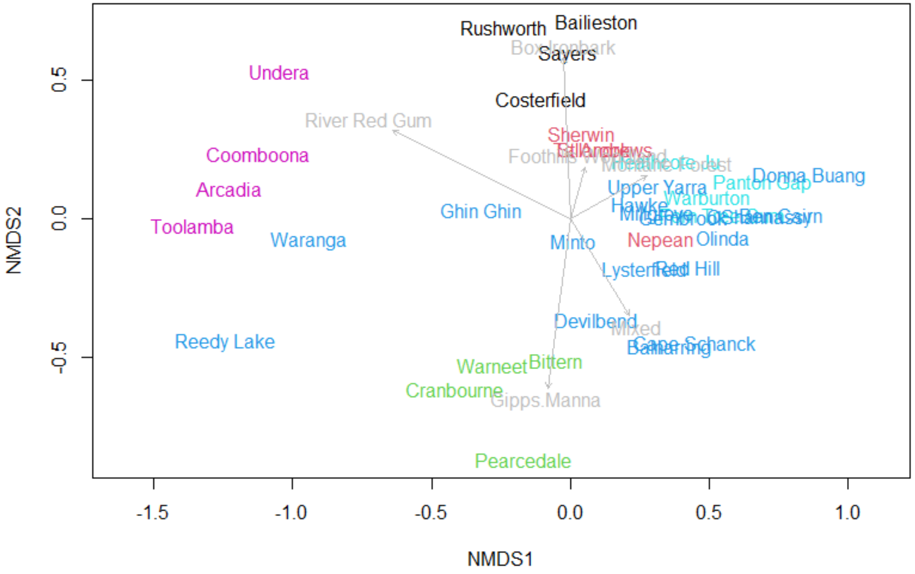


- Forests of the same type (colour label) tend to be closer together. Reflects that bird abundance is related to the forest type.

- Distances between sites reflect their similarity in relative species abundance (community composition)

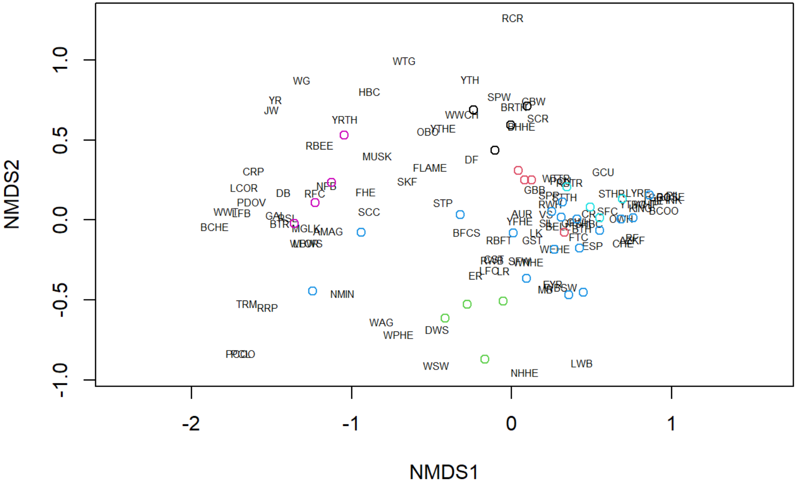
- Unlike in PCA, the dimensions in MDS are non-orthogonal
  - No longer have a potential interpretation of dimensions as *distinct* latent variables

MDS Step 3: Examine the subject (site) plot



- Environmental variables can be included in the plot in a similar way to supplementary variables for PCA (i.e. not part of the input, but correlation with each dimension calculated post-hoc)
- Could include variables such as soil pH, soil composition, altitude, etc.
- In this example we have just used each type of forest as a [binary] variable. The directions generally match the clusters of forest type

MDS Step 2: Examine the variables plot



- Variable (species) scores are included by taking a weighted average of the site scores

- Species that are closest to the sites in the ordination map are expected to have the highest abundances at that site

Step 4: Further summarisation

- We would like to (formally) test the hypothesis that the type of forest is associated with species composition, we can use PERMANOVA

```
> adonis2(macnally.dist~macnally$HABITAT)

Call:
adonis(formula = macnally.dist ~ macnally$HABITAT)

Terms added sequentially (first to last)

                 Df  SumsOfSqs MeanSqs
macnally$HABITAT  5       3.50   0.699
Residuals        31       4.60   0.148
Total            36       8.09
                 F.Model     R2  Pr(>F)
macnally$HABITAT    4.72 0.432   0.001
Residuals                0.568
Total                    1.000

macnally$HABITAT ***
Residuals
Total
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05
  '.' 0.1 ' ' 1
```

https://stats.stackexchange.com/questions/316590/what-is-the-reason-behind-the-name-adonis-for-permutational-manova

References

- Tutorial examples used:
  - PCA Decathlon example from FactomineR
  - CA Nobel Prize example from François Husson's github
  - nMDS Mac Nally example from Murray's R resources