STAT3401: Lab exercises concerning measures of distance

Paul Hewson

11th January 2007

1 Mahalanobis distance and multivariate normality

```
> USA.mah.dist <- mahalanobis(USArrests, center = mean(USArrests),
+     cov = var(USArrests))
> hist(USA.mah.dist, freq = FALSE)
> curve(dchisq(x, df = 4), add = TRUE, col = "red")
```

or consider qq plots:

```
> n <- 50
> p <- 4
> qqplot(USA.mah.dist, qchisq(ppoints(n), p))
```

- Do you think that the USArrests data can be considered to be multivariate normal. If so, repeat this with some other data, such as the simulated multivariate normal data you generated in week 1, or any other data we have met (e.g. iris data).
- For any data you consider, compare your findings on multivariate normality with what you find when examining univariate normality of either the margins or the linear combinations

More advanced work: there are some who think that a beta distribution should be used. There is a function `qqbetaM` in a file in the portal. Copy this into your workspace, `source` the function and see whether you think it makes a difference.

```
> source("qqbetaM.R")
> qqbetaM(USA.mah.dist, 4)
```

Compare the qqplots you obtain from this function with qqplots from the χ^2 based quantiles. Do you have any data where your impression of normality may be altered by a change in the comparison? From your reading material, when might it matter whether you use a beta distribution or a χ^2?

2 Gower's distance

This is quite an important exercise! Make sure you are happy calculating Gower's distance. Use the `class07.csv` data in the portal, and calculate the distance between a few individuals by hand. Then use `daisy` in `library(cluster)` to calcuate it in R (it's a slightly fiddly function to use).