**FORTRAN Programs and Data Files for:**

**S. Brudvig, M. J. Brusco, J. D. Cradit (2019). Joint selection of variables and clusters: recovering the underlying structure of marketing data. *Journal of Marketing Analytics*, 7 (1), 1-12.**

**Fortran programs:**

I. All-possible-subsets

aps.for – the Fortran source code for all possible subsets k-means clustering

aps.exe – the Fortran executable file

A. Input files

clusdat – first line contains three items: number of objects, number of variables, and number of clusters (the third argument is ignored by the program). The remaining lines contains the measurements for each object on each variable (row-by-row).

B. Output files

results – a summary output file containing, for each number of clusters on the interval 2 ≤ K ≤ 9, the best subset for each of the subset sizes on the interval 2 to V variables (along with the corresponding pseudo-R2 value. The computation time is also reported.

C. How to Run

After downloading the JMA\_PROGS folder to the hard drive (e.g., c:\jma\_progs), go to the hard command prompt, change the directory to c:\jma\_progs, and type aps at the command prompt:

>c:\jma\_progs> aps

For the current clusdat file (V = 13 variables), the program will take about 20 minutes to run on a standard microcomputer. A problem with fewer variables will take less time, and a problem with more than 15 variables could take a very long time. A user could also narrow the range of K in the source code to shorten the computation time, but this will require a Fortran compiler.

II. HINoV

hinov.for – the Fortran source code for selecting variables via HINoV

hinov.exe – the Fortran executable file

A. Input files

clusdat – first line contains three items: number of objects, number of variables, and number of clusters. The remaining lines contains the measurements for each object on each variable (row-by-row).

B. Output files

solclus – a summary output file containing, for the selected number of clusters, the V variables in their rank order of TOPRI values. The computation time is also reported.

C. How to Run

After downloading the JMA\_PROGS folder to the hard drive (e.g., c:\jma\_progs), go to the hard command prompt, change the directory to c:\jma\_progs, and type hinov at the command prompt:

>c:\jma\_progs> hinov

For the current clusdat file (V = 13 variables), the program will take just a few seconds.

III. K-means (100,000) restarts

kmeans.for – the Fortran source code 100,000 random restarts of k-means clustering

kmeans.exe – the Fortran executable file

A. Input files

cdat – first line contains three items: number of objects, number of variables, and number of clusters. The remaining lines contains the measurements for each object on each variable (row-by-row).

B. Output files

cents – a summary output file containing, for each cluster: the cluster size and cluster centroid.

teste – the cluster assignments for the best found partition

C. How to Run

After downloading the JMA\_PROGS folder to the hard drive (e.g., c:\jma\_progs), go to the hard command prompt, change the directory to c:\jma\_progs. A file cdat must be created. There are two options:

1. Copy clusdat to cdat to run on full data set.

2. Run the Fortran program extract at the command prompt, which will produce the file cdat from clusdat but using only the variables that are in the file vars.

extract.for – the Fortran source code to select only a subset of variables from the dataset

extract.exe – the Fortran executable file

The input files for extract are clusdat and a file named vars, which contains how many variables to select from clusdat and the indexes of those variables. The output file is cdat. To run extract, type:

>c:\jma\_progs> extract

After selecting either options 1 or 2, type the following at the command prompt to run 100,000 restarts of K-means clustering.

>c:\jma\_progs> kmeans

For the current clusdat file (V = 13 variables), the program should take no more than a couple of minutes on a standard microcomputer (although computations time is affected by the choice for the number of clusters).

D. Postprocessing the teste file with pseuodf

pseudof.for – the Fortran source code to compute pseudo-F and pseudo-R2 for a partition

pseudof.exe – the Fortran executable file

This program reads cdat and teste and computes the Calinski-Harabasz pseudo-F and the pseudo-R2 for the partition in teste and writes these values to the screen. The program, which runs instantaneously, is implemented as follows:

>c:\jma\_progs> pseuodf