Identification of patient classes in low back pain data using crisp and fuzzy clustering methods

Alexandre Gondeau and Vladimir Makarenkov

Abstract We performed a cluster analysis of the low back pain dataset in the framework of the IFCS-2017 data challenge. Because the original data contained missing values, the first part of our analysis concerned the imputation of missing values using the Fully Conditional Specification model. The Local Outlier Factor method was then used to detect and eliminate the outliers. After the data normalization, we removed highly correlated variables from the transformed dataset and carried out k-means clustering of the remaining variables based on their correlations, i.e., the variables with the highest mutual correlations were assigned to the same cluster. Once the variables were assigned to different clusters, one representative per cluster, i.e., the variable with the highest contribution score at the first principal component, was selected. Among the 13 selected variables, there are representatives of each of the 6 variable domains (contextual factor, participation, pain, psychological, activity and physical impairment), specified as important in the paper by Nielsen et al. (2016). Different clustering methods, including DAPC, k-means and k-medoids,

Alexandre Gondeau
Département d'informatique, Université du Québec à Montréal, Montreal, QC, H3C 3P8 Canada,
✉ gondeau.alexandre@courrier.uqam.ca

Vladimir Makarenkov
Département d'informatique, Université du Québec à Montréal, Montreal, QC, H3C 3P8 Canada
✉ makarenkov.vladimir@uqam.ca, corresponding author

Archives of Data Science, Series B
(Online First)
KIT Scientific Publishing
Vol. 1, No. 1, 2019

DOI 10.5445/KSP/1000085952/06
ISSN 2510-0564

were then carried out to cluster the reduced low back pain data. Consensus solutions, both crisp and fuzzy, were calculated using the GV3 method. The obtained crisp consensus clustering, including 5 classes, was described in detail and compared to the meta-data annotation.

1 Introduction

This paper presents the main steps and results of our analysis of the low back pain dataset originally described by Nielsen et al. (2016). The dataset provided by the organizers of the IFCS-2017 data challenge, containing the measurements on 928 objects (patients with low back pain) in rows and 121 variables in columns (plus an additional id variable in the first column), has been analysed.

First, we present our general data processing protocol that shows the main steps of our analysis. We then detail the results of our analysis for each of the main steps being performed:

- Imputation of missing values,
- Outlier elimination,
- Data normalization,
- Elimination of correlated variables,
- Variable selection using variable clustering and Principal Component Analysis (PCA),
- Data clustering using different partitioning algorithms,
- Computing a consensus solution for the best clusterings,
- Description of the obtained consensus clustering and the selected variables with respect to metadata.

2 Imputation of missing values

The original dataset: (928 patients recorded on 112 variables) was considered at this step. In total, 4865 values (4.68 % of the total number of data) in the original dataset were missing.

First, we used the *Mice* (Multivariate Imputation by Chained Equations) R package to impute missing values present in the original data. This package includes methods which allow for Multiple Imputation using Fully Conditional Specification (FCS), as described by Van Buuren and Groothuis-Oudshoorn (2011). The main advantage of this method is that each variable has its own imputation model. The imputation models available in *Mice* are provided for continuous data (predictive mean matching, normal), binary data (logistic regression), unordered categorical data (polytomous logistic regression) and ordered categorical data (proportional odds).

In total, 20 different (complete) datasets were imputed by *Mice*. The NRMSE (Normalized Root Mean Square Error) index (from the *hydroGOF* R package by Zambrano-Bigiarini 2014) was used to assess the imputation quality of each of the 20 datasets generated by *Mice*. Normalized Root Mean Square Error is computed between the vectors of estimated values and observed values representing the variables (see the documentation for the *hydroGOF* package for the exact formula). Two criteria were used: standard deviation of observations and the difference between the maximum and minimum of the observed values (max-min). Figure 1 shows the distribution of the NRMSE scores obtained for these 20 datasets. Two criteria were used: standard deviation of observations and the difference between the maximum and minimum of the observed values (max-min). According to both criteria, Dataset 16, providing the minimum of both curves presented in Fig.1, was selected as optimal.

Figure 2 shows examples of density maps for 10 original variables of the low back pain dataset obtained after the data imputation performed by *Mice*. The density of the original (incomplete) data is shown in black and the density of the imputed (complete) data is shown in yellow. Very close density curves can be observed for most of the variables. The *Stats* package of R was used to plot these density maps.

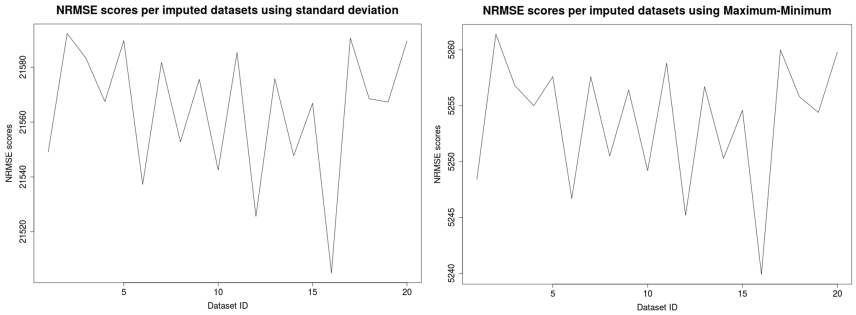

Figure 1: Distribution of the NRMSE scores for each of the 20 imputed datasets. Two criteria were used: standard deviation (left) and maximum-minimum (max-min) (right).

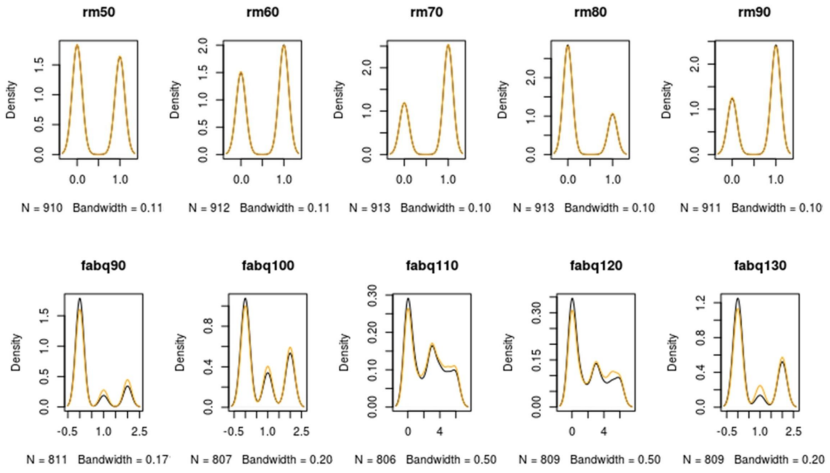

Figure 2: Density maps for 10 original variables in the selected dataset obtained after the variable imputation by *Mice*. The density of the original data is shown in black and the density of the complete data is shown in yellow.

3 Outlier elimination

At this step, the R implementation of the Local Outlier Factor (LOF) method available in the *Rlof* package (Hu et al, 2011) was used to detect the outliers. This method finds the local outlier factor (Breunig et al, 2000) of a given data matrix using the k neighbours method. The local outlier factor (LOF) is a measure of outlyingness, which was calculated for each object (i.e., patient) considered. The LOF method takes into account the density of the neighbourhood around the object to determine its outlierness. We used the Euclidean distance, which is the default option in this package.

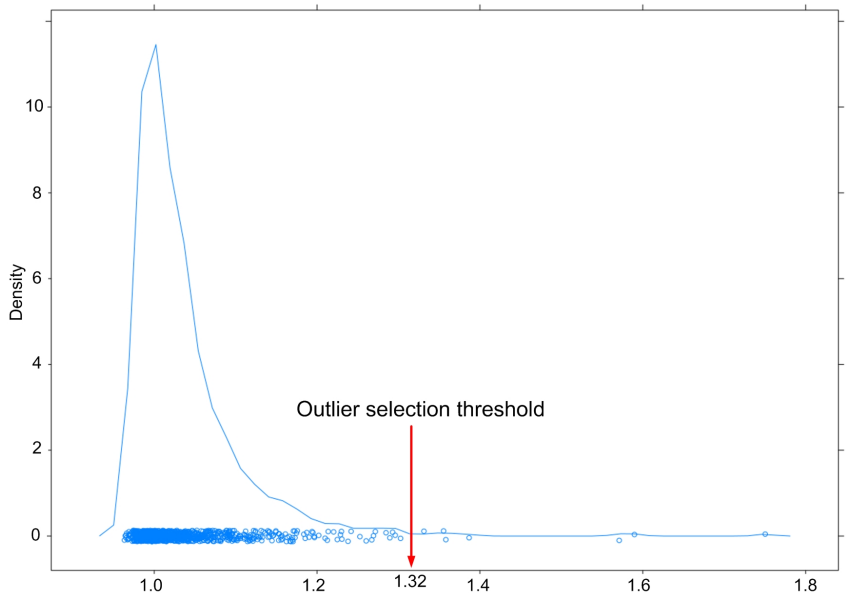

Figure 3: Density map for the LOF (Local Outlier Factor) scores obtained for the 928 objects (i.e., patients) of the low back pain dataset. The x-axis represents the LOF scores. The threshold of $Mean + 4SD$ was used to identify outliers.

The density scores for the 928 objects (i.e., patients) provided by LOF are shown in Fig. 3. We decided to use the threshold of $Mean + 4SD$ to identify outliers in our dataset. In fact, the threshold $Mean + 4SD$ is located within a larger gap than in the case of other commonly used outlier selection thresholds (e.g., $Mean + 3SD$ or $Mean + 2SD$). In total, 7 objects (0.8 % of all objects) have been classified as outliers and then removed from the dataset. The meta-data

description indicates that a few objects can be classified as outliers (a column specifying the outliers has been included in our class membership file available at: `http://www.info2.uqam.ca/ makarenkov_v/GM_IFCS2017_data.zip`). Thus, the data matrix under consideration was reduced to the size (921x112).

4 Data normalization

The normalization of data was achieved by using the traditional Z-score normalization method. The *scale* function of the base package of R was used for this purpose. The low back pain dataset contains binary, categorical and continuous variables, making the application of the robust Z-scores (Malo et al, 2005), as well as of the Z2 and Z3 normalizations proposed by Steinley (2004), impossible in this case.

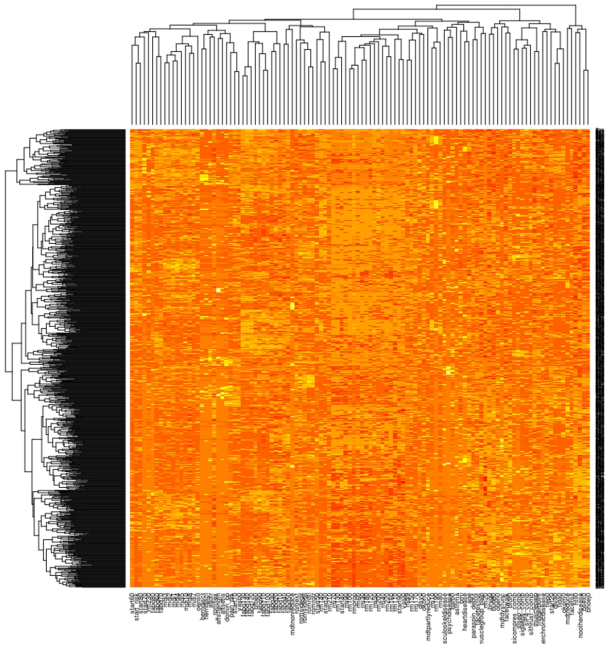

Figure 4: Heatmap of the standardized low back pain data containing 921 objects and 112 variables.

These methods were shown to outperform the traditional Z-score normalization, but cannot be applied in our case due to the presence of binary variables. Z-score normalization may not be always recommended for standardizing continuous variables (Stevens, 1946). Some authors argue however that the use of parametric statistics for categorical data may be allowed in certain cases in order to take advantage of a larger range of available statistical procedures (Cohen et al, 1996; Van Belle, 2011).

Figure 4 shows the heat map of the low back pain dataset (including 921 patients and 112 variables) obtained after Z-score normalization.

5 Elimination of correlated variables

At this step, we used the *findCorrelation* function of the *Caret* R package (Kuhn, 2015) to remove highly correlated variables from the low back pain dataset. The Spearman correlation was used to measure the degree of redundancy between the original variables. The Spearman correlation threshold of 0.4 was chosen, based on the correlation matrix heat maps before (Fig. 5a) and after (Fig. 5b) the application of the variable elimination procedure. The *findCorrelation* function searches through a correlation matrix and returns a vector of integers corresponding to variables to remove to decrease pairwise correlations. The absolute values of pairwise correlations are considered in *findCorrelation*. If two variables have a high correlation, *findCorrelation* calculates the mean absolute correlation of each variable and removes the variable with the largest mean absolute correlation. In total, 38 variables (34 % of the 112 original variables) were removed at this step. Thus, our dataset was reduced to the size of (921x74).

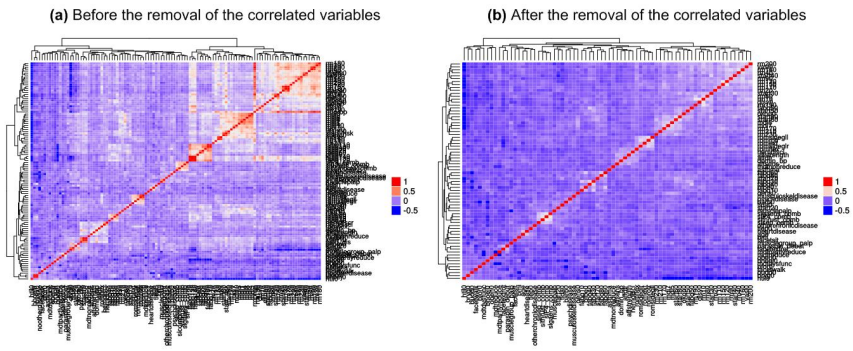

Figure 5: (a) Heat map of the correlation matrix of the 112 original variables. **(b)** Heat map of the correlation matrix after the removal of highly correlated variables.

6 Variable selection using variable clustering and Principal Component Analysis

At this step, we applied clustering to the remaining 74 variables of the low back pain dataset. The clustering of the variables was achieved by using the package *ClustOfVar* (Chavent et al, 2011) of R.

Here, we followed the recommendations presented in the well-known paper by Mitra et al (2002) concerning clustering of variables. We first used the function *hclustvar*, available in *ClustOfVar*, to create the hierarchy of the variables and then computed the bootstrap scores (i.e., stability indices) for different numbers of the variables partitions to assess their stability (Lord et al, 2017). Figure 6 shows the stability of the variables partitions with respect to the number of clusters (i.e., groups of variables) in terms of the mean score measure.

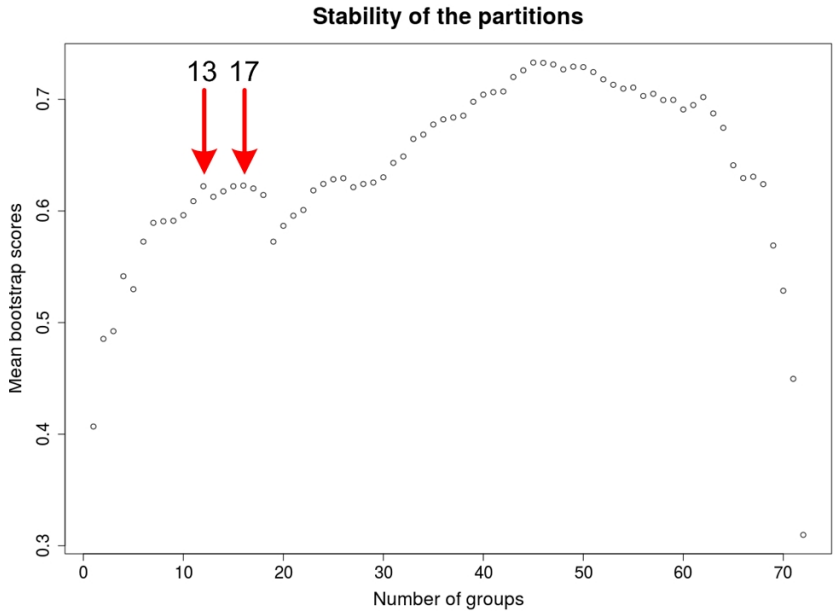

Figure 6: Stability of the variables partitions for the reduced low back pain dataset with respect to the number of clusters (i.e., groups of variables) in terms of the mean score measure. The first two local maxima found here correspond to clusters with 13 and 17 variables, respectively.

Using this graphical representation, we identified the first two peaks of the stability curve (i.e., partitions with 13 and 17 clusters, respectively). For both selected numbers of clusters (13 and 17), we first used the function *cuttrevar* to assign variables to initial clusters, and then the function *kmeansvar* with this initial partition to produce the final assignment of variables to different clusters. The *kmeansvar* function based on the popular k-means algorithm allowed us to assign the variables to different clusters based on their correlations (i.e., the variables with the highest mutual correlations were assigned to the same cluster). The squared loadings from the first principal component of the principal component analysis, performed variable-cluster-wise, were used. Once the variables were assigned to different groups, one representative per group (i.e., the variable with the highest contribution score to the first principal component) was selected.

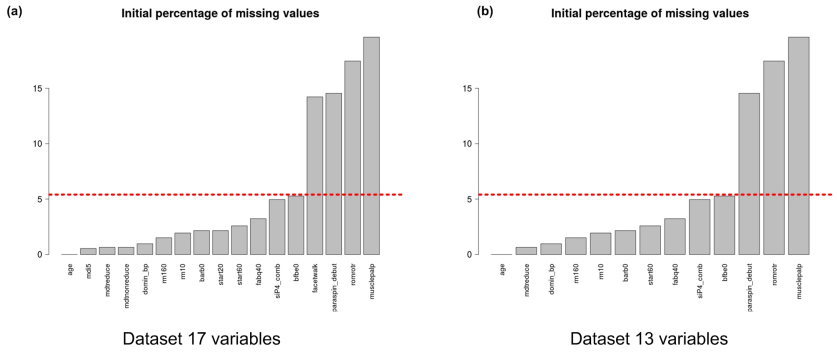

Figure 7: **(a)** Initial percentage of missing values for the variables included in the cluster with 17 variables; **(b)** Initial percentage of missing values for the variables included in the cluster with 13 variables.

Finally, we removed from the two selected datasets (i.e., with 13 and 17 variables, respectively) the variables with the highest *initial proportion* (i.e., calculated before the variable imputation step - see section 2) of missing values (i.e., the variables with the highest probability to introduce error) with a threshold of 10 %, which can be selected intuitively by observing the histograms in Fig. 7. The application of this threshold reduced the cluster of 13 variables to 10 variables, and the cluster of 17 variables to 13 variables. Thus, two reduced

datasets of sizes (921x10) and (921x13) were obtained. They were clustered using different partitioning algorithms as described in the next section.

7 Data clustering using different partitioning algorithms

At this step, we carried out a number of different partitioning algorithms based on various classification criteria in order to determine the optimal number of clusters in the low back pain dataset. Namely, the Discriminant Analysis of Principal Components (DAPC method by Jombart et al 2010, implemented in *find.clusters* function of the R package *Adegenet*), the traditional k-means (MacQueen et al, 1967) and k-medoids (Kaufman and Rousseeuw, 1987) algorithms as well as the Clara algorithm (Kaufman and Rousseeuw, 1990) were carried out.

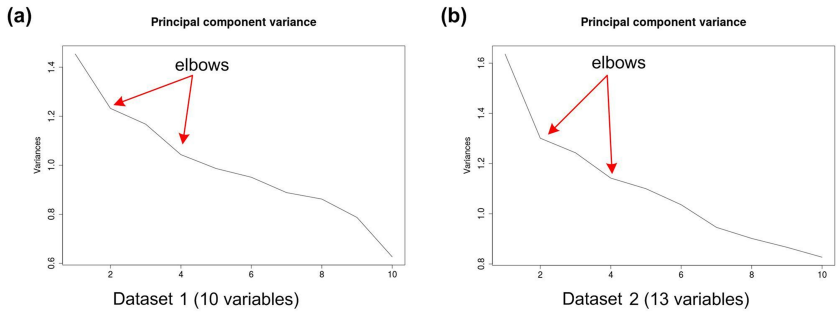

Figure 8: (a) Variation of variances for different numbers of principal components for the variant of the low back pain dataset with 10 variables; **(b)** Variation of variances for different numbers of principal components for the variant of the low back pain dataset with 13 variables. The first two "elbows" (i.e., local minima) of the two curves are indicated by the red arrows.

The clustering procedure used in Discriminant Analysis of Principal Components consists in successive runs of k-means with an increasing number of clusters after transforming data using a principal component analysis (PCA). In our analysis, the BIC index was used to measure the goodness of fit, and thus for selecting the optimal number of clusters in DAPC. In order to choose the optimal number of principal components to use as a parameter of the DAPC method,

we have drawn the variance curves with respect to the number of principal components. Using the "elbow" technique, we found that the first two "elbows" (i.e., the first two local minima of the variance function) occurred at 2 and 4 principal components, respectively, in both datasets (see Fig. 8a and Fig. 8b). Hence, the DAPC method was performed with 2 and 4 principal components in both cases.

Both the k-means and k-medoids algorithms were carried out using the Calinski-Harabasz (CH) and Silhouette (SI) cluster validity indices in order to determine the optimal number of clusters. These indices were among the top performers according to a number of comparative classification studies (e.g., see Milligan and Cooper, 1985 or Arbelaitz et al, 2013). Finally, the Clara algorithm (Kaufman and Rousseeuw, 1990), which is also based on the medoids computation, but works with subsets of original data in order to speed up the computation, has also been carried out.

The results of the application of the four above-mentioned methods along with the related CH and SI statistics are presented in table 1.

Table 1: The optimal numbers of clusters (nCl) and the corresponding values of the Calinski-Harabasz (CH) and Silhouette (SI) cluster validity indices obtained for the four clustering methods (DAPC, k-means, k-medoids and Clara), which were carried out for the two selected variants of the low back pain dataset (Dataset with 10 variables and Dataset with 13 variables).

Methods / Datasets	**DAPC 2 PC (BIC)**	**DAPC 4 PCs (BIC)**	**k-means CH**	**k-means SI**	**k-medoids CH**	**k-medoids SI**	**Clara CH**	**Clara SI**
10 variables	nCl= 3 CH = 82.4 SI = 0.08	nCl = 5 CH = 92.9 SI = 0.11	nCl = 4 CH = 123.1 SI = 0.18	nCl = 10 CH = 99.5 SI = 0.17	nCl = 3 CH = 115.2 SI = 0.16	nCl = 11 CH = 91.1 SI = 0.17	nCl = 5 CH = 97.4 SI = 0.13	nCl = 5 CH = 97.4 SI = 0.13
13 variables	nCl = 3 CH =70.9 SI = 0.06	**nCl = 5 CH =67.3 SI = 0.07**	**nCl = 5 CH = 92.1 SI = 0.15**	nCl = 5 CH = 92.1 SI = 0.15	nCl = 6 CH = 76.1 SI = 0.11	nCl = 6 CH = 76.1 SI = 0.11	nCl = 5 CH = 72.3 SI = 0.12	nCl = 5 CH = 72.3 SI = 0.12

It is worth noting that the values of the CH and SI cluster validity indices reported in table 1 are generally larger for the clusterings with 10 variables than for those with 13 variables. This is due to a well-known property that the values of these indices usually increase with the decrease in the number of variables in dataset. Based on the results presented in table 1, we decided to select for further analysis the dataset with 13 variables because it showed the highest clustering stability in terms of the number of clusters (i.e., the solutions with 5 clusters were provided by 5 out of 8 clustering methods). Hence, we also decided that 5 will be the optimal number of clusters for the low back pain dataset.

Among the 8 clusterings found for the dataset with 13 variables (see table 1), we selected for the consensus analysis (see section 8) the DAPC clustering obtained with 4 principal components because it was the only 5-cluster partition found with the BIC index as well as the k-means clustering found with CH because it provided the largest values of the CH and SI indices. Moreover, the selected k-means clustering had the average cluster stability of 0.86, which was the highest among the presented methods according to the stability analysis conducted with the *fpc* package by Hennig (2010). To perform the stability analysis, we selected the strategy that relies on the use of the Jaccard coefficient and the Bootstrap resampling technique ((Hennig, 2007)) applied to k-means clustering. The *clusterboot* function of the *fpc* package was ran with the default parameters; 100 bootstrap replicates were carried out.

8 Computing a consensus solution for the best clusterings

Finally, we calculated the consensus clustering between the Discriminant Analysis of Principal Components (DAPC) and k-means 5-class partitions that provided the best clustering performances among the selected methods (see section 7). This is a recommended practice when two or more good clusterings are available for a given dataset (Gordon and Vichi, 2001).

The GV3 method implemented in the *cl_consensus* function of the *CLUE* package of R (Hornik, 2005) was carried out. GV3 uses the SUMT algorithm (the "third model" according to Gordon and Vichi, 2001) for minimizing the Gordon and Vichi objective function based on a co-membership dissimilarity. The ARI index (Adjusted Rand Index, (Hubert and Arabie, 1985)) between the considered DAPC and k-means clusterings was equal to 0.19. After performing the consensus clustering using the GV3 algorithm (note that this algorithm implemented in *CLUE* returns as output both crispy and fuzzy consensus clusterings), we were able to calculate the ARIs between the consensus clustering and the DAPC clustering (ARI = 0.24) as well as between the consensus clustering and the k-means clustering (ARI = 0.92). This means that the k-means clustering based on CH was much closer to the obtained consensus clustering solution than the DAPC clustering based on BIC. This conclusion is confirmed by the overall Consensus, DAPC and k-means clustering results presented in Fig. 9.

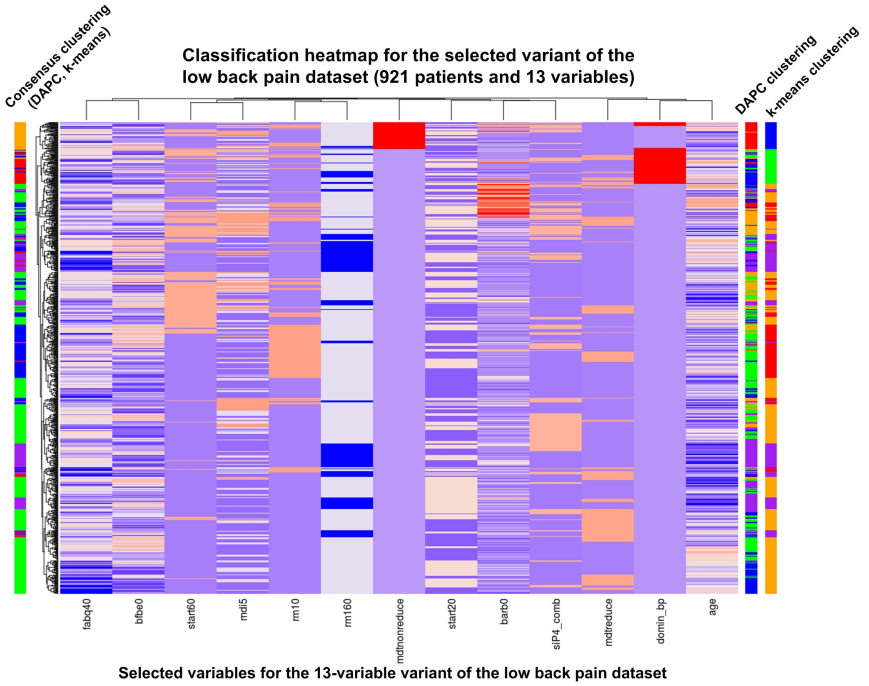

Figure 9: Classification heatmap for the selected variant of the reduced low back pain dataset (consisting of 921 patients and 13 variables). The consensus clustering for the Discriminant Analysis of Principal Components and k-means methods found by the GV3 algorithm is presented on the left. The individual DAPC and k-means clusterings are presented on the right.

9 Description of the obtained consensus clustering and the selected variables with respect to metadata

In this section, we describe the obtained crisp clustering (note that the alternative fuzzy clustering is also provided in our class membership file). While in crisp clustering each object always belongs to only one cluster, it can belong to more than one cluster in fuzzy clustering. Table 2 (below) reports the details on the variables and clusters classification provided by our crisp clustering.

Among the 13 selected variables of the low back pain dataset we can observe representatives of each of the 6 variable domains (contextual factor, participation, pain, psychological, activity and physical impairment), which were specified as

important in the metadata description as well as in the original paper by Nielsen et al. (2016). Each of these domains is represented by at least one variable. This finding confirms that all of the six major variable domains are important for assessing the condition of different groups of patients.

Let us now report the main characteristics of our clusters, which were inferred from the statistics presented in table 2.

Cluster 1 includes the oldest patients; they are mainly full-time workers with physical work load of sitting and walking; they do not stay at home; their back pain is not dominant; they do not suffer from non-paraspinal pain; they do not restraint their activities due to their condition and suffer from pain during muscle palpation; their test of sacro-illiac is mainly negative.

Cluster 2 includes full-time workers of middle age with a heavy physical work load; they stay at home and restraint their activities due to their condition; they avoid heavy jobs around their house; their back pain is dominant and they severely suffer from pain; their test of sacro-illiac is mainly negative.

Cluster 3, which is the largest of the obtained clusters, includes full-workers of middle age with physical work load of sitting and walking; they do not stay at home; their back pain is dominant; they mainly agree that they should not do activities due to their condition and mainly agree to avoid heavy jobs around their house; their test of sacro-illiac is mainly negative.

Cluster 4 includes full-workers of middle age with physical work load of sitting and walking, they do not stay at home; they mainly have non-paraspinal pain onset; their back pain is dominant; they agree that they should not do activities due to their condition and should avoid heavy jobs around their house; they mainly suffer from pain during muscle palpation; their test of sacro-illiac is mainly negative and they do not have reducible disk.

Cluster 5 is composed of the youngest patients; they are mostly male full-workers with physical work load of sitting and walking; they do not stay at home; they suffer greatly from their back pain; they are the least affected by psychological aspects; they do not avoid heavy jobs around their house and are mostly unsure about restricting their activities due to their condition; they mainly suffer from pain during muscle palpation; their test of sacro-illiac is mainly negative.

Table 2: Clustering statistics obtained for the consensus clustering based on the solutions provided by the DAPC and k-means methods. The variable domain is indicated for each of the 13 selected variables (+ Bsex0 (male/female percentage) variable shown for cluster description purposes). For the binary variables, the percentage of each value (0 and 1) is separated by a slash, the mean value is computed for continuous variables and the category with the highest frequency is shown for categorical variables.

Variable domains / Clusters	–	Contextual factor			Participation		Pain	
	Variables	Age	Bsex0	Barb0	Bfbe0	Rm10	Start20	Domin_bp
	Nb of patients	Mean age	Male/Female	Work situation	Physical work load	Stay at home? (Yes/No)	Shoulder/neck pain is not dominating? (Yes/No)	Back pain is not dominating? (Yes/No)
Cluster 1	73	50.74 (SD = 10.45)	49.32 / 50.68 %	full-time	sitting and walking	13.7 / 86.3 %	51.38889 / 48.61111 %	64.38 / 35.62%
Cluster 2	198	43.58 (SD = 11.29)	52.02 / 47.98 %	full-time	heavy physical load	88.39 / 11.61 %	37.5 / 62.5 %	6.06 / 93.94 %
Cluster 3	440	44.27 (SD = 11.21)	56.14 / 43.86 %	full-time	sitting and walking	0 / 100 %	48.86364 / 51.13636 %	0 / 100 %
Cluster 4	59	43.31 (SD = 11.49)	50.85 / 49.2 %	full-time	sitting and walking	30.51 / 69.49 %	52.77778 / 47.22222 %	28.81 / 71.19 %
Cluster 5	151	36.38 (SD = 10.24)	59.6 / 40.4 %	full-time	sitting and walking	3.31 / 96.69	53.59477 / 46.40523 %	0 / 100 %

Variable domains / Clusters	–	Psychological			Activity	Physical impairment		
	Variables	Start60	Fabq40	Mdi5	Rm160	Mdtnonreduce	sIP4_comb	Mdtreduce
	Nb of patients	Worrying? (Yes/No)	should not do physical activities?	Had a bad conscience?	avoid heavy jobs around the house? (Yes/No)	Non-reducible disc? (Yes/No)	SI-joint: Thigh thrust? (Yes/No)	Reducible disc? (Yes/No)
Cluster 1	73	27.40 / 72.60 %	completely disagree	at no time	50.68 / 49.32 %	0 / 100 %	12.33 / 87.67 %	35.62 / 64.38 %
Cluster 2	198	32.83 / 67.17 %	completely agree	at no time	87.88 / 12.12 %	0 / 100 %	26.26 / 73.74 %	23.74 / 76.26 %
Cluster 3	440	25.23 / 74.77 %	completely agree	at no time	100 / 0 %	0 / 100 %	26.36 / 73.64 %	26.14 / 73.86 %
Cluster 4	59	30.51 / 69.49 %	completely agree	at no time	88.14 / 11.86 %	73.61 / 26.39 %	35.59 / 64.41 %	5.08 / 94.92 %
Cluster 5	151	11.26 / 88.74 %	unsure	at no time	0.66 / 99.34 %	0 / 100 %	33.11 / 66.89 %	15.23 / 84.77 %

References

Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I (2013) An extensive comparative study of cluster validity indices. Pattern Recognition 46(1):243–256, Elsevier, DOI 10.1016/j.patcog.2012.07.021

Breunig MM, Kriegel HP, Ng RT, Sander J (2000) LOF: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, ACM, SIGMOD '00, vol. 29, p. 93–104, ISBN: 15-8113-217-4, DOI 10.1145/342009.335388

van Buuren S, Groothuis-Oudshoorn K (2011) MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software 45(3):1–68, DOI 10.18637/jss.v045.i03

Chavent M, Kuentz V, Liquet B, Saracco L (2011) ClustOfVar: an R package for the clustering of variables. Journal of Statistical Software, Articles, p. 1–16, DOI 10.18637/jss.v050.i13

Cohen RJ, Swerdlik ME, Phillips SM (1996) Psychological testing and assessment: An introduction to tests and measurement. Mayfield Publishing Co, Mountian View. ISBN: 15-5934-427-X

Gordon A, Vichi M (2001) Fuzzy partition models for fitting a set of partitions. Psychometrika 66(2):229–247, Springer, DOI 10.1007/BF02294837

Hennig C (2007) Cluster-wise assessment of cluster stability. Computational Statistics & Data Analysis 52(1):258–271, Elsevier, DOI 10.1016/j.csda.2006.11.025

Hennig C (2010) Fpc: Flexible procedures for clustering. URL `https://cran.r-project.org/web/packages/fpc/index.html`, R package

Hornik K (2005) A CLUE for CLUster Ensembles. Journal of Statistical Software 14(12):1–25, University of California, DOI 10.18637/jss.v014.i12

Hu Y, Murray W, Shan Y (2011) Rlof: R parallel implementation of Local Outlier Factor (LOF). URL `https://CRAN.R-project.org/package=Rlof`

Hubert L, Arabie P (1985) Comparing partitions. Journal of classification 2(1):193–218, Springer, DOI 10.1007/BF01908075

Jombart T, Devillard S, Balloux F (2010) DAPC: a new method for the analysis of genetically structured populations. BMC genetics 11(1):94, BioMed Central, DOI 10.1186/1471-2156-11-94

Kaufman L, Rousseeuw P (1987) Clustering by means of medoids. North-Holland, Amsterdam

Kaufman L, Rousseeuw PJ (1990) Partitioning around medoids (program pam), Wiley Online Library, p. 68–125. ISBN: 978-0-470316-80-1, DOI 10.1002/9780470316801.ch2

Kuhn M (2015) Caret: classification and regression training. URL `https://CRAN.R-project.org/package=caret`

Lord E, Willems M, Lapointe FJ, Makarenkov V (2017) Using the stability of objects to determine the number of clusters in datasets. Information Sciences 393:29–46, Elsevier, DOI 10.1016/j.ins.2017.02.010

MacQueen J, et al (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the 5th Berkeley symposium on mathematical statistics and probability, vol. 1, p. 281–297

Malo N, Hanley JA, Cerquozzi S, Pelletier J, Nadon R (2005) Statistical practice in high-throughput screening data analysis. Nature Biotechnology 24(2):167–175, Nature Publishing Group, DOI 10.1038/nbt1186

Milligan GW, Cooper MC (1985) An examination of procedures for determining the number of clusters in a data set. Psychometrika 50(2):159–179, Springer, DOI 10.1007/BF02294245

Mitra P, Murthy C, Pal SK (2002) Unsupervised feature selection using feature similarity. IEEE transactions on pattern analysis and machine intelligence 24(3):301–312, IEEE, DOI 10.1109/34.990133

Nielsen AM, Vach W, Kent P, Hestbaek L, Kongsted A (2016) Using existing questionnaires in latent class analysis: should we use summary scores or single items as input? A methodological study using a cohort of patients with low back pain. Clinical epidemiology 8:73, Dove Press, DOI 10.2147/CLEP.S103330

Steinley D (2004) Standardizing variables in K-means clustering. In: Classification, Clustering, and Data Mining Applications, Springer, Berlin, p. 53–60, ISBN: 978-3-642171-03-1

Stevens SS (1946) On the theory of scales of measurement. Science 103(2684):677–680, DOI 10.1126/science.103.2684.677

Van Belle G (2011) Statistical rules of thumb, vol. 699. John Wiley & Sons. ISBN: 978-0-470144-48-0, DOI 10.1002/9780470377963

Zambrano-Bigiarini M (2014) HydroGOF: Goodness-of-fit functions for comparison of simulated and observed hydrological time series. DOI 10.5281/zenodo.840087, URL `http://hzambran.github.io/hydroGOF/`, R package 0.3-8