Software for Cluster Size Estimation

Edward Susko
Department of Mathematics and Statistics, Dalhousie University

Installation

The sofware provides R functions for calculating three indeces for the number of clusters. A concise reference is Tibshirani, Walther and Hastie (2001). The three indeces are the CH index, the KL index and the gap statistic. To unpack the software, type

```
$ gzip -d clust_size.tar.gz
$ tar xvf clust_size.tar
```

This will create a directory `clust`. To create the executables type

```
$ cd clust
$ make
```

These programs will usually be called from an R function that assumes their names are `gap_stat` and `clust_size`. To make the R functions available, with a running R session, issue the command

```
> source("clust_size_fn.q")
```

Examples of use are contained in the file `example.q`. These can be run at the R command prompt with

```
> source("example.q")
```

Alternatively, lines within this file can be cut and pasted to the R command prompt.

R functions

The main R functions are `chkl.idx.k` and `gapc`. The function `chkl.idx.k` computes the CH and KL indeces for a sequence of cluster sizes and returns the cluster size that is deemed optimal according to these criterion. The function `gapc` computes the gap statistic for a sequence of cluster sizes and returns the cluster size that is deemed optimal

according to this criterion. An additional function `check.sim` is provided that allows you to simulate data from 4 normal clusters, which can be useful in testing things out.

Gap statistic calculation

`gapc(x, maxclust = 30, B = 100)` **description**: The gap statistic indece for the data in the x matrix

arguments:

`x`: data matrix. Each row gives a multivariate observation.

`maxclust`: The maximum number of clusters to calculate the indeces for.

value:

`idx`: `maxclust` $\times$ 4 matrix. The first column is the $log(SSW)$ for each of the cluster sizes, where SSW denotes the sum of squares within clusters; $log(W_k)$ in the notation of Tibshirani et al (2001). The second column is bootstrap estimate of the mean $log(SSW)$; $(1/B)sum_b log(W_{kb}*)$ in the notation of Tibshirani et al (2001). The third column gives the difference: $Gap(k) = (1/B)sum_b log(W_{kb}*) - log(W_k)$. The final column gives the standard error, s_k, for the bootstrap estimate of the mean $log(SSW)$.

`k`: the estimated cluster size based on the criterion in Tibshirani et al (2001): $k =$ smallest k such that $Gap(k) >= Gap(k+1) - s_k + 1$.

examples:

```
nclust <- 10 # maximum number of clusters of interest
g <- gapc(x, maxclust = nclust+1)
kg <- c(1:nclust)
# plot the gap criterion: Gap(k) - Gap(k+1) + s_k+1
# the smallest k for which this is >= 0 is the estimated k.
cat("The number of clusters estimated by the gap statistic is", g$k, "\n")
plot.wlines(kg, g$idx[1:10,3] - g$idx[2:11,3] - g$idx[2:11,4],
         xlab = "Number of Clusters", ylab = "Gap criterion")
abline(h = 0) # add the y = 0 line
```

CH and KL indice calculation

`chkl.idx.k(x, maxclust = 30)` **description**: The CH and KL indeces for the data in the x matrix

arguments:

`x`: data matrix. Each row gives a multivariate observation. maxclust - The maximum number of clusters to calculate the indeces for.

value:

`idx`: (`maxclust`1)- $\times$ 2 matrix. The first column gives the CL index and the second the KL index. indeces are for numbers of clusters ranging between 2 and `maxclust`.

`k`: 2 dimensional. The first entry is the CH estimate and the second entry is the KL estimate of cluster size.

examples:

```
x <- check.sim(100, 2, 5)
nclust <- 10 # maximum number of clusters of interest
kch <- c(2:nclust)
ch <- chkl.idx.k(x, maxclust = nclust)
cat("The number of clusters estimated by CH is", ch$k[1],
    "and the number of clusters estimated by KL is",  ch$k[2],
    "\n")
plot.wlines(kch, ch$idx[,1], xlab = "Number of Clusters", ylab = "CH index")
plot.wlines(kch, ch$idx[,2], xlab = "Number of Clusters", ylab = "KL index")
```

Generating clustered data

`check.sim(B, p, sd)` **description**: simulates data from 4 normal clusters

arguments:

`B`: A multiplier for the number of data points in `x`. Each of the `B` generations gives between 100 and 200 observations.

`p`: the dimension of the `x` matrix

`sd`: the standard deviation used in

value:

`x`: `B` $\times$ `p` matrix. Each row gives data from one of the 4 clusters

details: Four p-dimensional mean vectors are generated. These are generated so that, coordinate-wise the distance between the means is at least one. If too many simulations are required to obtain such mean vectors the routine stops with the warning "`nsim > 1000`". For each cluster, the number of data points generated from it is 25 or 50 with probability 1/2 of either sample size.

Warnings and Additional Comments

The functions `gapc` and `chkl.idx.k` call the programs `gap_stat` and `clust_size`. These programs are assumed to be in the working directory R was started in. If you would like to store these in another location, say `/home/myname/clustfns/`, so that you can call functions from a variety of different working directories, you will need to change the variable `clustdir` at the top of the file `clust_size_fn.q`:

```
clustdir <- "/home/myname/clustfns"
```

The program files `gapc` and `chkl.idx.k` create files in the working directory of R when they are called. The files `x.tmp` and `cl.tmp` are created by both routines. The file `gap_stat.out` is created by `gapc` and the file `clust_size.out` is created by `chkl.idx.k`.

All of the routines use hierarchical clustering to determine clusters. They do this with the "average" method. If you want to change the way in which hierarchical clustering is done, you should search for and make desired changes to the `hclust` lines in the file `clust_size_fn.q`.

The KL index is included but in my experience has not tended to give good estimation.

References

Tibshirani, R., Walther, G. and Hastie, T. (2000). Estimating the number of clusters in a data set via the gap statistic. *J. R. Statist. Soc. B.* **63**:411–423.