Paper SD-02

A SAS® Macro for Finding Optimal *k*-means Clustering in One Dimension with Size Constraints

Fengjiao Hu, Georgia Regents University; Robert E. Johnson, Vanderbilt University

ABSTRACT

Wang and Song (2011) proposed a *k*-means clustering algorithm in one dimension using exact dynamic programming, which guarantees optimality. Their algorithm solved the clustering problem by breaking it into smaller nested problems. The one-dimensional measure may, for example, be baseline measures related to a *before-after* study and subjects are grouped (clustered) on baseline before randomization. In this paper we extend their work by placing constraints on the cluster size, for example, each cluster must be no less than the number of study arms. A SAS macro will be presented which finds the optimal clustering given the constraint by minimizing the within cluster root mean squared error. An option that randomly allocates subjects to study arms is also included. An example will be given where a sample of primary care practices are to be allocated to treatment or control. The study measures the degree to which primary care physicians' deliver smoking cessation counseling. Prior to randomization, the practices are clustered on the baseline measure.

INTRODUCTION

A pre-post design can help control the within subject variance to achieve the maximum power of analysis, since each subject serves as its own control (Park and Johnson, 2005). The variance may be further controlled by stratification into blocks prior to randomization, where the blocks are relatively homogeneous on the baseline primary measures (Park and Johnson, 2005). Although Park and Johnson (2006) pointed out that treating the baseline as a covariable provided maximal control of the variance, with or without pair-matching, we are interested in understanding the effect of baseline clustering when considering an analysis of pre-post differences.

In order to stratify subjects into relatively homogeneous blocks, a partition of the subjects clustered on the baseline values may be formed so to minimize the root mean squared error (RMSE) of the baseline values. Since subjects within each block will be randomized to study arms, the subjects' baseline values must be clustered such that not only the minimal RMSE is achieved, but also each cluster must have at least the same number of subjects as arms of the study.

Hierarchical clustering methods, such as *k-means* (Lloyd, 1982), are often used to identify clusters. The *k-means* procedure does not always lead to the optimal clustering since the procedure employs a random path. These methods cannot easily constrain the result to have a minimum number of subjects for each cluster. Unconstrained, the solution may result in one or more clusters having less than the required number of subjects. Fusing neighboring clusters can resolve this, but may result in a non-optimal set of clusters (Park and Johnson, 2004).

In order to get an optimal set of clusters with constraints on number of subjects, a modified hierarchical clustering algorithm is developed here for identifying clusters of univariate baseline outcome data. It not only guarantees optimality, but also places the desired constraint on the minimum number of subjects in each cluster. The algorithm is implemented using SAS/IML® with the macro *ConClus* (Constrained Uni-dimensional Clustering).

Finally, we illustrate how to use this algorithm by allocating primary care practices into treatment and control groups. The practices are clustered on baseline physician delivery of smoking cessation counseling (Rothemich et. al, 2008).

METHOD

In order to cluster on baseline to achieve minimum RMSE with the constraint that every cluster size is no less than the number of study arms, we provide a new algorithm that is an improvement of the "CKmeans.1d.dp" algorithm (Wang and Song, 2011). Since, for a fix set of clusters, minimizing Euclidean sums of squares (ESS) is equivalent to minimizing RMSE when dealing with univariate data, we calculate RMSE by getting the square root of the ratio of *withinss* and degrees of freedom, where *withinss* represents the Euclidean sums of squares of within-cluster distances from each subject to its corresponding cluster mean and the degrees of freedom is the difference between total subjects and number of clusters.

Let $x_1,\ldots,x_n$ be the non-descending sorted baseline values of n subjects. We seek the cluster partition that minimizes RMSE subject to the constraint that every cluster has no less than the number of study arms r. Our method extends that developed by Wang and Song (2011) and is equivalent to their algorithm when $r=1$. Suppose we have arranged i subjects into m clusters with minimum *withinss*. We record the corresponding minimum *withinss*

in entry $D[i,m]$ of an $n \times \left\lfloor \frac{n}{r} \right\rfloor$ matrix D, where $\left\lfloor \frac{n}{r} \right\rfloor$ means the integer part of $\frac{n}{r}$. The second dimension of D is constrained since n subjects can be clustered into at most $\left\lfloor \frac{n}{r} \right\rfloor$ clusters. The last row of the matrix D indicates clustering all the subjects, thus the minimum value of the last row in matrix D corresponds to the number of clusters related to the cluster partition with the smallest *withinss* value, the solution to the original problem.

The matrix D is built dynamically. Suppose we want to find the *withinss* to place in $D[i,m]$ and the corresponding partition that leads to this value. Let j be the index of the smallest ordered data value in the last, m^{th}, cluster. The value of j cannot be less than $rm-r+1$ because it must have at least r subjects in each cluster, that is at least $r(m-1)$ totally for the first $m-1$ clusters. Further, j must be no greater than $i-r+1$ since otherwise the m^{th} cluster would not meet the constraint. Thus $r(m-1)+1 \le j \le i-r+1$. It is evident that $D[j-1,m-1]$ must be the optimal *withinss* for the first $j-1$ points in $m-1$ clusters for otherwise one would have a better solution to $D[i,m]$. This establishes the optimal substructure for dynamic programming and leads to the recurrence equation

$$D[i,m] = \min_{r(m-1)+1 \le j \le i-r+1} \{D[j-1,m-1] + d(x_j,\ldots,x_i)\}, 1 \le i \le n, 1 \le m \le \left\lfloor \frac{n}{r} \right\rfloor$$

where $d(x_j,\ldots,x_i)$ is the sum of squared distances from $x_j,\ldots,x_i$ to their mean. The matrix is initialized as $D[i,1] = d(x_1,\ldots,x_i)$, that is, *withinss* for clustering i subjects into one cluster, equals to the sum of squared distances from all i observations to their mean. Using the above recurrence, we can obtain $D[n,m]$ the minimum *withinss* if all n numbers are clustered into m groups, with minimum $RMSE[n,m] = \sqrt{\frac{D[n,m]}{n-m}}$.

In order to make the program more efficient, as suggested in Wang and Song (2011), the values $d(x_j,\ldots,x_i)$ in the recurrence $d(x_j,\ldots,x_i)$ can be computed progressively—and stored—based on $d(x_j,\ldots,x_{i-1})$. Using a general index from j to i, we iteratively compute

$$d(x_j,\ldots,x_i) = d(x_j,\ldots,x_{i-1}) + \frac{i-j}{i-j+1}(x_i - \mu_{j,i-1})^2, \text{ with } \mu_{j,i} = \frac{x_i + (i-j)\mu_{j,i-1}}{i-j+1},$$

where $\mu_{j,i}$ is the mean of $(x_j,\ldots,x_i)$. To find a clustering of data with minimum *withinss* of D, an auxiliary $n \times \left\lfloor \frac{n}{r} \right\rfloor$ matrix B is defined to record the index of the smallest number in cluster m

$$B[i,m] = \arg\min_{r(m-1)+1 \le j \le i-r+1} \{D[j-1,m-1] + d(x_j,\ldots,x_i)\}, 1 \le i \le n, 1 \le m \le \left\lfloor \frac{n}{r} \right\rfloor$$

Then we backtrack from $B[n,k]$ to obtain the starting and ending indices for all clusters and generate an optimal solution to the *k-means* problem.

EXAMPLE

Eighteen primary care practices were recruited as research sites under the auspices of the Virginia Ambulatory Care Outcomes Research Network (ACORN), a practice-based research network. Before conducting the intervention, each practice's baseline rate of providing cessation counseling by surveying a cross-sectional sample of visiting smokers was determined. The practices were then clustered according to their baseline rate. Since the study was a two-arm study with treatment and control, the clustering had the constraint that every cluster has at least two subjects. A group of 18 practices can be groups in 1 to 9 clusters as presented in Table 1. The within-cluster sum squares and RMSE are also presented. The results indicate that clustering practices into 8 clusters can achieve smallest RMSE, with 3 practices in first cluster and sixth cluster, and two practices in of the remaining clusters. Note that RMSE drops little from 5 to 8 clusters. For this study, 5 clusters were used. Also note that the 9th cluster's RMSE values are greater than those for the 8th cluster. Such a skewed-U shape in the cluster × RMSE plot is typical when cluster size is constrained (Figure 1).

			Number of Clusters								
			1	2	3	4	5	6	7	8	9
Cluster	*Withinss*	RMSE	Cluster Size								
1	0.3007	0.1330	18	.	.	.	.	.	.	.	.
2	0.0627	0.0626	6	12	.	.	.	.	.	.	.
3	0.0237	0.0398	5	5	8	.	.	.	.	.	.
4	0.0103	0.0271	5	4	7	2	.	.	.	.	.
5	0.0050	0.0195	5	2	3	6	2	.	.	.	.
6	0.0035	0.0171	5	2	2	2	5	2	.	.	.
7	0.0029	0.0161	5	2	2	2	3	2	2	.	.
8	0.0022	0.0150	3	2	2	2	2	3	2	2	.
9	0.0046	0.0226	2	2	2	2	2	2	2	2	2

Table 1: Withinss, RMSE, and Cluster Size for Optimal Partitions of the Study Baseline Data

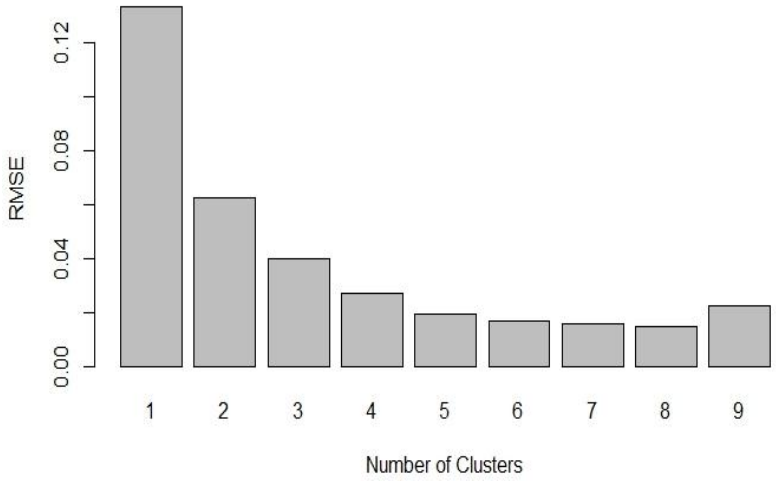

Figure 1: Example study: RMSE by the Number of Clusters.

		Number of Clusters							
		2	3	4	5	6	7	8	9
Baseline	ID	Cluster Number							
0.333	7	1	1	1	1	1	1	1	1
0.348	5	1	1	1	1	1	1	1	1
0.348	4	1	1	1	1	1	1	1	2
0.350	13	1	1	1	1	1	1	2	2
0.381	9	1	1	1	1	1	1	2	3
0.464	6	1	2	2	2	2	2	3	3
0.500	17	2	2	2	2	2	2	3	4
0.529	12	2	2	2	3	3	3	4	4
0.556	10	2	2	2	3	3	3	4	5
0.576	16	2	2	3	3	4	4	5	5
0.600	11	2	3	3	4	4	4	5	6
0.619	2	2	3	3	4	5	5	6	6
0.625	14	2	3	3	4	5	5	6	7
0.632	18	2	3	3	4	5	5	6	7
0.647	8	2	3	3	4	5	6	7	8
0.650	15	2	3	3	4	5	6	7	8
0.710	3	2	3	4	5	6	7	8	9
0.731	1	2	3	4	5	6	7	8	9

Table 2: Optimal Partitioning of 18 Primary Care Practices Into 1-9 Clusters

The details of how to assign each practice into clusters are presented in Table 2. The partitioning for a single cluster is obvious and is not presented. For example, the column of Cluster 5 shows how to cluster 18 practices into 5 clusters with the constraint that each cluster has at least 2 subjects: 5 practices with ID 7, 5, 4, 13 and 9 were

grouped into cluster 1, 2 practices with ID 6 and 17 were grouped into cluster 2, 3 practices with ID 12, 10 and 16 were grouped into cluster 3, 6 practices with ID 11, 2, 14, 18, 8 and 15 were grouped into cluster 4, and the remainder, ID 3 and 1, were grouped into cluster 5. This was the partition used in the example study.

THE MACRO

OVERVIEW

The algorithm shown above is implemented with SAS/IML® and provided as a macro, *ConClus*. The user supplies the input dataset name, variable on which clusters are desired (should be an ordinal numeric variable), variable that can help connect cluster result with original dataset (e.g., ID or post-treatment measures), and the constraint (e.g., number of study arms). Two datasets are created. *Cluster* contains the optimal cluster partitions given the constraint for all feasible number of clusters (example: Table 2). *RMSE* contains the *withinss*, RMSE, and the cluster sizes for each of the partitions (example: Table 1).

SYNTAX

%ConClus(*baseline*, <*options*>**)**

baseline
Name of numeric variable on which the cluster partition is desired. This is required and must be listed first. The variable should contain values of an ordinal measure.

Options

ID=< *id-variable*>
Name of variable used to help connect cluster result with original dataset.

DATA=<*SAS-data-set*>
Name of input data set. Defaults to the last data set, *_last_*.

CONSTR=<*r* >
Positive integer-valued constraint. Defaults to 1 and must be a divisor of the number of observations.

EXAMPLE

```
data Example;
       input Baseline ID @@;
       datalines;
0.600 11 0.522 6 0.625 14 0.731  1 0.647  8 0.333  7 0.576 16 0.348 4 0.348  5
0.529 12 0.619 2 0.710  3 0.500 17 0.632 18 0.350 13 0.650 15 0.381 9 0.563 10
;
run;

* Partition into clusters with two-arm study;
%ConClus(Baseline,id=ID,data=Example,constr=2)
```

CONCLUSION

A modified hierarchical clustering algorithm was provided as a SAS® macro for identifying clusters of univariate ordinal data that not only guarantees optimality but also places the desired constraint on the minimum number of subjects in each cluster. The macro provides two data sets: one describes how many subjects in each cluster and the minimum within-cluster sum squares and RMSE that can be achieved for each number of clusters, and the other one describes how to group the subjects into different number of clusters to achieve minimum RMSE. The algorithm presented here can only be applied to one-dimensional data. An algorithm that reaches optimality, is repeatable, and meets constraints on cluster size remains to be developed for multidimensional data. A future version of this macro will assist with randomization of subjects across study arms stratified by cluster.

REFERENCES

Lloyd, S.P. (1982), "Least Squares Quantization in PCM." *IEEE Transactions on Information Theory,* 28(2): 129-137.

Park, M and Johnson, RE (2004), "Methods for Matching Clusters on Baseline Outcome Measures Prior to Randomization." *2004 Proceedings of Journal of the American Statistical Association*, 423-426, Alexandria, VA: American Statistical Association.

Park, M and Johnson, RE (2005), "Stratification on Baseline Measure: The Variance Effect." *2005 Proceedings of Journal of the American Statistical Association*, Vol.4, Alexandria, VA: American Statistical Association.

Park, M and Johnson, RE (2006), "Design and Analysis Methods for Cluster Randomized Trials with Pair-Matching On Baseline Outcome: Reduction of Treatment Effect Variance." *2006 Proceedings of Journal of the American Statistical Association*, Alexandria, VA: American Statistical Association.

Rothemich, SF, Woolf, SH, Johnson RE et.al (2008) "Effect on Cessation Counseling of Documenting Smoking Status as a Routine Vital Sign: An ACORN Study." *Annals of Family Medicine*, Vol. 6, No.1, 60-68

Wang, H and Song, M (2011), "CKmeans.1d.dp: Optimal *k-means* Clustering in One Dimension by Dynamic Programming." *The R Journal*, Vol. 3/2, December 2011

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Name: Fengjiao Hu
Enterprise: Georgia Regents University
Address: 1120 15th Street, Department of Biostatistics
City, State ZIP: Augusta, GA, 30912
Work Phone:
Fax:
E-mail: fhu@gru.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.