

R-books: Successor to e-books

Paul Hewson

`paul.hewson@plymouth.ac.uk`

`www.plymouth.ac.uk/staff/phewson`

9th August 2007



- 1 Background
- 2 Rpad
- 3 Live Demo
- 4 Research lead teaching
- 5 Conclusions

From: r-help (25th April 2006)

Dear R People:

Are your undergraduate students receptive to learning R, as a rule?

Most of the time, mine really like it. But this semester, they act as though they are being eaten by rats when learning R. They are not trying at all. Any similar experiences? If anyone has any good ideas, I would be THRILLED to hear them, as I am using R in Summer School.

Thanks,

Sincerely,

Erin Hodgess

Associate Professor

Department of Computer and Mathematical Sciences

University of Houston - Downtown



From: r-help (25th April 2006)

Dear Erin,

I wrote the Rcmdr package because my undergrad intro stats students are much more comfortable with point-and-click interfaces. You're in a computer and math department, however, while I'm in sociology -- I would have thought that your students wouldn't have trouble with command-driven software.

Regards, John

John Fox

Department of Sociology

McMaster University



Considerable efforts to produce different user interfaces

- ESS
- The rest

Friendly interfaces

Include things like:

- Various script editors (WinEdt, Tinn-R, other)
- Tcl/Tk based interfaces
- Web-based interfaces, e.g. with Apache
- Java based interfaces
- DCOM interface



Try turning 180°

- Guided learning environments: UoP uses SharePoint and QM Perception; PMS uses Blackboard
- Numerous “widgets”, e.g. java applets demonstrate various aspects of statistics, cast.massey.ac.nz is particularly comprehensive
- Numerous textbooks available in e-Book format, such as www.xplore-stat.de/ebooks/ebooks.html from Springer containing illustrative “Quantlets”

Leveraging R

- I like R (so do many colleagues).
- We make our Maths and Stats students like R (Stockholm syndrome?)
- But we see a lot of non-M&S students for a short while and (rightly or wrongly) have never considered teaching them to use R
- But we think we can use R for some cute learning aids

Rpad

Rpad by Tom Short uses Tcl/Tk (available from within R) to provide a local webserver.

- Running locally sorts out security issues (I think)
- You actually run R on your own PC, not someone else's webserver
- Rpad uses DoJo JavaScripting library - quite clever
- Using something as powerful as R underneath means *any* statistical routine can be incorporated. So we can do all the basics, but have a carefully guided peek at the “state of the art” research interface



Using Rpad

```
<head>  
...  
<script type='textjavascript'  
src='guidojo.js'><script>  
<script type='textjavascript'  
src='guiRpad.js'><script>  
...  
</head>
```

Interfacing with R

```
<pre dojoType='Rpad' rpadRun = 'init'  
rpadHideSource='true' rpadOutput = 'html'>  
distmethod &lt;- c('euclidean', 'maximum',  
'manhattan', 'canberra', 'minkowski',  
'Gower')  
HTMLon()  
HTMLselect('DistMethod', distmethod)  
HTMLoff()  
</pre>
```

Sorry - I chickened out

- I've had a few glitches at home - don't know whether it was (tabbed) browser updates or something silly I've done. It seems happy without DoJo tabs but I wasn't sure what might happen here.
- There are two small glitches - the R console regains focus when the first graph is produced (it only needs to be minimised again) and there are some difficulties rendering the math (I used `ttH` to create base html from \LaTeX files)

A screenshot

R Books: An Introduction to Cluster Analysis with R

File Edit View Bookmarks Sidebar Tabs Sign In Tools Help

Home BT Yahoo! Mail BT.com Yahoo! Protection Photos News Music Movies Games Sport Shopping

http://127.0.0.1:8079/Rbook.Rpad

R Books: An Introducto... Multivariate Statistics with V

R Books: An Introduction to Cluster Analysis

Background

There is an online reading list associated with this topic: [Online Reading List for STAT3401](#). Cluster analysis is an extremely widely used set of techniques, you cannot fail to find some useful supplementary reading. In particular, it may be worth checking out these [Hierarchical Clustering java applets](#) which illustrate how three of the main hierarchical clustering algorithms work.

The first step in our exploration of cluster analysis is to select some data. The drop down menu will give you a choice of data objects already within R, as well as letting you load csv files currently residing in your workspace. You are cautioned that this is a very simplistic input routine, any files must be in an acceptable format. If anything goes wrong, it is quite likely to be because the data format isn't acceptable, full instructions are provided [HERE](#)

Please select a data set for further exploration:

Heptathalon.csv

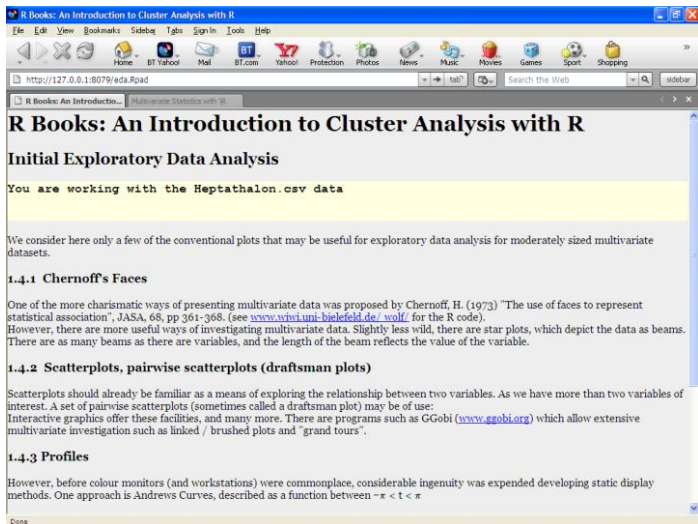
Mammalian Milk	X100mHurdles.s.	HighJump.m.	ShotPut.m.	X200m.sec.
Tinned Food	Min. :13.23	Min. :1.600	Min. :11.14	Min. :23.53
econ.csv	1st Qu.:13.52	1st Qu.:1.720	1st Qu.:12.81	1st Qu.:24.40
Heptathalon.csv	Median :13.73	Median :1.750	Median :13.54	Median :24.77
hotelling.csv	Mean :13.90	Mean :1.743	Mean :13.30	Mean :24.91
IND : 2	3rd Qu.:14.19	3rd Qu.:1.780	3rd Qu.:14.14	3rd Qu.:25.25
KZK : 2	Max. :15.12	Max. :1.840	Max. :15.55	Max. :27.27
UKR : 2				
(Other):13				
LongJump.m.	Javelin.m.	r800m.s.		
Min. :5.220	Min. :32.53	Min. :129.1		
1st Qu.:5.850	1st Qu.:41.48	1st Qu.:133.9		
Median :5.950	Median :43.92	Median :137.7		
Mean :5.978	Mean :43.47	Mean :137.2		

Done

Comments

- Unlike DIY java applets - loads of useful datasets already in R (or libraries such as `FlexClust`, `fpc` etc.)
- Simple R script also looks for *any* csv file in the R working directory

Another screenshot



R Books: An Introduction to Cluster Analysis with R

File Edit View Bookmarks Sidebar Tabs Sign In Tools Help

Home BT Yahoo! Mail BT.com Yahoo! Protection Photos News Music Movies Games Sport Shopping

http://127.0.0.1:8079/eda.Road

R Books: An Introducto... Multivariate Statistics with R

R Books: An Introduction to Cluster Analysis with R

Initial Exploratory Data Analysis

You are working with the `Heptathlon.csv` data

We consider here only a few of the conventional plots that may be useful for exploratory data analysis for moderately sized multivariate datasets.

1.4.1 Chernoff's Faces

One of the more charismatic ways of presenting multivariate data was proposed by Chernoff, H. (1973) "The use of faces to represent statistical association", *JASA*, 68, pp 361-368. (see www.wjvi.uni-bielefeld.de/~wolf/ for the R code). However, there are more useful ways of investigating multivariate data. Slightly less wild, there are star plots, which depict the data as beams. There are as many beams as there are variables, and the length of the beam reflects the value of the variable.

1.4.2 Scatterplots, pairwise scatterplots (draftsman plots)

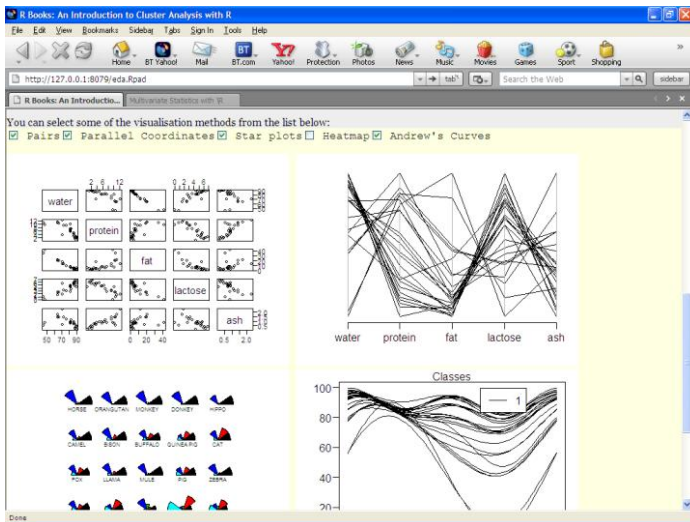
Scatterplots should already be familiar as a means of exploring the relationship between two variables. As we have more than two variables of interest. A set of pairwise scatterplots (sometimes called a draftsman plot) may be of use: Interactive graphics offer these facilities, and many more. There are programs such as GGobi (www.ggobi.org) which allow extensive multivariate investigation such as linked / brushed plots and "grand tours".

1.4.3 Profiles

However, before colour monitors (and workstations) were commonplace, considerable ingenuity was expended developing static display methods. One approach is Andrews Curves, described as a function between $-\pi < t < \pi$

Done

And here's one I made earlier



Gosh, isn't this exciting, another static screenshot

R Books: An Introduction to Cluster Analysis with R

http://127.0.0.1:8079/DevMat.Rpad

R Books: An Introductio... Multivariate Statistics with V...

interpreted as the physical distance between two p -dimensional points is also a convenient measure to understand. Formally, we can express this measure as:

$$d_{ij} = (\sum_{k=1}^p (x_{ik} - x_{jk})^2)^{1/2}$$

where we are trying to measure the distance between observations in row i and row j , in other words x_{ik} is the k th observation in row i , and x_{jk} is the corresponding k th observation in row j . Euclidean distance can be readily calculated in **R** using the `dist()` function with the default `method = "euclidean"`, as well as by `daisy()` with the default `metric = "euclidean"`, although in `daisy()` it is possible to standardise the data within the calculations by adding `stand = TRUE` to the function call.

City Block metric

The City Block metric, formally referred to as an l_1 norm, measures the absolute difference between two vectors. It is so-named because it measures the distance between two points in terms of movements parallel to the axis and therefore resembles the distance between two points in a city. [Krause, \[1975\]](#) (who had obviously never been in a London taxi) called this distance the *taxicab* distance, [Brandeau and Chiu, \[1988\]](#) used the term *rectilinear*, but perhaps the most common alternative name is *Manhattan*, suggested by [Larson and Sadiq, \[1983\]](#) reflecting the famous city block layout in Manhattan. Formally, we can express this distance as:

$$d_{ij} = (\sum_{k=1}^p |x_{ik} - x_{jk}|)$$

It can be calculated in **R** using the `dist()` function with `method = "manhattan"`

Minkowski metric

The Minkowski metric, or the l_p norm, is a generalisation of the Manhattan and Euclidean distances.

$$d_{ij} = (\sum_{k=1}^p |x_{ik} - x_{jk}|^\lambda)^{1/\lambda}$$

Where $\lambda = 1$ we have the Manhattan metric, where $\lambda = 2$ we have the Euclidean distance. It can be noted that increasing λ exaggerates dissimilar units relative to similar ones. This metric can be calculated in **R** using the `dist()` function with `method = "minkowski"` but additionally requires an argument to `p` to set λ , the power of this distance. Therefore, for example `dist(x, method = "minkowski", p=2)` gives the Euclidean distance for matrix x .

Done

Examining distances and clustering methods

R Books: An Introduction to Cluster Analysis with R

File Edit View Bookmarks Sidebar Tabs Sign In Tools Help

Home BT Yahoo! Mail BT.com Yahoo! Protection Photos News Music Movies Games Sport Shopping

http://127.0.0.1:8079/hclust.Rpad

You are working with the Mammalian Milk data

This page lets you try out different distance measures and different hierarchical clustering methods. Do note that we aren't making very good use of Gower's distance here (at the moment the webpage just guesses variable types) and so this should only be used with caution.

Select a distance method:

euclidean

Select an hierarchical clustering method

complete

This page uses the [cophenetic correlation](#) as a measure of the distortion in the dendrogram. Recall that a value of 1 implies no distortion (i.e. the dendrogram represents the distance matrix perfectly), a value of 0 is utterly distorted.

Cophenetic correlation is: 0.909504823725888

Cluster Dendrogram

Done

Distances and Clustering Algorithms

- Actually, it's quite tedious to explore the interplay between these two even for a fairly modest set of combinations (we tend to only consider four or five hierarchical clustering algorithms)



R graphics are superb . . .

R Books: An Introduction to Cluster Analysis with R

File Edit View Bookmarks Sidebar Tabs Sign In Tools Help

Home BT Yahoo! Mail BT.com Yahoo! Protection Photos News Music Movies Games Sport Shopping

http://127.0.0.1:8079/posteda.Rpad

Post Clustering Data Visualisation

You are working with the Mammalian Milk data

You need to decide how many groups you think exist in your data:

3

Please select any of the following plots that you think may be useful

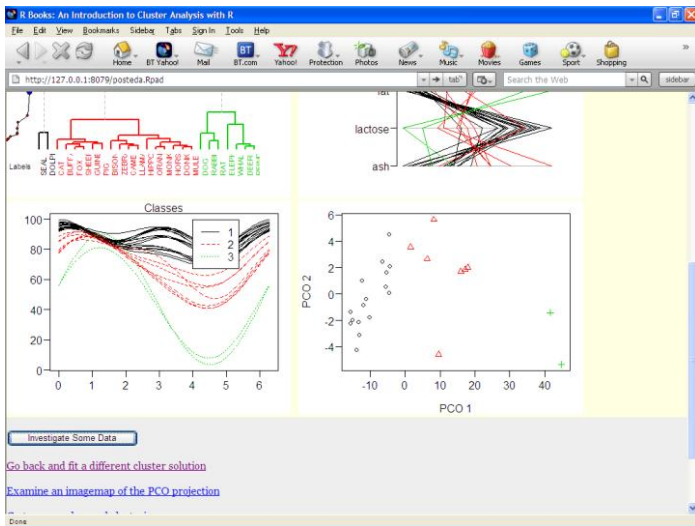
Pairs Parallel Coordinates Principal Component Projection
 Multidimensional Scaling Projection Andrew's Curves

Colored Dendrogram (3 groups)

water
protein
fat
lactose
ash

Done

Imagine having to write all this as custom java applets



Imagemaps allow a little interactivity

The screenshot shows a web browser window with the address bar displaying `http://127.0.0.1:8079/imagemappc.Rpad`. The page content includes a paragraph of text explaining scaling techniques, a dropdown menu set to '3', and a button labeled 'Plot the first two dimensions'. Below this is a plot titled 'RABBIT' showing PCO 1 on the x-axis and PCO 2 on the y-axis. The plot contains several data points: black circles, red triangles, and green crosses. A white arrow points to a red triangle at approximately (8, 5.5).

Scaling (or the particular type of metric scaling considered here also known as Principal Co-ordinates Analysis) is a useful projection technique in its own right which may reveal structure in the data. Partly for cosmetic reasons, you may wish to select a number of groups in the data before it is plotted.

3

Plot the first two dimensions

RABBIT

PCO 2

PCO 1

Done

The trouble with cluster analysis

- I wish I had the courage to follow Flury (1997) and never mention cluster analysis
- Learners brought up with computers are very trusting of their output
- Rather difficult to convey some of the limitations of things like cluster analysis
- Although it's beyond our syllabus, working with hybrid clustering really gets some discussion going on finding the "correct" solution



Hybrid clustering

R Books: An Introduction to Cluster Analysis with R

File Edit View Bookmarks Sidebar Tabs Sign In Tools Help

Home BT Yahoo! Mail BT.com Yahoo! Protection Photos News Music Movies Games Sport Shopping

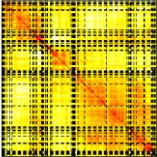
http://127.0.0.1:8079/genes.Rpad

The distance measures used vary according to whether we are clustering by gene or by array. Available options include the [cosine angle](#), also known as the uncentered correlation distance), or the [absolute cosine angle](#), the [correlation distance](#) and the [absolute correlation distance](#). Where appropriate, the euclidean distance is also available.

cosine angle

Searching for main clusters... Level 1 Level 2 Level 3 Level 4 Level 5 Level 6 Level 7 Level 8 Level 9 Level 10 Level 11 Level 12 Level 13 Level 14 Level 15 Identified 46 main clusters in level 4 with MSS = 0.04503999 Running down without collapsing from Level 4 Level 5 Level 6 Level 7 Level 8 Level 9 Level 10 Level 11 Level 12 <

Golub AMI/All Data (1999): Gene Distance Matrix



This is quite a slow process!! There should be some kind of output from the fitting algorithm, and the cluster map should appear some time before the bootstrap output.

Fit hybrid cluster

Done

Where it's been useful

- Cluster analysis
- Environmental science (three week session)
- Postgraduate medical training - very nice animations to help with hypothesis testing, the concept of power etc.
- Externally: Road Safety practitioners (illustrating regression to the mean)
- Leaping from basics to state-of-the-art and back again:
 - 1 A good way of asking questions about the basics;
 - 2 We show non-statistics specialists that this is a live and active subject.

Where (IMHO) it's *NOT* been useful

- Multivariate statistics (I do like to be able to use it as a linear algebra calculator)
- You do lose some key stuff, such as Ggobi
- Delaying the journey up the learning curve: too much good stuff to delay statisticians / scientists from getting to grips with it in a manner facilitating reproducible research

R books

- Yes, I know. It's an obvious use of R. Arguably we've taken it to an extreme and "hidden" it from the user entirely, but interactive statistics text books aren't new . . .
- . . .but you really can throw in some functionality with R
- R-books have been well received in some niche areas

R books

- Some cosmetic / navigation issues to sort out (easy, but it takes a little time), and there are a couple of weirdnesses, such as a focus shift when the first graph is produced (hard) and the math rendering (use latex2html?).
- Despite the glitches, it has been well received
- It might have made for a tighter interface if (one of) the java interfaces had been used. Possible one with pdf and java???
- Initially, we wanted to interface this with a corporate “Guided Learning environment”. That’s not quite there, currently, some simple Q&A stuff is also done with Rpad functions as well.
- Sorry about the oxymoron (mammalian milk?).



Thanks

- Thanks for listening and <HINT> not asking very awkward questions </HINT>
- Thanks to those concerned for R
- Thanks to Tom Short for Rpad
- Thanks for all the great contributed goodies
- Thanks to my students for keeping life interesting

