Visualisation

Dr Gordon Ross

Visualisation

In any statistics or machine learning task – not just stylometry – it is usually a good idea to start by visualizing the data that you have available.

This allows potentially interesting patterns to be discovered, which may inform the ultimate choice of model

It is usually a bad idea to just jump into fitting models before you understand the data!

Visualisation

Recall that the full Iris data has 4 features – Sepal Length, Sepal Width, Petal Length, Petal Width.

If we pick any two of these, then we can plot them against each other to form a 2 dimensional scatter plot.

Let's first pick Sepal Length and Sepal Width

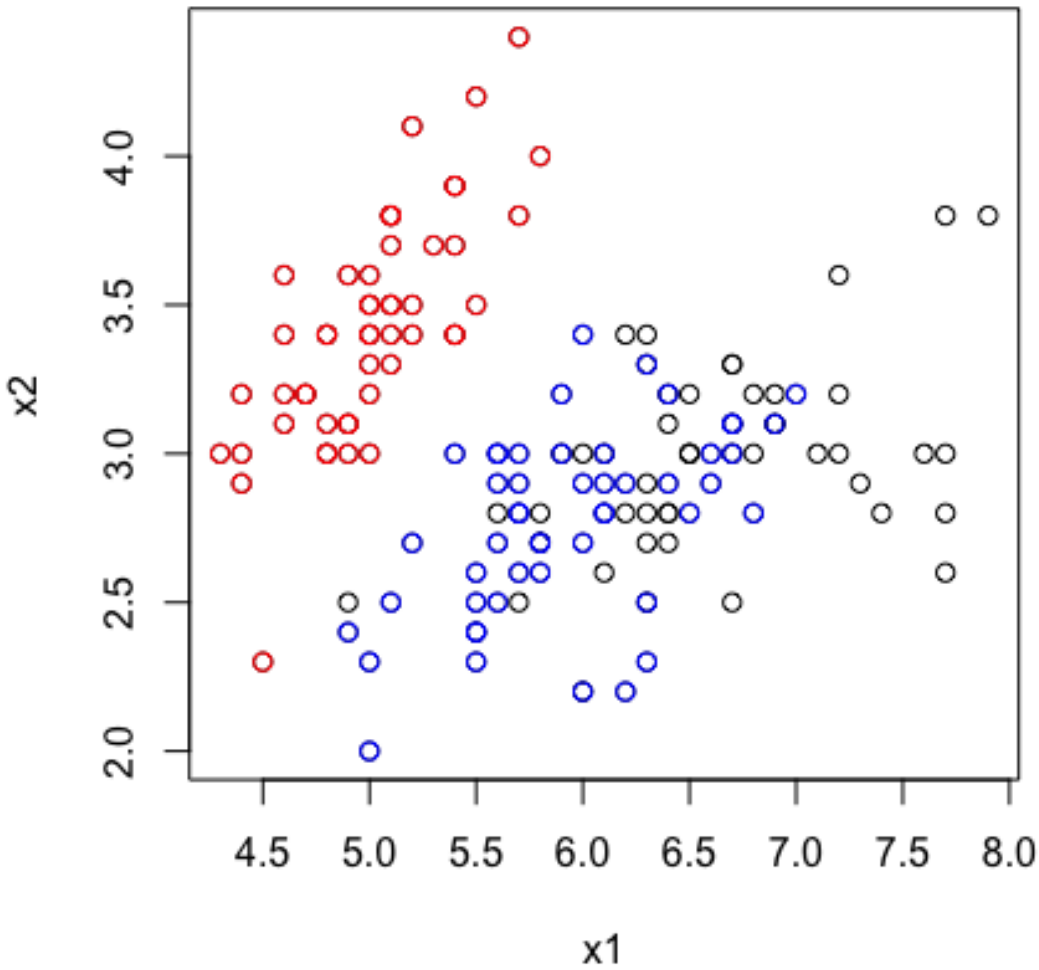
Visualisation

Example R code (note there are many other ways of doing this)

```
x1 <- iris$Sepal.Length;
x2 <- iris$Sepal.Width
plot(x1,x2)
inds <- which(iris$Species=='setosa')
points(x1[inds],x2[inds],col='red')
inds <- which(iris$Species=='versicolor')
points(x1[inds],x2[inds],col='blue')
```

The 'points' function is similar to plot() but it adds information to an existing plot. In this case, we are colouring the points based on the type of flower

Visualisation

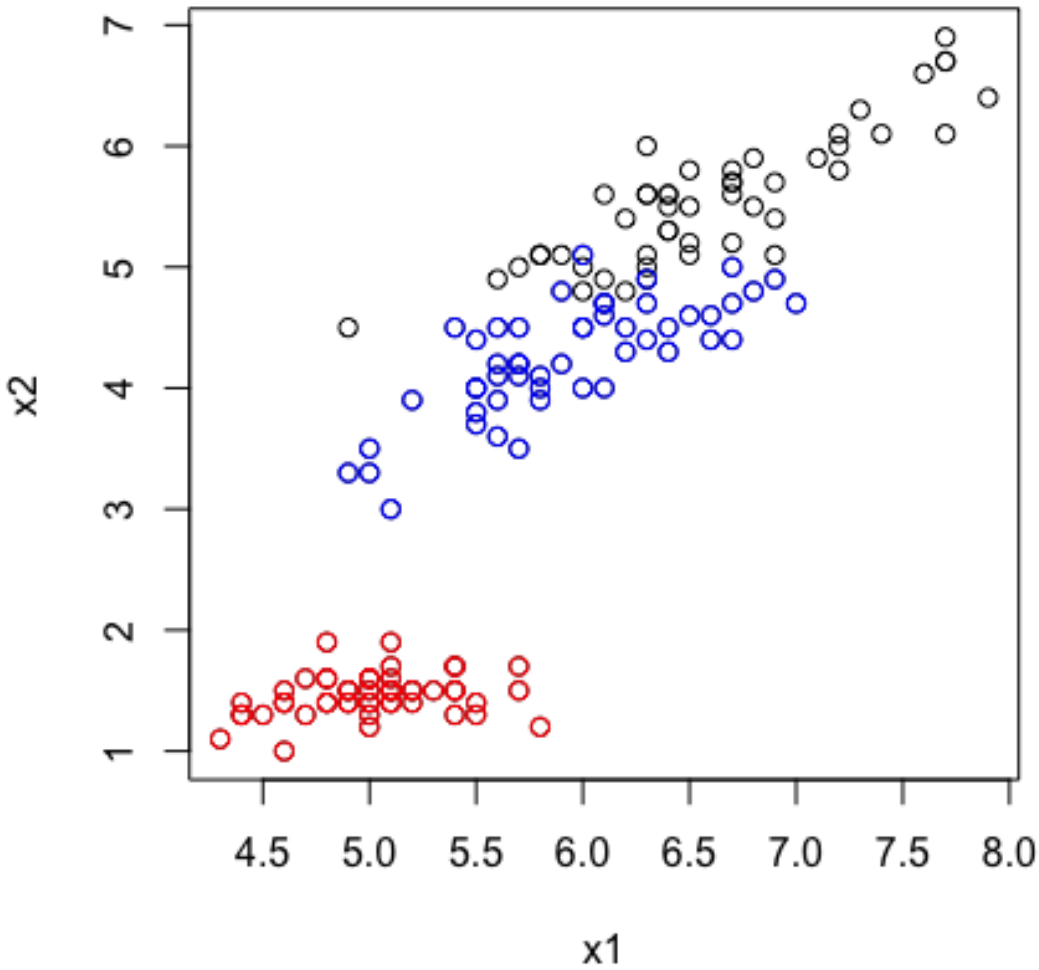


Visualisation

Or we could plot Sepal Length against Petal Length

```
x1 <- iris$Sepal.Length;
x2 <- iris$Petal.Length
plot(x1,x2)
inds <- which(iris$Species=='setosa')
points(x1[inds],x2[inds],col='red')
inds <- which(iris$Species=='versicolor')
points(x1[inds],x2[inds],col='blue')
```

Visualisation



Multidimensional Scaling

Most real-world data sets will have more than 2 features. For example, the full Iris data has 4 features, and the stylometry data has 71 features.

Only being able to plot two features at a time is awkward – ideally we want to plot all the features at once. For three features, we could create a 3 dimensional plot. But we cant go any higher than this, since humans cant visualise higher dimensional spaces.

One approach is to try and reduce the dimensionality of the data – to project it down into (e.g.) a 2 dimensinoal space

Multidimensional Scaling

Multi-dimensional scaling (MDS) is often used to represent high dimensional data in two dimensions. It is a type of **dimensionality reduction** method often used for visualisation.

Suppose we have n observations that live in a K-dimensional space (e.g. a $K = 71$ dimensional space for books in stylometry). Given two observations $\mathbf{p} = (p_1, \ldots, p_K)$ and $\mathbf{q} = (q_1, \ldots, q_K)$, let $d(\mathbf{p}, \mathbf{q})$ be a distance function that computes the distance between the points, for example the Euclidean distance:

$$d(\mathbf{p}, \mathbf{q}) = \sum_{k=1}^{K} \sqrt{(p_k - q_k)^2}$$

Multidimensional Scaling

For any two observations, we can hence compute the distance between them. We could then plot them in two dimensional space by simply arranging the observations in a way such that the distance between them (measured in 2 dimensions) is equal to the distance between them in K dimensions.

For example, suppose $d(\mathbf{p}, \mathbf{q}) = 5$. Then in our two dimensional plot, we would place $\mathbf{p}$ and $\mathbf{q}$ a distance of 5 apart

Multidimensional Scaling

This seems quite straightforwards. However in practice it is not possible to arrange the points in a way that fully preserves all the distances between points, i.e. some distances must be changed in order to fit the K-dimensional representation into two-dimensions.

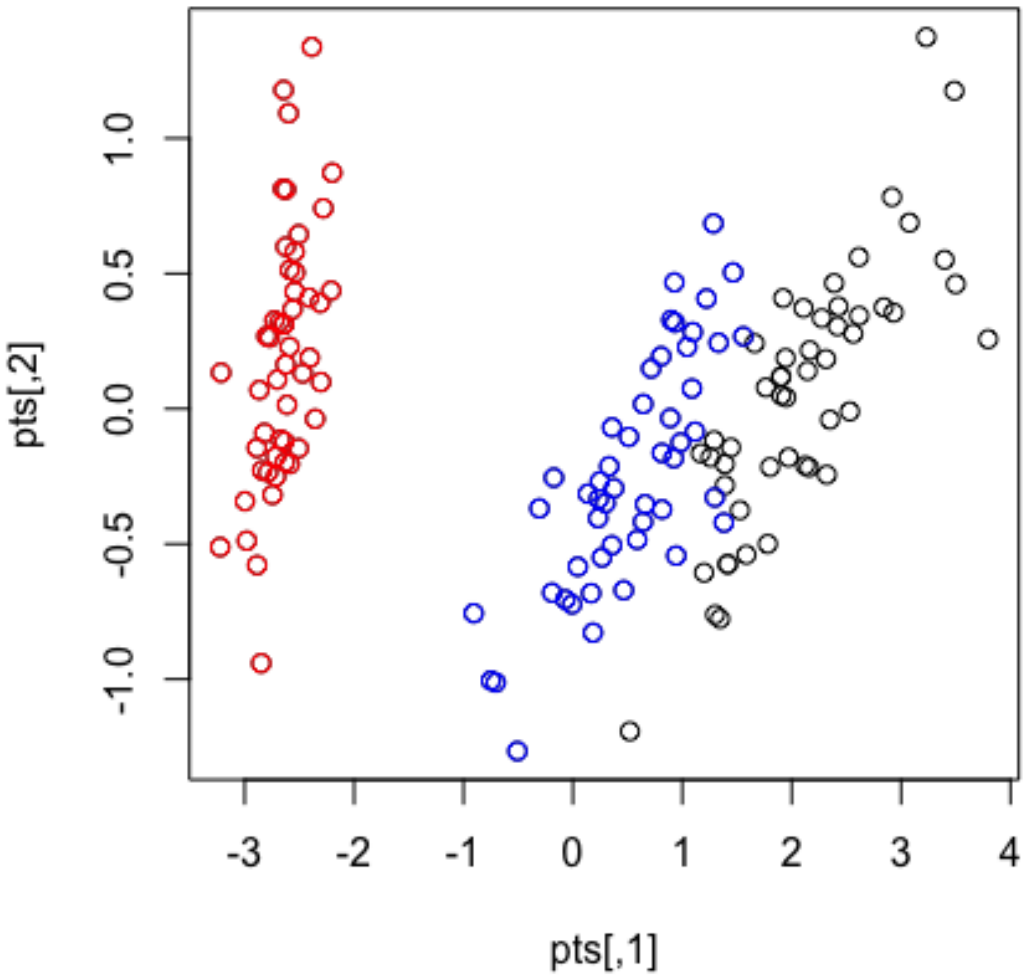
Multidimensional scaling uses an optimisation approach which tries to preserve the distances as best as possible. We will not go into the details, but it is quite similar to principal component analysis if you have encountered this before.

Instead, lets use R to perform multidimensional scaling. We use the cmdscale() function

Multidimensional Scaling

```
x <- iris[,1:4] #ignore class label
d <- dist(x) #create a distance matix
pts <- cmdscale(d) #perform MDS. Requires a distance matrix
plot(pts)
inds <- which(iris$Species=='setosa')
points(pts[inds,],col='red')
inds <- which(iris$Species=='versicolor')
points(pts[inds,],col='blue')
```

Visualisation



Multidimensional Scaling

Now let's apply this to stylometry. We could use MDS to visualise a corpus, to get an understanding of which books/authors are similar to each other. Lets try this on the corpus of classic authors and modern fantasy authors that we used previously.

As before, we load the corpus and create a single observation for each author by combining all their books together

```
M <- loadCorpus("~/Dropbox/Teaching/SCS/Data/
  Corpus/FunctionWords/", "frequentwords70")
x <- NULL
for (i in 1:length(M$features)) {
  x <- rbind(x, apply(M$features[[i]],2,sum))
}
```

Multidimensional Scaling

Next we standardise:

```
for (i in 1:nrow(x)) {
  x[i,] <- x[i,] / sum(x[i,])
}

for (j in 1:ncol(x)) {
  x[,j] <- (x[,j] - mean(x[,j]))/sd(x[,j])
}
```

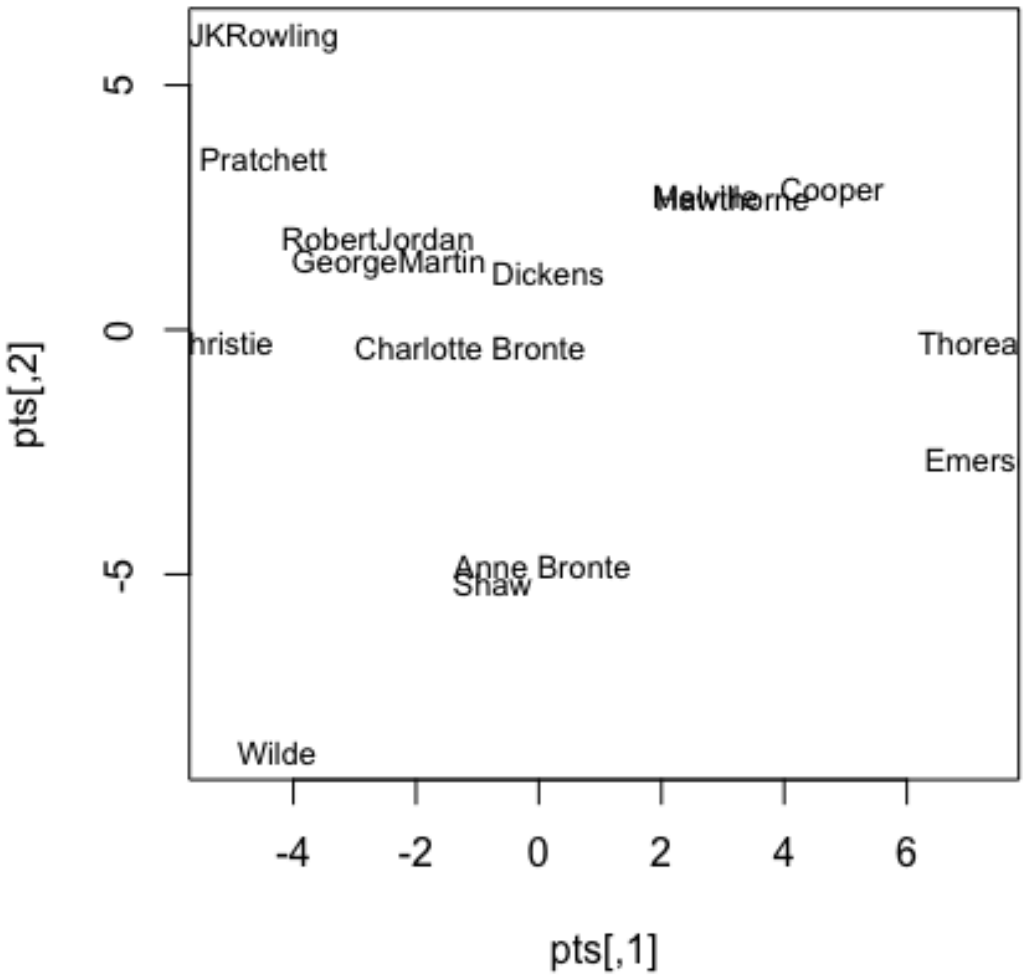
Multidimensional Scaling

Finally we create the MDS plot. For the visualisation, we use the text() function, which takes a vector of text strings (i.e. the author names) and the coordinates at which to plot them. The i^{th} text string gets plotted at the coordinates specified by the i^{th} row of the 'pts' matrix returned by the cmdscale() function.

```
plot(pts,type='n')
text(pts[,1],pts[,2],label=M$authornames,cex=0.8)
```

The 'cex' argument controls the font size used in plotting. Setting it to 0.8 shrinks the text to avoid making the plot too overcrowded.

Multidimensional Scaling



Multidimensional Scaling

Remember that the distance between the observations is based on how similar they are, so that more similar authors are closer together.

If we look at the top left of this MDS plot, we can see the modern fantasy authors such as J. K. Rowling, George R. Martin, Robert Jordan and Terry Pratchett. So these authors are more similar to each other than to the older (19th century) authors.

To the right, we see that Thoreau and Emerson are close. These are two American authors connected to the Transcendentalist movement, and had much in common

Multidimensional Scaling

We could also use MDS to study how the style of a single author changes over time.

Sir Terry Pratchett is a celebrated English fantasy novel who wrote the Discworld series of books, consisting of 41 novels written between 1983 and 2015.

During 2007, Pratchett was diagnosed with Alzheimer's disease, although he continued to write for several years after this diagnosis.

Question: did Alzheimer's cause a detectable change in his writing style?

Multidimensional Scaling

Pratchett is author 11 in the corpus. The corpus contains the 41 Discworld novels, and also 2 other non-Discworld books that he wrote later in life. We will look at his individual books rather than combining them together:

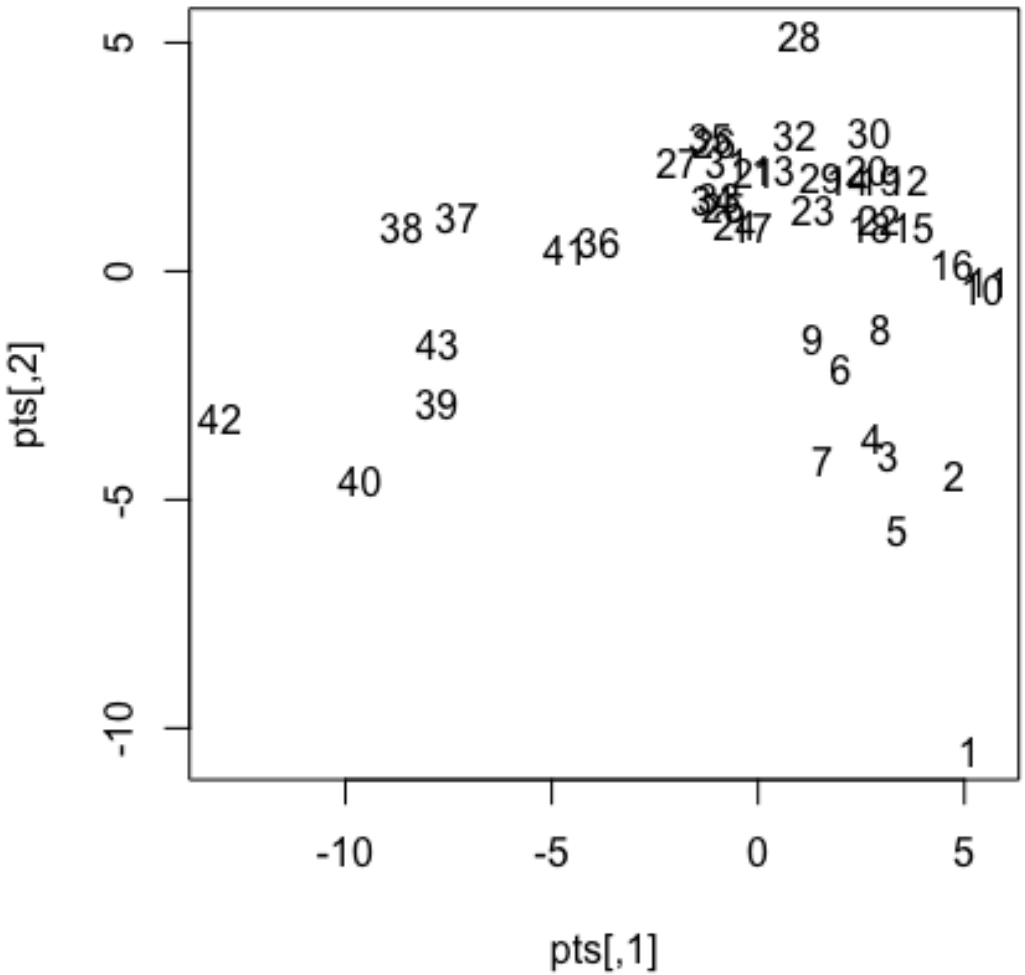
```
x <- M$features[[11]]
for (i in 1:nrow(x)) {
  x[i,] <- x[i,] / sum(x[i,])
}

for (j in 1:ncol(x)) {
  x[,j] <- (x[,j] - mean(x[,j]))/sd(x[,j])
}
```

Multidimensional Scaling

Now create the MDS plot. For the text labels on the plot, we will use the number of the book rather than their names, otherwise there will be too much text on the plot. These books are numbered in the order they were written, so book 1 is his first book, and book 43 is his last.

```
d <- dist(x)
pts <- cmdscale(d)
plot(pts,type='n')
text(pts[,1],pts[,2],label=1:nrow(x))
```

Multidimensional Scaling



Multidimensional Scaling

First note that in general, books that were written at around the same time (i.e. which have similar numbers) tend to be close together. This suggests a gradual evolution in writing style.

If we look at the bottom right, we see the early Discworld novels (1-9) which seem to have a slightly different style to Pratchett's mature work.

Towards the right side of the plot, we see that books 36 and later seem to be different to the earlier ones

Multidimensional Scaling

We can see the names of the books in the M$booknames field of the list. If we check, we find that Book 36 is called “Making Money”.

This book was published in September 2007. Pratchett’s Alzheimer’s diagnosis was made public in 2007. So it does seem like there is potentially a connection.

Of course, this is just a preliminary analysis based entirely on visualisations. A fuller analysis would require the use of statistical methods and models. But it is interesting that a simple visualisation can be so revealing!

Multidimensional Scaling

We can carry out a similar visualisation for JK Rowling. Unlike Pratchett, she did not suffer any major illnesses that would affect her writing style. But perhaps her style gradually changed when she started writing adult novels.

We first load and standardise the Rowling corpus (important note: I updated this corpus to add in her more recent novels. If you have downloaded it before, then please download it again!).

Multidimensional Scaling

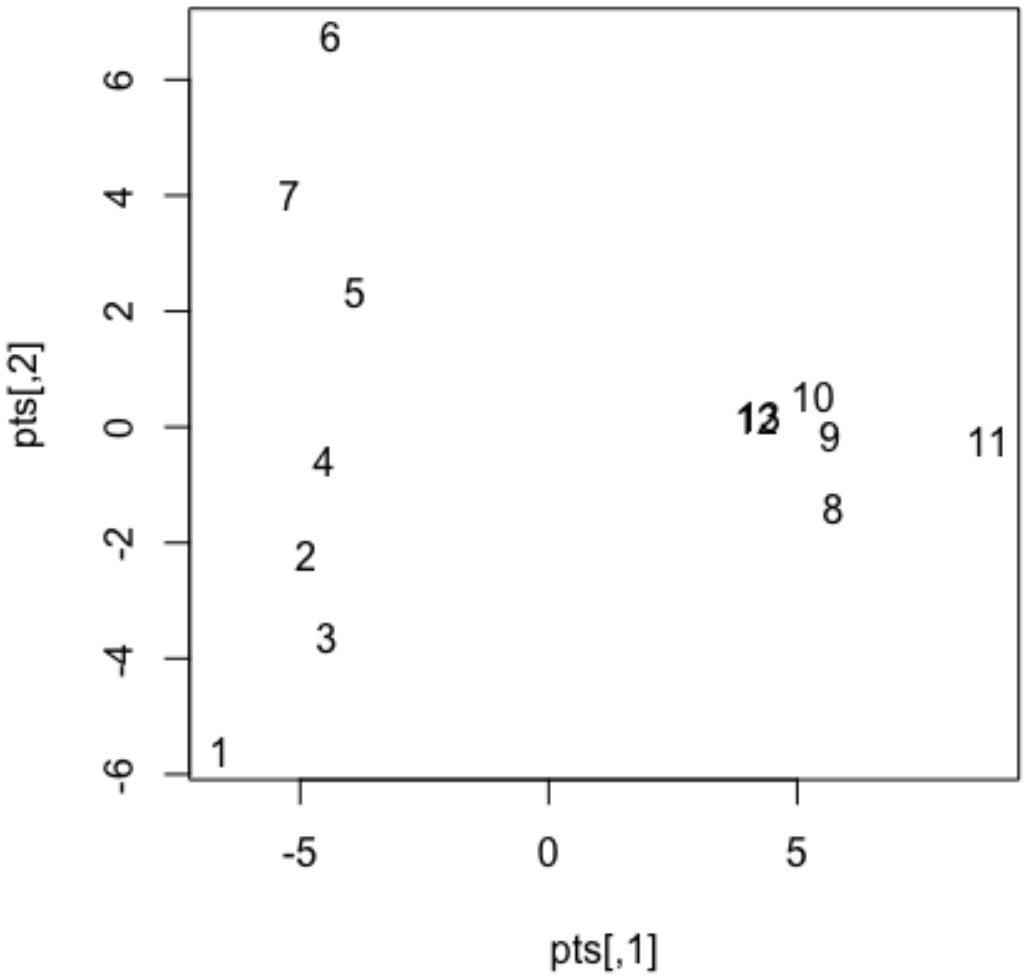
```
M <- loadCorpus("~/Dropbox/Teaching/SCS/Data/JKRowling/
  FunctionWords/", "frequentwords70")

x <- M$features[[1]]
for (i in 1:nrow(x)) {
  x[i,] <- x[i,] / sum(x[i,])
}

for (j in 1:ncol(x)) {
  x[,j] <- (x[,j] - mean(x[,j]))/sd(x[,j])
}
```

Multidimensional Scaling

Next we perform MDS as before:

```
d <- dist(x)
pts <- cmdscale(d)
plot(pts,type='n')
text(pts[,1],pts[,2],label=1:nrow(x))
```

Multidimensional Scaling



Multidimensional Scaling

Books 1-7 are the Harry Potter novels, while Books 8-13 are her crime novels. As before, books that were written during a similar time period (i.e. with numbers close together) tend to be more similar to each other.

However we can observe a clear switch in her writing style as she transitioned to writing crime novels. They seem quite distinct to the Harry Potter novels.

Remember that this analysis has only used function word counts, and no additional information! We are not making the (probably trivial?) point that her crime books may use longer words, more complicated grammar etc since they are aimed at adults. Instead, we are saying that there has been a structural change in how she used basic words such as ‘a’ and ‘of’. This is surely surprising!

Multidimensional Scaling

Despite this change, remember that the supervised learning methods were capable of correctly identifying that Rowling was the author of the Cuckoo's Calling and The Causal Vacancy.

So even though her style did change, these later books are clearly still more similar to the Harry Potter novels than they are to books written by different authors such as Pratchett or George R. R. Martin