STAT3019/G019/M019
In-Course Assessment 1 (Cluster Analysis) 2017/18, Term 2

Submission rules

The ICA has two parts. Part 1 is for group work and group submission, Part 2 is for individual work and individual submission. Every student needs to hand in their solution to Part 2 individually. Every group of students needs to submit a solution to Part 1, clearly indicating all group members. Every student needs to be either a group member or to submit Part 1 on their own. Every student can only participate in one solution to Part 1. The same ICA deadline holds for both Part 1 and Part 2. Part 1 and Part 2 have the same weight for the overall mark. See below for more information on group submissions.

General submission rules and information

- What you hand in for this assessment should be your own work (but see rules on group submission below). It is to be handed in by you to the Statistics Department Office by Thursday 8th March 2018, 4 pm.
- Write your name and student ID number on a cover sheet. To allow anonymous marking, provide your student ID number at the top of each of the answer sheets, but not your name.
- Before you hand in your work, complete and sign the slip below this rubric, cut it off and attach it firmly to your work.
- Please make sure that it is recorded on the list of students that you have handed in your work.
- Late work will not normally be accepted.
- Your submitted work needs to be printed. Handwriting can only be used for Part 2 Question 1 (a), (b), (c), and (e).
- The printed solutions for Part 1 should have at most 2000 words not including graphs, pictures and plain computer output (for which there is no length limit). This may be about four A4 pages with a letter size of 10pt and reasonable margins.

  For submissions that are longer, I will normally ignore for marking all material beyond the maximum length.

  There is no word count limit for Part 2, but I'd be very surprised if apart from computer output and mathematical formalism you'd need more than one page overall. Surely I don't expect discussion and interpretation to be extensive.
- Non-submission of in-course assessment may mean that your overall examination mark is recorded as "non-complete", i.e. you might not obtain a pass for the course.
- Any plagiarism will normally result in zero marks for all students involved, and may also mean that your overall examination mark is recorded as non-complete. Guidelines as to what constitutes plagiarism may be found in the Departmental Student Handbooks and will be provided together with this In-Course Assessment.

- A feedback sheet will be returned to you. The lecturer will keep the original, which the external examiner may wish to see, so make sure that you keep a copy of your work. You will receive a *provisional* grade – *grades are provisional until confirmed by the Statistics Examiners' Meeting in June 2018.*

Rules for group submission

The rules for the group submissions of Part 1 are as follows:

1. A single set of solutions can be handed in by groups of **up to 3 students**. This means that you can also submit your solutions either alone or as a pair or a group of three students.
2. **Students alone are responsible for forming groups.** The lecturer will neither decide about who works together, nor help with the organisation of such groups.
3. **All members of a group will get the same marks**, i.e., all will get marks for the whole set of solutions.
4. Working together and discussing solutions *within groups* is fine **for Part 1 only** and the usual plagiarism regulations do not apply to this. However, they do apply to plagiarism of work of other groups or other sources.
   Do not discuss your work with students that are not member of your group. Do not show your work to students that are not member of your group.
5. For part 2, plagiarism regulations apply to your individual work.
   Do not discuss your work on Part 2 with other students, not even those that are members of your Part 1 group. Do not show your work on Part 2 to other students, not even to those who are members of your Part 1 group.
6. All group members must be indicated (by their ID) on all pages of the submission.
7. Group solutions should be handed in as a single, stapled document.

Declaration

I am aware of the UCL Statistical Science Department's regulations on plagiarism for assessed coursework. I have read the guidelines in the student handbook and understand what constitutes plagiarism.

I hereby affirm that the work I am submitting for Part 2 of this in-course assessment is entirely my own, and that the work for Part 1 submitted in my name is entirely produced by the indicated group members.

Signed:

Please print your name:

Date:

Part 1 - Group work

Analyse the "`Dortmund_G3019ica.dat`" dataset.

Background: On the Moodle page of the course, a dataset named "`Dortmund_G3019ica.dat`" is provided. The data contains information in 30 variables regarding the 170 districts of the German city of Dortmund in the year 2002[1]. Your task is to produce a meaningful clustering of the 170 districts. Imagine that the clustering is to be used by the Dortmund city administration to characterise the districts[2].

The dataset can be read as follows:

```
dortmund <- read.table("Dortmund_G3019ica.dat",header=TRUE)
```

It should then look like this:

```
> str(dortmund)
'data.frame': 170 obs. of  30 variables:
 $ unemployed          : int  270 66 228 138 331 135 320 281 623 183 ...
 $ area_buildings      : int  106204 28592 144511 74849 307602 117753 205879 172003 225119 4
 $ births              : int  23 2 19 10 62 18 29 35 76 27 ...
 $ deaths              : int  26 57 33 39 112 26 78 42 41 9 ...
 $ moves_in            : int  317 60 312 356 431 138 204 280 462 173 ...
 $ moves_out           : int  223 52 251 241 393 109 186 241 331 105 ...
 $ cars                : int  1474 716 1370 1043 2719 894 1661 1335 1619 498 ...
 $ trucks              : int  249 450 80 86 140 40 33 51 46 41 ...
 $ motorbikes          : int  55 14 96 44 263 91 159 115 127 17 ...
 $ buildings_until_1900: int  7 7 16 0 24 1 37 19 91 60 ...
 $ buildings_1900.1918 : int  23 0 15 4 259 37 57 72 168 43 ...
 $ buildings_1919.1948 : int  33 9 17 2 130 129 174 50 30 19 ...
 $ buildings_1949.1957 : int  124 26 190 78 167 54 215 207 130 14 ...
 $ buildings_1958.1962 : int  52 18 30 40 40 19 37 22 28 5 ...
 $ buildings_1963.1972 : int  20 7 31 25 19 10 10 9 8 4 ...
 $ buildings_1973.1982 : int  17 9 15 4 19 5 0 2 13 3 ...
 $ buildings_1983.1992 : int  24 2 3 2 3 3 6 6 59 0 ...
 $ buildings_1993.2001 : int  11 0 0 2 13 3 6 2 3 0 ...
 $ households          : int  160 62 201 132 514 203 410 318 635 190 ...
 $ children            : int  237 98 281 202 756 286 574 468 1036 347 ...
 $ male                : int  682 190 692 428 1542 587 1105 884 1239 290 ...
 $ female              : int  320 75 409 203 1055 366 677 464 511 56 ...
 $ social_insurance    : int  983 217 1311 638 2539 947 1613 1491 2060 424 ...
 $ benefits            : int  249 66 138 172 158 58 217 203 706 185 ...
 $ age_under_26        : int  37 13 47 29 63 28 54 54 90 32 ...
 $ age_26.35           : int  233 39 284 150 632 207 336 349 437 120 ...
 $ age_36.45           : int  236 66 264 143 684 243 535 395 517 92 ...
 $ age_46.55           : int  217 59 227 126 570 195 297 238 288 56 ...
 $ age_56.65           : int  177 44 138 88 393 138 249 174 248 33 ...
```

[1]The dataset is extracted from a bigger dataset that was provided in the year 2003 by the administration of the German city of Dortmund to the German Classification Society "GfKl". The GfKl ran a competition about clustering these data with results presented at their annual conference in Dortmund 2004. I was in the competition jury.

[2]This had been the intention of the original competition

```
$ age_above_65        : int  102 44 141 95 255 142 311 138 170 13 ...
```

The names of the variables are mostly self explaining. Most of these are numbers, e.g., numbers of `unemployed` inhabitants, `births`, `deaths` etc., in 2002. An exception is `area_buildings`, which is the overall area of buildings (a measurement unit wasn't given in the original documentation; probably it's square meters). `moves_in`: number of people who moved into the district, analogously `moves_out`. The "buildings"-variables with year numbers give numbers of buildings by year of construction. `social_insurance`: number of inhabitants paying social insurance, `benefits`: number of inhabitants receiving benefits. The "age"-variables give numbers of inhabitants by age group. `row.names(dortmund)` shows the names of the districts.

What is expected: Produce at least two clusterings of the districts. Explain why you chose the specific clustering methods and motivate all the methodological decisions you make.

Produce at least one visualisation of each clustering.

Compare the clusterings and comment on how meaningful and useful you think they are.

Discuss potential issues with the data and the clusterings that you think are relevant and potentially informative regarding the data and your clustering.

Select the clustering that you prefer and interpret the clusters (referring to at least some of the variables). You can also use the file "Bezirke.pdf", which shows a map of Dortmund with the district numbers corresponding to row numbers in the dataset. On the map, as on most maps, the north is on top. The districts no. 1 and 2 ("City-Ost" and "City-West") are the centre of the city; the closer districts are to these, the more central they are. The interpretation of the clusters should be in a separate section and should be aimed at members of the city administration, who are interested in the clustering and know the dataset and variables, but do not have statistical knowledge.

Submit all R-code that you're using, with appropriate comments/explanations.

Hints: It is probably advisable to explore the dataset first, before actually clustering it or making data analytic decisions, by using one or more suitable visualisations, looking at value ranges etc..

You can standardise, transform, aggregate or leave out variables and obviously also compute dissimilarities as you see fit, but give justifications. Reasons for doing such things are normally that they make the representation of information by the used variables and/or dissimilarities more meaningful or useful to the city administration (e.g., making value ranges of variables comparable by standardisation or other means).

There is no single correct or best solution that I have in mind and want to see here. I'm very open to your suggestions.

Part 2 - Individual work

Submit all R-code that you're using, with appropriate comments/explanations.

1. Let $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a dataset with n objects, all from some object space $\mathcal{X}$. Let $d: \mathcal{X}^2 \mapsto \mathbb{R}_0^+$ be a dissimilarity on $\mathcal{X}$. Let $\mathcal{C} = \bigcup_{j=1}^n \mathcal{C}_j$, where $\mathcal{C}_j$, $j = 1, \ldots, n$ are partitions of $\mathcal{D}$ with $K_1 = n > \ldots > K_n = 1$, an agglomerative hierarchical clustering of $\mathcal{D}$, i.e., $\mathcal{C}$ is a hierarchy that has been obtained by the algorithm for agglomerative hierarchical clustering (AAHC) as introduced in Section 4.1 of the course notes (the notation there is also used here), and for all $C \in \mathcal{C}_j$, $j = 1, \ldots, n$: $C \subseteq \mathcal{D}$. Assume as in footnote 1 regarding the AAHC in the course notes that at every step there had been a unique pair of clusterings minimising D (as defined in the introduction of the AAHC). Assume also that the AAHC is monotonic, i.e., $0 = H_0 \leq \ldots \leq H_{K-1}$.

   The **cophenetic distance** $d_C: \mathcal{D}^2 \mapsto \mathbb{R}_0^+$ between objects is defined as follows. For $\mathbf{x}, \mathbf{y} \in \mathcal{D}$:

$$d_C(\mathbf{x}, \mathbf{y}) = \min\{H_k: \mathbf{x} \text{ and } \mathbf{y} \text{ are in the same cluster in } \mathcal{C}_{k+1}\}.$$

   In other words: $d_C(\mathbf{x}, \mathbf{y})$ is the merging level (height) at which clusters with $\mathbf{x}$ and $\mathbf{y}$ are merged in the AAHC.

   d_C can be interpreted as distance between the objects induced by the agglomerative hierarchical clustering $\mathcal{C}$, and as the "fit" provided by a hierarchical clustering to the original dissimilarity d.

   (a) Prove that d_C is a dissimilarity.

   (b) Prove that d_C is an **ultrametric**, i.e.,

$$\text{for } \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{D}: \; d_C(\mathbf{x}, \mathbf{z}) \leq \max[d_C(\mathbf{x}, \mathbf{y}), d_C(\mathbf{y}, \mathbf{z})]. \tag{1}$$

   (c) Prove that every ultrametric, i.e., every dissimilarity fulfilling (1), fulfills the triangle inequality.

   (d) A hierarchical clustering could be seen as "good" in the sense of representing the original dissimilarity d, if d_C is a good approximation of the original dissimilarity d. This is often measured by the **cophenetic correlation** c^*, defined as follows:

$$c^* = \frac{\sum_{1 \leq i < j \leq n}(d(\mathbf{x}_i, \mathbf{x}_j) - \bar{d})(d_C(\mathbf{x}_i, \mathbf{x}_j) - \bar{d}_C)}{\sqrt{\sum_{1 \leq i < j \leq n}(d(\mathbf{x}_i, \mathbf{x}_j) - \bar{d})^2 \sum_{1 \leq i < j \leq n}(d_C(\mathbf{x}_i, \mathbf{x}_j) - \bar{d}_C)^2}}.$$

   Here, "$\sum_{1 \leq i < j \leq n}$" is a sum over all pairs i, j so that $1 \leq i < j \leq n$.
   $\bar{d} = \frac{1}{(n-1)(n-2)/2} \sum_{1 \leq i < j \leq n} d(\mathbf{x}_i, \mathbf{x}_j)$ and $\bar{d}_C = \frac{1}{(n-1)(n-2)/2} \sum_{1 \leq i < j \leq n} d_C(\mathbf{x}_i, \mathbf{x}_j)$ are the means of the dissimilarities d, d_C. This means that c^* is the correlation coefficient between the vector of all $d(\mathbf{x}_i, \mathbf{x}_j)$ and the vector of all $d_C(\mathbf{x}_i, \mathbf{x}_j)$. Note that for an output object of the R-function `hclust`, the cophenetic distance d_C can be computed by the function `cophenetic`. c^* can be computed as (Pearson) correlation between the output of `cophenetic` and the dissimilarity d stored as `dist`-object.

   Consider the Olive Oil data as introduced in the course notes and provided on Moodle.

   - If the last number of your student ID is 0, 1, 2 or 3, use the Euclidean distance based on standardised variables as d.
   - If the last number of your student ID is 4, 5, or 6, use the Manhattan (L_1) distance based on standardised variables as d.

- If the last number of your student ID is 7, 8, or 9, use the Manhattan (L_1) distance based on unstandardised variables as d.

Based on the distance d, compute the Single Linkage, Complete Linkage and Average Linkage clustering, compute the corresponding cophenetic distances and correlations, and compare the three hierarchical clusterings based on their cophenetic correlations.

(e) **For BSc students only:** By cutting the dendrograms, obtain the partitions from all the hierarchical clusterings computed in part (d) with $K = 9$ clusters. Compute the adjusted Rand indexes to compare all these partitions with the 9 regions given in the dataset. Compare the "quality ranking" that you get from this with the one obtained from cophenetic correlations in part (d).

For MSc/MSci students only: For a general dataset $\mathcal{D}$ with dissimilarity d, let d_C^{single} and $d_C^{complete}$ be the cophenetic distances belonging to the Single Linkage and Complete Linkage hierarchical clustering. Decide which of the following statements is true in general, and prove it:

i. For all $\mathbf{x}, \mathbf{y} \in \mathcal{D}$: $d_C^{single}(\mathbf{x}, \mathbf{y}) \geq d_C^{complete}(\mathbf{x}, \mathbf{y})$.

ii. For all $\mathbf{x}, \mathbf{y} \in \mathcal{D}$: $d_C^{single}(\mathbf{x}, \mathbf{y}) \leq d_C^{complete}(\mathbf{x}, \mathbf{y})$.

2. (a) Run a simulation study based on data generated from the probability model used for generating Artificial Dataset 2 in the course notes in the following way.

i. Generate 20 datasets from this model. This can be done by function `art2` in the R-file on Moodle with code for the gap statistic simulation, discussed in the problem class on 25 January (generally you may want to use this file for "inspiration"; the `cvec`-vector defined in that file defines the "true" clusters for the sake of this study).

ii. Treat the number of clusters K as unknown and to be estimated, with $2 \leq K \leq 10$.

- For each of the 20 datasets, compute partitions by K-means, Single, Average and Complete Linkage and Partitioning Around Medoids. For all methods you can use the R-default settings[3].
- For all clustering methods, estimate K by the Average Silhouette Width. Note that this means that you need to compute partitions by these methods for all $2 \leq K \leq 10$, and select the one that gives the highest Average Silhouette Width.
- The dissimilarity-based methods Single, Average and Complete Linkage, Partitioning Around Medoids and the Average Silhouette Width (the latter even when applied with K-means) need to be based on a dissimilarity measure d.
  - If the last number of your student ID is 0, 3, 6 or 9, use the Euclidean distance based on standardised variables as d.
  - If the last number of your student ID is 1, 4, or 7, use the Manhattan (L_1) distance based on standardised variables as d.
  - If the last number of your student ID is 2, 5, or 8, use the Manhattan (L_1) distance based on unstandardised variables as d.

[3]The "default settings" are the pre-specified choices of input parameters of the R-functions, i.e., the parameter values that are used if a parameter is not specified by the user. If you want, you can change one or more of these, such as using `nstart=50` for K-means.

- For all methods, store the estimated value of K and the adjusted Rand index comparing the clusterings provided by the methods with the true clustering.

iii. Compare the methods by looking at the frequency of estimating $K = 3$, and by the average adjusted Rand index value over the 20 datasets.

Hints: The function `which.max` extracts the position of a maximum from a vector, i.e., if you have a vector of Average Silhouette Width values for $K = 1, \ldots, 10$ (you may set the value for $K = 1$ equal to -1 so that K cannot be estimated as 1), `which.max` will find you the estimate of K.

Keep in mind that results are for one specific way of generating the data and defining the "truth". They will not necessarily generalise to other kinds of data (in fact, what you find here may deviate quite a bit from what can be observed in many other situations).

If you have difficulties (e.g., you get some errors from running the R-code that you don't manage to remove, or you don't understand how to do a part of what's requested), submit what you have anyway. Marks are given for anything that you do that still makes sense. Don't ask other students to help you - that'd amount to plagiarism!

(b) Run the simulation study from part (a) (i.e., the same way of generating the data as in part (i) above, the same range of $2 \leq K \leq 10$, and the same evaluation, part (iii) above) with clustering by Gaussian mixtures using the R-function `Mclust`. Use the default settings to estimate K and the covariance matrix model by the BIC.

Hints: If you want, you can run parts (a) and (b) together in the same study. Note though that the `Mclust`-function can compute clusterings for all $2 \leq K \leq 10$ internally, including estimation of K, so this should not be handled in exactly the same way as the methods in part (a).

Marking scheme

This ICA (both parts tigether) counts 50% toward your overall mark for this module.

Part 1: 50 marks, assigned as follows:

- Correct application and presentation of the results of clustering and dissimilarity methodology: 10 marks
- Suitable visualisation: 6 marks
- Motivation of the choice of the clustering methods: 4 marks
- Motivation of further choices such as dissimilarity measure and number of clusters: 6 marks
- Comparison of clusterings (this refers to both a formal comparison and the interpretation of the result of such a comparison): 4 marks
- Interpretation of clustering for members of city administration: 10 marks
- General understanding and insight (which may potentially include good own ideas, relevant discussion and observations about the data): 10 marks

Part 2:

Question 1(a): 4 marks.

Question 1(b): 6 marks.

Question 1(c): 2 marks.

Question 1(d): 9 marks (correct running of methods 5, cophenetic correlations and discussion 4).

Question 1(e): 4 marks.

Question 2(a): 19 marks (correct running of clustering methods 5, correct simulation setup 6, presentation of results and discussion 5, quality of implementation 3).

Question 2(b): 6 marks (correct running of `Mclust` 2, correct simulation setup 2, presentation of results and discussion 2).

If in Questions 1(d) and 2 the final results are wrong or not existing, some marks can still be gained for submitted and explained code, even if this doesn't work properly.

The 25 marks assigned to Question 2 part (a) and (b) are handled flexibly. If parts (a) and (b) are run together, 25 marks can be earned for both parts combined, regardless of which marks can be clearly assigned to part (a) or (b). Also the marking scheme above for part (b) assumes that some work and ideas from part (a) are used; in case that part (b) is of better quality than part (a), I may assign some or all of the 3 marks on "quality of implementation" currently listed in part (a) to part (b).