Generalization Method when Manipulating Relational Databases

V. Cariou*,** L. Billard***

*LUNAM University, Oniris, Sensometrics and Chemometrics Laboratory,
Nantes Cedex F-44322 France
veronique.cariou@oniris-nantes.fr
**INRA, Nantes, F-44307, France
***Department of Statistics, University of Georgia, Athens, GA 30602, USA
lynne@stat.uga.edu

Abstract. Contemporary computers generate massive datasets. One way to handle these data is to aggregate them into smaller datasets (with the aggregation criteria dictated by meaningful scientific questions of interest). This paper focuses on aggregations that produce interval datasets. Algorithms are introduced both to build intervals which are typically homogeneous, and to test that such homogeneity pertains. They also test whether or not observations across the resulting intervals are mixtures of uniform distributions rather than the desired single distribution. These include consideration of outlier observations. The methods are illustrated on two datasets.

1 Introduction

Contemporary datasets can be enormous, too large for standard analytic methods to be used directly on the very same computers generating the datasets themselves. Thus, some form of data manipulation is needed in order to transform the original dataset into one that is more manageable for appropriate analyses.

There are many approaches that have been proposed to address this issue. Most have different strengths, most are more applicable to some settings and data types than others; all are useful. Data mining as a broadly based methodology identifies patterns in the dataset, and then delves more deeply into the part of the data responsible for those patterns, perhaps as one operation or perhaps by pattern type. See, e.g., Hand et al. (2001). Another approach is to take a sample of the dataset. One such technique is data squashing whereby the original data are sorted into clusters of like characteristics, with a “representative" pseudo-sample drawn from each cluster. The analysis is conducted on this scaled down sampled dataset. See, e.g., DuMouchel et al. (1999).

A third broad approach developed in the literature deals with aggregating data points, where the criteria for any particular aggregation vary depending on the nature of the scientific questions being asked. The resultant dataset then consists of lists, intervals, histograms, and the like, and fall under the general heading of symbolic data. See, e.g., Bock and Diday (2000)

and Billard and Diday (2006). While data squashing and symbolic data both arise from aggregation of an original dataset as a method to obtain one of a more manageable size, in data squashing a sample is drawn which hopefully is representative while for symbolic data typically all the data are retained.

The focus of this paper is on aggregating a large dataset into interval data, and how to handle rare events. Let us consider a classical data table, denoted by $\boldsymbol{X}$, where x_{ij} corresponds to the value observed on the unit i for the variable X_j. Thus, for given aggregation criteria, observations for each variable X_j, say, for each category k, become intervals $d_k = [a_k, b_k]$ where

$$a_k = \min_i \{x_{ij}\}, \quad b_k = \max_i \{x_{ij}\}$$

where $i = 1, \ldots, n_k$ are the individual observations making up that k^{th} category. Consider the particular case that a variable X_j takes values $\{17, 86, 82, 88, 90\}$. The usual aggregation into intervals gives $d_k = [17, 90]$. Even for only five observations, this interval is not as representative of the data as desired; if there are hundreds or thousands or, ... observations in the 80 to 90 range along with the outlier 17 value, then a more representative interval might be [80,90] rather than [17,90]. Similar arguments apply when aggregating categorical values.

The present work proposes to apply a generalization operator g and an associated reduction algorithm which formalize the process so as to aggregate into intervals which more closely reflect the original dataset. Criteria for selecting the bounds are established. In the example above for thousands of observations, intuition may confirm the selection of the [80,90] interval; but what about the case in which the value 17 is replaced by a value of 77? Statistical tests to affirm the validity of the resulting intervals are also considered. That is, we establish a formal process to tell us how and when are we justified in taking the $[80, 90]$ interval instead of the $[17, 90]$ interval. This work is presented in the context of the formulation of interval data. Since histogram data are in effect a weighted mixture of (sub)intervals, the same ideas can be extended to these data; likewise for categorical data. Note this generalization operator has been widely used within the symbolic data analysis area (see, e.g., Ichino and Yagushi, 1994, or Esposito, Malerba and Tamma, 2000, for the computation of dissimilarity measures between symbolic objects).

The first approach is primarily concerned with data aggregation where the notion of being cognizant of counter-examples is not relevant; see, e.g., Han et al. (1997). An example is when summarizing data in the presence of taxonomies. The second approach is more useful when interpreting the results of an analysis of the data, such as when factorial analyses or partitioning methodologies are applied.

In an unsupervised framework, we provide a specialized method which provides for the removal of atypical (rare) values with what is called the reduction algorithm.

The generalization and associated criteria are introduced in Section 2. In Section 3, the specialization method is developed; this section includes statistical tests relating to the adequacy or not of the method. It consists of making a balance between a concise description associated with the symbolic object and its capability to cover a great amount of the corresponding observations (coverage measure). The coverage measure has also been applied within a hierarchical agglomerative clustering (Esposito and d'Amato, 2007). Dealing with continuous data, a discretization based on a divisive partitioning is proposed. The underlying hypothesis considers the distribution of observations within each interval as a uniform one. We note that discretization issues have also been discussed in the scope of association rules mining in a number of

outlets, e.g., Ludl and Widner (2000), Srikant and Agrawal (1996), Miller and Yang (1997) with a review and extensive bibliography in Bay (2001). Unlike our approach, the latter author proposes a multivariate discretization taking account of the multivariate distribution of the data. Then, in Section 4, the discretization approach is presented while the reduction algorithm is considered in Section 5. An application is given in Section 6. More details can be found in Stéphan (1998). Generalization and reduction were discussed in Stéphan et al. (2000); this paper goes deeply into the details of the way intervals are discretized and how reduction is performed on the basis of this discretization.

2 The Generalization Process

2.1 The Basic Generalization Operator

Suppose a dataset contains p random variables $\boldsymbol{X} = (X_1, \cdots, X_p)$. Suppose for a specific individual i in $\Omega = \{1, \cdots, n\}$ and variable X_j, the observed value in $\mathcal{X}_j,\ j = 1, \cdots, p$, is x_{ij}, $i = 1, \cdots, n$, $j = 1, \cdots, p$. Note that $x_{ij} = \cdot$ is a missing value, and can occur; without loss of generalization assume there are no missing values. After suitable use of appropriate SQL (e.g.) components, these n individuals are grouped into K disjoint classes or categories, where typically $K \ll n$. Basically, the class membership is built up according to the value observed on one particular attribute of the SQL query. Let $G_1, \cdots, G_K$ be the K groups of individuals and $c_1, \cdots, c_K$ be the K categories, respectively, associated with the K clusters. Let the class G_k contain n_k individuals with $G_1 \cup \cdots \cup G_K = \Omega$, and $n_1 + \cdots + n_K = n$. Let $g = (g_1, \cdots, g_p)$ be the generalization operator on the class G_k with g_j representing the coordinate-wise operator on the observations in G_k for the variable X_j.

Then, the generalization operator g on $P(\Omega)$ is defined by:

$$\begin{array}{rcl} g & : & \mathcal{P}(\Omega) \to D = \mathcal{P}(\mathcal{X}_1) \times ... \times \mathcal{P}(\mathcal{X}_p) \\ & & g(G_k) = d_k = (d_{k1}, \cdots, d_{kj}, \cdots, d_{kp}) \end{array}$$

where for $j = 1, \cdots, p$, d_{kj} equals:

$$d_{kj} = \begin{cases} [\min\limits_{i \in G_k}(x_{ij}), \max\limits_{i \in G_k}(x_{ij})], & X_j \text{ quantitative,} \\ \{v \in \mathcal{X}_j | i, i' \in G_k, x_{ij} \leq v \leq x_{i'j}\}, & X_j \text{ ordinal,} \\ \{v \in \mathcal{X}_j | i \in G_k, x_{ij} = v\}, & X_j \text{ qualitative.} \end{cases} \tag{2.1}$$

If X_j is a taxonomy variable with a tree-structured range $\mathcal{X}_j$ and hierarchy $\mathcal{H}_j$, the union operator g_j is

$$d_{kj} = \{v \in \mathcal{X}_j | i \in G_k \text{ with } x_{ij} = v; \text{ and no } i' \in G_k \text{ with } x_{ij} \preceq x_{i'j}\} \tag{2.2}$$

where $\preceq$ is the partial ordering induced by $\mathcal{H}_j$. This constraint ensures coherence relating to the taxonomy tree structure is retained.

More formally, the family of operators g on $\mathcal{P}(\Omega)$ produces a description D. A particular vector $d_k = (d_{k1}, \cdots, d_{kp}) \in D$ is the description of those observations $\{i \in G_k\}$. That is, for $G_k \in \mathcal{P}(\Omega)$ and $d_k = g(G_k)$,

$$d_{kj} = \oplus(\{x_{ij} | i \in G_k\}).$$

Here, $\oplus$ is the union operator - also called a junction operator (see Michalski and Stepp, 1983, Ichino and Yaguchi, 1994) -

$$\oplus(\{x_{ij}|i \in G_k\}) = \begin{cases} [\min\limits_{i \in G_k} x_{ij}, \max\limits_{i \in G_k} x_{ij}], & Y \text{ quantitative,} \\ \cup_{i \in G_k}\{x_{ij}\}, & \text{otherwise.} \end{cases}$$

If we consider that the data table $\boldsymbol{X}$ is represented by a relational object called `MyDataset` (which can be either a view or a table from the database), the values d_{kj} from (2.1), or (2.2), can be obtained by appropriate SQL usage. Here, we assume that the partitioning of the n tuples into K groups is materialized into `MyDataset` through an attribute called C. To illustrate, if X_j is for example a quantitative variable, the required $d_{kj} = g_j(G_k)$ is found from:

```
select min(X_j), max(X_j)
from MyDataset
where C = c_k
```

If X_j is a qualitative variable, this becomes:

```
select X_j, count (X_j)
from MyDataset
where C = c_k
group by X_j
```

For other data types, there is no direct SQL instruction which leads to such a generalization.

2.2 Partial Order Over Generalizations

Considering the generalization process, we can define a partial order, denoted by $\preceq$, over the description space D. For the sake of simplicity, we assume here that each individual description $(x_{i1}, \cdots, x_{ip})$ may be rewritten as a vector of singletons, denoted by δ_i, from D, such as: $\delta_i = (\{x_{i1}\}, \cdots, \{x_{ip}\})$.

If we denote by d_1 and d_2 two generalizations from D, $d_1 \preceq d_2$ if and only if, for all $j = 1, \cdots, p$, $d_{1j} \subseteq d_{2j}$. Then, d_2 is called a generalization of d_1 while d_1 is a specialization of d_2. Thus, the generalization operator proposed above insures the following property:

$$\text{for all } i \in G_k, \delta_i \preceq g(G_k).$$

2.3 Improving the Generalization Operator

In some cases, the generalization performed may not be representative of the initial data belonging to the group since some extreme or rare values may occur. To overcome this limitation, there is a need to make a balance between a concise description and the fact that the generalization covers all the descriptions of the individuals belonging to the group. In order to perform this balance, we propose to introduce a criterion based on the coverage concept and a quality criterion.

Before introducing quality criteria, let us recall some notions of symbolic data analysis. Our work fits into the scope of symbolic data analysis by generalizing sets of units by a set of symbolic objects called assertions. In this context, assertions are based on the descriptions obtained through the generalization operator g. Formally, an assertion, denoted by a_k, refers to the generalization d_k of the class of units G_k and corresponds to the conceptual object:

$$a_k = \wedge_j [X_j \in d_{kj}].$$

Let $\mathcal{A}$ the set of assertions. The extension of an assertion $a \in \mathcal{A}$ is defined over an element of $\mathcal{P}(\Omega)$ such that:

$$\begin{array}{rcccl} ext & : & \mathcal{A} \times \mathcal{P}(\Omega) & \rightarrow & \mathcal{P}(\Omega) \\ & & ext(a; G) & = & \{i \in G | \delta_i \preceq d\}. \end{array}$$

In the particular case of a_k which is obtained through the generalization operator g over G_k:

$$G_k \subseteq ext(a_k, \Omega) = \{i \in \Omega | \delta_i \preceq d_k\}.$$

Let us recall our first example presented in the introduction. We have seen that the aggregation of the five numerical values $\{17, 86, 82, 88, 90\}$ gives $d_k = [17, 90]$. More generally, if G_k contains an extreme value or a rare event on one or several variables, the operator g produces an interval which cannot sustain any homogeneous-type property, such as an assumption that observations across the interval are (exactly, or approximately) uniformly distributed (which assumption is required for statistical methodology developed thus far for analyses of interval data). We seek a specialization step which improves the generalization process so that the properties of the resultant interval(s) are more reflective of the true characteristics of the observations in the interval(s); likewise, for counts if X_j is a qualitative variable.

Let d'_k represent the improved generalization of G_k; the description d'_k is included in d_k. For example, given the previous example, the description of G_k is $d_{kj} = [17, 90]$, while $d'_{kj} = [80, 90]$. The nature of the improvement to the generalization, represented by d'_k, will depend on the quality criteria adopted to evaluate it.

One quality criterion is the capacity of covering. Thus, an assertion a is a good generalization of a set of individuals G if it covers correctly those individuals, where we define coverage as:

$$Rec(a, G) = card(ext(a; G))/card(G) \tag{2.3}$$

where $ext(a; G)$ is the extension of the assertion a on the set of units G and consists of those elements in G for which d generalizes their description.

Another criterion deals with homogeneity of the individuals in G. As the name suggests, a description d provides a good generalization of G if the individuals induced by d give a uniform hypercube datapoint. Here, the corresponding descriptions assume that there is a uniform distribution across the values (x_{ij}) that make up G_k, and that $X_j, X_{j'}, j \neq j'$ are independent. Procedures to test this uniformity assumption are considered in Section 4 below.

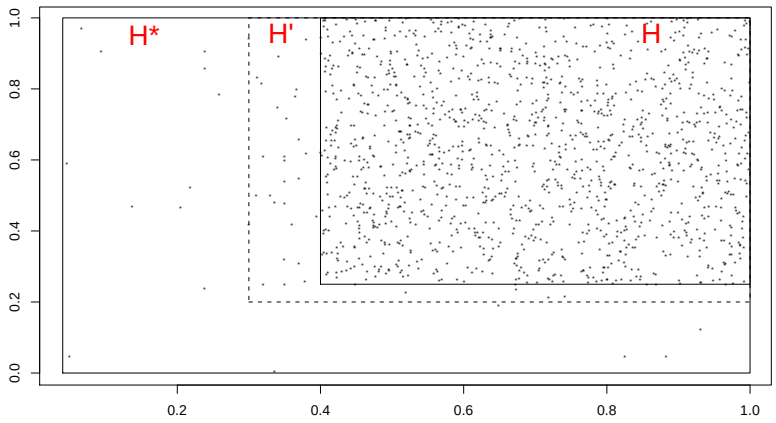

FIG. 1 – *Application of specialization method over a data hypercube.*

3 Specialization Method

3.1 Basics

Consider the example represented by Figure 1. Here, each point is a single data value based on two random variables X_1 and X_2. The hypercube (rectangle since $p = 2$) H^* is the result of the generalization process used to aggregate values into intervals for some specific category or class. The inner hypercube H contains most of the original datapoints for that class. The object is to improve the generalization to produce a hypercube H that is more reflective of the dataset. Whether the hypercube H or H' say (in Figure 1) is selected will depend on the level of homogeneity desired; this will give us an α-generalization.

To achieve this reduction of virtual descriptions selected, we first introduce the notion of the hypercube density.

Definition 1: Suppose $G \in \mathcal{P}(\Omega)$ and $a = \wedge_j [X_j \in d_j]$ is an assertion based on description d such that $d = (d_1, \cdots, d_p) = g(G)$. The density of the assertion a in G is defined as:

$$dens(a) = card(ext(a; G))/vol(d) \tag{3.1}$$

where $vol(d)$ is the volume of the description d.

One measure of the volume is that given by Brito (1994); specifically,

$$vol(d) = \prod_{j=1}^{p} \mu(d_j) \tag{3.2}$$

where

$$\mu(d_j) = \begin{cases} card(d_j), & \text{if } X_j \text{ is qualitative,} \\ \max\limits_{i \in G}\{d_j\} - \min\limits_{i \in G}\{d_j\}, & \text{if } X_j \text{ is quantitative.} \end{cases} \tag{3.3}$$

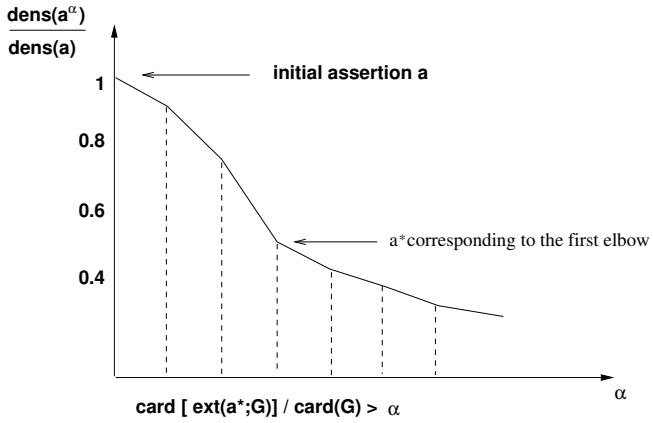

FIG. 2 – *Threshold α^* associated with the specialization process.*

The density corresponds to the ratio of the number of individuals satisfying the description of the assertion and the volume. As the density increases in its uniformity, then the quality of the generalization increases. The α-generalization method then is designed to find that hypercube H which has an appropriate minimized volume.

Definition 2: An α-generalization of G is the hypercube represented by the assertion (a^α) belonging to the set of assertions denoted by $\mathcal{A}$, which has a minimal volume such that:

$$a^\alpha = \arg\min_{a \in \mathcal{A}} \{vol(d) | card(ext(a; G)) \geq \alpha * card(G)\}. \tag{3.4}$$

The rationale of an α-generalization is to allow an assertion not to cover all the observations of G. More specifically, it makes it possible to tune α such that the assertion is a good compromise in terms of the volume of the description and the coverage of G.

In order to obtain such a compromise, the goal is to define an optimal threshold α^* which offers the best compromise between reducing the volume and the loss of information (as expressed, e.g., by the deletion of rare data points). This optimal α^* is that α value which produces a sharp bend in the curve of the plot of the inverse of the relative density $dens(a^\alpha)/dens(a)$ for assertion a producing the original hypercube (H^* in Figure 1); see Figure 2. This is achieved through the so-called reduction algorithm (presented in Section 5 below) and is based on two criteria. One criterion calculates the covering and volume of the assertion, taking into account whether variables are quantitative, qualitative or taxonomic. To account for differences in scale for quantitative (continuous) variables, a coding of the data is proposed by discretizing each interval during the original generalization step. The second criterion concerns the choice of the optimal reduction assertion. In this case, a curve of disconsonance is constructed from all admissible hypercubes and the one which offers the best compromise between the capacity of the covering and a small volume is selected. That is, for the set $\mathcal{A}_I \subseteq \mathcal{A}$ of assertions which correspond to the initial step, this criterion compares all the assertions in $\mathcal{A}_I$ by their density, to select the optimal assertion.

3.2 Construction of Generalization Intervals

To account for differences in scale in quantitative (continuous) variables versus qualitative variables, a coding or discretization process is constructed for each assertion and each interval. This transformation also optimizes the decomposition of the original intervals into intervals within which observations are uniformly distributed.

Suppose after aggregation (by, e.g., SQL query for a given assertion) the category or class G_k had the n_k individual description values $\{x_{1j}, \cdots, x_{n_k j}\}$ for the random variable X_j. These can be ordered to produce

$$a_j = x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n_k)} = b_j. \tag{3.5}$$

The goal is to subdivide the original interval $I(1, n_k) \equiv [a_j, b_j] = [x_{(1)}, x_{(n_k)}]$ into r subintervals $[x_{(1)}, x_{(u_1)}]$, $[x_{(u_1+1)}, x_{(u_2)}]$, $\cdots$, $[x_{(u_{r-1}+1)}, x_{(u_r)} = x_{(n_k)}]$, where the observations in each subinterval $[x_{(u_{s-1}+1)}, x_{(u_s)}]$, $s = 1, \cdots, r$, are uniformly distributed. The choice of the cut point is discussed in Section 4.3. It aims to find the cut point which produces two subintervals that are the closest to uniformity. We propose to perform a divisive partitioning algorithm to split the interval. The use of recursive partitioning has been widely applied within the context of the clustering of the variables (e.g., with VarClus) or more recently within the scope of spectral clustering (Dasgupta et al., 2006). It is particularly well suited when, as in our case, a stopping criterion is defined. The discretization procedure is done as follows:

Recursive Partitioning Algorithm:

Step 1: If the null hypothesis $H_0 : I(1, n_k)$ is uniformly distributed holds, then go to Step 5.

Step 2: Select the best cut point u from the $(n_k - 1)$ possible cut points.

Step 3: Apply this algorithm to the new intervals $I(1, u)$ and $I(u + 1, n_k)$ (Step 1 and 2).

Step 4: Repeat Steps 1-3 until Step 5 is reached.

Step 5: End.

This algorithm utilizes a test of uniformity across the subinterval $I(u_{s-1} + 1, u_s)$. It also involves the determination of a cut point u_s between $x_{(u_s)}$ and $x_{(u_s+1)}$, on an interval which is not uniformly distributed; it aims to find that u_s which produces two subintervals that are the closest (of all possible u_s values) to uniformity but which are not necessarily yet sufficiently uniform; hence the need for further divisions. See Section 4 below.

4 Tests of Uniformity

Three tests of uniformity are utilized, specifically, the goodness of fit test, the test of distance distributions, and the gap test. These are considered in turn.

4.1 Goodness of Fit Tests

Goodness of fit tests compare a theoretical distribution for the population with an empirical distribution based on the sample data. Many such tests exist. We focus on the Kolmogorov-Smirnov statistic for testing a theoretical uniform distribution.

For each $X \equiv X_j$, consider the n ordered observations $\{x_{(1)} \leq \cdots \leq x_{(n)}\}$ in the interval $[a, b]$. Then, if $[a, b] \notin [0, 1]$, the observations are transformed linearly by:

$$z_{(i)} = (x_{(i+1)} - x_{(1)})/(x_{(n)} - x_{(1)}), \qquad i = 1, \cdots, n-1, \tag{4.1}$$

to produce a sample of size $(n-1)$.

Then, under the hypothesis of uniformity, the theoretical cumulative distribution function is simply $F_0(z) = z$, $0 \leq z \leq 1$, while the empirical cumulative distribution function equals

$$\hat{F}(z) = card(\{i \in G | z_{(i)} \leq z\})/card(G). \tag{4.2}$$

From Saporta (1990), the largest absolute Kolmogorov-Smirnov distance,

$$ks_n = \sup_z |\hat{F}(z) - F(z)|, \tag{4.3}$$

is asymptotically distributed as

$$P\{ks_n\sqrt{n} < y\} \rightarrow \sum_{q=-\infty}^{\infty} (-1)^q exp(-2q^2y^2). \tag{4.4}$$

Hence, we can test

$$H_0 : F(z) = F_0(z) \text{ against } H_1 : F(z) \neq F_0(z),$$

which has critical region $c(n, \alpha)$ where $P\{ks_n > c(n, \alpha)\} = \alpha$. Note that from (4.1) and (4.3), the Kolmogorov-Smirnov distance (ks) is

$$ks = \sup_i |z_{(i)} - i/(n-1)|. \tag{4.5}$$

The Kolmogorov-Smirnov (KS) uniformity test algorithm, of complexity of order $O(n)$, for this test is simply:

KS uniformity test over [a,b]:

`Initialize` $ks = 0$

`Step` i, $i = 1, \cdots, n-2$,

`If` $|z_{(i)} - i/(n-1)| > ks$,

`then` $ks = z_{(i)}$ `and` $u = i$

`If` $ks \geq c(n-1, \alpha)$, `then partition recursively` $[a, z_u]$ `and` $[z_{u+1}, b]$ `constructed from the cut point` u.

Suppose now that the observations are those from two samples each with a different support. When there is a clear distinction between these supports, this test will not always reject the null hypothesis, since the empirical distribution based on the mixture of two distributions can indeed approach that of a uniform distribution. The null hypothesis of uniformity needs now to be compared with the alternative hypothesis that the distribution $F(z)$ is a mixture distribution. The gap test presented in Section 4.3 addresses this.

In order to improve the power of the test, the values

$$\{y_1 = z_{(1)}, y_i = z_{(i)} - z_{(i-1)}, i = 2, \cdots, n-1\} \tag{4.6}$$

are used instead of the ordered observations $\{z_{(1)}, \cdots, z_{(n-1)}\}$. The values y_i are exponentially distributed with mean $\beta = 1/(n-1)$, conditional on the constraint that $\Sigma y_i = 1$. Then, the cumulative distribution function of the distances between observations is:

$$F(x) = 1 - exp(-x/\beta).$$

If we denote the ordered distances by $\{y_{(i)}, i = 1, \cdots, n-1;\ y_{(0)} = 0\}$, we can construct new distance variables

$$y'_r = (n-r)(y_{(r)} - y_{(r-1)}), \qquad r = 1, \cdots, n-1. \tag{4.7}$$

The observations $\{y'_r, r = 1, \cdots, n-1\}$ constitute a random sample from the uniform distribution on $[0, 1]$ without regard to ordering (Sukhatme, 1937). We can then construct an ordered statistic (see Karr, 1986):

$$z'_{(j)} = \sum_{r=1}^{j} y'_r, \qquad j = 1, \cdots, n-1. \tag{4.8}$$

Durbin (1961) also showed that by using transformations that maximized (or minimized) collectively the distances between observations, the risks of error were reduced and hence the power of the test improved.

To accommodate this situation, we propose that the basic test for uniformity on the given sample of observation be first performed. If this is not rejected, then these transformed values can be found, and the basic hypothesis test can be performed on these transformed data.

4.2 Distance Tests

The distance test utilizes the conditional uniform test developed by Karr (1986) for Poisson point processes. This avoids the problem of finding an estimate of β in the (conditional on uniformity) exponential distribution, $Exp(0, \beta)$ of the distances y'_r of (4.7).

Therefore, the statistics,

$$V_i = y_{(i)}/y_{(n)}, \quad i = 1, \cdots, n-1,$$

where $y_{(i)}$ is the ordered distance (from $y_1, \cdots, y_n$), under the null hypothesis are statistics from a uniform distribution on [0, 1] independently of β.

Distance distributions have arisen often in the literature in the context of validation in classification analysis; see, e.g., Bock (1996) for a good review. The null hypothesis is that the distribution is uniform on the subspace of $\Re^p$ containing the objects of G. If G is not known, it is estimated by the convex envelope of the observations using the maximum likelihood method.

The test is used by comparing the empirical distribution of the initial distances

$$d_{kj} = \| x_k - x_j \| \tag{4.9}$$

to the theoretical distribution of the $\begin{pmatrix} n \\ 2 \end{pmatrix}$ distances between pairs of observations (y_k, y_j). Let

$$D^{(1)}_{(k)} = \min_{j \neq k} d_{kj}. \tag{4.10}$$

Then, under the assumption that these observations occur according to a Poisson process $P(\lambda)$ with mean λ, the $\{D^{(1)}_k, k = 1, \cdots, n\}$ are realizations from an exponential distribution $Exp(0, \lambda \times HV_p)$ where HV_p is the volume of a unit hypersphere in p-dimensional space, i.e.,

$$HV_p = \pi^{p/2}/\Gamma(1 + p/2). \tag{4.11}$$

A goodness of fit test can then be applied comparing the resulting empirical distribution with this theoretical exponential distribution.

As pointed out by Bock (1996), nearest neighbors need not be independent observations. When they are dependent, this test is unable to measure homogeneity directly. However, by modifying the test to one based on distances

$$\tilde{D}_j = \min_k \| x'_k - x'_j \| \tag{4.12}$$

where $\{x'_j, j = 1, \cdots, m\}$ are sampled with replacement from G, this problem is circumvented (see Bock, 1996, for details). In this case, instead of the distances $D^{(1)}_{(k)}$ of (4.9) and (4.10), we use the distances $\tilde{D}_j$ of (4.12).

4.3 The Gap Tests

Gap tests were developed in the context of classification and cluster analyses. We utilize this approach to compare the null hypothesis that the observations in the interval are uniformly distributed with the alternative hypothesis that the interval observations arise from a mixture of distributions. Intuitively, if there are “gaps" within the interval, then the uniformity hypothesis cannot hold. The origins of the test evolve from stationary Poisson processes $P(\lambda)$ where the observations are times at which events occurred. These times are by definition ordered time values. The times or distances between events are independent variables from an exponential distribution $Exp(0, \beta)$.

For each variable $X \equiv X_j$ with ordered realizations $x_{(i)}, i = 1, \cdots, n$, we are testing:

H_0: $\{x_{(1)}, \cdots, x_{(n)}\}$ has uniform distribution $F_0 = U(0,1)$, against
H_1: $\{x_{(1)}, \cdots, x_{(k)}\}, \{x_{(k+1)}, \cdots, x_{(n)}\}$ are ordered samples from $F_1(y)$ and $F_2(y)$, respectively.

Under H_1, the break point is at $x_{(k)}$ where

$$k = \arg \max_{1 \leq i < n} \{x_{(i+1)} - x_{(i)}\} \tag{4.13}$$

defines those observations $x_{(k)}$, $x_{(k+1)}$ which have the greatest distance between consecutive values.

Let $M_1 = \max_{i=1,\cdots,n-1} \{x_{(i+1)} - x_{(i)}\}$ be the maximum distance. Then, from Cox and Hinkley (1974), the test statistic becomes

$$Q(y) = \{1 - M_1/(x_{(n)} - x_{(1)})\}^n. \tag{4.14}$$

It follows therefore that the hypothesis of uniformity (H_0) is rejected if

$$P_{H_0}\{M_1/(x_{(n)} - x_{(1)}) \geq t(n,\alpha)\} = \alpha$$

where $t(n,\alpha)$ is some function of n and α. More details of this derivation can be found in Stéphan (1998); and more details of the test as applied to Poisson processes in general are in Kibushishi (1996).

Kibushishi (1996) has shown that, as $n \to \infty$, the distribution of

$$X^* = (nM_1)/(x_{(n)} - x_{(1)}) - log(n)$$

converges to a Gumbel distribution with $F(y^*) = exp(-exp(-y^*))$. Therefore, the critical region for H_0, for n sufficiently large, becomes:

$$P\{M_1/(x_{(n)} - x_{(1)}) > t(n,\alpha) | H_0\} = \alpha \tag{4.15}$$

where

$$t(n,\alpha) = (-log\{-log(1-\alpha)\} + log(n))/n. \tag{4.16}$$

At the end of the discretization process, each quantitative description of the set of assertions is transformed so that the split of the initial interval is taken into account.

In order to simplify the following formula with regard to the specialization step, we propose to associate a new ordinal variable with individual quantitative observations. Let us consider the initial variable X_j and let $V_j = \{v_1, \cdots, v_m\}$ be the output of the discretization process over the interval d_j. We define the recoded variable X_j^1 based on X_j and V_j as follows :

$$\begin{array}{rcl} X_j^1 & : & \Omega \to V_j \\ & & X_j^1(i) = v \text{ such that } x_{ij} \in v. \end{array}$$

For the sake of simplicity, we will denote identically the variable X_j and X_j^1. Similarly, the result of the split of an interval d_j is also denoted d_j. For each sub-interval v of d_j, the minimum value of v is written $\underline{v}$ while the maximum is denoted $\overline{v}$.

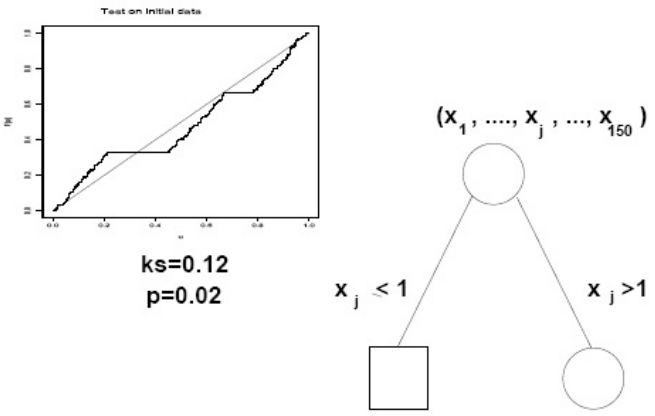

FIG. 3 – *Empirical distribution function over Sample 1.*

4.4 Illustration

The process is illustrated by the following two examples.

Sample 1

Suppose there are 150 observations in G; and suppose these were in fact aggregated from three separate simulated samples each of size 50 uniformly distribution populations $U(0,1)$, $U(2,3)$ and $U(3.5,4.5)$, respectively. The plot of the resultant empirical distribution is shown in Figure 3. The Kolmogorov-Smirnov distance from (4.5) is $ks = 0.12$. The gap test, that these 150 observations arise from one uniform distribution against the alternative hypothesis that they are from a mixture distribution, is rejected $(p = 0.02)$. The $x_{(u)}$ value at which the maximum distance occurs is $u = 50$. Therefore, the observations $P_1 = \{x_{(1)}, \cdots, x_{(50)}\}$ form one partition and the other observations $P_2 = \{x_{(51)}, \cdots, x_{(150)}\}$ form the second partition. The cut criterion is: If $x_{(i)} \leq 1$, then $x_{(i)} \in P_1$; otherwise, $x_{(i)} \in P_2$.

The uniformity test is then performed on each of P_1 and P_2. The plots of the resulting empirical distributions are shown in Figure 4 and Figure 5, respectively. Here (see the left-side plots), from Figure 4 for P_1, Test 1 gives $ks = 0.08$ and $p = 0.88$; and so the hypothesis that these data are uniformly distributed on $U(0,1)$ is not rejected. For those data in P_2 (in Figure 5), Test 1 gives $ks = 0.11$ and $p = 0.17$ which also does not reject the uniformity hypothesis. However, if the transformed data $\{z'_{(j)}, j = 1, \cdots, n\}$ of (4.8) (where here $n = 50, 100$ in P_1, P_2, respectively) are used instead of the $\{z_{(j)}, j = 1, \cdots, n\}$, then the corresponding empirical distributions are as shown in the Test 2 (right-side) plots of Figure 4 and Figure 5, and the Kolmogorov-Smirnov distances become $ks = 0.16$ (with $p = 0.07$) for P_1, and $ks = 0.18$ (with $p = 0.02$) for P_2. Therefore, the more powerful test based on the transformed values (i.e., Test 2), is able to identify the fact that the partition P_2 is not a single set of uniformly distributed observations but a mixture of distributions. The process is then repeated on P_2. In this case, P_2 is partitioned into $P_2^{(1)} = \{x_{(51)}, \cdots, x_{(100)}\}$ and $P_2^{(2)} = \{x_{(101)}, \cdots, x_{(150)}\}$. The

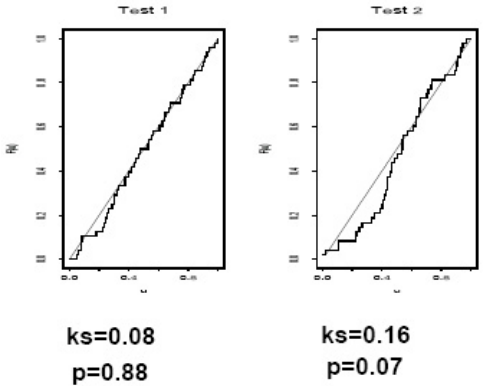

FIG. 4 – *Tests associated with the left leaf of Sample 1.*

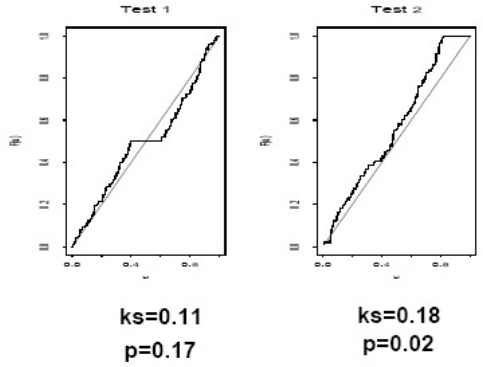

FIG. 5 – *Tests associated with the right node of Sample 1.*

Kolmogorov-Smirnov distances are $ks = 0.07$ with $p = 0.91$ (and $ks = 0.13$ with $p = 0.34$ for the transformed statistics) for $P_2^{(1)}$, and $ks = 0.11$ with $p = 0.50$ (and $ks = 0.10$ with $p = 0.59$ for the transformed statistics) for $P_2^{(2)}$. The cut point was $x_{(j)} \leq 3$ for $x_{(j)} \in P_2^{(1)}$. More details are in Stéphan (1998). For this partition, the Kolmogorov-Smirnov statistics and tests based on the original $\{z_{(j)}\}$ are confirmed by the transformed values $\{z'_{(j)}\}$. In this way, the three uniformly distributed partitions have been correctly identified.

Sample 2

The second example consists of observations drawn from four different uniform distributions, two each with large support and two with small supports. Suppose the full sample produces the empirical function as shown in Figure 6. The testing procedure elicited four uniformly distributed intervals $U(0, 0.96)$, $U(1.23, 2.17)$, $U(2.27, 3.3)$ and $U(3.41, 3.97)$ which compares favorably with the actual distributions ($U(0, 0.96)$, $U(1.23, 1.97)$, $U(2.13, 3.45)$ and

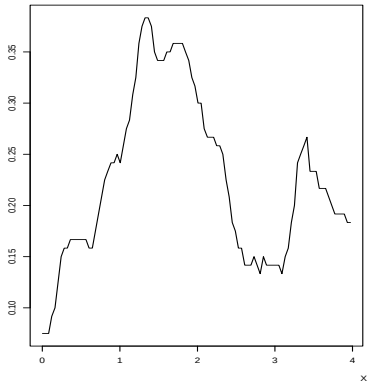

FIG. 6 – *Density function associated with Sample 2.*

$U(3.71, 3.97)$) from which the sample was drawn.

Henceforth, if the null hypothesis that the interval is homogeneous is not rejected, the gap test is applied, instead of subsequently testing the uniformity hypothesis using the transformed statistics.

5 The Reduction Algorithm

5.1 Some Terminology

In Section 3, the specialization method was outlined as one which provides for the possibility of removing outliers whereby a percentage α of the observations in the intervals (or hypercube) G are retained; the set G was obtained by the generalization process described in Section 2, based on the assertions which describe the classes or categories of interest. See Definition 2 and (3.4). This algorithm proceeds as follows.

Let the assertion $a = \wedge_j [X_j \in d_j]$ and let the maximum threshold reduction be α. Let S be the set of modalities resulting from d, with $S = \cup_{j=1}^{p} d_j$. We recall that d_j corresponds to a set of intervals when dealing with a quantitative variable X_j. For each modality $v \in S$, let $I(v) = j$ be the indicator that $v \in d_j$.

Then, if $M \in \mathcal{P}(S)$ is a set of modalities, the extension of M in the support of G is defined as

$$ext(M; G) = \{i \in G | v \in M, I(v) = j, x_{ij} = v\}. \tag{5.1}$$

In order to be conformed to the generalization process, the sets M of $\mathcal{P}(S)$ are restricted by the following constraint:

$$\begin{aligned} &\text{for all } M \in \mathcal{P}(S),\ v \in M \quad \Rightarrow \\ &(\text{for all } v_1 \in d_j,\ \overline{v_1} < \underline{v} \Rightarrow v_1 \in M) \quad \text{or} \quad \left(\text{for all } v_1 \in d_j,\ \underline{v_1} > \overline{v} \Rightarrow v_1 \in M\right). \end{aligned}$$

Proposition 1: Let $M \in \mathcal{P}(s)$ and let a be an assertion of description d obtained from the generalization G. It follows that if

$$card(ext(M;G))/card(G) = \beta,$$

then

$$card(ext(a^h;G))/card(G) = 1 - \beta \tag{5.2}$$

where assertion a^h corresponds to descriptions $d^h = (d_1^h, \cdots, d_j^h, \cdots, d_p^h)$ with

$$d_j^h = \oplus(\{x_{ij} \mid i \in G \setminus ext(G;M)\}, j = 1, \cdots, p) \tag{5.3}$$

where $A \setminus B$ is the set of elements of A which do not belong to B. Thus, if d_j^h is a discretized interval built-up from the set of V_j intervals, then

$$\oplus(\{x_{ij} | i \in G\}) = \{v \in V_j \mid \text{for all } i \in G,\ \underline{v} \geq \underline{x_{ij}} \text{ and } \overline{v} \leq \overline{x_{ij}}\}. \tag{5.4}$$

The proof is outlined in the Appendix.

It then follows that for all $M \in \mathcal{P}(S)$,

$card(ext(M;G))/card(G) \leq \alpha$ if and only if $card(ext(a^h;G))/card(G) \geq 1 - \alpha$

where a^h has a description d^h such that $d_j^h = \oplus(\{x_{ij} \mid i \in G \setminus ext(M;G)\}, j = 1, \ldots, p$.

The principle is to find those particular assertions which correspond to elements, generated by the generalization G, but not belonging to the extension of the complete set of modalities $M \in \mathcal{P}(S)$. It can be shown (by Proposition 2) that this is equivalent to constructing the assertion corresponding to the removal from the description the set of modalities M. This has the effect of reducing the number of operations needed, i.e., the complexity is decreased.

Let us define $M \in \mathcal{P}(S)$ to be a complete set of modalities if and only if the addition of a new modality not in M modifies the extension of M so that, for all

$$v \in S, v \notin M \Leftrightarrow ext(M \cup \{v\}; G) \neq ext(M; G). \tag{5.5}$$

Proposition 2: Let $a = \wedge_j [X_j \in d_j]$ be an initial assertion and let M be a complete set of modalities satisfying (5.5). Then, the assertion constructed by the generalization process G of those observations not in the extension of M has the same extension modalities of M in $(d_1, \cdots, d_p)$. This gives, for all $j \in \{1, \cdots, p\}$,

$$\oplus(\{x_{ij} \mid i \in G \setminus ext(M;G)\}) = d_j \setminus \{v \in M | I(v) = j\}. \tag{5.6}$$

The proof is outlined in the Appendix.

5.2 The Algorithm

After initiating the algorithm, an iteration stage constitutes the core. The set of solutions at stage k is denoted by L_k. Three subroutines (macros) are also involved. These procedures are related to the Apriori algorithm in the scope of mining association rules (Agrawal and Srikant, 1994). As for the Apriori algorithm, at each step, we seek sets of modalities which are candidates to be removed. The way we determine these sets is closely related to the Apriori strategy, achieved by taking advantage of its computation efficiency.

Step 1: Construction of L_1:

The first set of solutions, L_1, is defined as the set of singletons

$$L_1 = \{\{m\} \mid m \in S \text{ and } ext(\{m\}; G) \leq \alpha * card(G)\}$$

To construct L_1:

(i) Set $L_1 = \phi$

(ii) For $j = 1, \cdots, p$, do
if X_j is qualitative or taxonomic

$L \leftarrow L \cup \{v \in d_j \mid ext(\{v\}; G) \leq \alpha * card(G)\}$

else if Y_j is quantitative

$L \leftarrow L \cup \{[\underline{d_j}, \overline{v}] \mid v \in d_j \text{ and } ext([\underline{d_j}, \overline{v}]; G) \leq \alpha * card(G)\}$

$L \leftarrow L \cup \{[\underline{v}, \overline{d_j}] \mid v \in d_j \text{ and } ext([\underline{v}, \overline{d_j}]; G) \leq \alpha * card(G)\}$

(iii) end j; return to (ii) for $j = j + 1$

(iv) return L_1

(v) End.

Step 2: Construction of L_2:

The set L_2 consists of the set M_2 of pairs of elements from L_1. Each pair $M_2 = \{m_1, m_2\}$ with $\{m_i\} \in L_1, i = 1, ..., card(L_1)$, is admissible provided that:

(a) $m_1 \neq m_2$

(b) $card[ext(M_2; G)] \leq \alpha * card(G)$

(c) $card[ext(M_2; G)] > card[ext(\{m_i\}; G)]$ for both $i = 1$ and $i = 2$.

The condition (a) is used to guard against deleting the same element twice from a modality. Condition (b) ensures that the cardinality of the resulting set does not exceed the maximal reduction threshold α.

Condition (c) states that, if the elements covered by $\{m_i\}$ are also covered by $\{m_{i'}\}$ $(i' \neq i)$, we deduce that the assertion generated by the elimination of elements in the extension of $\{m_i\}$ is equivalent to the assertion generated by the elimination of elements from the extension of $\{m_1, m_2\}$. Thus, the solution $M_2 = \{m_1, m_2\}$ is not possible. Suppose we have

$$card[ext(M_2; G)] = card[ext(\{m_1\}; G)].$$

Then it follows that if G contains m_2, it also contains m_1.

Let m_1^C be the set of modalities which are complete in m_1. Initially, $m_1^C = \{m_1\}$. Then, m_1^C can be updated by merging it with $\{m_2\}$, viz.,

$$m_1^C \leftarrow m_1^C \cup \{m_2\}.$$

A consequence of this stage is that now all the elements $M_1 \in L_1$ are complete modalities.

Step 3: Construction of L_3:

The third step constructs L_3 by combining elements from each of M_2 and L_1. That is, for $M_2 \in L_2$ and $\{m\} \in L_1$, $L_3 = \{M_3 = m_2 \cup \{m\}\}$ subject to

(d) for all $v \in M_3, M_3 \backslash \{v\} \in L_2$

(e) $card[ext(M_3; G)] \leq \alpha * card(G)$

(f) for all $v \in M_3$, $card[ext(M_3; G)] > card[ext(M_3 \backslash \{v\}; G)]$.

These conditions parallel those ((a), (b), and (c)) for constructing L_2. This step constructs sets M_3 such that, for all $v \in M_3$, $M_3 \backslash \{v\} \in L_2$. This verifies that pairs of elements in L_2 unite with a new element to produce a triplet M_3. For example, suppose $M_3 = \{m_1, m_2, m_3\}$. Then adding this M_3 to L_3 necessitates the inclusion of the pairs $\{m_1, m_2\}$, $\{m_1, m_3\}$, $\{m_2, m_3\}$ in L_2. If one of these, $\{m_i, m_j\}$ say, is not in L_2, it follows that

$$card[ext(\{m_i, m_j\}; G)] > \alpha * card(G).$$

Therefore, the triplet M_3 cannot be added to L_3 since to do so has the consequence that $card[ext(M_3; G)] > \alpha * card(G)$ which violates condition (e). Further, if this pair $\{m_i, m_j\}$ is already not in L_2, there is no need to consider it again as its presence or otherwise was accounted for at the L_2 stage.

This last condition (f) is a general condition for sets with two or more elements, and protects against redundancy in L_2. This makes it possible to update all sets of modalities of L_2. If all conditions ((d)-(f)) are satisfied, then M_3 can be added to L_3. Therefore, if the complete set is initialized as M_3^C, $M_3^C \leftarrow \cup_{M_2 \in L_2, M_2 \subseteq M_3} M_2^C$ where M_2^C is the complete set deduced from M_2.

Step k+1: Construction of L_{k+1} from L_k:

Step 3 can be generalized to give a recursive construction of L_{k+1}, $k = 3, 4, \cdots$. The elements of L_{k+1} are $\{M_{k+1}\}$ such that for all $v \in M_{k+1}$, $M_{k+1} \backslash \{v\} \in L_k$ and subject to:

(k1) for all $v \in M_{k+1}, M_{k+1} \backslash \{v\} \in L_k$

(k2) $card[ext(M_{k+1}; G)] \leq \alpha * card(G)$

(k3) for all $v \in M_{k+1},\ card[ext(M_{k+1}; G)] > card[ext(M_{k+1} \backslash \{v\}; G)]$.

These modalities are found as follows:

A. Initialize a set J from L_k,
$J \leftarrow \{M_{k+1} |$ for all $v \in M_{k+1}, M_{k+1} \backslash \{v\} \in L_k\}$
= AutoJoin (L_k).

B. Initialize $L_{k+1} \leftarrow 0$.

C. For each $M_{k+1} \in J$, calculate $ext(M_{k+1}; G)$.

D. (Macro: CompleteUpdate (L_k, M_{k+1}) of Reduction Algorithm)

(i) If $card[ext(M_{k+1}; G)]/card(G) > \alpha$, then M_{k+1} is not included in L_{k+1}.

(ii) If there exists a modality $v \in M_{k+1}$ such that
$ext(M_{k+1}; G) = ext(M_{k+1} \backslash \{v\}; G)$, then M_{k+1} is not included in L_{k+1}, and the complete modality of M_k is updated, i.e.,
$M_k^c \leftarrow M_k^c \cup \{v\}$;

(iii) Otherwise,
$L_{k+1} \leftarrow L_{k+1} \cup \{M_{k+1}\}$,
$M_{k+1}^c \leftarrow \cup_{M_k \in L_k, M_k \subseteq M_{k+1}} M_k^c$.

E. Return to Step A with $k = k + 1$.

That this calculation is determined automatically is a consequence of the property of complete modality (see Proposition 2).

The elimination of a set of modalities (observations) is a consequence of the calculation of complete modalities. This is executed by reducing the volume of an assertion. For a complete set of modalities M, let $V_j = \{m \in M | I(m) = j\}$ where M is deduced from d_j.

Then, the volume reduction algorithm for modality in M and assertion a, is

```
Reduction Volume (a, M)
(macro component of Reduction Algorithm):
 Start
 1.  volume ← 1
 2.  For each j ∈ {1, ··· , p}, do
 3.      If X_j is qualitative or quantitative, volume ←
     volume×(μ(d_j) − μ(V_j))
 4.      Else (if Y_j is taxonomic), volume ← volume×μ(d_j\V_j)
 5.  End
 6.  Return to j = j + 1
 7.  If j = p, then output volume
     End
```

In step 4/3, the $\mu(\cdot)$ are as defined in (3.3).

This volume reduction stage makes it possible to choose from all possible assertions, thus optimizing the density of the final hypercube (see Definition 2). There are two possible optimality selection criteria. One is to use the scree test, in the density curve, as suggested in Section 3; see Figure 2. The optimal α is that value at which a change in sign of the second derivative occurs. Another criterion is to find that assertion which gives M^* for which

$$F(M^*) = \min_{M \in RF} \{F(M)\}$$

where

$$F(M) = \{card(G) - card[ext(M; G)]\} / \text{ Volume } (a, M),$$

RF is the final reduction set L_k, and Volume (a, M) is the volume calculated by the Reduction Volume algorithm. This stage is thence incorporated into the reduction algorithm. The complete reduction algorithm is given in Section 5.3.

5.3 Reduction Algorithm

For (a, G, α), the reduction algorithm is:

```
1.  Calculate L_1 from {d_j}_J as a function of G and α
2.  EnsAdm ← L_1
3.  k ← 2
4.  As long as (L_{k−1} ≠ φ), do
5.      J ← autojoin (L_{k−1})
6.      L_k ← φ
7.      For (M_k ∈ J), do
8.          if {card[ext(M_k; G)]/card(G)} ≤ α, then
```

```
9.            if (for all v ∈ M_k, ext(M_k;G)\{v}) ≠ ext(M_k;G),
   then
10.               L_k ← L_k ∪ M_k
11.            else
              MiseaJourComplete (L_{k-1}, M_k)
          end
       end
12.      EnsAdm ← EnsAdm ∪L_k
13.      k ← k + 1
14.  RedFinal ← φ
15.  For (M ∈ EnsAdm), do
16.      Vol (a\M) = Reduction Volume (a, M)
17.      If (for all M' ∈ RedFinal, card[ext(M';G)] >
           card[ext(M;G)] ⇒ Vol(a\M') > Vol(a\M)) then
18.          For (M' ∈ RedFinal) such that
           (card[ext(M';G)] ≥ card[ext(M;G)]∧Vol(a\M') ≤ Vol(a\M))
   do
19.              RedFinal ← RedFinal \{M'}
       else
           RedFinal ← RedFinal U{M}
20.      end
21.  choose the best a* from
       A = ∪_{M∈RedFinal}{a = ∧_j[X_j ∈ d_j\{m ∈ M}|I(m) = j]} .
```

5.4 Complexity associated with the reduction algorithm

As already said, the reduction step is related to the Apriori algorithm in the scope of mining association rules. Specifically, the first step consists of determining the set of singletons L_1 which verify the maximum threshold reduction criterion α. Among these singletons, those having an extension greater than $(1-\alpha)card(ext(a))$ are discarded from L_1. This step only requires one scan over the set of the observations belonging to the current assertion a. At the end of the first step, let us suppose that L_1 contains l singletons. If we do not consider a pruning strategy as proposed in the Aprioi algorithm, the number of solutions M (i.e. a set of modalities) to be generated equals 2^{l-1}.

One originality of our work is the use of the Apriori pruning strategy by removing solutions M which do not verify the maximum threshold criterion α. This strategy has been detailed in the previous paragraph. Let us consider step L_{k+1}. The elements of L_{k+1} are determined by merging a pair of elements of L_k. For this step, $card(L_k)^2$ merging operations are done over L_k to produce L_{k+1}. Then, the set of the n observations belonging to a is scanned in

order to update $card[ext(M_{k+1}; G)]$ for each M_{k+1} belonging to L_{k+1}. As previously, the elements M_{k+1} which do not verify the criterion α are discarded from L_{k+1}. With regard to the complexity, the generation of the sets of elements $L_1, ..., L_k$ associated with a scan of the observations is an NP-complete task.

6 Application

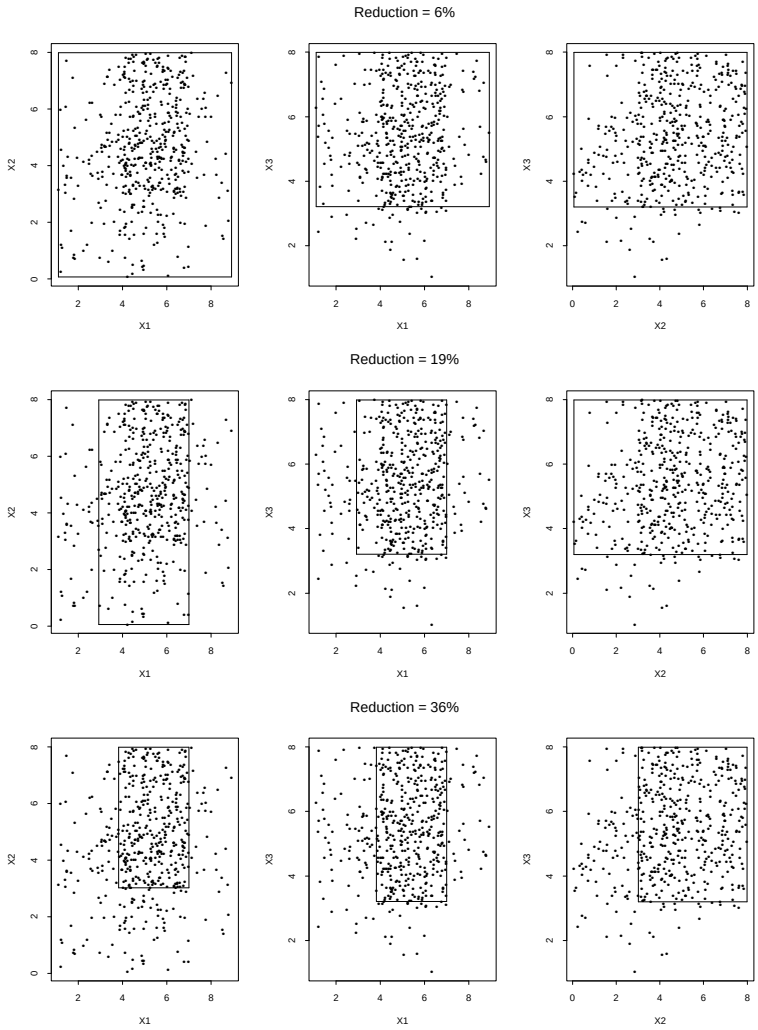

FIG. 7 – *Example of reduction with a dataset having 3 variables.*

Figure 7 represents the three pairwise plots of a dataset of 500 observations in three variables $\boldsymbol{X} = (X_1, X_2, X_3)$. The goal is to find the assertion(s) which describes the data as uniformly and as well as possible. Applying the generalization process directly to all the data and so finding the minimum and maximum values for each $X_j, j = 1, 2, 3$, (in effect $\alpha = 1$ with card $[ext(a; G)] = 500$) the assertion that results is

$$a = [X_1 \in [1.10, 8.93]] \wedge [X_2 \in [0.06, 7.99]] \wedge [X_3 \in [2.12, 7.99]] .$$

When the coverage is reduced to $\alpha = 0.97$, the assertion as it pertains to X_3 is $X_3 \in [2.9, 7.99]$ and when $\alpha = 0.94$, this assertion becomes $X_3 \in [3.21, 7.99]$. There is no change in the coverage space for X_1 or X_2. The lines marking "inner" rectangles in the top panel of Figure 7 shows the complete assertion for this 94% coverage.

However, when the focus shifts to X_1, coverage is reduced to $\alpha = 0.81$ (see the middle panel of Figure 7) with still no change in X_2 and no further change in X_3, the assertion as it relates to X_1 produces a more homogeneous region, with an overall assertion of

$$a = [X_1 \in [1.10, 6.99]] \wedge [X_2 \in [0.06, 7.99]] \wedge [X_3 \in [3.21, 7.99]] .$$

The final reduction is shown in the bottom panel of Figure 7, and was achieved at only five iterations. The coverage is $\alpha = 0.64$ with $card[ext(a; G)] = 320$; and the assertion is

$$a = [X_1 \in [3.82, 6.99]] \wedge [X_2 \in [3.02, 7.99]] \wedge [X_3 \in [3.21, 7.99]] .$$

This coverage space satisfies the hypothesis of uniformity. The progression from clearly a nonuniform scatter of observations in the top panel of Figure 7 to the homogeneous regions in the bottom panel of Figure 7 is evident.

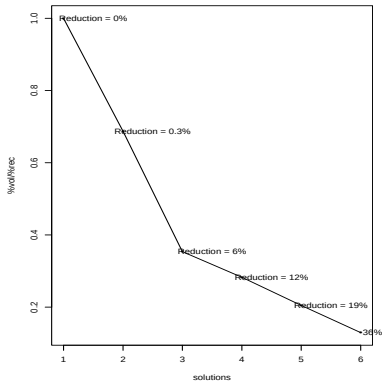

FIG. 8 – *Disconsonance curve associated with Figure 7.*

The disconsonance curve is shown in Figure 8. The elbow occurs at the second iteration concerned with X_3 with coverage $\alpha = 0.94$. That is, the assertion is, as shown in the top panel of Figure 7,

$$a = [X_1 \in [1.10, 8.93]] \wedge [X_2 \in [0.06, 7.99]] \wedge [X_3 \in [3.21, 7.99]] .$$

To illustrate that the reduction algorithm works well on a mixture of two distinct populations, consider the bivariate observations plotted in Figure 9. These are samples each of size 500 from two different bivariate normal distributions $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\mu}_1 = (5, 8) \text{ and } \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1.5 & 0.6 \\ 0.6 & 1.5 \end{bmatrix}$$

and

$$\boldsymbol{\mu}_2 = (9, 14) \text{ and } \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1.5 & -0.6 \\ -0.6 & 1.5 \end{bmatrix}.$$

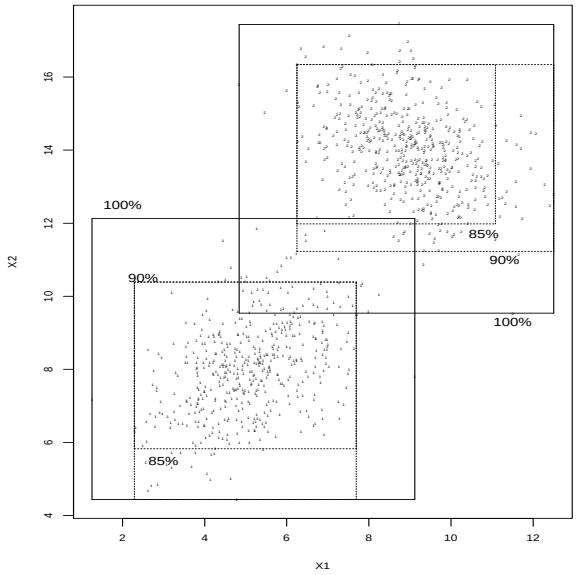

FIG. 9 – *Reduction on Example 2.*

The generalization process produces two hypercubes defined by the assertions,

$$\begin{aligned} \text{for } (\mu_1, \Sigma_1) : a_1 &= [X_1 \in [1.26, 9.12]] \wedge [X_2 \in [4.44, 12.13]], \\ \text{for } (\mu_2, \Sigma_2) : a_2 &= [X_1 \in [4.84, 12.5]] \wedge [X_2 \in [9.54, 17.44]]. \end{aligned}$$

The first stage of the reduction process results in a discretization of the X_1 and X_2 intervals in a_1 according to

$$\begin{aligned} a_1 = [X_1 \in (1.26, [2.29, 3.53], [3.55, 6.27], [6.28, 6.96], [6.96, 7.6], [7.71, 9.12])] \\ \wedge [X_2 \in ([4.44, 5.8], [5.83, 6.43], [6.44, 9.5], [9.55, 10.39], [10.42, 12.13])]. \end{aligned}$$

The reduction algorithm gave the α-generalization on the first sample the assertion, with coverage 0.95,

$$a_1(\alpha = 0.95) = [X_1 \in [3.2, 7.7]] \wedge [X_2 \in [4.4, 10.4]],$$

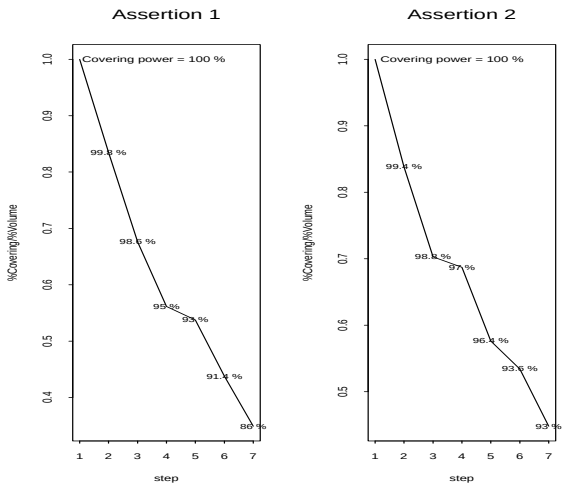

FIG. 10 – *Disconsonance curve associated with Figure 9.*

and for the second sample the α-generalization gave the assertion, with coverage 0.98,

$$a_2(\alpha = 0.98) = [X_1 \in [6.3, 12.5]] \wedge [X_2 \in [11.2, 17.4]].$$

The scree plots are shown in Figure 10. It is seen that the elbow in the plot for assertion a_1 occurs at the fourth (and fifth) iteration with coverage 0.95. For assertion a_2, there is an elbow at the third iteration with $\alpha = 0.989$, and at the fourth iteration for assertion a_1 where $\alpha = 0.95$.

7 Conclusion

In this paper, we advocate the use of a generalization and a specialization method to investigate in particular whether or not an aggregation of quantitative observations into intervals with those observations uniformly spread across those intervals. This is particularly important since so far methodologies for interval data have an underlying assumption that this uniformity feature pertains. If uniformity does not hold, then those methodologies can produce analytic results that are distorted; consider, e.g., how the example (in Section 1) of the interval $[17, 90]$ rather than the more representative interval $[80, 90]$ would distort the outputs of a symbolic principal component analysis. The analysts should be alert to the possibility that observations may not be necessarily homogeneous across the intervals, and that maybe two subintervals, e.g., should be used instead. The approach proposed herein helps to answer such questions. This has been undertaken with Georges Hébrail and Yves Lechevallier (Stéphan et al., 2000, Hébrail and Lechevallier, 2007) within the context of the European project SODAS (Diday and Noirhomme-Fraiture, 2008). To our knowledge further applications in symbolic data analysis do not take advantage of this specialization step, while it has been integrated in SODAS software. Nevertheless, differences can emerge when building symbolic objects from classical

clustering approaches rather than from rational databases as here. These issues are important; and our work is just a beginning on these issues.

There are a number of questions that arise. In this work, we have considered the case where rare or extreme observations arise. Another key problem is whether heterogeneity may lead to the determination of several assertions associated with the same group. In this context, we should stress the contributions of ElGolli (2004). We hope these two approaches to tackling the problem of over-generalization by either dividing or reducing an assertion can be more intensively used when generating symbolic objects.

References

Agrawal, R. and R. Srikant (1994). Fast algorithm for mining association rules. In: *Proceedings* 20^{th} *International Conference on Very Large Data Bases*, 487-499.

Bay, S. D. (2001). Multivariate discretization for set mining. *Knowledge and Information Systems* 3, 491-512.

Billard, L. and E. Diday (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley, Chichester.

Bock, H.-H. (1996). Probability models and hypothesis testing in partitioning cluster analysis. In: *Clustering and Classification* (eds. P. Arabie, G. De Soete and L. Hubert). World Science Publishers, 377-453.

Bock, H.-H. and E. Diday (eds.) (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer-Verlag, Berlin.

Brito, M. P. (1994). Use of pyramids in symbolic data analysis. In: *New Approaches in Classification and Data Analysis* (eds. E. Diday, Y. Lechevalier, M. Schader, P. Bertrand and B. Burtchy). Springer-Verlag, Berlin, 378-386.

Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics*. Chapman and Hall, London.

D'Agostino, R. B. and M. A. Stephens (1986). *Goodness-of-Fit Techniques*. Marcel Dekker.

Dasgupta, A., J. Hopcroft, R. Kannan, and P. Mitra (2006). Spectral clustering by recursive partitioning. In: *AlgorithmsÚESA*. Springer-Verlag, Berlin, 256-267.

Diday, E. and M. Noirhomme-Fraiture (2008). *Symbolic Data Analysis and the SODAS Software*. Wiley, Chichester.

DuMouchel, W., C. Volinsky, T. Johnson, C. Cortes, and D. Pregibon (1999). Squashing flat files flatter. In: *Proceedings of the Fifth ACM Conference on Knowledge Discovery and Data Mining*. ACM Press, 6-15.

Durbin, J. (1961). Some methods of constructing exact tests. *Biometrika* 48, 41-65.

ElGolli, A. (2004). Extraction de donnéees symboliques et cartes topologiques: Application aux donnéees ayant une structure complexe. Unpublished Doctoral Thesis, Université de Paris Dauphine.

Esposito, F. and C. d'Amato (2007). An agglomerative hierarchical clustering algorithm for improving symbolic object retrieval. In: *Selected Contributions in Data Analysis and Classification* (eds. P. Brito, P. Bertrand, G. Cucumel, and F. de Carvalho). Springer-Verlag, Berlin, 45-53.

Esposito, F., D. Malerba, and V. Tamma (2000). Dissimilarity measures for symbolic objects. In: *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data* (eds. H.-H. Bock and E. Diday). Springer-Verlag, 165-185.

Han, J., Y. Fu, W. Wang, J. Chiang, W. Gong, K. Koperski, D. Li, Y. Lu, and A. Xia (1997). Dbminer: a system for mining knowledge in large relational databases. In: *Proceedings* 2^{nd} *International Conference on Knowledge Discovery in Databases and Data Mining*, 250-255.

Hand, D. J., H. Manila, and P. Smyth (2001). *Principles of Data Mining*. MIT Press.

Hébrail, G. and Y. Lechevallier (2007). Building symbolic objects from data streams. In: *Selected Contributions in Data Analysis and Classification* (eds. P. Brito, P. Bertrand, G. Cucumel, and F. de Carvalho). Springer-Verlag, Berlin, 83-94.

Ichino, M. and H. Yaguchi (1994). General Minkowski metrics for mixed feature type data analysis. IEEE *Transactions on System, Man and Cybernetics* 24, 698-708.

Kibushishi, T. (1996). *On Some Applications on the Point Process Theory in Cluster Analysis and Pattern Recognition.* Unpublished Doctoral Thesis, Université Notre-Dame de la Paix, Namur, Belgique.

Karr, A. (1986). *Poisson Point Processes*. Marcel Dekker.

Ludl, M. C. and G. Widner (2000). Relative unsupervised discretization for association rule mining. In: *Proceedings* 4^{th} *European Conference on Principles and Practice of Knowledge Discovery in Databases*. Springer-Verlag, 148-158.

Michalski, R. S. and R. E. Stepp (1983). Learning from observations: conceptual clustering. In: *Machine Learning: An Artificial Intelligence Approach* (eds. R. S. Michalski, J. G. Carbonell and T. M. Mitchell). Morgan Kaufmann, 331-363.

Miller, R. J. and Y. Yang (1997). Association rules over interval data. In: *Proceedings ACM SIGMOD International Conference on Management of Data*. ACM Press, 452-461.

SAS/STAT 9.2 User's Guide: The VARCLUS Procedure.

Saporta, G. (1990). *Probabilités, Analyse des Données et Statistique*. Éditions Technip.

Srikant, R. and R. Agrawal (1996). Mining quantitative association rules in large relational tables. In: *Proceedings ACM SIGMOD International Conference on Management of Data*. ACM Press, 1-12.

Stéphan, V. (1998). *Construction d' Objects Symboliques par Synthèse des Résultats de Requêtes SQL*. Unpublished Doctoral Thesis, Université Paris-IX Dauphine, Paris, France.

Stéphan, V., G. Hébrail, and Y. Lechevallier (2000). Generation of symbolic objects from relational databases. In: *Analysis of Symbolic Data: Exploratory Methods for Extracting*

Statistical Information from Complex Data (eds. H.-H. Bock and E. Diday). Springer-Verlag, 78-105.

Sukhatme, P. V. (1937). Tests of significance for samples of the X^2 population with two degrees of freedom. *Annals Eugenics* 8, 52-56.

A APPENDIX

A.1 Proof of Proposition 1:

Let $a^h = \wedge_j[X_j \in \oplus(\{x_{ij} \mid i \in G \; and \; i \notin ext(M;G)\})]$. From the generalization property, we have

$$\text{for all } i \in G \backslash ext(M;G), \qquad \omega \in ext(a^h;G).$$

Therefore,

$$card(ext(a^h;G))/card(G) \geq 1 - \beta. \tag{*1}$$

We now show that

$$\text{for all } i \in ext(M;G), \qquad i \notin ext(a^h;G).$$

We prove this by contradiction. Suppose that there exists $i^* \in G$,

$$i^* \in ext(M;G) \wedge \omega^* \in ext(a^h;G).$$

Then, $i^* \in ext(M;G)$ implies there exists a $j \in \{1, \cdots, p\}$ with $x_{i^*j} \in M$.

First, suppose that X_j is not a quantitative variable. Since the modality $x_{i^*j} = v \in M$ tells us that, for all $i \in G, x_{ij} = v$ implies that $i \in ext(M;G)$ or, equivalently, $i \notin G \backslash ext(M;G)$. By the generalization property, it follows that

$$v \notin \{x_{ij} | i \in G \backslash ext(M;G)\},$$

or,

$$v \notin \oplus(\{x_{ij} | i \in G \backslash ext(M;G)\}).$$

Therefore, $i^* \notin ext(a^h;G)$.

Second, suppose that X_j is a quantitative variable. Then, $x_{i^*j} v \in M$ implies

$$\begin{aligned} I_1 \vee I_2 \equiv (&\text{for all } v' \in d_j, \; \overline{v}' < \underline{v} \text{ implies } v' \in M) \text{ or} \\ &(\text{for all } v' \in d_j, \; \underline{v}' > \overline{v} \text{ implies } v' \in M) \end{aligned}$$

where $\underline{v}$ and $\overline{v}$ denote the smallest and largest v value, i.e., $v = [\underline{v}, \; \overline{v}]$.

Consider the first of these two implications, I_1. It follows from I_1 that, for all $i \in G$,

$$\overline{x_{ij}} \leq \overline{v} \text{ implies } i \in ext(M;G),$$

and hence, for all $v' \in M$,

$$\overline{v}' \leq \overline{v} \text{ implies } v' \notin \oplus(\{x_{ij} \mid i \in G \backslash ext(M;G)\});$$

then

$$i^* \notin ext(a^h; G).$$

The same argument carries through for the second implication I_2, where we replace the largest by the smallest values (e.g., replace $\overline{v}$ by $\underline{v}$, etc.)

Together I_1 and I_2 tell us that, for all $i \in G$,

$$i \in ext(M; G) \text{ implies } i \notin ext(a^h; G).$$

Hence, we deduce that

$$card(ext(a^h; G))/card(G) \leq 1 - \beta. \quad (*2)$$

The two inequalities (*1) and (*2) give us

$$card(ext(a^h; G))/card(G) = 1 - \beta. \quad (*3)$$

This completes the proof.

A.2 Proof of Proposition 2:

Let M be the set of modalities. If M is complete, then, for all $v \in S$,

$$ext(M \cup \{v\}; G) \neq ext(M; G)$$

and vice versa. Let

$$d'_j = \oplus(\{x_{ij} \mid i \in G \backslash ext(M; G)\}).$$

First suppose that X_j is not a quantitative variable. Then, for all $v \in d'_j$, there exists $i \in G \backslash ext(M; G)$ where $x_{ij} = v$. It follows that there exists

$$i \in G \backslash ext(\{v\}; G) \wedge i \notin ext(M; G)$$

and vice versa, which in turn implies

$$ext(\{v\}; G) \nsubseteq ext(M; G)$$

and vice versa. This in turn implies

$$ext(M \cup \{v\}; G) \neq ext(M; G)$$

and conversely. However, M was assumed to be complete; so therefore, $v \notin M$. Therefore,

$$\oplus(\{Y_j(\omega) | \omega \in G \backslash ext(M; G)\}) = d_j \backslash \{v \in M | I(v) = j\}.$$

Secondly suppose that X_j is a quantitative variable. Now, for all $v \in d'_j$,

$$\begin{aligned} I_1 \vee I_2 \quad \equiv \quad & (\text{there exists } i \in G \backslash ext(M; G), x_{ij} > v) \text{ or} \\ & (\text{there exist } i_1, i_2 \in G \backslash ext(M; G), \overline{x_{i_1 j}} < \underline{v} \text{ and } \underline{x_{i_2 j}} > \overline{v}). \end{aligned}$$

If I_1 is true, then it is equivalent to the qualitative variable case. In contrast, the contradiction is supposed to be true and that

$$v \in d'_j \text{ and } v \in M.$$

This means that

$$\text{for all } \omega \in G \backslash ext(M;G), x_{ij} \neq v.$$

Also,

$$\text{for all } i_1, i_2 \in G \backslash ext(M;G),\ \overline{x_{i_1 j}} < \underline{v} \text{ and } \underline{x_{i_2 j}} > \overline{v}.$$

Or, we suppose $v \in M$. To construct M, we know that $v \in M$ implies that

$$(\text{for all } v' \in d_j,\ \overline{v}' < \underline{v} \text{ implies } v' \in M) \text{and } (v' > \overline{v} \text{ implies } v' \in M).$$

If

$$v' \in d_j,\ \overline{v}' < \underline{v},$$

if follows that

$$\text{for all } i \in G,\ \overline{x_{ij}} < \underline{v} \text{ implies } i \in ext(M;G).$$

This contradicts the condition that there exist

$$i_1, i_2 \in G \backslash ext(M;G),\ \overline{x_{i_1 j}} < \underline{v} \text{ and } \underline{x_{i_2 j}} > \overline{v}.$$

Therefore, the hypothesis that $v \notin M$ is false. In a similar way, we can show that the condition I_2 also contradicts the hypothesis that $v \notin M$.

We then can conclude that

$$\oplus(\{x_{ij} \mid i \in G \backslash ext(M;G)\}) = d_j \backslash \{v \in M \mid I(v) = j\}. \qquad (*4)$$

Thus, equivalence has been proven for both qualitative and quantitative variables. Therefore, (*4) holds for all $j \in \{1, \cdots, p\}$. This completes the proof.