ScottKnott: A Package for Performing the Scott-Knott Clustering Algorithm in R

Enio G. Jelihovschi and José Cláudio Faria

Abstract Scott-Knott is an exploratory clustering algorithm used in the ANOVA context, when the researcher is comparing treatment means, with a very important characteristic: it does not present any overlapping in its grouping results. We wrote a code, in R, that performs this algorithm starting from `vectors`, `data.frame`, `aov` or `aov.list` objects. The results are presented with letters representing groups as well as in graphical way using different colors to differentiate among the distinct groups.

Introduction

The Scott-Knott (SK) algorithm is a hierarchical clustering algorithm used as an exploratory data analysis tool. It was designed to help researchers working with an ANOVA experiment designed to compare treatment means, to find distinct homogeneous groups of those means whenever the situation leads to a significant *F-test.*

All multiple comparison procedures which are used to solve that problem usually divide the set of treatment means in groups which are not completely distinct, many treatments end up belonging to different groups simultaneously, this is called overlapping (Calinski and Corsten, 1985).

In fact, as the number of treatments increases, so do the number of overlapping making it difficult for the experimental users to distinguish the real groups to which the treatments should belong. The division of the treatments in completely distinct groups is the most important solution in this case for them. Even though the goal of multiple comparison methods is an all-pair comparison, not a division of the treatment means into groups, the biologists, plant breeders and many others expect those tests to do that for them.

The possibility of using cluster analysis in place of multiple comparison procedures was suggested by O'Neill and Wetherill (1971) since the results of cluster analysis type of solution would divide the treatments into distinct groups.

The SK algorithm is a hierarchical cluster analysis approach used to partition the treatments into distinct groups. Many other hierarchical cluster analysis approaches have been proposed since Scott, A.J. and Knott, M. (Scott and Knott, 1974) published their results, as for example Jollife (1975), Cox and Spjotvoll (1982), and Calinski and Corsten (1985). However, the SK approach has been the most widely used due to the simple intuitive appeal of its idea, and also the good results it always gives (Gates and Bilbro, 1978; Bony et al., 2001; Dilson et al., 2002; Jyotsna et al., 2003).

The SK procedure uses a clever algorithm of cluster analysis, where, starting from the whole group of observed mean effects, it divides, and keep dividing the sub-groups in such a way that the intersection of any two groups formed in that manner is empty. Using A. J. Scott and M. Knott own words: "we study the consequences of using a well-known method of cluster analysis to partition the sample treatment means in a **balanced design** and show how a corresponding likelihood ratio test gives a method of judging the significance of the difference among groups obtained" (Scott and Knott, 1974).

Simulation studies show that the performance of the SK procedure, compared to the multiple comparison procedures is very good (Da Silva et al., 1999; Borges and Ferreira, 2003).

We try to motivate the reader into the practice of the SK algorithm by bringing a real data example and compare the SK with other procedures, namely the Tukey test (package **agricolae**) and clustering (`hclust`, package **stats**).

This paper illustrates the use of the **ScottKnott** R package, which implements the SK procedure (Scott and Knott, 1974). The package is available on the Comprehensive R Archive Network (CRAN) website at `http://CRAN.R-project.org/package=ScottKnott`.

The R Package **ScottKnott** is composed of two methods, `SK` and `SK.nest`. The method `SK` performs the algorithm on treatments of main factors and `SK.nest` does the same on nested designs of factorial, split-plot and split-split-plot experiments. They return objects of class `SK`, and `SK.nest` containing the groups of means plus other variables necessary for `summary` and `plot`.

The generic functions `summary` and `plot` are used to obtain and print a summary and a plot.

Real data study

As a motivation we will use an experiment conducted at EMBRAPA Milho e Sorgo (The Brazilian Agricultural Research Corporation, Corn and Sorghum Section). It was published in Ramalho et al. (2000) page 167. The experiment consists of 16 treatments (cultivars) of sorghum conducted in a balanced squared lattice design and the yield by plot ($kg/plot$). For our purposes, it can be considered a incomplete randomized block design with 4 blocks, 16 treatments, and 5 repetitions, that is, the yield of each treatment is measured 5 times. This data is available in the **ScottKnott** package as `sorghum`.

The objective of this study is to compare the 16 treatment means, as a first step of the whole analysis, and the question is: are there groups of treatments which could be considered homogeneous? In other words, would it be possible to find groups for which the treatment means belonging to those groups represent cultivars yielding the same weight of sorghum and the differences in the observed results being due to random variability?

We understand that this way of questioning is very important for the agricultural researcher and might be even more important than testing for the difference between treatment means for every pair of those means. There are 120 of such pairs.

Even though this first study is just exploratory it will give the researcher the insight he(she) needs to continue the analysis further on.

This exploratory analysis was carried out using the SK algorithm whose results were compared to two other methods.

```
> av <- aov(y ~ r/bl + x, data = sorghum$dfm)
> sk <- SK(av, which = "x", sig.level = 0.05)
```

The first is the function `hclust` found in the package **stats** which performs hierarchical cluster analysis on a set of dissimilarities. The agglomeration method used was “ward”. We calculated the mean value of the 5 repetitions for each treatment and used it in the function `hclust`.

```
> cl.m <- hclust(dist(sorghum.cl.m),
+     "ward")
```

The second, was the Tukey’s HSD test which is a multiple comparison procedure but also used by researchers as a method to divide the treatment means in group. The package **agricolae** was used.

```
> tk.ag <- HSD.test(av, "x", group = TRUE,
+     alpha = 0.05)
> bar.group(tk.ag, ylim = c(0, 12),
+     density = 4, border = "blue")
```

Figure 1 shows the result of the SK algorithm. It divided the means in two groups. The two main groups found using the function `hclust` (Figure 2) are not exactly the same as those found using the SK algorithm, but are similar as it should be expected. The treatment means 14,8,5,7,9,3 belong to the same group in both figures, and the treatments 1,2,4 which belong to the same group of the above treatments in figure 1, but in figure 2 they belong to the second main group. Nevertheless, they are grouped together at the lowest level. The function `hclust` cuts at the big gap between treatments 9 and 3 and the function `SK` at the big gap between treatments 2 and 12.

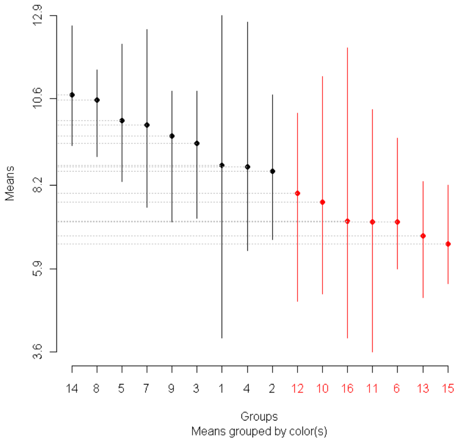

Figure 1: Yield of sorghum using Skott-Knott algorithm, $\alpha = 5\%$.

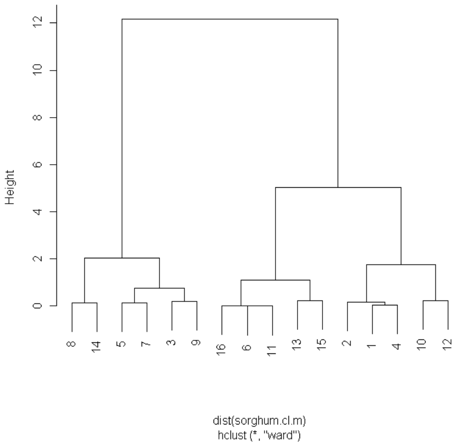

Figure 2: Yield of sorghum using hclust, method = ward.

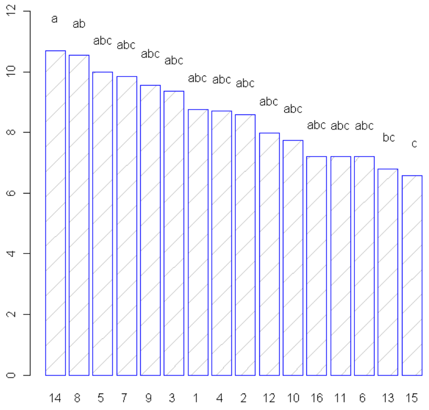

Figure 3: Yield of sorghum using Tukey, $\alpha = 5\%$.

Figure 3 shows the result by using the Tukey’s HSD test. It finds 3 groups, marked by the letters a,b

and c. Almost all the treatments are classified to the 3 groups and this overlapping makes it very difficult for the researcher to decide in which groups to sort those means.

What makes the SK algorithm more convenient for a user of ANOVA is that besides dividing the treatments in groups without overlapping, its result also uses a probabilistic approach in order to find the groups: the SK algorithm takes the maximum of between groups sum of squares which is used in a likelihood ratio test having an asymptotic χ^2 distribution. This approach is very useful when the number of treatments is large.

It is also possible to change the value of the parameter **sig.level** of the SK function, therefore getting diferent groupings, and so the researcher can check what makes sense in practice.

In a few words, the **sig.level** choice automates the group formation.

The SK clustering algorithm

Suppose we have a set of independent sample treatment means in the analysis of variance context, each treatment with the same number of replications, all normal variates, that is a **balanced design**. Furthermore, suppose that ANOVA leads to a significant *F-test* for the difference among the treatment means. Moreover, by rejecting the homogeneity of the treatment means there is a problem finding out how many homogeneous groups there are and which are the treatment means contained in each group.

It should be noted that we follow (Calinski and Corsten, 1985) in what we mean by homogeneity of treatments: “Once more it should be borne in mind that non rejection of equality is by no means equivalent to proving equality. We carefully defined homogeneity as non rejection of equality. Nor it should be inferred that treatments belonging to different ”homogeneous groups” are (significantly) different; treatments belonging to the same group, however, are not.”

The SK procedure is a hierarchical clustering algorithm which attempts to find out those groups without overlapping.

Let k be the number of treatments. As it starts, the SK procedure will either find two distinct groups dividing the treatment means or will declare those k treatment means a homogeneous belonging to just one group. To do so it should look at the $2^{k-1}-1$ possible partitions of the k means into two nonempty groups, but it is enough to look at the $k-1$ partitions formed by ordering the treatment means and dividing them between two successive ones (Scott and Knott, 1974). Let T_1 and T_2 be the totals of two of those groups with k_1 and k_2 treatments in each one, so that $k_1+k_2=k$, that is:

$$T_1=\sum_{i=1}^{k_1} y(i) \;\; T_2=\sum_{i=k_1+1}^{k_1+k_2} y(i)$$

Where $y(i)$, $i=1:m$ are the ordered treatment means and y the grand mean (Ramalho et al., 2000).

Also, let B be the between groups sum of squares. That is:

$$B=\frac{T_1^2}{k_1}+\frac{T_2^2}{k_1}-\frac{(T_1+T_2)^2}{k_1+k_2}$$

Let B_o be the maximum value, taken over the $k-1$ partitions of the k treatments into two groups, of the between groups sum of squares B. After finding out those groups we use the likelihood ratio test for the null hypothesis of equality of all means against the alternative that they belong to the two groups found above. If we reject this hypothesis then the two groups are kept, otherwise the group of k treatment means is considered homogeneous. We then repeat this procedure for each group separated and stop until all the groups formed up to then are homogeneous.

The statistics used for the likelihood ratio test is:

$$\lambda=\frac{\pi}{2(\pi-2)}\times\frac{B_o}{\sigma_o^2}$$

where σ_o^2 is the maximum likelihood estimator of $\frac{\sigma^2}{r}$.

Let $s^2=\frac{MSE}{r}$ be the unbiased estimator of $\frac{\sigma^2}{r}$, v be degrees of freedom associated with that estimator, then

$$\sigma_o^2=\frac{\sum_{i=1}^{k}(y(i)-y)^2+vs^2}{k+v}$$

λ is asymptotically a χ^2 distributed random variable with $v_o=\frac{k}{\pi-2}$ degrees of freedom. Therefore we can use that to set the cutoff point for a given α value each time we perform the test.

We can think the p-value of likelihood ratio test as a distance to be measured between the two selected groups and the chosen type I error (α value) to be the cut off. If the p-value is smaller than α the groups are too far away from each other and should be separated (they are heterogeneous) otherwise, they become just one group (homogeneous).

“Choosing an appropriate value for α is difficult. If α is too small, the splitting process will terminate too soon, while if α is too large, the process will go too far and split homogeneous sets of means” (Scott and Knott, 1974).

As we start dividing the first groups into other smaller groups, we repeat the same algorithm for each group. We keep doing that until every group formed in this way is either homogeneous or just contains one observed mean.

It is important to emphasize the fact that the α value defined above is not the nominal error rate of the type I error of the algorithm as a whole. If we set the α value to be 5% then every test the SK procedure performs to divide or not a sub-group has a type

I error rate of 5% but we cannot say that the former type I error rate is 5%. This α value is the parameter called **sig.level** in the SK function.

Comparative performance of SK method

In performance studies among statistical tests is often very difficult to obtain analytically their rate of type I error and power. The most usual way to get that information is through simulation using Monte Carlo methods. Boardman and Moffit (1971) show that the difference between analytical values and Monte Carlo's is very small therefore making its use an optimal way to get the necessary information. Their results are similar to those found by Bernhardson (1975).

In spite of the SK being a clustering procedure we can use simulation results to compare its performance to Tukey test and others, as if it were a multiple comparison procedure.

The two of the most common measures to compare "Multiple Comparison Procedures" found in the literature are:

- The ratio between the number of type I errors (reaching the result that $\mu_i \neq \mu_j$ when truly $\mu_i = \mu_j$) and the number of comparisons is defined as *comparisonwise error rate.*

- The ratio between the number of experiments with one or more type I errors and the total number of experiments is defined as *experimentwise error rate* (Carmer and Swanson, 1971; Steel and Torrie, 1980).

Two simulation studies (Boardman and Moffit, 1971; Bernhardson, 1975) conducted at the Universidade Federal de Lavras, Brazil, used Monte Carlo methods to evaluate the performance of the SK method. One Da Silva et al. (1999) has shown that it possesses high power and error rate almost always in accordance with the nominal levels using both comparisonwise and experimentwise error rates. That is, the rates are not far from α value cited above. The other, Borges and Ferreira (2003) evaluated the power and the type I error rates of the SK, Tukey and SNK test, in a wide variety of experimental situations, in conditions of normality and non-normality error distribution. They concluded that the SK is more powerful than the other two and is also robust against violations of normality assumptions. Both performed 2000 simulations for each experiment with 5, 10, 20 and 80 treatments with 4, 10 and 20 replications α value of 1% and 5% plus coefficients of variation 1%, 10%, 20% and 30%.

The ScottKnott package

The package **ScottKnott** was written in R language (R Development Core Team, 2010). It's results are objects of the class `list`, `SK` and `SK.nest`, which are used as input to the generic functions `summary` and `plot`.

The ScottKnott package performs the clustering algorithm on three designs and three experiments. It must be emphasized again that the two functions `SK` and `SK.nest` **work only on balanced designs**.

The designs are: Completely Randomized Design (CRD), Randomized Complete Block Design (RCBD) and Latin Squares Design (LSD). The experiments are: Factorial Experiment (FE), Split-Plot Experiment (SPE) and Split-Split-Plot Experiment (SSPE) .

The package ScottKnott has two main functions, `SK` and `SK.nest`. The function `SK` is used for clustering treatment means of a main factor. The function `SK.nest` is used for clustering treatment means of interaction among factors, that is whenever the treatment means belong to a factor nested in others. For example the treatment means of factor A for level 1 of factor B and level 1 of factor C. The function `SK.nest` supports at most two nesting as shown above.

The function `summary` generates an output where the different groups are shown by using letters of the alphabet. The `plot` function generates distinct groups differentiated by colors.

The main algorithm is the function `MaxValue` which builds groups of means according to the method of SK. Basically it is an algorithm for pre-order path in a binary decision tree. Every node of this tree, represents a different group of means and, when the algorithm reaches this node it takes the decision to either split the group in two, or form a group of means. At the end all the leaves of the tree are the groups of homogeneous means.

The functions `SK` and `SK.nest` are methods for objects of class `vector`, `matrix` or `data.frame` joined as default, and `aov` and `aovlist` for single experiments.

The main parameters used by those methods are:

- x: A design matrix, `data.frame` or an `aov` object.

- y: A vector of response variable. It is necessary to inform this parameter only if `x` represent the design matrix.

- which: The name of the factor to be used in the clustering. The name must be inside quoting marks.

- model: If x is a `data.frame` object, the model to be used in the aov must be specified.

- error: The error to be considered. Used only in case of *split-plot* or *split-split-plot* experiments.

- sig.level: Level of Significance, α value, used in the SK and SK.nest algorithms to create the groups of means. The default value is 0.05.

- fl_2: A vector of length 1 giving the level of the second factor in nesting order tested.

- fl_3: A vector of length 1 giving the level of the third factor in nesting order tested.

- id.trim: The number of characters to trim the label of the factor levels.

- ...: Further arguments (required by generic).

Split-Split-Plot Experiment (SSPE)

We show an example of how to use the **ScottKnott** package. An object of class aovlist will be used in the function SK.nest.

SSPE is the objet containing the data set of a Split-Split-Plot Experiment (SSPE). It is a simulated data to model a SSPE with 3 plots, each one split 3 times, each split, split again 5 times and 4 repetitions per split-split.

It can be called using the command below:

```
> data(SSPE)
> nav <- with(SSPE, aov(y ~ blk + ssp *
+     sp * p + Error(blk/p/sp), data = dfm))
```

The factor *ssp* is nested in factor *sp* which is nested in factor *p*. The value 1 of the parameter fl_2 and 1 of parameter fl_3 mean that the first level of factor *p* and factor *sp*, respectively, are chosen. The comparison is made only among levels (treatments) of factor *ssp* belonging to that particular combination of levels of factor *p* and factor *sp*. Look at the aov(model) and SK.nest (which) functions for the order at which the factors appear.

```
> nsk <- SK.nest(nav, which = "ssp:sp:p",
+     error = "Within", fl2 = 1, fl3 = 1)
> plot(nsk, rl.col = c(rep("black",
+     3), rep("red", 2)), title = "")
```

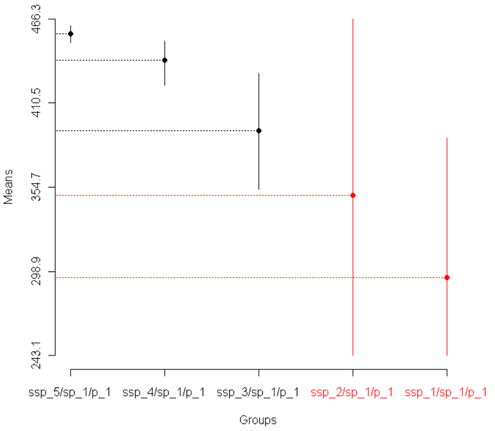

Figure 4: Split-Split-Plot Experiment (SSPE). Nested analysis (ssp/sp=1/p=1), $\alpha = 5\%$.

Further examples are documented in the folder *demo* of the R-package **ScottKnott**.

Bibliography

C.S. Bernhardson. Type I error rates when multiple comparison procedure follow a significant F test of ANOVA. *Biometrics*, **31**(1):337-340, 1975.

T.J. Boardman and D.R. Moffit. Graphical Monte Carlo Type I error rates for multiple comparison procedures. *Biometrics*, **27**(3):738-744, 1971.

S. Bony, N. Pichon, C. Ravel, A. Durix, F. Balfourier and J.J. Guillaumin. The Relationship between Mycotoxin Synthesis and Isolate Morphology in Fungal Endophytes of Lolium perenne. *New Phytologist*, **152**(1):125-137, 2001.

L.C. Borges, D.F. Ferreira. Power and type I errors rate of Scott-Knott, Tukey and Newman-Keuls tests under normal and no-normal distributions of the residues. *Revista de Matemática e Estatística*, **21**(1):67-83, 2003.

T. Calinski and L.C.A, Corsten. Clustering Means in ANOVA by Simultaneous Testing. *Biometrics*, **41**(1):39-48, 1985.

S.G. Carmer and M.R. Swanson. Detection of differences between means: a Monte Carlo study of five pairwise multiple comparison procedures. *Agronomy Journal*, **63**(6):940-945, 1971.

S.G. Carmer and M.R. Swanson. An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. *Journal American Statistical Association*, **68**:66-74, 1973.

D.R. Cox and E. Spjotvoll. On partitioning means into groups. *Scandinavian Journal of Statistics*, **9**:147-152, 1982.

E.C. Da Silva, D.F. Ferreira and E. Bearzoti. Evaluation of power and type I error rates of Scott-Knott's test by the method of Monte Carlo. *Ciências Agrotécnicas.*, **23**:687-696, 1999.

A.B. Dilson, S.D. David, J. Kazimierz and W.K. William. Half-sib progeny evaluation and selection of potatoes resistant to the US8 genotype of Phytophthora infestans from crosses between resistant and susceptible parents. *Euphytica*, **125**:129-138, 2002.

C.E. Gates and J.D. Bilbro. Illustration of a Cluster Analysis Method for Mean Separation. *Agron J*, **70**:462-465, 1978.

I.T. Jollife. Cluster analysis as multiple comparison method. *Applied Statistics.* RP Gupta (ed):159-168, North Holland, 1975.

S. Jyotsna, L.W. Zettler, J.W. van Sambeek, Ellersieck, C.J. Starbuck. Symbiotic Seed Germination and Mycorrhizae of Federally Threatened Platanthera Praeclara(Orchidaceae). *American Midland Naturalist*, **149**S(1):104-120, 2003.

R. O'Neill and G.B. Wetherill. The present state of multiple comparison methods (with discution). *Journal of the Royal Statistical Society Series B*, **33**:218-250, 1971.

M.A.P. Ramalho, D.F. Ferreira and Oliveira AC. *Experimentação em Genética e Melhoramento de Plantas.* Editora UFLA, Lavras, 2000.

R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0, URLhttp://www.R-project.org/.

A.J. Scott and M. Knott. A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, **30**:507-512, 1974.

R.G.D. Steel and J.H. Torrie. *Principles and procedures of statistics.* McGraw-Hill, New York, 1980.

Enio Jelihovschi
Departamento de Ciências Exatas e Tecnológicas - DCET
Universidade Estadual de Santa Cruz - UESC
Brazil
http://www.uesc.br/
eniojelihovs@gmail.com

José Cláudio Faria
Departamento de Ciências Exatas e Tecnológicas - DCET
Universidade Estadual de Santa Cruz - UESC
Brazil
http://www.uesc.br/
joseclaudio.faria@gmail.com