

# Exact Credal Treatment of Missing Data

Marco Zaffalon

IDSIA—Istituto Dalle Molle di Studi sull'Intelligenza Artificiale

Galleria 2, CH-6928 Manno, Switzerland

zaffalon@idsia.ch

## Abstract

This paper proposes an exact, no-assumptions approach to dealing with incomplete sets of multivariate categorical data. An incomplete data set is regarded as a finite collection of complete data sets, and a joint distribution is obtained from each of them, at a descriptive level. The tools to simultaneously treat all the possible joint distributions compatible with an incomplete set of data are given. In particular, a linear description of the set of distributions is formulated, and it is shown that the computation of bounds on the expectation of real-valued functions under such distributions is both possible and efficient, by means of linear programming. Specific algorithms are also developed whose complexity grows linearly in the number of observations. An analysis is then carried out to estimate population probabilities from incomplete multinomial samples. The descriptive tool extends in a straightforward way to the inferential problem by exploiting Walley's imprecise Dirichlet model.

*AMS Subject Classification:* Primary 62G05; secondary: 90C35.

*Keywords:* Incomplete data; Credal sets; Imprecise probabilities; Belief functions; Imprecise Dirichlet model; Linear optimization; Flow network

## 1 Introduction

This paper is largely motivated by the problem of learning discrete probability distributions from a database with missing data. Incomplete data are a major topic in statistical research and a pervasive problem of statistical practice. For example, many studies rely on large and complex surveys which contain missing values. To draw reliable conclusions from these surveys, the missing data must be handled properly.

One common approach to treating missing data from a statistical viewpoint is referred to as *imputation*. Imputation stands for the replacement of the missing data by plausible values. Schafer (1997) provides a comprehensive description of the state-of-the-art methods to deal with imputation in the multivariate case, while earlier work can be found in Little and Rubin (1987). Once the missing data are filled in, usual statistical procedures apply.

Imputation relies on the basic assumption that the missing-data mechanism is *ignorable*. For ignorability to hold, data must be missing at random, i.e. the probability that a value is missing must not depend on the other missing data. It is important to emphasize that whereas imputation techniques are quite sophisticated and generally accepted, the assumption that the data are missing at random is controversial, since it is clearly a strong assumption and it cannot be tested.

Apart from naive approaches that simply discard the observations with missing values, there do not appear to be alternatives to imputation if we restrict our attention to well-established methods. Some notable work is found in the framework of interval probabilities, e.g., work by Balke and Pearl (1994, 1997) on partially

related subjects, and in particular the recent work by Horowitz and Manski (1998, 2000) who show that, in many cases, missing data can be treated in a sound way without assuming ignorability. This approach aims to derive bounds on probabilities, or on quantities that depend on them, without assuming *anything* about the pattern of missing data. This is an *opposite* point of view to imputation: the approach does not try to reduce the missing data to the complete-data case, rather, it searches for all the possible consequences of the missing data. On one hand, this seems to be a basis for achieving consensus among researchers with different views; on the other hand, further assumptions, like the ‘missing completely at random’ assumption (a stronger assumption than missing-at-random), can be incorporated into the framework (Horowitz and Manski, 2000). In other words, this approach allows us to choose the strength of assumptions that are appropriate for a given problem, or to assume nothing at all. This is useful, for instance, in exploratory data analysis, where we often have many data and poor prior knowledge; under these conditions, it seems difficult to obtain reliable information about the missingness mechanism. The no-assumptions approach may also be useful since initially it is desirable to “let the (incomplete) data speak for themselves”; this also shortens the time of the analysis, which can be refined later when an estimate of the results is already available.

This paper proposes a treatment of missing data which is much in the spirit of Horowitz and Manski, with two main differences. Firstly, Horowitz and Manski work primarily at the population level. They tackle inference as a secondary matter, without discussing in depth the problem of computing estimates. In contrast, this paper focuses attention on the computation of descriptive quantities. Estimation is tackled at a secondary level in this paper too, by providing an inferential method that partially extends the descriptive development. Secondly, Horowitz and Manski impose some restrictions on the pattern of missing data. For instance, in Horowitz and Manski (2000), the values of different covariates in an

observation can only be either all unknown or all known. This allows us to obtain simple expressions for the bounds associated with missing data in special cases, but the analysis is not general. This paper, in contrast, places no restrictions on the pattern of missing data, and it obtains general results about computation.

Let us introduce the topic with an example. Suppose that the distribution of  $X$ , defined over  $\Omega_X = \{1, 2, 3\}$ , is obtained from data. That is, the distribution is the following vector of empirical probabilities (or probabilities, for short):  $[P[X = 1], P[X = 2], P[X = 3]]$ . Suppose that the observed data set is  $D = \langle 1, 1, 2, 3, 3, *, 1, 1 \rangle$ , where  $*$  denotes a missing value. If we know *nothing* about the missing data, any value in  $\Omega_X$  can replace the  $*$ , which is then called a *completely missing value*. Hence, three complete data sets originate from  $D$ ,  $D_1 = \langle 1, 1, 2, 3, 3, 1, 1, 1 \rangle$ ,  $D_2 = \langle 1, 1, 2, 3, 3, 2, 1, 1 \rangle$  and  $D_3 = \langle 1, 1, 2, 3, 3, 3, 1, 1 \rangle$ , and, correspondingly, three distributions for  $X$  are obtained,  $P_1 = [5/8, 1/8, 2/8]$ ,  $P_2 = [4/8, 2/8, 2/8]$  and  $P_3 = [4/8, 1/8, 3/8]$ . Since there is no reason to prefer one over another, we must deal with all of them together, as a set of *possible* distributions. Consider, for instance, the computation of  $P[X = 1]$ . Under the above conditions, all we can say is  $\underline{P}[X = 1] \leq P[X = 1] \leq \overline{P}[X = 1]$ , where  $\underline{P}[X = 1]$  and  $\overline{P}[X = 1]$  are called lower and upper probabilities, respectively, and are defined by

$$\underline{P}[X = 1] = \min_{i \in \{1, 2, 3\}} P_i[X = 1], \quad (1.1)$$

$$\overline{P}[X = 1] = \max_{i \in \{1, 2, 3\}} P_i[X = 1], \quad (1.2)$$

from which we have  $4/8 \leq P[X = 1] \leq 5/8$ . The procedure is similar for the other quantities.

This cautious approach is appealing because it does not require *any* assumption. At first sight, this benefit may seem to be overshadowed by the huge number of distributions that have to be dealt with, originating from the incomplete data set. If  $D$  contains 2 completely missing values,  $3^2$  complete data sets arise. In

general, if there are  $m$  completely missing values, the number of possible distributions is bounded by  $|\Omega_X|^m$ . However, as shown in this paper, such an exponential growth is not characteristic of the problem. The set of possible distributions can be synthesized by means of an exact linear formulation, whose size grows linearly with the number of observations for a fixed sample space. The linear description and the related optimization problems are the core of the paper. The ideas on which these results are based come from the field of *imprecise probabilities*.

Imprecise probabilities generalize probability theory by relaxing the assumption of precision. The requirement that probability values be precise, and therefore that there exists a single probability distribution, is one of the most controversial aspects of probability theory. In imprecise probability theory, uncertainty is modelled by a set of probability distributions called a *credal set* (Levi, 1980). Credal sets have great expressive power as they can represent a number of other models for uncertainty (e.g., in increasing order of generality: possibility measures, belief functions, Choquet capacities, coherent lower probabilities, coherent lower previsions), while providing a unifying view.

In this paper we basically work with credal sets at the level of belief functions (Shafer, 1976). In fact, an incomplete observation can be seen as a set-valued observation, i.e. as the set of all complete observations that are consistent with the incomplete observation (this is discussed in Section 2.1). Walley and Fine (1979, pp. 346–347) have shown that set-valued observations produce belief functions. This view is closely related to the multivalued mappings of Dempster (1967), and is well known in the field of belief functions (Chateauneuf and Jaffray, 1989; Jaffray, 1992, Section III.A). Nevertheless, to the author’s knowledge, the emphasis in the theory of belief functions is more on knowledge representation (Dubois et al., 1996) than on a frequentist interpretation. The statistical treatment of missing data does not appear to have benefited yet from belief function models, and the literature does not seem to provide a clear link with missing data in a statistical sense.

This seems to be true also for the literature on random sets (Molchanov, 1997; Stoyan, 1998) and on vague or fuzzy data (Kruse and Meyer, 1987), although some concepts expressed there are quite close to the statistical problem of incomplete data, e.g., the *imprecise observations* in Goodman et al. (1997, Section 7.1.1).

This paper aims to use credal sets in treating missing data, by showing that they allow a very general formulation to be realized but still permit exact and efficient computation. In the following, we adopt an optimization-based view that describes the problem in a natural way. (For that reason, we mostly speak in terms of credal sets rather than in terms of belief functions.) We are interested in the following optimization problems,

$$\min_{P \in \mathcal{P}} g(P), \quad (1.3)$$

$$\max_{P \in \mathcal{P}} g(P), \quad (1.4)$$

where the credal set  $\mathcal{P}$  is now regarded as the domain of the function  $g$  (i.e. the *feasible set* of the problem), which is also referred to as *objective function*:  $g$  is either a linear function or a ratio of linear functions (i.e. a *fractional linear function*) that describes how the quantity of interest is computed from the distribution  $P$ . This is quite a general view of the possible computations with credal sets. For instance, unconditional probabilities (and, more generally, expectations of real-valued functions under  $P$ ) are linear functions of  $P$ ; in this case the solutions of (1.3) and (1.4) are lower and upper probabilities. Fractional linear functions of  $P$  can be used to represent probabilities (and expectations) conditional on a given event.

The paper tackles problems (1.3) and (1.4) as follows. Section 2.1 considers a generic multivariate set of data of  $N$  units which may have missing values for some or all variables in each unit. The incomplete data set is regarded as a collection of complete data sets. Each of them gives rise to a joint distribution, so that we obtain a finite set of distributions which we call  $\tilde{\mathcal{P}}$ . Section 2.2 shows how the convex hull of  $\tilde{\mathcal{P}}$ , denoted by  $\mathcal{P}$ , can be described effectively by a linear formulation.

It is well known that the optimal solutions of (1.3) and (1.4) are equal to those obtained when  $g$  is defined over the feasible set  $\tilde{\mathcal{P}}$  (see Walley, 1991). Thus, the linear description of  $\mathcal{P}$  allows the sought optima to be efficiently computed by linear programming (Section 2.2). Section 3 investigates efficiency in more detail by developing fast specialized procedures for the above types of function  $g$ .

Finally, Section 4 regards the incomplete data set as a sample from a multinomial distribution. We show that Walley's *imprecise Dirichlet model* (1996) can be merged with the preceding framework in order to provide posterior lower and upper expectations of the unknown population probability of an event, and that our descriptive analysis can be extended to statistical inference in a straightforward way.

## 2 Credal treatment of missing values

### 2.1 Modelling incomplete sets of data

Let us first consider credal sets. We regard a credal set as the convex hull of a non-empty and finite family of discrete probability distributions. We recall that the convex hull is the set of all convex combinations of elements in the original set (hence a set is always included in its convex hull). In geometrical terms, a credal set is a *polytope*.

There are two equivalent ways to represent polytopes. The first way relies on *extreme points*. These are elements of the set that cannot be expressed as convex combinations of other elements. The finitely many extreme points are the *vertices* of the polytope. In the case of credal sets we also refer to them as *extreme distributions*. The set of extreme points is a synthetic representation of the polytope, which is the convex hull of this set.

The second way characterizes a credal set by means of linear constraints. In fact a polytope is a closed and bounded linear set, by definition. We can define

a credal set by linear constraints in different ways. For example, we can regard the probabilities of the elementary events as non-negative variables, and express the relationships among them by linear constraints (e.g., by expressing that a weighted sum of some probabilities is greater than or equal to a constant). These should include the constraint that the sum of all such variables is 1. More generally, the probabilities of the elementary events need not be explicitly represented by variables, but it must still be possible to express each of them as a linear function of some variables in the linear formulation. This paper is especially concerned with the linear-constraints view of credal sets and with the latter representation, as shown in Section 2.2.

Now we show how an incomplete data set gives rise to a set of distributions. Let  $D$  be a data set made up of a sequence of observations,  $\langle d_1, \dots, d_N \rangle$ ,  $N > 0$ . In a *complete* database, each unit is an instance of a set of  $k$  discrete variables; that is, the generic observation can be represented by the vector  $\mathbf{X} = (X_1, X_2, \dots, X_k) \in \Omega_{\mathbf{X}} = \times_{i=1}^k \Omega_{X_i}$ , where  $\Omega_{X_i}$  is a non-empty and finite set for all  $i$  in  $\{1, \dots, k\}$ . Let the symbol  $*$  denote a missing value; a unit from an *incomplete* database is an assignment of values to  $X_1, X_2, \dots, X_k$  from  $\times_{i=1}^k (\Omega_{X_i} \cup \{*\})$ .

An incomplete database can be transformed to a complete database by replacing each  $*$  with a value from the domain of the related variable. We use the word *completion* both to refer to this operation and also to refer to a particular complete database that originates from the incomplete one.

If we know nothing about the missing data, each missing element can take any value from the domain of the related variable. These are called *completely missing* values. Suppose that there are respectively  $n_1, n_2, \dots, n_k$  completely missing values for the variables  $X_1, X_2, \dots, X_k$  in  $D$ . The number of completions is  $\prod_{i=1}^k |\Omega_{X_i}|^{n_i}$ . This is also an upper bound for the number of different joint distributions  $P[X_1, X_2, \dots, X_k]$  that can be obtained from the completions of  $D$  (also referred to as distributions *compatible* with  $D$  or *originating* from  $D$ ).

In this paper we also allow the data to be *partially missing*. That is useful in modeling prior knowledge. We can think of two different types of partially missing values. A value is *partially missing of type 1* when the set of its replacements is a subset of the values of the related variable; a partially missing value of type 1 is said to be *partially missing of type 2* if its set of replacements is allowed to depend on the chosen replacements for other missing values in the same observation. In the first case, the prior knowledge disallows some values to be considered as possible replacements for the missing datum. The second case allows some types of dependency to be modeled (e.g., as when two missing values in the same observation must be replaced by the same value). This paper covers the two types of partially missing data.

We unify the view of completely and partially missing values as follows. We assume that there exists a map  $R : D \rightarrow 2^{\Omega_{\mathbf{x}}}$  such that  $R(d_j)$  is a non-empty set of possible replacements for unit  $d_j \in D$ . For example, consider the simple database given in Table 1, where  $\mathbf{X} = (X_1, X_2) \in \{1, 2, 3\} \times \{1, 2, 3\}$ . Since unit  $d_1$  is complete,  $R(d_1)$  is the observation  $\{(1, 1)\}$  itself; for the second unit,  $R(d_2)$  is generally a non-empty subset of  $\{(1, 1), (1, 2), (1, 3)\}$ . In the following, the map  $R$  is assumed to be given to allow possible knowledge about the missing data to be expressed. If there is no such knowledge,  $R$  should be taken as the particular map obtained by assuming that each  $*$  is a completely missing value.

\*\*\* TABLE 1 ABOUT HERE \*\*\*

Let us now refer to the data in Table 1 where we assume, for simplicity, that all the missing values are *completely* missing; the subsequent results also hold for the more general case of *partially* missing values, as they do not depend on the chosen map  $R$ . We depict the map  $R$  graphically in Figure 1. This represents the units of the database  $D$  as nodes in the leftmost column and the elements of  $\Omega_{\mathbf{x}}$  as nodes in the rightmost column. (Observe that  $D$  is simultaneously used to denote

a multivariate sequence of observations, the domain of the map  $R$ , and the set of leftmost nodes in the graph. We use the same symbol since these are just different representations of the same set of data. A similar remark applies to  $\Omega_{\mathbf{x}}$ .) For each  $d_j \in D$ , an arc leaves the node  $d_j$  and enters  $(x_1, x_2) \in \Omega_{\mathbf{x}}$  iff  $(x_1, x_2) \in R(d_j)$ . In other words, the arcs between the columns determine the map  $R$ . But observe that there are other arcs in the graph, which are external to the columns. This is because we want to regard the overall graph as a flow network, a well-known formalism in linear optimization theory. See Papadimitriou and Steiglitz (1982); Chateauneuf and Jaffray (1989) use a similar flow network to prove a proposition about Choquet capacities.

\*\*\* FIGURE 1 ABOUT HERE \*\*\*

More concretely, we regard the arcs as channels; a unitary flow is inserted into the network from each node in  $D$ , which is called a *source* node. All the flow inserted into the network reaches the rightmost nodes (called *sink* nodes) through the channels, following the arc directions. Notice that there are many different ways for the overall flow that enters the network to reach the sinks. We use a vector  $\mathbf{f}$  to represent a particular way to send the overall flow from sources to sinks. This is a vector of non-negative real numbers indexed by the arcs of the graph. Each element  $f(a, b)$  of the vector is the amount of flow on a generic arc  $a \rightarrow b$ , where  $a$  and  $b$  can be either nodes or the symbol ‘.’, which is used for the external arcs: we denote by  $f(\cdot, b)$  the flow into the source node  $b$  and by  $f(a, \cdot)$  the flow leaving the sink node  $a$ . The vector  $\mathbf{f}$  is called the *flow in the network*. A flow  $\mathbf{f}$  is said to be *integer* if all the elements of the vector are integers, otherwise it is said to be a *fractional* flow.

\*\*\* TABLE 2 ABOUT HERE \*\*\*

It is easy to see that there is a one-to-one correspondence between completions of  $D$  and integer flows in the network. A completion of  $d_j$ , for each  $d_j \in D$ , is

represented by redirecting its unitary flow as a whole to some element of  $R(d_j)$ . The converse is trivial to prove, by observing that integer flows are vectors whose elements are in  $\{0, 1\}$ . As an example, consider the completion given in Table 2. The observation  $d_2$  generates a unitary flow in the arc from node  $d_2$  to the right node  $(1, 2)$ , i.e.  $f(d_2, (1, 2)) = 1$ . The third observation is represented by a unitary flow from node  $d_3$  to the node  $(1, 3)$ , i.e.  $f(d_3, (1, 3)) = 1$ ; similarly for the others.

Now we show that the joint distribution from a given completion can be easily recovered from the related flow in the network. For a generic non-empty set of nodes  $S$ , let  $\mathbf{f}_S$  denote the subvector obtained from  $\mathbf{f}$  by only including the flows on the arcs leaving the nodes in  $S$  (we use the same notation for single nodes, too). For example, consider the node  $d_1$ ; since there is only one arc leaving it,  $f_{d_1}$  is just the flow on such an arc, i.e.  $f(d_1, (1, 1))$ . Consider the set of nodes  $\Omega_{\mathbf{X}}$ ;  $\mathbf{f}_{\Omega_{\mathbf{X}}}$  describes the configuration of the flow on the arcs that leave the nodes in  $\Omega_{\mathbf{X}}$ , that is,  $\mathbf{f}_{\Omega_{\mathbf{X}}} = [f((x_1, x_2), \cdot)]_{(x_1, x_2) \in \Omega_{\mathbf{X}}}$ .

The vector  $\frac{1}{N}\mathbf{f}_{\Omega_{\mathbf{X}}}$  is just the desired joint distribution. In fact, each element of the vector is the probability of an elementary event under the chosen completion. Notice that the formalism of the flow network does not directly represent the joint distribution, though this can be obtained by restricting the attention to the subvector  $\mathbf{f}_{\Omega_{\mathbf{X}}}$  and by normalizing it. With reference to Table 2, we have  $\mathbf{f}_{\Omega_{\mathbf{X}}} = [1, 1, 1, 0, 1, 0, 0, 2, 0]$ , where we consider the nodes in  $\Omega_{\mathbf{X}}$  in the sequence,  $(1, 1), (1, 2), \dots, (3, 3)$ , and therefore  $\frac{1}{N}\mathbf{f}_{\Omega_{\mathbf{X}}} = [1/6, 1/6, 1/6, 0, 1/6, 0, 0, 1/3, 0]$ .

## 2.2 Formal properties of the flow model

The example network in Figure 1 can be generalized in a straightforward way to allow any size  $N$  of the data set, any number of variables,  $X_1, X_2, \dots, X_k$ , as well as any map  $R$ . The present section addresses some formal properties of such a general type of flow network.

Let  $\tilde{\mathcal{P}}$  denote the finite set of joint distributions compatible with the incomplete data set  $D$ , and let  $\tilde{\mathcal{P}}_f$  denote the set of joint distributions produced by the integer flows in the network for  $D$ . The next lemma follows directly from the arguments in the preceding section.

**Lemma 2.1**  $\tilde{\mathcal{P}} = \tilde{\mathcal{P}}_f$ .

In order to derive other properties, we firstly note that the collection of feasible flows in the network is a linear set; that is, if we regard the vector  $\mathbf{f}$  as a point in a vector space, the set of all such points is a polytope, say  $\Psi$ . Such a polytope can be described by using linear constraints on the unknown vector  $\mathbf{f}$  (Papadimitriou and Steiglitz, 1982). Basically, the linear constraints express the *conservation of the flow*: for each node of the network, the total incoming flow must be equal to the total departing flow. For instance, with reference to node  $(2, 2)$  in Figure 1, we use the constraint  $f(d_4, (2, 2)) + f(d_6, (2, 2)) = f((2, 2), \cdot)$ . As well as the conservation of the flow, we must ensure the non-negativity of  $\mathbf{f}$  elementwise. More formally, the constraints are:

$$\sum_{\mathbf{x} \in R(d_i)} f(d_i, \mathbf{x}) = 1 \quad \forall d_i \in D, \quad (2.1)$$

$$\sum_{d_j \in D: \mathbf{x} \in R(d_j)} f(d_j, \mathbf{x}) = f(\mathbf{x}, \cdot) \quad \forall \mathbf{x} \in \Omega_{\mathbf{x}}, \quad (2.2)$$

$$\mathbf{f} \geq \mathbf{0}. \quad (2.3)$$

We do not require the elements of  $\mathbf{f}$  to be integers. Therefore, the set of feasible flows may contain fractional flows (for example, the unitary flow that enters  $d_2$  in Figure 1 is allowed to split into 3 flows of size  $1/3$  each, after leaving the node). Nevertheless, due to a property of flows in networks, called *total unimodularity* in Papadimitriou and Steiglitz (1982, pp. 316–318), the following lemma holds.

**Lemma 2.2** *The vertices (extreme points) of the polytope  $\Psi$  are integer flows.*

Now we can state the following theorem.

**Theorem 2.3** *let  $\mathcal{P}_f$  and  $\mathcal{P}$  respectively denote the convex hulls of  $\tilde{\mathcal{P}}_f$  and  $\tilde{\mathcal{P}}$ . Then  $\mathcal{P} = \mathcal{P}_f = \left\{ \frac{1}{N} \mathbf{f}_{\Omega_{\mathbf{x}}} \mid \mathbf{f} \in \Psi \right\}$ .*

**Proof.** Observe that  $\mathcal{P}_f = \mathcal{P}$  by Lemma 2.1. Let us focus on the set  $\left\{ \frac{1}{N} \mathbf{f}_{\Omega_{\mathbf{x}}} \mid \mathbf{f} \in \Psi \right\}$ ; this is the projection of  $\Psi$  onto a subspace, by definition of  $\mathbf{f}_{\Omega_{\mathbf{x}}}$ . For this reason it is a polytope and, by also taking Lemma 2.2 into account, its extreme points,  $ext \left\{ \frac{1}{N} \mathbf{f}_{\Omega_{\mathbf{x}}} \mid \mathbf{f} \in \Psi \right\}$ , are related to integer vectors  $\mathbf{f}_{\Omega_{\mathbf{x}}}$ .

Now notice that  $\tilde{\mathcal{P}}_f \subseteq \left\{ \frac{1}{N} \mathbf{f}_{\Omega_{\mathbf{x}}} \mid \mathbf{f} \in \Psi \right\}$  by construction. By taking the convex hull of both sets we have  $\mathcal{P}_f \subseteq \left\{ \frac{1}{N} \mathbf{f}_{\Omega_{\mathbf{x}}} \mid \mathbf{f} \in \Psi \right\}$ , since the set on the right side is already a polytope. Conversely, observe that  $ext \left\{ \frac{1}{N} \mathbf{f}_{\Omega_{\mathbf{x}}} \mid \mathbf{f} \in \Psi \right\} \subseteq \tilde{\mathcal{P}}_f$  because the elements of  $\tilde{\mathcal{P}}_f$  are all the distributions that are obtained by means of integer vectors  $\mathbf{f}_{\Omega_{\mathbf{x}}}$ . By taking the convex hull of both sides, we have  $\left\{ \frac{1}{N} \mathbf{f}_{\Omega_{\mathbf{x}}} \mid \mathbf{f} \in \Psi \right\} \subseteq \mathcal{P}_f$ .

■

By definition  $\mathcal{P}$  is a credal set, and it is the fundamental model to represent the uncertainty produced by the incomplete database (see Section 1). Theorem 2.3 is useful because it provides a description of  $\mathcal{P}$  in terms of the linear constraints (2.1)–(2.3). This gives an alternative representation, in addition to the one given by the extreme distributions. Of course, the two representations are formally equivalent, but they impact in quite different ways on the complexity of representation. In fact, the number of extreme points of a polytope is generally an exponential function of the size of its linear description in terms of variables and constraints. See Papadimitriou and Steiglitz (1982).

This observation also implies that algorithms based on the combinatorial approach of directly dealing with the extreme distributions have generally an exponential worst-case computational complexity. This is true, for instance, for the optimization of linear functions over a credal set. On the other hand, it is well known that the optimization of linear and fractional linear functions over a poly-

tope, like  $\mathcal{P}$ , is a polynomial task (Khachian, 1979). The linear representation of  $\mathcal{P}$  is the means to achieve such a polynomial worst-case complexity.

The preceding results also have an immediate practical implication. In order to describe  $\mathcal{P}$ , it suffices to write the flow network in terms of the constraints (2.1)–(2.3) and to restrict attention to the elements of  $\mathbf{f}_{\Omega_{\mathbf{x}}}$ . These are a subset of the optimization variables of the problem. By using linear or fractional linear functions of the elements of  $\mathbf{f}_{\Omega_{\mathbf{x}}}$ , we can solve problems (1.3) and (1.4) in a straightforward way by linear programming. We recall that fractional linear problems can be turned into linear problems by a result in Charnes and Cooper (1962), also reported in Schaible (1995, Section 2.2.2). In Section 3 we provide fast specialized algorithms that do not require linear programming to optimize the above functions.

### 2.3 A reduced flow network

In this section we investigate the possibility of representing  $\mathcal{P}$  by a reduced network so that the related linear formulation can be possibly expressed by fewer variables and constraints.

\*\*\* FIGURE 2 ABOUT HERE \*\*\*

Figure 2 shows a network equivalent to that in Figure 1. The network is obtained from that in Figure 1 by means of two simplifications: (i) it is unnecessary to use one source node for each  $d_j \in D$ , and it suffices to represent the distinct observations only; two units  $d_i, d_j \in D$  are said to be *distinct* if  $R(d_i) \neq R(d_j)$ . For example, the units  $d_2$  and  $d_3$  in Figure 1 are not distinct. They can be represented by a single node labeled  $(1, *)$ , with an incoming flow of size 2. In general, when a node in  $D$  represents  $m$  non-distinct observations, it becomes a source of a flow equal to  $m$ . (ii) It is also unnecessary to represent the nodes related to complete observations. For example, in Figure 1,  $d_1$  is unnecessary, because the flow passing

through it has no choice but to enter node  $(1, 1)$ . The only necessary nodes in  $D$  are the observations  $d_j$  such that  $|R(d_j)| > 1$ . Each unnecessary source node can be removed, and the arc from it to the sink node is turned into an external arc, which brings a fixed amount of flow to the right node.

The network obtained so far does equivalently represent  $\mathcal{P}$ , but it often has many fewer source nodes because there are only as many nodes as there are distinct incomplete observations in  $D$ . Since also the reduced network can be turned into a linear problem (as in Section 2.2), this problem has consequently a reduced size, being described by fewer variables and constraints. In the rest of the paper we do *not* use the reduced network given in this section, since the following results can be equivalently derived more simply from the original network.

### 3 Fast solution algorithms

The aim of the present section is to develop algorithms for special cases, in order to provide fast procedures to be used as alternatives to the general approach of linear programming. We express the computational complexity of the procedures by the notation  $O(\cdot)$  as in Graham et al. (1989); given the real-valued functions  $l, t : \mathbb{N}^+ \rightarrow \mathbb{R}$ , we write  $l(x) = O(t(x))$  if there exists a constant  $H$  such that  $|l(x)| \leq H|t(x)|$  for all  $x$ .

Both for linear objective functions and for ratios of linear functions representing probabilities under  $P \in \mathcal{P}$ , conditional on an event in  $\Omega_{\mathbf{X}}$ , the worst-case complexity of the exact optimizations is  $O(N|\Omega_{\mathbf{X}}|)$ . The optimum of the general fractional linear objective function can be approximated in time  $O(N|\Omega_{\mathbf{X}}|\lceil \log_2(L/\varepsilon) \rceil)$ , where  $\varepsilon$  is the required precision (i.e. the maximum absolute error),  $L$  is the length of an interval initially known to containing the optimum ( $L \geq \varepsilon$ ), and  $\lceil z \rceil$  denotes the smallest integer greater than or equal to a real number  $z$ .

### 3.1 Linear objective functions

We can represent probabilities and, more generally, expectations of real-valued functions under  $P \in \mathcal{P}$ , by letting the objective function of problems (1.3) and (1.4) be a linear combination of the flows in  $\mathbf{f}_{\Omega_{\mathbf{x}}}$ .

**Theorem 3.1** *Let  $g(\mathbf{f}_{\Omega_{\mathbf{x}}}) = \sum_{\mathbf{x} \in \Omega_{\mathbf{x}}} c_{\mathbf{x}} f_{\mathbf{x}}$ , where  $c_{\mathbf{x}} \in \mathbb{R}$  are given constants ( $\mathbf{x} \in \Omega_{\mathbf{x}}$ ). The optima in (1.3) and (1.4) are respectively  $\sum_{d_j \in D} \min_{\mathbf{x} \in R(d_j)} c_{\mathbf{x}}$  and  $\sum_{d_j \in D} \max_{\mathbf{x} \in R(d_j)} c_{\mathbf{x}}$ .*

**Proof.** The constant  $c_{\mathbf{x}}$  is regarded as a number associated with the node  $\mathbf{x}$  in the graph. Trivially, due to the structure of the graph, the local choices concerning where to direct the flow at the nodes  $d_j \in D$  have independent effects on the objective function. Therefore optimal local choices lead to the overall optima of the problems. ■

We can make the results of the Theorem 3.1 more explicit for the purpose of computing the lower and upper probabilities of an event.

**Corollary 3.2** *Let  $A$  be a generic event in  $\Omega_{\mathbf{x}}$ . Let  $\underline{D}_A = \{d_j \in D \mid R(d_j) \subseteq A\}$  and  $\overline{D}_A = \{d_j \in D \mid R(d_j) \cap A \neq \emptyset\}$ . Let  $\underline{n}(A)$  and  $\overline{n}(A)$  denote  $|\underline{D}_A|$  and  $|\overline{D}_A|$ , respectively. Then  $\underline{P}[A] = \underline{n}(A)/N$  and  $\overline{P}[A] = \overline{n}(A)/N$ .*

**Proof.** We represent  $A$  by letting  $c_{\mathbf{x}}$  be 1 for each  $\mathbf{x} \in A$  and  $c_{\mathbf{x}}$  be 0 otherwise. Let  $n(A)$  denote the number of occurrences of  $A$  for a generic completion of  $D$ , i.e. the number of source nodes in the graph which redirect their unitary flows to a node  $\mathbf{x}$  in  $A$ . Let  $\underline{n}(A)$  and  $\overline{n}(A)$  be the minimum and the maximum of  $n(A)$  obtained by taking into account all the completions of  $D$ . By Theorem 3.1 we have  $\underline{n}(A) = \sum_{d_j \in D} \min_{\mathbf{x} \in R(d_j)} c_{\mathbf{x}}$  and  $\overline{n}(A) = \sum_{d_j \in D} \max_{\mathbf{x} \in R(d_j)} c_{\mathbf{x}}$ . The corollary follows by observing that with the given constants, for any  $d_j \in D$ ,  $\min_{\mathbf{x} \in R(d_j)} c_{\mathbf{x}} = 1$  iff  $R(d_j) \subseteq A$  and, similarly,  $\max_{\mathbf{x} \in R(d_j)} c_{\mathbf{x}} = 1$  iff  $R(d_j) \cap A \neq \emptyset$ . ■

Let us investigate the worst-case complexity of an algorithm that implements Theorem 3.1. The algorithm takes as input the map  $R$  and the constants  $c_{\mathbf{x}}$  ( $\mathbf{x} \in \Omega_{\mathbf{x}}$ ), and outputs the desired optimum. The algorithm is defined below, where we use the notation  $\text{opt}_{\in \{\min, \max\}}$ :

1. Let *optimum* = 0; let  $j = 1$ .
2. Let *local\_optimum* =  $\text{opt}_{\mathbf{x} \in R(d_j)} c_{\mathbf{x}}$ .
3. Increase *optimum* by *local\_optimum*.
4. Increase  $j$  by 1; if  $j \leq N$  then go to step 2, else stop.

The complexity is determined by step 2 and by the loop in steps 2–4. Let us consider the computation of  $\text{opt}_{\mathbf{x} \in R(d_j)} c_{\mathbf{x}}$ . The optimum is computed by enumerating the constants  $c_{\mathbf{x}}$ ,  $\mathbf{x} \in R(d_j)$ . This takes time  $O(|R(d_j)|)$ . We bound such an expression by  $O(|\Omega_{\mathbf{x}}|)$ . Now consider that the loop is repeated  $N$  times. The overall complexity is thus  $O(N|\Omega_{\mathbf{x}}|)$ .

The term  $|\Omega_{\mathbf{x}}|$  of the complexity may look problematic for the efficiency of the algorithm, since it can be very large. However, the complexity is good because the term  $|\Omega_{\mathbf{x}}|$  originates from the input size. In fact both the worst-case size (number of arcs) of the network and the number of constants  $c_{\mathbf{x}}$  depend on  $|\Omega_{\mathbf{x}}|$ . Thus, the term  $|\Omega_{\mathbf{x}}|$  originates from the generality in the definition of the problem. But in practice the input size will not be very large, because it is unlikely that the user of the algorithm is interested in a problem that cannot be defined in a reasonable time. The user will generally define the input synthetically so that, for a large set  $\Omega_{\mathbf{x}}$ , such a structured input can be exploited both to avoid the explicit construction of the map  $R$  and to compute  $\text{opt}_{\mathbf{x} \in R(d_j)} c_{\mathbf{x}}$  without enumeration.

How this should be done depends on the specific computation required by the user (i.e., on the constants  $c_{\mathbf{x}}$  and on the type of missing data) and on the related definition of the input, and hence is not addressed in this paper. However,

providing a hierarchy of subcases seems an important issue for future work. An important branch of such a hierarchy is the problem of completely missing data. In this case the input structure can be exploited, to express the complexity in terms of the number  $k$  of variables, the number of missing values and the cardinality of the sets  $\Omega_{X_i}$  ( $i = 1 \dots k$ ).

### 3.2 Linear fractional objective functions representing conditional probabilities

This section shows that for the particular functions  $g$  representing  $P[A|B]$ ,  $A, B$  being generic events in  $\Omega_{\mathbf{X}}$ , it is still possible to obtain an exact  $O(N|\Omega_{\mathbf{X}}|)$  optimization algorithm. The argument that we use below is based on belief functions, because the proof follows more easily in this case as compared to the case of the flow model. Firstly, we need the following lemma.

**Lemma 3.3** *The lower probability function produced by  $\mathcal{P}$  is a belief function.*

**Proof.** Consider the function  $m : 2^{\Omega_{\mathbf{X}}} \rightarrow [0, 1]$  defined as  $m(B) = \sum_{d_j:R(d_j)=B} 1/N$  for each  $B \subseteq \Omega_{\mathbf{X}}$ . It is easy to see that  $m(\emptyset) = 0$  and  $\sum_{B \subseteq \Omega_{\mathbf{X}}} m(B) = 1$ . Let  $Bel : 2^{\Omega_{\mathbf{X}}} \rightarrow [0, 1]$  be defined by  $Bel(A) = \sum_{B \subseteq A} m(B)$  for each  $A \subseteq \Omega_{\mathbf{X}}$ .  $Bel$  is a belief function by definition (Shafer, 1976, p. 38). Observe that  $Bel(A) = \sum_{B \subseteq A} \sum_{d_j:R(d_j)=B} 1/N = \sum_{d_j:R(d_j) \subseteq A} 1/N$ , which is equal to  $\underline{P}[A]$ , as defined in Corollary 3.2. ■

Now we prove a theorem, where we use the notations introduced in Section 3.1.

**Theorem 3.4** *Let  $A, B$  be two generic events in  $\Omega_{\mathbf{X}}$  and let  $A^c$  denote the complement of  $A$ . Let  $\underline{P}[B]$  be positive. Then  $\underline{P}[A|B] = \underline{n}(A \cap B) / (\underline{n}(A \cap B) + \bar{n}(A^c \cap B))$  and  $\bar{P}[A|B] = \bar{n}(A \cap B) / (\bar{n}(A \cap B) + \underline{n}(A^c \cap B))$ .*

**Proof.** First, observe that we use assumption  $\underline{P}[B] > 0$  (whose validity can be checked by means of Theorem 3.1) because otherwise  $P[A|B]$  would not be defined for all the distributions in  $\mathcal{P}$ . Recall that  $\underline{P}$  is a belief function by Lemma 3.3. In this case, it is well known (Dempster, 1967, Eqs. 4.8) that  $\underline{P}[A|B] = \underline{P}[A \cap B] / (\underline{P}[A \cap B] + \overline{P}[A^c \cap B])$  and  $\overline{P}[A|B] = \overline{P}[A \cap B] / (\overline{P}[A \cap B] + \underline{P}[A^c \cap B])$ . The theorem follows by using Corollary 3.2 to compute the preceding upper and lower probabilities. ■

Let us apply Theorem 3.4 to  $P[X_1 = 1 | X_2 = 2]$ , with reference to the example in Table 1. We have  $\underline{n}(X_1 = 1 \cap X_2 = 2) = |\emptyset|$ ,  $\overline{n}(X_1 = 1 \cap X_2 = 2) = |\{d_2, d_3, d_4, d_6\}|$ ,  $\underline{n}(X_1 \neq 1 \cap X_2 = 2) = |\{d_5\}|$  and  $\overline{n}(X_1 \neq 1 \cap X_2 = 2) = |\{d_4, d_5, d_6\}|$ . It follows that  $\underline{P}[X_1 = 1 | X_2 = 2] = 0$  and  $\overline{P}[X_1 = 1 | X_2 = 2] = 4/5$ .

### 3.3 Linear fractional objective functions

This section deals with linear fractional functions, which are more general than those described in Section 3.2. Such generality is needed, for example, to compute the interval of expectations of real-valued functions under  $P \in \mathcal{P}$ , conditional on a given event in  $\Omega_{\mathbf{x}}$ .

**Theorem 3.5** *Let  $u(\mathbf{f}_{\Omega_{\mathbf{x}}}) = \sum_{\mathbf{x} \in \Omega_{\mathbf{x}}} c'_{\mathbf{x}} f_{\mathbf{x}}$  and  $v(\mathbf{f}_{\Omega_{\mathbf{x}}}) = \sum_{\mathbf{x} \in \Omega_{\mathbf{x}}} c''_{\mathbf{x}} f_{\mathbf{x}}$ , where  $c'_{\mathbf{x}}, c''_{\mathbf{x}} \in \mathbb{R}$  are given constants ( $\mathbf{x} \in \Omega_{\mathbf{x}}$ ). The function  $g$  is defined as  $u/v$ . The function  $v$  is required to be non-zero everywhere on the domain. Consider problems (1.3) and (1.4). Let  $L$  be the length of an interval initially known to contain the optimum (either the minimum or the maximum). Let  $\varepsilon \in \mathbb{R}^+$ ,  $\varepsilon \leq L$ . The required optimum can be computed in an approximate way, with precision  $\varepsilon$ , in time  $O(N |\Omega_{\mathbf{x}}| \lceil \log_2(L/\varepsilon) \rceil)$ .*

**Proof.** The proof follows from a well-known result in fractional programming, reported in Schaible (1995, Section 2.2.4). This requires that  $v$  is positive

in the domain; observe that we can always make it positive since  $v$  is already required to have constant sign, by continuity of  $v$  and assumption  $v(\mathbf{f}_{\Omega_{\mathbf{x}}}) \neq 0$ , and if it was negative we could multiply both  $u$  and  $v$  by  $-1$ . Let us consider the problem  $\text{opt}_{\mathbf{f} \in \Psi} g(\mathbf{f}_{\Omega_{\mathbf{x}}})$ , where  $\text{opt} \in \{\min, \max\}$ . Define the new function  $h(\gamma) = \text{opt}\{u(\mathbf{f}_{\Omega_{\mathbf{x}}}) - \gamma v(\mathbf{f}_{\Omega_{\mathbf{x}}}) : \mathbf{f} \in \Psi\}$ ,  $\gamma \in \mathbb{R}$ . Then  $h$  is continuous and strictly decreasing in  $\gamma$ , and  $h(\gamma) = 0$  iff  $\gamma = \gamma^*$ , where  $\gamma^*$  is the optimum value of  $g$ . We can approximate  $\gamma^*$  by a simple binary search that halves the interval containing the root of  $h$  at each stage. The number of evaluations of  $h(\gamma)$  is thus logarithmic in  $L/\varepsilon$ . Each evaluation is a linear problem that takes time  $O(N|\Omega_{\mathbf{x}}|)$  by Theorem 3.1; the overall procedure then works in time  $O(N|\Omega_{\mathbf{x}}| \lceil \log_2(L/\varepsilon) \rceil)$ . ■

Theorem 3.5 can be applied by the following procedure (see also Cozman, 2000):

1. Set  $i = 0$ . Compute an initial interval  $[\underline{g}_0, \bar{g}_0]$  that brackets the optimum value of  $g$ , so  $L = \bar{g}_0 - \underline{g}_0$ . This can be achieved by setting  $\underline{g}_0 = \min u(\mathbf{f}_{\Omega_{\mathbf{x}}}) / \max v(\mathbf{f}_{\Omega_{\mathbf{x}}})$  and  $\bar{g}_0 = \max u(\mathbf{f}_{\Omega_{\mathbf{x}}}) / \min v(\mathbf{f}_{\Omega_{\mathbf{x}}})$ , where the four optima are given by Theorem 3.1.
2. Let  $\gamma = (\underline{g}_i + \bar{g}_i) / 2$ . If  $\bar{g}_i - \underline{g}_i \leq \varepsilon$  then the procedure ends and the approximate value of  $\gamma^*$  is  $\gamma$ .
3. Compute  $h(\gamma)$  by means of Theorem 3.1.
4. Increase  $i$  by 1. If  $h(\gamma) > 0$  then let  $\underline{g}_i = \gamma$  and  $\bar{g}_i = \bar{g}_{i-1}$ , otherwise if  $h(\gamma) < 0$ , then let  $\underline{g}_i = \underline{g}_{i-1}$  and  $\bar{g}_i = \gamma$ . Whenever  $h(\gamma) = 0$ , let  $\underline{g}_i = \bar{g}_i = \gamma$ .
5. Go to step 2.

## 4 Estimating chances from incomplete samples

The preceding sections have shown that incomplete data sets have an intrinsic imprecision that can be properly described by credal sets. They have also provided tools to compute the intervals produced by such imprecision.

Now we consider a random sample from a multinomial distribution with parameters  $\theta_{\mathbf{x}}$  ( $\mathbf{x} \in \Omega_{\mathbf{x}}$ ). We can observe the sample only partially, as the incomplete data set  $D$ . For a generic event  $A \subseteq \Omega_{\mathbf{x}}$ , we denote by  $\theta_A$  the unknown chance of  $A$  under the multinomial distribution, i.e.  $\theta_A = \sum_{\mathbf{x} \in A} \theta_{\mathbf{x}}$ . We provide posterior lower and upper estimates of  $\theta_A$  by merging the results in Section 3.1 with the imprecise Dirichlet model (IDM) of Walley (1996).

The IDM models prior ignorance by a set of Dirichlet distributions and makes posterior inferences by combining this set with the observed likelihood function. The IDM is a well-founded model, with a number of important inferential properties (for example, inferences are independent of the definition of the sample space), and that allows us to naturally extend the preceding development while remaining in the field of imprecise probabilities. In the following we use the IDM in a way that is equivalent to considering a sample space of two events, i.e.  $A$  and  $A^c$ . This special case of the IDM (and with  $s = 1$ ; see below) was also proposed by Bernard (1996).

Let us consider a completion of  $D$ . By regarding the completion as a random sample, we obtain the following posterior lower and upper expectations of  $\theta_A$  by applying the IDM,

$$\frac{n(A)}{N + s} \tag{4.1}$$

and

$$\frac{n(A) + s}{N + s}, \tag{4.2}$$

where we use the notation  $n(A)$  introduced in Section 3.1. The hyperparameter  $s$

reflects the level of caution of the inferences; larger  $s$  gives greater caution. Walley provides arguments to choose  $s$  in  $[1, 2]$ .

Extending the IDM inferences to incomplete samples involves taking into account the imprecision due to the missing data. For this purpose, it suffices to consider the minimum of (4.1) and the maximum of (4.2), taken over all the completions of  $D$ . Notice that we can represent both  $n(A)/(N+s)$  and  $(n(A)+s)/(N+s)$  by linear functions  $g$ , in a similar way to that described in Corollary 3.2; so that the *exact* posterior lower and upper expectations of  $\theta_A$  inferred from the incomplete sample are respectively

$$\frac{\underline{n}(A)}{N+s} \tag{4.3}$$

and

$$\frac{\bar{n}(A)+s}{N+s}. \tag{4.4}$$

It is interesting to notice that we could have obtained the same result without explicitly referring to the IDM. It would be enough to augment the flow network for  $D$  with a new source node, with an incoming flow equal to  $s$ , and connected to all the sink nodes. On this basis, Corollary 3.2 would directly allow the desired lower and upper probabilities to be computed. It needs to be investigated whether this fact hold more generally, so that it could be exploited to extend the results in this section to other inferences.

## 5 Conclusions

The problem of incomplete data is an important one in statistics, and it cannot be neglected if one aims at drawing reliable conclusions from real data. The problem is usually tackled by using imputation techniques, but such an approach requires strong and untestable assumptions and it is often computationally demanding.

We believe that these difficulties originate from the conventional viewpoint on missing data, based on probability theory, which always models uncertainty by a single distribution. But when nothing is known about the missingness mechanism, as is often the case, incomplete data naturally give rise to a set of distributions, as shown in this paper. Instead of trying to reduce the set of distributions to a single one, recognizing the existence of the whole set is a first step in the direction of developing a more realistic basis for incomplete data.

The importance of choosing the proper framework to handle incomplete data is highlighted in this paper by the ease with which the formal results were obtained and by their generality. Moreover, the low computational complexity of the given procedures strongly supports the present approach.

The basic ideas of this paper seem general enough to be extended in several directions. We can identify two important lines of research, with the different purposes of specializing the results and of generalizing them. The first case is already described in Section 3.1. This aims to develop specializations of the algorithm, for linear objective functions and particular input structures. The second research direction aims to extend the methods of this paper to include additional knowledge about the missingness mechanism, whenever it is available. This would reduce imprecision because the credal sets involved would generally become smaller. Horowitz and Manski (2000) present an example of this type by showing how to treat the missing-completely-at-random assumption in their framework. Alternatively, we could try to incorporate knowledge that the missingness mechanism causes dependencies between different units (e.g., in such a way that the replacement for the missing value of a variable must be the same across different units), as a generalization of the dependencies modeled by partially missing values of type 2. In this case, more general networks of flow than those presented here might be necessary.

Furthermore, combining the imprecise Dirichlet model with our approach (Sec-

tion 4) seems a promising direction in which to develop methods of statistical inference. This might allow the properties of Walley's inferential model to be fully extended to the treatment of incomplete samples.

## Acknowledgements

I am grateful to Peter Walley, who provided me with many valuable comments and with relevant references concerning belief functions, random sets and vague data. He also helped to correct my English. I am grateful to Jean-Marc Bernard for his very detailed and helpful review, and to two anonymous referees for insightful comments. Thanks to Luca Maria Gambardella and Carlo Lepori for their kind attention and support. Thanks to Enrico Fagioli, Ivo Kwee and Fabio Stella for comments on an earlier version of the paper. This work was supported in part by SUPSI DIE, <http://www.supsi.ch>, under CTI grant # KTI 4217.1.

## References

- Balke, A., Pearl, J., 1994. Counterfactual probabilities: computational methods, bounds and applications. In: Lopez de Mantaras, R., Poole, D. (Eds.), *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Francisco, CA, pp. 46–54.
- Balke, A., Pearl, J., 1997. Bounds on treatment effects from studies with imperfect compliance. *J. Amer. Statist. Assoc.* **92**, 1171–1176.
- Bernard, J.-M., 1996. Bayesian interpretation of frequentist procedures for a Bernoulli process. *Amer. Statist.* **50**, 7–13.
- Charnes, A., Cooper, W.W., 1962. Programming with linear fractional functionals. *Naval Res. Logist. Quarterly* **9**, 181–186.
- Chateauneuf, A., Jaffray, J.-Y., 1989. Some characterizations of lower probabilities

- and other monotone capacities through the use of Möbius inversion. *Math. Social Sci.* **17**, 263–283.
- Cozman, F.G., 2000. Computing posterior upper expectations. *Internat. J. Approx. Reason.* **24**, 191–205.
- Dempster, A.P., 1967. Upper and lower probabilities induced by a multiple-valued mapping. *Ann. Math. Stat.* **38**, 325–339.
- Dubois, D., Prade, H., Smets, P., 1996. Representing partial ignorance. *IEEE Trans. Systems, Man and Cybernetics* **26**, 361–377.
- Goodman, I.R., Mahler, P., Nguyen, H.T., 1997. *Mathematics of Data Fusion*. Kluwer, Dordrecht.
- Graham, R.L., Knuth, D.E., Patashnik, O., 1989. *Concrete Mathematics: a Foundation for Computer Science*. Addison-Wesley, Reading, MA.
- Horowitz, J.L., Manski, C.F., 1998. Censoring of outcomes and regressors due to survey nonresponse: identification and estimation using weights and imputations. *J. Econometrics* **84**, 37–58.
- Horowitz, J.L., Manski, C.F., 2000. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *J. Amer. Statist. Assoc.*, to appear.
- Jaffray, J.-Y., 1992. Bayesian updating and belief functions. *IEEE Trans. Systems, Man and Cybernetics* **22**, 1144–1152.
- Khachian, L.G., 1979. A polynomial algorithm for linear programming. *Doklady Akad. Nauk. USSR* **244**, 1093–1096. Translated in *Soviet Math. Doklady* **20**, 191–194.
- Kruse, R., Meyer, K.D., 1987. *Statistics with Vague Data*. Reidel, Dordrecht.
- Levi, I., 1980. *The Enterprise of Knowledge*. MIT Press, London.
- Little, R.J.A., Rubin, D.B., 1987. *Statistical Analysis with Missing Data*. Wiley, New York.
- Molchanov, I., 1997. Statistical problems for random sets. In: Goutsias, J.,

- Mahler, R.P.S., Nguyen, H.T. (Eds.), *Random Sets: Theory and Applications*. Springer-Verlag, New York, pp. 27–45.
- Papadimitriou, H., Steiglitz, K., 1982. *Combinatorial Optimization: Algorithms and Complexity*. Prentice Hall, New York.
- Schafer, J.L., 1997. *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London.
- Schaible, S., 1995. Fractional programming. In: Horst, R., Pardalos, P.M. (Eds.), *Handbook of Global Optimization*. Kluwer, Dordrecht, pp. 495–608.
- Shafer, G., 1976. *A Mathematical Theory of Evidence*. Princeton University Press, Princeton.
- Stoyan, D., 1998. Random sets: models and statistics. *Internat. Statist. Review* **66**, 1–27.
- Walley, P., 1991. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York.
- Walley, P., 1996. Inferences from multinomial data: learning about a bag of marbles. *J. Roy. Statist. Soc. Ser. B* **58**, 3–57.
- Walley, P., Fine, T.L., 1979. Varieties of modal (classificatory) and comparative probability. *Synthese* **41**, 321–374.

Table 1: A simple example of an incomplete data set.

Observation	$X_1$	$X_2$
$d_1$	1	1
$d_2$	1	*
$d_3$	1	*
$d_4$	*	2
$d_5$	3	2
$d_6$	*	*

Table 2: A completion of the data set in Table 1.

Observation	$X_1$	$X_2$
$d_1$	1	1
$d_2$	1	2
$d_3$	1	3
$d_4$	2	2
$d_5$	3	2
$d_6$	3	2

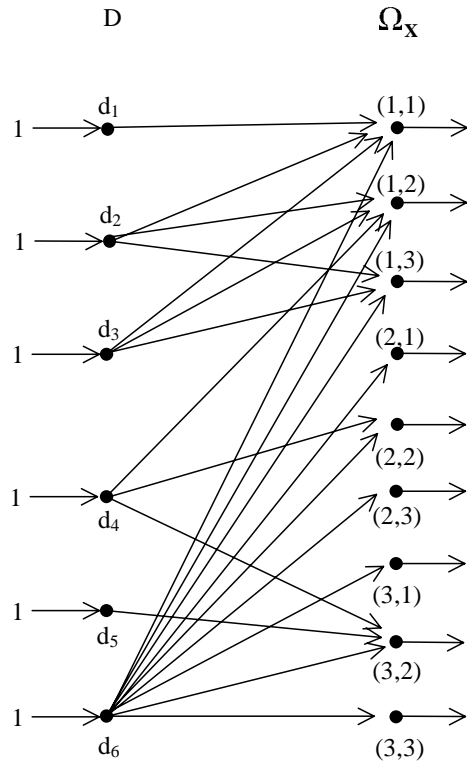


Figure 1: The flow network for the example of Table 1, where the missing data are regarded as completely missing.

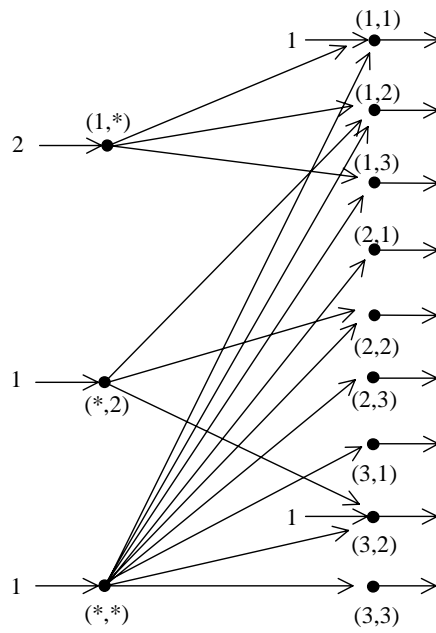


Figure 2: The reduced flow network for the example of Table 1.