Appendix to "Finding the Number of Groups in Model-Based Clustering via Constrained Likelihoods"

Andrea Cerioli, Luis A. García-Escudero,
Agustin Mayo-Iscar and Marco Riani

The materials in this document supplement the information presented in the manuscript "Finding the Number of Groups in Model-Based Clustering via Constrained Likelihoods". Section A provides the proof of Theorem 3.1 and a graphical illustration of it. Section B gives the optimal c values for each k, when using CLA$_c$-CLA and MIX$_c$-MIX, for the data set in Figure 1 and the associated "contour plot" for the three constrained clustering criteria. Section C shows the tables that have been applied to obtain the results presented in Section 5.2.1 and a graph with the first 4 discarded "spurious" solutions for the data set considered in that section. Section D provides the car-bike plot for the Hennig and Liao's type of data in Section 5.2.2. The application of the proposed methodology to the well-known "Iris data set" is given in Section E. Section F summarizes the three best ranked solutions obtained for the "road traffic data" in Section 6 by using functional boxplots. Finally, all the routines to obtain the results presented in this paper, and included in the FSDA toolbox for MATLAB, are briefly presented in Section G.

A Proof of Theorem 3.1 and graphical illustration

Proof of Theorem 3.1: In order to prove that result, let us first consider

$$B_t^* = \{(\lambda_1, ..., \lambda_D) : \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_D \leq c\lambda_1 \text{ and } 0 \leq \lambda_l \leq t\}.$$

We have

$$\begin{aligned}\text{Vol}(B_t^*) = \quad & \int_0^{t/c} \int_{\lambda_1}^{c\lambda_1} \int_{\lambda_2}^{c\lambda_1} ... \int_{\lambda_{D-1}}^{c\lambda_1} d\lambda_D d\lambda_{D-1} ... d\lambda_2 d\lambda_1 \\ & + \int_{t/c}^{t} \int_{\lambda_1}^{t} \int_{\lambda_2}^{c} ... \int_{\lambda_{D-1}}^{c} d\lambda_D d\lambda_{D-1} ... d\lambda_2 d\lambda_1.\end{aligned}$$

Given that

$$\int_{\lambda_{D-q}}^{t} ... \int_{\lambda_{D-1}}^{t} d\lambda_D d\lambda_{D-1} ... d\lambda_{D-q+1} = \frac{(t-\lambda_{D-q})^q}{q!},$$

we can see that

$$\begin{aligned} \mathrm{Vol}(B_t^*) &= \int_0^{t/c} \frac{(c\lambda_1 - \lambda_1)^{D-1}}{(D-1)!} d\lambda_1 + \int_{t/c}^{t} \frac{(b-\lambda_1)^{D-1}}{(D-1)!} d\lambda_1 \\ &= \frac{(c-1)^{D-1}(t/c)^D}{D!} + \frac{(t-t/c)^{D-1}}{D!} = \frac{t^D}{D!}\left(1-\frac{1}{c}\right)^{D-1}. \end{aligned}$$

There are $D!$ different orderings of $\lambda_1, ..., \lambda_D$ and, thus, we have (by considering obvious symmetry arguments) that

$$\mathrm{Vol}(B_t) = D! \times \mathrm{Vol}(B_t^*) = t^D \left(1-\frac{1}{c}\right)^{D-1}.$$

Thus, final result follows from the trivial fact that $\mathrm{Vol}(A_t) = t^D$. □

Figure 1 shows a graphical interpretation when $t = 1$, $D = 2$ (that is one group of two-dimensional observations) and $c = 4$. In this case $\mathrm{Vol}(A_t) = 1$ and the ratio $\mathrm{Vol}(B_t)/\mathrm{Vol}(A_t)$ equals the area of a square of side $[0, \sqrt{1-1/c}]$.

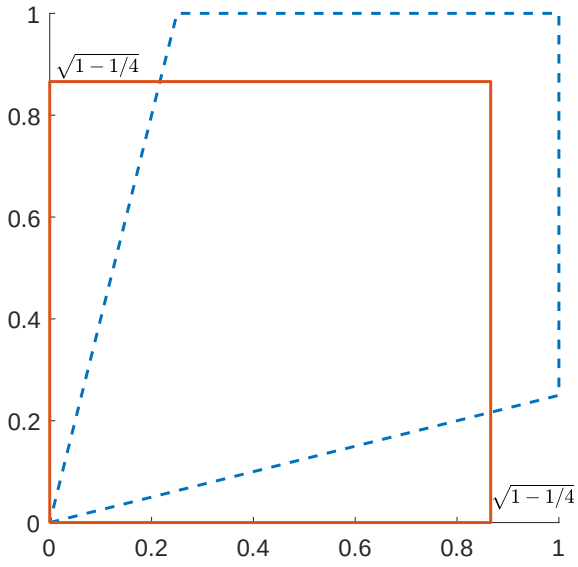

Figure 1: Illustration of Theorem 3.1 when $t = 1$, $D = 2$ and $c = 4$. The surface enclosed within dashed lines corresponds to B_1. Since $\mathrm{Vol}(A_1) = 1$, the ratio $\mathrm{Vol}(B_1)/\mathrm{Vol}(A_1)$ equals the area of the square $[0, \sqrt{1-1/4}] \times [0, \sqrt{1-1/4}]$ shown with solid lines.

Table 1: Optimal c values for each k when using CLA$_c$-CLA and MIX$_c$-MIX for the data set shown in Figure 1.

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
optimal c for CLA$_c$-CLA	4	16	8	128	1
optimal c for MIX$_c$-MIX	4	16	8	128	128

B Optimal c for each k and "contour plot" for the data set in Figure 1

The optimal c for each k is shown for the data set in Figure 1 in the first line of Table 1 when using CLA$_c$-CLA and in the second line when using MIX$_c$-MIX.

Figure 2 shows the associated contour plots that summarize the resulting monitoring process for the data set shown in Figure 1 and for our three constrained clustering criteria.

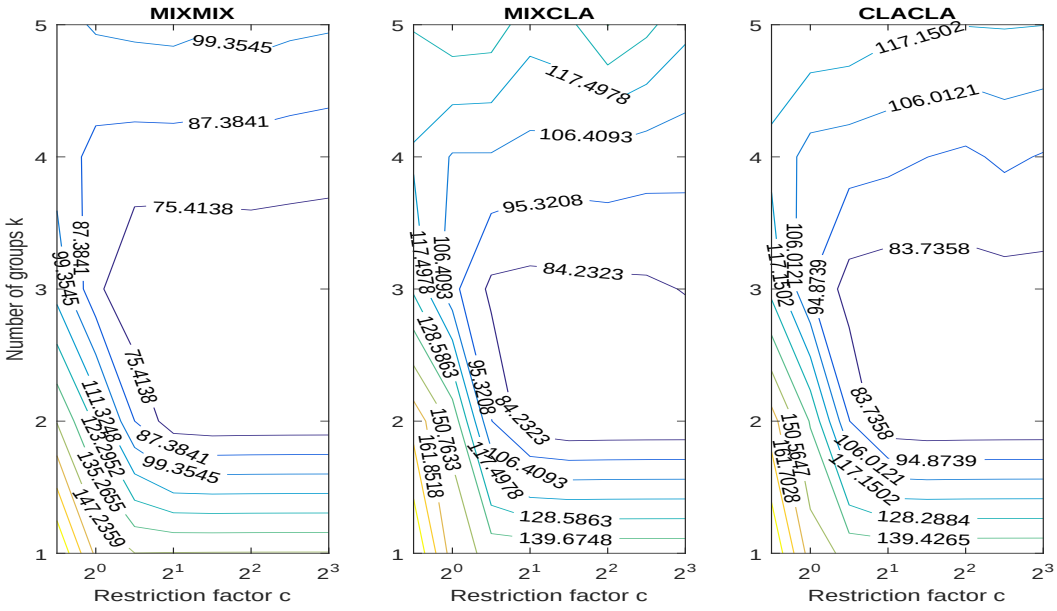

Figure 2: Contour plots for the $(k,c) \mapsto F_m(k,c)$ functions when the $m =$ MM, MC and CC criteria are applied.

C Tables for Section 5.2.1 and graph with the first 4 discarded "spurious" solutions

The starting point of the analysis done in Section 5.2.1 is the matrix which contains the values of CLA$_c$-CLA for all (k,c) pairs (given in Table 2). The

Table 2: Matrix of $K \times C$ possible of CLA$_c$-CLA (k, c) pairs to be explored.

	$c=1$	$c=2$	$c=4$	$c=8$	$c=16$	$c=32$	$c=64$	$c=128$
$k=1$	195.12	156.58	147.35	147.70	147.87	147.95	147.98	148.00
$k=2$	166.19	138.49	95.16	74.25	72.60	72.95	73.13	73.22
$k=3$	125.06	94.65	79.05	77.13	78.46	79.12	79.45	79.61
$k=4$	114.24	101.58	99.57	98.01	94.95	92.85	91.86	89.66
$k=5$	125.08	124.92	122.06	116.50	114.54	112.13	111.87	109.24

Table 3: Matrix of $K \times (C-1)$ containing the values of the ARI for two consecutive values of c (for fixed k).

$c:$	1-2	2-4	4-8	8-16	16-32	32-64	64-128
$k=1$	1	1	1	1	1	1	1
$k=2$	0.17	0.89	1	0.84	1	1	1
$k=3$	1	0.86	1	1	1	1	1
$k=4$	0.87	0.93	0.57	0.71	0.99	0.96	0.99
$k=5$	0.65	0.77	0.92	0.68	0.99	1	0.81

threshold given in equation (7) in the manuscript is obtained by considering the matrix which contains the ARI indexes for two consecutive values of c given k (see Table 3).

Figure 3 shows the first 4 discarded "spurious" solutions for the simulated data set in Section 5.2.1. We can see that these discarded solutions either include clusters made up with a few almost collinear or concentrated observations (solutions 3 and 5), or correspond to solutions close to one already detected "optimal" partition (solution 7).

D Car-bike plot for the Hennig and Liao's type of data

The car-bike plot (given in Figure 4) presents a nice summary of the solutions seen so far because it shows with a tall rectangle the first best ranked solution with 3 groups. The longest car is for the homoscedastic solution with 5 groups. The car-bike plot has the additional advantage of showing clearly that while the second best ranked solution with 4 groups is best just for a particular value of c, the homoscedastic solution is best in the interval c [1,16]. The height of the rectangle for the fourth best ranked solution is very small reflecting its low order in the ranking. The fifth best ranked solution is local and is shown as a "bike".

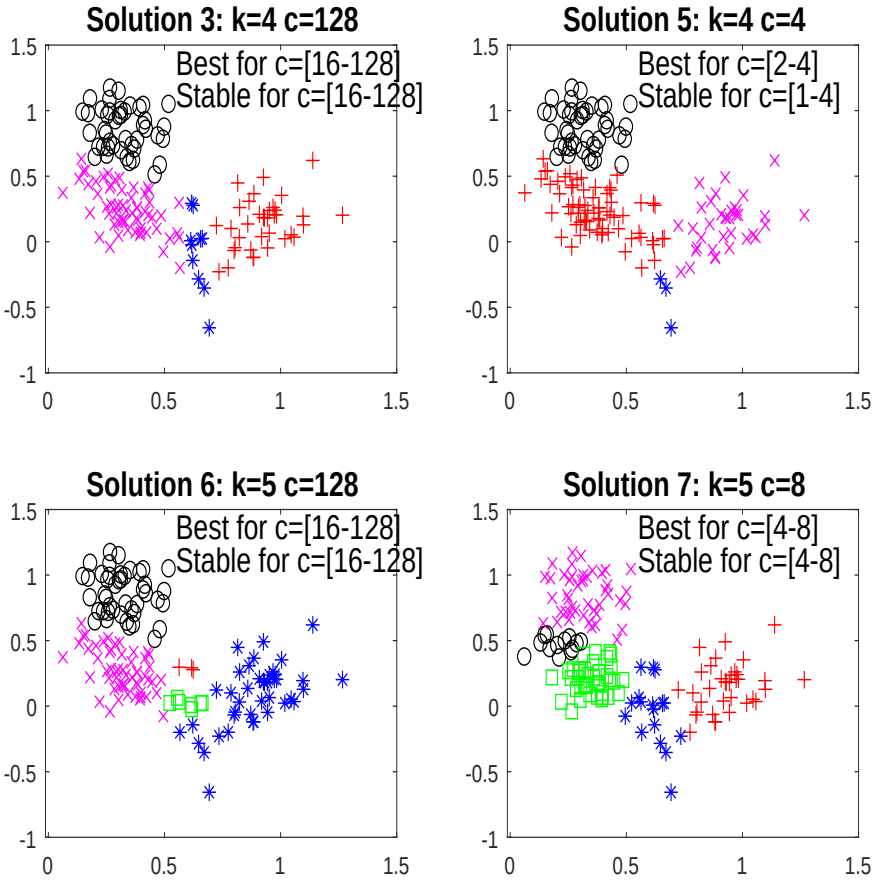

Figure 3: The first four discarded "spurious" solutions detected when using the *autCLA-CLA* procedure for the simulated data set displayed in Figure 1.

E Application to the "Iris data set"

The "Iris data set", originally collected by Anderson (1935) and first analyzed by Fisher (1936), is considered in this example. We have applied the proposed procedure to this well-known four-dimensional ($p = 4$) data set. Figure 5 shows the ranked list of "sensible" cluster partitions which are automatically found when using the *autMIXMIX* procedure. For purposes of clarity we show just the scatter plots of sepal width (SW) vs sepal length (SL), petal length (PL) vs sepal width (SW) and petal width (PW) vs petal length (PL).

We can see that the most clear two-component partition is the first offered by our method. In this partition "Iris setosa" is well-separated from "Iris virginica" and "Iris versicolor" (that are not so easy to separate). The second proposed partition essentially coincides with the three actual species.

With respect to the third best ranked solution, we recall that this "Iris data set" was initially collected by Anderson with the aim of seeing whether

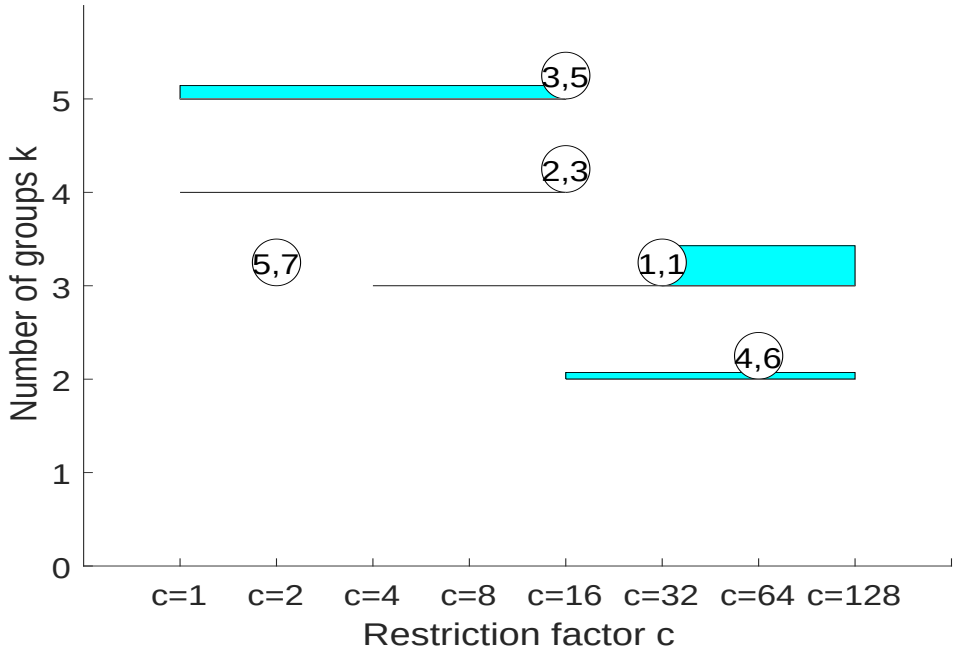

Figure 4: Car-bike plot for the when using the *autMIXMIX* for a data set similar to that in Hennig and Liao (2013).

there was "evidence of continuing evolution in any group of plants". Thus, it is interesting to evaluate whether "virginica" species should be split into two subspecies or not. In their Section 3.11, McLachlan and Peel (2000) focused only on the 50 virginica iris data and fitted a mixture of $k = 2$ normal components to them. They listed 15 possible local ML maximizers together with different quantities summarizing aspects as the separation between clusters, the size of the smallest cluster and the determinants of the scatter matrices corresponding to these solutions. After analyzing this information, the so-called "S1" solution is chosen as the most sensible one among the local ML maximizers. It is very nice to see that our third best ranked solution exactly detects a four-component partition where the "virginica" species is automatically split into 2 components in such a way that it coincides with the "S1" partition already proposed in McLachlan and Peel (2000).

F Functional boxplots for the three best ranked solutions for the "road traffic data"

Figures 6, 7 and 8 summarize the three best ranked solutions by using functional boxplots as introduced in Sun and Genton (2011) (we consider the `fbplot` function in the `fda` package with its default values; see Ramsay et al. (2014)).

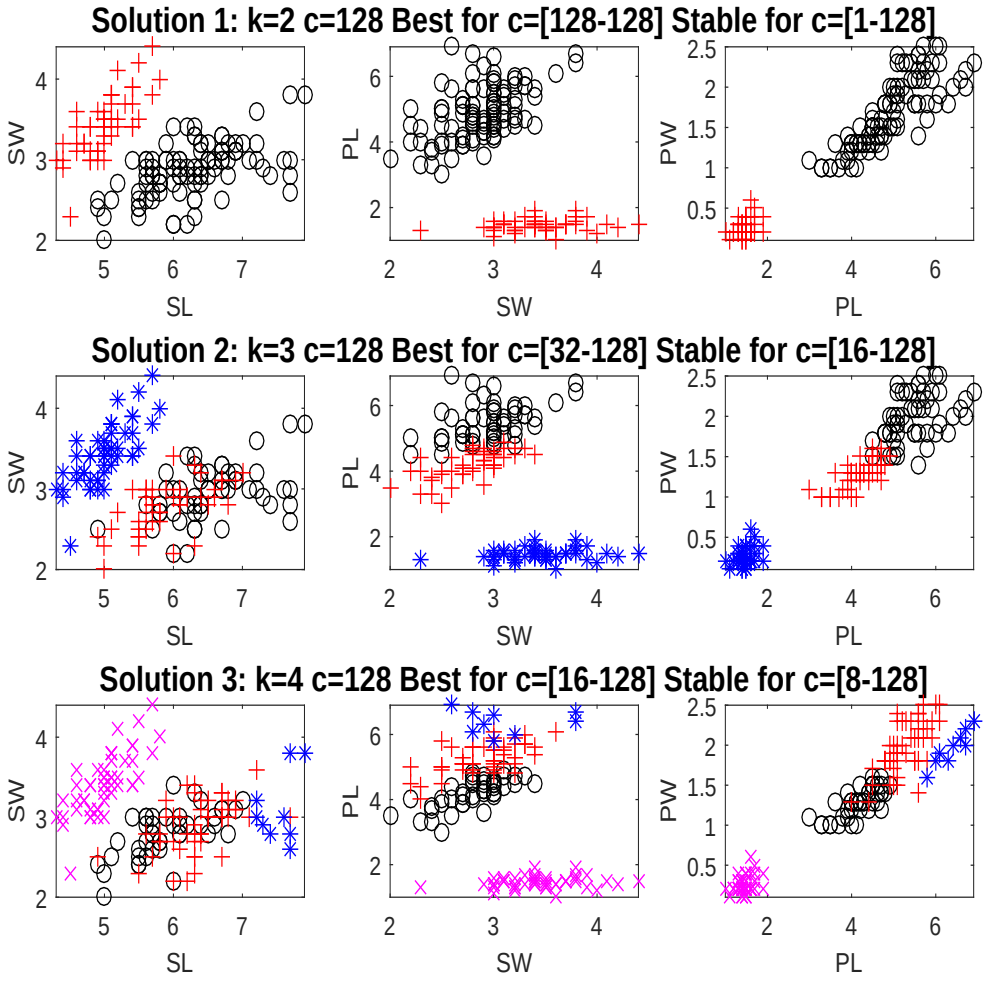

Figure 5: Best-ranked partitions when using *autMIXMIX* procedure criterion for the "Iris data set". Only some few pairs plots are shown for each cluster partition.

G Computer code

All the routines to obtain the results presented in this paper have been included in the FSDA toolbox for MATLAB which is freely downloadable from the web address

http://www.riani.it/MATLAB

or from

http://fsda.jrc.ec.europa.eu .

An explanation of the available routines is as follows:

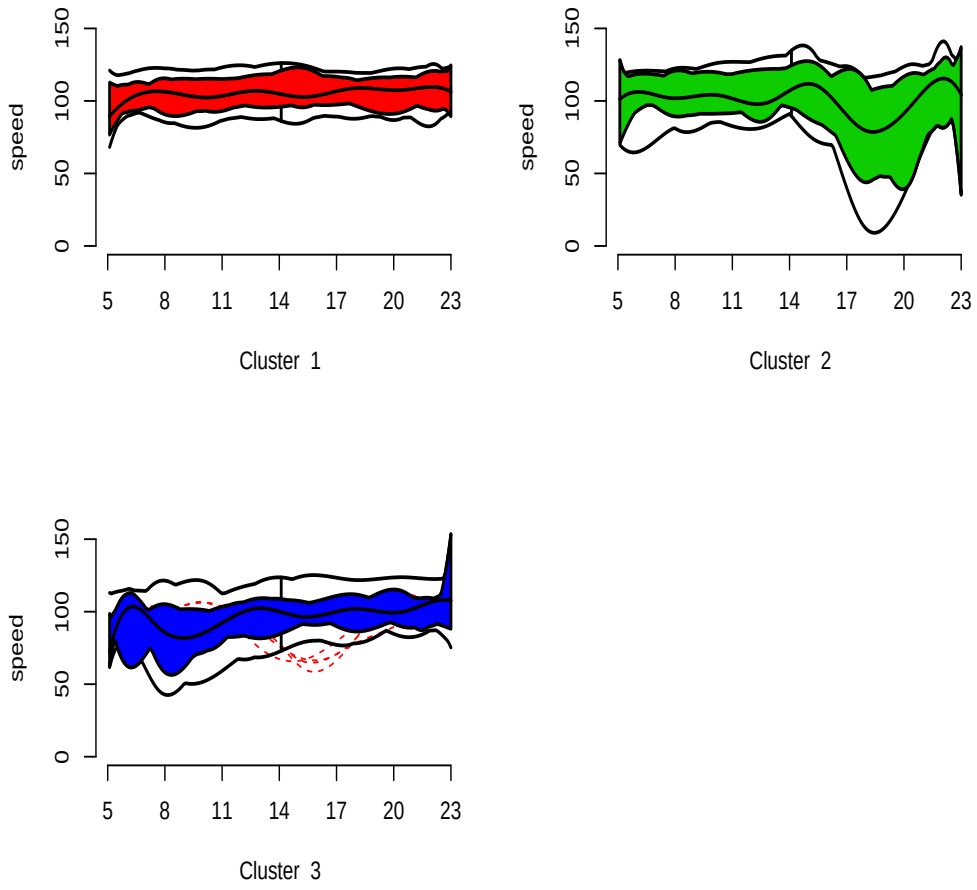

Figure 6: First of the three best-ranked partitions when using the *autMIXMIX* for the "road traffic data" with $k = 3$ and $c = 128$ represented by using functional Box-plots.

1. The routine `out=tclustIC(Y,varargin)` takes as input a data matrix containing n observations on p variables and computes the values of BIC (MIXMIX), ICL (MIXCLA) or CLA (CLACLA), for different values of k (number of groups) and different values of c (constraint factor). In `varargin` it is possible to specify the range of mixture components, the values of the constraint factor, the information criteria to use, the trimming level, the number of subsamples to extract, the number of refining iterations, the tolerance for the refining steps, the number of cores to use in parallel computing, and another series of small options. The output of this routine is a structure which contains a series of matrices which for each combination of values of k and c

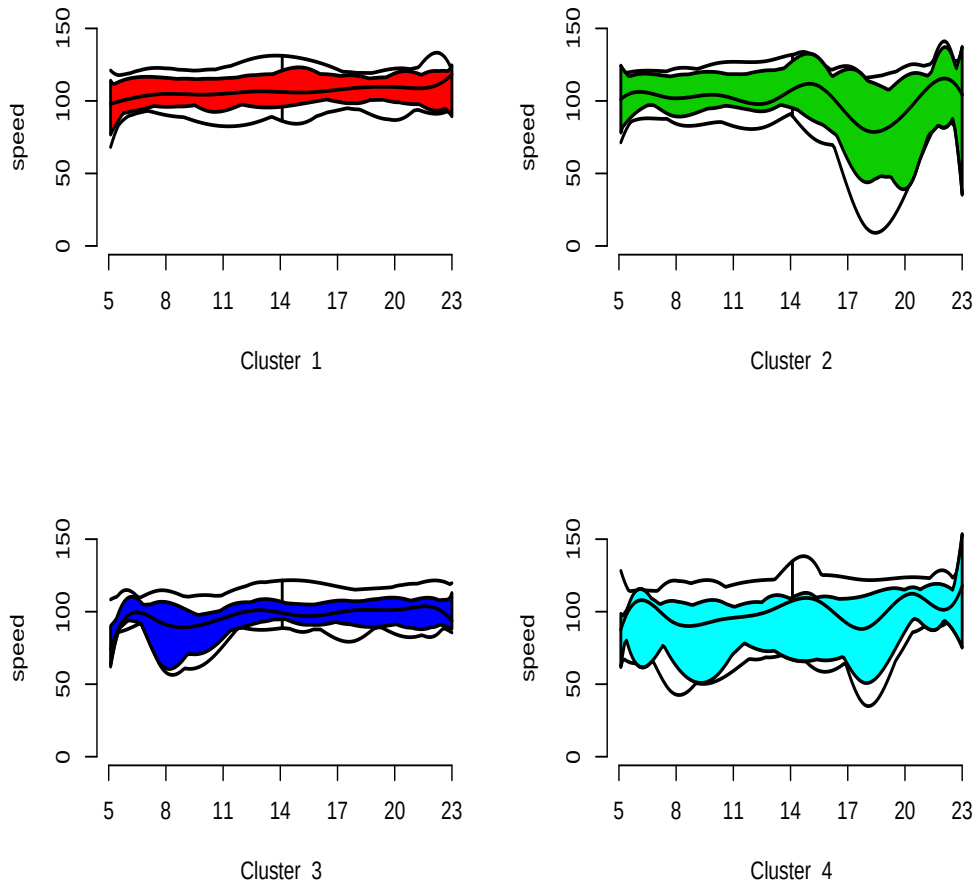

Figure 7: Second of the three best-ranked partitions when using the *autMIXMIX* for the "road traffic data" with $k = 4$ and $c = 128$.

gives the associated information criterion.

2. The routine `out = tclustICsol(IC,varargin)` takes as input the output of function `tclustIC` and extracts the first best solutions. In `varargin` it is possible to specify the information criterion to use, the number of solutions (`NumberOfBestSolutions`) to consider, the threshold to identify spurious solutions and another series of small options. The output of this routine is a structure which contains a MATLAB cell of size `NumberOfBestSolutions`-×-5 with the details of the best solutions and a matrix of adjusted Rand indexes among the best solutions associated with the requested information criteria.

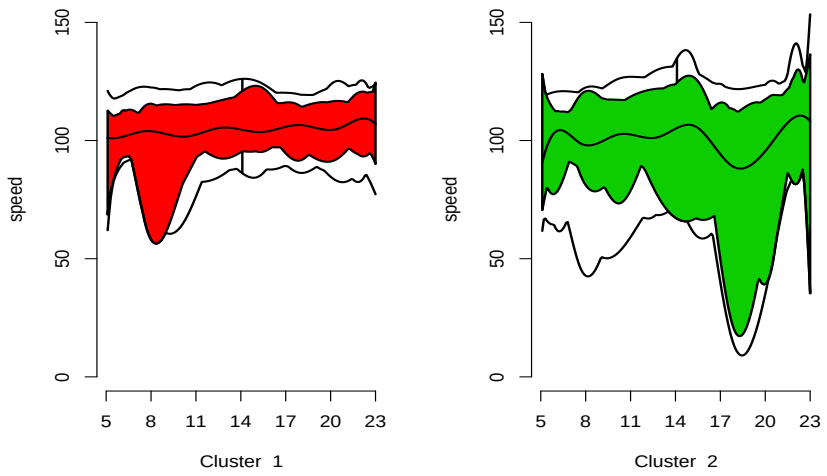

Figure 8: Third of the three best-ranked partitions when using the *autMIXMIX* for the "road traffic data" with $k = 2$ and $c = 128$.

3. The routine `tclustICplot(IC,varargin)` plots information criteria as a function of c and k. In other terms, `tclustICplot` takes as input the output of function tclustIC (that is a series of matrices which contain the values of the information criteria BIC/ICL/CLA) and plots them as a function of c or of k. Similarly to many of the other graphical routines included inside FSDA, the plot enables interaction in the sense that, if option `databrush` has been activated, it is possible to click on a point in the plot and to see the associated classification in the scatter plot matrix.

   At the end of the preamble of each .m file (and also inside the corresponding .html file) there are a series of examples containing chunks of code which can reproduce all the figures shown in the current paper.

4. The routine `carbikeplot` takes as input the output of function `tclustICsol` and enables us to create the car-bike plot.

Finally, in agreement with all the other routines present inside FSDA toolbox, the above procedures have an extensive documentation both inside the .m file and in the corresponding .html file. The help system of the FSDA toolbox is completely integrated with that of MATLAB and is almost indistinguishable from that of the official toolboxes provided by Mathworks.

References

Anderson, E. (1935). The irises of the gaspe peninsula. *Bulletin of the American Iris Society*, 59:25.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188.

Hennig, C. and Liao, T. (2013). How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification,. *J. Roy. Statist. Soc. Ser. C*, 62:309–369.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley Sons, Ltd.

Ramsay, J., Wickham, H., Graves, S., and Hooker, G. (2014). fda: Functional data analysis. *available at https://cran.r-project.org/web/packages/fda/*.

Sun, Y. and Genton, M. G. (2011). Functional boxplots,. *J. Comput. Graph. Stat.*, 20:316334.