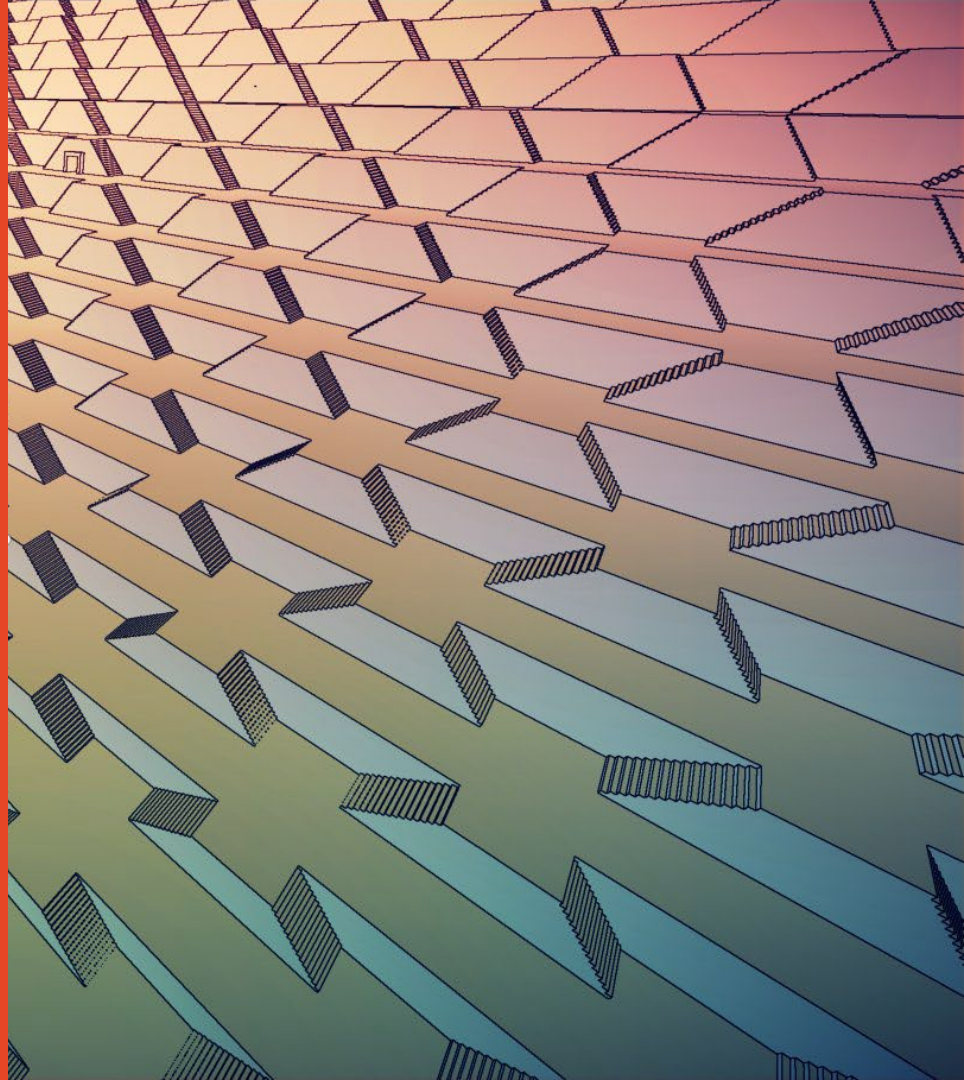
Multivariate Statistical Analysis 1: Dimension Reduction Methods

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

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."

During the workshop

Ask short questions or clarifications during the workshop (either by Zoom chat or verbally). There will be breaks during the workshop for longer questions.

Slides with this blackboard icon are mainly for your reference, and the material will not be discussed during the workshop.

Challenge questions will be encountered throughout the workshop.

Learning Objectives

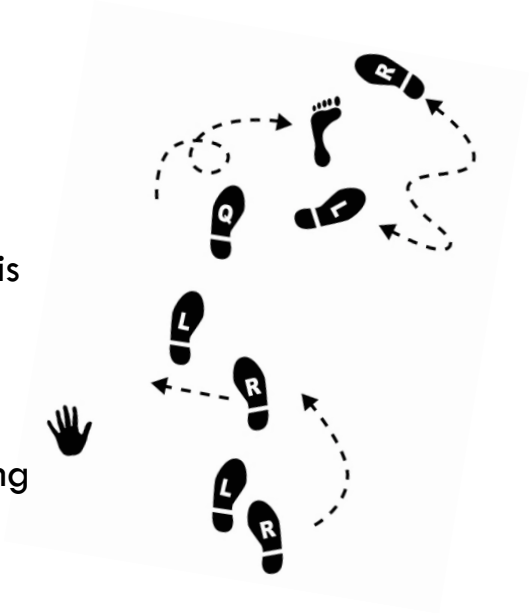
- Understanding the motivation for using multivariate methods
- Understanding the goals of multivariate methods
- A workflow for dimension reduction methods
- Introduction to major dimension reduction methods:
  - Principal Components Analysis and Factor Analysis
  - Correspondence Analysis
  - nMDS

- <u>Not covered in this workshop:</u>

- Multivariate regression
- Multivariate hypothesis tests

What is a workflow?

- Every statistical analysis is different, but all follow similar paths. It can be useful to know what these paths are
- We have developed practical, step-by-step instructions that we call '*workflows*', that can you can follow and apply to your research
- We have a general research workflow that you can follow from hypothesis generation to publication
- And statistical workflows that focus on each major step along the way (e.g. experimental design, power calculation, model building, analysis using linear models/survival/multivariate/survey methods)

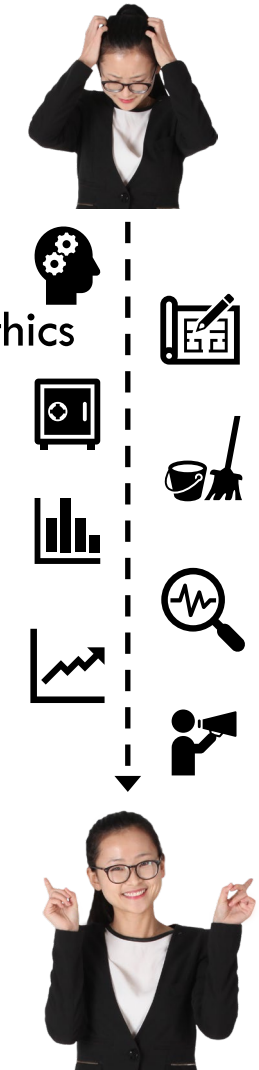


Statistical Workflows

- Our **statistical workflows** can be found within our workshop slides
- **Statistical workflows** are software agnostic, in that they can be applied using any statistical software
- There may also be accompanying **software workflows** that show you how to perform the statistical workflow using particular software packages (e.g. R or SPSS). We won't be going through these in detail during the workshop. If you are having trouble using them, we suggest you attend our monthly Hacky Hour where SIH staff can help you.

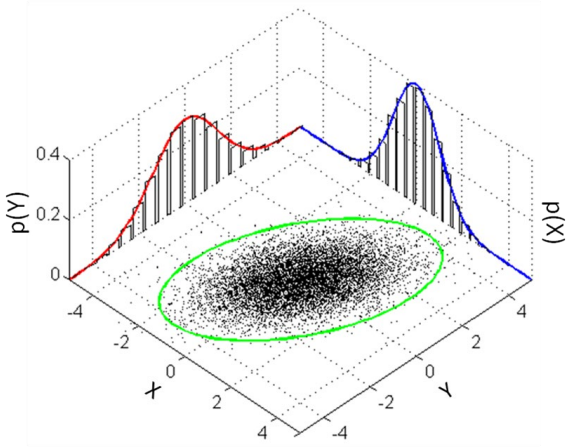
General Research Workflow

1. **Hypothesis Generation** (Research/Desktop Review)
2. **Experimental and Analytical Design** (sampling, power, ethics approval)
3. **Collect/Store Data**
4. **Data cleaning**
5. **Exploratory Data Analysis (EDA)**
6. **Data Analysis aka inferential analysis**
7. **Predictive modelling**
8. **Publication**



Multivariate Statistics

- Multivariate statistics applied when you are modelling the distribution of more than one variable (usually an outcome variable).
- We are usually interested in the relationships between such variables, i.e. how the variables vary together (their joint distribution). Otherwise, we could just perform [a simpler] analysis on each variable separately.
- It is common to have more than one variable in your study, but we often treat some variables as 'fixed' (i.e. not modelling their distribution). *E.g. in multiple linear regression, all of the explanatory variables are fixed.*
- Multivariate statistics applies when you are considering multiple random variables in combination e.g. *in linear models with multiple outcome variables.*



https://en.wikipedia.org/wiki/Multivariate_normal_distribution

Why multivariate statistics?

- **Multivariate statistics** is a very broad topic with a huge array of different techniques. The motivation for using one of these techniques may include:
  - **Investigation of dependence**: relationships among variables
  - **Sorting and grouping**: identify groups among the entities/subjects under study
  - **Data reduction**: summarise multiple variables through a smaller set of 'synthetic variables'
  - **Hypothesis testing**: group differences for multiple outcome variables
  - **Prediction**: based on a multivariate model

Which of these has brought you to this workshop?

Special cases of multivariate statistics

- Multivariate statistics also come in to play in situations where we have repeated measurements (of the same thing) on the same subjeccts, especially longitudinal data
- Although these are indeed important applications of multivariate statistics, they are a special case and outside the scope of this workshop
- In this workshop, our examples will be limited to those where different things have been measured on the same individuals
- See our Linear Models series of workshops and/or book a consult with us to discuss analysis of repeated measures data

Dimensionality Reduction Techniques

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

Why do we need dimension reduction?

- We want to understand the relationships between variables (dependence) in situations where we have a lot of variables:
  - Are there patterns in associations/correlations across multiple variables?
  - Do these patterns yield insight into **latent variables,** these are unobserved variables that are a type of synthetic variable representing meaningful concept that can't be observed directly (e.g. personality, stage of disease progression)

- We want to understand the relationships between subjects:
  - Are there groups (**clusters**?) of subjects
  - How similar are our subjects to each other? Do they fall along some gradient that correlates with some other variable that can be directly measured (e.g. different field sites w/ an environmental variable)

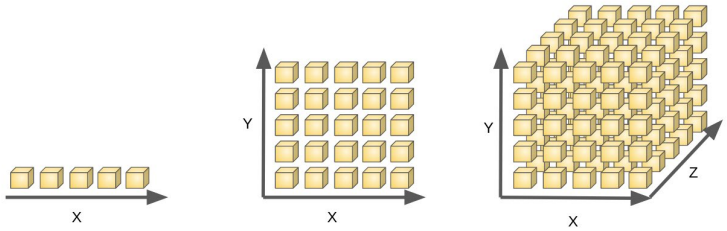
- We just want to have fewer variables going into some analysis:
  - Is there a way to summarise all of the measurements into a few measurements (e.g. what are groups of genes that exhibit a similar response to drug treatment?)

Why do we need dimension reduction for linear models?

- Multicollinearity – situations where predictors are very similar to each other
  - Consequence: Difficulty accurately estimating effect of each predictor on the outcome. See **Model Building Workshop**

- Situations where k ≈ n or k > n, we have as many measurements on each subject as we have subjects (or more measurements). **Sometimes referred to as the 'curse of dimensionality'** (Common in high throughput biological assays gene expression, neuroimaging etc)
  - Consequence: Can't fit a model

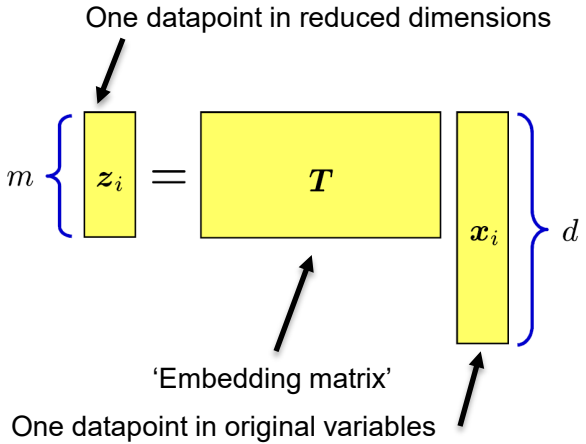
"Even if the number of collected data points is large, they remain sparsely submerged in a voluminous high-dimensional space that is practically impossible to explore exhaustively"

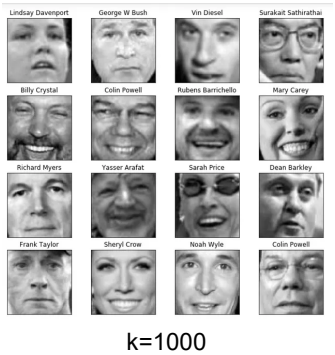


Nguyen LH, Holmes S (2019) Ten quick tips for effective dimensionality reduction. PLOS Computational Biology 15(6): e1006907. https://doi.org/10.1371/journal.pcbi.1006907

Dimension Reduction (DR): what is it?

- You can think of dimension reduction as using the relationships between variables or subjects to extract information dispersed throughout your original dataset and 'concentrate' it in a lower number of dimensions
- Every datapoint in the space of original variables can be transformed to a point on in the reduced dimensional space, which facilitates exploration of your 'sparse' data
- Depending on the method, you may end up with as many dimensions as you had in your original dataset, but by virtue of concentrating the information in smaller dimensions, you can discard or ignore dimensions that contain relatively little information





k=1000

k=250

https://towardsdatascience.com/eigenfaces-recovering-humans-from-ghosts-17606c328184

What are the dimensions we get as output?

- Synthetic Variables: We reduce the number of variables in our analysis by replacing original variables with a smaller number of synthetic variables. By 'synthetic variable', I mean one that *synthesises* information from multiple 'original variables'

- **Body Mass Index (BMI)** is a simple example of a synthetic variable:

$$\frac{\text{Weight in kg}}{(\text{Height in m})^2}$$

- Dimension reduction techniques combine information across variables* to create useful synthetic variables, these may or may not be interpretable as latent variables

*For simplicity and consistency I will mostly call the original variables just 'variables' and the synthetic variables 'dimensions'.

A workflow for dimension reduction

A workflow for dimension reduction

0. Identify your variable types and perform appropriate Exploratory Data Analysis
1. Choose a dimension reduction method and 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

Step 0 – Exploratory Data Analysis

- EDA is an important step in our general research workflow
  - See the EDA step in our Research Essentials workshop
  - Useful for choosing the right Dimension Reduction method

- Once you have chosen, there is some further EDA required – partly from the output of the method:
  - Want to make sure that the method is appropriate
  - Want to make sure that we use the right settings for our data and analysis questions

Step 1 – Run Dimension Reduction Analysis

- The purpose of this workshop is to introduce the theory behind dimension reduction and show some basic examples
- You may have a single dimension reduction method you want to use, or you may want to run multiple methods and compare outputs
- Dimension reduction analysis can be performed in several statistical software packages
- Often the best learning resource is to find a tutorial using the software package you intend to use for your own data

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

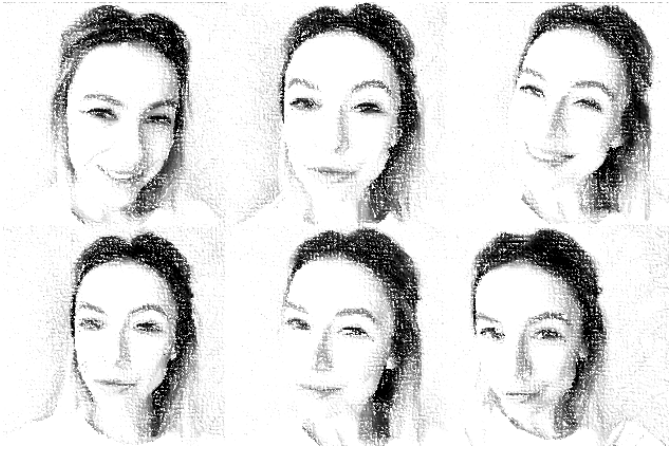
Nguyen LH, Holmes S (2019) Ten quick tips for effective dimensionality reduction. PLOS Computational Biology 15(6): e1006907. https://doi.org/10.1371/journal.pcbi.1006907
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006907

Dimensionality Reduction Techniques: PCA-like

Principal Component Analysis (PCA)

Principal Components Analysis (PCA): Ever taken a selfie?

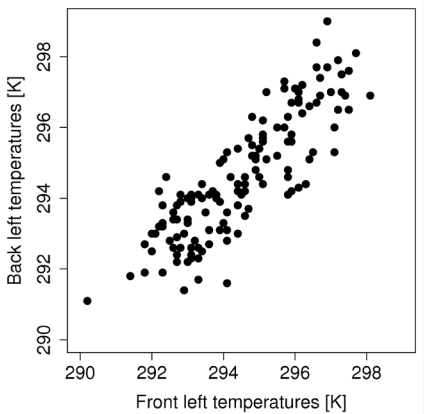
- There are many sites that advise on how to take the best selfie
- One element that is clearly important is, which angle? You're reducing from a 3D person to a 2D image so you are losing information
- We might choose the angle to highlight certain features of our face
- PCA is like taking a photo. Your data points are in high dimensions. PCA chooses an angle in this space that captures as much information as possible in a low number of dimensions.

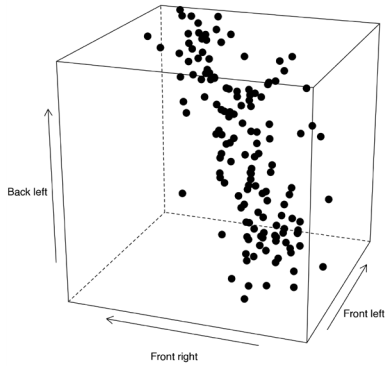


https://macpaw.com/how-to/take-best-selfie

Mechanics of Dimension Reduction

- To get a geometric intuition of what PCA-like dimension reduction methods are doing, we need to think of our original measurements as a 'cloud' of points in space. Each original variable defines an axis, and each observation is located at a point in space defined by its measurements





Can we describe these points in fewer dimensions while still retaining as much as possible of the shape of this point cloud?

https://learnche.org/pid/latent-variable-modelling/principal-component-analysis/visualizing-multivariate-data

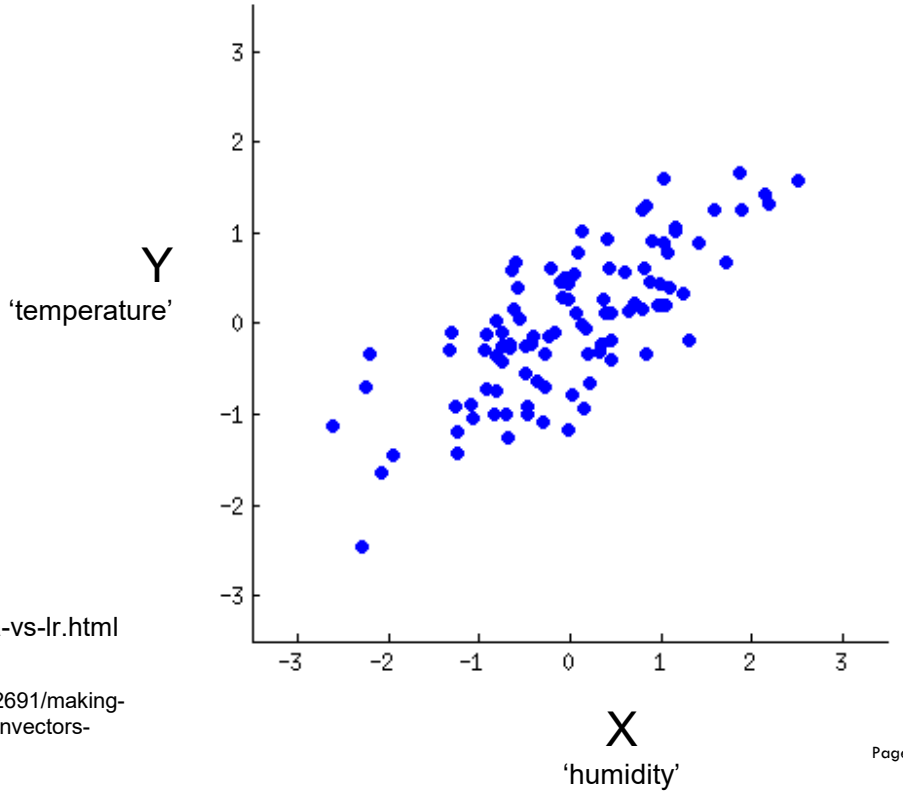
Challenge question: Dimension reduction

The scatter plot on the right represents measurements on two variables (X and Y). For illustrative purposes, lets say that Y is temperature and X is humidity at midday for some set of days. The variables have been mean centred, so each measurement is +ve if above the mean and –ve if below.

We want to create a new scale that measures 'stickiness', some combination of X and Y that will preserve the most information, so that points that had relatively high measurements on both *original variables have higher numbers in the new scale.*

Draw a line through the points that will form a new axis (or measurement scale). All points will 'fall on to the line' by following a perpendicular path to this new axis.

Is this line of best fit, like that in a simple linear regression? No. See: https://shankarmsy.github.io/posts/pca-vs-lr.html



https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

The PCA approach

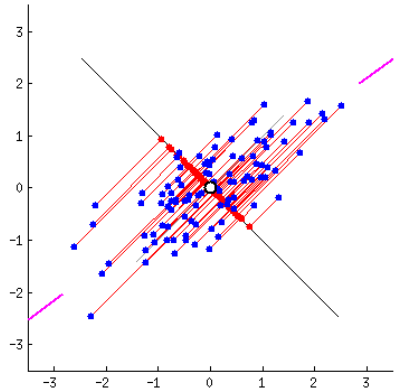
- PCA chooses a new coordinate system, by choosing a series of 'principal directions'. Projection of your original data points along these directions form Principal Components (PCs)

- How are the principal directions chosen?
  - Maximises variation in the initial principal component (PC), PC1

  - All subsequent principal components are orthogonal/uncorrelated, while maximising variation sequentially

- **3D example here**

- **2D example:**



https://stats.stackexchange.com/questions/2691/making-sense-of-principal-component-analysis-eigenvectors-eigenvalues

What are the outputs from a PCA?

- Mathematically, each principal component is a linear combination or weighted sum of the original variables.
  - E.g. PC1 = Ax + By + Cz … where A, B and C are scalar coefficients/weights and x, y and z are points in the original coordinate system

- The **loadings** (coefficients)* allow us to calculate the projection of our data from the original variables to a **PC score** on each individual principal component

- The PCA outputs are mathematically related to your covariance/correlation matrix
  - The principal directions are **eigenvectors** of the covariance/correlation matrix
  - The corresponding **eigenvalues** quantify the variation captured in the relevant principal component

- Technical note: Different R packages provide either unscaled (eigenvectors), or scaled (loadings) as a summary. Terminology can vary, see https://stats.stackexchange.com/questions/143905/loadings-vs-eigenvectors-in-pca-when-to-use-one-or-another

An example dataset: Decathlon

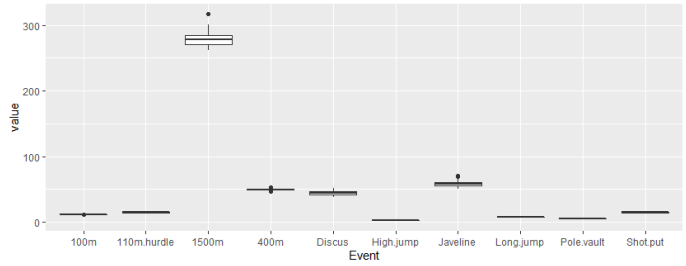
- Some of the concepts are best explained using a concrete example
- Let's use the **decathlon** data set, which contains the performance of athletes in the 10 events of the decathlon
  - Track events: seconds to complete distance
  - Field events: distance travelled in metres
- So we have data from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

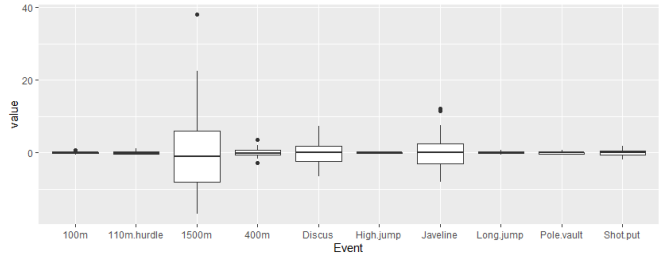
```
library(FactoMineR)
library(factoextra)
data(decathlon)
```

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

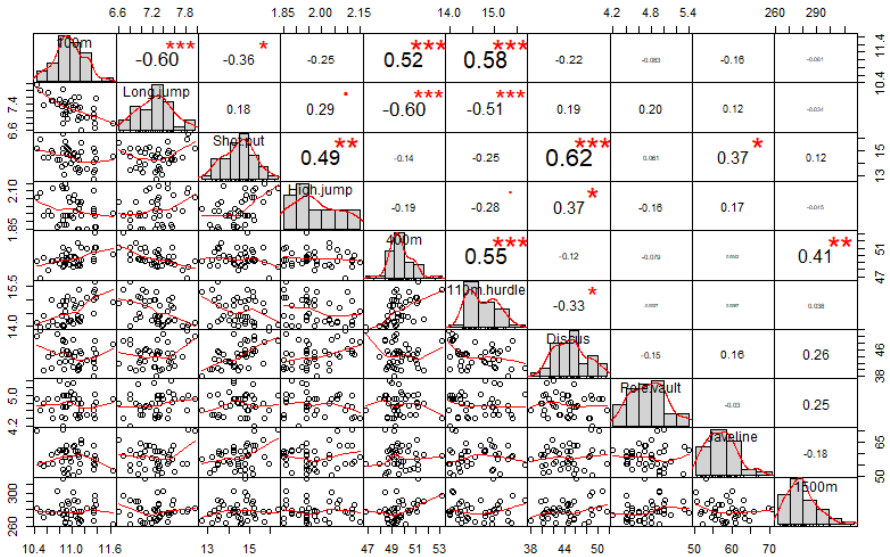




- We always mean centre the data (subtract the mean value from each variable) before PCA. Remember that we are using correlation, so the mean must be subtracted otherwise biases are introduced.
- It is also clear in this example that the different variables are on a different scale of measurement (different 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))
```

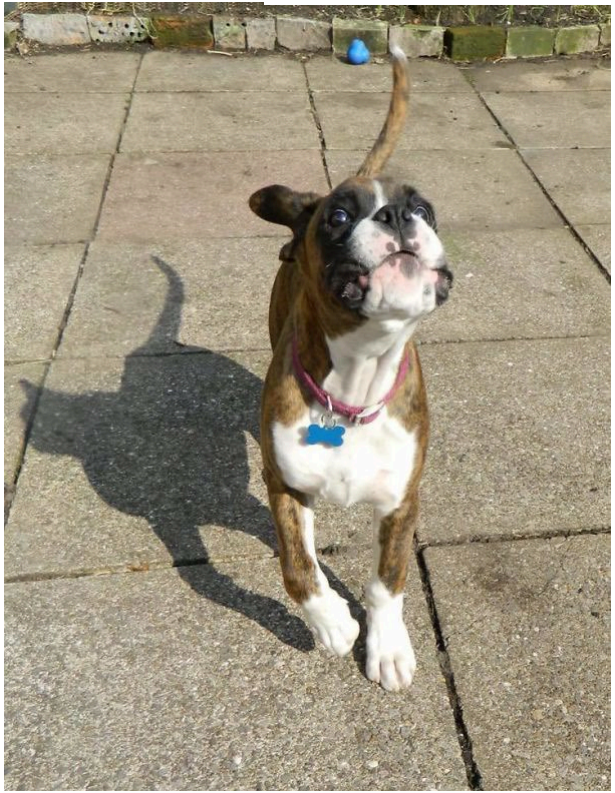
- Clear correlations between the similar field events, and similar track events
- Negative correlations reveal the track events have a different polarity to the field events
  - A bigger number is a good thing in long jump (you jumped further)
  - A bigger number is a bad thing in 100m (you took longer)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

Step 2 and 3 examine plots

- We can't see in more than 3 dimensions, but if you have a 3D space we can visualise what projections of different vectors look like onto a 2D plane.
- The length of some vectors (D, E) is close to the length of their projections, while for others (C), the projection length is much shorter. This depends on the size of the vector and the angle between the vector and the plane
- The closer the actual vector length is to the length of its projection, the more accurately it is being represented by its projection

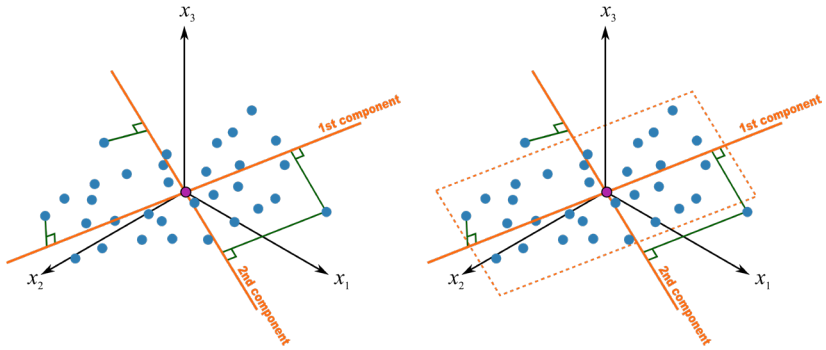


https://www.youtube.com/watch?v=msGKYXVdMxc

Step 2 and 3 examine plots

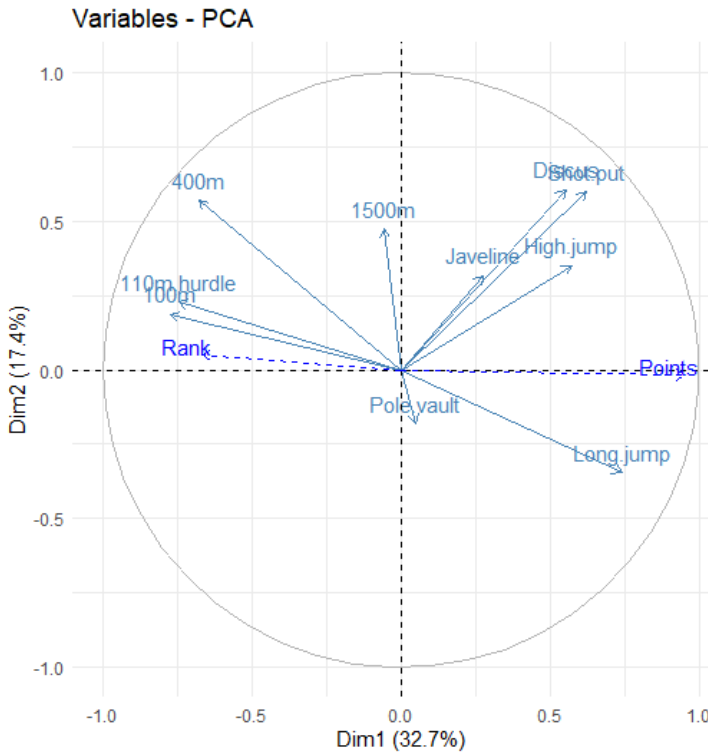
- Projection is an important geometrical concept to understand in dimension reduction

- In a (PCA-style) loading plot:
  - The vectors represent original variables
  - The orientation of the plane is the PCA solution (principal directions)
  - Projection of the variables (shown as vectors) shows us the **loadings plot**
  - Subjects can also be projected onto the plane (shown as points) **subjects plot**
  - Or both variables and subjects **biplot**



https://learnche.org/pid/latent-variable-modelling/principal-component-analysis/geometric-explanation-of-pca

Decathlon Dataset. PCA step 2: Examine the variable plot



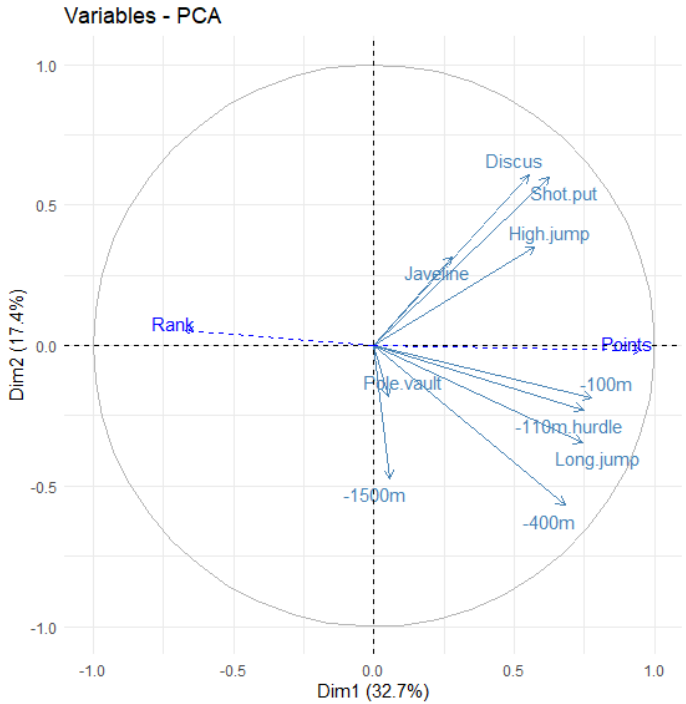
- Let's run the PCA and look at the loading plot
- This loading plot shows what happens if we didn't decide to multiply all track times by -1 (something we picked up in EDA). The 'polarity' problem separates track events and field events on this plot, which makes this plot more difficult to interpret
- 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).

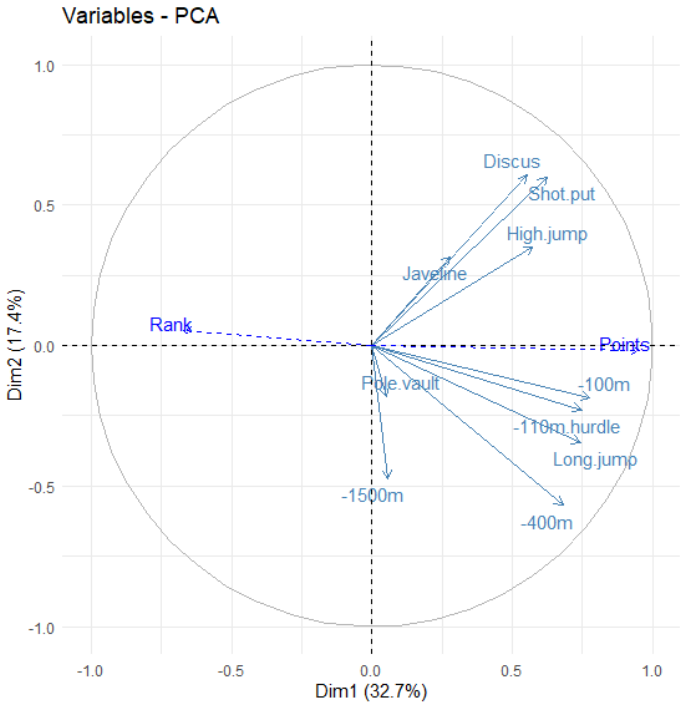
- Those original variables with shorter vectors (e.g. Pole Vault) are less loaded on to the PCs. Recall that the principal directions chosen in PCA preserve as much information as possible, so the variables poorly represented are less important for explaining variability overall in your data



PCA Step 2: Examine the variable plot

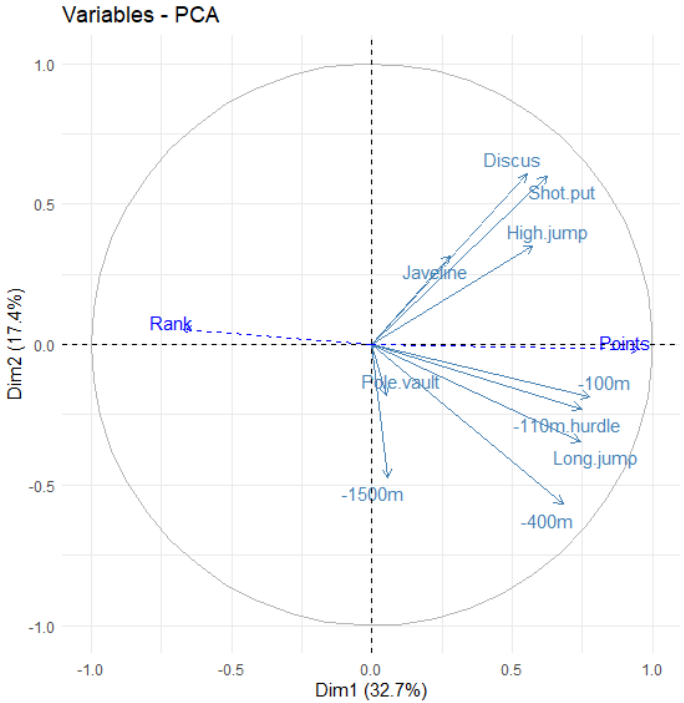
- Sometimes you may have supplementary variables available that are not used in PCA analysis but potentially correlate with PCs. In this case we can use the 'Points' and 'Rank' which give the number of points achieved and the Rank of that athlete. These can be added to the loading plot by calculating 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.



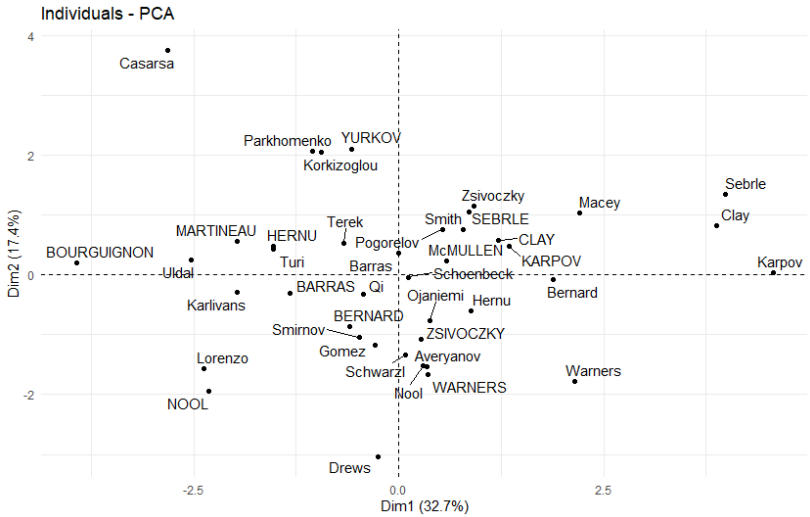
PCA Step 2: Examine the variable plot

- 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, PC2 has a more interesting interpretation *as a latent variable*. If specialisation is not present, this PC will not appear reliably.



*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.

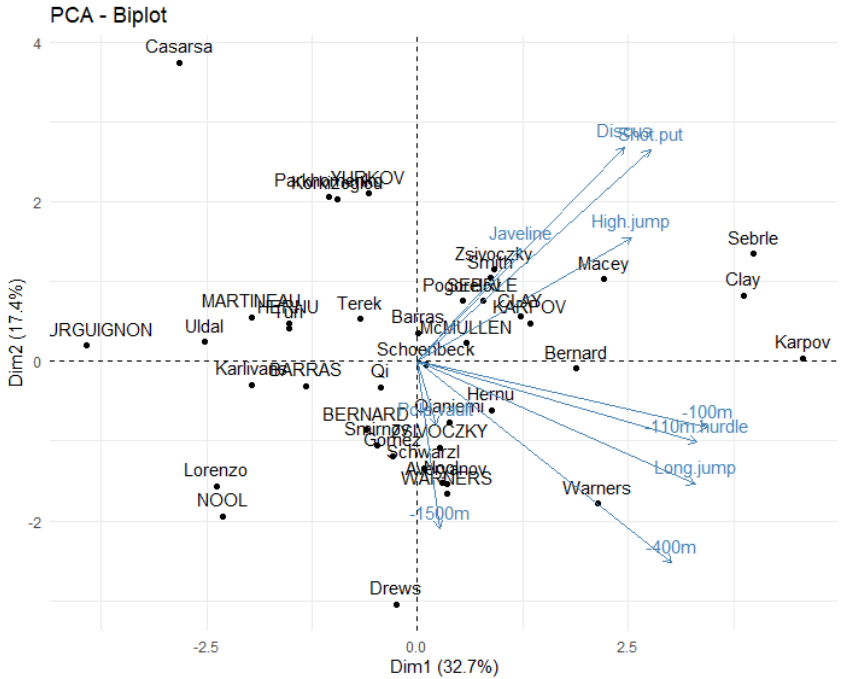
PCA Step 3: Examine the subject plot



- We can plot the individual observations plotted in PC space, often called the score plot. Their PC scores for 2 dimensions at a time
- This is again a projection into PC space, but not of the variables, but instead the individual measurements
- 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.

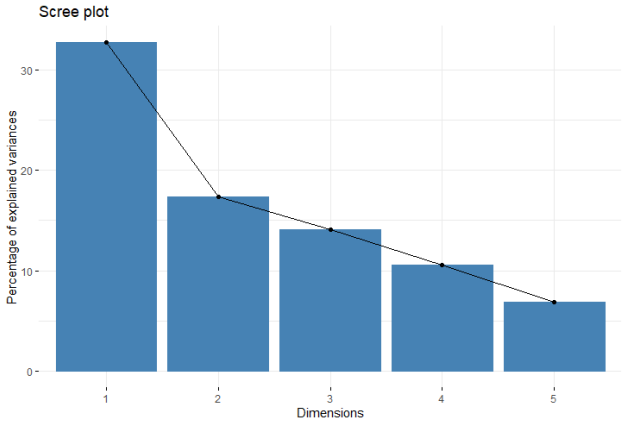
PCA Step 2 & 3: Examine the variable and subject plot

- 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

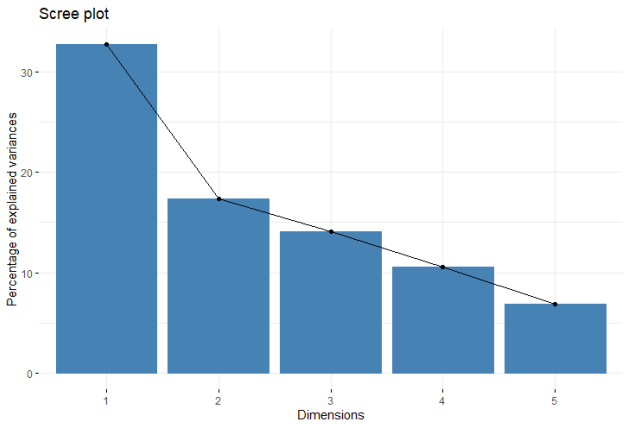
How many PCs should we consider?



- We have seen that the first 2 PCs capture about 50% of the total variance and we have plausible latent variable interpretation for each
- There are as many PCs as there were variables from the output
- Each PC captures some amount of the total variance (across all original variables = across all dimensions), and this decreases for each subsequent PC
- The size of the eigenvalue corresponds to the variance of each PC. We can divide by the total variance to find the % total variance of each PC.

```
res.pca = PCA(decathlon[,1:10], scale.unit=TRUE, ncp=5)
fviz_screeplot(res.pca, ncp=5)
```

How do you choose how many PCs?

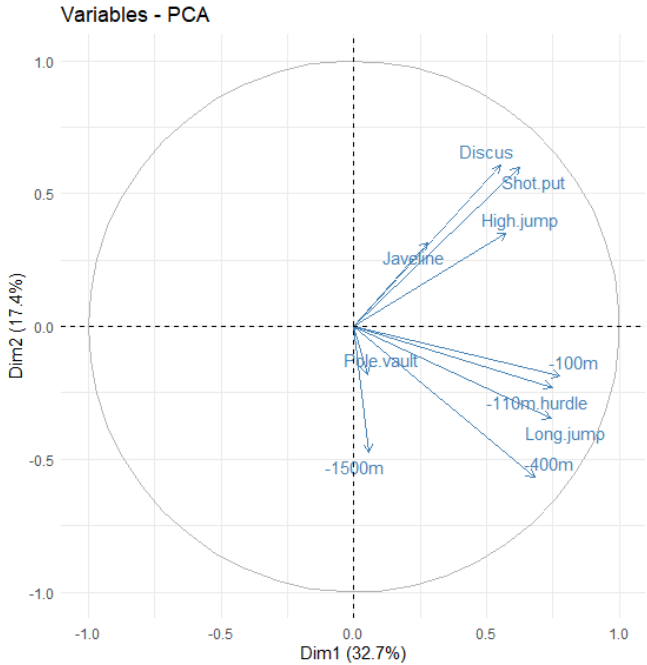


- A 'scree plot' is a plot of % of total variance captured in each PCA
- The steeper the drop between bars, the less information is being captured in the kth PCs compared to the first (k – 1) PCs. We can consider dropping the kth and higher PCs
- Rules of thumb:
  - Smallest # of PCs that together hold 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
- 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'.

```
fviz_screeplot(res.pca, ncp=5)
```

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?

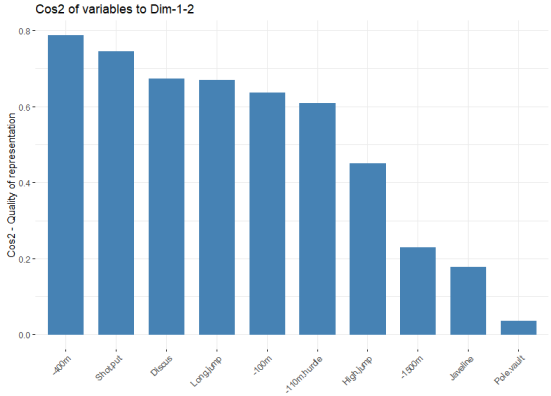


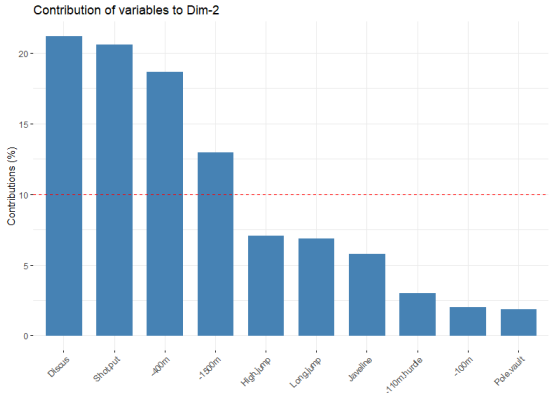
Challenge question – correlation between PCs

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)





PCA Downstream analysis

- Possible uses of PCA downstream:
  1. Pre-processing for clustering algorithms
  2. Looking for outliers or time dependency in subjects
  3. As predictors in multiple regression (PCR-Regression)
  4. As a pre-processing step for further dimension reduction

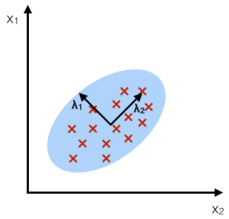
Method Extension: Linear Discriminant Analysis (LDA)

- PCA can be modified towards a different goal by imposing a different constraint. *Recall that in [classic] PCA we want to maximise variability on each principal component*

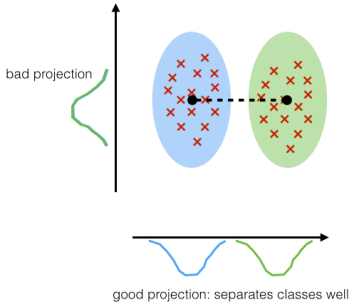
- If your goal is to find principal components that separate two or more different classes of subject, you could use LDA

- It is a modified version of PCs. LDA calculates PCs that maximise the separability between your classes

PCA:
component axes that maximize the variance



LDA:
maximizing the component axes for class-separation

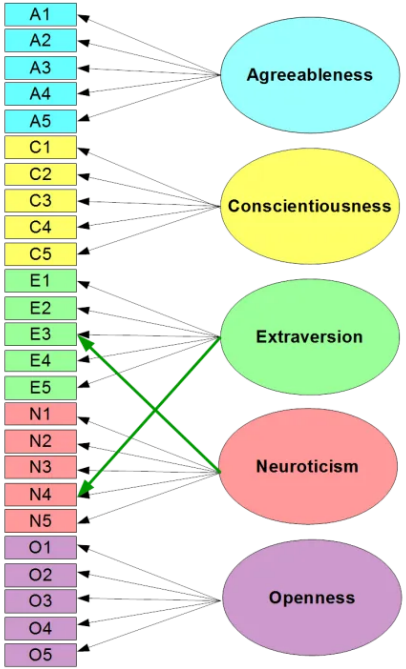


https://sebastianraschka.com/Articles/2014_python_lda.html

Factor Analysis (FA)

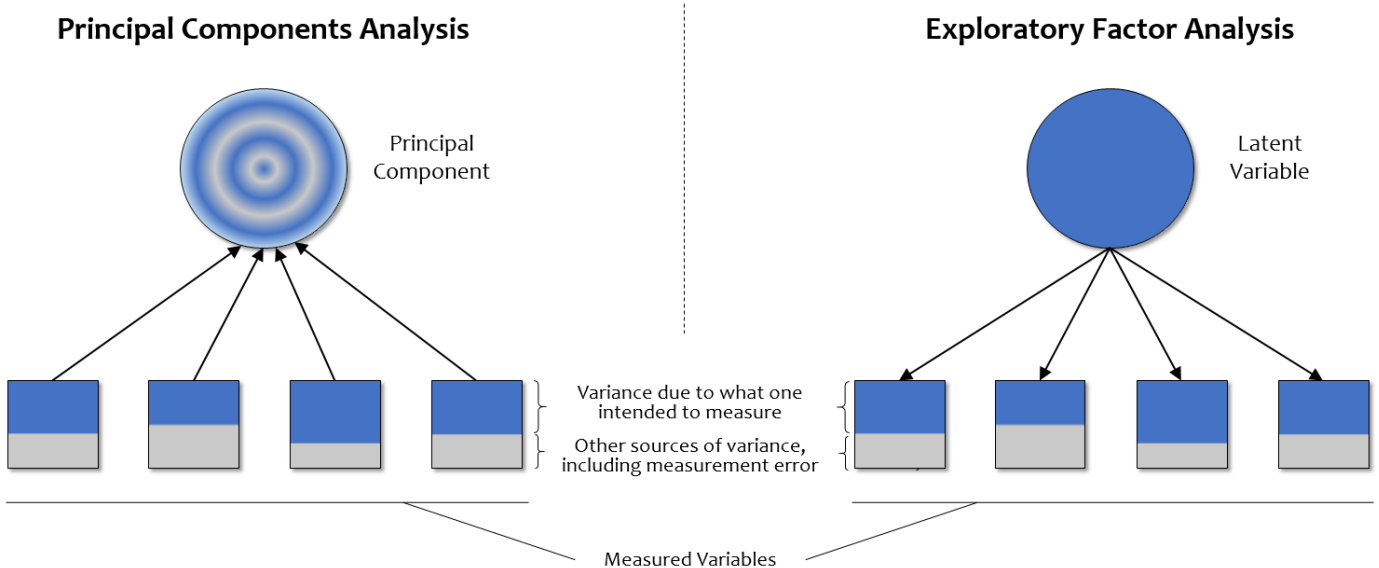
Factor Analysis

- Another major dimension reduction technique with continuous variables as input
- Similar to PCA in many respects, but differences in the methods reflect different aims and applications
- Most often used to define latent variables with the aim of producing an explicit, and testable model for how these interact with each other and measured variables
- Defining meaningful latent variables is usually more important when using this technique than mere data reduction
- Common in psychometrics. Things like personality traits can't be measured directly but are of interest. Think building a theoretical/functional model using a probabilistic model.



https://www.r-bloggers.com/2018/10/new-course-factor-analysis-in-r/

PCA vs. Factor Analysis



- PCA identified principal components, which capture patterns of variance in measured variables. You can think of any measurement error as being distributed throughout the PCs (but hopefully relegated mostly to PCs that are discarded)

- In factor analysis you will model error in your measured variables, and this is separate from the shared variance captured by each factor

https://community.jmp.com/t5/JMP-Blog/Principal-components-or-factor-analysis/ba-p/38347

PCA vs. Factor Analysis – a terminology note

- In SPSS the Factor Analysis menu lists PCA as a method. But most analysts would not consider PCA to be true Factor Analysis. Factor Analysis would only be when using a 'factor extraction method' that partitions variance of a measured variable into unique and common (see previous slide)

- Other differences:
  - Factor analysis uses an iterative algorithm to find an optimal solution, PCA uses direct calculation to find the single solution
  - Factor analysis loading plots do not resemble PCA loading plots, but instead show only the loading for each measured (observed) variable onto each factor
  - Factor analysis usually involves a rotation of the initial solution to maximise interpretability of the factors

- Some people refer to Factor Analysis as Common Factor Analysis, or Exploratory Factor Analysis (EFA) as distinct to CFA (Confirmatory Factor Analysis), or SEM (Structural Equation Modelling) where a defined model is tested for model fit

Factor Analysis Downstream analysis

- The goal of FA is to come up with an explicit and testable model for how measured variables and factors interact. So FA is usually the first step in a longer process.

- Typical downstream steps from Factor Analysis (FA):
  - Confirmatory Factor Analysis
  - Structural Equation Modelling

- These are referred to in our Surveys 2 workshop, as surveys are commonly used as the input for factor analysis

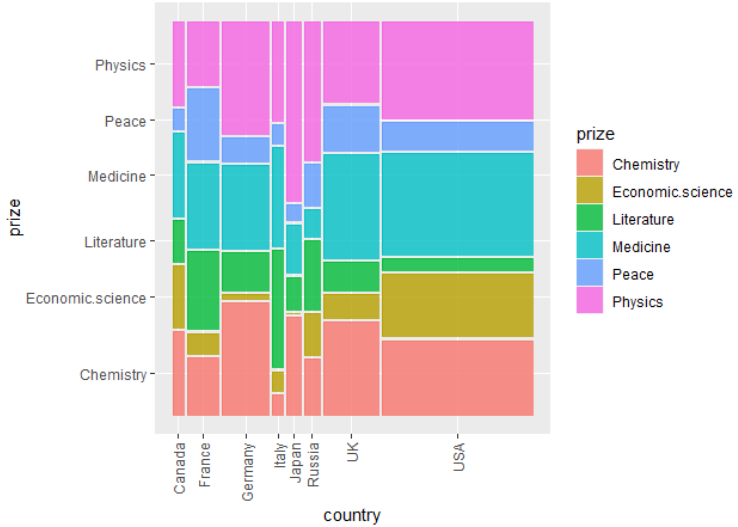
Correspondence Analysis (CA)

Correspondence Analysis (CA)

- Very common in surveys with categorical responses. Can also be used for abundance measurements in ecology.
- 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

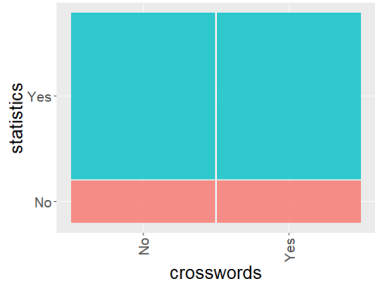
- Nobel prize winners by country from 1901 to 2015. Just G8 countries, and excluding mathematics.
- We could use a contingency table, or a mosaic plot as shown below.
  - The width of the columns represents the column proportions (country)
  - The area of the blocks represents the proportion of all prizes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



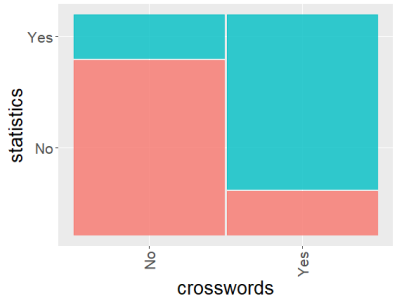
https://husson.github.io/MOOC.html

Association

- In an underlying population, there are two possible relationships between the two variables: independence, or association. Say we have two variables: whether the subject likes/dislikes statistics, and likes/dislikes crosswords.



INDEPENDENCE



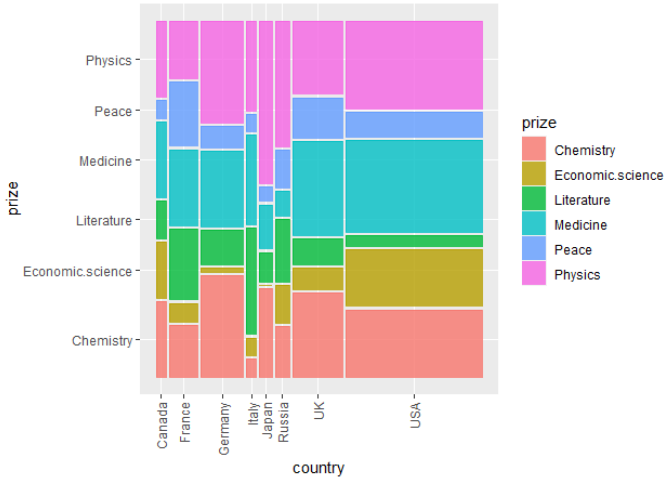
DEPENDENCE /association

- 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.

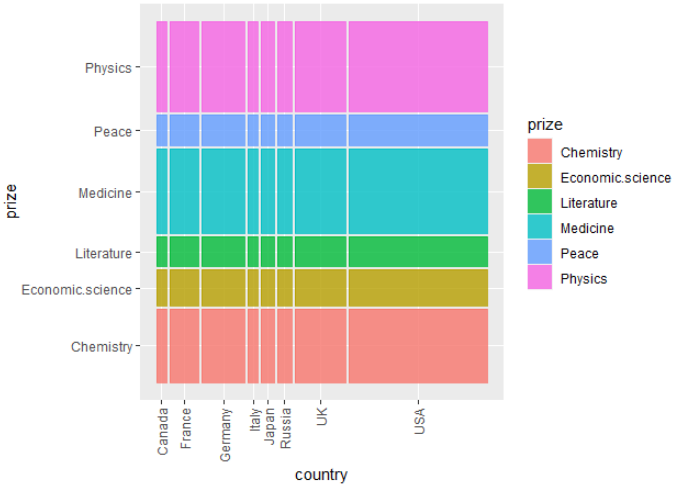
- But what about if there are more than two categories in either or both variables?

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 each combination of country and prize is calculated from the overall (marginal) proportion of each variable



Observed data

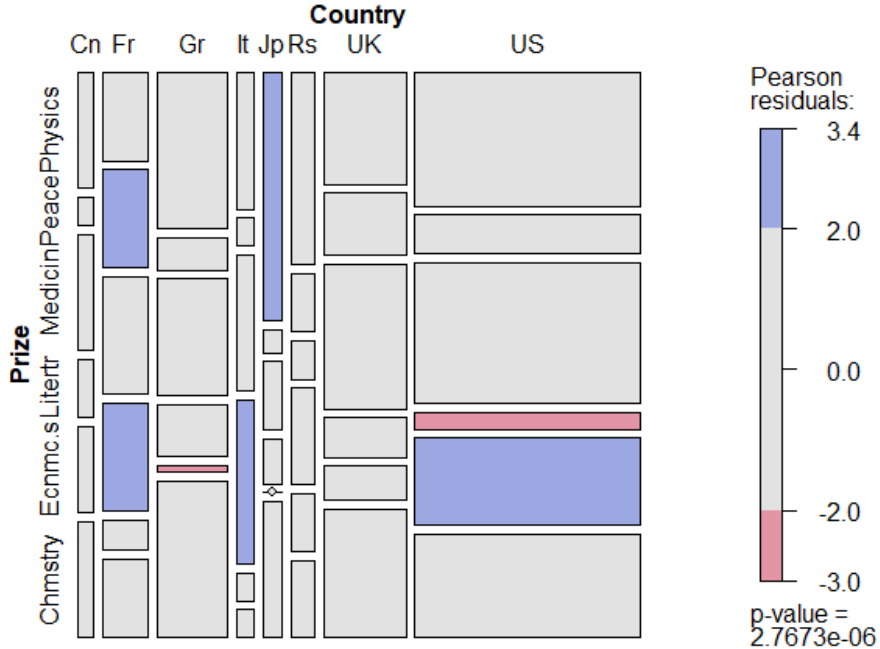


Expected under null hypothesis
INDEPENDENCE

Nobel Prize data Step 0: EDA

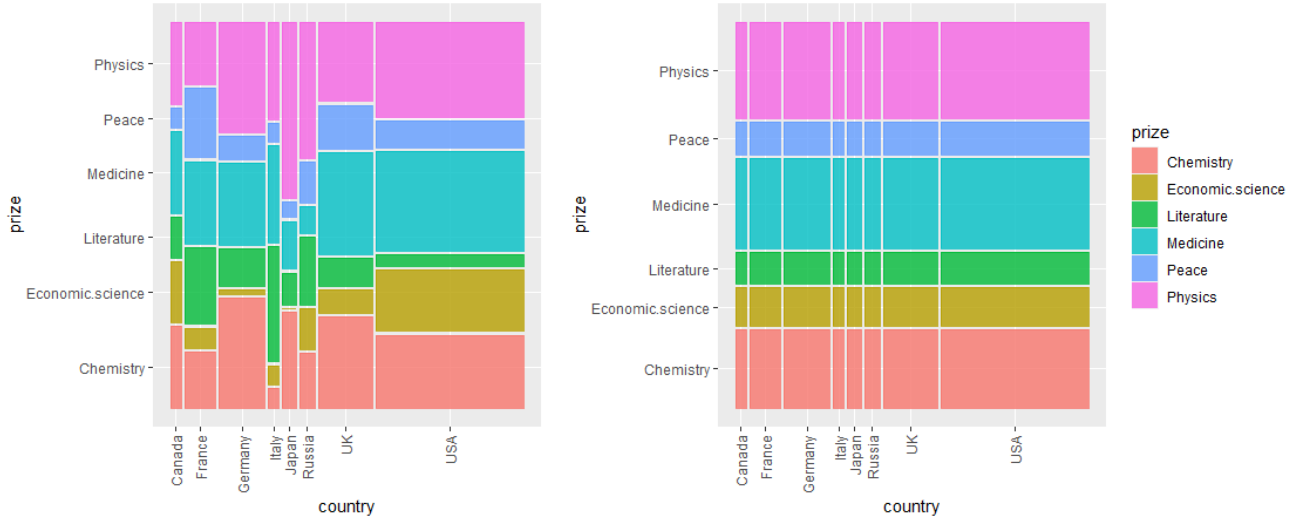
- We can perform a Chi-squared test for independence of the two variables (very strong evidence of association for these data)

- 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?

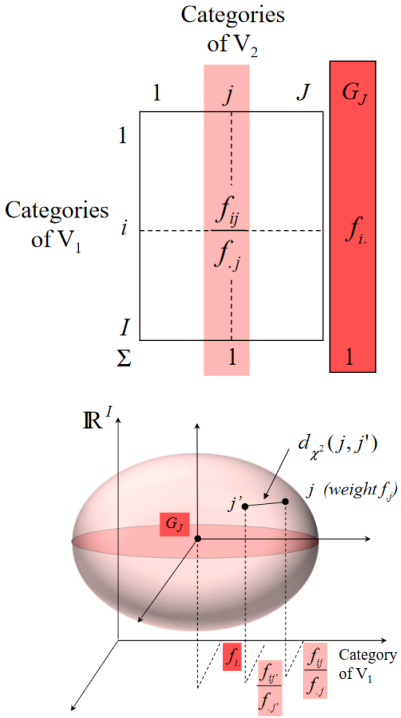
- *Can these frequencies be treated as continuous data?*
- *For each country (column), what is its profile? (Relative frequency of each prize)*

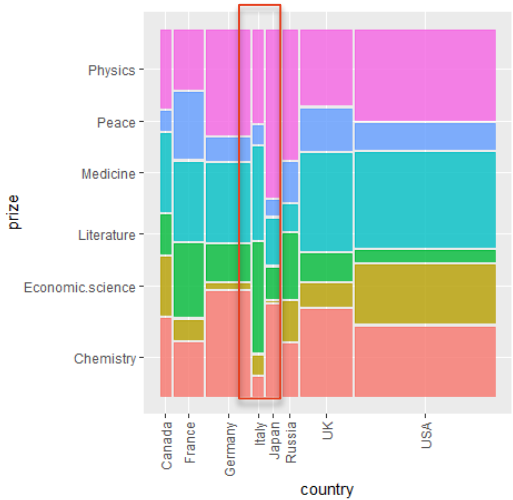



```
nobel_sums_long %>% mutate(Country=fct_relevel(Country,"Total",after = Inf)) %>% ggplot(aes(x=Country,fill=name,y=value)) +
geom_bar(position="fill",stat="identity") + scale_fill_brewer(palette = "Set2")
```

How does correspondence analysis work?

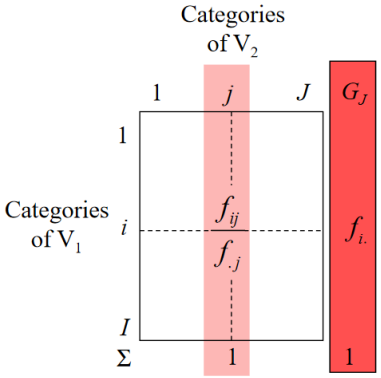
- Each column profile (country) is represented as a point in J-dimensional space, where J is the number of rows (prizes). The location depends on the relative frequency of each row.
- The mean column profile is represented as another point (G_i) that serves as a reference to observed column profiles
- 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)

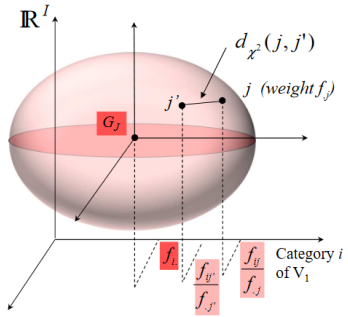


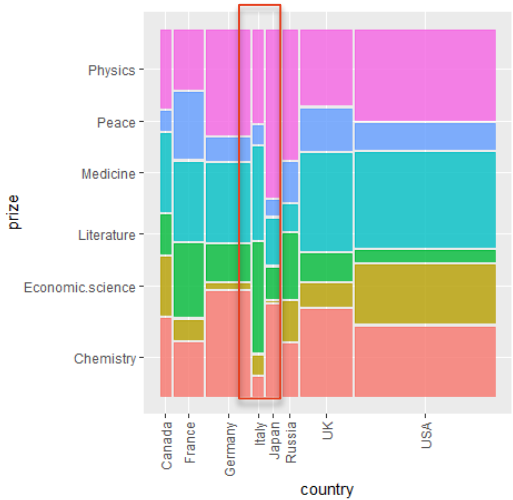


How does correspondence analysis work?

- 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.





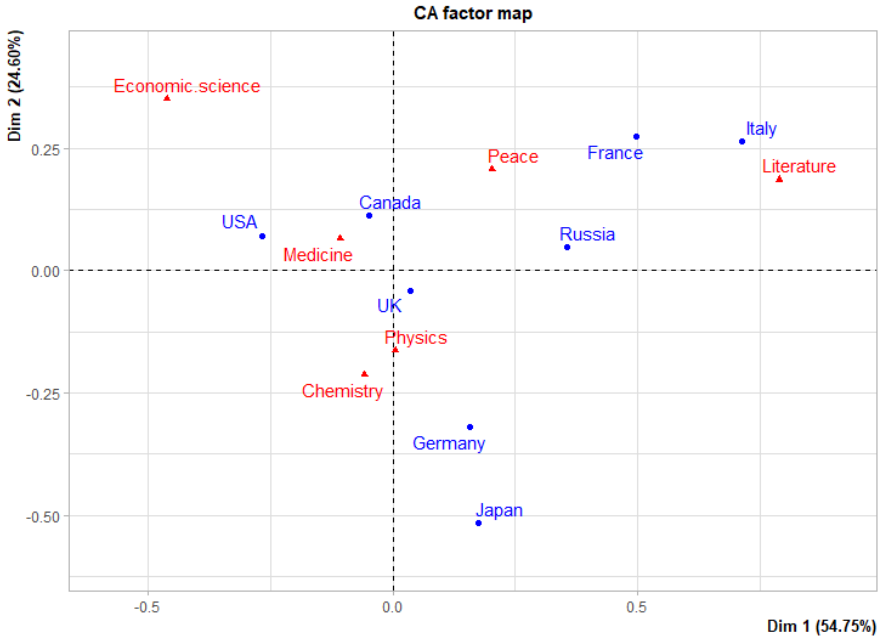


CA Steps 2&3: Correspondence Map

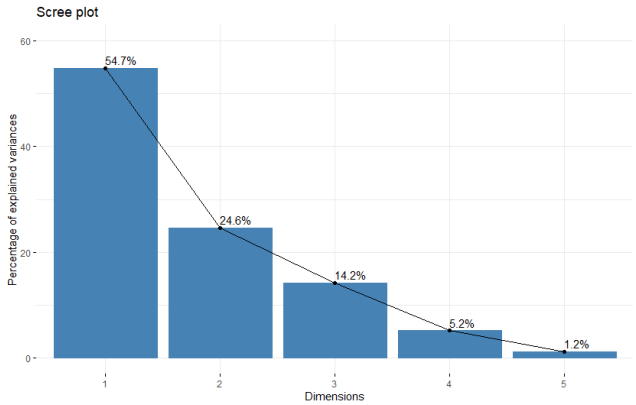
- This is the correspondence map for the Nobel prize data

- 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
```

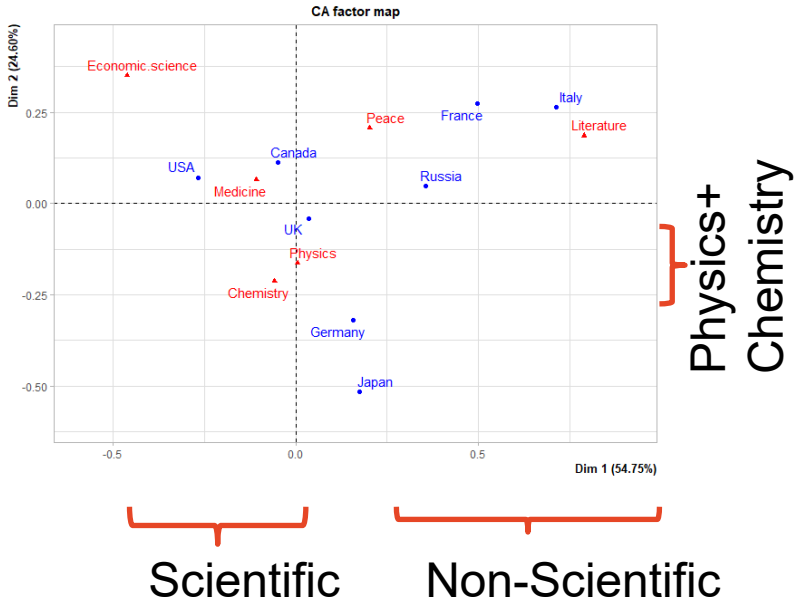
- Scree plot shows the percentage of inertia captured by each dimension
- The first two dimensions capture ~80% of total inertia

- 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]
```

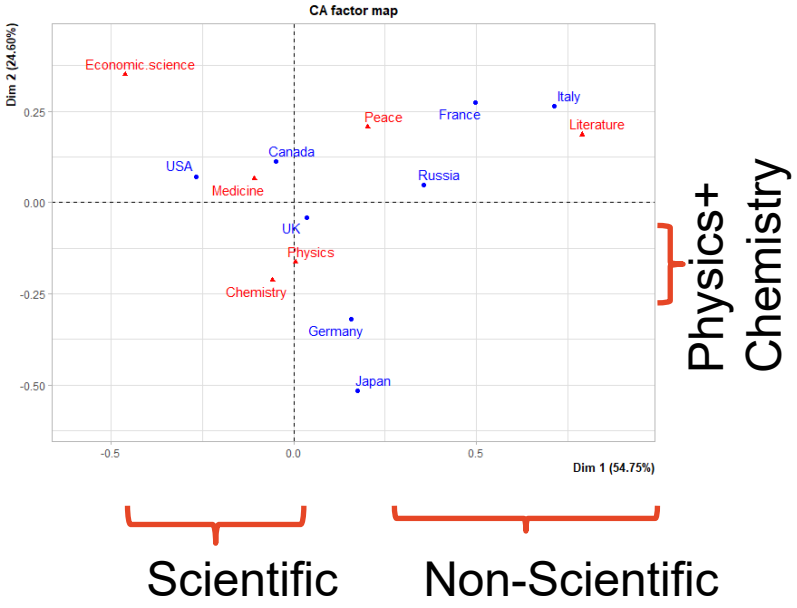
CA Step 2: Examine the 'variable' plot

- Dimension 1 appears to contrast the scientific prizes from non-scientific
- Dimension 2 appears to contrast 'hard sciences' Physics and Chemistry from the others



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
O	O	O	O

Do you have sugar in your tea?

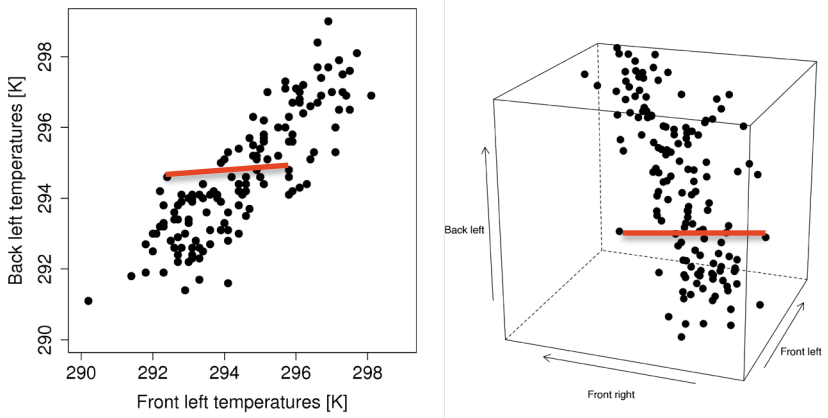
Yes	No
O	O

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

Dimensionality Reduction Techniques: Distance based

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 (ordination/embedding) that most accurately visualises the distances between the subjects, i.e. preserves as much as possible the distance information from higher dimensions



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)

- 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 subjects (e.g. site C is more dissimilar to site A than site C is dissimilar to site B).

ACTUAL DISTANCE

SAME DISSIMILARITY RANK

DIFFERENT DISSIMILARITY RANK

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
- 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)
- 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)

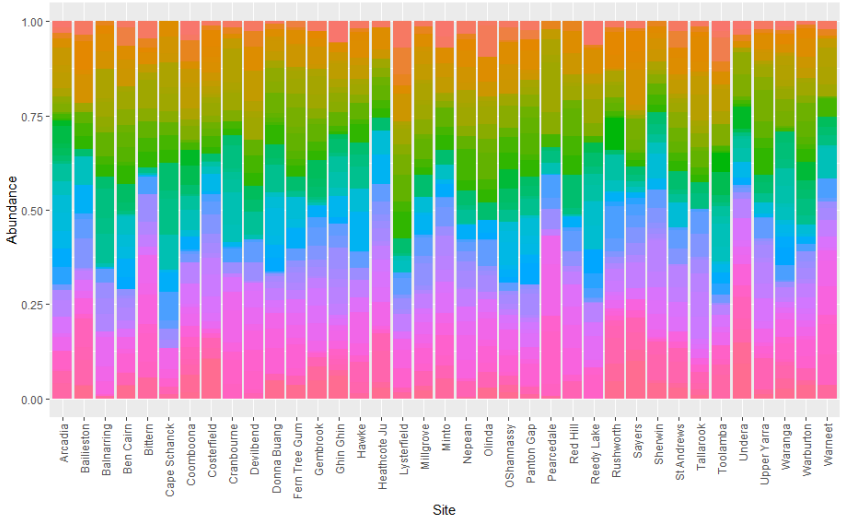
- Research question: Do the bird assemblages differ between forest types?

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

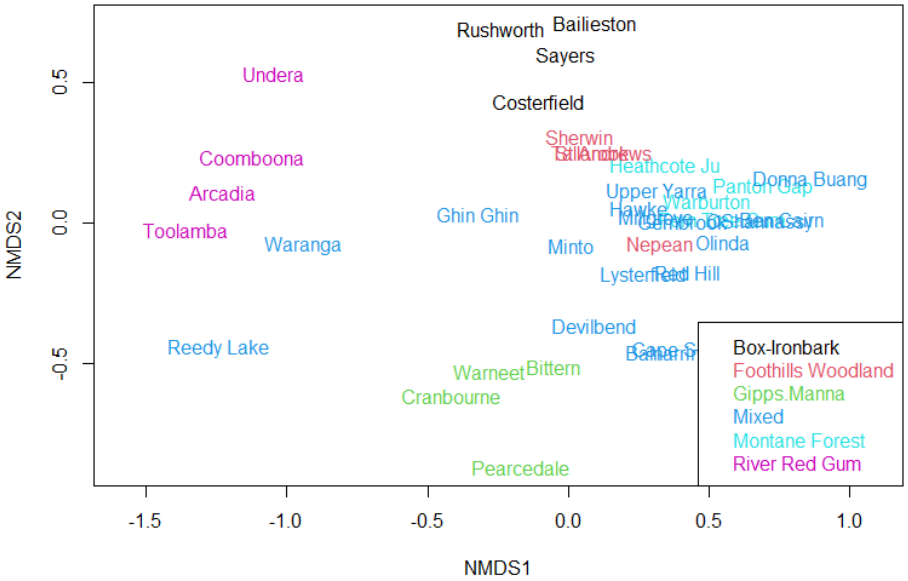
Bird abundance Step 0: EDA

- 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

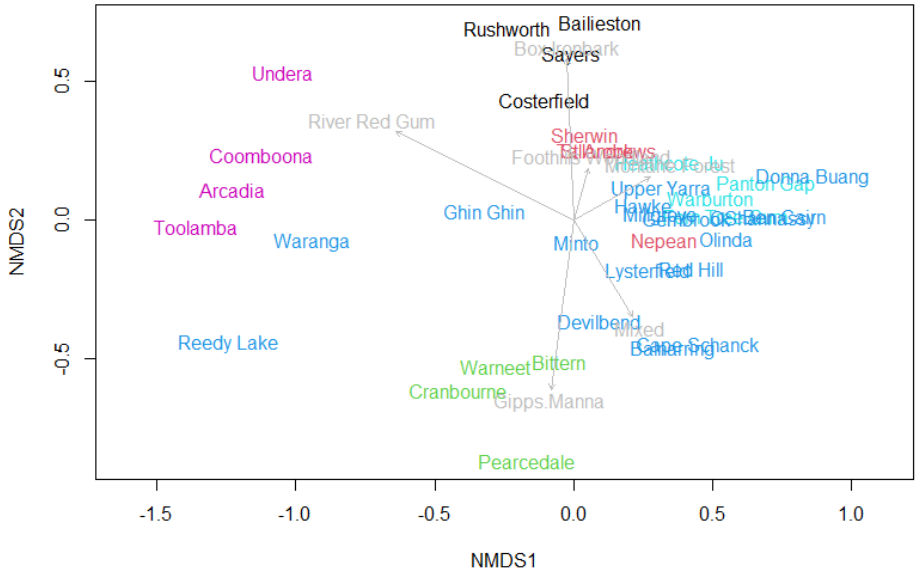


MDS Step 2: Examine the subject (site) plot



- 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 2: 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 subject (site) plot

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

```
> adonis(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
```

- Which forest types are different to 'mixed forest'

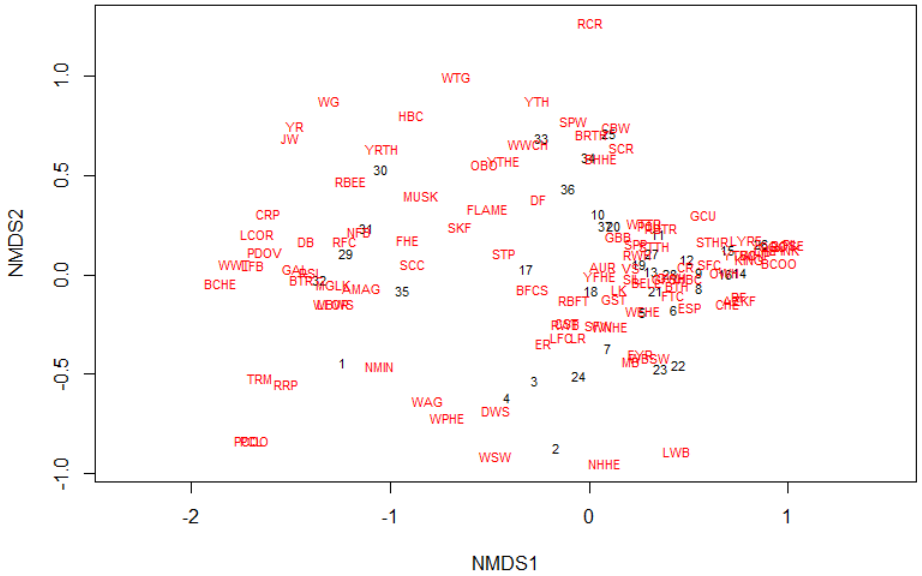
```
Call:
adonis(formula = macnally.dist ~ Box.Ironbark +
Foothills.Woodland +      Gipps.Manna +
Montane.Forest + River.Red.Gum, data = mm,
contr.unordered = "contr.treat")

Terms added sequentially (first to last)

                  Df SumsOfSqs MeanSqs
Box.Ironbark       1      0.70   0.702
Foothills.Woodland 1      0.34   0.336
Gipps.Manna        1      0.72   0.722
Montane.Forest     1      0.36   0.363
River.Red.Gum      1      1.37   1.372
Residuals         31      4.60   0.148
Total             36      8.09
                  F.Model    R2 Pr(>F)
Box.Ironbark         4.74 0.087  0.001
Foothills.Woodland   2.27 0.042  0.030
Gipps.Manna          4.87 0.089  0.003
Montane.Forest       2.45 0.045  0.016
River.Red.Gum        9.25 0.170  0.001
Residuals                 0.568
Total                     1.000
```

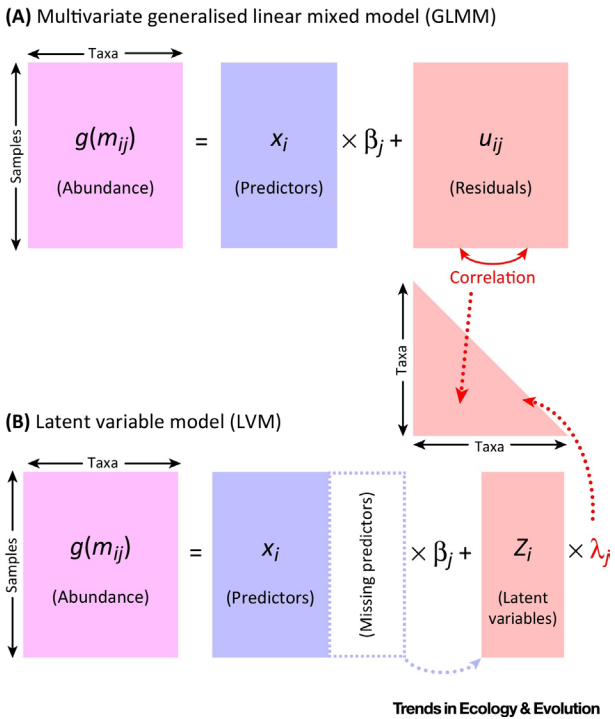

https://stats.stackexchange.com/questions/316590/what-is-the-reason-behind-the-name-adonis-for-permutational-manova

MDS Step 3: 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

Method extension: Joint species distribution models



- Model-based ordination methods have recently been developed
- These methods handle the challenges of abundance data (non-normality, 0 observations) and allow multivariate [joint] regression models to be fit to this type of data, with environmental variables as predictors
- These methods go beyond dimension reduction techniques, incorporating dimension reduction within a modelling approach

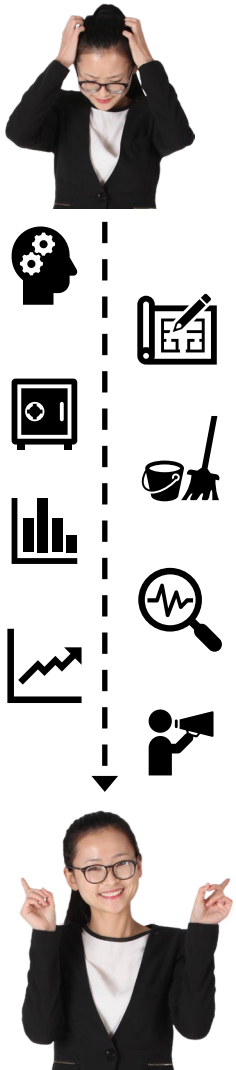
Warton DI, Blanchet FG, O'Hara RB, Ovaskainen O, Taskinen S, Walker SC, Hui FKC. So Many Variables: Joint Modeling in Community Ecology. Trends Ecol Evol. 2015 Dec;30(12):766-779. doi: 10.1016/j.tree.2015.09.007. Epub 2015 Oct 28. PMID: 26519235.

Conclusions

Dimension Reduction in the General Research Workflow

1. **Hypothesis Generation (Research/Desktop Review)**
2. Experimental and Analytical Design (sampling, power, ethics approval)
3. Collect/Store Data
4. Data cleaning
5. Exploratory Data Analysis (EDA)
6. Data Analysis aka inferential analysis
7. Predictive modelling
8. Publication

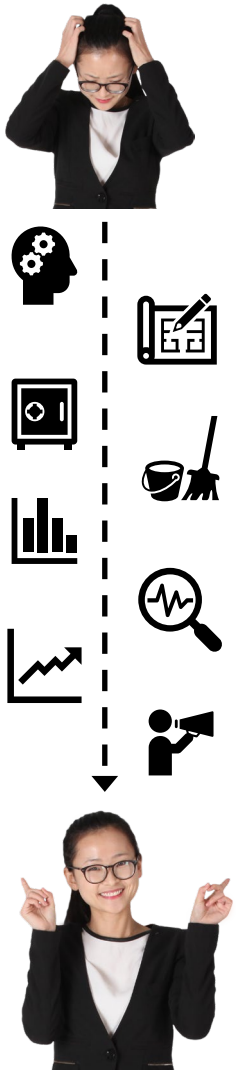
Discovery projects to generate new hypotheses e.g. new cell populations for further characterisation from single-cell expression data



Dimension Reduction in the General Research Workflow

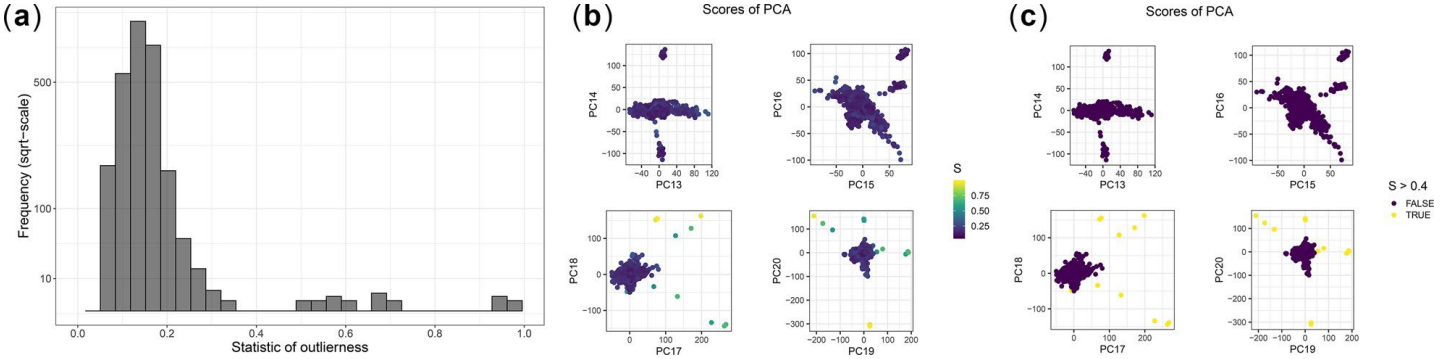
1. Hypothesis Generation (Research/Desktop Review)
2. Experimental and Analytical Design (sampling, power, ethics approval)
3. Collect/Store Data
4. **Data cleaning**
5. Exploratory Data Analysis (EDA)
6. Data Analysis aka inferential analysis
7. Predictive modelling
8. Publication

Multivariate techniques are an important part of QC (Quality Control) in high-throughput biology



Dimension Reduction in the C

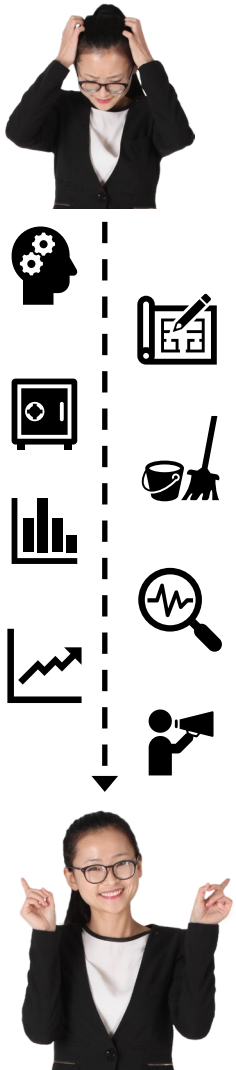
Fig. 2. Outlier detection in the 1000 Genomes (1000G) project, using prob_dist (Section 3.4)



Bioinformatics, Volume 36, Issue 16, 15 August 2020, Pages 4449–4457, https://doi.org/10.1093/bioinformatics/btaa520

Dimension Reduction in the General Research Workflow

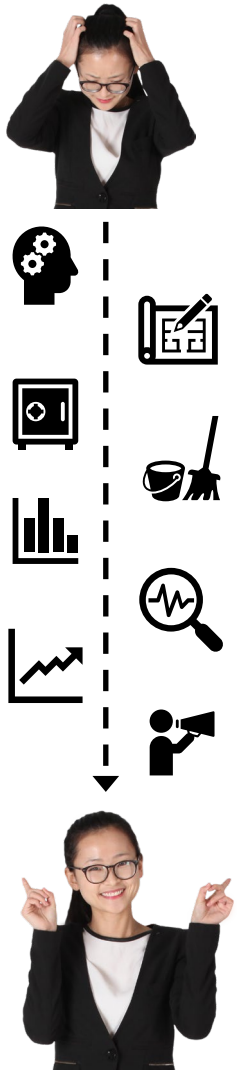
1. Hypothesis Generation (Research/Desktop Review)
2. Experimental and Analytical Design (sampling, power, ethics approval)
3. Collect/Store Data
4. Data cleaning
5. **Exploratory Data Analysis (EDA)**
6. Data Analysis aka inferential analysis
7. Predictive modelling
8. Publication



An important component of the research workflow. When there are multiple predictors, characterising the higher-order relationships between them is useful. See **Model Building Workshop.**

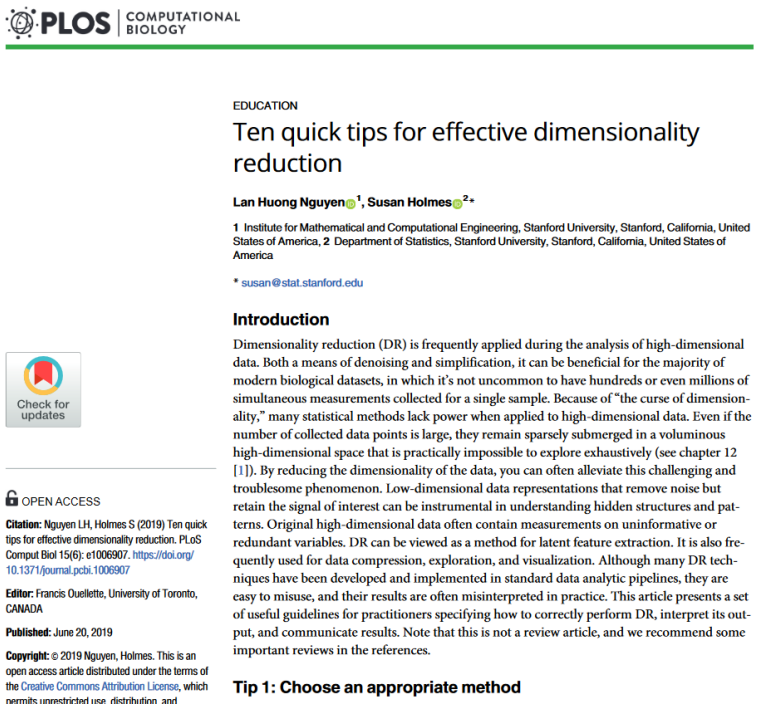
Dimension Reduction in the General Research Workflow

1. Hypothesis Generation (Research/Desktop Review)
2. Experimental and Analytical Design (sampling, power, ethics approval)
3. Collect/Store Data
4. Data cleaning
5. Exploratory Data Analysis (EDA)
6. **Data Analysis aka inferential analysis**
7. **Predictive modelling**
8. Publication



PCA regression: perform linear regression in the presence of a high level of multicollinearity (see **Model Building Workshop**)

Choosing a method



PLOS COMPUTATIONAL BIOLOGY

EDUCATION

Ten quick tips for effective dimensionality reduction

Lan Huong Nguyen[1], **Susan Holmes**[2]*

1 Institute for Mathematical and Computational Engineering, Stanford University, Stanford, California, United States of America, 2 Department of Statistics, Stanford University, Stanford, California, United States of America

* susan@stat.stanford.edu

Introduction

Dimensionality reduction (DR) is frequently applied during the analysis of high-dimensional data. Both a means of denoising and simplification, it can be beneficial for the majority of modern biological datasets, in which it's not uncommon to have hundreds or even millions of simultaneous measurements collected for a single sample. Because of "the curse of dimensionality," many statistical methods lack power when applied to high-dimensional data. Even if the number of collected data points is large, they remain sparsely submerged in a voluminous high-dimensional space that is practically impossible to explore exhaustively (see chapter 12 [1]). By reducing the dimensionality of the data, you can often alleviate this challenging and troublesome phenomenon. Low-dimensional data representations that remove noise but retain the signal of interest can be instrumental in understanding hidden structures and patterns. Original high-dimensional data often contain measurements on uninformative or redundant variables. DR can be viewed as a method for latent feature extraction. It is also frequently used for data compression, exploration, and visualization. Although many DR techniques have been developed and implemented in standard data analytic pipelines, they are easy to misuse, and their results are often misinterpreted in practice. This article presents a set of useful guidelines for practitioners specifying how to correctly perform DR, interpret its output, and communicate results. Note that this is not a review article, and we recommend some important reviews in the references.

Tip 1: Choose an appropriate method

Check for updates

OPEN ACCESS

Citation: Nguyen LH, Holmes S (2019) Ten quick tips for effective dimensionality reduction. PLoS Comput Biol 15(6): e1006907. https://doi.org/10.1371/journal.pcbi.1006907

Editor: Francis Ouellette, University of Toronto, CANADA

Published: June 20, 2019

- I've given you a broad overview of basic dimension reduction methods
- There are many extensions to these methods and many other methods available
- Many excellent tutorials and other resources are available for implementing a particular method
- Also consider the choice of method after doing some basic EDA

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

Further Assistance at Sydney University

SIH

- Statistical Consulting website: containing our workshop slides and our favourite external resources (including links for learning R and SPSS)
- Hacky Hour an informal monthly meetup for getting help with coding or using statistics software
- 1on1 Consults can be requested on our website (click on the big red 'contact us' link)

SIH Workshops

- Create your own custom programmes tailored to your research needs by attending more of our Statistical Consulting workshops. Look for the statistics workshops on our training page.
- Other SIH workshops
- Sign up to our mailing list to be notified of upcoming training

Other

- Open Learning Environment (OLE) courses
- Linkedin Learning

How to use our workshops

Workshops developed by the Statistical Consulting Team within the Sydney Informatics Hub form an integrated modular framework. Researchers are encouraged to choose modules to ***create custom programmes tailored to their specific needs.*** This is achieved through:

- **Short 90 minute workshops,** acknowledging researchers rarely have time for long multi day workshops.
- Providing **statistical workflows appliable in any software,** that give **practical step by step instructions which researchers return to when analysing and interpreting their data or designing their study** e.g. workflows for designing studies for strong causal inference, model diagnostics, interpretation and presentation of results.
- Each one focusing on a specific statistical method while also integrating and referencing the others to give a **holistic understanding of how data can be transformed into knowledge from a statistical perspective** from hypothesis generation to publication.

For other workshops that fit into this integrated framework refer to our training link page under statistics https://www.sydney.edu.au/research/facilities/sydney-informatics-hub/workshops-and-training.html#stats

We recommend our Experimental Design and Sample Size Workshops

Experimental Design Workshop

- Far too many researchers think they know all they need to in this area. We commonly see designs that could be substantially improved for stronger causal inference and improved results which leads to publication in higher impact journals (amongst other benefits).
- Even if you have already collected your data it is well worth attending since it may improve your write up and analysis e.g. we had a client who didn't realise they had a very strong Before/After Control/Impact (BACI) design.

Sample and Power Workshop

- Shows the steps and decisions researchers need to make when designing an experiments to ensure sufficient sample e.g. Power, minimum required to fit the necessary model, etc.
- Also how much Power the study has i.e. does it have sufficient power to detect the effects you expect to see, or is your study a complete waste of time and resources.

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

A reminder: 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."

We value your feedback

We want to hear about you and whether this workshop has helped you in your research. What **worked** and what **didn't work.**

We actively use the feedback to improve our workshops.

Completing this survey really does help us and we would appreciate your help! It only takes a few minutes to complete (*promise!*)

You will receive a link to the anonymous survey by email

Glossary of Terms

- **Variance** is simply variability of a single variable. Similar to modelling, unexplained variation is a bad thing because it contributes to the error of our estimates, but explained variation is a good thing because we can attribute that variability to a particular factor, and turn 'noise' into a meaningful 'signal'. When we try to "maximise variability" in PCA, you can think about this as trying to maximise explained variability or "information".

- **Total Variance/Total inertia** usually refers to the variance or inertia summed across all variables/dimensions

- **Dimension/Components/Factors** Dimension reduction techniques employ linear algebra, which often uses terminology from geometry. Mathematically, we can think about each original variable, or synthetic variable as being a dimension (I use 'dimension' to refer exclusively to the dimensions identified by the various dimension reduction methods These dimensions have differing names depending on the technique being used.)