Alicja Ciok

On the number of clusters - a grade approach

Instytut Podstaw Informatyki
Polskiej Akademii Nauk

Alicja Ciok

On the number of clusters - a grade approach

Instytut Podstaw Informatyki
Polskiej Akademii Nauk

Warszawa 2004

http://rbc.ipipan.waw.pl

Redaktor Naukowy
Instytut Podstaw Informatyki PAN
Prof. dr hab. Antoni Mazurkiewicz

Pracę do druku opiniował:
Prof. dr hab. Jacek Koronacki

Adres autora:
Alicja Ciok
Instytut Podstaw Informatyki PAN
ul. Ordona 21
01-237 Warszawa

© Copyright by Instytut Podstaw Informatyki PAN, 2004

ISBN 83-910948-9-8

http://rbc.ipipan.waw.pl

Contents

http://rbc.ipipan.waw.pl

1

Introduction

According to the title the main subject of this paper is the determination of the proper number of clusters. The specialist literature provides plenty of various methods in this area but this variety creates also problems. The most serious problem follows from the fact that the determination of the number of clusters can not be separated from the definition of clusters and that there are plenty of such definitions, given directly or hidden inside the procedures of cluster analysis. These problems were raised by many researchers - let us quote a few opinions.

"Anyone who has carried out cluster analysis will be aware of the fact that markedly different results can be obtained when the same data set is analysed using different clustering strategies; it is thus important to give thought to the problem of selecting clustering criteria that are appropriate for analyzing the data being investigated. The problem is analogous to the one of specifying relevant models for data in parametric statistical inference. It is mistake to believe that, because of its more informal approach, classification does not involve a model for the data; the clustering criteria employed implicitly make various assumptions about the data. Use of an inappropriate method of analysis can thus misrepresent the structure in the data." - ([15]).

"Already in 1964, Bonner argued that there could not be a universal definition of cluster and that it was too late to impose one ([5]). Much more recently, we find that clusters and outliers are in the eye of the beholder: one person's noise could be another person's signal ([19])" - ([18]).

"A fundamental, and largely unsolved, problem in cluster analysis is the determination of the "true" number of groups in a data set. Numerous approaches to this problem have been suggested over the years. Unfortunately, many of the approaches that have been suggested for choosing the number of clusters were developed for a specific problem and are somewhat ad hoc. Those methods that are more generally applicable tend either to be model-based, and hence require strong parametric assumptions, or to be computation-intensive, or both" - ([31]).

The immediate conclusion is that the direct comparison of various clustering methods and consequently the comparison of the methods which determine the number of clusters are generally not possible (or limited to narrow areas at the most). It seems that the only way of methods evaluation is to show that they yield satisfactory results when they are applied to many various data sets.

http://rbc.ipipan.waw.pl

It is obvious now that the presentation of the grade methods of cluster analysis must start from a cluster definition. Since in our case this definition is imposed by the grade framework, the paper starts from the short presentation of this framework. The grade methods and procedures were already presented in several papers, e.g. [12], [11], [13], [25]. Referring interested readers to them for more detals, we recall now only a few ideas which are necessary to understand this paper.

Chapter 2 includes the basic ideas of the grade correspondence analysis (GCA) and the clustering method which is based on it (GCCA). Also the method of visualization of the grade methods results, so called overrepresentation map, is presented.

The core of all grade mehods is the concept of concentration of one random variable with respect to the other. This concentration is expressed by two concentration indices (the concentration curves and the numerical concentration index which is based on concentration curve). Both indices measure diversity of one random variable with respect to the other. The concentration curves are exploited in the procedure of simultaneous discretization of two random variables. This procedure requires a number of new categories as an input parameter. It is described in Sec. 2.2.

Thanks to the equivalence between the concentration curves and the grade correlation curves this discretization can be applied in the case of two-dimensional data tables. The results are the clusters of rows or columns of the data table; their numbers must be provided at the start. In the grade framework the input data tables must have the form of bivariate probability tables, yet this does not mean that the GCCA is restricted to such tables only. On the contrary, it is discussed and illustrated by various examples of real and randomly generated data that the grade methods can be applied to many kinds of data tables, for example to tables which include values of variables (Chapters 2 and 3). Of course in any case the proper interpretation of results needs appropriate adjustments according to the data type.

The further generalization can be made when the input probability table with infinite number of rows and columns is taken into account. In such a case the bivariate probability table becomes the density table. It should be emphsized that the used formalism is applicable for finite as well as for infinite data tables.

Irrespective of the data meaning all results are expressed in the language of statistical dependence measured by the grade correlation coefficient ρ^* for the row and column variables which characterize the data table. The grade correspondence analysis (GCA) maximizes this dependence (the value of ρ^*) in the set of all permutations of rows and columns of the data table. The optimal arrangement of the data table results in the clear description of the GCCA clusters (of course if the cluster structure exists), what is very helpful in their interpretation.

http://rbc.ipipan.waw.pl

The same clustering procedure is applicable for the rows as well as for the columns of data tables. Therefore, this procedure can provide the separate clusterings of rows or columns but also it can be exploited in the procedure of simultaneous clustering of rows and columns. In this paper only separate clusterings are taken into account, a detailed description of the grade simultaneous clustering can be found in [11].

The concept of natural cluster is discussed in Chapter 3; it follows directly from the grade discretization framework. The natural clusters are determined by linear segments of the respective correlation curves; these segments correspond to the constant values of the respective grade regression functions (the standardized integral of the grade regression function is equal to the correlation curve).

The very important property of the GCA is that it permutes the rows and columns of data table in such a way that they are ordered according to the values of respective regression functions. This is the reason why the GCCA forms clusters only from adjacent rows (columns) - adjacent in the GCA optimal permutations. Consequently, the differences between regression values for particular rows (columns) can serve as dissimilarity measures. The matrix which includes these measures has a particularly desirable property - it is anti-Robinson. In other words the rows (columns) are arranged according to their similarities.

In Sections 3.2 and 3.3 the results of the GCCA applied to several data tables (real or artificially created) are shown and compared. All of them include the values of variables instead of real probabilities. Those randomly generated tables are specially selected to show that the quality of the GCCA results depends chiefly on the separability of the data values among particular clusters irrespectively of the data meaning. As these tables consists of two columns (variables) it is possible to calculate the probability table corresponding to the pairs of variables (after the preliminary discretization of the variables). The results of the GCCA applied to these tables are very similar to the results obtained for the initial tables; once more it turns out that their similarity depends on the data separability. The comparison of results of these analyses shows also the difference between the well-known approaches to clustering: the random-partition and the mixture models.

Chapter 4 is wholy devoted to the determination of the proper number of clusters in the grade framework. The problems with this determination were signaled already in Sec. 3.3, when the real questionaire data which characterize the living conditions of Polish households are analysed. In this chapter the concept of ideal regularity is introduced and exploited to determine the number of GCCA clusters. First, a regularity is defined for the concentration curves, next it is generalized for two-dimensional data tables. Roughly speaking the regularity can be interpreted as infinite divisibility, that is there is not "true" number of clusters (in other words, any number

http://rbc.ipipan.waw.pl

of clusters may be accepted). The clusters obtained for the chosen number describe well the monotone trends generated by the GCA. On average these clusters differ among themselves, but they are not well separated. The example from Sec. 3.3 is a good illustration of such a data structure.

The first step in determining the number of clusters is checking whether the data table is regular and consequently whether there are no well separated clusters in the data. The problem is that the direct checking via the definition can be troublesome. The proposed solution consists of two parts. Both exploit the paraboloidal family of data tables (it is the generalization of the univariate distributions with the cdfs in the form of parabola for two-dimensional data tables) which is ideally regular and hence without any cluster structure present. This family is treated as an ideal model of regular data. The choice of this family as a reference point is not accidental. The segments of correlation curves in the form of parabola correspond to the natural clusters discussed in Chapter 3.

If the data table is very similar to a member of the paraboloidal family, the data should be considered as almost ideally regular. If such a paraboloidal data are not found, then we may use the method which is based on the colinearity of so called regression points. This method is descibed in Sec. 4.2.3. It is designed for finite data tables and helps to determine whether the table derives from the discretized paraboloidal family. The colinearity of the regression points confirms this hypothesis. Let us note that finite tables can not have the property of infinite divisibility, hence they can be only approximately regular. Almost regular finite tables also do not possess separated clusters.

The regression point method is used for evaluation whether the generated clusters are the natural clusters and consequently whether the chosen number of clusters is appropriately chosen. This evaluation is based on a few specially designed measures of clusters' homogeneity. The idea met in varius statistical domains and adopted here is to use the various measures of clusters' homogeneity simultaneously to assess the quality of clusters. Such a procedure helps to express different aspects of possible cluster structures and makes the inference more reliable.

Finally, the method of cluster evaluation which is based on the regression points is used in determination of the number of clusters. The choice of this number is based on the comparisons of the clusters' homogeneity measures calculated for various possible numbers of clusters with the analogous measures calculated for randomly generated data.

In Chapter 5 the relatons between the GCCA and the other clustering procedures are discussed. As the reasons presented above prevent us from the direct comparison of the GCCA with other respective methods, the location of the grade methods in the selected typologies of clustering procedures is discussed.

http://rbc.ipipan.waw.pl

2

Grade correspondence-cluster analysis (GCCA)

2.1 Grade correspondence analysis (GCA) and overrepresentation map

Grade methods of data analysis require input data in the *form of bivariate probability tables.* Such tables are very well known and commonly used in the analysis of pairs of discrete random variables whose categories correspond to particular rows and columns of the data table. When the numbers of rows and columns goes to infinity then the probabilities become the densities. Hence bivariate densities can be expressed as infinite probability tables for pairs of continuous variables, whose rows and columns correspond to particular values of these variables.

Let us note that this restriction does not imply that the grade methods are applicable only to this kind of data. On the contrary, as any two-dimensional data table can be easily transformed into this probability form, the grade methods are applicable to a broad spectrum of data types in the form of two-dimensional tables, for example to the tables including values of attributes. Data values are just treated as probabilities irrespectively of their meaning. Consequently the grade methods can be also applied to multivariate data.

Irrespective of the contents, the tables must meet the following additional conditions:

- the marginal sums of table rows and columns (eventually after appropriate normalization) must be greater than zero,
- the values included in these tables should reflect some order, for example they can express intensity of levels of features. If the data table includes the values of a nominal variable, this variable should be replaced by dummy variables corresponding to particular categories. Let us note that the pairs of nominal variables are acceptable without restrictions when the data tables include their respective bivariate probabilities.

Due to this bivariate probability form, data structures can be expressed in terms of stochastic dependence between marginal (row and column) variables (say X and Y) irrespectively of the data meaning. Let us note that this universal language does not imply that the interpretation of results

http://rbc.ipipan.waw.pl

does not depend on data. Obviously the interpretation must be appropriately adjusted for the data tables which include values of variables.

The data in the form of bivariate probability table will be denoted by $P = (p_{i,j})$. It corresponds to a pair of row and column variables (X, Y). If these variables are discrete then the table is finite. When they are continuous, the probabilities become densities, which can be treated as a generalized two-dimensional table. This table is infinite and the row index i (the column index j) takes real values. The grade formalism is applicable in both cases.

Instead of initial variables X and Y (discrete for finite tables), the pair of *continuous* variables (X^*, Y^*) defined on the unit square $[0,1] \times [0,1]$ is considered. If the marginal variables X and Y are continuous then X^* and Y^* can be expressed by the transformation: $X^* = F_X(X)$ and $Y^* = F_Y(Y)$, where F_X and F_Y are respective cumulative distribution functions (cdfs) of X and Y. This transformation is called the grade transformation, hence the name of the methods. For discrete variables the grade transformation is followed by randomization, that is *the initial categories are replaced by the intervals of values.* These values are distributed uniformly inside the respective intervals. Therefore the bivariate distribution of the new continuous variables (X^*, Y^*) is characterized by the density h which is constant and equal to

$$h_{ij} = p_{ij}/(p_{i\bullet}p_{\bullet j}) \tag{2.1}$$

on any rectangle

$$\{(u,v) : S^X_{i-1} < u \leq S^X_i \text{ and } S^Y_{j-1} < v \leq S^Y_j\} \tag{2.2}$$

where $S^X_i = \sum_{l=1}^{i} p_{l\bullet}$, $S^Y_j = \sum_{l=1}^{j} p_{\bullet l}$ and $p_{i\bullet} = \sum_{l=1}^{k} p_{il}$, $p_{\bullet j} = \sum_{l=1}^{m} p_{lj}$ for $i = 1, ..., m$; $j = 1, ..., k$. This density is called the *randomized grade density of* (X, Y) (cf. [25]).

Each of the variables X^* and Y^* are uniform on $[0,1]$. The bivariate distributions on $[0,1] \times [0,1]$ with uniform marginals are known in literature as copulas (literature on copulas is enormous - see e.g. [29]). Therefore the joint distribution of X^* and Y^* is the *copula of* (X, Y). Let us note that any change in the permutations of rows and columns of the data table (generally, in the arrangements of values of X and Y) affects the values of S^X_i and S^Y_j and consequently changes the copula.

The grade density can be visualised by so called overrepresentation map (cf. [25]). This map consists of the unit square partitioned into rectangles. These rectangles are defined by formula (2.2); they represent particular values from the input data table. The widths of rows (columns) of the map reflect respective marginal sums of rows (columns) of the data table.

http://rbc.ipipan.waw.pl

The value range of the grade density h is partitioned into several subintervals (categories) represented by various colours. The rectangles of the unit square marked by these colours form the overrepresentation map of the data table. Let us note that according to the definition the values of the grade density h_{ij} measure deviation from statistical dependency between the row and column variables. Hence, if a value h_{ij} is high we say that there is a high overrepresentation corresponding to the cell (i, j) of the data table.

In this paper 5 categories for discretization of h are used; the respective partition points are the following: $\frac{2}{3}$, 0.99, $\frac{1}{0.99}$, $\frac{3}{2}$. The adopted convention is that dark colours mark high magnitudes, light colours correspond to the low. As an illustration let us compare the data shown in Tab. 2.1 and the corresponding overrepresentation map shown in Fig. 2.1.

TABLE 2.1. Example of data

		Y	
X	1	2	3
1	0.1	0.2	0
2	0.3	0.1	0.1
3	0.05	0	0
4	0	0.1	0.05

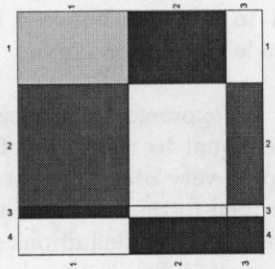


FIGURE 2.1. Overrepresentation map for data from Tab. 2.1

Now let us return to the general formalism. The standard correlation coefficient between X^* and Y^*

$$\rho^*(X, Y) = corr(X^*, Y^*) \tag{2.3}$$

measures dependence between variables X and Y. It is called the *randomized grade correlation coefficient*. For discrete variables, it is equal to

http://rbc.ipipan.waw.pl

Schriever's extension of Spearman's ρ (cf. [30]). The coefficient ρ^* may be expressed by various equivalent formulas. In the correspondence and cluster analysis the following formula is the most convenient one.

$$\rho^*(X,Y) = 6\int_0^1 (u - C^*(Y:X)(u))du = 6\int_0^1 (u - C^*(X:Y)(u))du, \tag{2.4}$$

where

$$C^*(Y:X)(t) = 2\int_0^t r^*(Y:X)(u)du;\ t \in [0,1] \tag{2.5}$$

is called the *randomized grade correlation curve* and

$$r^*(Y:X)(t) = E(Y^* \mid X^* = t);\ t \in [0,1] \tag{2.6}$$

is the *randomized grade regression function.*

Grade correlation curves are always continuous and lie in the unit square. If X (Y) is discrete then $r^*(Y:X)$ ($r^*(X:Y)$ respectively) is constant in the intervals defined by formula (2.2), hence $C^*(Y:X)$ ($C^*(X:Y)$) consists of linear segments linking point $(0,0)$, the points determined by the categories of X (Y respectively) and point $(1,1)$. Let us note that each row or column of the initial data table (each value of X or Y) is characterized by the value of respective grade regression function r^* and the corresponding linear segment of correlation curve C^*. In any case the grade correlation coefficient is proportional to the area between the diagonal of the unit square ($y = x$) and the grade correlation curve.

Correlation curves can be expressed as concentration curves and the correlation coefficient ρ^* is equal to the value of concentration index ar. This property turns out to be very useful, and it is exploited in the grade clusterting procedure described in the next section.

To proceed further let us recall the definitions of both concentration indices. Both measure the diversity between two random variables. Let V and Z be random variables with the common value range S, the respective cdfs F_V, and F_Z, and the densities f_V, and f_Z. If variables V and Z are continuous, the concentration curve of Z on V, denoted by $C(Z:V)$, consists of points $(F_V(t), F_Z(t))$, where $t \in S$. If these variables are discrete, the curve $C(Z:V)$ consists of linear segments linking points: $(0,0)$, $(F_V(t), F_Z(t))$, $t \in S$ and point $(1,1)$. The concentration index $ar(Z:V)$ is based on the concentration curve $C(Z:V)$ and is defined as:

$$ar(Z:V) = 2\int_0^1 (u - C(Z:V)(u))du \tag{2.7}$$

http://rbc.ipipan.waw.pl

According to this formula the concentration index ar is equal to twice the area between the diagonal of the unit square and the concentration curve. As concentration curves are also cdfs of variables defined on $[0,1]$, it is justified to write shortly $ar(K)$, where $K = C(Z : V)$, instead of $ar(Z : V)$.

Let us consider the pair of variables: X^* and $R_{Y:X}$, where $R_{Y:X}$ has values from $[0,1]$ and the density equal to $2r^*(Y : X)$. The correlation curve $C^*(Y : X)$ (cf. (2.5) and (2.6)) is the cdf of the variable $R_{Y:X}$. As X^* is uniform, the correlation curve $C^*(Y : X)$ is also the concentration curve $C(R_{Y:X} : X^*)$ and ρ^* is proportional to the respective concentration index ar. The pair of variables: Y^* and $R_{X:Y}$ determines the analogous concentration curve which is equal to the second correlation curve $C^*(X : Y)$.

Figure 2.2 shows an example of the regression function $r^*(Y : X)$, where this function is calculated for the data table from Tab. 2.1. Its segments correspond the categories of the initial row variable X.

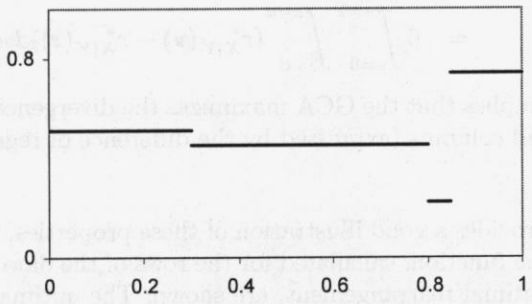


FIGURE 2.2. Grade regression function $r^*(Y : X)$ calculated for the data from Tab. 2.1

The grade correspondence analysis (GCA) maximizes the positive dependence between X and Y (measured by $\rho^*(X,Y)$) in the set of all permutations of rows and columns of the data table (categories of X and Y). This maximal value of ρ^* will be denoted here by ρ^*_{max}. Thanks to the aforementioned equivalence of correlation and concentration curves, maximization of ρ^* is equivalent to maximization of the concentration index ar, and consequently is equivalent to minimization of the area below the correlation curve.

The GCA has many useful properties which are exploited in the procedure of cluster analysis described further:

http://rbc.ipipan.waw.pl

- identical rows (columns) always occupy adjacent places in the respective GCA optimal permutations and their aggregation (sum) does not change the value of ρ^*,
- *both regressions* $r^*(Y : X)$ *and* $r^*(X : Y)$ *are nondecreasing for the GCA optimal permutations* (cf. [13], [25]) *and consequently both correlation curves are convex.* Thanks to this property the difference between the values of regression function r^* corresponding to the rows (columns) of the initial data table (categories of the initial row and column variables) can serve as a dissimilarity measure between these rows (columns). In this sense the GCA always arranges *rows (as well as columns) according to their similarity* and consequently the matrix of these dissimilarities is anti-Robinson.

Links between coefficient ρ^* and these dissimilarities can be expressed by the following formula (cf. [25]):

$$\begin{aligned}\rho^*(X,Y) &= 6\int_{u=0}^{u=1}\int_{s=0}^{s=u}(r^*_{Y|X}(u) - r^*_{Y|X}(s))dsdu = \\ &= 6\int_{u=0}^{u=1}\int_{s=0}^{s=u}(r^*_{X|Y}(u) - r^*_{X|Y}(s))dsdu\end{aligned}$$

This formula implies that the GCA maximizes the divergence between particular rows and columns (expressed by the difference of regression values)

.

Figure 2.3 provides a good illustration of these properties, where the values of regression function, calculated for the rows of the data from Tab. 2.1 but after its optimal rearrangement, are shown. The optimal permutation for rows is : 4, 1, 2, 3; and for columns: 2, 3, 1. The comparison of Figs 2.2 and 2.3 shows very clearly how much the GCA can change the values of regression functions.

2.2 Grade correspondence-cluster analysis (GCCA)

2.2.1 Procedure scheme

The grade cluster analysis (GCCA) is based on optimal permutations provided by the GCA and it does not consider overlapping clusters. Assuming that the numbers of clusters are given, the rows and/or columns of the data table (categories of X and/or Y) are optimally aggregated. The respective probabilities (data values in the input table) are the sums of component probabilities, and they form a new data table. The *optimal aggregation means that* $\rho^*(X,Y)$ *is maximal in the set of these aggregations of rows and/or columns, which are adjacent in the GCA optimal permutations.*

http://rbc.ipipan.waw.pl

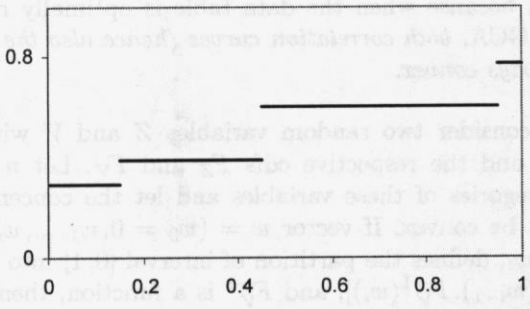


FIGURE 2.3. Grade regression function $r^*(Y : X)$ calculated for the data from Tab. 2.1 rearranged according to the GCA

The monotonicity of regressions (the values of $r^*(Y : X)$ and $r^*(X : Y)$ corresponding to the initial rows or columns of the data table are nondecreasing if the respective permutations are optimal) provides the reason why clusters include only rows (columns), which are adjacent in the GCA optimal permutations. This property is also exploited in the procedure of ρ^* maximization (cf [12], [7], [8]).

The rows and columns of the initial data table may be aggregated either separately (i.e. we maximize ρ^* for aggregated X and nonaggregated Y or for nonaggregated X and aggregated Y), or simultaneously. The first type of clustering is called the single clustering, the simultaneous clustering for X and Y is called the double clustering. The comparison of both methods can be found in [11]. As the double clustering is based on the sequence of single clusterings the problem of cluster determination is common to both methods.

2.2.2 Maximization framework

The clustering maximization procedure is based on the idea of simultaneous discretization of two random variables via their concentration curve (cf. [12], [7]). In this clustering framework the equivalence between the grade correlation curves ($C^*(Y : X)$ or $C^*(X : Y)$) and the concentration curves for the respective pairs of variables ($C(R_{Y:X} : X^*)$ or $C(R_{X:Y} : Y^*)$ respectively) are exploited. The clusters of rows or columns of the initial data table are determined by the simultaneous discretization of pair $(X^*, R_{Y:X})$ or $(Y^*, R_{X:Y})$ discussed in the previous section. In the clustering of rows the first pair of variables is taken into account, if columns are clustered then the second pair should be considered.

The discretization procedure which maximizes $\rho^*(X, Y)$ requires convex-

http://rbc.ipipan.waw.pl

ity of the concentration curves (the correlation curves). This condition is always fulfilled because when the data table is optimally rearranged according to the GCA, *both correlation curves (hence also the concentration curves) are always convex.*

Now let us consider two random variables Z and V with a common value range S and the respective cdfs F_Z and F_V. Let n be the number of new categories of these variables and let the concentration curve $C = C(Z : V)$ be convex. If vector $w = (w_0 = 0, w_1, ..., w_{n-1}, w_n = 1)$, where $w_{i-1} \leq w_i$, defines the partition of interval $[0, 1]$ into n consecutive intervals $(F_V^{-1}(w_{i-1}), F_V^{-1}(w_i)]$, and F_V^{-1} is a function, then the intervals $(F_V^{-1}(w_{i-1}), F_V^{-1}(w_i)]$ define the partition of the common range S into non-overlapping subsets. If F_V^{-1} is not a function then the following intervals are considered $(\min_{F_V^{-1}(t)=w_{i-1}} t, \max_{F_V^{-1}(t)=w_i} t]$. These intervals determine n categories of the new, discrete variables.

According to the definition, the concentration curve corresponding to these new variables, consists of linear segments linking points $(w_i, F_Z(w_i))$ $(i = 0, ..., n)$. These points are common with the initial concentration curve C. The immediate conclusions is that for convex curves the discretization can only decrease values of the concentration index.

It is obvious that a discretization should provide a minimal loss of information about the joint distribution of considered variables. Discretizations which distort the characteristics of this distribution may lead to a wrong inference if it is based on the discretized variables. In our framework, this requirement is expressd in the form of maximization of the correlation coefficient ρ^* for discretized row or column variables (equivalently maximization of the concentration index for the special pair of variables).

For a given number of categories the simultaneous discretization of two random variables is called optimal if the respective value of the concentration index is maximal. The procedure which generates an optimal partition is based on the necessary conditions for optimality provided by the following theorem (cf. [12], [7]).

Theorem 1 *Let C be a convex concentration curve $C(Y : X)$. If partition vector w determines the optimal simultaneous discretization of variables (X, Y) into n categories, then the following condition must be satisfied:*

$$\frac{dC}{dx}(w_i - 0) \leq \frac{C(w_{i+1}) - C(w_{i-1})}{w_{i+1} - w_{i-1}} \leq \frac{dC}{dx}(w_i + 0) \tag{2.8}$$

If C has a continuous derivative (denoted by c) then condition (2.8) is transformed into the equality:

$$c(w_i) = \frac{C(w_{i+1}) - C(w_{i-1})}{w_{i+1} - w_{i-1}} \tag{2.9}$$

http://rbc.ipipan.waw.pl

Very often the necessary condition is also sufficient. For example, if the derivative c of C is continuous and increasing then the solution of condition (2.8) is unique. Moreover, even if more solutions exist, they are equivalent in the sense that they have identical values of the concentration index (cf. [7]).

Figure 2.4 shows three concentration curves: one corresponds to the pair of continuous variables (Z, U), where U is uniformly distributed. The interval $[0, 1]$ is the common support of both variables, the cdf of Z is the parabola: $F_Z(x) = x^2$. The remaining two concentration curves corresponds to the optimally discretized counterparts of Z and U; one of the curve corresponds to discretization into 2 categories, the second corresponds to 4 categories. In both cases the discretization points are $w_i = \frac{i}{n}$, where $i = 0, ..., n$ and $n = 2$ or 4 respectively. As the support S is equal to the interval $[0, 1]$, the intervals $(\frac{i-1}{n}, \frac{i}{n}]$ determine the new categories of variables Z and U.

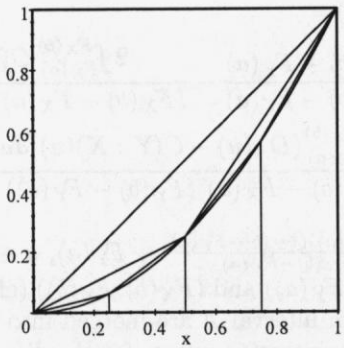


FIGURE 2.4. Concentration curve in the form of parabola and its two discretized counetrparts

Generally, the discretization points w_i for curve $C(Z : V)$ are the quantiles of order $F_V(w_i)$ for variable V. In the case of correlation curve $C^*(Y : X)$ points w_i are equal to $S^X_{k_i}$ which correspond to the initial categories of X and are defined by formula (2.2).

The discretization determined by $w_i = \frac{i}{n}$ will be called the uniform quantile discretization. The uniform discretization is optimal for the concentration curves in the form of parabola. As this property is valid for every n, it is clear that the optimal discretization into kn categories only adds new partition points to those defining the discretization into n categories. This property has great implications for determination of clusters - this problem will be discussed in the next chapter.

http://rbc.ipipan.waw.pl

Usually in cluster analysis the requirement of homogeneity of generated clusters is formulated. Many various indices are used in the literature to measure this homogeneity. It is natural that in the concentration framework the cluster homogeneity should be measured by the concentration index ar.

Let us consider variables X and Y with the common support S and respective differentiable cdfs: F_X and F_Y. X^A and Y^A are the same variables, but restricted to the subinterval $A = [a, b] \subset S$. Their cdfs are denoted respectively by F_X^A and F_Y^A. Assuming that F_X^{-1} is a function, it is easy to calculate a so called restricted concentration curve.

The restricted concentration curve $C(Y^A : X^A)$ is equal to

$$C(Y^A : X^A)(u) = \frac{F_Y[F_X^{-1}[u(F_X(b) - F_X(a)) + F_X(a)]] - F_Y(a)}{F_Y(b) - F_Y(a)} =$$
$$= \frac{C(Y : X)[u(F_X(b) - F_X(a)) + F_X(a)] - F_Y(a)}{F_Y(b) - F_Y(a)}, \tag{2.10}$$

where $u \in [0, 1]$. Consequently

$$ar(C(Y^A : X^A)) =$$
$$= \frac{F_Y(b) + F_Y(a)}{F_Y(b) - F_Y(a)} - \frac{2\int_{F_X(a)}^{F_X(b)} C(Y : X)(u)du}{(F_X(b) - F_X(a))(F_Y(b) - F_Y(a))} =$$
$$= \frac{2\int_{F_X(a)}^{F_X(b)} (D_A(u) - C(Y : X)(u))du}{(F_X(b) - F_X(a))(F_Y(b) - F_Y(a))}$$

where $D_A(u) = \frac{(u - F_X(a))(F_Y(b) - F_Y(a))}{F_X(b) - F_X(a)} + F_Y(a)$, $u \in A$, is a linear segment linking points $(F_X(a), F_Y(a))$ and $(F_X(b), F_Y(b))$ (cf Fig. 2.5). If the initial values of variables from interval A are merged into one category then D_A is a segment of the concentration curves for the discretized variables which corresponds to this new category (value). Hence the area between D_A and the initial concentration curve $C(Y : X)$ measures a loss of information caused by this discretization. This loss can be expressed by the restricted concentration curve and the corresponding concentration index.

Consequently, if the intervals $A_i = (a_i, a_{i+1}]$, $i = 1, ..., n$, form the partition of the common range (and consequently determine the discretization) then

$$ar(C(Y : X)) - ar(C(\widehat{Y} : \widehat{X})) = \sum_{i=1}^{n-1} ar(C(Y^{A_i} : X^{A_i}))\, \Delta_X(A_i)\, \Delta_Y(A_i)$$

where $\widehat{X}$ and $\widehat{Y}$ denote the discretized counterparts of the initial variables; the discretization is defined by A_i, $i = 1, ..., n$, and $\Delta_X(A_i) = F_X(a_{i+1}) - F_X(a_i)$, $\Delta_Y(A_i) = F_Y(a_{i+1}) - F_Y(a_i)$.

http://rbc.ipipan.waw.pl

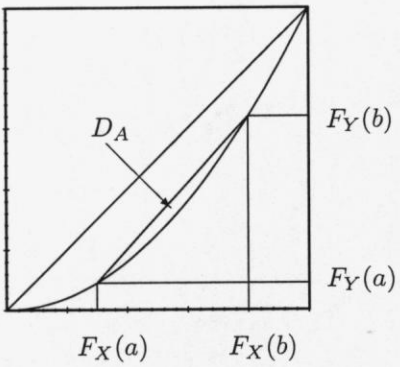


FIGURE 2.5. Example of discretization

Hence the difference between the concentration indices for the initial variables and their discretized counterparts mesures the diversities "within" new categories (that is within the intervals which form these categories). Of course, thanks to the equivalence between the concentration index and the correlation coefficient, the same statement can be formulated in the case of grade cluster analysis for data tables and the corresponding correlation coefficients ρ^*. Therefore the GCCA minimizes the overall, within cluster diversity. It resambles in this point many other clustering methods (for example the well known k-means method), which also minimizes within diversity, however this diversity can be expressed in many different ways. The definitions of homogeneity measures usually depends on the accepted notion of cluster and consequently they must have a great impact on determination of the proper number of clusters.

http://rbc.ipipan.waw.pl

3

Natural clusters

3.1 General remarks

An enormous volume of literature is devoted to the problem of "real" cluster detection which is a good measure of the problem difficulty. The well known, classic (although not formal) definition of cluster analysis says that it should divide the data into blocks of similar elements, with blocks differing as greatly as possible. The problem is that there are many ways of expression of similarity as well as of diversity concepts, what leads to many clustering procedures and generates problems with comparison of results.

A part of the solution is provided by the following citation from ([14]). "Any classification (clustering) is a division of the objects into groups based on a set of rules - it is neither true nor false and should be judged on the usefulness of results". As it is hard to imagine useful results without practical interpretation, hence interpretability becomes a main criterion for the evaluation and comparison of clustering procedures.

In this section it will be shown what kinds of clusters are considered in the grade framework. The grade discretization framework provides a natural definition of clusters (categories of row or column variables). According to the definition of concentration curves, they include linear segments if the variables are discrete. On the other hand, continuous variables can also produce concentration curves which include linear segments. This occurs when the common range of the variables is partitioned into consecutive intervals where both variables have constant or proportional densities. The idea of inner uniformity as a cluster distinguishing feature is not new (for example, it was discussed in [3]). In the grade discretization framework this idea is a straightforward generalization of natural categories of discrete variables.

The equivalence between grade correlation curves and concentration curves provides immediate generalization of natural clusters in the case of two-dimensional data tables. The intervals of values for row (column) variables (determining the subsets of rows or columns) define natural clusters if they are characterized by the same values of the grade regression function. Identical vectors of conditional densities or probabilities (or their formal counterparts for the data other than bivariate probabilities) imply that the corresponding values of the row (column) variable belong to the same cluster. Of course, the equality of regressions can coexist with the diversity of the corresponding data vectors. In any case, when the probabilities (data

http://rbc.ipipan.waw.pl

values) inside clusters of rows (or columns) with constant regressions are aggregated (summed), the regressions and the value of ρ^* calculated for the aggregated data table remain unchanged. In other words, in such a case the optimal aggregation does not cause any loss of information except granulation. If all regressions are constant, then there are no clusters and $\rho^*_{\max}$ for the initial table is equal to zero, but the opposite inference is not true.

These properties have a great implications for cluster characterization. The value of ρ^* for the data aggregated according to the GCCA measures the diversity of clusters. The difference between the value of $\rho^*_{\max}$ for the initial table and the analogous value for the aggregated table (according to the GCCA) measures the clusters' homogeneity. The difference equal to zero means that the clustering restores the real cluster structure present in the data. An interesting question arises: does small difference always imply a good approximation of real clusters?

On the other hand, as it was discussed above, the differences between values of the grade regression functions play the role of the dissimilarity measures among the values of row (column) variable. Since the GCA always arranges these values in nondecreasing order of the grade regression functions, we are able to say that they are ordered according to their similarities and the diversity of regressions inside particular clusters measures also the lack of clusters' homogeneity. Let us note that if the data table contains values of variables (say $Z_1, .., Z_k$) which characterize the set of objects, then the values of regression function calculated for rows (objects) are just the weighted sums of values of Z_i. Hence the differences in the values of regression function work similarly to any disimilarity measure commonly used in cluster analysis. That is, for two objects characterized by similar values of variables Z_i, the difference in regression function (dissimilarity) is near zero. When the data table of this kind is characterized by a strong dependence between the row and the column variables, then consecutive groups of objects in the optimal ordering (and consequently in the GCCA clusters) are characterized by particularly (albeit relatively) large values corresponding to the consecutive groups of variables. If the values of variables are treated as "measurements of similarity to some standard" (i.e., not in the sense of mathematical theory of measurement), this interpretation becomes almost identical with the interpretation for probability data. This provides additional support for the argument that the GCCA are applicable to tables various kinds of data. The analysis of several examples of data tables presented in the next section confirms that this point of view is right.

http://rbc.ipipan.waw.pl

3.2 Grade approach versus random-partition and mixture models

In this section relations between the grade approach to clustering and the two classic approach: the random-partition model and the mixtures models (cf. [2]) will be discussed. To this purpose several radomly generated data examples will be analysed and compared. Let us remind that in the first (random-partition) model clusters are considered as the samples drawn from different distributions specific for particular clusters. In the second model clusters are also samples but they correspond to particular components of the mixure distribution.

All examples are generated in the same way. The tables contain values of variables, where rows correspond to objects and columns correspond to the variables. As data tables are always transformed to the form of bivariate probability table, then we may also assume that the value range for all variables is $[0,1]$. Let $m_i = (m_i^1, ..., m_i^k)$, $m_i^l \in [0,1]$, $i = 1, ..., s$ be fixed vectors and let $U = (U_1, ..., U_k)$, where U_j, $j = 1, ..., k$, be random variables uniform on $[-1,1]$. The generated data are realizations of the k-dimensional random variables $G_i = m_i + \alpha U$, where α is a constant and $i = 1, ..., s$; for each i the block of k-dimensional data were generated. The part of data table obtained by one generator will be called a data block.

Let $T_1, ..., T_5$ denote the generated data tables. In all examples two variables ($k = 2$) and three generators ($s = 3$) are considered, the sample sizes are identical and equal to 100 for each data block. Therefore each table T_i has 2 columns and 300 rows and includes 3 data blocks. All generated data are shown in Fig. 3.1a-d and in Fig. 3.2.

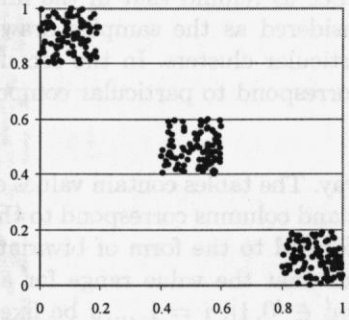
The GCCA procedure was applied to each table T_i, $i = 1, ..., 5$, a priori knowledge about number of clusters ($n = 3$) was used. In the case of tables T_1, T_2 and T_3 (Fig. 3.1a, b, c) all initial data blocks are restored completely, that is the blocks of data corresponding to different generators are identical with the GCCA row clusters.

In the case of table T_4 (Fig. 3.1d) there are 11 misclassifications (3.66(7)% of the sample), that is for 11 data vectors the assignments to the blocks do not agree with the assignments to the clusters. However such an effect in this case must be expected. The data blocks corresponding to particular generators G_i ($i = 1, 2, 3$) overlap, whereas the GCCA does consider overlapping clusters. Let us note that these misclassifications change only slightly the cluster centroids; the respective means inside clusters are equal to $\widehat{m}_1 = (0.2813, 0.7012)$, $\widehat{m}_2 = (0.4933, 0.5067)$, $\widehat{m}_3 = (0.7014, 0.3105)$.

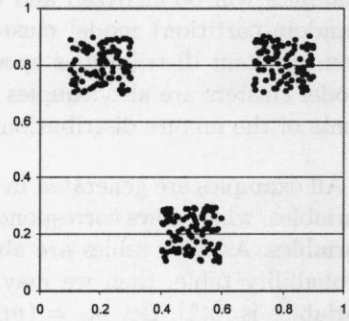
It seems that the data from table T_5, shown in Fig. 3.2, are very similar to those shown in Figs 3.1a and c. All of them are generated in the same way but the important difference is that the mean vectors m_i are proportional in

http://rbc.ipipan.waw.pl

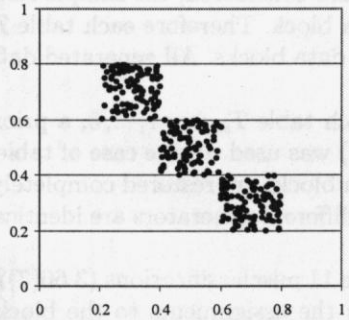
(a) - T_1
$m_1 = (0.1, 0.9)$
$m_2 = (0.5, 0.5)$
$m_3 = (0.9, 0.1)$
$\alpha = 0.1$
$\rho^*_{\max} = 0.568$



(b) - T_2
$m_1 = (0.2, 0.8)$
$m_2 = (0.8, 0.8)$
$m_3 = (0.5, 0.2)$
$\alpha = 0.1$
$\rho^*_{\max} = 0.317$



(c) - T_3
$m_1 = (0.3, 0.7)$
$m_2 = (0.5, 0.5)$
$m_3 = (0.7, 0.3)$
$\alpha = 1/10$
$\rho^*_{\max} = 0.290$



(d) - T_4
$m_1 = (0.3, 0.7)$
$m_2 = (0.5, 0.5)$
$m_3 = (0.7, 0.3)$
$\alpha = 1/7$
$\rho^*_{\max} = 0.306$

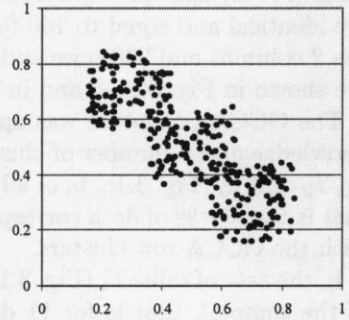


FIGURE 3.1. Data tables T_1, T_2, T_3 and T_4

this case. The consequence is that they become identical when transformed into the conditional probability form and consequently the corresponding rows are assigned to the same clusters. For data tables which contain real probabilities this is how it should be. When the tables include the values of variables this phenomenon has two different implications.

Proportionality for columns means that the corresponding variables differ only in their scales. In this case they should be reckoned as similar and

http://rbc.ipipan.waw.pl

$$m_1 = (0.25, 0.25)$$
$$m_2 = (0.50, 0.50)$$
$$m_3 = (0.75, 0.75)$$
$$\alpha = 0.1$$

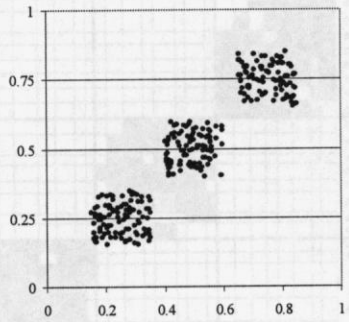


FIGURE 3.2. Data table T_5

placed in the same cluster. In the case of rows (objects) it depends on data whether such an effect is admissible or not. Sometimes (for example for questionaire data) only values which are higher (or lower) than the average level, characteristic for a particular respondent, are significant (not the real values). In this case proportional rows should belong to the same cluster. When real magnitudes are important, the problem can be solved for example by inserting an additional variable (column) into the data table. This variable is equal to the sum of the initial variables but with the scale reversed. For table T_5 the function $2-x$ was used as the reversion function. After this modification of the data the GCCA restores completely all three data blocks.

When the data table includes many columns-variables then the probability of appearance of proportional vectors significantly decreases. It practically reduces to zero when the table includes strongly but negatively correlated variables. The practical implication of this observation is that to avoid the undesirable proportionality effect, the scales of some variables from the set of highly correlated variables should be reversed.

Obviously the data tables which include values of variables can be easily transformed into probability tables. As the GCCA can be applied to the both tables, the question arises: are the results similar? Let us compare the pairs of results for tables T_1 and T_4. To estimate the probabilities the interval $[0, 1]$ was partitioned into 40 subintervals of equal lengths. Each subinterval is represented by its center (mean); these centers are values of the new, discretized variables (all variables are discretized in the same way). Figures 3.3 and 3.4 show the overrepresentation maps for the optimally permuted probability tables corresponding to the discretized variables from

http://rbc.ipipan.waw.pl

T_1 and T_4.

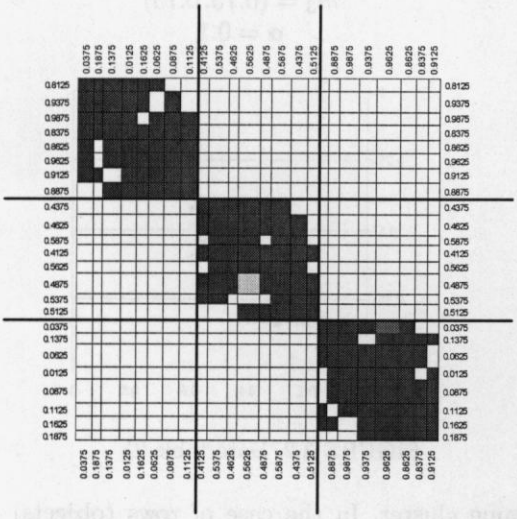


FIGURE 3.3. Overrepresentation map for data table T_1

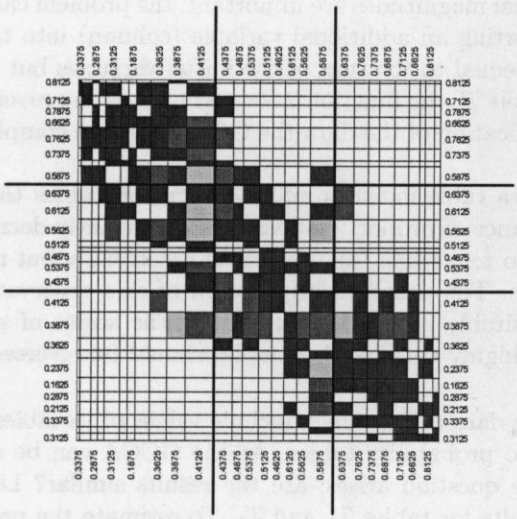


FIGURE 3.4. Overrepresentation map for data table T_4

The horizontal and vertical lines mark boundaries of clusters for rows and columns respectively (3 was assumed as the number of clusters for

http://rbc.ipipan.waw.pl

rows as well for columns). These clusters were determined by the GCCA applied separately for rows and for columns.

Now the problem arises how to compare the results of clusterings for the initial data tables and for the corresponding probability tables constructed as descibed above. Let the row clusters obtained for the initial data tables be called value clusters. The pairs of clusters (i, j) generated for rows and columns of the respective probability table are called probability clusters. In these pairs the first element corresponds to the clusters of rows, the second corresponds to the clusters of columns.

If (v_1, v_2) denotes a value vector which correspond to the pair of initial variables (it forms a row of the initial data table); and $(\widehat{v}_1, \widehat{v}_2)$ denotes the analogous vector which correspond to the discretized variables (it forms also a row but in the transformed data table), then by agreement between value and probability clusters we mean that (v_1, v_2) belongs to the i-th value cluster if and only if $(\widehat{v}_1, \widehat{v}_2)$ belongs to the (i, i)-th probability cluster. Let us note that the complete restoration of initial data blocks by the GCCA does nor imply that the natural order of values is retained by the GCA for the probability table.

The clusters generated for the probability data corresponding to table T_1 agree ideally with the respective clusters generated for the initial data.

The analogous compatibility between the value and probability clusters for table T_4 is characterized by the frequencies shown in Tab. 3.1. This table includes the numbers of discretized data vectors $(\widehat{v}_1, \widehat{v}_2)$ (these vectors correspond to particular rows of the initial data table) which are assigned to the i-th value cluster and simultaneously to the (j, l)-th probability cluster. The fact that the highest frequency values correspond to the pairs (i-th value cluster, (i, i)-th probability cluster), $i = 1, 2, 3$, confirms that the results of both clusterings agree to a high degree. In this evaluation we must take into account the granularity caused by discretization, the small sample sizes which distort the probability estimators, and the fact that the data blocks corresponding to different generators overlap. In the case of overlapping blocks the values of variables can not determine clusters exactly (like it was for table T_1).

The presented observations are a good illustration of the discrepancy between two basic probabilistic models of clusters known in the literature. The first is the random-partition model, the second is the mixture model (cf for example [2]). The random model assumes that a global population of objects is divided into m subpopulations. m clusters of objects are obtained by random, independent sampling from these subpopulations. The classic definition adds here that the subpopulations should be "homogeneous" and differ among themselves. The mixture model assumes that the sample data are drawn according the distribution which is a mixture of several distributions corresponding to particular clusters. Looking at Fig. 3.1d one easily agrees that in the case of non-overlapping clusters "the mixture model

http://rbc.ipipan.waw.pl

TABLE 3.1. Frequency table for probability and value clusterings

Probability clusters	Value clusters 1	2	3	Total
(1,1)	82	0	0	82
(1,2)	16	22	0	38
(1,3)	0	4	0	4
(2,1)	2	4	0	6
(2,2)	0	51	1	52
(2,3)	0	10	21	31
(3,2)	0	5	4	9
(3,3)	0	1	77	78
Total	100	97	103	300

provides no clustering approach in the strong sense" (cf [2]).

Table T_4 generates also other difficulties. The three blocks obtained by different generators overlap and there are no natural boundaries between particular clusters like for tables T_1, T_2 or T_3. Consequently, if they are separated by any clustering method, it is arguable whether the generated groups are different clusters according to the classic cluster definition because there are no gaps between them. Therefore it is reasonable to accept that there is only one cluster. On the other hand this one cluster is not very homogeneous. The many clusters solution provides clusters which on average differ among themselves and are more homogeneous. The GCCA produces often such cluster structures when a strong monotone trend is present in the data. This trend is described well by these ordered clusters, so the cluster structure can serve as a tool of description of monotone trends and therefore the separability of particular clusters are not necessary.

3.3 Living condition of Polish households - analysis of data example

Let the analysis of a real data serve as an illustration of the problems with cluster determination. The table includes the questionaire data which describe the living conditions of 658 Polish households, where the rows correspond to the households, the columns correspond to the questions. The data come from the research of Ms K. Kuśmierczyk (1999) of the Institute of Home Market and Consumption. Table 3.2 includes the desciption of all questions used in this analysis. Most of the questions characterize housholds' satisfaction with respect to various aspects of their living conditions. Several questions concern the assessments of actual or future financial conditions of households or the prediction of changes in Polish economy. All

http://rbc.ipipan.waw.pl

corresponding variables have integer values whose ranges vary from $[1,2]$ to $[1,6]$.

According to the formal requirements, the input data for the GCA must be transformed into the form of bivariate probability table. In this case the normalization consists of two steps. First, all columns are transformed into the form of conditional probability vectors, that is the values in each column are divided by the total sum in this column. This way a possible influence of different variables' scales is avoided. Next, all values are divided by the total sum in the table after the first normalization.

The GCA was applied to the normalized table, its results are shown in Tab. 3.2 and also in Fig. 3.5, which presents the overrepresentation map of our normalized data. The table as well as the figure show the question (columns) and the households (rows) in the optimal permutation according to the GCA.

TABLE 3.2. Analyzed variables characterising the living condition data

No	Name	Description
1	Q6_E	Are household needs for holiday rest met?
2	Q6_F	Are household needs for culture goods and services met?
3	Q4	Assessment of the actual financial situation of household
4	Q6_B	Are household needs for clothing and footwear met?
5	Q6_C	Are household needs for forniture, radio and television, household appliances met?
6	Q6_A	Are household needs for food met?
7	Q6_D	Are household needs for housing condition met?
8	Q10	Does financial situation force the reduction of household expenditures?
9	Q22	How will general economic situation change in the next 2-3 years, will it improve or worsen?
10	Q9	Is it possible to save something with the actual income?
11	Q21	How will the standard of household living change in the next 2-3 years?
12	Q1	Did the financial situation of household improve or worsen comparing with the last year?
13	Q7	Does household own saving?
14	Q2	Assessment of actual financial situation of household compared with that from before 10 years
15	Q20	Did the standard of household living change recently?
16	Q5	Does household have enough money for its needs?

The overrepresentation map reveals that there is a natural partition of the variable set. Remembering that dark rectangles in the map correspond to large values of variables and light rectangles correspond to low values, one can see that there are two natural cluster of variables. The scale direction

http://rbc.ipipan.waw.pl

of variables from one cluster is opposite to the scale direction for the other cluster. The GCCA applied to the columns, where the number of clusters is equal to 2 provides the clusters which ideally coincides with these natural partition. The vertical line in Fig. 3.5 and the horizontal line in Tab. 3.2 mark the boundaries of these clusters.

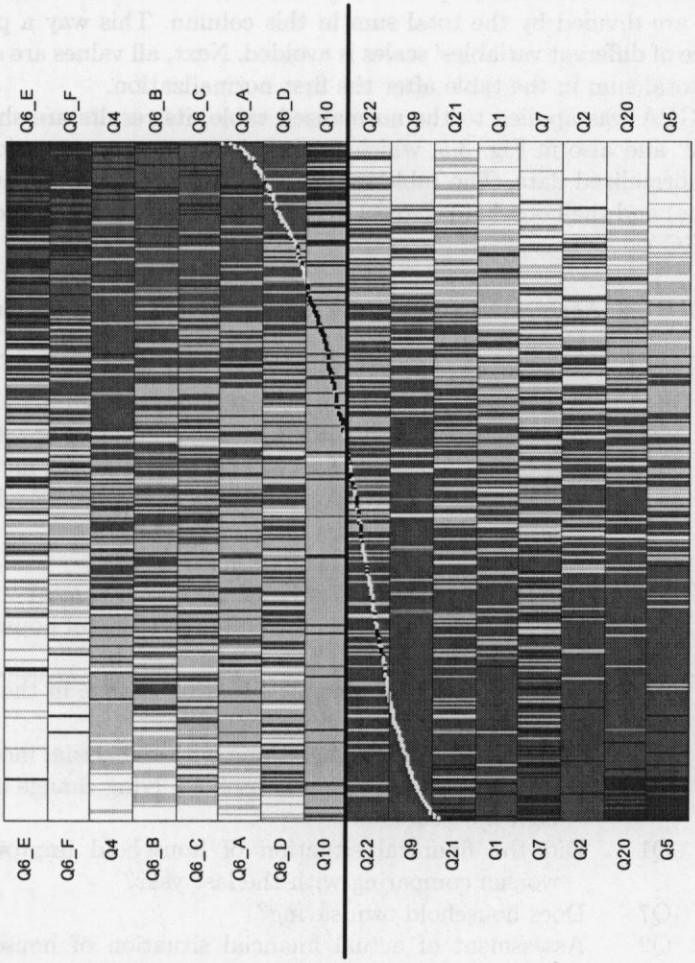


FIGURE 3.5. Overrepresentation map for the living condition data

There is not a cluster pattern with natural boundaries visible for rows (households). This does not mean that the rows are not differentiated. On the contrary, the monotone trend generated by the GCCA can be interpreted as a quality of living conditions. According to it, the households are arranged is such a way that these whose various needs are met concentrate

http://rbc.ipipan.waw.pl

at the one end of the trend scale (left upper corner of the overrepresentation map). The households which declare that they are not satisfied with respect to their various needs are grouped at the opposite end of the scale. This trend strengthens if the financial situation improved in last ten years for the households from the first cluster, or if this situation worsen for households from the second cluster.

As the cluster determination is based on the values of grade regression function, let us compare these values calculated for our data. These values (for rows as well as for columns) are shown in the overrepresentation map but they are better visible in Figs 3.6 and 3.7. Let us note that instead of linear segments like those shown in Fig. 2.2, the regressions function are represented here by the points: the regression values corresponding to particular rows or columns on vertical axis and the center of the respective intervals on the horizontal axis (the regressions are constant in these intervals). This kind of respresentation turn out to be useful in cluster determination and will be discussed in detail in the next chapter. Here it just shows how much the regression values are differentiated.

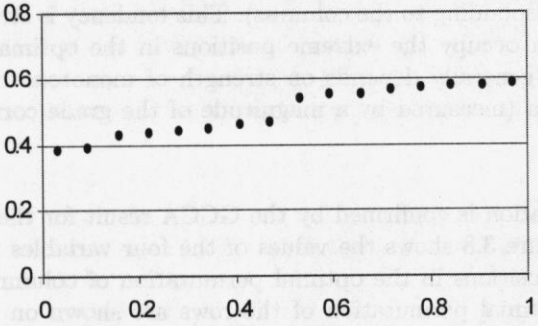


FIGURE 3.6. Values of the grade regression function calculated for the columns of living condition data table

There is a natural partition of the regression values for columns (Fig. 3.6), it is determined by the gap between 8-th and 9-th column. This gap coincides with the boundaries of the optimal clusters provided by the GCCA when the assumed number of clusters is equal to 2.

In the case of row regression the situation is different. The regression values (Fig. 3.7) go smoothly without significant gaps. They form almost linear curve; only small groups of the lowest and the highest values deviate from it. Consequently there is no natural partition point like that for the row clustering. On the other hand the regression values are differentiated.

http://rbc.ipipan.waw.pl

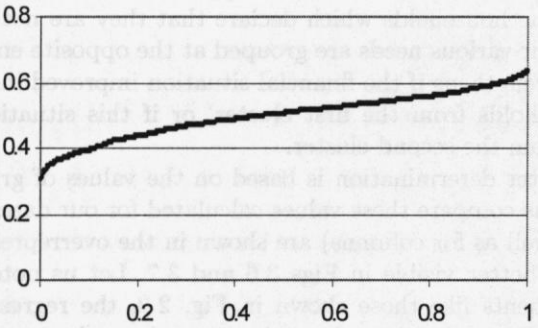


FIGURE 3.7. Values of the grade regression function calculated for the rows of living condition data table

Usually in the optimaly permuted data tables different regression values for particular clusters coincide with different values of data (values of variables corresponding to the columns). This tendency is strongest for the columns which occupy the extreme positions in the optimal GCCA permutation and generally depends on strength of monotone trend which is present in data (measured by a magnitude of the grade correlation coeficient).

This observation is confirmed by the GCCA result for the living condition data. Figure 3.8 shows the values of the four variables which occupy the extreme positions in the optimal permutation of columns (from both ends). The optimal permutation of the rows are shown on the horizonal axes, the corresponding values of variables - on the vertical axes. The vertical line in Fig. 3.8 marks the boundaries of the two row clusters generated by the GCCA. It is easy to observe that the distributions of the variables in these clusters differ significantly. This observation is confirmed by the respective frequencies calculated for both clusters and shown in Tab. 3.3. These frequencies show also that the difference between clusters is of statistical nature, that is there are no gaps in data values between particular clusters.

The discussion of the data examples presented above (the living data table as well as the randomly generated data tables) leads to the following conclusions.

- The quality of the GCCA results depends on the separability of data among distinct blocks of values, irrespective of the data meaning.

http://rbc.ipipan.waw.pl

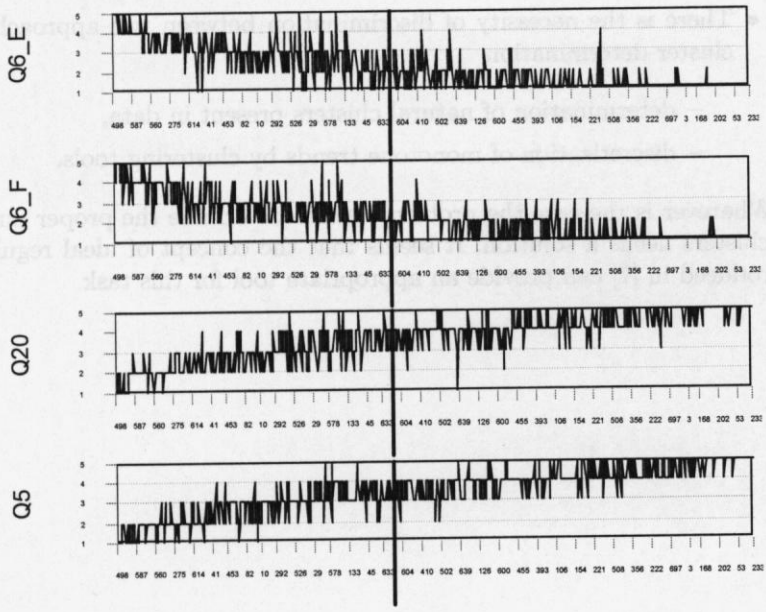


FIGURE 3.8. Values of the selected variables for the living condition data when the rows are optimaly permuted according to the GCA

TABLE 3.3. Frequencies of categories for the selected variables in two optimal clusters

Cluster			Q6_E		
	1	2	3	4	5
1	19	63	115	72	17
2	251	100	18	3	0
			Q6_F		
	1	2	3	4	5
1	11	82	114	61	18
2	226	126	19	1	0
			Q20		
	1	2	3	4	5
1	15	106	106	52	7
2	1	2	42	179	148
			Q5		
	1	2	3	4	5
1	24	99	118	42	3
2	0	3	56	197	116

http://rbc.ipipan.waw.pl

- There is the necessity of discrimination between two approaches to cluster determination
  - determination of natural clusters present in data,
  - discretization of monotone trends by clustering tools.

Whatever is the case the problem how to determine the proper number of clusters needs a solution. It seems that the concept of ideal regularity introduced in [7] can provide an appropriate tool for this task.

http://rbc.ipipan.waw.pl

4

Regularity concept in grade cluster analysis

4.1 Regularity concept in cluster modeling

4.1.1 Ideally regular concentration curves

There are random variables which have convex concentration curves with extraordinary properties. Let us consider the optimal discretization of variables X and U defined on $[0,1]$ into n categories. Variable U is uniform and the cdf of X is the parabola x^2. As it was mentioned above, the optimal discretization in this case is equal to the uniform quantile discretization, that is, the discretization points are equal to $\frac{i}{n}$, $i = 0, ..., n$. As this property is valid for every n, it is clear that the optimal discretization into kn categories only adds new partition points to those defining the discretization into n categories. This property gives rise to the following definition.

Definition 2 *Let vector* $w = (w_0 = 0, w_1, ..., w_{n-1}, w_n = 1)$ *define the optimal discretization of the convex concentration curve* C *into* n *categories. Vectors* $w^i = (w_0^i = w_{i-1}, w_1^i, ..., w_{k-1}^i, w_k^i = w_i)$ *determine the partitions of the subintervals* $[w_{i-1}, w_i]$ $(i = 1, ..., n)$ *according to the condition (2.8). Moreover,* $w_j^i \neq w_{j+1}^i$ *for all* $i = 0, ..., n$ *and* $j = 0, ..., k-1$. *If for every* n *and* k *the discretization into* nk *categories defined by* w^i $(i = 1, ..., n)$ *is optimal, then the curve* C *is called ideally regular.*

Ideal regularity can be interpreted as infinite divisibility. The following properties provide additional arguments for this interpretation.

- The optimal simultaneous discretization of the variables restricted to a chosen interval is determined by the optimal partition of the respective interval of the initial concentration curve.
- The immediate conclusion is that if a concentration curve is ideally regular, then its restricted counterparts are also ideally regular. Moreover, this property repeats on any partition level.

The following example is a good illustration of these properties. Fig. 4.1a presents the ideally regular curve in the form of parabola and two optimally discretized counterparts generated for 3 and 6 categories. The six subcategories are derived from separate partitions (according to the condition (2.8)) of the previously obtained three categories. This is a good

http://rbc.ipipan.waw.pl

illustration of the ideal regularity concept. Each category can be partitioned separately into further subcategories according to the condition (2.8) and the discretization still remains optimal.

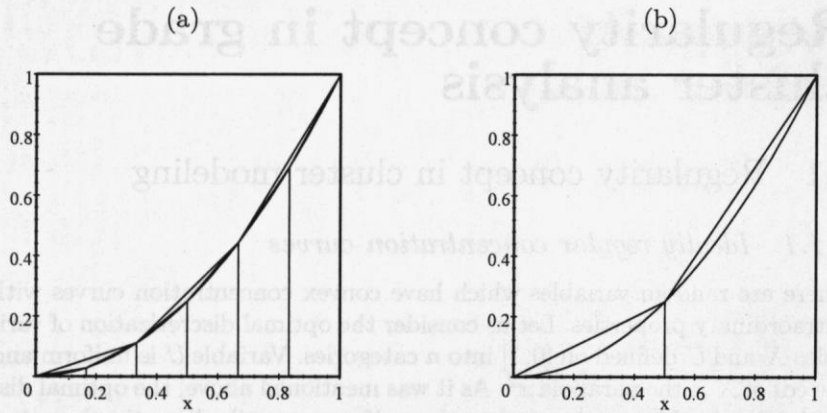


FIGURE 4.1. Parabola curve and its two discretized counterparts (a), the corresponding restricted concentration curve and its discretized counterpart (b)

Figure 4.1b shows the restricted concentration curve $C(X^A : U^A)$, where the restriction interval A for both variables X and U is equal to $[0, \frac{1}{3}]$. Figure 4.1b shows the restricted concentration curve which is discretized into two categories. This restricted curve is generated by the linear transformation given by formula (2.10); the same function transforms the partition point $\frac{1}{6}$ for $C(X : U)$ into the partition point $\frac{1}{2}$ for $C(X^A : U^A)$.

Currently, two families of ideally regular curves are known: parabolas and ellipses. The regularity of ellipses follows from the fact that they are restricted concentration curves calculated for curves in the form of circles. The infinite divisibility of circles follows from the fact that the optimal partitions generate the circle segments of identical lengths. It can be proven that curves which are symmetric w.r.t. the diagonal $y = x$ or $y = 1 - x$ are also ideally regular (cf. [7]). Consequently, we have the additional families of ideally regular curves symmetric to parabolas and ellipses.

The parabolas are the ideal examples of concentration curves where the only one natural category is present - its derivative is constant on the whole interval $[0, 1]$.

4.1.2 Ideally regular two-dimensional data tables

The concept of ideally regular concentration curves can be generalized on two-dimensional data tables (cf. [9]). Like the family of parabolas is considered as the models of ideally regular curves, analogous family of models

http://rbc.ipipan.waw.pl

for generalized regularity of two-dimensional data tables is needed. Such a family should be characterized by the lack of any cluster structure like in the one-dimesional case.

Let us consider a parametrized family of pairs of continuous variables. These variables are defined on $[0,1]$ and their joint densities f_β are given by the following formula:

$$f_\beta(x,y) = \beta + 2(1-\beta)(x+y-2xy) \quad x,y \in [0,1]. \tag{4.1}$$

The respective cdfs F_β are equal to:

$$F_\beta(x,y) = xy(\beta + (1-\beta)(x+y-xy)) \quad x,y \in [0,1].$$

These pairs of variables will be called the paraboloidal and denoted by (X_β, Y_β), where $\beta \in [0,2]$. If $\beta = 1$ then the density f_β is constant and consequently the distribution is uniform on the unit square ($F_1(x,y) = xy$). If the scale order of variable X (or Y) is changed into its reverse, the density preserves its form, but the parameter β changes into $2-\beta$:

$$f_\beta(1-x,y) = f_\beta(x,1-y) = f_{2-\beta}(x,y)$$

The important property of the paroboloidal family is that the corresponding conditional densities are given by the linear formulas, consequently the conditional cdfs are parabolas, which are ideally regular curves.

Since the grade regression functions, in the case of continuous variables defined on the unit square are equal to the common regression: $E(Y_\beta|X_\beta = x)$ then

$$r^*_{Y_\beta|X_\beta}(x) = \frac{1}{6}(4-\beta+2x(\beta-1)), x \in [0,1]$$

These regressions are increasing (decreasing) functions of x iff $\beta > 1$ (iff $\beta < 1$). If $\beta = 1$ then $r^*_{Y_\beta|X_\beta} \equiv 0.5$. The relationship between parameter β and $2-\beta$ can be expressed in the form:

$$r^*_{Y_\beta|X_\beta}(x) = 1 - r^*_{Y_{2-\beta}|X_{2-\beta}}(x), x \in [0,1].$$

Due to the symmetry, the analogous formula can be calculated for $r^*_{X_\beta|Y_\beta}$.

The grade correlation curve is given by the formula:

$$2C^*_{cor(Y_\beta:X_\beta)}(x) = \frac{x}{3}\left(4-\beta+(\beta-1)x\right), x \in [0,1]. \tag{4.2}$$

This correlation curve is a parabola; if $\beta > 1$ it is convex (if $\beta < 1$ it is concave). Of course the identical calculation can be made for $C^*_{cor(X_\beta:Y_\beta)}$. Due to symmetry, the repective formula is identical to formula (4.2).

http://rbc.ipipan.waw.pl

The known correlation curves enable us to calculate the grade correlation coefficient $\rho^*(X_\beta, Y_\beta)$, which in this case reduces to the simple formula:

$$\rho^*(X_\beta, Y_\beta) = \frac{\beta - 1}{3}$$

Note that in this case $\rho^*(X_\beta, Y_\beta) = \rho^*_{\max}(X_\beta, Y_\beta)$. Since $\beta \in [0, 2]$, the values of $\rho^*(X_\beta, Y_\beta)$ must belong to the interval $[-\frac{1}{3}, \frac{1}{3}]$. The value of $\rho^*(X_\beta, Y_\beta) > 0$ iff $\beta > 1$. Moreover $\rho^*(X_{2-\beta}, Y_{2-\beta}) = -\rho^*(X_\beta, Y_\beta)$.This equality is in agreement with the fact described above that changing from β to $2 - \beta$ is equivalent to the reversal of value order for X_β or Y_β.

Let us consider the GCCA procedure. According to its rules the new categories of row (column) variables are determined by the optimal discretization of the respective grade correlation curves. The first step in this procedure is to arrange the values of both variables in nondecreasing order of the respective grade regressions r^*. Paraboloidal variables have their regression functions properly arranged if $\beta > 1$. In the following we will consider only the paraboloidal variables with such parameters. Generally, in the family of paraboloidal variables, both normalized correlation curves are parabolas which are convex for $\beta > 1$.

The definition of a paraboloidal family can be generalized to bivariate distributions whose value ranges are not restricted to the interval $[0, 1]$. The set of bivariate distributions whose grade distributions are defined by formula (4.1) will be called the paraboloidal family. These generalized family retains all properties discussed above.

The comparison of the clusters for paraboloidal variables provided by the GCCA procedure, with m and mk as the numbers of clusters for rows (columns), reveals behaviour identical to the behaviour of ideally regular concentration curves. The clustering for mk does not change the cluster boundaries, but only divides further the clusters obtained previously for m clusters. When this property does not depend on the values of m and k, such a property can be interpreted as *infinite divisibility*. Consequently, the concept of ideal regularity can be analogously introduced for two-way data tables.

Definition 3 *Two-way data tables will be called ideally regular if both corresponding normalized grade correlation curves are ideally regular. If only one correlation curve is ideally regular then the data table will be called ideally semi-regular.*

Since parabolas are ideally regular, the pairs of paraboloidal variables are also ideally regular.

Let us compare the results of two clusterings for rows (as well as for columns) of the paraboloidal pair of variables (X_β, Y_β) for 20 and 10 clusters, where $\beta = 1.5$. The value of ρ^* for the initial continuous distribution

http://rbc.ipipan.waw.pl

is equal to 0.166(6); for the discretized variables the value of ρ^* decreases only to 0.1658 and to 0.1633 respectively for 20 and 10 categories. Generally, the values of ρ^* decrease smoothly and monotonically while numbers of clusters increase. In other words, there are not such numbers of clusters which cause a sudden jump in the values of ρ^* and which consequently can help to determine the proper numbers of clusters.

Like parabolas in the case of curves, paraboloidal bivariate distributions have no natural clusters. Generally, all ideally regular data tables have this property. As the ideal regularity concept and deviations from it were exploited for determination of the proper number of variables' categories, a similar procedure will be repeated now for two-way data tables.

4.2 Regularity and cluster detection

4.2.1 Discrete almost regular data tables

The single clustering procedure can always provide a discretization of one variable (row or column); the double clustering discretizes both variables simultaneously. If the data table is (semi) regular, then both homogeneity measures discussed above increase monotonically when the number of clusters increases, and they never reach zero. This implies that any number of clusters is fine on the condition that the values of these measures are acceptable.

When the table is not ideally regular and the grade regressions (the grade correlations as well) and the value of ρ^* do not change after aggregation according to the clustering, then the real structure (the proper number of clusters and the proper clusters) is revealed.

Like the parametrized family of ideally regular curves in the univariate case, the parametrized family of paraboloidal distributions can be used to check that there is no natural granularity (clusters) in the data. *Instead of checking the definition for the formula of ideal regularity, it is much easier to find the paraboloidal table which is most similar to the given table. If the similarity is great, the data can be regarded as approximately ideally regular.*

Unfortunately the data tables one usually works with are finite and non-symmetric. These tables are transformed into copulas which are continuous distributions; obviously these copulas have the natural cluster structure corresponding to the initial data. Therefore, they can not be ideally regular, they can only approximate ideally regular (or semi regular) tables.

Let f be the joint density of a bivariate distribution corresponding to the pair of continuous variables (X, Y) defined on $[0,1] \times [0,1]$. Let us

http://rbc.ipipan.waw.pl

assume that each of these variables is uniform (after transformation into the copula they are always uniform) and that the density has a more general form than the densities of the paraboloidal tables (in particular it can be nonsymmetric).

$$f(x,y) = a_1xy + a_2x + a_3y + a_4,\ x,y \in [0,1].$$

For a fixed x (y) this density as well as the regression function $r^*_{Y|X}$ ($r^*_{X|Y}$) is a linear function of y (x); the corresponding correlation curve is a parabola and consequently this curve is ideally regular.

Let $(\widehat{X},\widehat{Y})$ be the same variable pair, but after independent discretizations of both variables, these discretizations being determined by respective vectors $(x_1,...,x_n)$ and $(y_1,...,y_k)$. Let us note that both variables remain uniform after discretization. The regression function $r^*_{\widehat{Y}|\widehat{X}}$ is given by the following formula:

$$2r^*_{\widehat{Y}|\widehat{X}}(x_{i+1}) = \sum_{j=1}^{k} \frac{(y_{j+1}+y_j)p_{i+1,j+1}}{p_{i+1,\bullet}},$$

where p_{ij} $(i=1,...,n;\ j=1,...,k)$ are the probabilities of this new discrete distribution and $p_{i+1,\bullet} = \sum_{j=1}^{k} p_{i+1,j} = x_{i+1} - x_i$. According to definition, the correlation curve consists of linear segments determined by the points $(x_j, 2\sum_{i=1}^{j} r^*_{Y|X}(x_i)p_{i\bullet})$. Then the slopes of particular segments are equal to $2r^*_{Y|X}(x_{i+1})$. For our distribution

$$\begin{aligned} p_{i+1,j+1} &= \frac{1}{2}(x_{i+1}-x_i)(y_{j+1}-y_j)[a_1(x_{i+1}+x_i)(y_{j+1}+y_j) + \\ &\quad + a_2(x_{i+1}+x_i) + a_3(y_{j+1}+y_j) + 2a_4] \end{aligned}$$

and

$$\begin{aligned} 2r^*_{\widehat{Y}|\widehat{X}}(x_{i+1}) &= \frac{1}{2}\sum_{j=1}^{k}(y^2_{j+1}-y^2_j)[a_1(x_{i+1}+x_i)(y_{j+1}+y_j) + \\ &\quad + a_2(x_{i+1}+x_i) + a_3(y_{j+1}+y_j) + 2a_4]. \end{aligned}$$

Then $2r^*_{\widehat{Y}|\widehat{X}}(x_{i+1})$ is a linear function of $\frac{1}{2}(x_{i+1}+x_i)$ for fixed $y_j, j=1,...,k$. This provides a simple rule for determining whether the correlation curve is derived from a parabola via discretization (in other words, whether it can be regarded as an approximation of a parabola). If points

$$(\frac{1}{2}(x_{i+1}+x_i), r^*_{\widehat{Y}|\widehat{X}}(x_{i+1})) \text{ for } i=1,...,n \tag{4.3}$$

lie on a line then we can say that the corresponding correlation curve is approximately ideally regular and the rows do not form clusters. Of course the same determination rule can be created for columns of data tables.

http://rbc.ipipan.waw.pl

4.2.2 Analysis of simulation examples

The figures presented below (Figs 4.2-4.5) show the regression points defined by formula (4.3) for the data tables T_1, T_2, T_3 and T_4 considered in Sec. 3.2. The points shown in Figs 4.2 and 4.3 form three almost linear segments, each clearly distinguished from the other. In Fig. 4.4 also three almost linear segments are visible, but they are not strongly differentiated. In Fig. 4.5 all points lie approximately on one line; there are not distinguished linear segments.

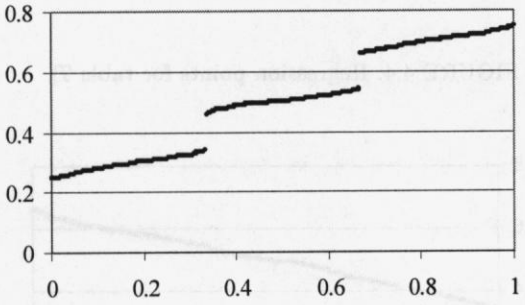


FIGURE 4.2. Regression points for table T_1

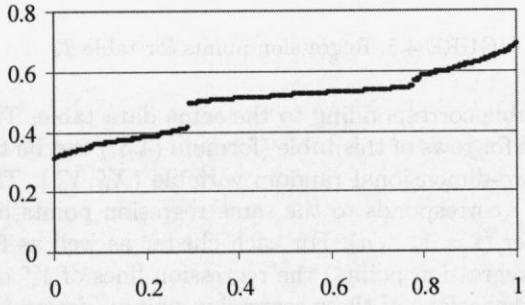


FIGURE 4.3. Regression points for table T_2

Many indices can be used to measure deviations from linearity. In this paper the well known measure is applied: the *variance of regression residuals*. Let n be the number of clusters for the row variable X and let Y be

http://rbc.ipipan.waw.pl

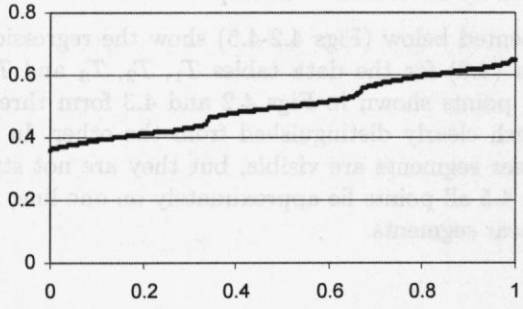


FIGURE 4.4. Regression points for table T_3

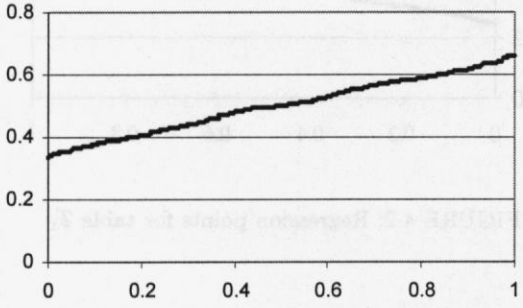


FIGURE 4.5. Regression points for table T_4

the column variable corresponding to the same data table. The regression points calculated for rows of this table (formula (4.3)) can be treated as the realizations of two-dimensional random variable (X_0^r, Y_0^r). The analogous variable (X_i^r, Y_i^r) corresponds to the same regresion points but restricted to the i-th cluster $(i = 1, ..., n)$. For each cluster as well as for the whole sample (all row regression points) the regression lines of Y_i^r on X_i^r can be calculated. The deviations of these regression points (determined by (4.3)) from the respective regresion lines are measured by the standard residual variance $var_{res}(Y_i^r : X_i^r)$. This variance can be expressed by various formulas, one of them is the following:

$$var_{res}(Y_i^r : X_i^r) = (1 - corr^2(X_i^r, Y_i^r))var(Y_i^r),\ i = 0, 1, ..., n, \qquad (4.4)$$

where $corr$ denotes the Pearson correlation coefficient, and $var(Z)$ is the variance of variable Z.

http://rbc.ipipan.waw.pl

Table 4.1 includes the residual variances calculated for tables $T_1, ..., T_4$. In each case the variances are generated for all points and for particular clusters (like in Sec. 3.2, the considered number of clusters n is equal to 3). It is interesting that the variances in particular clusters remain on the similar level for all data tables, whereas the analogous variances for the merged clusters are strongly differentiated. Moreover, the values of the total residual variance are ordered according to the number of data table, that is this variance is highest for T_1 and it is smallest for T_4. This means that the residual variances reflect the separability of particular data blocks.

Table 4.2 shows relative residual variances for particular clusters, which are the ratios:

$$\frac{var_{res}(Y_i^r : X_i^r)}{var_{res}(Y_0^r : X_0^r)}, \quad i = 1, ..., n$$

Once more tables $T_1, ..., T_4$ are ordered, this time the order is determined by the maximal ratios magnitudes. The smallest maximal ratio corresponds to table T_1, the highest correspond to table T_4. This observation provides a simple rule which helps to determine whether the obtained by the GCCA clusters are the natural or they are just the effects of discretization If the relative variances (at least several of them) approach one then the clusters do not correspond to natural clusters. *There are two possible reasons: there are not natural clusters in the data or the number of clusters is not properly chosen.*

TABLE 4.1. Residual variances for tables $T_1, ..., T_4$

Cluster	T_1	T_2	T_3	T_4
1	$0.470E-05$	$1.483E-05$	$0.524E-05$	$0.232E-05$
2	$0.931E-05$	$0.143E-05$	$0.947E-05$	$2.063E-05$
3	$0.605E-05$	$1.765E-05$	$0.757E-05$	$1.912E-05$
All clusters	$116.841E-05$	$52.332E-05$	$6.735E-05$	$2.175E-05$

TABLE 4.2. Relative residual variances for tables $T_1, ..., T_4$

Cluster	T_1	T_2	T_3	T_4
1	0.004	0.028	0.078	0.107
2	0.008	0.003	0.141	0.948
3	0.005	0.034	0.112	0.879

Undoubtedly ideal examples of data without a cluster structure are samples drawn from uniform distributions. Such data are often used in the context of determination of cluster numbers (cf. for example [22]). Let us

http://rbc.ipipan.waw.pl

assume once more that the variables' range is the interval $[0, 1]$. In the simulation experiment $m \times k$ data tables are randomly generated, each value is drawn according to the distribution uniform on $[0, 1]$. Figures 4.6, 4.7 and 4.8 show the regression points calculated for rows of three such tables. These tables have the same number of rows $m = 300$, the number of columns are 100, 30 and 2 respectively.

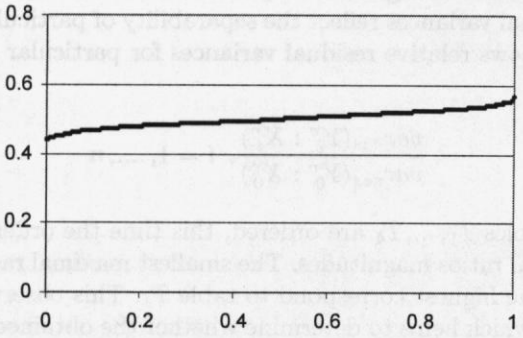


FIGURE 4.6. Regression points calculated for rows of 300 × 100 data table

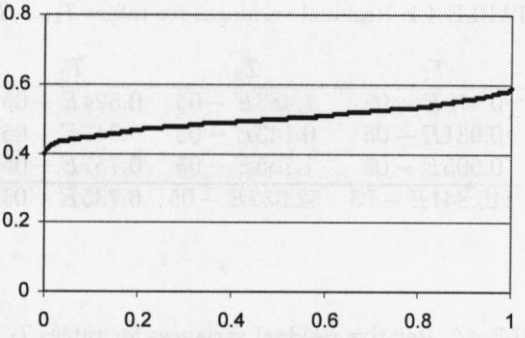


FIGURE 4.7. Regression points calculated for rows of 300 × 30 data table

The linearity of the regression points for the two first tables are clearly visible, what is confirmed by the residual variances equal to $1.53095E-05$ and $3.58333E-05$, respectively. For the third table the regression points are far from linearity; this agrees with the much higher value of residual variance $64.6304E-05$. The corresponding values of $\rho^*_{\max}$ are equal to

http://rbc.ipipan.waw.pl

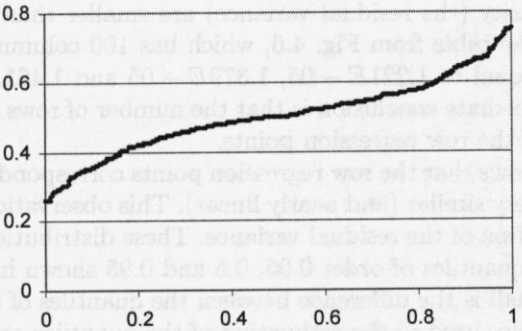


FIGURE 4.8. The regression points calculated for rows of 300 × 2 data table

0.086, 0.130, 0.350 respectively. The trend is clear: the linearity of regression points and independence between rows and columns (measured by $\rho^*_{\max}$) increase while the number of columns increases.

This effect of columns' numbers is confirmed by further simulation experiments. This time 200 data tables are generated analogously but they have different sizes: 30 rows and 300 columns. Figure 4.9 shows the regression points calculated for the rows of three tables; the first two tables are randomly chosen from the generated sample, the third corresponds to the maximal value of the residual variance.

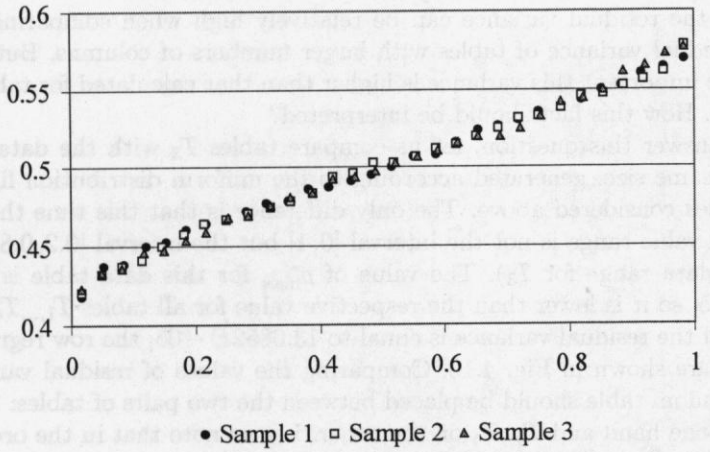


FIGURE 4.9. Regression points calculated for rows of 30 × 300 data table

http://rbc.ipipan.waw.pl

The row regression points of these tables are almost linear. The deviation from linearity (the residual variance) are smaller than the analogous deviation for the table from Fig. 4.6, which has 100 columns, the residual variances are equal to $1.221E-05$, $1.379E-05$ and $1.461E-05$ respectively. The immediate conclusion is that the number of rows does not affect the linearity of the row regression points.

Figure 4.9 shows that the row regression points corresponding to particular tables are very similar (and nearly linear). This observation is confirmed by the distribution of the residual variance. These distributions are characterized by the quantiles of order 0.05, 0.5 and 0.95 shown in Tab. 4.3. Let us note how small is the difference between the quantiles of order 0.05 and 0.95. The samples used to the estimation of the quantiles are not independent, the sample of size s_1+s_2 includes the sample of size s_2. This provide an additional evaluation of the stability of this distribution. The minimal value of the residual variance in the largest sample is equal to $1.175E-05$, the maximal is equal to $1.461E-05$. All these calculations confirm that the deviation from linearity depends on the size of data tables.

TABLE 4.3. Quantile of residual variances

Sample sizes	Quantiles of order 0.05	0.5	0.95
100	$1.223E-05$	$1.281E-05$	$1.374E-05$
150	$1.211E-05$	$1.278E-05$	$1.374E-05$
200	$1.200E-05$	$1.278E-05$	$1.376E-05$

Now let us return to the analysis of $m \times k$ tables where $k=2$. As it was shown the residual variance can be relatively high when comparing with the residual variance of tables with larger numbers of columns. But what is more important this variance is higher than that calculated for tables T_3 and T_4. How this fact should be interpreted?

To answer this question, let us compare tables T_3 with the data table of the same size, generated according to the uniform distribution like the examples considered above. The only difference is that this time the considered value range is not the interval $[0,1]$ but the interval $[0.2,0.8]$ (this is the data range for T_3). The value of $\rho^*_{\max}$ for this data table is equal to 0.215, so it is lower than the respective value for all tables $T_1,..,T_4$. The value of the residual variance is equal to $13.0852E-05$, the row regression points are shown in Fig. 4.10. Comparing the values of residual variance, this random table should be placed between the two pairs of tables: T_1, T_2 on the one hand and T_3, T_4 on the other. Let us note that in the ordering according to the decreasing values of $\rho^*_{\max}$ this table occupies the last place; hence the ordering according to the linearity and the ordering according to dependence can be different.

http://rbc.ipipan.waw.pl

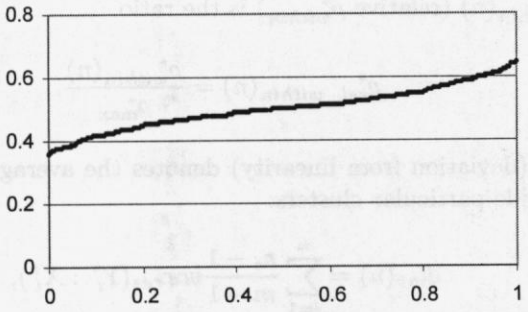


FIGURE 4.10. Regression points calculated for rows of 300×2 data table restricted to the range $[0.2, 0.8]$

The analysis of the various examples of data tables presented above implies a classification rule which helps to determine whether data tables have natural clusters. Let us choose a statistics which characterize the distribution of the residual variance. If the value of residual variance calculated for the data is smaller than the value of this statistics then we assume that there is no cluster structure in the data table. The estimation of this distribution must be based on randomly generated data tables which have the identical sizes as the initial table and which are generated according to the uniform distribution with the same value range as the original data. If the value of the residual variance for the random table discussed above (see Fig. 4.10) will be used as a rough characteristic of the distribution of the residual variance (as it was shown above such distributons have small diversity) then this rule indicates that table T_3 and T_4 have no cluster structures.

4.2.3 Determination of the proper number of clusters

In the previous sections we tried to evaluate whether the chosen numer of clusters is good and which measures are usefull in this task. Now the results will be exploited for determination of the proper number of clusters. Let us consider once more the data tables T_1 and T_4. Tables 4.4 and 4.5 show several indices which help to compare the results of the GCCA clusterings for different number of clusters obtained for tables T_1 and T_4 (as previously clusters of rows are considered). Let us introduce the following notation:

- $\rho^*_{within}(n)$ denotes the diffference between $\rho^*_{\max}$ for the initial data and $\rho^*_{agrr}(n)$ which is equal to ρ^* calculated for the table aggregated (summed) in each of n clusters according to the GCCA clustering. If

http://rbc.ipipan.waw.pl

number of clusters is equal to 0 then ρ^*_{within} is equal to $\rho^*_{\max}$,

- $\rho^*_{rel\_within}(n)$ (relative ρ^*_{within}) is the ratio

$$\rho^*_{rel\_within}(n) = \frac{\rho^*_{within}(n)}{\rho^*_{\max}}$$

- $dl_{av}(n)$ (deviation from linearity) denotes the average residual variance inside particular clusters:

$$dl_{av}(n) = \sum_{i=1}^{n} \frac{n_i - 1}{m - 1} var_{res}(Y_i^r : X_i^r),$$

where n denotes the number of clusters, n_i is the size of i-th cluster, $m = \sum_{i=1}^{n} n_i$ is a number of rows in the data table, the residual variance in i-th cluster $var_{res}(Y_i^r : X_i^r)$ is given by formula (4.4). If number of clusters is equal to 0 then dl_{av} is equal to the residual variance for the all data.

- $dl_{rel}(n)$ (relative deviation from linearity) is the ratio of deviation from linearity to the total residual variance

$$dl_{rel}(n) = \frac{dl_{av}(n)}{dl_{av}(0)}$$

Comparison of the values of these four measures calculated for both tables T_1 and T_4 reveals two facts. The measures of clusters' heterogeneity - ρ^*_{within}, relative or simple - diminish while the number of clusters increases. This effect is typical for various clustering methods (cf. [22]) which exploite intracluster homogeneity (heterogeneity) measures. Let us note that the diminishing trends for both data tables are very similar, however these values decrease faster for T_1 than for T_4. Moreover the gap between the values corresponding to two and three clusters are much greater for T_1 than that for T_4.

TABLE 4.4. Indices characterizing data table T_1

Number of of clusters	ρ^*_{within}	$\rho^*_{rel\_within}$	dl_{av}	dl_{rel}
	0.568	1	$116.841E-05$	1
2	0.145	0.2553	$116.424E-05$	0.9964
3	0.026	0.0458	$0.664E-05$	0.0057
4	0.019	0.0335	$0.610E-05$	0.0052
5	0.012	0.0211	$0.535E-05$	0.0046
6	0.007	0.0123	$0.478E-05$	0.0041

http://rbc.ipipan.waw.pl

TABLE 4.5. Indices characterizing data table T_4

Number of of clusters	ρ^*_{within}	$\rho^*_{rel\_within}$	dl_{av}	dl_{rel}
	0.306	1	$2.175E-05$	1
2	0.080	0.2614	$1.314E-05$	0.6043
3	0.033	0.1078	$1.392E-05$	0.6399
4	0.021	0.0686	$1.144E-05$	0.5259
5	0.012	0.0392	$0.870E-05$	0.4001
6	0.009	0.0294	$0.474E-05$	0.2178

The measures of deviation from linearity retain the similar diminishing trend while number of clusters increases, however there is one exception from it for table T_4. For table T_1 there is a huge gap between the results for two and three clusters. For greater numbers of clusters the values of these measures are low and decrease very slowly. For table T_4 the analogous values are much greater and there is no a large gap betwen two and three clusters. On the contrary the deviation from linearity is slightly greater for three clusters than that calculated for two clusters.

Tables 4.6 and 4.7 show the sizes of particular clusters obtained by the GCCA when various numbers of clusters were assumed. As all clusters are formed by adjacent intervals of the optimal GCCA orderings, these sizes characterize the memberships for particular clusters

When the number of clusters is 2 or 3 the clusters of both tables have very similar sizes. For the greater number of clusters the clustering procedure generates totaly different results. For table T_1 each of the three cluster (as it was shown before, the natural clusters corresponding to the different generators are restored) is partitioned into further subclusters. The analogous clusters for table T_4 are totally different; they have similar sizes and there are not stable boundaries which remain unchanged when the GCCA is applied with different numbers of clusters. This comparison provides additional arguments that the proper number of clusters is 3 for T_1, but for T_4 there is not such a natural choice.

TABLE 4.6. Sizes of clusters for T_1

Clusters	Numbers of clusters				
	2	3	4	5	6
1	147	100	100	53	53
2	153	100	100	47	47
3		100	47	100	52
4			53	47	48
5				53	47
6					53

http://rbc.ipipan.waw.pl

TABLE 4.7. Sizes of clusters for T_4

Clusters	Numbers of clusters				
	2	3	4	5	6
1	153	100	86	60	58
2	147	97	78	40	63
3		103	65	84	54
4			71	61	43
5				55	49
6					44

It should be emphasized that very similar results as presented above were obtained for other random data tables generated by the same generators as the previously used. In other words *the general trends are stable (not random) and depend on the used generators.*

The analyses presented above lead to the formulation of several rules, which help to determine the proper number of clusters. The most important conclusion of these results as well as of the many others obtained for random or real data tables is that the choice should be based on several measures, which describe various characteristics of cluster structures present in data tables. This agrees with recent trends in classification, where several criterions are applied simultaneously instead of seeking the best one (cf [23], [6]). In the GCCA framework the following measures should be taken into consideration:

- $\rho^*_{\max}$ which measures the strenght of monotone structure present in the data (formally it is the strenght of dependence between row and column variables of the data table).
  According to the definition, the stronger diversity between rows (columns), the higher value of $\rho^*_{\max}$. On the other hand the values of $\rho^*_{\max}$ are influenced by various factors, unrelated to cluster structures. For example low values of $\rho^*_{\max}$ may be caused by the lack of monotone trend as well as by the presence of two different monotone trends (this problem is discussed in [10], [8]). Undoubtedly, $\rho^*_{\max}$ alone can not express complexity of data structures

- $\rho^*_{within}(n)$ and $\rho^*_{rel\_within}(n)$; they measure how much the data differ from the set of natural clusters - hence they also measure intracluster homogeneity.
  Similarly to the value of $\rho^*_{\max}$ also the values of ρ^*_{within} are biased by factors unrelated to cluster structures. Therefore it is better to use the relative measure which is more robust.

- $dl_{av}(n)$ and $dl_{rel}(n)$ - the measures of deviation from linearity of the regression points. In the GCCA context they are also intracluster homogeneity measures.

http://rbc.ipipan.waw.pl

Many well known clustering procedures use some kind of intraclaster homogeneity measures, but usually only one is used. Usually the values of this measure decrease while the number of clusters increases. Hence the choice of the proper number of clusters is based on a big gap between respective values of the homogeneity measure calculated for two nearest cluster numbers. The problem is how to determine whether the gap is big or not. The recent propositions suggest to compare the value of the homogeneity measure with its counterpart calculated for random data (cf [22], [31]). This leads to another problems: which distribution should be used as a reference, which statistics (mean, median, other quantile etc) should be chosen to characterize this distribution. Uniform and normal distributions are the most popular in the context of cluster determination. Whatever is the choice, the next step consists in testing consequtive pairs of numbers $(n, n+1)$. If the gap between the measures calculated for n-th and $(n+1)$-th cluster is bigger than those calculated for the random data then $n+1$ are recognized as the "true" number of clusters.

In the GCCA framework the first homogeneity measure $(\rho^*_{rel\_within})$ is monotone with respect to the increasing numbers of clusters. The second (dl_{rel}) does not posses this property, nevertheless the gap rule is also applicable here. In agreement with the concept of natural clusters introduced above, the natural choice of the reference point is a uniform distribution. Undoubtedly uniform distributions are the best models for data without cluster structures, whereas multinormal distributions correspond to the structures which is characterized by one big cluster and many single points. Moreover the one of the few assumption in the GCCA is that data values should be nonnegative, hence samples from normal distributions need to be transformed into the form admissible to the GCCA. An additional argument that multinormal distributions seem not appropriate as universal references is that they are rare in questionaire data which are common in sociology, psychology or market research.

The proposed multicriterial procedure which helps to determine the proper number of clusters is the following.

- Let

$$g_{\rho^*}(n-1,n) = \begin{cases} \rho^*_{rel\_within}(n-1) - \rho^*_{rel\_within}(n) & \text{for } n > 2 \\ \rho^*_{rel\_within}(n) & \text{for } n = 2 \end{cases}$$

and

$$g_{dl}(n-1,n) = \begin{cases} dl_{rel}(n-1) - dl_{rel}(n) & \text{for } n > 2 \\ dl_{rel}(n) & \text{for } n = 2 \end{cases}$$

denote gaps calculated for the homogeneity measures. Let $g^{\text{rand}}_{\rho^*}(n-1,n)$ and $g^{\text{rand}}_{dl}(n-1,n)$ denote chosen statistics which characterize

http://rbc.ipipan.waw.pl

the distributions of respective gaps. These statistics are estimated on the sample drawn from the uniform distribution with the same value range as the initial data table.

- The best candidate is the smallest n such that the both gaps are big, that is the gaps are greater than their random counterparts $g_{\rho^*}^{\text{rand}}(n-1,n)$ and $g_{dl}^{\text{rand}}(n-1,n)$ (greater than the values of a chosen statistic which characterize the respective distributions). If the clustering for $n+1$ clusters divides only one cluster from those obtained for n clusters' solution and the others remain almost unchanged, then this is an additional confirmation that n is a good choice.

- If there is no such n in the considered sequence of numbers and $g_{dl}(n-1,n)$ are all small then the data table is close to ideally regular (paraboloidal) and any number of clusters can be chosen. If $g_{dl}(n-1,n)$ are relatively big than the data may be regular but not paraboloidal type. Hence the next step is to check whether the data table is regular by comparing the correlation curve with the ideally regular family of curves introduced above. If the answer is negative - the curve is not similar to any member of this family, than the definition of ideal regularity should be used and various number of clusters $n = n_1 n_2$ should be tested. Let us note that if each of n_1 clusters is optimally partitioned into n_2 clusters and the result remains optimal for the number of cluster $n = n_1 n_2$, but this occurs only for particular numbers n_1 and n_2 than a hierarchy is present in the data, not ideal regularity.

Now let us use this set of rules to the data from table T_3 (Sec. 4). Table 4.8 shows the value of both gaps g_{ρ^*} and g_{dl} calculated for a few numbers of clusters. As the value range for both variables from table T_3 is the same and equal to the interval $[0.2, 0.8]$, the distribution uniform on this interval is used as the reference. A sample of the same size as in T_3 is drawn and used for estimation of distributions of both gaps. Table 4.9 includes quantiles of order 0.1, 0.5 and 0.9 of these distributions.

TABLE 4.8. Two kinds of gaps calculated for table T_3

n	$g_{\rho^*}(n-1,n)$	$g_{dl}(n-1,n)$
2	0.255	0.890
3	0.171	0.818
4	0.025	-0.241
5	0.021	0.081

The value of $g_{\rho^*}(2,3)$ attains the level of quantile of order 0.9, but the value of gap $g_{dl}(2,3)$ is undoubtedly big. As the respective values of gaps

http://rbc.ipipan.waw.pl

TABLE 4.9. Quantiles for both gaps

n	$g_{\rho^*}(n-1,n)$ quantiles of order			$g_{dl}(n-1,n)$ quantiles of order		
	0.1	0.5	0.9	0.1	0.5	0.9
2	0.285	0.292	0.302	0.541	0.650	0.719
3	0.154	0.162	0.172	0.438	0.511	0.583
4	0.052	0.056	0.060	-0.044	0.018	0.116
5	0.023	0.027	0.030	0.010	0.024	0.069

calculated for other number of clusters are much worse comparing with the respective quantiles then the natural conclusion is that three clusters is acceptable for table T_3, however the values of g_{ρ^*} indicates that this clustering structure is not very strong. Taking into account the inference which is based on the total residual variance (presented in the previous section) and summing up all results one may say that two numbers of clusters can be chosen for table T_3: $n = 0$ (no clusters) or $n = 3$.

http://rbc.ipipan.waw.pl

5

GCCA versus other clustering methods

There are two reasons of serious difficulties with comparison of any clustering method with the other ones. The first is technical and lies in an enormous number of clustering methods. The second is methodological and it consists in the lack of one generally accepted definition of clusters. In consequence there is a variety of existing definitions and even greater variety of clustering methods. The algorithms which are based on different definitions are hardly comparable; the area of comparisons should be practicaly limited to the methods with similar underlying assumptions. The determination of groups of similar clustering methods leads to a typology of clustering methods where the method - subject of comparison should be properly located.

Unfortunately the same methodological reason caused that there are several taxonomies or more general typologies which are discussed and criticised by specialists (cf [18], [27]). Let us remind that the term "taxonomy" is usually understood as a hierarchical classification of a given set of objects (in our case this is the set of clustering procedures). So taxonomies can be considered as the results of hierarchical clusterings.

The typologies of clustering procedures exploites various features of the clustering methods as classification criterions. As this paper is not intendended to provide a review of typologies, only a few of them, the best known, will be considered here. Let us start from the classic taxomomy shown in Fig. 4.10 (cf. [27])

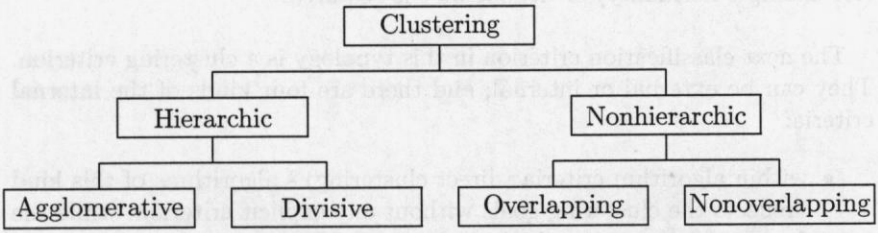


FIGURE 5.1. Classical taxonomy of clustering methods

In this taxonomy the GCCA belongs to the group of nonhierarchic methods which generate nonoverlapping clusters. Thanks to the rapid develop-

http://rbc.ipipan.waw.pl

ment of clustering methods many of the recent ones do not fit into this classification. The newer and more general typology, proposed in [27], takes into account three features which form the base of the classification:

- kind of input data,
- kind of criterion,
- kind of output cluster structure.

This typology takes into account only two-dimensional input data tables - identically as the GCCA. Its author distinguishes the three kinds of input data:

- column conditional data (i.e. object - variable tables),
- comparable data (i.e. proximity/dissimilarity tables including distance and correlation matrices, also object - variable tables but all values across the table should be comparable),
- aggregable data (i.e. contingency data or category to category data)

Most clustering methods are designed for one kind of input data, what's worse many of them are meant only for very specific data (for instance Boolean variables). In this aspect the GCCA is absolutely different; its basic underlying assumption is universality; as it was shown in this paper and also in [8], this method is applicable for any kind of input data from those considered in this typology. The only exception are the data which consist of the values of nominal variables, then each variable category should be expressed by the respective separate variables. It should be emphasized that the same procedure can be applied to various kinds of data, instead of developing many narrow-specialized procedures. This universality is attained thanks to the lack of assumptions - no special forms of distributions (for example normality) or dependency is required.

The next classification criterion in this typology is a clustering criterion. They can be external or internal; and there are four kinds of the internal criteria:

- within algorithm criteria (direct clustering) - algorithms of this kind "reflects the clustering goals without any explicit criterion, while the formal criteria are used within its particular iterative steps",
- optimization (there is a variety of optimization criteria, strongly related to the underlying definitions of clusters),
- definition (explicit definitions of clusters; this approach involves concepts defined to fit perfectly into any feasible data),

http://rbc.ipipan.waw.pl

- consensus ("in this approach a classification method is considered as a mapping $F : D \to C$, where D is the set of all feasible data and C is the set of all classification structures of a given kind"; "mapping F is reffered to as a consensus function if it satisfies some natural properties").

Since the grade correlation coefficient is maximized in the GCCA, the method belongs to the second category. This coefficient expresses within-cluster diversity as well as differences among clusters (cf. Sec. 2.2.2 and Sec. 3.1). In this aspect the GCCA resembles the well known k-means method. The main difference between these two methods consists in diversity measures (measures of clusters' homogeneity) and consequently in optimization criteria.

Let us note that however combinatorial problems are unseparable from cluster analysis (cf. [24], [26]), the optimization methods are especially prone to generate problems unfeasible for large data tables. In the case of GCCA this combinatorial problem is solved by decomposition into two optimization procedures (the GCA and the grade clustering based on the GCA), which have not this drawback..

The methods which exploited the formalism of the classic correspondence analysis are known in the clustering literature (cf. [16], [17]). They belong to the class of hierarchical clustering methods and use the formulas specific to this correspondence analysis as the local criteria of procedures' steps. Therefore these methods belong to the other class of this typology than the GCCA - to the class "within algorithm criteria". Other difference between these methods and the GCCA follows from the difference between the classic and the grade correspondence analysis. The classic approach exploites decomposition of χ^2 statistic, the grade approach is based on the concentration indices and the grade correlation coefficient (cf. Chapter 2).

The following major categories of cluster structures are considered in the discussed typology:

- subset,
- partition,
- hierarchy,
- association structure (this group includes also ordered partitions),
- biclustering structure (they correspond to two-mode clusterings, that is when the rows and columns of data table are of different nature),
- nonstandard clusters (i.e. fuzzy subsets and partitions, overlapping clusters, extended hierarchies like pyramids, standard point typology - for instance Kohonen maps),

http://rbc.ipipan.waw.pl

- concept (or conceptual cluster or classification tree or decision tree),
- separating surface (discriminant function),
- neural network,
- probabilistic distribution.

For most clustering methods these groups of output structures are disjoint, in the case of the GCCA they overlap. As it was shown above the clusters generated by the GCCA are determined by the partitions of the set of rows (or columns) of the input data table. As these partitions possess an order structure the GCCA results should be classified also as an association structure.

The ordered data structures arouses much interest from many years.(cf. [1]). Some specialists even believe that "there are the signs of a mature discipline including its own specialized journal" (cf. [1]). A particularly great attention is focused on one-dimensional orderings (usually the term seriation is used in this context). "In the last several decades the methods of seriation have been developed most aggresively by archeologists" (cf. [1]).

The determination of GCCA clusters is based on the grade regression functions, these functions can be treated as the discriminat functions, hence the GCCA output structure can be also considered as a separating surface.

The variant of the GCCA (so called double clustering - cf. [11], [8]) is specially designed for two-way clustering that is for the simultaneous clustering of rows and columns of two-dimensional data tables. "Starting with the pioneering work of Hartigan (cf. [20], [21]) who significally contributed to the development of the simulaneous clustering domain both conceptually and algorithmically, and with some decision - theoretical investigation by Bock (cf. [4]), a broad range of simultaneous clustering methods has been developed by various authors. In the past, a few attempts have been made to structure the whole of the resulting methods but a comprehensive overview of the domain is still lacking, and the taxonomic efforts in the area has been criticized (cf. [28])". It is hard to disagree with this quotation from [32], which characterizes well the problems with clustering typologies. Its authors propose their own typology of the two-way clustering methods (cf. [32]). Also in this case several criteria are chosen which help to structure a variety of methods:

- kind of input data
  - case by variable type, categorical predictor type, proximity type
  - column conditional, row conditional, matrix conditional
- kind of clusters
  - elements of clusters (Cartesian products of row and column clusters

http://rbc.ipipan.waw.pl

or not)
- set-theoretical relations between clusters (partitions, nested clusterings, overlapping clusterings)

- level of modeling and optimization
  - procedural level
  - deterministic level
  - stochastic level (fixed-partition, random-partition)

The two-way GCCA procedure does not fit ideally into this classification, thanks mostly to its applicability to the various types of input data (the two-way GCCA retains all properties of the one-way procedure which are discussed above). There are no problems with the classification of GCCA according to the second criterion, only partitions and only Cartesian products of row and column clusters are taken into account.

http://rbc.ipipan.waw.pl

References

[1] Arabie P, Hubert L.J. (1996) An overview of combinatorial data analysis. In: Arabie P, Hubert L. J, De Soete G (Eds) Clustering and Classification. World Scientific Publ., River Edge, NJ, 5-63.

[2] Bock H-H. (2002) Clustering methods: from classical models to new approaches. Statistics in Transition, Vol. 5 No. 5, 725–758.

[3] Bock H-H. (1996) Probability models and hypotheses testing in partitioning cluster analysis. In: Arabie P, Hubert L.J, De Soete G. (Eds) Clustering and Classification. World Scientific Publ., River Edge, NJ, 377-453.

[4] Bock H-H. (1968) Stochastiche Modelle fur die einfache und doppelte Klassification von Normalverteilten Beobachtungen. Dissertation, University of Freiburg.

[5] Bonner R. (1964) On some clustering techniques. IBM Journal of Research and Development, 8, 22-32.

[6] Breiman L. (1996) Bagging predictors. Machine Learning 24(2), 123-140.

[7] Ciok A. (2004) Discretization and regularity. In: Kowalczyk T, Pleszczyńska E, Ruland F. (Eds) Grade models and methods for data analysis: with applications for the analysis of data population. Studies in Fuzziness and Soft Computing, vol. 151, Springer-Verlag, Berlin Heidelberg New York, 167-184.

[8] Ciok A. (2004) Cluster analysis based on GCA. In: Kowalczyk T, Pleszczyńska E, Ruland F. (Eds) Grade models and methods for data analysis: with applications for the analysis of data population. Studies in Fuzziness and Soft Computing, vol. 151, Springer-Verlag, Berlin Heidelberg New York, 297-324.

[9] Ciok A. (2004) Regularity and the number of clusters. In: Kowalczyk T, Pleszczyńska E, Ruland F. (Eds) Grade models and methods for data analysis: with applications for the analysis of data population. Studies in Fuzziness and Soft Computing, vol. 151, Springer-Verlag, Berlin Heidelberg New York, 325-338.

http://rbc.ipipan.waw.pl

[10] Ciok A. (2002) Grade correspondence-cluster analysis applied to separate components of reversely regular mixtures. In: Jajuga K, Sokołowski A, Bock H-H. (Eds), Classification, Clustering and Data Analysis, Recent Advances and Applications, Springer, 211–218.

[11] Ciok A. (2000) Double versus optimal grade clusterings. In: Kiers H.A.L, Rasson J-P, Groenen P.J.F, Schader M. (Eds.), Data Analysis, Classification, and Related Methods, Springer, 41-46.

[12] Ciok A. (1998) Discretization as a tool in cluster analysis. In: Rizzi A, Vichi M, Bock H-H. (Eds.), Advances in Data Science and Classification, Proceedings of the 6th Conference of the International Federation of Classification Societies (IFCS-98), Rome, July 21-24 1998, Springer, 349-354.

[13] Ciok A, Kowalczyk T, Pleszczyńska E, Szczesny W. (1995) Algorithms of grade correspondence-cluster analysis. The Collected Papers of Theoretical and Applied Computer Science vol. 7, no. 1-4, 5-22.

[14] Everitt B.S. (1993) Cluster analysis. Edward Arnold, London.

[15] Gordon A.D. (1996) Hierarchical classification. In: Arabie P, Hubert L.J, De Soete G. (Eds) Clustering and Classification, World Scientific Publ., River Edge, NJ, 65-121.

[16] Greenacre M.J. (1984) Theory and application of correspondence analysis. Academic Press.

[17] Greenacre M.J. (1988) Clustering the rows and columns of a contingency table. Journal of Classification 5, 39-51.

[18] Estivill-Castro V. (2002) Why so many clustring algorithms - a position paper. SIGKDD Explorations, vol. 4, issue 1, 65-75.

[19] Han J, Kamber M. (2000) Data Mining: concepts and techniques. Morgan Kaufmann Publishers, San Marteo, CA.

[20] Hartigan J.A. (1975) Clustering algorithms, J. Wiley, New York.

[21] Hartigan J.A. (1972) Direct clustering of a data matrix, Journal of the American Statistical Association, 67, 123-129.

[22] Hastie T, Tibshirani R, Walther G. (2000) Estimating the number of data clusters via the Gap statistics, Technical Report, also published in JRSSB 63, 411-423.

[23] Hothorn T, Lausen B. (2002) Bagging tree classifiers for glaucoma diagnosis. In: Hardle W, Ronz B. (Eds) COMPSTAT Preceedings in Computational Statistics, 15th Symposium Held in Berlin, 2002. Physica-Verlag.

http://rbc.ipipan.waw.pl

[24] Hubert L, Arabie P, Meulman J. (2001) Combinatorial data analysis. Optimization by dynamic programming. SIAM, Philadephia.

[25] Kowalczyk T, Pleszczyńska E, Ruland F. (2004) (Eds) Grade models and methods for data analysis: with applications for the analysis of data population. Studies in Fuzziness and Soft Computing, vol. 151, Springer-Verlag, Berlin Heidelberg New York.

[26] Mirkin B, Muchnik I. (1998) Combinatorial optimization in clustering. In: Du D-Z, Pardalos P.M. (Eds) Handbook of combinatorial optimization. Kluwer Academic Publishers, Dordrecht, Boston, Londyn, 261-329.

[27] Mirkin B. (1996) Mathematical classification and clustering. Nonconvex optimization and clustering. Kluwer Academic Publishers, Dordrecht, Boston, Londyn.

[28] Mirkin B, Arabie P, Hubert L.J. (1995) Additive two-mode clustering: the error variance approach revisited. Journal of Classification, 12, 243-263.

[29] Nelsen R.B. (1999) An Introduction to Copulas. Lecture Notes in Statistics No. 139, Springer-Verlag, New York.

[30] Schriever B.F. (1985) Order Dependence. Ph.D. dissertation, Centrum voor wiskunde en informatika, Vrije Universiteit te Amsterdam.

[31] Sugar C.A, James G.M. (2003) Finding the number of clusters in a data set: An information theoretic approach, Journal of the American Statistical Association 98, 750-763.

[32] Van Mechelen I, Bock H-H, De Boeck P. (2004) Two-mode clustering methods: A structured overview. Statistical Methods in Medical Research (to appear).

http://rbc.ipipan.waw.pl