UNIVERSIDAD CARLOS III DE MADRID working papers

Working Paper 09-85
Statistics and Econometrics Series 26
December 2009

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-49

RECOMBINING DEPENDENT DATA: AN ORDER STATISTICS APPROACH

Adolfo Álvarez* and Daniel Peña**

Abstract

This article discusses the problem of forming groups from previously split data. Algorithms for Cluster Analysis like SAR proposed by Peña, Rodriguez and Tiao (2004), divide the sample into small very homogeneous groups and then recombine them to form the definitive data configuration. This kind of splitting leads to dependent data in the sense that the groups are disjoint, so no traditional homogeneity of means or variances tests can be used.

We propose an alternative by using Order Statistics. Studying the distribution and some moments of linear combination of Order Statistics it is possible to recombine disjoint data groups when they merge into a sample from the same population.

Keywords: SAR, Cluster Analysis, Order Statistics, L-statistics, Bootstrapping.

* Álvarez Pinto, Adolfo, Departamento de Estadística, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), e-mail: aaapinto@est-econ.uc3m.es

* *Peña Sánchez de Rivera, Daniel, Departamento de Estadística, Universidad Carlos III de Madrid, C/ Madrid 126, 28903 Getafe (Madrid), e-mail: dpena@est-econ.uc3m.es

Aknowledgements: This research was supported by Spanish Ministry of Science and Innovation, project SEJ2007-64500. Adolfo Álvarez is supported by the F.P.I. program from Ministry of Science and Innovation, reference BES-2008-009290.

Recombining dependent data: An Order Statistics Approach

Adolfo Álvarez P. and Daniel Peña S.

Abstract.

This article discusses the problem of forming groups from previously split data. Algorithms for Cluster Analysis like SAR proposed by Peña, Rodriguez and Tiao (2004), divide the sample into small very homogeneous groups and then recombine them to form the definitive data configuration. This kind of splitting leads to dependent data in the sense that the groups are disjoint, so no traditional homogeneity of means or variances tests can be used.

We propose an alternative by using Order Statistics. Studying the distribution and some moments of linear combination of Order Statistics it is possible to recombine disjoint data groups when they merge into a sample from the same population.

Keywords: SAR, Cluster Analysis, Order Statistics, L-statistics, Bootstrapping.

Introduction: Model Heterogeneity

In statistical analysis, we speak of "model heterogeneity" when not all the data points in the sample can be explained by the same model. For example, one of the applications of model heterogeneity is the problem of outliers, where most of the data points come from the same distribution but a few of them have been generated by one or several distributions which differ from the previous one.

The existence of model heterogeneity can bring significant complications when

performing inference, because biased estimates of the parameters can be obtained, with the consequent loss of efficiency in estimation and a bad prediction.

In multivariate analysis, model heterogeneity has been studied mainly under the name of “cluster analysis”. In particular, Peña (2002) define it as the analysis which has as a main objective to arrange the observations into homogeneous groups by means of defining similarities between them. Commonly Cluster Analysis is used to join data points but also is possible to apply it to arrange variables.

These methods are also known as Automatic Unsupervised Classification Methods or Unsupervised Pattern Recognition Methods. The name of “unsupervised” is used to distinguish them from discriminant analysis, where the researcher possess labels or classifyers to identify the groups where the observation belongs.

According to Peña (2002) Cluster analysis deals with three kind of problems:

- Partition of the data. In which available data are suspected to be heterogeneous and want to divide them into a fixed number of clusters (MacQueen, 1967; Anderberg, 1973; Hartigan and Wong, 1979; Dubes, 1987) so that (1) Each element belongs to one and only one of the groups; (2) Each item is classified and (3) Each group is internally homogeneous.
- Construction of hierarchies. In which the aim is to structure hierarchically the elements of a data set by their similarity. Strictly speaking, these methods don't define groups, but they show the structure of chain association that may exist between the elements, however, the hierarchy obtained, also allows a partition of the data into groups (King, 1967; Ward, 1963; Murtagh, 1984) .
- Classification of variables. In presence of many variables, it is interesting to make an initial exploratory study to divide the variables into groups. Such studies may be useful as a guide prior to the application of formal models to reduce dimensionality (Gnanadesikan et al, 1995; Raftery and Dean, 2006).

As can be seen, Cluster Analysis, as a particular case of the use of model heterogeneity, covers a wide variety of problems, which in turn can be approached from several viewpoints. Good references for this can also be founded in Peña (2002), Hartigan (1975), Kaufmann and Rosseeuw (1990), Jain, Murty and Flynn (1999), and Gan et al (2007).

The SAR Process

Peña, Rodriguez and Tiao (2004) propose a new exploratory approach to address the problem of identifying clusters in particular, and model heterogeneity in general. The method, named by authors as SAR (split and recombine), divide the sample into smaller subgroups and then recombine them to form the final clusters. However, as mentioned above, this methodology is general enough to encompass also problems of identification of outliers, both in multivariate cluster analysis (Peña, Rodriguez and Tiao, 2004) and in regression (Peña, Rodriguez and Tiao, 2003)

The SAR procedure is based on the concept of Model Heterogeneity as follows:

Let M be the model adjusted to a set of n observations $Y=(y_1, y_2, \cdots, y_n)$, where y_i is a vector of dimension m. The procedure is based on defining a measure $H(y, Y)$ of heterogeneity between an observation y and the data set Y , and iteratively use this measure to cover the following steps: To identify outliers and eventually delete them from the sample; to split the sample into more homogeneous groups and finally recombine the observations to form the final clusters.

To this end, the authors argue that the natural way to test whether a new observation is homogeneous with respect to the rest of data set is to see whether this element is close to its prediction based on Y , and the model M , with p-dimensional vector of parameters θ . Then assuming that for certain θ , observations Y and y are independent, the distribution of the prediction for a new data point y given Y is equal to:

$$p(y/Y)=\int p(y/\theta)\,p(\theta/Y)\,d\theta,$$

where $p(y/Y)$ is the distribution of the data point y, while $p(\theta/Y)$ is the posterior distribution for parameter θ. Thus, if the density of the observed value is small, there is reason to believe that this value is heterogeneous with respect to the sample Y.

However, it is not always easy to obtain these distributions, so the authors propose an alternative by normalizing the predictive density over the modal value $\hat{y}$, which yields the following measure of heterogeneity:

$$H(y,Y)=C_0(y)=-2\ln\frac{p(y/Y)}{p(\hat{y}/Y)}.$$

Assuming a set of independent observations coming from an univariate normal distribution $N(\mu,\sigma^2)$, where distribution parameters (μ,σ) has non-informative a priori distribution $p(\mu,\sigma)\propto\sigma^{-1},$, then $\hat{y}=E(y/Y)$ and the measure of heterogeneity is defined as:

$$C_0(y)=N\ln\left\{1+\frac{t^2}{\nu}\right\}$$

where $\nu=N-1, t^2=\left(\frac{N}{N+1}\right)\frac{(y-\bar{y})^2}{s^2}$, $\bar{y}$ is the sample mean of the N observations on Y, and $s^2=\nu^{-1}\sum_j(y_j-\bar{y})^2$, is the corresponding sample variance. Finally t^2 has a F distribution with 1 and N-1 degrees of freedom.

Splitting Process

They define y_l as the discriminator of y_i if the latter observation appears as most discrepant (using the heterogeneity measures) with respect to the rest of the data set when the discriminator is deleted from the sample. In this way: If two observations are identical, they must have the same discriminator, thus, if they are sufficiently close to each other, they should still have the same discriminator. Finally, the splitting process consist into:

- Identify and eliminate outliers, based on the heterogeneity measure

- Points sharing a common discriminator are put in the same group. (Discriminators are considered as isolated observations)
- Then, each group is now considered as a new sample and the procedure is continued until splitting further the sample will lead to groups that all of them are of size smaller than some minimum size n_0 .
- When a group can not be split again, is called a "basic set". The minimum size is proposed as $n_0 = p + h$, where $h \geqslant 2$ and p is the number of coefficients of the fitted model.

Recombination process

Since the partitioning stage will tend to define many groups, it is important to have a procedure for recombining the observations after the split. The more we split the sample, the smaller the internal variability of the resulting groups, so it requires a process that increases the internal variability of homogeneous groups, incorporating new observations, but at the same time avoiding the inclusion of observations that are clearly heterogeneous with respect to the group. So, recombination is established as follows:

- Calculate $C_0(y_i)$ for each point outside the core set.
- Find the nearest point y_l to the basic set, i.e. one that satisfy $C_0(y_l) = min_{y_i} C_0(y_i)$
- If $C_0(y_l)$ is below a certain cutoff value, c_N , which depends on the size of the basic set, N, the point is incorporated into the basic set to form a new group of size $N+1$, and the process repeats until the closest point to the group exceeds the cut-off value. Then the basic set is considered as an homogeneous group.

After applying the recombination process to all basic sets, there are two possible situations:

a) All basic sets are increased to include the entire sample, or constitute a single

partition of the sample in a set of disjoint groups and some outliers.

b) After eliminating redundancies, some enlarged basic groups overlap with others. In this case we again apply the three steps of eliminating outliers, splitting and recombination to the supplementary part of a group, treating this data as a new sample. The process continues for each basic group, creating a branch structure until the entire sample is split into several disjoint subsets. Each different form of splitting is then regarded as a Possible Data Configuration data (PCD). When more than one PCD is found, the problem of choosing the best can be solved by some model selection procedure.

Dependent Data Recombination

The implementation of the SAR algorithm proposed by Peña, Rodriguez and Tiao (2004) may be high time consuming when you have a large sample size n, high dimensionality and / or when you have too many basic groups. Moreover, the recombination process does not take into account the information obtained by the splitting process, because the observations that belong to the same basic group are regarded as isolated points when the algorithm is trying to enlarge another one.

A possible improvement for these drawbacks is to make the process of recombination not by observations, but considering each of the basic groups as a unit to recombine. Thus, the process becomes more efficient in time and, moreover, has the advantage of considering the information obtained in the partition, because the data points which were already united in this first stage will remain together in the second, and an unique solution can be founded.

In this manner, the usual way to check if two groups come from the same population is by performing an hypothesis test like equality of means, or equality of variances test, or both at the same time (Mardia et al, 1979). However, in this case, the basic groups doesn't hold with the condition of independence, because they are not independent samples from a population, but disjoint partitions of samples.

By construction, the partition of a sample by some criteria involves defining certain order. And in particular, if we are interested in check if two disjoint groups forms a partition (or sub partition) of the same sample we can study the distribution of the order statistics or a linear combination of them, to perform a test.

Linear Combination of Order Statistics / Bootstrap Approach

Let $X_{1:n} \leqslant X_{2:n} \leqslant ... \leqslant X_{n:n}$ be an ordered sample, if we split the sample into two groups of sizes n_1 and n_2 such that $n_1 + n_2 = n$:

$$\underbrace{X_{1:n} X_{2:n} X_{3:n} ... X_{n_1:n}}_{n_1} \underbrace{X_{n_1+1:n} X_{n_1+2:n} ... X_{n:n}}_{n_2}$$

then, the difference between the means of these two groups will be given by:

$$D = \bar{X}_2 - \bar{X}_1 = \frac{X_{n_1+1} + X_{n_1+2} + ... + X_n}{n_2} - \frac{X_1 + X_2 + ... + X_{n_1}}{n_1}$$

This is a linear combination of order statistics, also called “L-statistic” in this form:

$$T_n = \sum_{i=1}^{n} c_i X_{i:n}$$

then is possible to write the difference D as follows:

$$D = \left(\frac{-1}{n_1}\right) X_1 + \left(\frac{-1}{n_1}\right) X_2 + ... + \left(\frac{-1}{n_1}\right) X_{nI} + \left(\frac{1}{n_2}\right) X_{n_1+1} + \left(\frac{1}{n_2}\right) X_{n_1+2} + ... + \left(\frac{1}{n_2}\right) X_n$$

and in this case vector of constants will be $c = \left[-\frac{1}{n_1}; -\frac{1}{n_1} ... -\frac{1}{n_1}; \frac{1}{n_2}; \frac{1}{n_2}; ... \frac{1}{n_2}\right]$

So, if two groups come from the same population, it is possible to study the

distribution and moments of the difference between the means of the groups in order to make a Test to merge them.

However, how to estimate moments of this statistics? This kind of expressions are unsolved for main distributions, and only approximations have been made(Stigler, 1969; Balakrishnan et al, 2003; Rychlik, 2004; Kaluszka and Okolewski, 2005). Lately, another approaches has been attempted, like Bootstrap (Hutson and Ernst, 2000), Jackknife (Parr and Shucany, 1982), or B-splines (Agarwal and Pant, 2008).

In this paper we propose the use of moments of these difference of means, based on bootstrap methodology. In particular, Hutson and Ernst (2000) propose exact bootstrap mean and variance of L-estimators based on exact bootstrap mean, variances and covariances of the whole set of order statistics from a sample, with this formulae:

$$E^*(X_{r:n}) = \sum_{j=1}^{n} w_{j(r)} X_{j:n}$$

$$Var^*(X_{r:n}) = \sum_{j=1}^{n} w_{j(r)} (X_{j:n} - \hat{\mu}_{r:n})^2$$

$$Cov^*(X_{r:n}, X_{s:n}) = \sum_{j=2}^{n} \sum_{i=1}^{j-1} w_{ij(rs)} (X_{i:n} - \hat{\mu}_{r:n})(X_{j:n} - \hat{\mu}_{s:n}) + \sum_{j=1}^{n} v_{j(rs)} (X_{j:n} - \hat{\mu}_{r:n})(X_{j:n} - \hat{\mu}_{s:n})$$

where:

$$w_{j(r)} = r \binom{n}{r} \left[B\left(\frac{j}{n}; r, n-r+1\right) - B\left(\frac{j-1}{n}; r, n-r+1\right) \right]$$

$$w_{ij(rs)} = nCrs \sum_{k=0}^{s-r-1} \binom{s-r-1}{k} \frac{(-1)^{s-r-1-k}}{s-k-1} \left[\left(\frac{i}{n}\right)^{s-k-1} - \left(\frac{i-1}{n}\right)^{s-k-1} \right]$$
$$\times \left[B\left(\frac{j}{n}; k+1, n-s+1\right) - B\left(\frac{j-1}{n}; k+1, n-s+1\right) \right]$$

$$v_{j(rs)} = nCrs \sum_{k=0}^{s-r-1} \binom{s-r-1}{k} \frac{(-1)^{s-r-1-k}}{s-k-1}$$
$$\left\{ B\left(\frac{j}{n}; s, n-s+1\right) - B\left(\frac{j-1}{n}; s, n-s+1\right) - \left(\frac{j-1}{n}\right)^{s-k-1} \left[B\left(\frac{j}{n}; k+1, n-s+1\right) - B\left(\frac{j-1}{n}; k+1, n-s+1\right) \right] \right.$$

$B(x,a,b)=\int_0^x t^{a-1}(1-t)^{b-1}dt$ (is the incomplete beta function) and finally,

$$nCrs=\frac{n!}{(r-1)!(s-r-1)!(n-s)!}$$

Proceeding by this way, the error due to bootstraping resampling is eliminated, and the expectation and variance of any linear combination of order statistics can be obtained. Let $c=(c_{1,}c_{2,\ldots},c_n)'$ the vector of constants corresponding to a specific L-estimator, and let:

$$\hat{\mu}=(\hat{\mu}_{1:n},\hat{\mu}_{2:n},\ldots,\hat{\mu}_{n:n})'$$

be the exact bootstrap mean vector of the order statistics. Therefore, let

$$\hat{\Sigma}=\begin{pmatrix}\hat{\sigma}^2_{1:n} & \hat{\sigma}_{12:n} & \cdots & \hat{\sigma}_{1n:n}\\ \hat{\sigma}_{21:n} & \hat{\sigma}^2_{2:n} & \cdots & \hat{\sigma}_{2n:n}\\ \vdots & \vdots & \ddots & \vdots\\ \hat{\sigma}_{n1:n} & \hat{\sigma}_{n2:n} & \cdots & \hat{\sigma}^2_{n:n}\end{pmatrix}$$

be the bootstrap variance-covariance matrix whose elements are obtained as showed before.

Thus, the bootstrap mean of the L-statistics T_n is given by:

$$\hat{\mu}_{T_n}=c'\hat{\mu}=\sum_{i=1}^{n}c_i\hat{\mu}_{i:n}$$

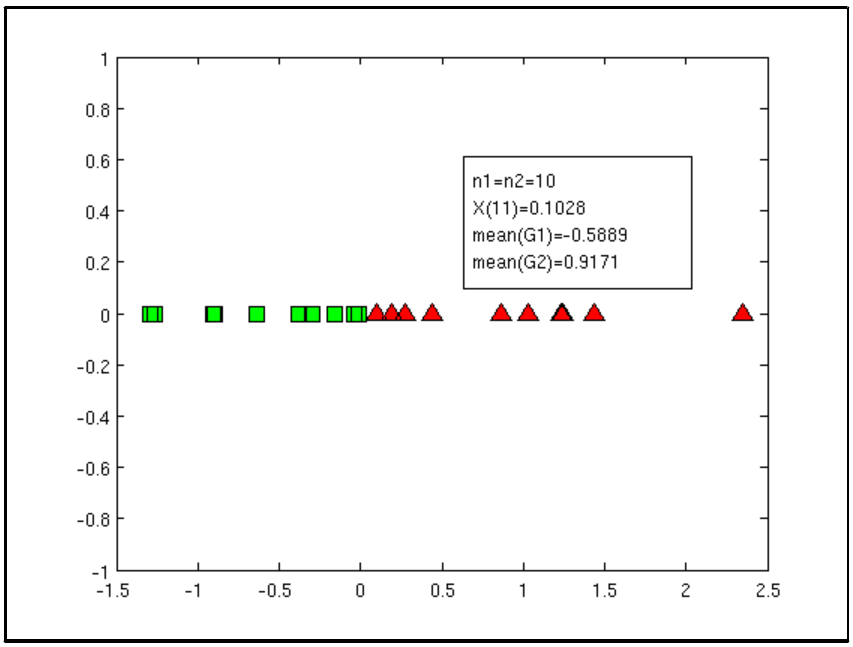
and the bootstrap variance will be:

$$\hat{\sigma}^2_{T_n}=c'\hat{\Sigma}c=\sum_{i=1}^{n}c_i^2\hat{\sigma}^2_{i:n}+2\sum\sum_{i<j}c_ic_j\hat{\sigma}_{ij:n}$$

But, how can bootstrap help us to test if two dependent sample come from the same population?

Example:

Let X be a sample of size n=20 coming from a normal distribution:



X	
G1	**G2**
-1,2926	0,1028
-1,2649	0,1924
-0,9082	0,2769
-0,8987	0,4447
-0,6361	0,8625
-0,3797	1,0288
-0,2996	1,2372
-0,1624	1,2447
-0,0346	1,4309
-0,0119	2,3496

Figure 1: sample of size 20, from a N(0,1) split into two groups

Now consider n bootstrap samples taken from the second group:

B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	…
0,1028	0,1028	0,1924	0,1924	0,1028	0,1028	0,1028	0,1924	0,2769	0,1028	…
0,8625	0,1924	0,1924	0,1924	0,1028	0,2769	0,1924	0,4447	0,4447	0,1028	…
0,8625	0,2769	0,1924	0,8625	0,1924	0,2769	0,2769	0,4447	0,8625	0,1924	…
0,8625	0,2769	0,2769	0,8625	0,1924	0,4447	0,4447	0,4447	0,8625	0,2769	…
0,8625	0,8625	0,2769	0,8625	0,1924	0,8625	0,8625	0,4447	0,8625	0,2769	…
1,0288	0,8625	1,2372	1,4309	0,4447	1,4309	1,2372	0,8625	1,0288	1,2372	…
1,0288	1,2372	1,2447	1,4309	0,8625	1,4309	1,2372	1,0288	1,2372	1,2372	…
1,2447	1,2372	1,2447	1,4309	1,0288	2,3496	1,2447	1,0288	1,2372	1,2447	…
1,2447	1,2447	1,2447	2,3496	1,2447	2,3496	1,2447	1,0288	1,4309	2,3496	…
2,3496	2,3496	2,3496	2,3496	1,4309	2,3496	1,2447	1,2447	2,3496	2,3496	…

Table 1: Bootstrap samples obtained from the second group of an ordered N(0,1) sample

The bootstrap distribution of the first element from the second group will be:

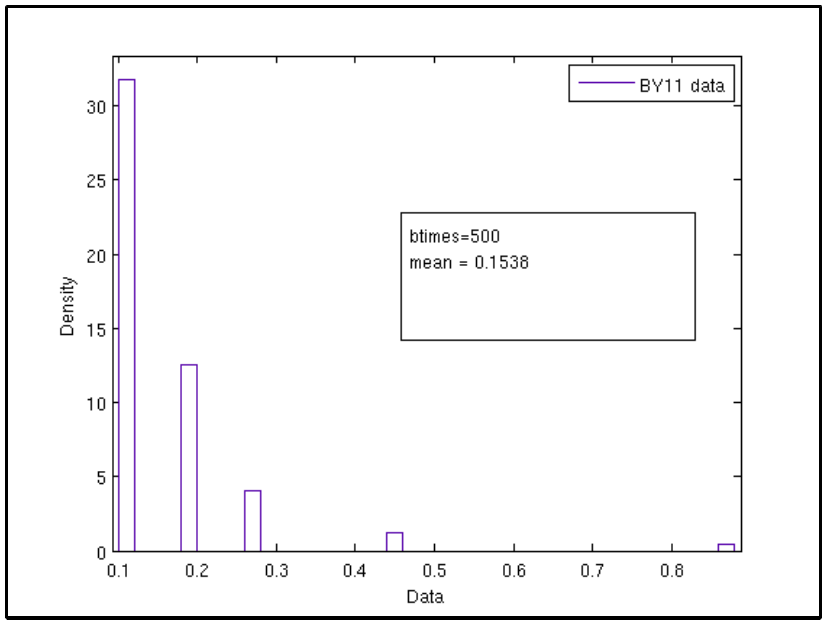


Figure 2: Bootstrap distribution of the first element of second group

Now, we consider n bootstrap samples taken from the entire sample:

	B1	**B2**	**B3**	**B4**	**B5**	**B6**	**B7**	**B8**	**B9**	**B10**	**…**
1	-1,2926	-1,2926	-1,2926	-1,2649	-0,9082	-1,2926	-1,2926	-1,2926	-1,2926	-1,2926	…
2	-1,2649	-0,9082	-0,8987	-0,9082	-0,8987	-1,2649	-1,2926	-1,2649	-1,2926	-0,9082	…
3	-1,2649	-0,8987	-0,3797	-0,9082	-0,8987	-1,2649	-0,6361	-1,2649	-1,2926	-0,8987	…
4	-0,3797	-0,8987	-0,2996	-0,8987	-0,6361	-0,8987	-0,6361	-0,6361	-1,2649	-0,8987	…
5	-0,2996	-0,6361	-0,2996	-0,3797	-0,6361	-0,3797	-0,6361	-0,6361	-0,6361	-0,6361	…
6	-0,1624	-0,6361	-0,1624	-0,3797	-0,3797	-0,3797	-0,3797	-0,3797	-0,3797	-0,6361	…
7	-0,1624	-0,3797	-0,1624	-0,3797	-0,2996	-0,3797	-0,2996	-0,2996	-0,2996	-0,6361	…
8	-0,1624	-0,2996	-0,0346	-0,2996	-0,2996	-0,2996	-0,1624	-0,0346	-0,1624	-0,3797	…
9	-0,1624	-0,1624	0,1028	-0,2996	-0,2996	-0,2996	-0,0346	-0,0119	-0,0346	-0,3797	…
10	-0,0119	-0,1624	0,1028	-0,0346	-0,0346	-0,2996	-0,0346	0,2769	-0,0119	-0,3797	…
11	**-0,0119**	**-0,1624**	**0,2769**	**0,1924**	**-0,0119**	**-0,0346**	**-0,0346**	**0,2769**	**0,1028**	**-0,2996**	**…**
…	…	…	…	…	…	…	…	…	…	…	…

Table 2: Bootstrap samples obtained from the entire sample from a N(0,1)

And the bootstrap distribution of the 11th element from the entire sample is:

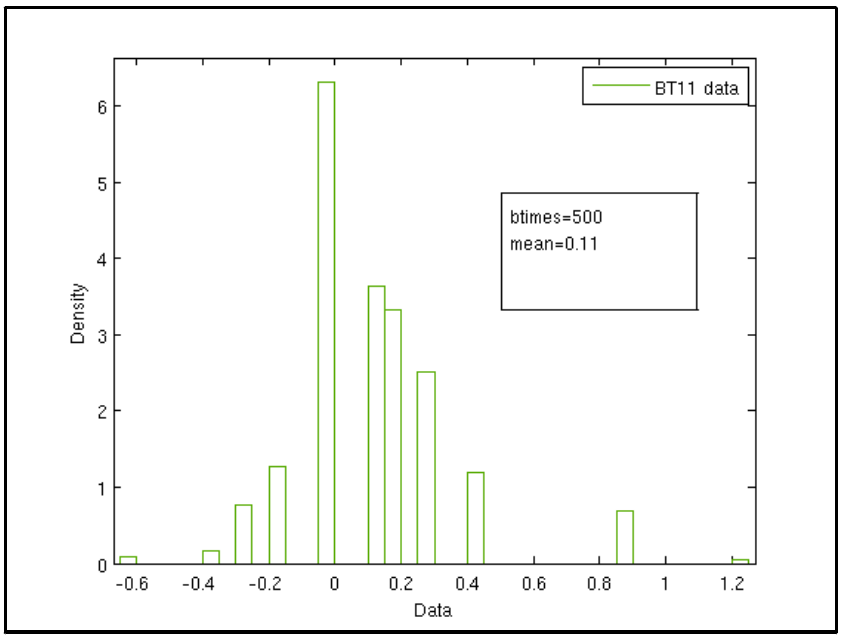


Figure 3: Bootstrap distribution of 11th element of the sample

So, if the groups are close enough, i.e. if both together make a sample from the same distribution, then most of the times of the bootstrap process, the 11th element bootstraped from the entire sample and the 1st element bootstraped from the second group should be close enough. Then , $E(X^{total}_{(11:20)}) \sim E(X^{2nd}_{(1:10)})$ and finally, the difference between them will be:

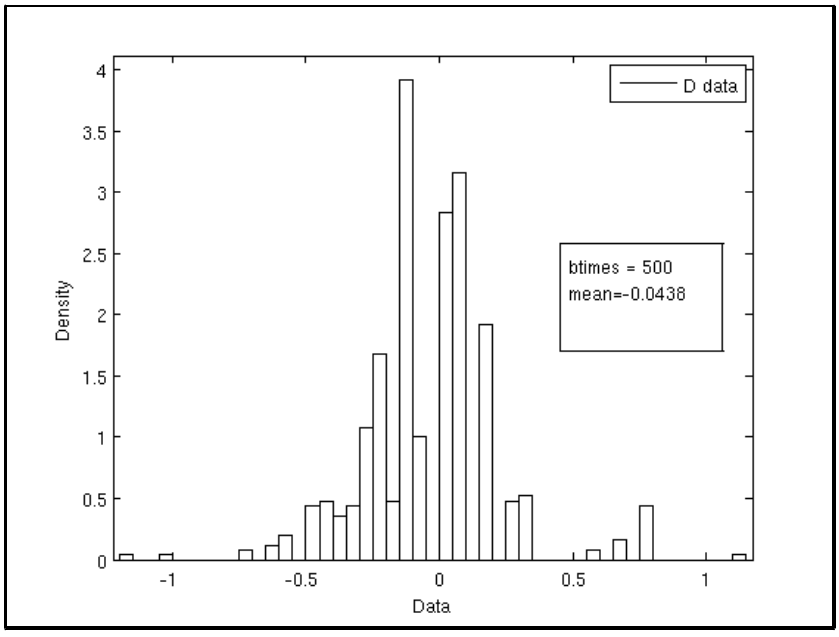


Figure 4: Distribution of the difference between the bootstrap first element of group 2 and 11th of the total sample.

We can see that the difference is centred on zero, but the procedure is not robust enough, because is based in the bootstrap of one element. But following the same structure also it is possible to consider the difference between the means of the second group (bootstraping only from it) and the second part obtained bootstraping from the entire sample.

Example:

We generate 1000 (sorted) samples from a standard Normal distribution(n=40), and then split it into 4 groups of 10 observations each from lowest to highest values in this way:

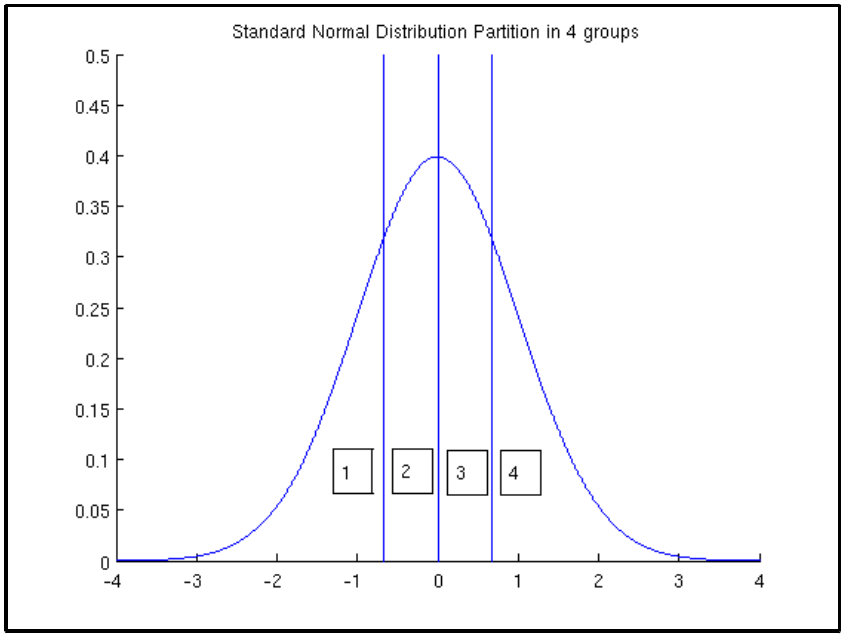


Figure 5: Partition methodology of a Normal Distribution in four groups

For each sample we bootstrap 10.000 times from the second group and 10.000 times from the entire sample. Then for each bootstrap resample we calculate the difference between the mean of the second group and the mean of the last n_2 observations from the entire sample.

$$\bar{X}_{(1:n_2)} - \bar{X}_{(nl+1:n)}; n_1 + n_2 = n$$

Finally, the bootstrap distribution of the difference between the groups 1 and 2 is:

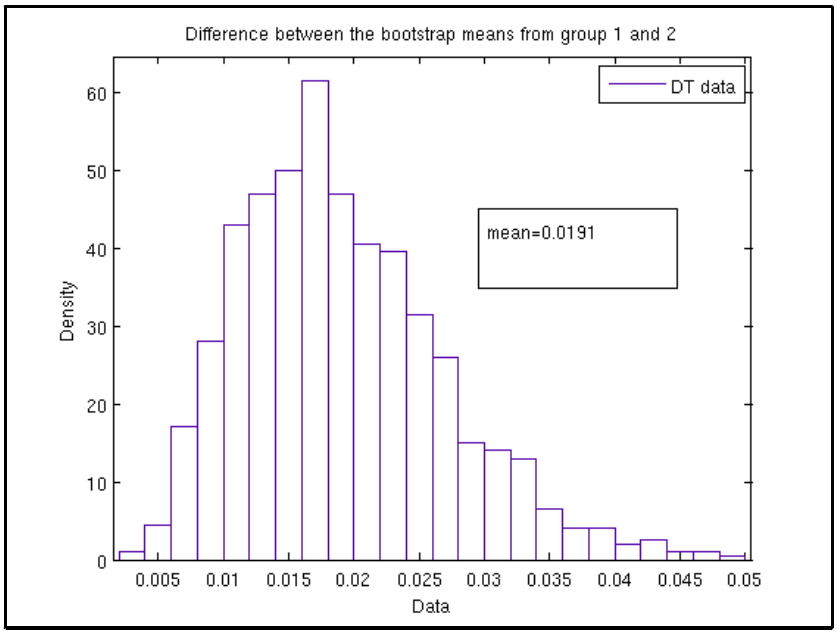


Figure 6: Distribution of the difference between bootstrap means 1 and 2.

Because of the construction of the bootstrap methodology applied here, it is impossible for this difference to be centred at zero (one group is always greater than other), but if both groups are close enough, forming part of the same split, the expectation of the difference of the bootstrap means will be small (In this case, the mean is 0.0191.)

In the case of groups 1 and 4 (the two tales of the distribution), the bootstrap distribution of the differences between the two means will be:

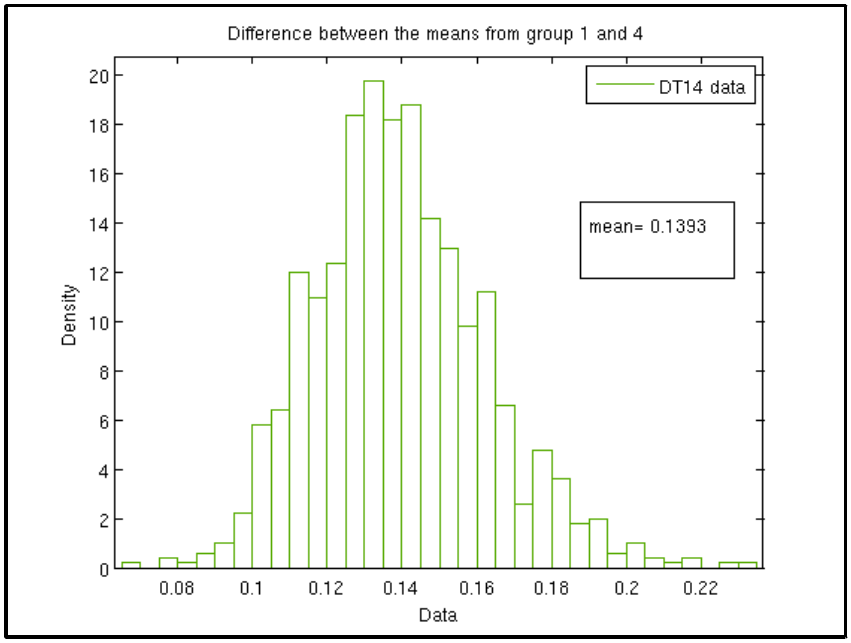


Figure 7: Distribution of the difference between bootstrap means 1 and 4.

and this mean is much bigger than the previous one.

Extending this results, we simulated 1.000 samples from a Standard Normal distribution of size n=100, and then split them into 2, 3, 4, and 5 groups each time, and then we calculate the bootstrap expectations for the mean of group 2 and the mean of second part of entire sample, with the methodology of Hutson and Ernst (2000) presented above, in order to not generate bootstrap samples and avoid the resampling error. The following table show the means and standard deviation of the 1000 samples for each splitting process:

	number of groups			
groups	**2**	**3**	**4**	**5**
1 – 2	**0.0062 (0.0018)**	**0.0069 (0.0023)**	**0.0079 (0.0028)**	**0.0091 (0.0032)**
1 – 3		0.0488 (0.0064)	0.0453 (0.0068)	0.0445 (0.0078)
2 – 3		**0.0070 (0.0023)**	**0.0063 (0.0022)**	**0.0064 (0.0024)**
1 – 4			0.0838 (0.0088)	0.0762 (0.0091)
2 – 4			0.0448 (0.0068)	0.0381 (0.0066)
3 – 4			**0.0079 (0.0026)**	**0.0063 (0.0023)**
1 – 5				0.1140 (0.0114)
2 – 5				0.0759 (0.0094)
3 – 5				0.0441 (0.0079)
4 – 5				**0.0090 (0.0033)**

Table 3: Means of the difference between bootstrap expectations of split samples. Standard deviation in parenthesis

Then, when the groups are close each other and constitute a partition we get results less than 0.01 in all cases, and this doesn't depend on what part of the distribution we are, either over the tales (like in groups 1-2, 4-5, for the 5 groups example) or in the middle (like in 2-3 for the 4 groups example). From this results it is possible to construct cutoffs which allow us to recombine previously split data set, where we don't know from what part of the distribution the split group come. Therefore, by using bootstrap methodology, large sizes of data samples are not needed, and also the exact moments proposed by Hutson and Ernst (2000) implemented here, allow us to avoid the resampling error.

However, more research is needed in order to extend this results to multivariate data sets, and although bootstrap methods are easily implemented in p>2 dimensions, is necessary to attempt different approaches like reduction of dimensionality through the use of projections (Peña and Prieto, 2001), or defining some Multivariate Linear Combination of Order Statistics (Fraiman and Meloche, 1999) .

Bayesian Clustering Hypothesis Test

A recent alternative to merge disjoint groups for multivariate data is given by Fuentes and Casella (2009). They propose a new methodology to test the hypothesis $H_0: \kappa = 1$ vs. $H_1: \kappa = 2$, where k denotes the number of clusters existing in a sample. The procedure is based on a methodology of Bayesian model selection, using the "Bayes factor" to obtain an explicit hypothesis testing for the existence of groups, obtaining the posterior probabilities for the null hypothesis, and a frequentist p-value. One advantage of this formulation is that it is not based on distance, which avoids the use of metrics to identify groups within the sample.

In order to evaluate this test, the authors focus their methodology in a bayesian approach, using the following Bayes Factor associated with the hypothesis:

$$BF_{10} = \frac{m(Y|\kappa = k)}{m(Y|\kappa = 1)}$$

where $m(Y|\kappa = k)$ denotes the distribution of the data, Y , given that there are exactly k clusters.

Considering the total number of all the possible partitions ω of n elements in k clusters, given by $S_{n,k}$, the Bayes factor can be written as:

$$BF_{10} = \sum_{\omega \in S_{n,k}} \frac{m(Y|\omega)}{m(Y|\omega_1)} \frac{\pi(\omega)}{\pi(\omega_1)}$$

where $\pi(\omega)$ denotes prior probability for the partition ω . Since the sum over the set of all possible partitions is typically large even with small numbers of observations and clusters, they estimate the value of Bayes factor through an importance sampling sum (See Fuentes and Casella, 2009 for details).

Finally, the posterior probability of H_0 is given by:

$$P(H_0|Y) = \frac{1}{1 + BF_{10}}$$

and will provide evidence against H_0 when is small.

The recombination process is proposed as follows:

- Order basic groups by size. Start with the one with largest number of observations, and calculate the Mahalanobis distance between $\bar{y}_1$ and the average of all other groups. A group whichever has the shortest distance, is then selected as candidate to recombine, i.e. the group i compliant with:

$$i = arg \min_{G \geq j \geq 1} (\bar{y}_1 - \bar{y}_j) S_1^{-1} (\bar{y}_1 - \bar{y}_j)'$$

- Check for groups 1 and i can be combined. This is done by hypothesis test proposed by Fuentes and Casella (2009). It is necessary to establish a minimum number of iterations for convergence of the Metropolis-Hastings algorithm which is based on the method. Although the authors recommend a minimum of 500,000 iterations, similar results are achieved with 25,000. On the other hand, are made at least 4 repetitions of the algorithm, obtaining the posterior probability of H0 and the corresponding p-value.
- After performing the test, 2 options are possible:

  a) If the test concludes that the two groups should not be merged, then group 1 stays the same, defined as a homogeneous group. Then it goes on to group 2, which is the second largest group, and calculate the distances between the remaining groups and group 2, and so on until the test suggests that groups should be merged.

  b) If the test concludes that the basic sets should be combined, then form a new group and the Mahalanobis distance from the other groups to this new group is calculated. The candidate group will be the closest to recombine.

  c) The process is repeated for all groups, until they can be increased by combining with others, then the algorithm stops.

Example

Using the known data set "Old Faithful Geyser" (Azzalini and Bowman, 1990), the

process of partitioning of the SAR algorithm detects 18 basic groups. Starting with the biggest one, and labeling it by “1”, the distance between this groups and the rest is calculated, labeling the closest group as “2”, the next one as “3” and so on.

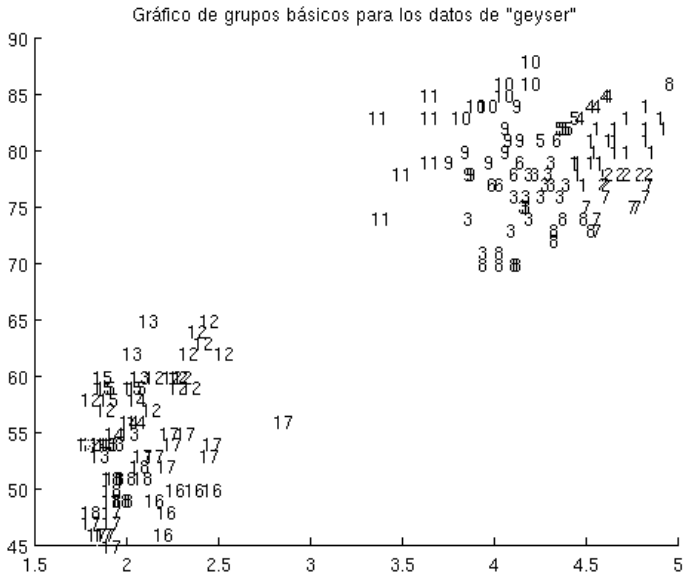


Figure 8: Basic groups obtained by SAR process from “geyser” data set

Hypothesis Test

Using the methodology of Fuentes and Casella (2009) it is possible to test whether these two groups can be recombined by the hypothesis:

$$H_0: \kappa = 1 \quad \text{vs.} \quad H_1: \kappa = 2$$

Where k represents the number of groups.

In the case of groups 1 and 2, by applying the test we obtain the following results:

```
Cluster test conducted on data object data1, with 25000 iterations.

Num. observations       : 31
Min cluster size        : 6
p                       : 2
H0                      : k = 2
```

```
**************************************
Final Empirical Posterior Probability:
**************************************
      Post.Probs
data1     0.1863

        emp.prob pvalue
data1 0.1863407 0.0709
```

The posterior probability of H_0 (There are no groups within the data), is 0.1863, with a p-value of 0.0709. Despite being a small p-value is above $\alpha = 0.05$ so do not reject the null hypothesis and the two groups can be recombined.

Nevertheless, the authors recommend to apply this algorithm performing replications, so we repeat the procedure, and this time with four replications each, for the rest of the basic groups, the results obtained are as follows:

added group	P(H0)	P-value
1		
2	0,1077	0,0610
3	0,1270	0,1866
4	0,1190	0,1877
5	0,4830	0,2860
6	0,7128	0,1240
7	0,6887	0,0950
8	0,5696	0,0600
9	0,7299	0,0830
10	0,9188	0,2230
11	0,9199	0,2030
12	**0,0000**	**0,0010**
13	0,0000	0,0010
14	0,0011	0,0010
15	0,0094	0,0090
16	0,0261	0,0210
17	0,3330	0,2050
18	0,4412	0,2830

Table 4: Testing results adding one by one the basic groups obtained by SAR from "geyser" data.

In this way, we started by testing groups 1 and 2, not rejecting $H_0 : \kappa = 1$ but with small posterior probability and p-value, then we add group 3 to groups 1 and 2, and so on. The conclusion of the tests seem to be not too robust at the beginning on

the procedure in joining groups from 1 to 11, which actually conform the superior group of the geyser data (Figure 7). But when we add group number 12, the posterior probability of H_0 and the corresponding p-value jumps to values close to zero, which strongly suggest the existence of two groups, so basic group number 12 and the previous ones should not be merged, and cluster number one (basic groups from 1 to 11) is detected.

When more basic groups are incorporated, the posterior probabilities and p-values tend to increase again, for example when we add group number 17 to the rest. This is because the algorithm only considers the existence of one versus 2 clusters inside the data, and incorporating more variability to it, maybe could suggest the existence of more than 2 groups. So, in the context of SAR basic groups, the algorithm should only be used to test $H_1 : \kappa = 2$ groups, and stop when it strongly reject H_0 (In this case, in incorporating group number 12).

Conclusions and further research.

Some approaches on recombining dependent data were presented here. The motivation to work with this type of data arises from the SAR algorithm (Peña, Rodriguez and Tiao, 2004), in which a set of observations is divided into small disjoint groups each, based on a measure of heterogeneity with the ultimate objective of detecting outliers and finally, clusters. This type of partitions obtained from the first part of the algorithm (Splitting) can not be combined by the usual equality of means and variances hypothesis tests since they are not independent.

An alternative to solve this problem is studied using linear combinations of order statistics, but given the difficulty of finding simple expressions for the moments of these statistics, we chose to implement computationally the proposal of Hutson and Ernst (2000), who present exact bootstrap moments for L-statistics, allowing to incorporate the advantages of the bootstrap methodology, and at the same time avoiding the resampling error. Finally, for the multivariate case, using the alternative proposed by Fuentes and Casella (2009) who, using MCMC methods and the use of Bayes Factor, achieve to test the null hypothesis of non-existence of groups within a

sample.

Using exact bootstrap moments for linear combinations of order statistics, the results allow to find cut-off values to discriminate whether two data sets come from the same partition, regardless of the location of the partition into the original sample. Moreover, for the multivariate case, it is possible to combine two 'basic groups' or sub-partitions, by testing hypothesis of the existence of two groups versus only one.

However, is still necessary to delve into this and other issues related to the recombination of data from a partition. In particular, in the case of the use of bootstrap methodology, although so far the results do not depend on the original distribution of data, it is possible to find more accurate cut-off values assuming normality and a certain confidence level, and other wider which hold under more general conditions. With respect to the use of Fuentes and Casella (2009) hypothesis test, will be necessary to further refine the procedure so as to find appropriate simulation parameters for the case $\kappa = 1$ vs. $\kappa = 2$ according to the size of the groups, that give more robustness to the results.

References

Agarwal, G. and Pant, R. (2008). *Moments of L-statistics: A Divided Differences Approach.* Communications in Statistics- Simulation and Computation, 37(5). 829-843.

Anderberg, M. (1973). *Cluster Analysis for Applications.* Academic Press, Inc., New York.

Azzalini, A. and Bowman, A. (1990). *A Look at Some Data on the Old Faithful Geyser* . Journal of the Royal Statistical Society. Series C (Applied Statistics), 39, 357-365.

Balakrishnan, N.; Charalambides, C.; and Papadatos, N. (2003) B*ounds on expectations*

of order statistics from a finite population. Journal of Statistical planning and Inference 213, 569-588.

Dubes, R. (1987). *How many clusters are best? - an experiment.* Journal of Pattern Recognition, 20, 645–663.

Fraiman, R. and Meloche, J. (1999). *Multivariate L-estimator. Sociedad de Estadística e Investigación Operativa Test*, 8, 255-317.

Fuentes, C. and Casella, G. (2009) *Testing for the existence of clusters,* SORT journal, 33, Forthcoming article.

Gan, G.; Ma, C. and Wu, J. (2007). Data Clustering: Theory, Algorithms and Applications. SIAM, Society for Industrial and Applied Mathematics.

Gnanadesikan, R.; Kettenring, J. and Tsao, S. (1995). *Weighting and selection of variables for cluster analysis*. Journal of Classication 12, 113-136.

Hartigan, J. (1975). *Clustering Algorithms.* Editorial Wiley, Nueva York.

Hartigan, J. and Wang, M. (1979) Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 28, pp. 100-108.

Hutson, A. and Ernst, M.D. (2000). *The exact bootstrap mean and variance of an L-estimator.* Journal of the Royal Statistical Society, Series B (Statistical Methodology) 62(1),

69-94.

Jain, Anil; Murty, M. Narashima and Flynn, Patrick. (1999). *Data Clustering: A review.* ACM Computing Surveys, 31, 264-322.

Kaluszka, M. and Okolewski, A. (2005). *Bounds for L-statistics from weakly dependent samples of random length.* Communications in Statistics-Theory and Methods 34(9).1899-1910.

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data: an Introduction to Cluster Analysis.* Editorial Wiley, Nueva York.

King, B. (1967). *Step-Wise Clustering Procedures*. Journal of the American Statistical Association, 62, 86-101.

Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate Analysis*. Academic Press, New York.

MacQueen, J. (1967). *Some methods for classification and analysis of multivariate observations*. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 281–297.

Murtagh, F. (1984). *A survey of recent advances in hierarchical clustering algorithms which use cluster centers.* The Computer Journal, 26, 354-359.

Parr, W. and Schucany, W. (1982). Jackknifing L-Statistics with Smooth Weight Functions. Journal of the American Statistical Association, 77. 629-638.

Peña, D. (2002). *Análisis de datos multivariantes.* Editorial Mc Graw Hill, Madrid.

Peña, D.; Rodriguez, J. and Tiao, G. (2004). *A general partition cluster algorithm*, Proceedings in Computational Statistics, J. Antoch (editor), Phsyca-Verlag, 371-380.

Peña, D.; Rodriguez, J. and Tiao, G. (2003). *Identifying Mixtures of Regression Equations by the SAR procedure*, Bayesian Statistics, 7, Bernardo et al. (eds), 327-347.

Peña, D. and Prieto, J. (2001). *Cluster Identification using projections.* The Journal of the American Statistical Association, 96, 1433-1445.

Raftery, A. and Dean, N. (2006). *Variable Selection for Model-Based Clustering.* Journal of the American Statistical Association, 101, 168-178.

Rychlik, T. (2004). *Optimal bounds on L-statistics based on samples drawn with replacement from finite populations.* Statistics 38(5). 391-412.

Stigler, S. (1969). Linear Functions of Order Statistics. The Annals of Mathematical Statistics, 40, 770-788.

Ward, J. (1963). Hierarchical grouping to optimize an objective function. The Journal of the American Statistical Association, 58, 236-244.