ARTICLE IN PRESS

Available online at www.sciencedirect.com

ScienceDirect

Theoretical Computer Science ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Theoretical Computer Science

www.elsevier.com/locate/tcs

# An analytical approach to the inference of summary data of additive type

Francesco M. Malvestuto*, Mauro Mezzini, Marina Moscarini

*"La Sapienza" University of Rome, Italy*

Received 16 September 2006; received in revised form 8 May 2007; accepted 12 July 2007

Communicated by O. Watanabe

**Abstract**

Summary data take the form of a triple that consists of a summary attribute, a category and a numeric value. The inference problem of summary data consists in deciding whether or not a summary data of interest is evaluable (i.e., can be computed) from a given set of summary data. We address the special case of the inference problem with homogeneous summary data (i.e., summary data with the same summary attribute), where the summary attribute is of additive nature. Owing to additivity, one can model the information content of the given summary data by a linear equation system whose variables are constrained to take their values from the domain of the summary attribute. We state two evaluability criteria, one for a real or integral summary attribute, and the other for a nonnegative-real or nonnegative-integral summary attribute. Using the two evaluability criteria, we show that our inference problem can be solved in strongly polynomial time for a real or integral or nonnegative real summary attribute, and is co*NP*-complete for a nonnegative-integral summary attribute. Moreover, we prove that, given a summary data of interest that is not evaluable, even in the (simplest) case that the summary attribute is of a real type, finding an evaluable summary data whose category is maximally contained in (or minimally contains) the category of the summary data of interest is an *NP*-hard problem.

*Keywords:* Category hierarchy; Summary database; Evaluability

## 1. Introduction

Aggregate data appear in many environments such as statistical databases [3,4,6,7,10,26,32–35,39,44,46], data warehouses [14,38,48] and multidimensional databases (or OLAP systems) [2,14,15,19,40], and several data models (e.g., cubes [2,11,47] and multidimensional tables [16,17,29]) have been proposed. Independently of the data model in use, a key problem for getting an efficient implementation of aggregate data [19] as well as for protecting sensitive information [32] is that of characterising the aggregate data that can be inferred from a given set of aggregate data [12, 13,18,34]. Two distinct approaches to the inference problem have been proposed in literature: one is based only on the

* Corresponding address: Universita "La Sapienza" di Roma, Dipartimento di Informatica, Facolta di Scienze Matematiche, Fisiche e Naturali, Via Salaria 113, III piano, 00198 Rome, Italy. Tel.: +39 0649918310; fax: +39 068541842.

*E-mail address:* malvestuto@di.uniroma1.it (F.M. Malvestuto).

0304-3975/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2007.07.004

use of the conceptual definitions of the input aggregate data and is known as the problem of "query rewriting" using materialised views [8,12,13,28,41,45], and the other takes into account their values too and is known as the problem of "query answering" using materialised views [9,21–23,25,29,32,39].

We shall consider only aggregate data (henceforth called *summary data*) that provide information about groups of individuals of statistical interest [26,40,46]. The assumption is that an arbitrarily chosen group of individuals may not have a statistical meaning, and that statistical information about a group of individuals should convey a "meaningful aspect" from a statistical point of view. To achieve this, a set of so-called "category attributes" is pre-defined, and summary data are obtained from raw data by computing aggregate functions (such as *sum*, *count*, *avg*, *min*, *max*) over groups of individuals selected using category attributes only. As pointed out by some authors [5–7,32,39], the choice of category attributes should be made accurately in order to avoid disclosing sensitive information.

In this paper, we focus our attention on summary data that are all obtained by applying the aggregate function *sum* to a numeric attribute of interest (the "aggregation attribute"). Moreover, we assume that summary data are obtained from a single relation of raw data, whose scheme contains the aggregation attribute. Although this restriction might seem drastic, these summary data are fundamental in many applications in data warehousing and OLAP [13]. For example, consider a relation of name Employee with schema {NAME, SSN, ADDRESS, GENDER, AGE, SALARY}. A possible set of category attributes is {GENDER, AGE-CLASS}, where AGE-CLASS has domain {Young, Middle-aged, Old} and is defined as a "re-coding" of the attribute AGE. By taking SALARY as aggregation attribute, an example of summary data might then be the sum of salaries of middle-aged employees. According to the terminology used in statistical databases, we call *sum*(SALARY) the *summary attribute* (elsewhere called "measure attribute" [11]) of the summary data, and the following relation on the category attributes GENDER and AGE-CLASS

| GENDER | AGE-CLASS |
|---|---|
| Male | Middle-aged |
| Female | Middle-aged |

the *category* of the summary data.

The inference problem we address is to decide whether or not a  summary data of interest can be evaluated (i.e., computed) from a given set of summary data (e.g., an arbitrary subset of a data cube) with the same summary attribute, say $A$, whose domain, say $D$, is either the set of reals ($R$) or the set of integers ($Z$) or the set of nonnegative reals ($R^+$) or the set of nonnegative integers ($Z^+$). After modelling the information content of the input summary data by a linear equation system whose variables are constrained to take their values from $D$, we state two evaluability criteria, one for $D = R$ or $D = Z$, and the other for $D = R^+$ or $D = Z^+$. Using the two evaluability criteria, we shall prove that our inference problem above can be solved in polynomial time if $D \neq Z^+$, and is co*NP*-complete if $D = Z^+$. Previous works on aggregate data only consider $D = R$ [5–7,34] and $D = R^+$ [32] in the general case, and $D = Z^+$ only in some very special very cases (e.g., 2-dimensional tables [21–23]).

The second problem we address concerns the case that the summary data of interest is not evaluable. As in the problem of "query containment" [8,18] for materialised views, we search for a nontrivial evaluable summary data whose category is maximally contained in the category of the summary data of interest. We shall prove that this problem is *NP*-hard.

The work is organised as follows. Section 2 contains the basic notions of categories, of summary statistics and of a summary database. Section 3 focuses on category hierarchies and their algebraic properties. In Section 4, we introduce the notion of the "evaluability" of a summary statistic from a summary database, and state two evaluability criteria, one for a summary attribute of general type $D = R$ or $D = Z$) and the other for a summary attribute of nonnegative type ($D = R^+$ or $D = Z^+$). In Section 5, we state an algebraic characterisation of evaluable summary statistics. The complexity of the inference problem is discussed in Section 6. Section 7 deals with the containment problem and contains the proof of its *NP*-hardness. Section 8 contains some closing notes and directions for future research.

## 2. Preliminaries

Let $R$ be a relation of raw data. A set $C$ of attributes (which need not belong to the scheme of $R$) is a *categorisation* of $R$ if there exists a mapping $\kappa$ from $R$ onto the domain of $C$, denoted by dom$(C)$; that is, for every tuple $t$ in $R$, there is exactly one element $c$ of dom$(C)$ such that $\kappa(t) = c$ and, for every $c \in$ dom$(C)$, there is at least one tuple $t$ in $R$ such that $\kappa(t) = c$. Typically, dom$(C)$ has a relatively small cardinality (with respect to the size of $R$) and, like the

Please cite this article in press as: F.M. Malvestuto, et al., An analytical approach to the inference of summary data of additive type, Theoretical Computer Science (2007), doi:10.1016/j.tcs.2007.07.004

value set of an enumerated type in programming languages, the elements of dom($C$) are known in advance. We call an element of dom($C$) a *cell*, and a set of cells a *category*; thus, every category is an ordinary relation with scheme $C$. For any category $K$, we denote the set $\{t \in R\colon \kappa(t) \in K\}$ by $R[K]$. Note that, owing to the above properties of the mapping $\kappa$, one has that $R[K] = \varnothing$ if and only if $K = \varnothing$.

A *summary attribute* of $R$ is a numeric attribute whose domain is taken to be the codomain of a function $f$ defined on the set of nonempty subsets of $R$. Typically, the function $f$ is defined by an aggregate function (such as *sum*, *count*, *avg*, *min*, *max*) and a numeric attribute (the aggregation attribute) in the scheme of $R$: e.g., *sum*(Salary) or *avg*(Salary). A summary attribute is *additive* if, for every two disjoint nonempty subsets $R'$ and $R''$ of $R$, one has $f(R' \cup R'') = f(R') + f(R'')$, where '+' stands for the ordinary addition of real numbers. In this paper, we only consider additive summary attributes that are defined by the aggregate function *sum* and whose domains are the set of reals ($R$), or the set of integers ($Z$), or the set of nonnegative reals ($R^+$), or the set of nonnegative integers ($Z^+$).

A *summary statistic* on $R$ is a couple $(A, K)$, where $A$ is a summary attribute and $K$ is a nonempty category. The quantity $a = f(R[K])$ is the *actual value* of the summary statistic, and the triple $(A, K, a)$ is a *summary data*. Note that a "data cube" with "measure attribute" $A$ [11] is the set of the summary data $(A, K, a)$, where $K$ is either dom($C$) or a singleton $\{c\}$, where $c$ is any cell, or the Cartesian product dom($C'$)× $\{c\}$, where $C'$ is any nonempty proper subset of $C$ and $c$ is an element of dom($C\backslash C'$).

Let $A$ be an additive summary attribute, and let $K_1$ = dom($C$), $K_2, \ldots, K_n$ ($n \geq 1$) be distinct nonempty categories. The set of summary data $\mathtt{S} = \{(A, K_1, a_1), \ldots, (A, K_n, a_n)\}$ will be referred to as a *summary database* with summary attribute $A$, and the attribute set $C$ as the *frame* of S. The inference problem consists in deciding, for a given nonempty category $T$, whether or not the actual value of the summary statistic $(A, T)$ can be computed from the contents of S. If this is the case, then (and only then) $(A, T)$ is said to be "evaluable" from S. The formal definition of evaluability will be given in Section 4. To this end, we need some preliminary notions about categories which will be stated in the next section.

**Example 1.** Let $R$ be a relation of raw data and let $C$ = {GENDER, AGE-CLASS} be a categorisation of $R$, where GENDER and AGE-CLASS have domains {Male, Female} and {Young, Middle-aged, Old}, respectively. Consider the three categories $K_1$, $K_2$ and $K_3$ where

$K_1$

| GENDER | AGE-CLASS |
|---|---|
| Male | Young |
| Male | Middle-aged |
| Male | Old |
| Female | Young |
| Female | Middle-aged |
| Female | Old |

$K_2$

| GENDER | AGE-CLASS |
|---|---|
| Male | Young |
| Male | Middle-aged |
| Female | Young |
| Female | Middle-aged |

$K_3$

| GENDER | AGE-CLASS |
|---|---|
| Male | Young |
| Male | Old |
| Female | Young |
| Female | Old |

Let $A$ be an additive summary attribute with domain $D$, let the actual values of the summary statistics $(A, K_1)$, $(A, K_2)$ and $(A, K_3)$ be 30, 20 and 10, respectively. We shall discuss the evaluability of summary statistics such as $(A,$

Please cite this article in press as: F.M. Malvestuto, et al., An analytical approach to the inference of summary data of additive type, Theoretical Computer Science (2007), doi:10.1016/j.tcs.2007.07.004

$T$), where $T$ is any nonempty subset of dom($C$) ($=K_1$), from the summary database $\mathtt{S} = \{(A, K_1, 30), (A, K_2, 20), (A, K_3, 10)\}$. ■

## 3. Category hierarchies

Let $C$ be a categorisation of a relation of raw data. A *category hierarchy* with *frame* $C$ is a set $\mathcal{F}$ of nonempty categories such that there is a member $U$ of $\mathcal{F}$ (called the *root* of $\mathcal{F}$) that contains every member of $\mathcal{F}$. A *$\sigma$-algebra containing* $\mathcal{F}$ is a set $\mathcal{A}$ of subsets of $U$ such that the following holds:

(i) if $K \in \mathcal{F}$, then $K \in \mathcal{A}$;
(ii) if $K, K' \in \mathcal{A}$, then $K \cup K' \in \mathcal{A}$;
(iii) if $K \in \mathcal{A}$, then $U \backslash K \in \mathcal{A}$.

Note that, since

$$K \cap K' = U \backslash ((U \backslash K) \cup (U \backslash K')) \quad \text{and} \quad K \backslash K' = K \cap (U \backslash K')$$

for every two subsets $K$ and $K'$ of $U$, one has that $\mathcal{A}$ is also closed under intersection and difference. An *atom* of $\mathcal{A}$ is a nonempty category of $\mathcal{A}$ such that none of its proper nonempty subsets is in $\mathcal{A}$. The set of atoms of $\mathcal{A}$ will be referred to as the *basis* of $\mathcal{A}$. Note that the basis of $\mathcal{A}$ is a partition of $U$. By the closure of $\mathcal{A}$ under intersection and difference, for every nonempty member $K$ of $\mathcal{A}$, an atom of $\mathcal{A}$ has a nonempty intersection with $K$ if and only if it is contained in $K$, so that $K$ is a union of one or more atoms of $\mathcal{A}$.

The *$\sigma$-algebra generated by* $\mathcal{F}$, denoted by $\mathcal{F}^*$, is the smallest $\sigma$-algebra containing $\mathcal{F}$, that is, $\mathcal{F}^*$ is the intersection of the $\sigma$-algebras containing $\mathcal{F}$. Let $F = \{K_1, \ldots, K_n\}$. Then, every atom of $\mathcal{F}^*$ is of the form $\cap_{i=1,\ldots,n} H_i$, where $H_i$ is either $K_i$ or $U \backslash K_i$, and the basis of $\mathcal{F}^*$ is the coarsest of the partitions $\mathcal{X}$ of the root $U$ of $\mathcal{F}$ such that each member of $\mathcal{F}$ can be expressed as a union of one or more classes of $\mathcal{X}$.

**Lemma 1** ([27]). *Given a category hierarchy $\mathcal{F}$ with root $U$, the atoms of $\mathcal{F}^*$ can be obtained by grouping together the cells in $U$ that are contained in exactly the same members of $\mathcal{F}$.*

By Lemma 1, there are not two atoms of $\mathcal{F}^*$ that are contained in the same members of $\mathcal{F}$. Let $\mathcal{X} = \{X_1, \ldots, X_m\}$ be the basis of $\mathcal{F}^*$. (Note that $m \leq 2^{n-1}$.) Let $K$ be a member of $\mathcal{F}^*$. We know that $K$ is a union of zero or more atoms of $\mathcal{F}^*$, say

$$K = \bigcup_{j \in J} X_j$$

for some (possibly empty) subset $J$ of $\{1, \ldots, m\}$. Henceforth, the set $J$ will be referred to as the *support* of $K$ *in* $\mathcal{F}^*$. Let $J_i$ be the support of $K_i$ ($1 \leq i \leq n$) in $\mathcal{F}^*$; that is,

$$K_i = \bigcup_{j \in J_i} X_j \quad (1 \leq i \leq n).$$

Since each member $K_i$ of $\mathcal{F}$ is a nonempty category, no $J_i$ is empty. It is convenient to summarise the above set-theoretic relationships among members of $\mathcal{F}$ and members of $\mathcal{X}$ by a binary matrix $\mathbf{B} = (b_{ij})$ of size $n \times m$, where $b_{ij} = 1$ if and only if $j \in J_i$ (or, equivalently, $X_j \subseteq K_i$). In what follows, $\mathbf{B}$ will be referred to as the *incidence matrix* of $\mathcal{F}^*$.

**Remark 1.** The incidence matrix of the $\sigma$-algebra generated by a category hierarchy contains an all-one row which corresponds to the root of the category hierarchy, and contains no all-zero rows. Moreover, by Lemma 1, its columns are all distinct.

**Example 2.** Consider the category hierarchy $\mathcal{F} = \{K_1, K_2, K_3\}$ of Example 1. Recall that $K_1 = \text{dom}(C)$. By Lemma 1, the basis of $\mathcal{F}^*$ is $\mathcal{X} = \{X_1, X_2, X_3\}$ where

$X_1$

| GENDER | AGE-CLASS |
|---|---|
| Male | Young |
| Female | Young |

$X_2$

| GENDER | AGE-CLASS |
|---|---|
| Male | Middle-aged |
| Female | Middle-aged |

$X_3$

| GENDER | AGE-CLASS |
|---|---|
| Male | Old |
| Female | Old |

Accordingly, the supports of $K_1$, $K_2$ and $K_3$ in $\mathcal{F}^*$ are respectively:

$$J_1 = \{1, 2, 3\} \qquad J_2 = \{1, 2\} \qquad J_3 = \{1, 3\},$$

and the incidence matrix of $\mathcal{F}^*$ is as follows:

$$\mathbf{B} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}. \quad ■$$

Let $\mathcal{F}$ be a category hierarchy with root $U$, and let $T$ be a nonempty subset of $U$. We denote the category hierarchy $\mathcal{F}\cup \{T\}$ by $\mathcal{F} + T$. Consider the $\sigma$-algebra $(\mathcal{F} + T)^*$ generated by $\mathcal{F} + T$, and let $\mathcal{Y}$ be the basis of $(\mathcal{F} + T)^*$. Of course, $(\mathcal{F} + T)^*$ is a $\sigma$-algebra containing $\mathcal{F}$, so that $\mathcal{Y}$ is a partition of $U$ that is finer than the basis $\mathcal{X}$ of $\mathcal{F}^*$; moreover, $\mathcal{Y} = \mathcal{X}$ (i.e., $(\mathcal{F} + T)* = \mathcal{F}^*$) if and only if $T$ is a member of $\mathcal{F}^*$. We now show how to construct $\mathcal{Y}$ from $\mathcal{X}$, which will lead to a useful characterisation of the support of $T$ in $(\mathcal{F} + T)^*$. By Lemma 1, there are no two atoms of $(\mathcal{F} + T)^*$ that are contained in exactly the same members of $\mathcal{F} + T$. However, there may exist two atoms $Y$ and $Y'$ of $(\mathcal{F} + T)^*$ that are contained in exactly the same members of $\mathcal{F}$; if this is the case, then we say that $Y$ and $Y'$are *mates*. By Remark 1, one has that, if two atoms $Y$ and $Y'$ of $(\mathcal{F} + T)^*$ are mates, then either $Y$ or $Y'$ is contained in $T$, and that every atom of $(\mathcal{F} + T)^*$ has at most one mate.

**Example 2** (*Continued*). Consider the following category:

$T$

| GENDER | AGE-CLASS |
|---|---|
| Male | Young |
| Male | Middle-aged |
| Female | Middle-aged |

The basis of $(\mathcal{F} + T)^*$ is $\mathcal{Y} = \{Y_1, Y_2, Y_3, Y_4\}$ where

$Y_1$

| GENDER | AGE-CLASS |
|---|---|
| Male | Young |

$Y_2$

| GENDER | AGE-CLASS |
|---|---|
| Male | Middle-aged |
| Female | Middle-aged |

$Y_3$

| GENDER | AGE-CLASS |
|---|---|
| Male | Old |
| Female | Old |

$Y_4$

| GENDER | AGE-CLASS |
|---|---|
| Female | Young |

Please cite this article in press as: F.M. Malvestuto, et al., An analytical approach to the inference of summary data of additive type, Theoretical Computer Science (2007), doi:10.1016/j.tcs.2007.07.004

There is only one pair of mates: $Y_1$ and $Y_4$ (both are contained in each of the three categories in $F$). ■

Given the basis $\mathcal{X} = \{X_1, \ldots, X_m\}$ of $\mathcal{F}^*$, let

$$H = \{j : X_j \cap T \neq \varnothing \ \& \ X_j \backslash T \neq \varnothing\} \qquad J = \{j : X_j \subseteq T\}. \tag{1}$$

From the foregoing, it follows that the basis $\mathcal{Y}$ of $(\mathcal{F} + T)^*$ can be obtained from $\mathcal{X}$ by splitting each $X_j$, $j \in H$, into the pair of mates $X_j \cap T$ and $X_j \backslash T$; explicitly one has

$$\mathcal{Y} = \{X_j \cap T : j \in H\} \cup \{X_j : j \notin H\} \cup \{X_j \backslash T : j \in H\}. \tag{2}$$

Moreover, since

$$T = \left( \bigcup_{j \in H} X_j \cap T \right) \cup \left( \bigcup_{j \in J} X_j \right),$$

the support of $T$ in $(\mathcal{F} + T)^*$ is given by $H \cup J$.

**Example 2** (*Continued*). One has $X_1 \cap T \neq \varnothing$ and $X_1 \backslash T \neq \varnothing$, $X_2 \subseteq T$ and $X_3 \cap T = \varnothing$. Therefore, $H = \{1\}$ and $J = \{2\}$ so that the basis of $(\mathcal{F} + T)^*$ is given by

$$\mathcal{Y} = \{Y_1 = X_1 \cap T,\ Y_2 = X_2,\ Y_3 = X_3,\ Y_4 = X_1 \backslash T\}.$$

Finally, the support of $T$ in $(\mathcal{F} + T)^*$ is $H \cup J = \{1, 2\}$. ■

We now prove a technical lemma which will be used later. Let $V$ be a (possibly empty) member of $\mathcal{F}^*$, $V \neq U$, and let $V = \cup_{j \in N} X_j$ for some (possibly empty) proper subset $N$ of $\{1, \ldots, m\}$. We denote the category hierarchy $\{K \backslash V : K \in F\}$ by $\mathcal{F} - V$. Note that the basis of $(\mathcal{F} - V)^*$ is the partition $\mathcal{Z} = \mathcal{X} \backslash \{X_j : j \in N\}$ of the root $U \backslash V$ of $\mathcal{F} - V$, and the incidence matrix of $(\mathcal{F} - V)^*$ can be obtained from the incidence matrix of $\mathcal{F}^*$ by deleting the columns indexed by $N$ and the all-zero rows (if any).

**Lemma 2.** *The category $T \backslash V$ belongs to $(\mathcal{F} - V)^*$ if and only if $H \subseteq N$.*

**Proof.** Since $T = (\cup_{j \in H} X_j \cap T) \cup (\cup_{j \in J} X_j)$, one has

$$T \backslash V = \left( \bigcup_{j \in H \backslash N} X_j \cap T \right) \cup \left( \bigcup_{j \in J \backslash N} X_j \right)$$

and, hence, $T \backslash V$ belongs to $(\mathcal{F} - V)^*$ if and only if $\cup_{j \in H \backslash N} X_j \cap T = \varnothing$. Since $X_j \cap T \neq \varnothing$ for all $j \in H$, $T \backslash V$ belongs to $(\mathcal{F} - V)^*$ if and only if $H \subseteq N$. □

By taking $V = \varnothing$ in Lemma 2, we obtain the following.

**Corollary 1.** *$T$ is a member of $\mathcal{F}^*$ if and only if $H = \varnothing$ (i.e., if and only if, for all $j$, $X_j \cap T \neq \varnothing$ implies $X_j \subseteq T$).*

## 4. Evaluability

In this section, we give the formal definition of evaluability of a summary statistic from a summary database $\mathrm{S} = \{(A, K_1, a_1), \ldots, (A, K_n, a_n)\}$. Let $\mathcal{F} = \{K_1, \ldots, K_n\}$, $C$ the frame of $\mathcal{F}$ and $T$ a nonempty subset of dom$(C)$. Let $\mathcal{Y} = \{Y_1, \ldots, Y_p\}$ be the basis of $(\mathcal{F} + T)^*$. Of course, the support of $K_i$ $(1 \leq i \leq n)$ in $(\mathcal{F} + T)^*$ need not be the same as the support of $K_i$ in $\mathcal{F}^*$ (which was denoted by $J_i$ in the previous section). Henceforth, we denote the support of $K_i$ in $(\mathcal{F} + T)^*$ by $L_i$. For each $l$ $(1 \leq l \leq p)$, let us introduce a variable $y_l$ which stands for a feasible value of the summary statistic $(A, Y_l)$. By the additivity of the summary attribute $A$, the relationships between the summary data in $\mathrm{S}$ and the variables $y_l$ can be described by the following constraint system:

$$\begin{cases} \sum\limits_{l \in L_i} y_l = a_i & (1 \leq i \leq n) \\ y_l \in D & (1 \leq l \leq p) \end{cases} \tag{3}$$

where $D$ denotes the domain of the summary attribute $A$. Note that the coefficient matrix of the equation system in system (3) is the incidence matrix of $(\mathcal{F} + T)^*$. Moreover, system (3) is consistent since it has the "true" solution $(s_1, \ldots, s_p)$ with $s_l = f(R[Y_l])$ $(1 \le l \le p)$, where $R$ is the relation of raw data.

Given a nonempty subset $L$ of $\{1, \ldots, p\}$, we say that the sum-expression

$$\sum_{l \in L} y_l$$

is a *sum-invariant* [21–23] (an *invariant*, for short) of system (3) if, for every two solutions $(s_1, \ldots, s_p)$ and $(s'_1, \ldots, s'_p)$ of the system (3), one has $\sum_{l \in L} s_l = \sum_{l \in L} s'_l$. If this is the case, then by the *value* of the invariant $\sum_{l \in L} y_l$, we mean the sum $\sum_{l \in L} s_l$ where $(s_1, \ldots, s_p)$ is any solution of system (3). The sum-expression $\sum_{l \in L} y_l$ is called a *zero-invariant* of system (3) if it is an invariant with value zero.

Let $L$ be the support of $T$ in $(\mathcal{F} + T)^*$. We say that the summary statistic $(A, T)$ is *evaluable* from the summary database $\mathtt{S}$ if the sum-expression $\sum_{l \in L} y_l$ is an invariant of system (3).

**Remark 2.** If the summary statistic $(A, T)$ is evaluable from $\mathtt{S}$, then the value of the invariant $\sum_{l \in L} y_l$ is $\sum_{l \in L} s_l$ where $(s_1, \ldots, s_p)$ is the true solution of system (3) and, hence, coincides with the actual value of $(A, T)$.

**Example 1** (*Continued*). Let $T$ be the category of Example 2. The supports of $K_1$, $K_2$ in $K_3$ in $(\mathcal{F} + T)^*$ are $L_1 = \{1, 2, 3, 4\}$, $L_2 = \{1, 2, 4\}$ and $L_3 = \{1, 3, 4\}$, respectively. System (3) reads

$$\begin{cases} y_1 + y_2 + y_3 + y_4 = 30 \\ y_1 + y_2 + y_4 = 20 \\ y_1 + y_3 + y_4 = 10 \\ y_1, y_2, y_3, y_4 \in D. \end{cases}$$

For $D = R$ (or $Z$), the general solution is of the form

$$(y_1 = \lambda, y_2 = 20, y_3 = 10, y_4 = -\lambda)$$

where $\lambda$ is any real (integer, respectively).

For $D = R^+$ (or $Z^+$), there is exactly one solution:

$$(y_1 = 0, y_2 = 20, y_3 = 10, y_4 = 0).$$

Recall that the support of $T$ in $(\mathcal{F} + T)^*$ is $\{1, 2\}$ (see Example 2). By definition, the summary statistic $(A, T)$ is evaluable from the summary database $\mathtt{S}$ if (and only if) the sum-expression $y_1 + y_2$ is an invariant of the constraint system above. Therefore, $(A, T)$ is not evaluable from $\mathtt{S}$ if $D$ is $R$ or $Z$, and is evaluable from $\mathtt{S}$ if $D$ is $R^+$ or $Z^+$. ■

We shall state two criteria for the evaluability of the summary statistic $(A, T)$ from the summary database $\mathtt{S}$, one if $D = R$ or $D = Z$, and the other if $D = R^+$ or $D = Z^+$. Both evaluability criteria make use of a constraint system which, unlike system (3), is independent of the category $T$ and hence, can be constructed once and for all when the summary database is created. In what follows, the category $T$ will be referred to as the *target category*.

Let $\mathcal{X} = \{X_1, \ldots, X_m\}$ be the basis of $\mathcal{F}^*$, and let $J_i$ be the support of $K_i$ $(1 \le i \le n)$ in $\mathcal{F}^*$. For each $j$ $(1 \le j \le m)$, let us introduce a variable $x_j$ which stands for a feasible value of the summary statistic $(A, X_j)$. Again, by the additivity of the summary attribute $A$, the relationships between the summary data in S and the variables $x_j$ can be described by the following constraint system:

$$\begin{cases} \sum_{j \in J_i} x_j = a_i & (1 \le i \le n) \\ x_j \in D & (1 \le j \le m) \end{cases} \tag{4}$$

where, as usual, $D$ denotes the domain of the summary attribute $A$. Note that the coefficient matrix of the equation system in system (4) is the incidence matrix of $\mathcal{F}^*$.

Given the target category $T$ and the basis $\mathcal{X}$ of $\mathcal{F}^*$, let $H$ and $J$ be the two subsets of $\{1, \ldots, m\}$ defined by (1). Recall that the support of $T$ in $(\mathcal{F} + T)^*$ equals $H \cup J$ so that the sum-expression associated with $T$ is $\sum_{l \in H \cup J} s_l$. Then, by Corollary 1, $T$ is a member of $\mathcal{F}^*$ if and only if $H = \varnothing$. If this is the case, then $\mathcal{Y} = \mathcal{X}$ and systems (3) and

(4) are the same (up to a renaming of variables) so that the summary statistic $(A, T)$ is evaluable from the summary database S if and only if the sum-expression $\sum_{j \in J} x_j$ is an invariant of system (4). Suppose that $T$ is not a member of $\mathcal{F}^*$. By Corollary 1, $H \neq \emptyset$ and, by (2), $|H| = |\mathcal{Y}| - |\mathcal{X}| = p - m$. Without loss of generality, we can assume that $H = \{1, \ldots, p - m\}$. Then, by (2), the categories in $\mathcal{Y}$ can be expressed in function of the categories in $\mathcal{X}$ as follows:

$$\begin{array}{llll}
Y_1 = X_1 \cap T & Y_2 = X_2 \cap T & \ldots & Y_{p-m} = X_{p-m} \cap T \\
Y_{p-m+1} = X_{p-m+1} & Y_{p-m+2} = X_{p-m+2} & \ldots & Y_m = X_m \\
Y_{m+1} = X_1 \backslash T & Y_{m+2} = X_2 \backslash T & \ldots & Y_p = X_{p-m} \backslash T.
\end{array}$$

Note that there are exactly $p - m$ pairs of mates in $\mathcal{Y}$: $Y_1$ and $Y_{m+1}$, $Y_2$ and $Y_{m+2}, \ldots, Y_{p-m}$ and $Y_p$. Moreover, for each pair of mates $Y_h$ and $Y_{m+h}$, one has that $Y_h \cup Y_{m+h} = X_h$. Finally, since $Y_h$ and $Y_{m+h}$ are contained in the same members of $\mathcal{F}$, the variables $y_h$ and $y_{m+h}$ occur in exactly the same equations of system (3). Therefore, system (3) can be obtained from system (4) by replacing each variable $x_j$

- by the sum-expression $y_j + y_{m+j}$ if $1 \leq j \leq p - m$, and
- by $y_j$ if $p - m + 1 \leq j \leq m$.

From the foregoing, the following holds.

**Lemma 3.** *A sum-expression is an invariant of system* (4) *if the corresponding sum-expression obtained with the substitution rule above is an invariant of system* (3) *; moreover, if this is the case, then the values of the two sum-expressions are the same.*

**Lemma 4.** *If* $(A, T)$ *is evaluable from* S *then either* $H = \emptyset$ *or, for each* $h$ *(*$1 \leq h \leq p - m$*), the variable* $x_h$ *is a zero-invariant of system* (4)*.*

**Proof.** Suppose by contradiction that, for some $h \leq p - m$, the variable $x_h$ is not a zero-invariant of system (4). By Lemma 3, the sum-expression $y_h + y_{m+h}$ is not a zero-invariant of system (3) so that there exists a solution $(s_1, \ldots, s_p)$ of system (3) such that $s_h + s_{m+h} = v$ where $v \neq 0$. For each $l$ $(1 \leq l \leq p)$, let

$$s'_l = \begin{cases} s_l & l \neq h,\ m+h \\ 0 & l = h \\ v & l = m+h \end{cases} \qquad s''_l = \begin{cases} s_l & l \neq h,\ m+h \\ v & l = h \\ 0 & l = m+h \end{cases}$$

Of course, $(s'_1, \ldots, s'_p)$ and $(s''_1, \ldots, s''_p)$ are both solutions of system (3) and, since $\sum_{l \in H \cup J} s'_l \neq \sum_{l \in H \cup J} s''_l$, the sum-expression $\sum_{l \in H \cup J} y_l$ is not an invariant of system (3) and, hence, the summary statistic $(A, T)$ is not evaluable from S (contradiction). □

In order to get a necessary and sufficient condition for evaluability, we distinguish the following two cases: (i) $D$ is $R$ or $Z$, and (ii) $D$ is $R^+$ or $Z^+$.

Case (i): $D$ is $R$ or $Z$.

**Lemma 5.** *For* $D = R$ *or* $D = Z$*, if* $(A, T)$ *is evaluable from* S*, then* $H = \emptyset$*.*

**Proof.** Suppose by contradiction that $H \neq \emptyset$. Let $h \in H$; that is, $1 \leq h \leq p - m$. By Lemma 4, the variable $x_h$ is a zero-invariant of system (4), and then by Lemma 3, the sum-expression $y_h + y_{m+h}$ is a zero-invariant of system (3). Let $(s_1, \ldots, s_p)$ be a solution of system (3). Then, $s_h + s_{m+h} = 0$. Arbitrarily choose an integer $v > 0$, for each $l$ (1 $\leq l \leq p$), let

$$s'_l = \begin{cases} s_l & l \neq h,\ m+h \\ -v & l = h \\ v & l = m+h \end{cases} \qquad s''_l = \begin{cases} s_l & l \neq h,\ m+h \\ v & l = h \\ -v & l = m+h. \end{cases}$$

Of course, $(s'_1, \ldots, s'_p)$ and $(s''_1, \ldots, s''_p)$ are both solutions of system (3) and, since $\sum_{l \in H \cup J} s'_l < \sum_{l \in H \cup J} s''_l$, the sum-expression $\sum_{l \in H \cup J} y_l$ is not an invariant of system (3) and, hence, the summary statistic $(A, T)$ is not evaluable from S (contradiction). □

Finally, we are in a position to state our first evaluability criterion.

**Theorem 1.** *For $D = R$ or $D = Z$, a summary statistic $(A, T)$ is evaluable from $\mathcal{S}$ if and only if*

(a) *$H = \varnothing$, and*
(b) *the sum-expression $\sum_{j \in J} x_j$ is an invariant of system* (4) .

*Moreover, if the summary statistic $(A, T)$ is evaluable from* S*, then its actual value is given by the value of the invariant $\sum_{j \in J} x_j$.*

**Proof.** (*if*) By Corollary 1, from (a) it follows that $T$ is a member of $\mathcal{F}^*$ so that system (3) is the same as system (4). Moreover, $J$is the support of $T$ in $\mathcal{F}^*$. The statement then follows from the definition of evaluability and Remark 2. (*only if*) By Lemma 5, $H = \varnothing$ and, again by Corollary 1, $T$ is a member of $\mathcal{F}^*$ so that system (3) is the same as system (4) and the statement follows from the definition of evaluability and Remark 2. □

Case (ii): $D$ is $R^+$ or $Z^+$. By Lemma 4, we have soon our second evaluability criterion.

**Theorem 2.** *For $D = R^+$ or $D = Z^+$, a summary statistic $(A, T)$ is evaluable from $\mathcal{S}$ if and only if the following two conditions are both satisfied:*

(a) *if $H \neq \varnothing$ then, for each h ($1 \le h \le p - m$), the variable $x_h$ is a zero-invariant of system* (4)*;*
(b) *if $J \neq \varnothing$, then the sum-expression $\sum_{j \in J} x_j$ is an invariant of system* (4).

*Moreover, if the summary statistic $(A, T)$ is evaluable from* S*, then its actual value is zero if $J = \varnothing$ and is given by the value of the invariant $\sum_{j \in J} x_j$ otherwise.*

**Example 1** (*Continued*). System (4) reads

$$\begin{cases} x_1 + x_2 + x_3 = 30 \\ x_1 + x_2 = 20 \\ x_1 + x_3 = 10 \\ x_1, x_2, x_3 \in D \end{cases}$$

and has exactly one solution:

$$(x_1 = 0, x_2 = 20, x_3 = 10)$$

so that each variable is an invariant. Since $H \neq \varnothing$, by Theorem 1 the summary statistic $(A, T)$ is not evaluable from the summary database S if $D = R$ or $D = Z$; moreover, since $H = \{1\}$ and the variable $x_1$ is a zero-invariant, by Theorem 2, the summary statistic $(A, T)$ is evaluable from S if $D = R^+$ or $D = Z^+$. Recall that we have already arrived at the same conclusions using the definition of evaluability. ■

## 5. Algebraic categories

In this section, we state a characterisation of the categories that satisfy the evaluability criterion for $D = R$ or $D = Z$ (see Theorem 1 above). We call them "algebraic categories" and also prove that they are tightly related to the categories that satisfy the evaluability criterion for $D = R^+$ (see Theorem 2 above). To this end, we need some standard definitions of linear algebra [43], which are now recalled.

Let $\mathbf{A}$ be a real matrix of size $n \times m$. The *row space* of $\mathbf{A}$ is the linear space (over $R$) spanned by the rows of $\mathbf{A}$, that is, the set of the $m$-dimensional vectors that can be expressed as linear combinations of rows of $\mathbf{A}$ with real coefficients. A *vector base* of the row space of $\mathbf{A}$ is a maximal linearly independent set of its vectors. The dimension of the row space of $\mathbf{A}$ (that is, the cardinality of every vector base) is equal to the rank of $\mathbf{A}$. The *kernel* (or "null space") of $\mathbf{A}$ is the solution set of the homogenous system $\mathbf{A}\ \mathbf{v} = \mathbf{0}$, and turns out to be the orthogonal complement of the row space of $\mathbf{A}$, that is,

- for every vector $\mathbf{u}$ of the row space of $\mathbf{A}$ and for every vector $\mathbf{v}$ of the kernel of $\mathbf{A}$, the inner product $\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{j=1,\ldots,m} u_j v_j$ is zero;
- if $\mathbf{A}$ has rank $r$, then the dimension of the kernel of $\mathbf{A}$ is $m - r$ (that is, every vector base of the kernel of $\mathbf{A}$ has cardinality $m - r$).

Please cite this article in press as: F.M. Malvestuto, et al., An analytical approach to the inference of summary data of additive type, Theoretical Computer Science (2007), doi:10.1016/j.tcs.2007.07.004

It follows that, if $\{\mathbf{v}_1, \ldots, \mathbf{v}_{m-r}\}$ is a vector base of the kernel of $\mathbf{A}$, then a real $m$-dimensional vector $\mathbf{u}$ belongs to the row space of $\mathbf{A}$ if and only if

$$\langle \mathbf{u}, \mathbf{v}_1 \rangle = \cdots = \langle \mathbf{u}, \mathbf{v}_{m-r} \rangle = 0.$$

Finally, a vector base of the kernel of $\mathbf{A}$ can be found using the Gaussian elimination method and, if $\mathbf{A}$ is a rational matrix, then such a base vector of the kernel of $\mathbf{A}$ is rational (e.g., see [43]). Therefore, the following holds.

**Lemma 6.** *Let* $\mathbf{A}$ *be a rational matrix of size $n \times m$. An $m$-dimensional vector* $\mathbf{u}$ *belongs to the row space of* $\mathbf{A}$ *if and only if the inner product of* $\mathbf{u}$ *with every rational vector of the null space of* $\mathbf{A}$ *is zero.*

Let $\mathcal{F} = \{K_1, \ldots, K_n\}$ be a category hierarchy, $\mathcal{X} = \{X_1, \ldots, X_m\}$ be the basis of $\mathcal{F}^*$, and $\mathbf{B}$ be the incidence matrix of $\mathcal{F}^*$. Thus, $\mathbf{B}$ has $n$ rows and $m$ columns. Let $T$ be a category in $\mathcal{F}^*$ with support $J$; the *support vector* of $T$ is the charateristic vector of $J$, that is, the $m$-dimensional binary vector $\mathbf{u} = (u_1, \ldots, u_m)$ with

$$u_j = \begin{cases} 1 & \text{if } j \in J \\ 0 & \text{else} \end{cases} \qquad (1 \le j \le m).$$

We say that a category $T$ in $\mathcal{F}^*$ is an *algebraic category* (of $\mathcal{F}^*$) if the support vector $\mathbf{u}$ of $T$ belongs to the row space of $\mathbf{B}$, that is, if there exists a real-valued solution of the following linear system with unknowns $z_1, \ldots, z_n$:

$$\sum_{i=1,\ldots,n} z_i \mathbf{b}_i = \mathbf{u} \tag{5}$$

where $\mathbf{b}_i$ is the $i$-th row of $\mathbf{B}$.

We denote the set of algebraic categories of $\mathcal{F}^*$ by $\mathcal{F}^a$. Of course, $\mathcal{F} \subseteq \mathcal{F}^a \subseteq \mathcal{F}^*$. Moreover, it is easy to see that $\mathcal{F}^a$ is closed under disjoint union and proper difference [3,4,36]; that is:

if $K, K' \in \mathcal{F}^a$ and $K \cap K' = \varnothing$, then $K \cup K' \in \mathcal{F}^a$;

if $K, K' \in \mathcal{F}^a$ and $K' \subseteq K$, then $K \backslash K' \in \mathcal{F}^a$.

Let $U$ be the root of $\mathcal{F}$. Since $U \in \mathcal{F} \subseteq \mathcal{F}^a$, one has that $U \backslash K \in \mathcal{F}^a$ for every $K \in \mathcal{F}^a$; that is, $\mathcal{F}^a$ is closed under complementation too.

Let $\mathtt{S} = \{(A, K_1, a_1), \ldots, (A, K_n, a_n)\}$ be a summary database with frame $C$, $\mathcal{F} = \{K_1, \ldots, K_n\}$, where $K_1 = \text{dom}(C)$, and $\mathbf{B}$ the incidence matrix of $\mathcal{F}^*$. We call the vector $\mathbf{a} = (a_1, \ldots, a_n)$ the *summary vector* of S. We now apply the evaluability criteria stated in Theorems 1 and 2 to derive two evaluability tests, one for $D = R$ or $D = Z$, and the other for $D = R^+$.

($D = R$ or $D = Z$) The following lemma is a re-statemement of a well-known result of linear algebra (e.g., see Corollary 3.1d in [43]).

**Lemma 7.** *Let $J$ be a nonempty subset of $\{1, \ldots, m\}$ and let* $\mathbf{u}$ *be the characteristic vector of $J$. For $D = R$, the sum-expression $\sum_{j \in J} x_j$ is an invariant of system* (4) *if and only if* $\mathbf{u}$ *belongs to the row space of the matrix* $\mathbf{B}$. *Moreover, if the sum-expression $\sum_{j \in J} x_j$ is an invariant of system* (4)*, then its value is given by the inner product of the summary vector of* $\mathtt{S}$ *with any solution of system* (5).

The following theorem states that the algebraic categories of $\mathcal{F}^*$ are all and the only categories that pass the evaluability test for $D = R$ or $D = Z$.

**Theorem 3.** *For $D = R$ or $D = Z$, a summary statistic $(A, T)$ is evaluable from* $\mathtt{S}$ *if and only if $T$ is an algebraic category of $F^*$. Moreover, if the summary statistic $(A, T)$ is evaluable from* $\mathtt{S}$*, then its actual value is given by the inner product of the summary vector of* $\mathtt{S}$ *with any solution of system* (5)*, where* $\mathbf{u}$ *is the support vector of $T$.*

**Proof.** For $D = R$, the statement follows from Theorem 1 and Lemma 7. Consider now the case $D = Z$. The "if" part follows from the fact that $Z \subseteq R$. We now prove the "only-if" part. By condition (a) of Theorem 1, $T$ is a member of $\mathcal{F}^*$. Suppose, by contradiction, that $T$ is not algebraic. Then, the support vector $\mathbf{u}$ of $T$ does not belong to the row space of $\mathbf{B}$. By Lemma 6, there exists a rational vector $\mathbf{v}$ of the kernel of $\mathbf{B}$ such that $\langle \mathbf{u}, \mathbf{v} \rangle \neq 0$. Let $v_j = \frac{q_j}{r_j}$ with $r_j > 0$ $(1 \le j \le m)$, and let $r$ be the least common multiple of $r_1, \ldots, r_m$. Let $\mathbf{s}$ be an integral solution of system (4). Then $\mathbf{s}' = \mathbf{s} + r\mathbf{v}$ is another integral solution of system (4). Finally, since neither $r$ nor $\langle \mathbf{u}, \mathbf{v} \rangle$ is zero, one has

$$\sum_{j \in J} s'_j = \langle \mathbf{u}, \mathbf{s}' \rangle = \langle \mathbf{u}, \mathbf{s} \rangle + r \langle \mathbf{u}, \mathbf{v} \rangle \neq \langle \mathbf{u}, \mathbf{s} \rangle = \sum_{j \in J} s_j$$

so that the sum-expression $\sum_{j \in J} x_j$ is not an invariant (contradiction). Therefore, $T$ must be algebraic.
Finally, if the sum-expression $\sum_{j \in J} x_j$ is an invariant of system (4) then, since $Z \subseteq R$, its value over $Z$ is the same as over $R$. □

**Example 1** (*Continued*). Consider again the summary database with summary vector $\mathbf{a} = (30, 20, 10)$ and category hierarchy $\mathsf{F} = \{K_1, K_2, K_3\}$. Recall that $\mathsf{F}^*$ has incidence matrix

$$\mathbf{B} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

and the equation system $\mathbf{B}\,\mathbf{x} = \mathbf{a}$ has exactly one solution:

$$\mathbf{x} = (x_1 = 0, x_2 = 20, x_3 = 10)$$

so that the actual values of the summary statistics $(A, X_1)$, $(A, X_2)$ and $(A, X_3)$ are 0, 20 and 10, respectively. The same values can be obtained using Theorem 3 as follows. Let $\mathbf{u}_j$ be the support vector of $X_j$ in $\mathsf{F}^*$ ($j$= 1, 2, 3). Each $X_j$ is an algebraic category, since

$$\mathbf{u}_1 = -\mathbf{b}_1 + \mathbf{b}_2 + \mathbf{b}_3 \qquad \mathbf{u}_2 = \mathbf{b}_1 - \mathbf{b}_3 \qquad \mathbf{u}_3 = \mathbf{b}_1 - \mathbf{b}_2$$

so that $(-1\ 1\ 1)$, $(1\ 0 - 1)$ and $(1 - 1\ 0)$ are solutions of system (5) when $\mathbf{u} = \mathbf{u}_1$, $\mathbf{u} = \mathbf{u}_2$ and $\mathbf{u} = \mathbf{u}_3$ respectively. By Theorem 3, the actual values of the summary statistics $(A, X_1)$, $(A, X_2)$ and $(A, X_3)$ are the inner products

$$-30 + 20 + 10 = 0 \qquad 30 - 10 = 20 \qquad 30 - 20 = 10$$

which are the same as above. ■

($D = R^+$) Consider the variables $x_j$ of system (4) that are zero-invariants. We denote their set by $\{x_j \colon j \in N\}$ for some (possibly empty) subset $N$ of $\{1, \ldots, m\}$. Let $\mathbf{B}'$ be the matrix obtained from $\mathbf{B}$ by deleting the columns indexed by $N$.

**Lemma 8** ([32]). *Let $J$ be a nonempty subset of $\{1, \ldots, m\}$, and let $\mathbf{u}$ be the characteristic vector of $J - N$. For $D = R^+$, the sum-expression $\sum_{j \in J} x_j$ is an invariant of system (4) if and only if $\mathbf{u}$ belongs to the row space of $\mathbf{B}'$. Moreover, if $\sum_{j \in J} x_j$ is an invariant of system (4), then its value is given by the inner product of the summary vector of S with any solution of the following equation system*

$$\sum_{i=1,\ldots,n} z_i \mathbf{b}'_i = \mathbf{u} \tag{6}$$

*where $\mathbf{b}'_i$ is the i-th row of $\mathbf{B}'$.*

**Theorem 4.** *Let $V = \cup_{j \in N} X_j$. For $D = R^+$, a summary statistic $(A, T)$ is evaluable from $\mathsf{S}$ if and only if $T \backslash V$ is an algebraic category of $(\mathsf{F} - V)^*$. Moreover, if $(A, T)$ is evaluable from S, then its actual value is given by the inner product of the summary vector of $\mathsf{S}$ with any solution of system (6) where $\boldsymbol{u}$ is the support vector of $T \backslash V$.*

**Proof.** First of all, observe that by Theorem 2, the summary statistic $(A, T)$ is evaluable from $\mathsf{S}$ if and only if $H$ is a (possibly empty) subset $N$ and, if $J \neq \varnothing$, then the sum-expression $\sum_{j \in J} x_j$ is an invariant of system (4).
(*only if*) Assume that $(A, T)$ is evaluable from $\mathsf{S}$, so that $H \subseteq N$ and, if $J \neq \varnothing$, then the sum-expression $\sum_{j \in J} x_j$ is an invariant of system (4). Since $H \subseteq N$, by Lemma 2, $T \backslash V$ is an algebraic category of $(\mathsf{F} - V)^*$ and the statement follows from Lemma 8.

(*if*) Let us assume that $T \backslash V$ is an algebraic category of $(\mathsf{F} - V)^*$. By Lemma 2, $H \subseteq N$. Moreover, by Lemma 8 the sum-expression $\sum_{j \in J \backslash N} x_j$ is an invariant of the system (4) and, since $H \subseteq N$, the sum-expression $\sum_{j \in J} x_j$ is also an invariant of system (4). Therefore, by Theorem 1, $(A, T)$ is evaluable from $\mathsf{S}$.

Finally, if the summary statistic $(A, T)$ is evaluable from $\mathsf{S}$, then its actual value is zero if $J \subseteq N$ and is given by the value of the sum-expression $\sum_{j \in J \backslash N} x_j$ otherwise. In both cases, by Lemma 8 it equals the inner product of the summary vector of $\mathsf{S}$ with any solution of system (6). □

Please cite this article in press as: F.M. Malvestuto, et al., An analytical approach to the inference of summary data of additive type, Theoretical Computer Science (2007), doi:10.1016/j.tcs.2007.07.004

**Example 1** (*Continued*). Recall that $H = \{1\}$ and $J = \{2\}$. Suppose that $D = R^+$. Then, $N = \{1\}$, $V = \{X_1\}$ and $T\backslash V$ is as follows:

| GENDER | AGE-CLASS |
|---|---|
| Male | Middle-aged |
| Female | Middle-aged |

The basis of $(\mathcal{F} - V)^*$ is $Z = \{X_2, X_3\}$. By Theorem 4, the summary statistic $(A, T)$ is evaluable from the summary database S since $T\backslash V = X_2$ is trivially an algebraic category of $(\mathcal{F} - V)^*$. ■

## 6. Computational aspects

Given a summary database $\mathrm{S} = \{(A, K_1, a_1), \ldots, (A, K_n, a_n)\}$ with frame $C$ and a target category $T$, we now discuss the computational issues connected with the evaluability criteria given in Section 4.

Let $\mathcal{F} = \{K_1, \ldots, K_n\}$, $\mathcal{X} = \{X_1, \ldots, X_m\}$ be the basis of $\mathcal{F}^*$ and $\mathbf{B}$ be the incidence matrix of $\mathcal{F}^*$; thus, $\mathbf{B}$ is an $n \times m$ matrix. We first consider the cases $D = R$, $D = Z$ and $D = R^+$ and, then, the case $D = Z^+$.

($D = R$ or $D = Z$) By Theorem 3, we need to check that $T$ is an algebraic category of $\mathcal{F}^*$. To this end, we need to compute the sets $H$and $J$ defined by (1), which can be done simply as follows.

```
(1) $H := \varnothing$, $J := \varnothing$.
(2) For $j = 1, \ldots, m$
        if $X_j \cap T \neq \varnothing$    then
        if$X_j \backslash T \neq \varnothing$    then $H := H \cup \{j\}$ else $J := J \cup \{j\}$.
```

If subsets of dom$(C)$ are represented by Boolean vectors of size $k = |\mathrm{dom}(C)|$, finding $H$ and $J$ requires performing $O(mk)$ Boolean operations. At this point, by Corollary 1, we have to check that $H = \varnothing$, which takes $O$ (1) time. If $H \neq \varnothing$ then, by Corollary 1, we can soon conclude that $(A, T)$ is not evaluable from $S$. Otherwise, $T$ is a member of $\mathcal{F}^*$ and we test $T$ for algebraicity by checking the consistency of system (5), where $\mathbf{u}$ is the support vector of $T$ in $\mathcal{F}^*$ (that is, the characteristic vector of $J$). We can do it using the "LU-factorization method" [43] (also called the "Cholesky decomposition method"), which is based on the representation of the coefficient matrix of system (5) (i.e., the transpose of $\mathbf{B}$) in the form of a product of two matrices. Given these two matrices, which can be computed once and for all when the summary database is created, finding a solution (if any) of system (5) requires performing $O(mn)$ arithmetic operations. To sum up, our evaluability test for $D = R$ or $D = Z$ is strongly polynomial, since it requires $O(mk)$ Boolean operations and $O(mn)$ arithmetic operations.

($D = R^+$) In order to apply Theorem 4, we need to know the set of variables of system (4) that are zero-invariants. Since a variable of system (4) is a zero-invariant if and only if the maximum of its feasible values is zero, deciding whether or not variable $x_j$ is a zero-invariant of system (4) requires solving the linear-programming problem

maximise $x_j$ subject to system (4)

which can be solved in polynomial time using Karmarkar's algorithm [1,24]. Since the set of the variables of system (4) that are zero-invariants can be solved once and for all, by Theorem 4, the complexity of the evaluability test for $D = R^+$ is the same as for $D = R$.

($D = Z^+$) We can apply Theorem 2, which requires testing the invariance of the $p - m$ variables $x_h$ $(h \in H)$ and of the sum-expression $\sum_{j \in J} x_j$. Testing the invariance of such sum-expressions may be a hard problem since integer-linear programming methods are needed, and we shall prove the following result.

**Theorem 5.** *For $D = Z^+$, the problem of deciding if there exists at least one invariant variable of system* (4) *is coNP-complete.*

In order to prove Theorem 5,[1] We make use of the following decision problem, which is known to be *NP*-hard [10].

[1] Our proof is essentially the same as the proof of the co*NP*-completeness of the *Boolean Auditing Problem* (see Theorem 2.1 in [25]). Note that the proof of Theorem 2.1 given in [25] contains a flaw, and the correct proof can be found in our Lemma 10. (In a private communication, the authors of [25] agreed with us on the flaw and on the correctness of our proof.)

Please cite this article in press as: F.M. Malvestuto, et al., An analytical approach to the inference of summary data of additive type, Theoretical Computer Science (2007), doi:10.1016/j.tcs.2007.07.004

Given a positive integer $m$ and a collection $\{J_1, \ldots, J_n\}$ of nonempty subsets of $\{1, \ldots, m\}$ each of which has size less than 4, does the following system

$$\begin{cases} \sum_{j \in J_i} z_j = 1 & (1 \le i \le n) \\ z_j \in Z^+ & (1 \le j \le m) \end{cases} \tag{7}$$

have a solution?

We now provide a polynomial reduction of this problem to an instance of the problem of deciding that there is no invariant variable of system (4) where $D = Z^+$. The reduction is performed in five steps. Henceforth, all the variables we shall introduce are assumed to be defined on $Z^+$.

*Step* 1. For each $j$ $(1 \le j \le m)$, let us introduce fifteen variables $x_{j,1}, \ldots, x_{j,15}$ and the nine equations

$$x_{j,1} + x_{j,2} + x_{j,3} + x_{j,7} + x_{j,8} + x_{j,15} = 3 \tag{8a}$$

$$x_{j,4} + x_{j,5} + x_{j,6} + x_{j,9} + x_{j,10} + x_{j,15} = 3 \tag{8b}$$

$$x_{j,1} + x_{j,4} = 1 \qquad x_{j,2} + x_{j,5} = 1 \qquad x_{j,3} + x_{j,6} = 1 \tag{8c}$$

$$x_{j,7} + x_{j,11} = 1 \qquad x_{j,8} + x_{j,12} = 1 \tag{8d}$$

$$x_{j,9} + x_{j,13} = 1 \qquad x_{j,10} + x_{j,14} = 1 \tag{8e}$$

**Lemma 9.** *Every solution of the equation system* (8a)–(8e) *is binary.*

**Proof.** By Eqs. (8c)–(8e), in every solution $\mathbf{x}_j = (x_{j,1}, \ldots, x_{j,15})$, one has that $x_{j,1}, \ldots, x_{j,14}$ can only assume the values 0 and 1. We now prove that also $x_{j,15}$ can only assume the values 0 and 1. To this aim, we first show that the feasibilty range of the sum-expression

$$x_{j,1} + x_{j,2} + x_{j,3}$$

is $\{1, 2\}$. If $x_{j,1} + x_{j,2} + x_{j,3} = 0$, then one would have $x_{j,4} = x_{j,5} = x_{j,6} = 1$ by (8c) so that $x_{j,15} = 0$ by (8b) and, since $x_{j,7}$ and $x_{j,8}$ are binary, equation system (8a) would be inconsistent. If $x_{j,1} + x_{j,2} + x_{j,3} = 3$, then one would have $x_{j,15} = 0$ by (8a), $x_{j,4} = x_{j,5} = x_{j,6} = 0$ by (8c) and (8b) would be inconsistent. Therefore, the sum-expression $x_{j,1} + x_{j,2} + x_{j,3}$ can only assume the values 1 and 2. Finally, if $x_{j,1} + x_{j,2} + x_{j,3} = 1$, then $x_{j,4} + x_{j,5} + x_{j,6} = 2$ by (8c), and $x_{j,15}$ can only assume the values 0 and 1 by (8b); otherwise (that is, if $x_{j,1} + x_{j,2} + x_{j,3} = 2$), then $x_{j,15}$ can only assume the values 0 and 1 by (8a). □

It is easy to see that there are always twelve solutions of the equation system (8a)–(8e). To see this, distinguish two cases depending on the value of the sum-expression $x_{j,1} + x_{j,2} + x_{j,3}$, which is either 1 or 2 (see the proof of Lemma 9).

Case 1: $x_{j,1} + x_{j,2} + x_{j,3} = 1$. The solution set is given by the Cartesian product of

| $x_{j,1}$ | $x_{j,2}$ | $x_{j,3}$ | $x_{j,4}$ | $x_{j,5}$ | $x_{j,6}$ |
|---|---|---|---|---|---|
| 0 | 0 | 1 | 1 | 1 | 0 |
| 0 | 1 | 0 | 1 | 0 | 1 |
| 1 | 0 | 0 | 0 | 1 | 1 |

with

| $x_{j,7}$ | $x_{j,8}$ | $x_{j,9}$ | $x_{j,10}$ | $x_{j,11}$ | $x_{j,12}$ | $x_{j,13}$ | $x_{j,14}$ | $x_{j,15}$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |

Please cite this article in press as: F.M. Malvestuto, et al., An analytical approach to the inference of summary data of additive type, Theoretical Computer Science (2007), doi:10.1016/j.tcs.2007.07.004

Case 2: $x_{j,1} + x_{j,2} + x_{j,3} = 2$. The solution set is given by the Cartesian product of

| $x_{j,1}$ | $x_{j,2}$ | $x_{j,3}$ | $x_{j,4}$ | $x_{j,5}$ | $x_{j,6}$ |
|---|---|---|---|---|---|
| 0 | 1 | 1 | 1 | 0 | 0 |
| 1 | 0 | 1 | 0 | 1 | 0 |
| 1 | 1 | 0 | 0 | 0 | 1 |

with

| $x_{j,7}$ | $x_{j,8}$ | $x_{j,9}$ | $x_{j,10}$ | $x_{j,11}$ | $x_{j,12}$ | $x_{j,13}$ | $x_{j,14}$ | $x_{j,15}$ |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 |
| 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |

**Remark 3.** Even if the value of $x_{j,1}+x_{j,2}+x_{j,3}$ is fixed, none of the fifteen variables of the equation system (8a)–(8e) is an invariant, and the feasibility range of the sum-expression $\sum_{h=1,\ldots,15} x_{j,h}$ is {7, 8}.

*Step* 2. For each $j$ $(1 \leq j \leq m)$, we replace each (binary) variable $z_j$ in system (6) by the sum-expression $x_{j,1} + x_{j,2} + x_{j,3} - 1$ and add the nine Eqs. (8a)–(8e). Let

$$\begin{cases} \sum\limits_{j \in J_i} x_{j,1} + x_{j,2} + x_{j,3} = 1 + |J_i| & (1 \leq i \leq n) \\ \begin{cases} \text{Eqs. (8a)–(8e)} \\ x_{j,1}, \ldots, x_{j,15} \in Z^+ \end{cases} & (1 \leq j \leq m). \end{cases} \tag{9}$$

Note that each constant term of system (9) is not greater than 4. The solution set of system (9) is the set of (binary) vectors $(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ where

— each $\mathbf{x}_j$ is a solution of the corresponding equation system (8a)–(8e), and
— the vector $\mathbf{z} = (z_1, \ldots, z_m)$, where $z_j = x_{j,1} + x_{j,2} + x_{j,3} - 1$, is a solution of system (7).

Therefore, system (9) is inconsistent if and only if system (7) is inconsistent and, if system (7) is consistent then, by Lemma 9, every solution of system (9) is binary. Moreover, for every solution $\mathbf{z}$ of system (7), there are exactly $12^m$ solutions of system (9) so that, if system (7) is consistent, then by Remark 3 no variable of system (9) is an invariant and the feasibility range of the sum-expression

$$\sum_{j=1,\ldots,m} \left( \sum_{h=1,\ldots,15} x_{j,h} \right)$$

is $\{7m, 7m + 1, \ldots, 8m - 1, 8m\}$.

*Step* 3. Let us introduce seven more variables: $x_{0,1}, \ldots, x_{0,7}$ and the following system of six equations

$$\begin{cases} x_{0,1} + x_{0,5} = 1 & x_{0,2} + x_{0,5} = 1 \\ x_{0,2} + x_{0,6} = 1 & x_{0,3} + x_{0,6} = 1 \\ x_{0,3} + x_{0,7} = 1 & x_{0,4} + x_{0,7} = 1 \end{cases} \tag{10}$$

System (10) has exactly two (binary) solutions:

$$\mathbf{x}_0 = (0, 0, 0, 0, 1, 1, 1) \quad \text{and} \quad \mathbf{x}_0 = (1, 1, 1, 1, 0, 0, 0),$$

in correspondence of which the sum-expression $\sum_{h=1,\ldots,7} x_{0,h}$ takes on the values 3 and 4, respectively.

*Step* 4. For each $i$ $(1 \leq i \leq n+9m)$, add to the left-hand side of the $i$-th equation in system (9), the sum-expression

$$\sum_{h=1,\ldots,c_i} x_{0,h}$$

Please cite this article in press as: F.M. Malvestuto, et al., An analytical approach to the inference of summary data of additive type, Theoretical Computer Science (2007), doi:10.1016/j.tcs.2007.07.004

where $c_i$ is the constant term of the $i$-the equation of system (9); moreover, add the six equations of system (10). Thus, the first $n$ equations of the resulting system are like

$$\sum_{h=1,\dots,1+|J_i|} x_{0,h} + \sum_{j \in J_i} x_{j,1} + x_{j,2} + x_{j,3} = 1 + |J_i| \quad (1 \le i \le n).$$

Let

$$\mathbf{Hx} = \mathbf{a} \tag{11}$$

be the resulting equation system, where $\mathbf{x} = (\mathbf{x}_o, \mathbf{x}_1, \dots, \mathbf{x}_m)$ and $\mathbf{a}$ is a vector of size $n + 9m + 6$.

**Lemma 10.** *If system (7) is consistent, then system (11) is consistent, no variable of system (11) is an invariant and the sum-expression*

$$\sum_{h=1,\dots,7} x_{0,h} + \sum_{j=1,\dots,m} \left( \sum_{h=1,\dots,15} x_{j,h} \right)$$

*is not an invariant. Otherwise, system (11) has exactly one solution and the above sum-expression is an invariant with value* 4.

**Proof.** If system (7) is consistent, then system (11) is consistent since, given a solution $\mathbf{z}$ of system (7), we can obtain $12^m$ solutions $\mathbf{x} = (\mathbf{x}_o, \mathbf{x}_1, \dots, \mathbf{x}_m)$ of system (11) by taking $\mathbf{x}_o = (0, 0, 0, 0, 1, 1, 1)$ and setting $\mathbf{x}_j$, $1 \le j \le m$, to each of the twelve solutions of the equation system (8a)–(8e) having $x_{j,1} + x_{j,2} + x_{j,3}$ equal to $z_j + 1$. If system (7) is inconsistent, then system (11) has exactly one solution $\mathbf{x} = (\mathbf{x}_o, \mathbf{x}_1, \dots, \mathbf{x}_m)$ with $\mathbf{x}_o = (1, 1, 1, 1, 0, 0, 0)$ and $\mathbf{x}_j = \mathbf{0}$ for all $j$ $(1 \le j \le m)$. Moreover, in the former case, no variable of system (11) is an invariant and the feasibility range of the sum-expression

$$\sum_{h=1,\dots,7} x_{0,h} + \sum_{j=1,\dots,m} \left( \sum_{h=1,\dots,15} x_{j,h} \right)$$

is $\{4, 7m+3, \dots, 8m+3\}$; in the latter case, each variable of system (11) is an invariant and the above sum-expression is an invariant with value 4. □

By Lemma 10, the problem of deciding that there exists at least one invariant variable of an equation system such as system (11) is co*NP*-complete. Moreover, as pointed out in [25], it is co*NP*-hard to distinguish between the case in which all variables are invariant and the case in which none is; as a consequence, telling whether a specific variable is an invariant is also a co*NP*-complete problem.

*Step* 5. Let us introduce one more variable $w$ and consider the constraint system obtained from system (11) by adding the equation

$$\sum_{h=1,\dots,7} x_{0,h} + \sum_{j=1,\dots,m} \left( \sum_{h=1,\dots,15} x_{j,h} \right) + w = 8m + 3.$$

Let

$$\mathbf{B}(\mathbf{x}, w) = (\mathbf{a}, 8m + 3) \tag{12}$$

be the resulting equation system, where $\mathbf{B} = \begin{bmatrix} \mathbf{H} & 0 \\ \mathbf{1} & \end{bmatrix}$.

**Proof of Theorem 5.** First of all, observe that, since $\mathbf{B}$ is a binary matrix and contains the all-one row, system (12) can be viewed as being an instance of system (4) for $D = Z^+$. By Lemma 10, we have the following:

— if system (7) is consistent, then system (12) is consistent and no variable of system (12) is an invariant (the feasible values of the variable $w$ are $0, 1, \dots, m, 8m - 1$).

— if system (7) is inconsistent, then system (12) has exactly one solution so that each variable is an invariant (the value of $w$ is $8m - 1$).

Please cite this article in press as: F.M. Malvestuto, et al., An analytical approach to the inference of summary data of additive type, Theoretical Computer Science (2007), doi:10.1016/j.tcs.2007.07.004

Since deciding the consistency of system (7) is an *NP*-hard problem, deciding that there exists at least one invariant variable of system (12) is a co*NP*-complete problem. □

## 7. Nonevaluability

Let $\mathsf{S} = \{(A, K_1, a_1), \dots, (A, K_n, a_n)\}$ be a summary database whose summary attribute $(A)$ has domain $R$ or $Z$ or $R^+$. In this section, we discuss the case that a summary statistic $(A, T)$ is not evaluable from S, and how to answer to the question: what about the actual value of $(A, T)$?

Let $(a', a'')$ be the range of feasible values of $(A, T)$. If the domain of $A$ is $R^+$, then $(a', a'')$ is a closed interval so that it is an informative answer to the question above. But, if the domain of $A$ is $R$ or $Z$, then $a' = -\infty$ and $a'' = +\infty$, so that issuing the feasibility range of $(A, T)$ gives no information about the actual value of $(A, T)$. In this case, knowing a nonempty subset $T'$ of $T$ such that the summary statistic $(A, T')$ evaluable from $\mathsf{S}$ would allow us to reduce the amount of numeric computation in that, after evaluating the summary statistic $(A, T\backslash T')$ over the raw relation, the actual value of $(A, T)$ could be obtained by adding the actual values of $(A, T')$ and $(A, T\backslash T')$. On the other hand, if the raw data are not longer available then, as suggested by some authors [8,18,41,42], it could be convenient to know the actual values of summary statistics that are evaluable from $\mathsf{S}$ and are "close" to $(A, T)$, e.g., the actual values of two evaluable summary statistics $(A, T')$ and $(A, T'')$, where $T'$ is a maximal subset of $T$and $T''$ is a minimal superset of $T$. We shall prove that in general, the problem of finding categories such as $T'$ and $T''$ is *NP*-hard. However, after the submission of this paper, one of the authors [37] proved that, for incidence matrices with up to two nonzero entries per column (that is, for incidence matrices of graphs), the problem can be solved in linear time.

Let $\mathsf{F} = \{K_1, \dots, K_n\}$, let $\mathsf{X} = \{X_1, \dots, X_m\}$ be the basis of $\mathsf{F}^*$ and let $\mathbf{B}$ be the incidence matrix of $\mathsf{F}^*$. Let $H$ and $J$ be the subsets of $\{1, \dots, m\}$ defined by (1). By Theorem 3, one has that:

— $T'$ is a maximal subset of $T$ for which $(A, T')$ is evaluable from $\mathsf{S}$ if and only if $T'$ is an algebraic category of $\mathsf{F}^*$ that is maximally contained in the category $\cup_{j \in J} X_j$;

— $T''$ is a minimal superset of $T$ for which $(A, T'')$ is evaluable from $\mathsf{S}$ if and only if $T''$ is an algebraic category of $\mathsf{F}^*$ that minimally contains the category $\cup_{j \in H \cup J} X_j$.

Therefore, the problem of finding $T'$ (or $T''$) consists in searching for a maximal algebraic subset (a minimal algebraic superset, respectively) of a given category in $\mathsf{F}^*$. Note that, since dom$(C)$ is an algebraic category of $\mathsf{F}^*$, by the closure under complementation, one has that

—$T''$ is an algebraic category of $\mathsf{F}^*$ that minimally contains the category $\cup_{j \in H \cup J} X_j$ if and only if dom$(C)\backslash T''$ is an algebraic category of $\mathsf{F}^*$ that is maximally contained in the category $\cup_{j \notin H \cup J} X_j$.

Therefore, we can limit our considerations to the problem of finding a maximal algebraic subset of the category $\cup_{j \in J} X_j$. Of course, this problem is not less hard than the following problem, we call the *Nonempty Algebraic Subset problem* (the NAS problem, for short):

Given a binary matrix $\mathbf{B}$ of size $n \times m$ that contains an all-one row and given a nonempty subset $J$ of $\{1, \dots, m\}$, find a nonempty subset $J'$of $J$ such that the characteristic vector $\mathbf{u}$ of $J'$ is a linear combination of rows of $\mathbf{B}$.

We shall prove that the NAS problem is *NP*-complete. Therefore, generally speaking, the only efficient choice for $T'$ and $T''$ are the empty set and dom$(C)$, respectively.

Our proof of *NP*-completeness is obtained by providing a polynomial reduction of the *NP*-complete problem that in the literature (see [10]) is referred to as the *Subset Sum problem* (the (SS) problem, for short). Before proving the statement, we first recall the SS problem: (SS) Given an integer $k \geq 2$, and an integral vector $\mathbf{s} = (s_1, \dots, s_k)$, where $s_1, \dots, s_{k-1}$ are all positive and $s_k$ is negative, is there a nonzero binary vector $\mathbf{q} = (q_1, \dots, q_k)$ such that $\sum_{h=1,\dots,k} q_h\, s_h = 0$?

First of all, we give a characterisation of solutions of the SS problem which will be used to prove the *NP*-completeness of the NAS problem.

Let $n > k$, let $J = \{i_1, \dots, i_k\}$ be a subset of $\{1, \dots, n\}$, let $\mathbf{a} = (a_1, \dots, a_n)$ be a row vector such that $a_{i_h} = s_h$ for each $h$ $(1 \leq h \leq k)$, and let $\mathbf{E}$ be the $n \times (n + 1)$ matrix obtained from the identity $\mathbf{I}_n$ matrix of size $n$ by adding one more column given by the transpose of $\mathbf{a}$, that is,

$$\mathbf{E} = \begin{bmatrix} 1 & 0 & \dots & 0 & a_1 \\ 0 & 1 & \dots & 0 & a_2 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & a_n \end{bmatrix} = \begin{bmatrix} \mathbf{I}_n & \mathbf{a}^{\mathrm{T}} \end{bmatrix}.$$

Note that the rank of $\mathbf{E}$ (and, hence, the dimension of the row space of $\mathbf{E}$) is $n$.

**Lemma 11.** *Let $\mathbf{q} = (q_1, \dots, q_k)$ be a nonzero binary vector and let $\mathbf{u} = (u_1, \dots, u_n, u_{n+1})$ be the binary vector with $u_{i_h} = q_h$ for each $h$ $(1 \le h \le k)$, and $u_i = 0$ for each $i \notin \{1, \dots, n+1\} \backslash J$. Then, $\mathbf{q}$ is a solution of the SS problem if and only if $\mathbf{u}$ belongs to the row space of $\mathbf{E}$.*

**Proof.** Recall that $\mathbf{u}$ belongs to the row space of $\mathbf{E}$ if and only if $\mathbf{u}$ is orthogonal to the kernel of $\mathbf{E}$, that is, if and only if $\mathbf{u}$ is orthogonal to every solution $\mathbf{v} = (v_1, \dots, v_n, v_{n+1})$ of the homogeneous equation system $\mathbf{E}\,\mathbf{v} = \mathbf{0}$, which consists of the $n$ equations

$$v_i + a_i v_{n+1} = 0 \quad (1 \le i \le n).$$

The general solution of $\mathbf{E}\,\mathbf{v} = \mathbf{0}$ has

$$v_i = -a_i v_{n+1} \quad (1 \le i \le n)$$

and, hence, the kernel of $\mathbf{E}$ (has dimension 1 and) is spanned by the vector

$$(v_1 = a_1, \dots, v_n = a_n, v_{n+1} = -1).$$

Therefore, $\mathbf{u}$ belongs to the row space of $\mathbf{E}$ if and only if $\mathbf{u}$ is orthogonal to $(a_1, \dots, a_n, -1)$, that is, if and only if

$$\sum_{i=1,\dots,n} u_i a_i - u_{n+1} = 0.$$

On the other hand, $u_i = 0$ for each $i \notin J$ so that $\mathbf{u}$ belongs to the row space of $\mathbf{E}$ if and only if

$$\sum_{h=1,\dots,k} u_{i_h} a_{i_h} = 0.$$

Finally, since $u_{i_h} = q_h$ and $a_{i_h} = s_h$ for each $h$ $(1 \le h \le k)$, one has that $\mathbf{u}$ belongs to the row space of $\mathbf{E}$ if and only if

$$\sum_{h=1,\dots,k} q_h s_h = 0,$$

that is, if and only if $\mathbf{q}$ is a solution of the (SS) problem. □

In what follows, the matrix $\mathbf{E}$ and the vector $\mathbf{u}$ will be referred to as the matrix *associated* with $\mathbf{a}$ and the *extension* of $\mathbf{q}$, respectively. Thus, by Lemma 11, $\mathbf{q}$ is a solution of the SS problem if the extension of $\mathbf{q}$ belongs to the row space of the matrix associated with $\mathbf{a}$.

In order to prove the *NP*-completeness of the NAS problem, we first construct a vector $\mathbf{a}$ such that the row space of the matrix $\mathbf{E}$ associated with $\mathbf{a}$ enjoys the following two properties:

(i) the all-one vector ($\mathbf{1}$) belongs to the row space of $\mathbf{E}$,
(ii) the row space of $\mathbf{E}$ has a vector base $\{\mathbf{b}_1, \dots, \mathbf{b}_n\}$ where each $\mathbf{b}_i$ is a binary vector.

The vector $\mathbf{a}$ is defined in terms of $k+1$ vectors $\mathbf{a}_1, \dots, \mathbf{a}_k, \mathbf{a}_{k+1}$ we now introduce. We first define $\mathbf{a}_1, \dots, \mathbf{a}_{k-1}$, next $\mathbf{a}_k$ and, finally, $\mathbf{a}_{k+1}$. In what follows, the number of components of $\mathbf{a}_l$ $(1 \le l \le k+1)$ and the sum of its components will be denoted by $|\mathbf{a}_l|$ and $\|\mathbf{a}_l\|$, respectively.

$(1 \le l \le k-1)$. Let $n_l = \lfloor \log_2 s_l \rfloor$, where $\lfloor x \rfloor$ denotes the highest integer that is not greater than $x$. If $n_l = 0$ (that is, if $s_l = 1$), then $|\mathbf{a}_l| = 1$ and $\mathbf{a}_l = (s_l)$; otherwise, $|\mathbf{a}_l| = 3n_l + 1$ and

$$\mathbf{a}_l = (s_l, -\lfloor s_l/2 \rfloor, -\lfloor s_l/2 \rfloor, \lfloor s_l/2 \rfloor, \dots, -\lfloor s_l/2^{n_l} \rfloor, -\lfloor s_l/2^{n_l} \rfloor, \lfloor s_l/2^{n_l} \rfloor)$$

$(l = k)$. Let $n_k = \lceil \log_2 |s_k| \rceil$, where $\lceil x \rceil$ denotes the least integer that is not less than $x$. If $n_k = 0$ (that is, if $s_k = -1$), then $|\mathbf{a}_k| = 3$ and $\mathbf{a}_k = (s_k, 1, 1)$; otherwise (that is, if $s_k < -1$), $|\mathbf{a}_k| = 3n_k$ and

$$\mathbf{a}_k = (s_k, \lceil |s_k|/2 \rceil, \lceil |s_k|/2 \rceil, -\lceil |s_k|/2 \rceil, \dots, \lceil |s_k|/2^{n_k} \rceil, \lceil |s_k|/2^{n_k} \rceil)$$

$(l = k + 1)$. Let $r = \sum_{l=1,\dots,k} \|\mathbf{a}_l\|$. Then $|\mathbf{a}_{k+1}| = r + 1$ and

$$\mathbf{a}_{k+1} = (-1 - 1 \dots - 1\ 1).$$

Note that $\|\mathbf{a}_{k+1}\| = 1 - r$.

At this point, we are in a position to define the vector $\mathbf{a}$. The vector $\mathbf{a}$ is obtained by collecting the components of $\mathbf{a}_1, \dots, \mathbf{a}_k, \mathbf{a}_{k+1}$ in this order. Thus, the number of components of $\mathbf{a}$ is

$$n = \left( \sum_{l=1,\dots,k-1} 3n_l + 1 \right) + 3n_k + r + 1.$$

Accordingly, the matrix $\mathbf{E}$ associated with $\mathbf{a}$ has size $n \times (n + 1)$. Explicitly, one has

$$\mathbf{E} = \begin{bmatrix} \mathbf{I}_n & \begin{matrix} \mathbf{a}_1^\top \\ \dots \\ \mathbf{a}_k^\top \\ \mathbf{a}_{k+1}^\top \end{matrix} \end{bmatrix}.$$

**Example 3.** Consider the SS problem with $k = 3$ and $s_1 = s_2 = 1$ and $s_3 = -2$. Trivially, it has exactly one solution: $\mathbf{q} = (1, 1, 1)$. Then $\mathbf{a}_1 = (1)$, $\mathbf{a}_2 = (1)$ and $\mathbf{a}_3 = (-2, 1, 1)$. Since $|\mathbf{a}_1| = |\mathbf{a}_2| = 1$ and $|\mathbf{a}_3| = 0$, one has $\mathbf{a}_4 = (-1, -1, 1)$ and the matrix $\mathbf{E}$ associated with the vector $\mathbf{a} = (\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4)$ is

$$\mathbf{E} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -2 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

The kernel of $\mathbf{E}$ is spanned by the 9-dimensional vector $(1, 1, -2, 1, 1, -1, -1, 1, -1)$, so that the extension $\mathbf{u} = (1, 1, 1, 0, 0, 0, 0, 0, 0)$ of $\mathbf{q}$ is orthogonal to the kernel of $\mathbf{E}$ and hence, belongs to the row space of $\mathbf{E}$. ■

The next lemma shows that the matrix $\mathbf{E}$ enjoys property (i).

**Lemma 12.** *The $(n + 1)$-dimensional vector $\mathbf{1}$ belongs to the row space of* $\mathbf{E}$.

**Proof.** Let $\mathbf{E} = (\mathbf{e}_i)_{i=1,\dots,n}$. We now show that $\sum_{i=1,\dots,n} \mathbf{e}_i = \mathbf{1}$. For each $j$ $(1 \le j \le n)$, one has trivially $\sum_{i=1,\dots,n} e_{i,j} = 1$. For $j = n + 1$, one has

$$\sum_{i=1,\dots,n} e_{i,n+1} = \sum_{i=1,\dots,k+1} \|\mathbf{a}_l\| = \sum_{l=1,\dots,k} \|\mathbf{a}_l\| + \|\mathbf{a}_{k+1}\| = r + (1 - r) = 1,$$

which proves the statement. □

In order to prove that the matrix $\mathbf{E}$ enjoys property (ii), we construct the vectors $\mathbf{b}_1, \dots, \mathbf{b}_n$ as follows. For each $l$ $(1 \le l \le k + 1)$, let $\mathbf{e}_{m(l)}, \dots, \mathbf{e}_{M(l)}$ be the rows of $\mathbf{E}$ corresponding to $\mathbf{a}_l$. Explicitly, one has

$$\begin{array}{ll} m(1) = 1 & M(1) = |\mathbf{a}_1| \\ m(2) = |\mathbf{a}_1| + 1 & M(2) = |\mathbf{a}_1| + |\mathbf{a}_2| \\ \dots & \dots \\ m(k+1) = |\mathbf{a}_1| + \dots + |\mathbf{a}_k| + 1 & M(k+1) = |\mathbf{a}_1| + \dots + |\mathbf{a}_k| + |\mathbf{a}_{k+1}| (= n) \end{array}$$

Please cite this article in press as: F.M. Malvestuto, et al., An analytical approach to the inference of summary data of additive type, Theoretical Computer Science (2007), doi:10.1016/j.tcs.2007.07.004

The vectors $\mathbf{b}_{m(l)} \ldots, \mathbf{b}_{M(l)}$ are the output of the following three procedures, one for $l \leq k-1$, the other for $l = k$ and the last for $l = k+1$.

$(1 \leq l \leq k-1)$

```
(1) $i := m(l)$
(2)     if $M(l) = m(l)$, then $\mathbf{b}_i := \mathbf{e}_i$;
        otherwise, while $i < M(l)$ do
            begin
            $\mathbf{b}_i := \mathbf{e}_i + \mathbf{e}_{i+1} + \mathbf{e}_{i+2}$
            $\mathbf{b}_{i+1} := \mathbf{e}_{i+1} + \mathbf{e}_{i+3}$
            $\mathbf{b}_{i+2} := \mathbf{e}_{i+2} + \mathbf{e}_{i+3}$
            $i := i + 3$
            end
```

$(l = k)$

```
(1)     $i := m(k)$
(2)     while $i < M(k)$ do
            begin
            $\mathbf{b}_i := \mathbf{e}_i + \mathbf{e}_{i+1} + \mathbf{e}_{i+2}$
            if $i + 3 \leq M(k)$ then do
                begin
                $\mathbf{b}_{i+1} := \mathbf{e}_{i+1} + \mathbf{e}_{i+3}$
                $\mathbf{b}_{i+2} := \mathbf{e}_{i+2} + \mathbf{e}_{i+3}$
                end
            $i := i + 3$
            end
```

$(l = k+1)$

```
(1)     For $i = m(k+1)$ to $n-1$, $\mathbf{b}_i := \mathbf{e}_i + \mathbf{e}_n$ ;
(2)     $\mathbf{b}_n := \mathbf{e}_n$.
```

It is easy to check that the vectors $\mathbf{b}_1, \ldots, \mathbf{b}_n$ are all binary vectors.

**Example 3** (*Continued*). Starting from $\mathbf{E}$, we obtain the following eight binary vectors:

$\mathbf{b}_1 = (1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1)$
$\mathbf{b}_2 = (0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 1)$
$\mathbf{b}_3 = (0\ 0\ 1\ 1\ 1\ 0\ 0\ 0\ 0)$
$\mathbf{b}_4 = (0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 1)$
$\mathbf{b}_5 = (0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 1)$
$\mathbf{b}_6 = (0\ 0\ 0\ 0\ 0\ 1\ 0\ 1\ 0)$
$\mathbf{b}_7 = (0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 0)$
$\mathbf{b}_8 = (0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1)$. ■

The next lemma shows that the matrix $\mathbf{E}$ enjoys property (ii) by proving that $\{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ is a vector base of the row space of $\mathbf{E}$.

**Lemma 13.** *The set* $\{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ *is a vector base of the row space of* $\mathbf{E}$.

**Proof.** First of all, note that the rank of $\mathbf{E}$ is $n$ (i.e., the dimension of the row space of $\mathbf{E}$ is $n$). On the other hand, by construction, the rows of $\mathbf{B}$ belong to the row space of $\mathbf{E}$ so that $\{\mathbf{b}_1, \ldots, \mathbf{b}_n\}$ is a vector base of the row space of $\mathbf{E}$

Please cite this article in press as: F.M. Malvestuto, et al., An analytical approach to the inference of summary data of additive type, Theoretical Computer Science (2007), doi:10.1016/j.tcs.2007.07.004

if and only if $\mathbf{b}_1, \ldots, \mathbf{b}_n$ are linearly independent. Owing to the structure of **E**, it is sufficient to prove that, for each vector $\mathbf{a}_l$ the corresponding vector set $\{\mathbf{b}_{m(l)}, \ldots, \mathbf{b}_{M(l)}\}$ is linearly independent, that is, that the following equation with unknowns $c_p, \ldots, c_q$

$$c_{m(l)}\mathbf{b}_{m(l)} + c_{m(l)+1}\mathbf{b}_{m(l)+1} + \cdots + c_{M(l)}\mathbf{e}_{M(l)} = \mathbf{0}$$

is satisfied only for $c_{m(l)} = \cdots = c_{M(l)} = 0$. Again, we first consider the case $1 \le l \le k-1$, then the case $l = k$ and, finally, the case $l = k+1$.

$(1 \le l \le k-1)$. Using the definition of $\mathbf{b}_i$, $m(l) \le i \le M(l)$, the equation above can be re-written as

$$c_{m(l)}(\mathbf{e}_{m(l)} + \mathbf{e}_{m(l)+1} + \mathbf{e}_{m(l)+2}) + c_{m(l)+1}(\mathbf{e}_{m(l)+1} + \mathbf{e}_{m(l)+3}) + \cdots = \mathbf{0}$$

and hence

$$c_{m(l)}\mathbf{e}_{m(l)} + (c_{m(l)} + c_{m(l)+1})\mathbf{e}_{m(l)+1} + (c_{m(l)} + c_{m(l)+2})\mathbf{e}_{m(l)+2} + \cdots = \mathbf{0}.$$

Since the rows of **E** are linearly independent, one has

$$c_{m(l)} = 0 \qquad c_{m(l)} + c_{m(l)+1} = 0 \qquad c_{m(l)} + c_{m(l)+2} = 0 \qquad \ldots$$

and, hence, $c_{m(l)} = c_{m(l)+1} = c_{m(l)+2} = \cdots = 0$. Therefore, the vectors $\mathbf{b}_{m(l)}, \ldots, \mathbf{b}_{M(l)}$ are linearly independent.

$(l = k)$. The proof is analogous to the previous one.

$(l = k+1)$. Using the definition of $\mathbf{b}_i$, $m(k+1) \le i \le n$, the equation above can be re-written as

$$c_{m(k+1)}(\mathbf{e}_{m(k+1)} + \mathbf{e}_n) + c_{m(k+1)+1}(\mathbf{e}_{m(k+1)+1} + \mathbf{e}_n) + \cdots + c_{n-1}(\mathbf{e}_{n-1} + \mathbf{e}_n) + c_n\mathbf{e}_n = \mathbf{0}$$

and hence

$$c_{m(k+1)}\mathbf{e}_{m(k+1)} + c_{m(k+1)+1}\mathbf{e}_{m(k+1)+1} + \cdots + (c_{m(k+1)} + \cdots + c_n)\mathbf{e}_n = \mathbf{0}.$$

Since the rows of **E** are linearly independent, one has

$$c_{m(k+1)} = 0 \qquad c_{m(k+1)+1} = 0 \ldots \qquad c_{m(k+1)} + \cdots + c_n = 0$$

which entail $c_{m(k+1)} = c_{m(k+1)+1} = \cdots = c_{n-1} = c_n = 0$. Therefore, the vectors $\mathbf{b}_{m(k+1)}, \ldots, \mathbf{b}_n$ are linearly independent. □

**Theorem 6.** *The NAS problem is NP-complete.*

**Proof.** First of all, note that the NAS problem is in *NP* (see Section 3). Consider now the index set $J$ and the binary matrix

$$\mathbf{B} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \ldots \\ \mathbf{b}_n \\ \mathbf{1} \end{bmatrix}$$

obtained by reduction of the SS problem (see above). By Lemmas 12 and 13, the matrix **E** have the same row space of **B** . Therefore, by Lemma 11, the SS problem has a solution if and only if there exists a binary vector of the row space of **B** such that **u** is the characteristic vector of a nonempty subset of $J$. Finally, our reduction is polynomial, since the sizes of **a** and **B** are polynomial in the size of the SS problem, which completes the proof of *NP*-completeness of the NAS problem. □

## 8. Conclusions

We have proposed a general framework for the inference problem and the containment problem in a summary database whose summary attribute is additive. In our framework, we have proved that the inference problem can be solved in polynomial time (using standard linear-algebra algorithms) if the summary attribute is of real type, of integer type or of nonnegative-real type; but, it becomes intractable if the summary attribute is of nonnegative-integer type.

Moreover, even in the simplest case (i.e., the summary attribute is of real type), the containment problem turns out to be *NP*-hard.

The complexity of the inference problem and the containment problem can be reduced by imposing some special structure on the underlying summary database. For example, with a "graphical" summary database [32,33] (as for a 2-dimensional table with suppressed cells [23,30]), one has that

— if the summary attribute is of real type, then the complexity of the inference problem is linear [34];
— if the summary attribute is of nonnegative-real type, then there exists a linear algorithm to find the set of all zero-invariant variables [30] and a network-flow algorithm to find the feasibility range of every variable [31].

These results are related to general properties of matrices with up to two nonzero entries per column [20], and how to exploit them to deal with the case that the summary attribute is of nonnegative-integer type is an open question, which has been answered only if the matrix is totally unimodular [43].

Another open question is the existence of some wider class of summary databases (or constraint matrices) for which either problem can be solved efficiently. Finally, another direction for future research is represented by "semiadditive" summary databases [4,5] (as when the summary data are the results of the application to some aggregation attribute of the aggregate functions *min* and *max*).

## Acknowledgments

The authors thank the referees for their helpful suggestions and comments.

## References

[1] I. Adler, N. Karmarkar, M.G.C. Resende, G. Veiga, An implementation of Karmarkar's algorithm for linear programming, Math. Program. 44 (1989) 279–335 (Errata Math. Program. 50 (1991) 41).
[2] R. Agrawal, A. Gupta, S. Sarawagi, Modeling multidimensional databases, in: Proc. ICDE 97, IEEE Comp. Soc, pp. 435–452.
[3] M.C. Chen, L. McNamee, M.A. Melkanoff, A model of summary data and its applications in statistical databases, in: Proc. IV Int. Working Conf. on "Statistical & Scientific Database Management", in: M. Rafanelli, J.C. Klensin, P. Svensson (Eds.), Lecture Notes in Computer Sciences, vol. 339, Springer-Verlag, 1989, pp. 354–372.
[4] M.C. Chen, L. McNamee, On the data model and access method of summary data management, IEEE Trans. Knowl. Data Eng. 1 (1989) 519–529.
[5] F. Chin, Security problems on inference control for SUM, MAX, and MIN queries, J. ACM TODS 33 (1986) 451–464.
[6] F.Y. Chin, G. Özsoyoglu, Statistical database design, ACM TODS 6 (1981) 113–139.
[7] F.Y. Chin, G. Özsoyoglu, Auditing and inference control in statistical databases, IEEE Trans. Softw. Eng. 8 (1982) 574–582.
[8] S. Cohen, W. Nutt, A. Serebrenik, Rewriting aggregate queries using views, in: Proc. XVIII ACM Symp. on "Principles of Database Systems", 1999, pp. 155–166.
[9] A. Dobra, S.E. Fienberg, Bounds for cell entries in contingency tables given marginal totals and decomposable graphs, Proc. Natl. Acad. Sci. USA 97 (2000) 11885–11892.
[10] M.R. Garey, D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, W. H. Freeman, San Francisco, 1979.
[11] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, H. Pirahesh, Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals, Data Mining Knowl. Discovery 1 (1997) 29–53.
[12] S. Grumbach, M. Rafanelli, L. Tinini, Querying aggregate data, in: Proc. XVIII ACM Symp. on "Principles of Database Systems", 1999, pp. 174–184.
[13] S. Grumbach, L. Tininini, On the content of materialized aggregate views, J. Comput. Syst. Sci. 66 (2003) 133–168. A preliminary version appeared in Proc. XIX ACM Symp. on "Principles of Database Systems", 2000.
[14] A. Gupta, Selection of views to materialize in a data warehouse, in: Proc. VI Int. Conf. on "Database Theory", vol. 1186, Springer, 1997, pp. 98–112.
[15] A. Gupta, I.S. Mumik, Selection of views to materialize under a maintenance cost constraint, in: Proc. VII Int. Conf. on "Database Theory", 1999, pp. 453–470.
[16] M. Gyssens, L.V.S. Lakshmanan, A foundation form multi-dimensional databases, in: Proc. Int. Conf. on "Very Large Data Bases", 1996, pp. 453–470.
[17] M. Gyssens, L.V.S. Lakshmanan, I.N. Subramanian, Tables as a paradigm for querying and restructuring, in: Proc. XV ACM Symp. on "Principles of Database Systems", 1996, pp. 93–103.
[18] A.Y. Halevy, Answering queries using views: A survey, VLDB J. 10 (2001) 270–294.
[19] V. Harinayan, A. Rajaraman, J.D. Ullman, Implementing data cubes efficiently, in: Proc. ACM Int. Conf. on "Management of Data", 1996, pp. 205–216.
[20] D.S. Hochbaum, Monotonizing linear programs with up to two nonzeroes per column, Oper. Res. Lett. 32 (2004) 49–58.
[21] M.Y. Kao, Efficient detection and protection of information in cross-tabulated tables II: Minimal linear invariants, J. Combin. Optim. 1 (1997) 187–202.

[22] M.Y. Kao, Total protection of analytic-invariant information in cross-tabulated tables, SIAM J. Comput. 26 (1997) 231–242.

[23] M.Y. Kao, D. Gusfield, Efficient detection and protection of information in cross-tabulated tables I: Linear invariant test, SIAM J. Discrete Math. 6 (1993) 460–473.

[24] N. Karmarkar, A new polynomial-time algorithm for linear programming, Combinatorica 4 (1984) 373–392.

[25] J.M. Kleinberg, C.H. Papadimitriou, P. Raghavan, Auditing Boolean attributes, J. Comput. System Sci. 66 (2003) 244–253. A preliminary version appeared in Proc. XIX ACM Symp. on "Principles of Database Systems", 2000.

[26] H.-J. Lenz, A. Shoshani, Summarizability in OLAP and statistical data bases, in: Proc. IX Int. Conf. on "Scientific and Statistical Database Management", 1997, pp. 132–143.

[27] D. Maier, The Theory of Relational Databases, Computer Science Press, Rockville, 1983.

[28] F.M. Malvestuto, The derivation problem of summary data, in: Proc. ACM Int. Conf. on "Management of Data", 1988, pp. 82–89.

[29] F.M. Malvestuto, A universal-scheme approach to statistical databases containing homogeneous summary tables, ACM Trans. Database Syst. 18 (1993) 678–708.

[30] F.M. Malvestuto, M. Mezzini, A linear algorithm for finding the invariant edges of an edge-weighted graph, SIAM J. Comput. 31 (2002) 1438–1455.

[31] F.M. Malvestuto, M. Mezzini, Auditing sum attributes, in: Proc. IX Int. Conf. on "Database Theory", 2003, pp. 126–142.

[32] F.M. Malvestuto, M. Mezzini, M. Moscarini, Auditing sum-queries to make a statistical database secure, ACM Trans. Inf. Syst. Security 9 (2006) 31–60.

[33] F.M. Malvestuto, M. Moscarini, Minimal invariant sets in a vertex-weighted graph, Theoret. Comput. Sci. 362 (2006) 140–161.

[34] F.M. Malvestuto, M. Moscarini, Query evaluability in statistical databases, IEEE Trans. Knowl. Data Eng. 2 (1990) 425–430.

[35] F.M. Malvestuto, M. Moscarini, An audit expert for large statistical databases, in: Proc. Conf. on "Statistical Data Protection", EUROSTAT, 1999, pp. 29–43.

[36] F.M. Malvestuto, C. Zuffada, The classification problem with semantically heterogenous data, in: Proc. IV Int. Working Conf. on "Statistical & Scientific Database Management", in: M. Rafanelli, J.C. Klensin, P. Svensson (Eds.), Lecture Notes in Computer Sciences, vol. 339, Springer-Verlag, 1989, pp. 157–176.

[37] M. Mezzini, A linear time algorithm to find a nonempty algebraic subset of an edge set in graphs, J. Graph Algorithms Appl. 2006 (in press).

[38] I.S. Mumik, D. Quass, B.S. Mumik, Maintenance of data cubes and summary tables in a warehouse, in: Proc. ACM Int. Conf. on "Management of Data", 1997, pp. 110–111.

[39] W.K. Ng, C.V. Ravishankar, Information synthesis in statistical databases, in: Proc. IV Int. Conf. on "Information & Knowledge Management", 1995, pp. 355–361.

[40] T.B. Pedersen, C.S. Jensen, C.E. Dyreson, Extending practical pre-aggregation in on-line analytical processing, in: Proc. XXV Int onf. on "Very Large Data Bases", 1999.

[41] H. Sato, Handling summary information in a database: Derivability, in: Proc. ACM Int. Conf. on "Management of Data", 1981, pp. 98–107.

[42] H. Sato, Statistical data models: From a statistical table to a conceptual approach, in: Z. Michalewicz (Ed.), Statistical and Scientific Databases, in: ACM Int. Conf. on "Management of Data", 1981, Ellis Horwood, West Sussex, 1991, pp. 167–200.

[43] A. Schrijver, Theory of Linear and Integer Programming, Wiley, 1986.

[44] A. Shoshani, OLAP and statistical databases; similarities and differences, in: Proc. XVI ACM Symp. on Principles of Database Systems, 1997, pp. 185–196.

[45] D. Srivastava, S. Dar, H.V. Jagadish, A. Levy, Answering queries with aggregation using views, in: Proc. XXII Conf. on "Very Large Data Bases", 1996, pp. 318–329.

[46] J.D. Ullman, Principles of Database Systems, Computer Science Press, Rockville, MD, 1982.

[47] P. Vassiliadis, Modeling mutidimensional databases, cubes and cube operations, in: Proc. X Int. Conf. on "Scientific & Statistical Database Management", 1998, pp. 53–62.

[48] J. Widom, Research problems in data warehousing, in: Proc. IV Int. Conf. on "Information and Knowledge Management", 1995, pp. 25–30.

Please cite this article in press as: F.M. Malvestuto, et al., An analytical approach to the inference of summary data of additive type, Theoretical Computer Science (2007), doi:10.1016/j.tcs.2007.07.004