Extracting partially ordered clusters from ordinal polytomous data

Abstract

In practical applications of knowledge space theory, knowledge states can be conceived as partially ordered clusters of individuals. Existing extensions of the theory to polytomous data lack methods for building "polytomous" structures. To this aim, an adaptation of the k-median clustering algorithm is proposed. It is an extension of k-modes to ordinal data in which the Hamming distance is replaced by the Manhattan distance, and the central tendency measure is the median, rather than the mode. The algorithm is tested in a series of simulation studies and in an application to empirical data. Results show that there are theoretical and practical reasons for preferring the k-median to the k-modes algorithm, whenever the responses to the items are measured on an ordinal scale. This is because the Manhattan distance is sensitive to the order on the levels, while the Hamming distance is not. Overall, k-median seems to be a promising data-driven procedure for building polytomous structures.

Keywords: knowledge space theory, polytomous KST, k-modes, k-median, clustering algorithms

1 Introduction

Knowledge space theory (KST; Doignon & Falmagne, 1985, 1999; Falmagne & Doignon, 2011) was developed in its Boolean fashion with the aim of achieving reliable, effective and efficient adaptive knowledge assessment. In dichotomous KST the knowledge of an individual is represented by her *knowledge state*, which is the subset K of the problems in a given knowledge domain Q that the specific individual can solve. A *knowledge structure* is formally defined as the pair $(Q, \mathcal{K})$ where $\mathcal{K}$ is a collection of knowledge states which contains at least $\emptyset$ and Q. In practical applications of KST, a knowledge structure can be conceived as a collection of partially ordered clusters of individuals. Each cluster in the structure is a knowledge state.

Since the very beginning of the development of the theory, one of the most challenging issues in KST has been the construction of the knowledge structure. This task is, evidently, crucial in the application of the theory in practice. Therefore a considerable amount of research explored this topic ending up with three different approaches.

The first one is the *query to experts* (Dowling, 1993; Kambouri, Koppen, Villano, & Falmagne, 1994; Koppen, 1993; Koppen & Doignon, 1990; Müller, 1989; Schrepp & Held, 1995). The basic idea behind it is that a structure can be defined on the basis of the prerequisite relation that can be found among the items of the domain. Such relation is supposed to be detectable by means of a systematic interrogation of experts.

The second method is the *skill map construction* (Albert & Lukas, 1999; de Chiusole & Stefanutti, 2013; Doignon, 1994; Heller, Augustin, Hockemeyer, Stefanutti, & Albert, 2013; Lukas & Albert, 1993; Stefanutti & de Chiusole, 2017). In this approach, a set of skills that are needed to solve the items of the domain is defined. Different items require different subsets of skills. Once the skill assignment is carried out for each

item, the structure is built by including all the states that are coherent with the skill assignment. This second method can also be found, with the name of Q-matrix specification in the cognitive diagnostic models (CDM; Bolt, 2007; DiBello & Stout, 2007; Junker & Sijtsma, 2001; Tatsuoka, 1990). Similarities and differences between KST and CDM are pointed out by Heller, Stefanutti, Anselmi, and Robusto (2015).

The third class of methods is *data-driven construction* of a knowledge structure (de Chiusole, Stefanutti, & Spoto, 2017; Falmagne, Albert, Doble, Eppstein, & Hu, 2013; Robusto & Stefanutti, 2014; Sargin & Ünlü, 2009; Schrepp, 1999a, 1999b, 2003; Spoto, Stefanutti, & Vidotto, 2016; Villano, 1991). This last group of methods has recently received attention and it relies on the extraction of knowledge structures from data. Data-driven approaches can be divided into two categories. Those belonging to one category impose specific properties to the knowledge structure underlying the data (e.g., closure under union). These methods can infer knowledge states that are not observed in the data set (e.g., Sargin & Ünlü, 2009; Schrepp, 1999b, 2003; Spoto et al., 2016). Methods in the other category do not impose any restriction. In Robusto and Stefanutti (2014) and Schrepp (1999a), only observed response patterns can be states of the constructed structure. On the other side, the k-states approach proposed by de Chiusole et al. (2017) can construct a knowledge structure by neither imposing restrictions on it, nor assuming that only observed patterns can be states. The k-states procedure is an incremental extension of the k-modes algorithm (Chaturvedi, Green, & Caroll, 2001; Huang & Ng, 1999) to knowledge structure extraction.

The core element shared by all the aforementioned methods is that they all refer to the construction of a structure on items that have dichotomous answer format. This answer format appears to be very well suited for assessing knowledge, but it might be inadequate when KST is applied to different fields such as psychological assessment,

attitude evaluations or medical diagnosis. These last applications have recently received an increased attention and they represent one of the core applicative developments of the theory (Bottesi, Spoto, Freeston, Sanavio, & Vidotto, 2015; Donadello et al., 2017; Serra, Spoto, Ghisi, & Vidotto, 2015, 2017; Spoto, Bottesi, Sanavio, & Vidotto, 2013; Spoto, Serra, Donadello, Granziol, & Vidotto, 2018; Spoto, Stefanutti, & Vidotto, 2010). Nonetheless, the possibility of a generalization of KST to the polytomous case has been explored so far only from a theoretical perspective (Bartl & Belohlavek, 2011; Schrepp, 1997; Stefanutti, Anselmi, de Chiusole, & Spoto, under review). The earliest attempt to generalize KST to polytomous items was carried out by Schrepp (1997). Unfortunately, his work had no sequel for almost twenty years and, mostly, it was never applied to real data. One of the possible reasons is that, for applying the theory in practice, some tools for building polytomous structures should be available.

No procedures are yet available based neither on the query to experts methodology, nor on the skill map approach, nor on the data-driven approach. In this article, a data-driven approach to the construction of a polytomous structure is proposed that is in line with the procedure recently developed by de Chiusole et al. (2017). The procedure is applicable to polytomous items, whose responses belong to an ordered set. Therefore, compared to the dichotomous case, a higher amount of information is contained in the data and this information can be used also in the construction of the structure. More specifically, while in the classical KST the k-modes procedure represented the only available solution for this kind of clustering, a polytomous structure (with levels defined on an ordinal scale) could refer to more informative clustering statistics, such as the *median*. The k-modes procedure can, with no harm, be applied to polytomous data, after having extended the number of categories from 2 to $n > 2$. However it cannot account for the presence of an eventual order on the response categories. Thus,

the present research is aimed at verifying the applicability and testing the performance of a k-median clustering method for extracting polytomous structures from data.

The article is organized as follows. In Section 2 the state of the art of the polytomous extension of KST, of the different dissimilarity measures that can be used as clustering statistics and of the k-modes algorithm is given. In Sections 3 an adaptation of the k-median clustering algorithm for extracting polytomous structures by the data is described. In Section 4 the relationships between the k-median and the k-modes algorithms are examined. In Sections 5 and 6 a simulation study and an empirical application are presented, respectively. In both of them the aim was to compare the performances of k-modes and k-median algorithms in extracting a structure from polytomous data. Final remarks are given in Section 7.

2 Backgrounds

A brief description of the notions needed in the subsequent sections are given. In particular, Section 2.1 introduces the basic concepts of polytomous KST. Section 2.2 describes some dissimilarity measures, that is the Hamming and the Manhattan distances. Finally, in Section 2.3 the k-modes algorithm for dichotomous data sets is presented.

2.1 Polytomous structures

As mentioned in the introduction, a very first attempt to extend KST to polytomous items was carried out by Schrepp (1997). In his proposal, for each item $q \in Q$ he introduced more than two response levels, taken from a linearly ordered set $(L, \leqslant)$, and indicating the "quality of the solution" for q. According to Schrepp, a *knowledge*

state is a mapping $K : Q \to L$, assigning to each item $q \in Q$ an element in L. Only cases where L is finite are considered in this article. The collection of all the mappings $K : Q \to L$ is denoted L^Q, and a *polytomous structure* is any nonempty subset $\mathcal{K} \subseteq L^Q$. The mappings in L^Q are partially ordered by the *pointwise* order $\sqsubseteq$ such that, given any two mappings $K_1, K_2 \in L^Q$,

$$K_1 \sqsubseteq K_2 \text{ iff } K_1(q) \leqslant K_2(q) \text{ for all } q \in Q.$$

In the subsequent sections, each mapping $K \in L^Q$ is represented as an n-tuple$(x_1, x_2, \ldots, x_n)$, where positions represent items in Q so that, for $v \in L$, $K(q) = v$ if and only if $x_q = v$.

2.2 Dissimilarity measures

Both the k-modes and the k-median algorithms rely upon appropriate dissimilarity measures for the data to which they apply. A metric that is often used to measure dissimilarity between sets in the dichotomous KST is the so-called Hamming distance. For any two subsets $A, B \subseteq Q$, it is defined as the cardinality of their symmetric difference:

$$d_H(A, B) = |A \Delta B| = |(A \setminus B) \cup (B \setminus A)|,$$

where $A \Delta B$ is the symmetric distance between the two sets A and B. Of course, the Hamming distance is also defined in the polytomous case: for any two mappings $X, Y \in L^Q$ let

$$d_H(X, Y) = \sum_{q \in Q} \delta_H(X(q), Y(q)), \tag{1}$$

where, $\delta_H : L \times L \to \mathbb{R}$ is such that, for $i, j \in L$,

$$\delta_H(i,j) = \begin{cases} 1 & \text{if } i \neq j \\ 0 & \text{if } i = j. \end{cases}$$

The Hamming distance is a measure of dissimilarity which is independent of how the elements in the set L are ordered. If the order changes (thus changing also the order on the elements in L^Q), this does not affect the Hamming distance among the elements in L^Q.

A discrepancy measure that depends on the order on the elements in L is the Manhattan distance. Given two mappings $X, Y \in L^Q$, the Manhattan distance equals the length of the shortest path connecting X to Y in the Hasse diagram of the partially ordered set $(L^Q, \sqsubseteq)$. The Manhattan distance between X and Y is obtained as

$$d_M(X,Y) = \sum_{q \in Q} \delta_M(X(q), Y(q)), \tag{2}$$

where, $\delta_M : L \times L \to \mathbb{R}$ is such that, for $i, j \in L$,

$$\delta_M(i,j) = |i^{\downarrow} \Delta j^{\downarrow}|,$$

and, for $l \in L$, $l^{\downarrow} = \{i \in L : i \leqslant l\}$ is the down set generated by l in the totally ordered set L. From this definition it follows that the Manhattan distance depends on the order on L.

It is worth noticing that, given a fixed set of classes, each represented by a state in the structure $\mathcal{K}$, the two distances may lead to different partitions of the same set of observed patterns into the classes. This difference strictly depends on the ordered set

L of levels. Such order is taken into account by the Manhattan distance to calculate the absolute value of the difference between the elements of the vectors, while it is not considered by the Hamming distance which aims at calculating the number of different elements between two vectors, irrespectively of the value of their distance.

For example, consider two states $K_1, K_2 \in \mathcal{K}$ and two response patterns $R_1, R_2 \in \mathcal{R}$, such that $K_1 = (1, 1, 1, 2)$, $K_2 = (1, 2, 1, 0)$, $R_1 = (0, 2, 1, 0)$, $R_2 = (1, 2, 0, 3)$. Then we have:

$$d_H(R_1, K_2) = 1 < d_H(R_1, K_1) = 3$$

and

$$d_H(R_2, K_2) = 2 < d_H(R_2, K_1) = 3.$$

The Hamming distance assigns both R_1 and R_2 to the class of K_2, because this last is at a minimum distance from both. On the other hand,

$$d_M(R_1, K_2) = 1 < d_M(R_1, K_1) = 4,$$

and

$$d_M(R_2, K_2) = 4 > d_M(R_2, K_1) = 3.$$

Hence the Manhattan distance assigns R_1 to the class of K_2 and R_2 to the class of K_1. Comparing the two dissimilarity measures, the Manhattan distance classifies the two response patterns into two separate classes, whereas the Hamming distance does not discriminate between them (see Figure 1).

INCLUDE FIGURE 1 ABOUT HERE

The opposite situation is possible too, as illustrated in the following example. Let

$R_1 = (0, 2, 1, 0)$, $R_2 = (0, 1, 0, 1)$, $K_1 = (1, 2, 1, 0)$, $K_2 = (0, 1, 0, 6)$. Then we have:

$$d_H(R_1, K_1) = 1 < d_H(R_1, K_2) = 3$$

and

$$d_H(R_2, K_1) = 4 > d_H(R_2, K_2) = 1.$$

Therefore, the Hamming distance assigns pattern R_1 to the class of K_1 and pattern R_2 to the class of K_2. On the other side

$$d_M(R_1, K_1) = 1 < d_M(R_1, K_2) = 8$$

and

$$d_M(R_2, K_1) = 4 < d_M(R_2, K_2) = 5.$$

Therefore, both R_1 and R_2 are assigned to the class of K_1 by using a Manhattan distance. In this case, the Hamming distance discriminates the two response patterns, whereas the Manhattan distance does not (see Figure 2).

INCLUDE FIGURE 2 ABOUT HERE

These examples show that pattern classification strongly depends on the type of dissimilarity measure that is used.

2.3 The k-modes algorithm for dichotomous data

The standard k-modes algorithm is an adaptation of the k-means (Hartigan & Wong, 1979) paradigm to the categorical data case. The k-modes algorithm is characterized

by: (i) the use of a simple matching dissimilarity measure, (ii) the use of mode statistics to center clusters according to the observed frequencies in the data set.

In this section, the adaptation of the k-modes algorithm developed by de Chiusole et al. (2017) is described. This procedure was provided for extracting a knowledge structure from a set of responses of N individuals to dichotomous items. Denote with the pair $(\mathcal{R}, F)$ the observed data set, where $\mathcal{R} = 2^Q$ is the power set of Q and $F : \mathcal{R} \to \mathbb{R}$ is a function assigning observed frequencies to response patterns, with $F(R) \geq 0$ for all $R \in \mathcal{R}$, and $\sum_{R \in \mathcal{R}} F(R) = N$.

Given a starting knowledge structure $\mathcal{K}_1$ on Q, this version of the k-modes algorithm operates in $m > 0$ iterations. In each of them the algorithm accomplishes two tasks (let $i = 1, 2, \ldots, m$ indicate iteration number):

(KM1) given $\mathcal{K}_i$, classify the N observed response patterns into $|\mathcal{K}_i|$ different clusters, each of which is uniquely represented by a state $K \in \mathcal{K}_i$;

(KM2) adjust each state $K \in \mathcal{K}_i$ in order to minimize the mean discrepancy between K and the patterns in the class designated by K. At this point $\mathcal{K}_{i+1}$ is the collection of the adjusted states.

Step (KM1) is called "pattern classification", whereas step (KM2) is called "centroid adjustment". Concerning (KM1), given a knowledge structure $\mathcal{K}$ on Q, the *partition function* $f : \mathcal{R} \times \mathcal{K} \to \mathbb{R}$ assigns the N observed patterns to $|\mathcal{K}|$ classes and satisfies the following two conditions:

(C1) $f(R, K) \geq 0$ for all $R \in \mathcal{R}$ and $K \in \mathcal{K}$,

(C2) $\sum_{K \in \mathcal{K}} f(R, K) = F(R)$ for all $R \in \mathcal{R}$.

Therefore, given each $K \in \mathcal{K}$ and each $R \in \mathcal{R}$, f assigns $f(R,K)$ out of $F(R)$ occurrences of response pattern R to the class represented by K. Moreover, the condition C2 assures that every occurrence of R is assigned to some state K.

The main task of step (KM1) is to detect the partition function f that minimizes the sum of the within-class dissimilarities

$$D_f(\mathcal{R},\mathcal{K}) = \sum_{K\in\mathcal{K}} \sum_{R\in\mathcal{R}} f(R,K) d_H(R,K). \tag{3}$$

This is obtained by setting $f(R,K) > 0$ if and only if $d_H(R,K) = \hat{d}_H(R,\mathcal{K})$, where

$$\hat{d}_H(R,\mathcal{K}) = \min\{d_H(R,K') : K' \in \mathcal{K}\} \tag{4}$$

is the minimum discrepancy between R and the states in $\mathcal{K}$ (de Chiusole et al., 2017). We observe that the discrepancy is measured by using the Hamming distance.

Concerning step (KM2), that is the knowledge state adjustment, for $i > 0$, $\mathcal{K}_i$ is the knowledge structure obtained at iteration i of the algorithm, and f_i is any minimum discrepancy partition function for $\mathcal{K}_i$. A state $K_i \in \mathcal{K}_i$ is adjusted as follows. If the class of K_i is empty or $K_i \in \{Q, \emptyset\}$ then $K_{i+1} = K_i$. Otherwise, a new state K_{i+1} is obtained as the element-wise mode in the class represented by K_i.

Let $\mathcal{K}_{i+1}$ be the collection of all adjusted states, and let f_i be any minimum discrepancy partition function for $\mathcal{K}_i$. It is proven that $D_{f_i}(\mathcal{R},\mathcal{K}_{i+1})$ is minimal in the sense that there is no knowledge structure $\mathcal{K}'$ on Q, with $|\mathcal{K}'| = |\mathcal{K}_{i+1}|$ such that $D_{f_i}(\mathcal{R},\mathcal{K}') < D_{f_i}(\mathcal{R},\mathcal{K}_{i+1})$.

It has to be stressed how in both (KM1) and (KM2) the overall discrepancy $D_f(\mathcal{R},\mathcal{K})$ is minimized. More specifically, at each iteration $i > 0$, in (KM1) the knowl-

edge structure $\mathcal{K}_i$ is set fixed and the partition function f_{i-1} is replaced by a minimum discrepancy partition function f_i for $\mathcal{K}_i$, thus obtaining $D_{f_i}(\mathcal{R}, \mathcal{K}_i) \leq D_{f_{i-1}}(\mathcal{R}, \mathcal{K}_i)$. Moreover, in (KM2) the partition function f_i is set fixed and the knowledge structure $\mathcal{K}_i$ is adjusted so that $D_{f_i}(\mathcal{R}, \mathcal{K}_{i+1}) \leq D_{f_i}(\mathcal{R}, \mathcal{K}_i)$. The algorithm terminates when this difference is zero, or below some tolerance value.

This version of the k-modes approach is well suited for the case of unordered categorical data and it applies to dichotomous data. In this case, in fact, the mode is the only admissible clustering statistics. In the next section the case in which the measurement level of the item responses is ordinal is explored.

3 A k-medians algorithm for polytomous structures

The clustering algorithm described in this section is an adaptation of k-modes to ordinal data in which the Hamming distance is replaced by a Manhattan distance, and the central tendency measure is the median, rather than the mode. Like the k-modes algorithm it consists of an iteration of two fundamental steps: the *pattern classification step*, and the *centroid adjustment step*.

Let L be a finite set of linearly ordered levels. The pattern classification step consists of partitioning the whole set of observed patterns into the classes represented by the centroids in $\mathcal{K} \subseteq L^Q$ so that the intraclass dissimilarity is minimum. Given an observed frequency distribution $F : L^Q \to \mathbb{R}$, with $F(R) \geq 0$ for all $R \in L^Q$, a partition of L^Q is any function $f : L^Q \times \mathcal{K} \to \mathbb{R}$ such that $f(R, K) \geq 0$ for all $R \in L^Q$ and all $K \in \mathcal{K}$ and

$$\sum_{K \in \mathcal{K}} f(R, K) = F(R),$$

for all $R \in L^Q$. The goal is to find a partition f for which the sum of the intraclass dissimilarities

$$D_f(\mathcal{K}) = \sum_{K \in \mathcal{K}} \sum_{R \in L^Q} f(R, K) d_M(R, K)$$

is minimum.

Pattern classification is based on a suitable minimum discrepancy partition function, which is defined as follows. The Manhattan-based minimum discrepancy between an observed pattern $R \in L^Q$ and a whole structure $\mathcal{K} \subseteq L^Q$, is

$$\hat{d}_M(R, \mathcal{K}) = \min\{d_M(R, K) : K \in \mathcal{K}\}. \tag{5}$$

A *Manhattan-based minimum discrepancy partition function* is any partition function $f_M : L^Q \times \mathcal{K} \to \mathbb{R}$ satisfying, for all $R \in L^Q$ and $K \in \mathcal{K}$,

$$f_M(R, K) > 0 \iff d_M(R, K) = \hat{d}_M(R, K)$$

Thus $f_M(R, K)$ distributes the observed frequency of R among states in $\mathcal{K}$ that are at a minimum discrepancy from it, when this discrepancy is measured by a Manhattan distance.

The objective of the centroid adjustment step is to update the class centroids $K \in \mathcal{K}$, so that the updated set $\mathcal{K}'$ of centroids is such that the difference $D_f(\mathcal{K}) - D_f(\mathcal{K}')$ is maximum among all possible choices of $\mathcal{K}'$ having the same size as $\mathcal{K}$. For every single centroid $K \in \mathcal{K}$ the updating consists of replacing K with the "elementwise median" $\hat{K} \in L^Q$ of the class represented by K. The elementwise median is defined as follows. Denote with

$$\mathcal{C}^K = \{R \in L^Q : f_M(R, K) > 0\}$$

the class represented by state K. Moreover, given any level $l \in L$, and any item $q \in Q$, let

$$\mathcal{C}_{ql}^{K} = \{R \in \mathcal{C}^{K} : R(q) = l\}$$

the collection of all the patterns in $\mathcal{C}^{K}$ assigning level l to item q. Finally, define

$$f_M(\mathcal{C}^{K}) = \sum_{R \in \mathcal{C}^{K}} f_M(R, K),$$

and

$$f_M(\mathcal{C}_{ql}^{K}) = \sum_{R \in \mathcal{C}_{ql}^{K}} f_M(R, K).$$

We observe that $f_M(\mathcal{C}^{K})$ is the size of (number of response patterns in) class $\mathcal{C}^{K}$, whereas $f_M(\mathcal{C}_{ql}^{K})$ is the total number of patterns in $\mathcal{C}^{K}$ assigning level l to item q.

Then level $l \in L$ is *median* at item $q \in Q$ in the class $\mathcal{C}^{K}$ if both the following conditions hold true:

$$\sum_{i=0}^{l} f_M(\mathcal{C}_{qi}^{K}) \geq \frac{1}{2} f_M(\mathcal{C}^{K})$$

and

$$\sum_{i=0}^{l-1} f_M(\mathcal{C}_{qi}^{K}) < \frac{1}{2} f_M(\mathcal{C}^{K}),$$

where $l-1$ is the lower neighbor of l in the linearly ordered set L. Then, the elementwise median $\hat{K}$ is obtained by taking the median at q in $\mathcal{C}_K$ for all $q \in Q$.

An algorithm based on these two steps has been implemented in MATLAB and the code can be found in the Supplementary Materials.

4 Relationships between k-median and k-modes

Recalling that L is a finite linearly ordered set, let $L^+ = L \setminus \{\min L\}$. In this section, it is shown that there exists a precise isometry between the "polytomous" metric space (d_M, L^Q) and the "dichotomous" metric space $(d_H, 2^{Q\times L^+})$. By way of such an isometry, subsets of "polytomous patterns" in L^Q can be represented by appropriate subsets of dichotomous patterns in $2^{Q\times L^+}$. In practice, any polytomous data set can be bijectively converted into a dichotomous one, in which distances are preserved. The isometry is then used to show that an application of the k-median algorithm to a polytomous data set and an application of the k-modes algorithm to the dichotomous representation of the same data set will essentially produce the same results.

4.1 Polytomous versus dichotomus representation

A *dichotomous representation* of a mapping $K \in L^Q$ is any injective function $r : L^Q \to 2^{Q\times L^+}$. The *nominal* and *ordinal* dichotomous representations are considered here. The nominal dichotomous representation is a function *nom* such that, for $K \in L^Q$,

$$nom(K) = \{(q, l) \in Q \times L^+ : l = K(q)\}.$$

The ordinal dichotomous representation is a function *ord* such that

$$ord(K) = \{(q, l) \in Q \times L^+ : l \leqslant K(q)\}.$$

Both functions are injective. Let $nom(L^Q) = \{nom(K) : K \in L^Q\}$ denote the image of L^Q under *nom*, and $ord(L^Q)$ its image under *ord*. Then the two functions $n : L^Q \to nom(L^Q)$ and $o : L^Q \to ord(L^Q)$ such that, for $K \in L^Q$, $n(K) = nom(K)$ and

$o(K) = ord(K)$ are bijections.

Theorem 1. *The ordinal dichotomous representation ord is an order-homomorphism of L^Q into $2^{Q\times L^+}$. In particular, the function o is an order-isomorphism.*

Proof. It suffices to show that, for $K_1.K_2 \in L^Q$ the equivalence

$$K_1 \sqsubseteq K_2 \iff o(K_1) \subseteq o(K_2)$$

holds true. Necessity: suppose $K_1 \sqsubseteq K_2$. Then $K_1(q) \leqslant K_2(q)$ for all $q \in Q$. From $(q, K_1(q)) \in o(K_1)$, $K_1(q) \leqslant K_2(q)$ and $(q, K_2(q)) \in o(K_2)$ we infer $(q, K_1(q)) \in o(K_2)$. Thus $o(K_1) \subseteq o(K_2)$. Sufficiency: suppose $o(K_1) \subseteq o(K_2)$ and let $(q, l) \in o(K_1)$ be such that $l' \leqslant l$ for all $(q, l') \in o(K_1)$, so that $K_1(q) = l$. Since $o(K_1) \subseteq o(K_2)$ we have $(q, l) \in o(K_2)$, entailing that $K_1(q) = l \leqslant K_2(q)$. Since this holds for any arbitrary $q \in Q$, we conclude that $K_1 \sqsubseteq K_2$. □

The nominal dichotomous representation is not a homomorphism in general. To give an example, let $K_1 = (1, 2, 0)$ and $K_2 = (1, 3, 0)$. Then $nom(K_1) = \{(1, 1), (2, 2)\}$ whereas $n(K_2) = \{(1, 1), (2, 3)\}$. Thus we see that $K_1 \sqsubseteq K_2$, but $n(K_1) \not\subseteq n(K_2)$.

We are here interested in dichotomous representations that preserve distances in a well defined sense. A dichotomous representation r is an *isometry* (or distance preserving) if there are two metrics μ_1 on L^Q and μ_2 on $2^{Q\times L^+}$ such that

$$\mu_1(K_1, K_2) = \mu_2(r(K_1), r(K_2))$$

for all $K_1, K_2 \in L^Q$. The function r is a *global isometry* if it is bijective.

Theorem 2. *The ordinal dichotomous representation ord is an isometry of L^Q into $2^{Q\times L^+}$. Moreover, the function o is a global isometry. More precisely, for $K_1, K_2 \in L^Q$ it holds that*

$$d_H(o(K_1), o(K_2)) = d_M(K_1, K_2).$$

Proof. For $K_1, K_2 \in L^Q$, let $O_1 = o(K_1)$ and $O_2 = o(K_2)$. For $q \in Q$ define $q^* = \{q\} \times L^+ = \{(q, l) : l \in L^+\}$. If $h(K_1(q)) \leq h(K_2(q))$, then $O_1 \cap q^* \subseteq O_2 \cap q^*$, with $|(O_2 \cap q^*) \setminus (O_1 \cap q^*)| = h(K_2(q)) - h(K_1(q))$ and $|(O_1 \cap q^*) \setminus (O_2 \cap q^*)| = 0$. Thus

$$|(O_2 \cap q^*) \setminus (O_1 \cap q^*)| + |(O_1 \cap q^*) \setminus (O_2 \cap q^*)| = |h(K_2(q)) - h(K_1(q))|.$$

It should be observed that the equality holds even in the case $h(K_1(q)) \geq h(K_2(q))$. Summing up for all $p \in Q$,

$$\sum_{p\in Q} (|(O_2 \cap p^*) \setminus (O_1 \cap p^*)| + |(O_1 \cap p^*) \setminus (O_2 \cap p^*)|) = \sum_{p\in Q} |h(K_2(p)) - h(K_1(p))|$$

which gives

$$\sum_{p\in Q} |((O_2 \cap p^*) \setminus (O_1 \cap p^*)) \cup ((O_1 \cap p^*) \setminus (O_2 \cap p^*))| = d_M(K_1, K_2).$$

Observing that $\{O_1 \cap p^* : p \in Q\}$ is a partition of O_1, $\{O_2 \cap p^* : p \in Q\}$ is a partition of O_2 we can write

$$\left| \bigcup_{p\in Q} ((O_2 \cap p^*) \setminus (O_1 \cap p^*)) \cup \bigcup_{p\in Q} ((O_1 \cap p^*) \setminus (O_2 \cap p^*)) \right| = d_M(K_1, K_2)$$

and then

$$|(O_1 \setminus O_2) \cup (O_2 \setminus O_1)| = d_M(K_1, K_2).$$

But the left-hand side term of this last equation is the Hamming distance between $O_1 = o(K_1)$ and $O_2 = o(K_2)$, hence $d_H(o(K_1), o(K_2)) = d_M(K_1, K_2)$. □

We observe that the nominal dichotomous representation n is not an isometry in general. Considering again $K_1 = (1, 2, 0)$ and $K_2 = (1, 3, 0)$ we have $d_M(K_1, K_2) = 1$, whereas $d_H(n(K_1), n(K_2)) = d_H(\{(1, 1), (2, 2)\}, \{(1, 1), (2, 3)\}) = 2$. We do not obtain an isometry even when d_M is replaced by d_H for measuring distances in L^Q. In fact $d_H(K_1, K_2) = 1$.

As long as the binary representation is o, the isometry is only obtained if the Manhattan distance is mapped to the Hamming distance. A Hamming to Hamming mapping would not work: $d_H(o(K_1), o(K_2)) \neq d_H(K_1, K_2)$ in general. This observation follows at once from Theorem 2 and from the fact that $d_H(K_1, K_2) \neq d_M(K_1, K_2)$ in general.

4.2 k-modes mimics k-median

For $X \in ord(L^Q)$ and $\mathcal{O} = ord(\mathcal{K})$, define the minimum discrepancy function

$$\hat{d}_H(X, \mathcal{O}) = \min\{d_H(X, Y) : Y \in \mathcal{O}\}.$$

A Hamming-based minimum discrepancy partition is any function $f_H : ord(L^Q) \times ord(\mathcal{K}) \to \mathbb{R}$ satisfying:

(1) $f_H(X, Y) > 0$ iff $d_H(X, Y) = \hat{d}_H(X, \mathcal{O})$ for all $X \in ord(L^Q)$ and $Y \in \mathcal{O}$;

(2) $\sum_{Y \in ord(\mathcal{K})} f_H(X, Y) = F(X)$ for all $X \in ord(L^Q)$,

where $F(X) = F(o^{-1}(X))$ is the observed frequency of X.

Theorem 3. *The function f_H defined by $f_H(o(R), o(K)) = f_M(R, K)$ for all $R \in L^Q$ and all $K \in \mathcal{K}$, is a Hamming-based minimum discrepancy partition.*

Proof. The fact that $\hat{d}_H(o(R), ord(\mathcal{K})) = \hat{d}_M(R, \mathcal{K})$ for all $R \in L^Q$ follows at once from Theorem 2. Then, for $X \in ord(L^Q)$ and $Y \in ord(\mathcal{K})$ we have $f_H(X, Y) > 0$ iff $f_M(o^{-1}(X), o^{-1}(Y)) > 0$ iff $d_M(o^{-1}(X), o^{-1}(Y)) = \hat{d}_M(o^{-1}(X), \mathcal{K})$ iff $d_H(X, Y) = \hat{d}_H(X, ord(\mathcal{K}))$. Hence f_H is a (Hamming-based) minimum discrepancy partition. □

The theoretical results obtained so far show that, in the pattern classification step, a Manhattan distance based pattern classification and its ordinal dichotomous representation are equivalent. In the remainder of the section we show that an equivalence exists also in the centroid adjustment step. Let $o(L^Q)$ be the image of L^Q in $2^{Q \times L^+}$ under the transformation o. Let moreover $(\mathcal{O}, G)$ be such that $\mathcal{O}$ is any nonempty subset of $o(L^Q)$ and $G : \mathcal{O} \to \mathbb{R}^+$ is a frequency distribution. For $q \in Q$ and $l \in L^+$, denote with

$$\mathcal{O}_{ql} = \{X \in \mathcal{O} : (q, l) \in X\}$$

the collection of all subsets in $\mathcal{O}$ containing the pair (q, l). Then the pair $(q, l) \in Q \times L^+$ is *modal* in $(\mathcal{O}, G)$ if it belongs to the majority of the subsets in $\mathcal{O}$, that is, when the condition

$$\sum_{X \in \mathcal{O}_{ql}} G(X) \geq \frac{1}{2} \sum_{Y \in \mathcal{O}} G(Y).$$

is satisfied. Then we have the following

Theorem 4. *Let $\mathcal{C}$ be any nonempty subset of L^Q. Then the pair $(q, m) \in Q \times L^+$ is modal in $o(\mathcal{C})$ if and only if m is median at q in $\mathcal{C}$.*

Proof. It follows from the definition of the function o that, for $R \in L^Q$ and $(q,l) \in Q \times L^+$, $(q,l) \in o(R) \iff l \leqslant R(q)$. Therefore, recalling that o is a bijection, and letting $\mathcal{O} = o(\mathcal{C})$,

$$o^{-1}(\mathcal{O}_{ql}) = \{R \in o^{-1}(\mathcal{O}) : l \leqslant R(q)\}.$$

Hence

$$o^{-1}(\mathcal{O}_{ql}) = \{R \in \mathcal{C} : l \leqslant R(q)\}.$$

Defining $l^\uparrow = \{i \in L^+ : l \leqslant i\}$, this can be written as

$$o^{-1}(\mathcal{O}_{ql}) = \bigcup_{i \in l^\uparrow} \mathcal{C}_{qi}.$$

Thus we can write:

$$\begin{aligned}\sum_{X \in \mathcal{O}_{ql}} F(o^{-1}(X)) &= \sum_{R \in o^{-1}(\mathcal{O}_{ql})} F(R) \\ &= \sum_{R \in \bigcup_{i \in l^\uparrow} \mathcal{C}_{qi}} F(R) \\ &= \sum_{i \in l^\uparrow} \sum_{R \in \mathcal{C}_{qi}} F(R).\end{aligned}$$

Thus, letting $G = F \circ o^{-1}$, the pair (q,l) is modal in $(\mathcal{O}, G)$ if and only if

$$\sum_{i \in l^\uparrow} \sum_{R \in \mathcal{C}_{qi}} F(R) \geq \frac{1}{2} \sum_{Y \in o(\mathcal{C})} G(Y) = \frac{1}{2} \sum_{Z \in \mathcal{C}} F(Z).$$

Let $m \in L^+$ be such that, for $q \in Q$,

$$\sum_{X \in \mathcal{O}_{qm}} G(X) \geq \frac{1}{2} \sum_{Y \in \mathcal{O}} G(Y),$$

that is, m is the maximum level in L^+ at which (q, m) is modal in $(\mathcal{O}, G)$. Furthermore, denote by $m+1$ the upper neighbor of m in the linearly ordered set L. Then

$$\begin{aligned}\sum_{X\in\mathcal{O}_{qm}} G(X) \geq \frac{1}{2}\sum_{Y\in\mathcal{O}} G(Y) &\iff \sum_{i\in m^{\uparrow}}\sum_{R\in\mathcal{C}_{qi}} F(R) \geq \frac{1}{2}\sum_{R\in\mathcal{C}} F(R)\\ &\iff \sum_{i\in m^{\downarrow}\setminus\{m\}}\sum_{R\in\mathcal{C}_{qi}} F(R) < \frac{1}{2}\sum_{R\in\mathcal{C}} F(R)\end{aligned}$$

and

$$\begin{aligned}\sum_{X\in\mathcal{O}_{q,m+1}} G(X) < \frac{1}{2}\sum_{Y\in\mathcal{O}} G(Y) &\iff \sum_{i\in(m+1)^{\uparrow}}\sum_{R\in\mathcal{C}_{qi}} F(R) < \frac{1}{2}\sum_{R\in\mathcal{C}} F(R)\\ &\iff \sum_{i\in m^{\downarrow}}\sum_{R\in\mathcal{C}_{qi}} F(R) \geq \frac{1}{2}\sum_{R\in\mathcal{C}} F(R).\end{aligned}$$

showing that (q, m) is modal in $\mathcal{O}$ iff m is median at q in $o^{-1}(\mathcal{O})$. □

Suppose the polytomous data set $(\mathcal{R}, F)$ is transformed into a dichotomous one by means of the ordinal dichotomous representation o, obtaining the data set $\mathcal{O} = o(\mathcal{R})$. Taken together, Theorems 3 and 4 imply that an application of the k-median algorithm to $(\mathcal{R}, F)$ is equivalent to an application of the k-modes algorithm to the ordinal dichotomous representation $(\mathcal{O}, F)$.

5 Simulation Study: Comparison between k-modes and k-median algorithms

In this study, the performance of k-modes and k-median algorithms were compared in extracting a structure from polytomous data. The aim was to investigate in which

conditions (i.e., error in the data, sample size, etc.) they perform equally well and in which others they perform differently. In particular, the interest was to find if there are circumstances in which one algorithm performs better than the other.

5.1 Aims and hypotheses

Theorems 3 and 4 defined in Section 4.1 imply that an application of the k-median algorithm to a polytomous data set $(\mathcal{R}, F)$ is equivalent to an application of the k-modes algorithm to the ordinal dichotomous representation $(\mathcal{O}, F)$ of those data. Thus, the first aim was to provide evidence for this theoretical result.

One of the core assumptions of the probabilistic framework of the dichotomous KST, is that the response pattern of a subject and her knowledge state can differ in some items. In particular, it is plausible to assume that some *careless errors* (i.e., the item belongs to the state but not to the response pattern) and/or some *lucky guesses* (i.e., the item belongs to the response pattern but not to the state) might arise.

In the polytomous framework of KST, such type of errors have to be redefined. Given any item q and any two levels $i, j \in L$, let $p_{qij} = P(R(q) = j|K(q) = i)$ be the conditional probability that in the pair $\langle R, K\rangle$, response pattern R assigns level j to item q, given that state K assigns level i to q. Thus, for each item q and for each level $i \in L$, $\sum_{j\in L} p_{qij} = 1$. It is worth noticing that when $R(q) = K(q)$ no item error is observed, when $R(q) > K(q)$ the observed response is an *overrate* of the true response, whereas when $R(q) < K(q)$ the observed response is an *underrate* of the true response. Since the item response scale L is assumed to be ordinal, it is plausible that, as the "distance" of the observed level from the true one increases, its probability decreases. We call this assumption *δ–monotonicity*.

Let $\delta : L \times L \to \mathbb{R}$ be a metric on the set L of levels. A function $f : L \times L \to (0, 1)$ is said to be δ–*monotone* if the implication

$$\delta(i, j) \leq \delta(i, k) \implies f(i, j) \geq f(i, k)$$

holds true for all $i, j, k \in L$. The form of δ–monotonicity much depends on the chosen metric δ. When L is a finite set, the valid metric is the Hamming distance. If, furthermore, L is totally ordered then the Manhattan distance can also be applied. In the sequel, a function f is δ_H–monotone if the distance is Hamming and it is δ_M–monotone if the distance is Manhattan.

Given any metric δ, we say that the probabilities p_{qij}, $q \in Q, i, j \in L$, respect δ–monotonicity if the following implication holds for all $q \in Q$ and all $i, j, k \in L$:

$$\delta(i, j) \leq \delta(i, k) \iff p_{qij} \geq p_{qik}.$$

Furthermore, we say that a population respects δ–monotonicity if this last is respected by the probabilities p_{qij} in that population.

The hypothesis of the simulation study is that, when the data set is extracted from a population respecting δ_M–monotonicity, the k-median algorithm performs better than k-modes. This is based on the fact that the median is a central tendency index that minimizes the Manhattan distance, whereas the mode minimizes the Hamming distance.

5.2 Data set generation process and simulation design

Let Q be the set of items and $(L, \leqslant)$ be a linearly ordered set, where $L = \{0, 1, \cdots, l\}$ is the set of the response categories of the items $q \in Q$. Each response pattern R of a data set was generated as follows. Let $\perp_{|Q|}$ be the $|Q|$-elements vector of zeros, and $\top_{|Q|} = (l, l, \cdots, l)$. A structure $\mathcal{K}_t$ and a probability distribution $\pi_{\mathcal{K}_t}$ on $\mathcal{K}_t$ were assumed. The structure $\mathcal{K}_t$ was obtained by computing $\{\perp_{|Q|}, \top_{|Q|}\} \cup \mathcal{P}$, where $\mathcal{P}$ was generated at random, using a sampling without replacement on the collection $L^{|Q|} \setminus \{\perp_{|Q|}, \top_{|Q|}\}$. It was assumed that $\pi_{\mathcal{K}_t}$ was the uniform distribution, so that each state $K \in \mathcal{K}_t$ had the probability $1/|\mathcal{K}_t|$ to be observed.

Thereafter, a certain number N of pairs $\langle R, K \rangle$ were generated. For each pair, the state K was sampled with replacement from the collection $\mathcal{K}_t$ of states, by using $\pi_{\mathcal{K}_t}$ as the multinomial distribution; whereas, the response pattern R was obtained through the introduction of some amount of error in K. Given any item q and any two levels $i, j \in L$, let $p_{qij} = P(R(q) = j | K(q) = i)$ be the conditional probability that in the pair $\langle R, K \rangle$, response pattern R assigns level j to item q, given that state K assigns level i to q.

The data were generated in two different scenarios. In the former, both δ_H– and δ_M–monotonicity were respected by all probabilities p_{qij}. In the latter, only δ_H–monotonicity was respected. For each item q and each level i, the probabilities p_{qij} were generated by using the following procedure. First the probability p_{qii} was randomly sampled from a uniform distribution in the interval (a, b), $0 < a < b < 1$. Then, a $(|L| - 1)$-element vector $\mathbf{x}$ was generated by randomly sampling each of its elements from the uniform distribution in the interval $(0, 1)$. Then the $(|L| - 1)$-element vector

of the probabilities p_{qij}, for $i \neq j$ was obtained by the linear transformation

$$\frac{(1 - p_{qii})\mathbf{x}}{\sum_{t=1}^{|L|-1} \mathbf{x}_t}.$$

In this way, for each level $i \in L$, $\sum_{j \in L} p_{qij} = 1$. Concerning the probabilities p_{qij}, $i \neq j$, in the former scenario, they were assigned to the different levels j in such a way that as the distance $\delta_M(i, j)$ increases, the probability p_{qij} decreases. Whereas, in the latter scenario they were assigned to the different levels j in such a way that as the distance $\delta_M(i, j)$ increases, the probability p_{qij} increases. Hence, in the former scenario the δ_H-monotonicity assumption held in the data, whereas, in the latter scenario the δ_M-monotonicity held.

In both scenarios, a domain Q of 15 items and an item response scale L of four levels were assumed. The data sets were generated by manipulating three different variables:

- the number of states in the true structure: $|\mathcal{K}_t| \in \{500, 1000\}$;
- the probability p_{qii}, that was at least $\{.95, .85, .75\}$ for each item $q \in Q$ and for each level $i \in L$;
- the sample size N: $2,000$ or $5,000$.

In the whole, a number of 2 (structures) $\times$ 3 (p_{qii} probabilities) $\times$ 2 (sample sizes) $\times$ 2 (monotonicity) = 24 different conditions were considered. Table 1 summarizes the simulation design.

INCLUDE TABLE 1 ABOUT HERE

For each condition, 100 different samples were generated.

5.3 Performance Measures

The k-modes and k-median algorithms were applied to each of the $24 \times 100 = 2,400$ samples, for recovering the true structures $\mathcal{K}_t$. Furthermore, the k-modes algorithm was applied to the ordinal dichotomous representation of each sample. In this way, for each sample, three different structures were extracted: $\mathcal{K}_e^1$ was extracted by k-modes applied to polytomous data, $\mathcal{K}_e^2$ was extracted by k-medians applied to polytomous data, and $\mathcal{K}_e^3$ was extracted by k-modes applied to the ordinal dichotomous representation of the polytomous data (see Figure 6).

INCLUDE FIGURE 6 ABOUT HERE

Recalling that the algorithms improve an initial set $\mathcal{I}$ of states, this set was the same for both algorithms and it was obtained for each of the $2,400$ simulated samples by a random selection of a number of $|\mathcal{K}_t|$ response patterns. This choice was made for allowing the algorithms to operate in the same situation.

The comparison between the performances of the two algorithms was accomplished by using two different indexes. Both of them compare the true structure $\mathcal{K}_t$ with the structure $\mathcal{K}_e^i$, with $i \in \{1, 2, 3\}$, extracted from the data by a particular algorithm. These indexes were:

1. The *true positive rate* (TPR), that is the proportion of true states $K \in \mathcal{K}_t$ belonging to the extracted structure $\mathcal{K}_e^i$. Formally:

$$\text{TPR} = \frac{|\mathcal{K}_e^i \cap \mathcal{K}_t|}{|\mathcal{K}_t|}; \tag{6}$$

2. The *Manhattan* and the *Hamming average discrepancies* between the set of the

states $K \in \mathcal{K}_e^i$ and the true structure $\mathcal{K}_t$, that is:

$$d(\mathcal{K}_e^i, \mathcal{K}_t) = \frac{1}{|\mathcal{K}_e^i|} \sum_{K \in \mathcal{K}_e^i} d(K, \mathcal{K}_t), \tag{7}$$

where $d(K, \mathcal{K}_t) = \min_{K' \in \mathcal{K}_t} \delta(K, K')$ is the minimum of the distance between the state $K \in \mathcal{K}_e^i$ and the true structure $\mathcal{K}_t$. Such distance can be either Manhattan ($\delta = \delta_M$) or Hamming ($\delta = \delta_H$).

In each condition of the study, average values (and the corresponding standard deviations) of both indexes were computed across the 100 simulated samples. Then, the performance indexes obtained by using the two algorithms were compared to one another.

5.4 Results

Table 2 displays the results for all the 24 conditions obtained by applying the k-modes (top 12 rows of the table) and the k-median (the middle 12 rows of the table) algorithms to the polytomous data, and by applying the k-modes algorithm to the ordinal dichotomous transformation of the polytomous data (the bottom 12 rows of the table).

[INSERT TABLE 2 ABOUT HERE]

Column 1 displays the number of the first 12 conditions in which the scenario was the δ_M-monotonicity, whereas Column 5 displays the number of the second 12 conditions in which the scenario was the δ_H-monotonicity. For each index of TPR, $d_M(\mathcal{K}_e^i, \mathcal{K}_t)$ (denoted in the table as d_M) and $d_H(\mathcal{K}_e^i, \mathcal{K}_t)$ (denoted in the table as d_H), the average values (standard deviations) computed across the 100 simulated samples are displayed,

respectively, in Columns 2 to 4, for the δ_M-monotonicity scenario, and in Columns 6 to 8 for the δ_H-monotonicity scenario.

In order to facilitate comparisons between the performances of the two algorithms, Figure 3 and Figure 4 have been produced from results displayed in Table 2. Figure 3 shows the variation of the TPR as the sample size, the cardinality of the true structure and the amount of error increase. Each panel of the figure displays the amount of error on the x-axis. Solid, dashed and dotted lines refer, respectively, to k-modes and k-median algorithms applied to polytomous data and to k-modes applied to the ordinal dichotomous transformation of the polytomous data. Each of the two groups of curves represents the results obtained with the two cardinalities $|\mathcal{K}_t| = 500$ or 1,000. Left panels show the results obtained when δ_M-monotonicity holds in the data (that is, Condition numbers 1 to 12 of Table 2), whereas right panels show the results when δ_H-monotonicity holds in the data (that is, Condition numbers 13 to 24 of Table 2). Upper panels are referred to conditions in which the sample size was $2,000$, instead bottom panels are referred to conditions in which the sample size was $5,000$.

[INSERT FIGURE 3 ABOUT HERE]

From the figure, several interesting results can be pointed out. By comparing the performance of the same algorithm in the two different scenarios δ_M-monotonicity (panels on the left) and δ_H-monotonicity (panels on the right), it can be noted that the k-modes algorithm (solid lines) performs equally well, obtaining same percentages of TPR in most conditions. This is because the modality condition (that is the only requirement for the k-modes algorithm) is respected in both the δ_H- and δ_M-monotonicity scenarios. Differently, the k-median algorithm (dotted lines) improves, as expected, its performance in passing from a δ_H-monotonicity to a δ_M-monotonicity scenario, espe-

cially when the amount of error in the data increases.

Concerning the k-modes algorithm applied to the ordinal dichotomous transformation of polytomous data (dashed lines), it obtains, as expected, performances that are essentially equal to those obtained by the k-median algorithm (dotted lines) applied to polytomous data. This is true regardless of the amount of error in the data, of the sample size and of the scenario.

By comparing the k-modes (solid lines) and the k-median (dotted lines) algorithms, it is clear that, when the δ_H-monotonicity holds in the data, k-modes performs better than k-median. The opposite is observed in the δ_M-monotonicity scenario, where the k-median algorithm performs better, in almost all the conditions. The difference between the two is negligible with sample size $N = 2,000$, but it is considerable with sample size $N = 5,000$, especially when the amount of error in the data increases.

Figure 4 is obtained by fixing the sample size to $2,000$ and it displays the results on the average discrepancies $d_M(\mathcal{K}_e^i, \mathcal{K}_t)$ and $d_H(\mathcal{K}_e^i, \mathcal{K}_t)$, where $i \in \{1, 2, 3\}$, between the structures $\mathcal{K}_e^i$ extracted by the algorithms and the true structures $\mathcal{K}_t$. In the figure, upper panels refer to the Manhattan discrepancy, bottom panels refer to the Hamming average discrepancy, left panels refer to the δ_M-monotonicity scenario and right panels refers to δ_H-monotonicity scenario. In each panel, the discrepancy is plotted as a function of the amount of error in the data. The reference of the curve types is the same of Figure 3.

INCLUDE FIGURE 4 ABOUT HERE

From the results obtained by the discrepancy measures, other interesting comments arise. By looking at the Manhattan average discrepancy (upper panels), the results on the performance of the algorithms are coherent with those obtained for the TPR index.

Indeed, when the δ_M-monotonicity holds in the data, the k-median algorithms (dotted lines) obtains a better performance than the one of k-modes (solid lines), extracting structures that have a smaller distance to the true one. This is particularly true when the amount of error is high. This trend is reversed in the δ_H-monotonicity scenario, in which k-modes (solid lines) performs better than k-median (dotted lines). Moreover, the k-modes algorithm applied to the ordinal transformation of the polytomous data (dashed lines) obtains, as expected, very similar results as those obtained by k-median.

By looking at the Hamming average discrepancy (bottom panels), it can be noted that k-modes algorithm obtains very similar results in the two scenarios. Instead, the other two obtain systematically lower distances, in passing from the δ_H-monotonicity to δ_M-monotonicity. Again, when the data are δ_M-monotone, the k-median algorithms and that of k-modes applied to the ordinal dichotomous transformation of polytomous data improve their performances.

By fixing the monotonicity "type", the Manhattan and the Hamming discrepancies for the k-modes algorithm applied to the ordinal dichotomous transformation of the polytomous data are equal one another. Moreover, the Hamming discrepancy for the k-modes algorithm applied to the ordinal dichotomous transformation of the polytomous data is very close to the Manhattan discrepancy obtained for the k-median algorithm. The former result only confirms that when data are dichotomous the Manhattan and the Hamming distance are exactly the same. The latter results confirms what stated by Theorems 3 and 4.

There seems to be a disagreement between the upper- and the lower-left panels concerning the performances of the two algorithms. In the upper panel the performance of k-median is slightly better than that of k-modes. Whereas, in the lower panel very similar performances are obtained. The only difference between the two panels is the

metric used for assessing performances, that is, Manhattan in the upper panel and Hamming in the lower panel. The disagreement suggests that one of the two has to be wrong. The Hamming distance is not affected by the order on the levels, in the sense that any permutation of this order would result in the same Hamming average discrepancy. This is not the case for the Manhattan discrepancy. While the order in the δ_H-monotonicity is arbitrary, in the δ_M-monotonicity there is only one order. Thus, the correct metric is the Manhattan distance.

5.5 Conclusion

Concerning the aims and the hypothesis formulated at the beginning of this section, both of them are reached. In particular, by applying the k-modes algorithm to the ordinal dichotomous transformation of polytomous data, similar results to those obtained by applying the k-median algorithm to the polytomous data are obtained by using both the TPR and the discrepancy measures. Furthermore, simulation results showed that when the δ_M-monotonicity holds in the data, the k-median algorithm seems to overperform the k-modes algorithm, confirming our hypothesis. Summarizing, it seems that the scenario in which the δ_M-monotonicity condition is respected by the data is the more natural one for applying the k-median algorithm.

6 Empirical Application

The main aim was to apply both the k-modes and the k-median algorithms for extracting a polytomous structure from a pre-existing data set. It has to be highlighted that the two algorithms can only be applied if the cardinality of the structure (number of clusters) is known. In an empirical application this is usually not the case. Therefore,

a set of increasing cardinalities were evaluated and a "best" one was selected according to distance measures. Moreover, a cross-validation procedure has been used in order to avoid issues of overfit of the extracted structures. The details of all these aspects are described below.

6.1 The questionnaire and the data

The application was carried out starting from a set of answers to the Italian version of the reduced form of the State Trait Anxiety Inventory form Y (STAI-Y; Pedrabissi & Santinello, 1989). It is a psychological self-report questionnaire investigating the "state" anxiety and it consists of 10 items rated on a 4-point Likert scale, ranging from "not at all" (coded as 0) to "very much" (coded as 3), so that $L = \{0, 1, 2, 3\}$. Six items of the test have a positive wording (i.e., the higher the level, the higher the state anxiety), whereas the others have a negative wording (i.e., the higher the level, the lower the state anxiety). The responses to the latter items were re-scored prior to analyses.

The sample was composed of $N = 3,673$ participants that signed the informed consent and were asked to answer to all the items of the questionnaire. No time limit was imposed.

6.2 Methods

Following cross-validation techniques, the sample was randomly partitioned into two complementary sets: a *training set* $\mathcal{T}$, composed of about the two thirds of the participants and used for extracting the structure; a *validation set* $\mathcal{V}$, composed by the remaining participants and used for testing the extracted structure.

One hundred different bipartitions $\{\mathcal{V}_j, \mathcal{T}_j\}$, $j \in \{1, \ldots, 100\}$ of the original data set were generated and for each of them twenty different conditions were considered. What varied across the twenty conditions was the cardinality of the initial set $\mathcal{I}_j$ of states on which the algorithms started, which varied in the interval $\{50, 100, \ldots, 1000\}$ with a step of 50. The two algorithms k-modes and k-median where applied to each training set $\mathcal{T}_j$.

In each of the 100 bipartitions and in each of the twenty conditions, the two polytomous structures $\mathcal{K}_{k\text{-modes}}$ and $\mathcal{K}_{k\text{-median}}$ were extracted from the data through the two algorithms. A total number of 100 (bipartitions) $\times$ 20 (conditions) $\times$ 2 (algorithms) $= 4,000$ different structures were obtained.

The performances of the algorithms and, thus, the corresponding extracted structures can be compared by using the two statistics $D(\mathcal{R}, \mathcal{K})$ and $D(\mathcal{K}, \mathcal{R})$. They allow to compute the "distance" between the validation set $\mathcal{V}_i$ and the extracted structures. A description of how the two distances $D(\mathcal{R}, \mathcal{K})$ and $D(\mathcal{K}, \mathcal{R})$ can be computed follows.

For each $R \in \mathcal{R}$, the minimum Hamming discrepancy $\hat{d}_H(R, \mathcal{K})$ and the minimum Manhattan discrepancy $\hat{d}_M(R, \mathcal{K})$ were computed as in Equations (4) and (5), respectively. Performing this computation for all patterns in $\mathcal{R}$, a frequency distribution of the minimum discrepancies is obtained. This distribution is referred to as the *Hamming discrepancy distribution* (or, respectively, as the *Manhattan discrepancy distribution*) for the data set $\mathcal{R}$ and the polytomous structure $\mathcal{K}$.

Finally, a *discrepancy index* from $\mathcal{R}$ to $\mathcal{K}$ is obtained by computing the mean of this discrepancy distribution. For $i \in \{M, H\}$,

$$D_i(\mathcal{R}, \mathcal{K}) = \frac{1}{N} \sum_{R \in \mathcal{R}} \hat{d}_i(R, \mathcal{K}) F(R). \tag{8}$$

The $D_i(\mathcal{R},\mathcal{K})$ discrepancy equals zero whenever $\mathcal{R} \subseteq \mathcal{K}$. For the Hamming distance, the theoretical maximum of this index is $\lfloor |Q|/2 \rfloor$. For the Manhattan distance the theoretical maximum distance for each item $q \in Q$ between $R(q)$ and $K(q)$ is $\lfloor (|L|-1)/2 \rfloor$. The theoretical maximum value of the discrepancy $D_i(\mathcal{R},\mathcal{K})$ is obtained when the structure is $\{\perp_{|Q|}, \top_{|Q|}\}$, and its value depends on both the number of levels $l \in L$ and the number of items $q \in Q$ according to the following general formula:

$$D_i(\mathcal{R},\mathcal{K})_{max} = \left\lfloor \frac{|Q|}{2} \right\rfloor \left(\left\lfloor \frac{|L|-1}{2} \right\rfloor + \left\lceil \frac{|L|-1}{2} \right\rceil \right) + \left(|Q| - 2 \left\lfloor \frac{|Q|}{2} \right\rfloor \right) \left\lfloor \frac{|L|-1}{2} \right\rfloor .$$

It has to be noted that in general $\lfloor (|L|-1)/2 \rfloor + \lceil (|L|-1)/2 \rceil = |L|-1$. Therefore, the formula could be rewritten as

$$D_i(\mathcal{R},\mathcal{K})_{max} = \left\lfloor \frac{|Q|}{2} \right\rfloor (|L|-1) + \left(|Q| - 2 \left\lfloor \frac{|Q|}{2} \right\rfloor \right) \left\lfloor \frac{|L|-1}{2} \right\rfloor . \tag{9}$$

Let $|Q|$ be an even number and $|L|$ be an odd number. Then the state at maximum distance from $\{\perp_{|Q|}, \top_{|Q|}\}$ has all the items at distance $(|L|-1)/2$ from any of the two extreme states. Thus

$$D_i(\mathcal{R},\mathcal{K})_{max} = |Q| \frac{|L|-1}{2}.$$

In this case, the second addend of Equation (9) is 0, since whenever $|Q|$ is even we have that $|Q| = (2\lfloor (|Q|)/2 \rfloor)$. Along these lines, if both $|Q|$ and $|L|$ are even numbers, then the state at a maximum distance would have half of its states at a distance $(|L|-1)/2$ from $\perp_{|Q|}$ and half of the items at the same distance from $\top_{|Q|}$. The resulting formula is

$$D_i(\mathcal{R},\mathcal{K})_{max} = \frac{|Q|}{2} \left(\left\lfloor \frac{|L|-1}{2} \right\rfloor + \left\lceil \frac{|L|-1}{2} \right\rceil \right),$$

that is again

$$D_i(\mathcal{R},\mathcal{K})_{max} = |Q|\frac{|L|-1}{2}.$$

Moreover, if both $|L|$ and $|Q|$ are odd numbers, the state at maximum distance from $\{\perp_{|Q|}, \top_{|Q|}\}$ has a number $(|Q|-1)$ of items that are at distance $(|L|-1)/2$ from any of the extreme states. Moreover, since the number of items is odd we have that $|Q| - (2\lfloor(|Q|)/2\rfloor) = 1$. Thus, there must be a further item necessarily at a distance $(|L|-1)/2$ from any of the two extreme states. Thus,

$$D_i(\mathcal{R},\mathcal{K})_{max} = (|Q|-1)\frac{|L|-1}{2} + \frac{|L|-1}{2}.$$

Finally, when $|L|$ is even and $|Q|$ is odd we have that

$$D_i(\mathcal{R},\mathcal{K})_{max} = \left\lfloor \frac{|Q|}{2} \right\rfloor \left(\left\lfloor \frac{|L|-1}{2} \right\rfloor + \left\lceil \frac{|L|-1}{2} \right\rceil \right) + \left\lfloor \frac{|L|-1}{2} \right\rfloor,$$

which is Equation (9).

The statistic $D_i(\mathcal{K},\mathcal{R})$, comparing the extracted structure $\mathcal{K}$ with the set $\mathcal{R}$ of response patterns, is obtained by:

$$D_i(\mathcal{K},\mathcal{R}) = \frac{1}{|\mathcal{K}|}\sum_{K\in\mathcal{K}} \hat{d}_i(K,\mathcal{R}), \tag{10}$$

where $\hat{d}_i(K,\mathcal{R}) = \min\{d_i(K,R) : R \in \mathcal{R}, F(R) > 0\}$.

For each of the $4,000$ extracted polytomous structures the statistics $D_i(\mathcal{K},\mathcal{R})$ and $D_i(\mathcal{R},\mathcal{K})$ have been computed by using both the Manhattan and the Hamming discrepancy indexes.

With the aim of selecting the "best structure", a discrepancy index based on the

two discrepancies $D_i(\mathcal{R}, \mathcal{K})$ and $D_i(\mathcal{K}, \mathcal{R})$ was computed as follows (in the sequel we drop the index i for notational convenience). Let $\mathbf{K}$ be the collection of all the knowledge structures obtained by one of the k-modes and k-median algorithms. For each knowledge structure $\mathcal{K} \in \mathbf{K}$ define the discrepancy

$$V(\mathcal{K}, \mathcal{R}) = \max\{D(\mathcal{R}, \mathcal{K}), D(\mathcal{K}, \mathcal{R})\}.$$

Furthermore, given any $z \geq 0$, define the set

$$\mathbf{H}_z = \{\mathcal{K} \in \mathbf{K} : V(\mathcal{K}, \mathcal{R}) \leq z\}.$$

This set contains all knowledge structures $\mathcal{K} \in \mathbf{K}$ for which both distances $D(\mathcal{R}, \mathcal{K})$ and $D(\mathcal{K}, \mathcal{R})$ are less or equal to a given z. We seek at finding the smallest value z_{min} for which the set $\mathbf{H}_{z_{min}}$ is nonempty:

$$z_{min} = \min\{z \in \mathbb{R} : \mathbf{H}_z \neq \emptyset\}.$$

The structures contained in $\mathbf{H}_{z_{min}}$ are regarded as "best" in the sense that for any other structure not in $\mathbf{H}_{z_{min}}$ at least one of the two distances $D(\mathcal{R}, \mathcal{K})$ or $D(\mathcal{K}, \mathcal{R})$ is higher than z_{min}.

Proposition 1. *The smallest value z_{min} such that $\mathbf{H}_{z_{min}}$ is nonempty is the smallest value of $V(\mathcal{K}, \mathcal{R})$ across all $\mathcal{K} \in \mathbf{K}$, that is*

$$\begin{aligned} z_{min} &= \min_{\mathcal{K} \in \mathbf{K}} V(\mathcal{K}, \mathcal{R}) \\ &= \min_{\mathcal{K} \in \mathbf{K}} \max\{D(\mathcal{R}, \mathcal{K}), D(\mathcal{K}, \mathcal{R})\}. \end{aligned}$$

Proof. Let $v_{min} = \min_{\mathcal{K}\in\mathbf{K}} V(\mathcal{K},\mathcal{R})$ and suppose $z_{min} > v_{min}$. Then there must be $\mathcal{K} \in \mathbf{K}$ with $V(\mathcal{K},\mathcal{R}) = v_{min}$ and hence $\mathbf{H}_{v_{min}} \neq \emptyset$, which contradicts the definition of z_{min}. Hence $z_{min} \leq v_{min}$. Suppose now $v_{min} > z_{min}$. Then there is $\mathcal{K} \in \mathbf{H}_{z_{min}} \subseteq \mathbf{K}$ with $V(\mathcal{K},\mathcal{R}) < v_{min}$, which contradicts the definition of v_{min}. Hence $v_{min} \leq z_{min}$. Thus $v_{min} = z_{min}$. □

Proposition 1 provides a practical way of obtaining z_{min}. It is just the minimum of the discrepancy $V(\mathcal{K},\mathcal{R})$. The knowledge structures $\mathcal{K} \in \mathbf{K}$ for which $V(\mathcal{K},\mathcal{R}) = z_{min}$ are regarded as "best". According to a parsimony principle, if more than a single knowledge structure satisfy this condition, the one with the smallest cardinality is selected.

The procedure described so far was applied by using both $\mathbf{K}_{k\text{-modes}}$ and $\mathbf{K}_{k\text{-median}}$ and both the Manhattan and the Hamming distance, obtaining thus four different "best" structures.

6.3 Results

Figure 5 shows the results obtained by the two algorithms for the z_{min} index. The left panel of the figure displays the results obtained by using the Manhattan discrepancy index, whereas right panel shows those obtained by using the Hamming discrepancy index. Solid and dotted lines refer to the performances of k-modes and k-median algorithms applied to the (original) polytomous data, respectively. Dashed line refers to the performance of k-modes applied to the ordinal dichotomous transformation of the data. The cardinality of the initial set of states $\mathcal{I}$ are on the x-axis.

[INSERT FIGURE 5 ABOUT HERE]

The results obtained on the two discrepancy indexes are quite similar, with the exception of the results obtained with k-modes applied to the dichotomous transformation of the data (i.e., for the Manhattan distance its performance is quite similar to those of k-median, for the Hamming distance its performance is the worst). This is an expected result since the Manhattan and the Hamming distances, computed on the dichotomous data are the same.

Comparing the results of the algorithms when they are applied to the original data, the one extracting the structure with the smallest value of z_{min} is k-median, whereas k-modes obtained a worse performance. Nevertheless, both the algorithms extracted a structure having the same cardinality, that is 150. Finally, it should be highlighted that the value obtained for the z_{min} is remarkably low (i.e., less than 1.5) for both the algorithms.

Interpreting these results in view of what found in the simulations, it seems that the δ_M-monotonicity holds in the data. This means that there is evidence about the assumption of a linear order among the response categories of the item response scale L of the test (i.e., 4-point Likert scale, from "not at all" to "very much").

7 Final remarks

In the present research, an adaptation of the k-median clustering algorithm for extracting (knowledge) structures from polytomous data has been proposed. It is an extension of k-modes to ordinal data in which the Hamming distance is replaced by the Manhattan distance, and the central tendency measure is the *median*, rather than the mode. Like the k-modes algorithm, it consists of an iteration of two steps. The former is the *pattern classification step* and it consists of partitioning the whole set of observed

patterns into the classes represented by centroids that minimize the intraclass dissimilarity. In case of ordinal polytomous data the suitable minimum discrepancy function is based on the Manhattan distance. The latter is the *centroid adjustment step*, that consists in updating the centroids in order to minimize the Manhattan discrepancy from the data. The best centroid is the "elementwise" median of the class.

The algorithm has been implemented in MATLAB and it was used for running simulations. The performances of k-median and k-modes algorithms were compared with respect to their capabilities of reconstructing structures from polytomous data. In comparing their performances, two different scenarios were considered. In the former, the data were simulated assuming that the "true" level was the most likely. In the latter, it was moreover assumed that, the probability of a response decreases as it moves away from the true level. Results showed that in the latter scenario the k-median algorithm performed better than k-modes, confirming our hypothesis.

The k-modes and the k-median algorithms were applied also to a real polytomous data-set in which the responses to the items were given on an ordinal scale. Both the algorithms extracted a structure of the same size, but the one extracted by k-median was the closest to the data.

Overall, there are both theoretical and practical reasons for preferring the k-median to the k-modes algorithm whenever the responses to the items are measured on an ordinal scale. The central role is played by the distance, since between the Hamming and the Manhattan distances, only the latter is sensitive to the order on the levels.

Moreover, it seems that a way for testing the ordinal nature of an item response scale could consist in applying both k-modes and k-median to the data and checking whether k-median extracts the best structure. Future research should investigate this possibility.

To conclude, k-medians seems to be a promising data-driven procedure for building polytomous structures, providing polytomous KST with one of the tools necessary for its application to a wider range of fields.

Acknowledgments

The research developed in this article was carried out under the research project CPDR152105 "Learning how students learn. Mathematical modeling of learning processes in intelligent tutoring system navigation", funded by the Department of Philosophy, Sociology, Pedagogy and Applied Psychology, University of Padua, Italy.

Open practices statement

The data used in the empirical application (State Trait Anxiety Inventory form Y) are available from the authors upon request. The MATLAB algorithms implementing the k-modes and k-median procedures are also available from the authors upon request.

References

Albert, D., & Lukas, J. (Eds.). (1999). *Knowledge spaces: Theories, empirical research, and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bartl, E., & Belohlavek, R. (2011). Knowledge spaces with graded knowledge states. *Information Sciences*, *181*(8), 1426–1439.

Bolt, D. (2007). The present and future of IRT-based cognitive diagnostic models (ICDMs) and related methods. *Journal of Educational Measurement*, *44*(4), 377–383.

Bottesi, G., Spoto, A., Freeston, M. H., Sanavio, E., & Vidotto, G. (2015). Beyond the score: clinical evaluation through formal psychological assessment. *Journal of personality assessment*, *97*(3), 252–260.

Chaturvedi, A., Green, P. E., & Caroll, J. D. (2001). K-modes clustering. *Journal of Classification*, *18*(1), 35–55.

de Chiusole, D., & Stefanutti, L. (2013). Modeling skill dependence in probabilistic competence structures. *Electronic Notes in Discrete Mathematics*, *42*, 41–48.

de Chiusole, D., Stefanutti, L., & Spoto, A. (2017). A class of k-modes algorithms for extracting knowledge structures from data. *Behavior research methods*, *49*(4), 1212–1226.

DiBello, L. V., & Stout, W. (2007). Guest editors' introduction and overwiew: Irt-based cognitive diagnostic models and related methods. *Journal of Educational Measurement*, *44*(4), 285–291.

Doignon, J.-P. (1994). Knowledge spaces and skill assignments. In G. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics and methodology* (p. 111-121). New York: Springer-Verlag.

Doignon, J.-P., & Falmagne, J.-C. (1985). Spaces for the assessment of knowledge. *International Journal of Man-Machine Studies*, *23*, 175-196.

Doignon, J.-P., & Falmagne, J.-C. (1999). *Knowledge spaces*. New York: Springer.

Donadello, I., Spoto, A., Sambo, F., Badaloni, S., Granziol, U., & Vidotto, G. (2017). Ats-pd: an adaptive testing system for psychological disorders. *Educational and psychological measurement*, *77*(5), 792–815.

Dowling, C. E. (1993). On the irredundant generation of knowledge spaces. *Journal of Mathematical Psychology*, *37*(1), 49–62.

Falmagne, J.-C., Albert, D., Doble, C., Eppstein, D., & Hu, X. (2013). *Knowledge spaces: Applications in education*. Springer Science & Business Media.

Falmagne, J.-C., & Doignon, J.-P. (2011). *Learning spaces*. New York: Springer.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108.

Heller, J., Augustin, T., Hockemeyer, C., Stefanutti, L., & Albert, D. (2013). Recent developments in competence-based knowledge space theory. In *Knowledge spaces* (pp. 243–286). Springer.

Heller, J., Stefanutti, L., Anselmi, P., & Robusto, E. (2015). On the link between cognitive diagnostic models and knowledge space theory. *Psychometrika*, *80*(4), 995–1019.

Huang, Z., & Ng, M. K. (1999). A fuzzy k-modes algorithm for clustering categorical data. *Fuzzy Systems, IEEE Transactions on*, *7*(4), 446–452.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*, 258–272.

Kambouri, M., Koppen, M., Villano, M., & Falmagne, J.-C. (1994). Knowledge assessment: Tapping human expertise by the query routine. *International Journal of Human-Computer Studies*, *40*(1), 119–151.

Koppen, M. (1993). Extracting human expertise for constructing knowledge spaces: An algorithm. *Journal of mathematical psychology*, *37*(1), 1–20.

Koppen, M., & Doignon, J.-P. (1990). How to build a knowledge space by querying

an expert. *Journal of Mathematical Psychology*, *34*(3), 311–331.

Lukas, J., & Albert, D. (1993). Knowledge assessment based on skill assignment and psychological task analysis. In G. Strube & K. Wender (Eds.), *The cognitive psychology of knowledge* (p. 139-160). Amsterdam: North-Holland.

Müller, C. E. (1989). A procedure for facilitating an expert's judgements on a set of rules. In *Mathematical psychology in progress* (pp. 157–170). Springer.

Pedrabissi, L., & Santinello, M. (1989). Inventario per l'ansia di "stato" e di "tratto": nuova versione italiana dello stai forma y: Manuale. *Organizzazioni Speciali, Firenze*, 44.

Robusto, E., & Stefanutti, L. (2014). Extracting a knowledge structure from the data by a maximum residuals method. *TPM: Testing, Psychometrics, Methodology in Applied Psychology*.

Sargin, A., & Ünlü, A. (2009). Inductive item tree analysis: Corrections, improvements, and comparisons. *Mathematical Social Sciences*, *58*(3), 376–392.

Schrepp, M. (1997). A generalization of knowledge space theory to problems with more than two answer alternatives. *Journal of mathematical psychology*, *41*(3), 237–243.

Schrepp, M. (1999a). Extracting Knowledge Structures from Observed Data. *British Journal of Mathematical and Statistical Psychology*, *52*, 213—224.

Schrepp, M. (1999b). On the empirical construction of implications between bi-valued test items. *Mathematical social sciences*, *38*(3), 361–375.

Schrepp, M. (2003). A method for the analysis of hierarchical dependencies between items of a questionnaire. *Methods of Psychological Research Online*, *19*, 43–79.

Schrepp, M., & Held, T. (1995). A simulation study concerning the effect of errors on the establishment of knowledge spaces by querying experts. *Journal of Mathe-*

matical Psychology, *39*(4), 376–382.

Serra, F., Spoto, A., Ghisi, M., & Vidotto, G. (2015). Formal psychological assessment in evaluating depression: a new methodology to build exhaustive and irredundant adaptive questionnaires. *PloS one*, *10*(4), e0122131.

Serra, F., Spoto, A., Ghisi, M., & Vidotto, G. (2017). Improving major depressive episode assessment: A new tool developed by formal psychological assessment. *Frontiers in psychology*, *8*, 214.

Spoto, A., Bottesi, G., Sanavio, E., & Vidotto, G. (2013). Theoretical foundations and clinical implications of formal psychological assessment. *Psychotherapy and psychosomatics*, *82*(3), 197–199.

Spoto, A., Serra, F., Donadello, I., Granziol, U., & Vidotto, G. (2018). New perspectives in the adaptive assessment of depression: the ats-pd version of the queds. *Frontiers in Psychology*, *9*, 1101.

Spoto, A., Stefanutti, L., & Vidotto, G. (2010). Knowledge space theory, formal concept analysis, and computerized psychological assessment. *Behavior Research Methods*, *42*(1), 342-350.

Spoto, A., Stefanutti, L., & Vidotto, G. (2016). An iterative procedure for extracting skill maps from data. *Behavior research methods*, *48*(2), 729–741.

Stefanutti, L., Anselmi, P., de Chiusole, D., & Spoto, A. (under review). On the polytomous generalization of knowledge space theory. *Journal of Mathematical Psychology*.

Stefanutti, L., & de Chiusole, D. (2017). On the assessment of learning in competence based knowledge space theory. *Journal of Mathematical Psychology*, *80*, 22–32.

Tatsuoka, K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In *Monitoring skills and knowledge acquisition* (pp. 453–488).

Hillsdale: Lawrence Erlbaum Associates.

Villano, M. (1991). Computerized knowledge assessment: Building the knowledge structure and calibrating the assessment routine.

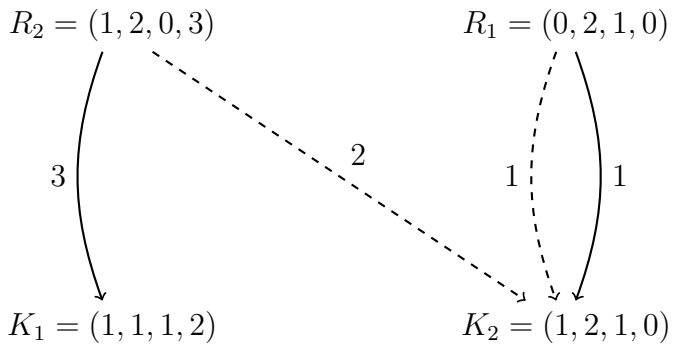

Figure 1: Comparison between the two dissimilarity measures Hamming (dashed lines) and Manhattan (solid lines) distances in classifying response patterns R_1 and R_2 into the two classes of states K_1 and K_2. The Manhattan distance classifies them into two separate classes, whereas the Hamming distance does not discriminate between them.

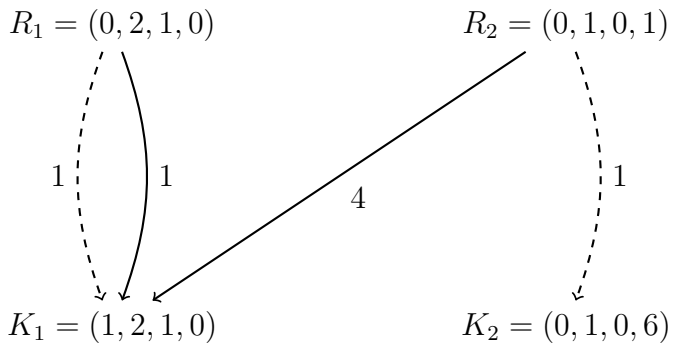

Figure 2: Comparison between the two dissimilarity measures Hamming (dashed lines) and Manhattan (solid lines) distances in classifying the response patterns R_1 and R_2 in to the two classes of states K_1 and K_2. The Manhattan distance does not discriminate between the two response patterns, whereas the Hamming distance classifies them into two different classes.

Table 1: Sample design used in the simulation study leading to 24 different conditions. Columns 1, 2 and 3 report, respectively, the values of the sample size N, the true structure's cardinality $|\mathcal{K}_t|$ and the maximum error Err used for generating the data. Columns 4 and 6 display a number identifying each of the 24 conditions, and columns 5 and 7 display the type of monotonicity considered, between the δ_H- and the δ_M-monotonicity.

N	$\lvert\mathcal{K}_t\rvert$	Err	Cond.	Monotonicity	Cond.	Monotonicity
2,000	500	.05	1	δ_M	13	δ_H
2,000	500	.15	2	δ_M	14	δ_H
2,000	500	.25	3	δ_M	15	δ_H
2,000	1000	.05	4	δ_M	16	δ_H
2,000	1000	.15	5	δ_M	17	δ_H
2,000	1000	.25	6	δ_M	18	δ_H
5,000	500	.05	7	δ_M	19	δ_H
5,000	500	.15	8	δ_M	20	δ_H
5,000	500	.25	9	δ_M	21	δ_H
5,000	1000	.05	10	δ_M	22	δ_H
5,000	1000	.15	11	δ_M	23	δ_H
5,000	1000	.25	12	δ_M	24	δ_H

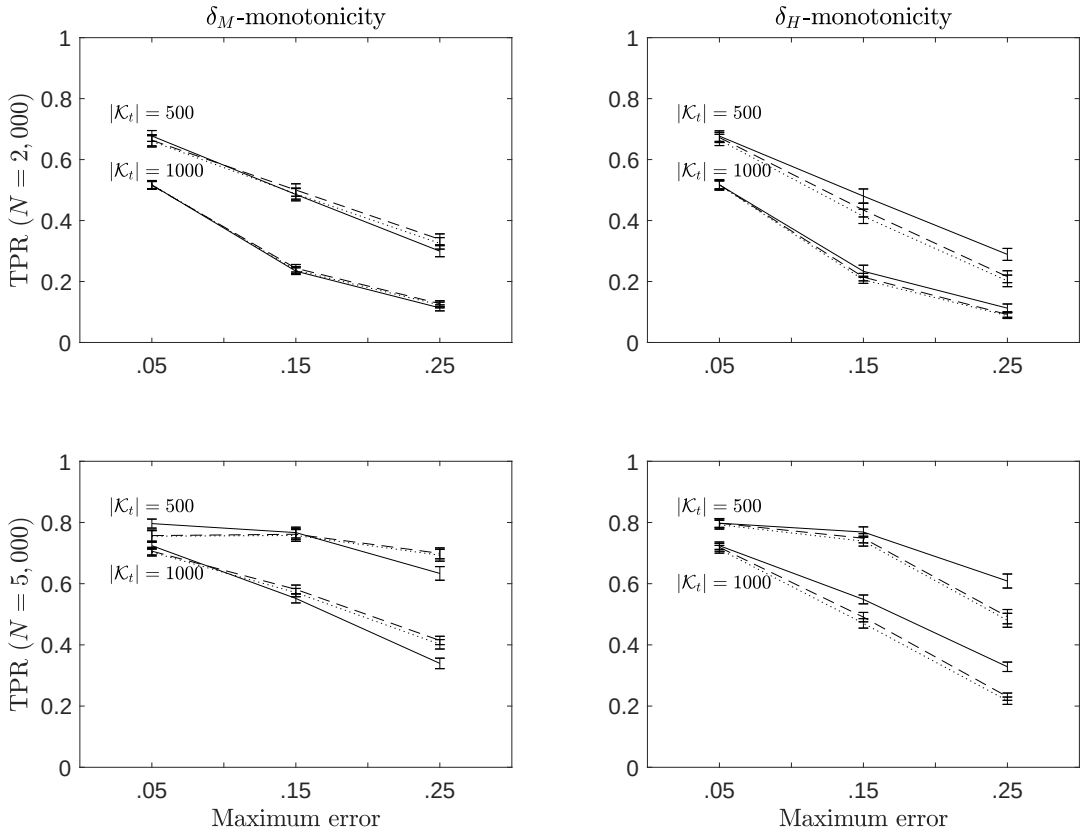

Figure 3: Variation of the TPR as the sample size (row panels of the figure) and the amount of error (x-axis) increase. Solid, dashed and dotted lines refer, respectively, to k-modes and k-median algorithms applied to polytomous data and to k-modes applied to the ordinal dichotomous transformation of the polytomous data. The two groups of curves represent the results obtained with a different cardinality of the true structure $|\mathcal{K}_t|$. Left panels show the results obtained when the δ_M-monotonicity holds in the data, whereas right panels show the results obtained when the δ_H-monotonicity holds in the data.

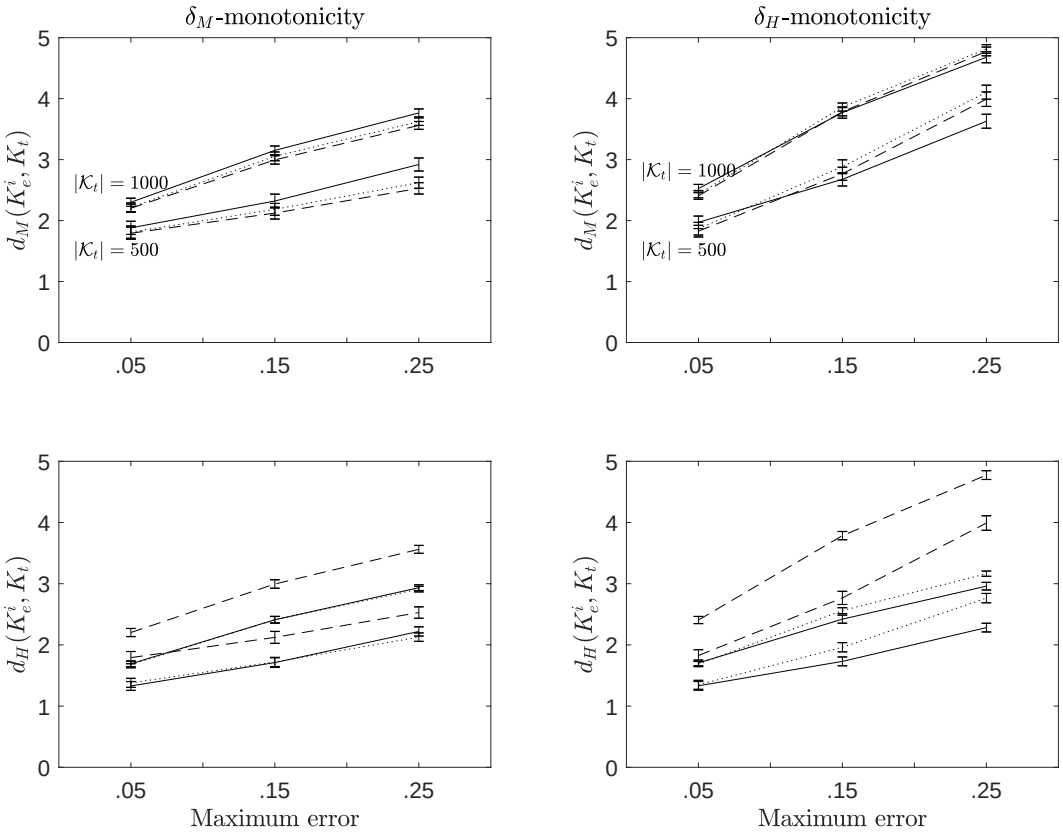

Figure 4: Variation of the average discrepancies $d_M(\mathcal{K}_e^i, \mathcal{K}_t)$ (upper panels) and $d_H(\mathcal{K}_e^i, \mathcal{K}_t)$ (bottom panels) as the amount error (x-axis) increases. The sample size was fixed to $2,000$. Left panels refer to the δ_M-monotonicity scenario, whereas right panels refers to δ_H-monotonicity scenario. Solid, dashed and dotted lines refer, respectively, to k-modes and k-median algorithms applied to polytomous data and to k-modes applied to the ordinal dichotomous transformation of the polytomous data. The two groups of curves represent the results obtained with a different cardinality of the true structure $|\mathcal{K}_t|$.

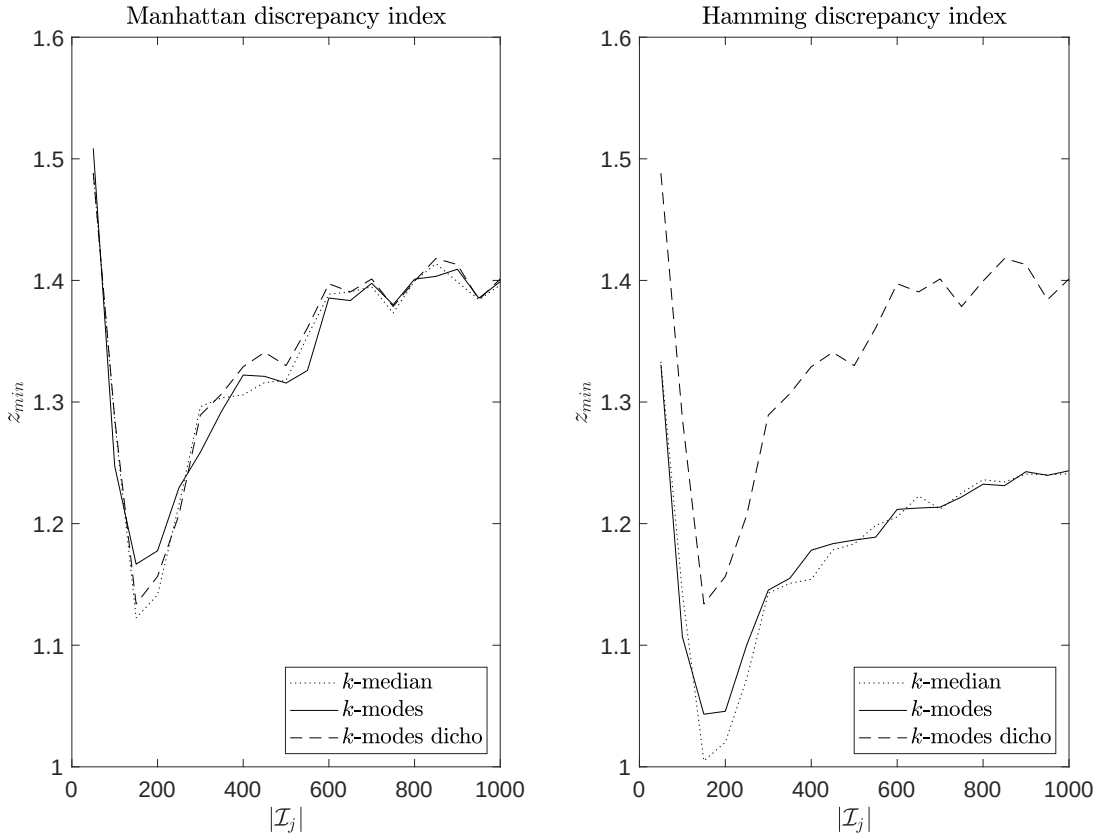

Figure 5: Results of the empirical application obtained by the k-modes and the k-median algorithms for the z_{min} index. Left panel of the figure displays the results obtained by using the Manhattan discrepancy index, whereas the right panel shows those obtained by using the Hamming discrepancy index. Solid and dashed lines refer to the performances of k-modes and k-median algorithms applied to polytomous data, respectively. The dashed line refers to k-modes applied to the ordinal dichotomous transformation of polytomous data (called "k-modes dicho" in the legend). The cardinalities of the initial set $\mathcal{I}_j$ are on the x-axis.

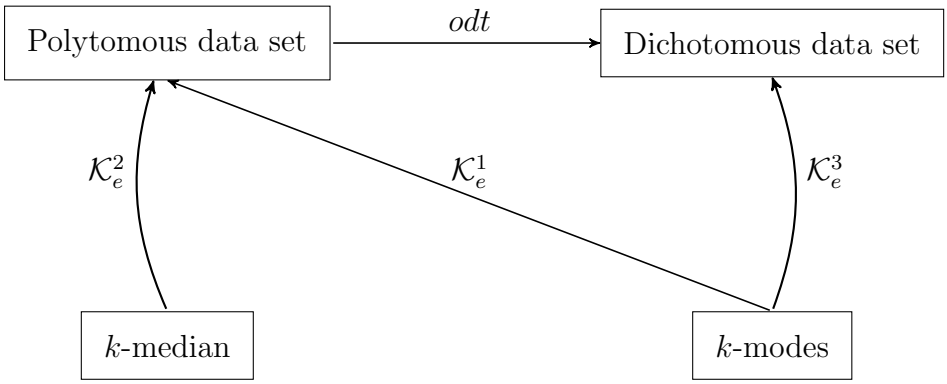

Figure 6: Design used in the simulation study (Section 5) for applying k-modes and k-median algorithms to the poltytomous data and for applying the k-modes algorithm to the ordinal dichotomous transformation (*odt*) of them. The three structures $\mathcal{K}_e^1$, $\mathcal{K}_e^2$ and $\mathcal{K}_e^3$ have been extracted.

Table 2: Results of the 24 conditions of the simulation study obtained by applying the k-modes (first 12 rows of the table) and the k-median (second 12 rows of the table) algorithms to the polytomous data, and by applying the k-modes algorithm to the ordinal dichotomous transformation of the polytomous data (last 12 rows of the table). See text for the details.

	δ_M-monotonicty				δ_H-monotonicty		
C	TPR	d_M	d_H	C	TPR	d_M	d_H
1	.68 (.02)	1.88 (.11)	1.33 (.07)	13	.68 (.02)	1.97 (.10)	1.33 (.07)
2	.49 (.02)	2.32 (.11)	1.71 (.08)	14	.48 (.02)	2.68 (.11)	1.73 (.07)
3	.30 (.02)	2.92 (.11)	2.22 (.08)	15	.29 (.02)	3.63 (.11)	2.28 (.07)
4	.52 (.01)	2.30 (.07)	1.69 (.05)	16	.52 (.02)	2.53 (.07)	1.71 (.05)
5	.23 (.01)	3.15 (.07)	2.41 (.05)	17	.23 (.02)	3.77 (.09)	2.42 (.07)
6	.11 (.01)	3.77 (.07)	2.93 (.05)	18	.11 (.01)	4.68 (.09)	2.96 (.06)
7	.80 (.01)	1.43 (.10)	.98 (.07)	19	.80 (.01)	1.42 (.09)	.96 (.06)
8	.77 (.02)	1.42 (.11)	1.00 (.07)	20	.77 (.02)	1.44 (.10)	.98 (.07)
9	.63 (.02)	1.74 (.12)	1.28 (.08)	21	.61 (.02)	2.06 (.13)	1.36 (.09)
10	.72 (.01)	1.59 (.07)	1.12 (.05)	22	.72 (.01)	1.64 (.06)	1.12 (.04)
11	.55 (.01)	2.00 (.07)	1.47 (.05)	23	.55 (.02)	2.25 (.08)	1.49 (.05)
12	.34 (.01)	2.68 (.07)	2.05 (.05)	24	.33 (.02)	3.31 (.09)	2.13 (.06)
1	.66 (.02)	1.81 (.10)	1.38 (.07)	13	.66 (.02)	1.87 (.10)	1.35 (.07)
2	.49 (.02)	2.18 (.10)	1.72 (.07)	14	.41 (.02)	2.88 (.11)	1.96 (.08)
3	.33 (.02)	2.62 (.09)	2.13 (.07)	15	.20 (.02)	4.11 (.11)	2.76 (.08)
4	.52 (.01)	2.22 (.07)	1.68 (.05)	16	.52 (.01)	2.43 (.06)	1.69 (.04)
5	.24 (.01)	3.05 (.07)	2.41 (.05)	17	.21 (.01)	3.86 (.07)	2.56 (.05)
6	.12 (.01)	3.62 (.06)	2.91 (.04)	18	.09 (.01)	4.81 (.07)	3.16 (.04)
7	.76 (.02)	1.47 (.10)	1.13 (.08)	19	.79 (.01)	1.28 (.08)	.96 (.06)
8	.76 (.02)	1.31 (.10)	1.02 (.07)	20	.74 (.02)	1.43 (.08)	1.03 (.06)
9	.69 (.02)	1.35 (.09)	1.08 (.07)	21	.48 (.02)	2.56 (.12)	1.77 (.08)
10	.70 (.01)	1.54 (.05)	1.18 (.04)	22	.71 (.01)	1.52 (.06)	1.12 (.04)
11	.57 (.01)	1.77 (.06)	1.41 (.05)	23	.47 (.01)	2.47 (.07)	1.72 (.05)
12	.41 (.01)	2.19 (.07)	1.79 (.05)	24	.22 (.01)	3.81 (.08)	2.62 (.05)
1	.66 (.02)	1.79 (.10)	1.79 (.10)	13	.67 (.02)	1.83 (.10)	1.83 (.10)
2	.50 (.02)	2.12 (.10)	2.12 (.10)	14	.43 (.02)	2.77 (.11)	2.77 (.11)
3	.34 (.02)	2.53 (.09)	2.53 (.09)	15	.22 (.02)	3.99 (.12)	3.99 (.12)
4	.52 (.01)	2.20 (.07)	2.20 (.07)	16	.52 (.01)	2.41 (.06)	2.41 (.06)
5	.24 (.01)	3.00 (.07)	3.00 (.07)	17	.21 (.01)	3.79 (.07)	3.79 (.07)
6	.13 (.01)	3.56 (.06)	3.56 (.06)	18	.09 (.01)	4.77 (.07)	4.77 (.07)
7	.76 (.02)	1.46 (.10)	1.46 (.10)	19	.80 (.02)	1.26 (.09)	1.26 (.09)
8	.76 (.02)	1.29 (.09)	1.29 (.09)	20	.75 (.02)	1.38 (.08)	1.38 (.08)
9	.70 (.02)	1.33 (.09)	1.33 (.09)	21	.49 (.02)	2.49 (.12)	2.49 (.12)
10	.71 (.01)	1.53 (.06)	1.53 (.06)	22	.72 (.01)	1.48 (.06)	1.48 (.06)
11	.58 (.01)	1.74 (.06)	1.74 (.06)	23	.49 (.02)	2.37 (.07)	2.37 (.07)
12	.42 (.01)	2.11 (.07)	2.11 (.07)	24	.23 (.01)	3.73 (.08)	3.73 (.08)