Conservative Rules for Predictive Inference with Incomplete Data

Marco Zaffalon
IDSIA
Galleria 2, CH-6928 Manno (Lugano)
Switzerland
zaffalon@idsia.ch

Abstract

This paper addresses the following question: how should we update our beliefs after observing some incomplete data, in order to make credible predictions about new, and possibly incomplete, data? There may be several answers to this question according to the model of the process that creates the incompleteness. This paper develops a rigorous modelling framework that makes it clear the conditions that justify the different answers; and, on this basis, it derives a new conditioning rule for predictive inference to be used in a wide range of states of knowledge about the incompleteness process, including near-ignorance, which, surprisingly, does not seem to have received attention so far. Such a case is instead particularly important, as modelling incompleteness processes can be highly impractical, and because there are limitations to statistical inference with incomplete data: it is generally not possible to learn how incompleteness processes work by using the available data; and it may not be possible, as the paper shows, to measure empirically the quality of the predictions. Yet, these depend heavily on the assumptions made.

Keywords. Predictive inference, statistical inference, incomplete data, missing data, conservative inference rule, imprecise probability, conditioning, classification, data mining.

1 Introduction

Suppose you are given a multivariate set made of $N + 1$ units of categorical data. Each unit is categorized according to some *attribute variables* and a *class variable*. The data set is incomplete, in the sense that there are missing values, or, more generally, the data set can be regarded as a collection of complete data sets. You want to predict the (probability of the) class of the $(N + 1)$-th unit in the data set, such a class being missing. How can you do it?

There are two basic components in the sketched problem: a process that produces complete data, which are not observable, and another that turns complete into incomplete, and observable, data, namely, an *incompleteness process*[1] (IP). Statistically speaking, these are fundamentally different processes: the data may help us learn about the former, but they generally do not help with respect to the latter. Therefore, we may not be able to test our assumptions about the IP, but nevertheless our inferences rely heavily on the specific assumptions that we make about it. This suggests that we should model the IP carefully, by discussing in particular the assumptions underlying each approach.

The most popular approach to incomplete data in the literature and in the statistical practice is based on the so-called *coarsening-at-random* assumption (or CAR [3]); this is called *missing at random* (or MAR [5]) in the special case of missing data. Consider the latter case, for the sake of explanation. MAR allows missing data to be treated in a relatively simple way: they can be neglected or replaced by specific values, thus turning the incomplete data problem into one of complete data. But CAR/MAR embodies the idea that the incompleteness process is not *selective*, which is something widely recognized to definitely narrow the scope of such an assumption in applications (see in particular [4] on this point). Other approaches are also used: some treat the symbol of missing value as another possible value; imprecise-probability minded methods often regard incomplete data as set-based data, yielding set-based predictions and lower and upper probabilities and expectations [6, 8, 12]. Criteria to select one approach among those mentioned do not always appear to be clear. The situation is further complicated by the fact that the several methods proposed may treat the $(N+1)$-th unit of the data set is a different way from the rest. The focus of the methods is typically on the first N units, the so-called *learning set*, while the possible incompleteness of the $(N + 1)$-th unit (the so-called *unit to classify*, in terms of *pattern classification*) due to missing attribute values, is often dealt with the traditional conditioning rule that simply neglects

[1] This paper uses the word 'process' rather than the more common 'mechanism', as the latter seems to be misleading: it is often humans who create the incompleteness, or other complex entities that are not *mechanisms* in a proper sense.

such missing values after the conditioning bar.

This *state of affairs* is particularly unfortunate as incompleteness of data is pervasive of real applications, and improper treatments of incomplete data may yield highly misleading conclusions, defeating the efforts taken in modelling the process generating complete data.

This paper is an attempt to provide a rigorous approach to incomplete data, while providing a general, credible, and flexible tool that generalizes predictive inference to such an important case. It does so by following a top-down approach. Initially, Section 2 states assumptions that formalize the overall process producing the data in a very general way. The assumptions logically lead to a predictive inference rule for incomplete data, called *conservative inference rule* (CIR). This can be regarded as a new type of conditioning, which generalizes both the traditional conditioning rule and the *conservative updating rule* [2]. The latter has been recently proposed to address the case of near-ignorance about the IP in the context of expert systems. CIR generalizes the conservative updating rule firstly to statistical inference, which involves learning a model from data, rather than being given a model; and also to more general states of knowledge about the IP, including near-ignorance, but also up to knowing that the IP is not selective. Surprisingly, CIR seems to be the first proposal to address the near-ignorance case in the statistical setting, despite its theoretical and practical importance.

Section 3 strengthens the assumptions stated previously, by assuming that the process producing complete data is independently and identically distributed (IID) and that the IP is independently distributed (ID). The new assumptions are shown to yield CIR again. Section 5 discusses whether the IP should be assumed to be also identically distributed. While it shows that it is exactly the IID assumptions for the IP that provides a justification to treating the missing data symbol as another value, it argues that such an assumption is strong and hence not often tenable in practice.

Overall, the results show that the conservative inference rule is suited to be applied to a very wide range of situations. In achieving this, the paper also yields a framework that permits to clearly sort out the existing approaches proposed for incomplete data, according to the assumptions that underlie them. Two further conclusions of the paper are worth mentioning. First, the developed framework allows an old controversial question to be clarified in Section 4: how should we treat data that are completely unobserved? Second, Section 6 argues that in many real cases it is not possible to measure empirically the accuracy of the predictive inferences with incomplete data. This appears to be a further limitation of statistical inference with incomplete data that, again, suggests using care in modelling the IP.

2 Predictive Inference in a General Setting

This section considers a very general process that generates complete data, that is not assumed to be IID, and a very general model for the IP. This is regarded as a mix of two processes, one that is nearly unknown and a CAR one. The basic modelling units are introduced in Section 2.1. Section 2.2 states all the needed assumptions, which are discussed in Section 2.3, and re-worked in Section 2.4. Finally, Section 2.5 derives the conservative inference rule.

2.1 Sets and Variables

Consider a finite set $\mathcal{C}$, called set of *classes* and two finite sets, $\mathcal{X}$ and $\hat{\mathcal{X}}$, called sets of *attributes*. They are also referred to more generically as sets of *complete*, or *ideal*, observations. Let $\mathcal{D} := \mathcal{C} \times \mathcal{X}$ (the symbol ':=' denotes a definition). Also, consider a statistical random parameter θ taking generic value ϑ from the continuous set Θ.[2]

Consider the so-called *ideal variables* C_i, X_i, $D_i := (C_i, X_i)$, and $\hat{X}_i$, $i = 1, \ldots, N+1$, taking values in the sets $\mathcal{C}$, $\mathcal{X}$, $\mathcal{D}$, and $\hat{\mathcal{X}}$, respectively. Generic values of C_i, X_i, D_i, and $\hat{X}_i$, are denoted by c_i, x_i, d_i, and $\hat{x}_i$, respectively. The convention to denote generic values of variables and their domain by the corresponding small, resp. script capital, letters is maintained throughout the paper. Also, to simplify notation, we denote $M := N + 1$ and often group variables into vectors. For the moment, consider the following ones:[3] $\boldsymbol{C}^{\text{-}} := (C_1, \ldots, C_N)$, $\boldsymbol{C}^{+} := (C_1, \ldots, C_M)$, $\boldsymbol{X}^{\text{-}} := (X_1, \ldots, X_N)$, $\boldsymbol{X}^{+} := (X_1, \ldots, X_M)$, $\hat{\boldsymbol{X}}^{\text{-}} := (\hat{X}_1, \ldots, \hat{X}_N)$, $\hat{\boldsymbol{X}}^{+} := (\hat{X}_1, \ldots, \hat{X}_M)$. Let also $\boldsymbol{D}^{\text{-}} := (D_1, \ldots, D_N)$, $\boldsymbol{D}^{+} := (D_1, \ldots, D_M)$, $\boldsymbol{D} := (\boldsymbol{C}^{\text{-}}, \boldsymbol{X}^{+})$. Observe that $\boldsymbol{d}^{\text{-}} \in \boldsymbol{\mathcal{D}}^{\text{-}} \subseteq \mathcal{D}^N$, $\boldsymbol{d}^{+} \in \boldsymbol{\mathcal{D}}^{+} \subseteq \mathcal{D}^M$, $\boldsymbol{d} \in \boldsymbol{\mathcal{D}} \subseteq \mathcal{D}^N \times \mathcal{X}$, $\hat{\boldsymbol{x}}^{\text{-}} \in \hat{\boldsymbol{\mathcal{X}}}^{\text{-}} \subseteq \hat{\mathcal{X}}^N$, and $\hat{\boldsymbol{x}}^{+} \in \hat{\boldsymbol{\mathcal{X}}}^{+} \subseteq \hat{\mathcal{X}}^M$. The focus of our interest in this paper is on the random matrix

$$(\boldsymbol{D}^{+}, \hat{\boldsymbol{X}}^{+}) = \begin{bmatrix} C_1 & X_1 & \hat{X}_1 \\ C_2 & X_2 & \hat{X}_2 \\ \vdots & \vdots & \vdots \\ C_i & X_i & \hat{X}_i \\ \vdots & \vdots & \vdots \\ C_N & X_N & \hat{X}_N \\ C_M & X_M & \hat{X}_M \end{bmatrix}.$$

[2] Θ can be also countable. In this case the integrals in the following should be replaced by summations.

[3] Informally speaking, the vectors with empty superscript refer to the entire block of data that one usually has, i.e., the learning sample together with the attributes of the unit to classify; the superscript '-' refers to the entire block of data *minus* the attributes of the unit to classify (i.e., to the first N units, the so-called learning sample); and the superscript '+' refers to the entire block of data *plus* the class of the unit to classify (i.e., to the first $N + 1$ units).

Realizations of $(\boldsymbol{D}^{+}, \hat{\boldsymbol{X}}^{+})$ represent the possible complete data sets, which we cannot observe directly. Note that realizations of $(\boldsymbol{D}^{-}, \hat{\boldsymbol{X}}^{-})$ represent the possible complete learning sets, while those of $(X_M, \hat{X}_M)$ represent the possible complete observations to classify.

Now assume that, besides the ideal variables, there are *actual* variables. Actual variables represent observations of ideal variables (and so they will always be denoted using the letter O), in the sense that we can be informed about ideal variables only indirectly through actual variables. Each actual variable refers to a specific ideal variable. We will use frequently the following actual variables: $\boldsymbol{O}^{-}$, $\boldsymbol{O}^{+}$, $\boldsymbol{O}$, $\hat{\boldsymbol{O}}^{-}$, $\hat{\boldsymbol{O}}^{+}$, O_{C_M}, which denote the observations of $\boldsymbol{D}^{-}$, $\boldsymbol{D}^{+}$, $\boldsymbol{D}$, $\hat{\boldsymbol{X}}^{-}$, $\hat{\boldsymbol{X}}^{+}$, and C_M, respectively. Actual variables always take values in the powerset of the possibility space for the related ideal variables. With reference to the above variables, we have that $\boldsymbol{o}^{-} \in \mathfrak{O}^{-} \subseteq \wp(\mathfrak{D}^{-})$, $\boldsymbol{o}^{+} \in \mathfrak{O}^{+} \subseteq \wp(\mathfrak{D}^{+})$, $\boldsymbol{o} \in \mathfrak{O} \subseteq \wp(\mathfrak{D}^{-} \times \mathfrak{X})$, $\hat{\boldsymbol{o}}^{-} \in \hat{\mathfrak{O}}^{-} \subseteq \wp(\hat{\mathfrak{X}}^{-})$, and $\hat{\boldsymbol{o}}^{+} \in \hat{\mathfrak{O}}^{+} \subseteq \wp(\hat{\mathfrak{X}}^{+})$. The possibility spaces of the actual variables are called sets of *actual*, or *incomplete*, observations.

In other words, the realizations of actual variables model set-based observations. For example, the realizations of $(\boldsymbol{O}^{+}, \hat{\boldsymbol{O}}^{+})$ represent the possible incomplete data sets that we can observe, intended as sets of complete data sets, and those of $(\boldsymbol{O}^{-}, \hat{\boldsymbol{O}}^{-})$ represent the possible incomplete learning sets that we can observe.

It is important to understand from the very beginning that actual observations play a double role. For instance, $\boldsymbol{o}^{+}$ can be regarded as a set of observations of $\boldsymbol{D}^{+}$, but when $\boldsymbol{o}^{+}$ is interpreted as a value of $\boldsymbol{O}^{+}$, it is only a symbol of the alphabet $\mathfrak{O}^{+}$.

2.2 Measures and Assumptions

We assume that the joint density over the variables of interest is expressed by the following relation:

$$\begin{aligned} &p(\theta, \boldsymbol{D}^{+}, \hat{\boldsymbol{X}}^{+}, \boldsymbol{O}^{+}, \hat{\boldsymbol{O}}^{+}) = \\ &= p(\theta) p(\boldsymbol{D}^{+}, \hat{\boldsymbol{X}}^{+} | \theta) p(\boldsymbol{O}^{+} | \boldsymbol{D}^{+}) p(\hat{\boldsymbol{O}}^{+} | \hat{\boldsymbol{X}}^{+}). \end{aligned} \quad (1)$$

Here $p(\theta)$ is an imprecise prior density for θ, in the sense that $p(\theta)$ is known to belong to $\mathcal{P}(\theta)$, which is a non-empty set of precise prior densities for θ. [The assumptions in this section should be intended for all $p(\theta) \in \mathcal{P}(\theta)$, when it is the case.] $p(\boldsymbol{D}^{+}, \hat{\boldsymbol{X}}^{+} | \theta)$ is a *sampling model*. Together, they are required to satisfy a technical condition of positivity:

$$p(\boldsymbol{D}^{+}, \hat{\boldsymbol{X}}^{+}) > 0. \quad (2)$$

The conditional mass functions $p(\boldsymbol{O}^{+} | \boldsymbol{D}^{+})$ and $p(\hat{\boldsymbol{O}}^{+} | \hat{\boldsymbol{X}}^{+})$ represent independent incompleteness processes that act on different parts of the data. The first is called the *unknown IP*, and the second is the *CAR IP*, for reasons that will become clear shortly after. Note that the factorization of the density implies that the IPs depend on the complete data and do not depend on the random parameter θ, which is a way to express that the generation of actual data is made of two serial steps. A technical condition about the IPs is that the specific incomplete data observed can be practically observed:

$$p(\boldsymbol{o}^{+}, \hat{\boldsymbol{o}}^{+}) > 0. \quad (3)$$

More substantially, the IPs are assumed to be *truthful*, or *perfect*, in the following sense:

$$p(\boldsymbol{o}^{+} | \boldsymbol{d}^{+}) = 0 \text{ if } \boldsymbol{d}^{+} \notin \boldsymbol{o}^{+} \text{ and } p(\hat{\boldsymbol{o}}^{+} | \hat{\boldsymbol{x}}^{+}) = 0 \text{ if } \hat{\boldsymbol{x}}^{+} \notin \hat{\boldsymbol{o}}^{+}. \quad (4)$$

Note that this assumptions is basically what connects ideal with actual observations. The assumption may be perhaps made clearer by using a metaphor, in which the IPs are regarded as vision devices that may vary the focus on an (ideal) object that we want to see. When they produce singletons, they are perfectly in focus; when they produce the entire set of possible ideal observations, the devices are totally out of focus and we cannot see anything. All the intermediate states are also possible. But whatever their state, being truthful implies that they must point to the object of our interest, i.e., they cannot exclude it from our view.

The process $p(\hat{\boldsymbol{O}}^{+} | \hat{\boldsymbol{X}}^{+})$ is assumed, in addition, not to be *selective* (or *malicious*), in the sense that it acts as a CAR process on ideal observations:

$$p(\hat{\boldsymbol{o}}^{+} | \hat{\boldsymbol{x}}^{+}) = \alpha \quad \forall \hat{\boldsymbol{x}}^{+} \in \hat{\boldsymbol{o}}^{+}, \quad (5)$$

where α is a positive constant. This condition is the most frequently imposed on IPs in the literature.

Finally, consider two additional assumptions that are specific to predictive problems, and that concern the unknown IP. The first states that there is no way to observe directly the value of the variable to predict, i.e., C_M:

$$p(O_{C_M} = \mathcal{C}) = 1. \quad (6)$$

The second states that the incompleteness process $p(\boldsymbol{O}^{+} | \boldsymbol{D}^{+})$ does not directly depend on the variable to predict:

$$p(\boldsymbol{O}^{+} | \boldsymbol{D}^{+}) = p(\boldsymbol{O}^{+} | \boldsymbol{D}). \quad (7)$$

This is an important assumption for the following development. It embodies the idea that unknown IP generates incompleteness without knowledge of C_M.

The next section does some additional discussion about the important assumption (5) and (7). For the moment, note that nothing else is assumed about the IPs. This means, in particular, that we are nearly ignorant about $p(\boldsymbol{O}^{+} | \boldsymbol{D}^{+})$, on which we have stated weak assumptions only.

2.3 Assumptions Discussed

This section discusses more widely Assumptions (5) and (7). The other assumptions are very weak and relatively easy to accept. Assumption (4) may be an exception

as it makes sense to consider IPs that are imperfect, or that can lie, but this is out of the scope of the present paper.

Let us start with Assumption (5). This models a process that does not coarsen ideal data with a specific purpose. Assumption (5) excludes in this way many common and important processes. Consider the medical domain, for example, focusing on a diagnostic application. The ideal attribute variables in this case might describe information about a patient, such as gender, age, life style, and also the results of medical tests. The classes would represent the possible diseases. The IP in this case (at least part of it) is often originated by the interaction between the doctor and the patient themselves: indeed, there exists usually a systematic bias in reporting, and asking for, symptoms that are present instead of symptoms that are absent; and a bias to report, and ask for, urgent symptoms over the others [7]. Furthermore, a doctor typically prescribes only a subset of the possible diagnostic tests, according to personal views and cost/benefit criteria.[4] Overall, the process described is non-CAR by definition, because the incompleteness arises following patterns that do depend on the specific values of the ideal variables.

Descriptions such as the one above support the idea that CAR is strong by means of informal arguments. But also on the formal level, recent research has shown in a specific sense that CAR appears to be the exception rather than the rule in practice [4]. This should be taken into account in order to put CAR in a more balanced perspective, especially with regard to its very frequent use. This is not to say that CAR should be rejected a priori, as there are situations when CAR is completely justified. Consider one notable example: the case when we know that $\hat{X}_i$ is missing for any i. In this case the related IP clearly satisfies MAR: the probability of observing $\hat{\boldsymbol{O}}^+ = \hat{\mathcal{X}}^M$ is one, irrespectively of the value of $\hat{\boldsymbol{X}}^+$. More broadly speaking, MAR holds when some data are lost in an unintentional way. In these cases, not assuming MAR would lead to results much too weak. The CAR IP in the modelling framework presented in Section 2.2 is designed just to account for these situations. In other words, it is designed to provide one with some flexibility in stating the available knowledge about a domain, without having to adopt necessarily a worst-case approach (as opposed to the best case embodied by CAR/MAR), of the kind of the unknown IP.

Indeed, the unknown IP is designed so as to model one's ignorance about an incompleteness process. Let us first remark that it makes sense to adopt a conservative approach to model IPs in practice, for two specific reasons: first, IPs may be very difficult processes to model. They are a special case of observational processes, which are often originated by human-to-human interaction, or by other complex factors; they can actually be regarded as the result of a communication protocol. The medical example above is intended to illustrate just this, also if it is not saying anything new: the difficulty in modelling IPs has been pointed out strongly already long ago [11, 4]. But IPs are difficult objects to handle also for a second reason: IP models can be tightly dependent on specific situations. Re-consider the medical example: different doctors typically do different questions to diagnose a disease, even in the same hospital. By changing hospital, one can find entirely different procedures to diagnose the same disease, as the procedures depend on the local culture, on the money available to make the tests, and which ones, or the local time constraints. In other words, even if one is able to model an IP for a specific situation, that model may no longer be realistic when another doctor is in charge of doing the diagnosis, or when one tries to apply the IP model to another hospital, perhaps in another country. In summary, modelling the IP (especially in a precise way) may present serious practical difficulties, as, in contrast with domain knowledge (e.g., medical knowledge), the way information can be accessed may well depend on the particular environment where a system will be used; and this means that models of the IP may not be easily re-usable, and may therefore be costly.

These arguments support considering a conservative approach to model the IP that can be effectively implemented, and this is the reason for introducing the unknown IP in Section 2.2. As a side remark, note that the unknown IP is only *nearly* unknown, especially because we require that (7) holds. On the other hand, observe that by dropping it we could only draw vacuous conclusions about C_M. To see this, suppose that you want to predict the probability of $C_M = c_M$ given a certain incomplete observation $\boldsymbol{o}$ (assume, to make things easier, that there is no CAR IP). In these conditions, and without assuming (7), we could not exclude that the IP produces $\boldsymbol{o}$ if and only if $C_M \neq c_M$, so that the probability of $C_M = c_M$ is zero. In the same way, we could not exclude that the IP produces $\boldsymbol{o}$ if and only if $C_M = c_M$, so that the probability of $C_M = c_M$ is one. Of course also all the intermediate randomized cases would be possible, so that the probability would be always vacuous. This is to emphasize, perhaps not surprisingly, that complete ignorance about an IP is not consistent with the possibility to draw useful conclusions.

Having said this, it is still useful to wonder whether (7) is reasonable in the present setup. To this extent, let us re-write (7) in a somewhat more natural form thanks to Bayes' rule, Assumptions (2) and (6): $p(C_M|\boldsymbol{O},\boldsymbol{D})p(\boldsymbol{O}|\boldsymbol{D}) = p(C_M|\boldsymbol{D})p(\boldsymbol{O}|\boldsymbol{D})$. Focusing now only on pairs $(\boldsymbol{O},\boldsymbol{D})$ that are practically possible, Assumption (7) looks finally as follows: $p(C_M|\boldsymbol{O},\boldsymbol{D}) = p(C_M|\boldsymbol{D})$. This means to assume that when we know the

[4] This may perhaps clarify the reason why we will hardly get ever rid of incomplete data: often, incompleteness does not happen by mistake, rather, it is generated *deliberately*. In these cases it actually represents *patterns of knowledge* (indeed, one could say what disease a patient was, or was not, suspected to have by only looking at the medical tests that a doctor did, and did not, prescribed). In this sense, it seems that incompleteness is doomed to be deeply rooted to many, if not most, real problems; as such, it appears to be a fundamental, and indissoluble, component of uncertain reasoning, rather than a mere accident.

complete data $\boldsymbol{D}$, knowing in addition the incomplete data $\boldsymbol{O}$ that are produced from them, is completely superfluous to predict the class. In the latest form, in other words, Assumption (7) appears to be nothing else but the precise characterization of the problems of incomplete or missing data: these problems are characterized by the fact that when something that can be missing is actually measured, the problem of missing data disappears. It this were not the case, the observation $\boldsymbol{O}$ would not only carry information about C_M via its implications on $\boldsymbol{D}$, but it would say something about C_M also on its own. This does not look like a problem of missing, or incomplete, data, rather it appears to be a modelling issue: the actual observations that could be produced from $\boldsymbol{D}$ should be better modelled as further attributes that are relevant to predict the class, for instance by including the symbol of missing value in the set of ideal attributes, and treating it accordingly {see Reference [2] for further discussion about (7)}.

2.4 Assumptions Revisited

This section re-formulates some of the assumptions stated in Section 2.2, in a form that is more suitable for the subsequent derivations.

Lemma 1. *Assumptions* (1)– (4) *imply the following ones:*

$$p(\theta, \boldsymbol{D}^+, \hat{\boldsymbol{X}}^+, \boldsymbol{O}^+, \hat{\boldsymbol{O}}^+) =$$
$$p(\theta)p(\boldsymbol{D}^+, \hat{\boldsymbol{X}}^+|\theta)p(\boldsymbol{O}|\boldsymbol{D})p(\hat{\boldsymbol{O}}^+|\hat{\boldsymbol{X}}^+) \quad (8)$$
$$p(\boldsymbol{D}, \hat{\boldsymbol{X}}^+) > 0 \quad (9)$$
$$p(\boldsymbol{o}, \hat{\boldsymbol{o}}^+) > 0 \quad (10)$$
$$p(\boldsymbol{o}|\boldsymbol{d}) = 0 \text{ if } \boldsymbol{d} \notin \boldsymbol{o} \text{ and } p(\hat{\boldsymbol{o}}^+|\hat{\boldsymbol{x}}^+) = 0 \text{ if } \hat{\boldsymbol{x}}^+ \notin \hat{\boldsymbol{o}}^+. \quad (11)$$

Proof. First, consider $p(\boldsymbol{O}^+|\boldsymbol{D}^+)$. This is equal to $p(\boldsymbol{O}^+|\boldsymbol{D})$ thanks to (7). It follows by (6) that $p(\boldsymbol{O}^+|\boldsymbol{D}) = p(O_{C_M} = \mathcal{C}, \boldsymbol{O}|\boldsymbol{D}) = p(\boldsymbol{O}|\boldsymbol{D})$, whence $p(\boldsymbol{O}^+|\boldsymbol{D}^+) = p(\boldsymbol{O}|\boldsymbol{D})$. Whence (1) implies (8).
Assumption (9) is an immediate consequence of Assumption (2).
Now observe that $\boldsymbol{O}^+ = (o_{C_M}, o)$ if and only if $\boldsymbol{O} = \boldsymbol{o}$, again by (6). Thus $p(\boldsymbol{o}^+, \hat{\boldsymbol{o}}^+) = p(\boldsymbol{o}, \hat{\boldsymbol{o}}^+)$, and (10) follows from (3). Note that, also taking into account (5), Condition (10) implies that there exists $\boldsymbol{d} \in \boldsymbol{o}$ such that $p(\boldsymbol{o}|\boldsymbol{d}) > 0$.
Finally, consider $p(\boldsymbol{o}^+|\boldsymbol{d}^+)$. Write it as $p(o_{C_M}, \boldsymbol{o}|c_M, \boldsymbol{d})$. This is equal to $p(o_{C_M}, \boldsymbol{o}|\boldsymbol{d})$ by (7), and then to $p(\boldsymbol{o}|\boldsymbol{d})$ by (6). Note also that since $o_{C_M} = \mathcal{C}$, $(c_M, \boldsymbol{d}) \notin (o_{C_M}, \boldsymbol{o})$ if and only if $\boldsymbol{d} \notin \boldsymbol{o}$. Whence Assumption (4) implies (11). □

2.5 Derivation

In order to formulate the predictive problem in a sufficiently general way, let us focus on the problem of updating beliefs about a generic function $g : \mathcal{C} \to \mathbb{R}$, to posterior beliefs conditional on $(\boldsymbol{o}^+, \hat{\boldsymbol{o}}^+)$. If the prior density on θ and the incompleteness processes were precisely known, we would update beliefs on g by computing the following expectation:

$$E(g|\boldsymbol{o}^+, \hat{\boldsymbol{o}}^+) = E(g|\boldsymbol{o}, \hat{\boldsymbol{o}}^+) =$$
$$= \frac{\sum_{c_M \in \mathcal{C}} g(c_M) p(c_M, \boldsymbol{o}, \hat{\boldsymbol{o}}^+)}{\sum_{c_M \in \mathcal{C}} p(c_M, \boldsymbol{o}, \hat{\boldsymbol{o}}^+)}.$$

Note that $E(g|\boldsymbol{o}^+, \hat{\boldsymbol{o}}^+)$ is well defined as $p(\boldsymbol{o}, \hat{\boldsymbol{o}}^+) > 0$ by Assumption (10). The following lemma provides a more explicit form for $E(g|\boldsymbol{o}^+, \hat{\boldsymbol{o}}^+)$.

Lemma 2.

$$E(g|\boldsymbol{o}^+, \hat{\boldsymbol{o}}^+) =$$
$$\frac{\sum_{\boldsymbol{d} \in \boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d}) \sum_{c_M \in \mathcal{C}} g(c_M) p(c_M, \boldsymbol{d}, \hat{\boldsymbol{x}}^+ \in \hat{\boldsymbol{o}}^+)}{\sum_{\boldsymbol{d} \in \boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d}) p(\boldsymbol{d}, \hat{\boldsymbol{x}}^+ \in \hat{\boldsymbol{o}}^+)} =$$
$$= E(g|\boldsymbol{o}, \hat{\boldsymbol{x}}^+ \in \hat{\boldsymbol{o}}^+).$$

Proof. By focusing on $p(c_M, \boldsymbol{o}, \hat{\boldsymbol{o}}^+)$, it follows that

$$p(c_M, \boldsymbol{o}, \hat{\boldsymbol{o}}^+) =$$
$$\sum_{\boldsymbol{d} \in \boldsymbol{o}} \sum_{\hat{\boldsymbol{x}}^+ \in \hat{\boldsymbol{o}}^+} \int_\Theta p(\vartheta, c_M, \boldsymbol{d}, \hat{\boldsymbol{x}}^+, \boldsymbol{o}, \hat{\boldsymbol{o}}^+) d\vartheta =$$
$$\sum_{\boldsymbol{d} \in \boldsymbol{o}} \int_\Theta p(\vartheta) \sum_{\hat{\boldsymbol{x}}^+ \in \hat{\boldsymbol{o}}^+} p(c_M, \boldsymbol{d}, \hat{\boldsymbol{x}}^+|\vartheta) p(\boldsymbol{o}|\boldsymbol{d}) p(\hat{\boldsymbol{o}}^+|\hat{\boldsymbol{x}}^+) d\vartheta$$
$$= \alpha \sum_{\boldsymbol{d} \in \boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d}) p(c_M, \boldsymbol{d}, \hat{\boldsymbol{x}}^+ \in \hat{\boldsymbol{o}}^+),$$

where the sums are over $\boldsymbol{d} \in \boldsymbol{o}$ resp. $\hat{\boldsymbol{x}}^+ \in \hat{\boldsymbol{o}}^+$ thanks to (11), the second passage is due to (8), and the last to (5). The statement of the lemma is an immediate consequence of the expression derived. □

In order to extend the analysis done so far to the imprecise case, we need first to express our knowledge about $p(\boldsymbol{o}|\boldsymbol{d})$. This is done by Assumptions (10) and (11). These lead to the linear inequalities

$$\begin{cases} \sum_{\boldsymbol{d} \in \boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d}) > 0 \\ 0 \leq p(\boldsymbol{o}|\boldsymbol{d}) \leq 1, \ \boldsymbol{d} \in \boldsymbol{o}, \end{cases}$$

which define an open linear set called $\mathcal{P}(\boldsymbol{o}|\boldsymbol{D})$. The elements of this set are vectors, denoted by $p(\boldsymbol{o}|\boldsymbol{D})$, whose elements are $p(\boldsymbol{o}|\boldsymbol{d})$, $\boldsymbol{d} \in \boldsymbol{o}$. Note the restriction on $\boldsymbol{d} \in \boldsymbol{o}$, which is just the way to embody Assumption (11); also, note that the set $\mathcal{P}(\boldsymbol{o}|\boldsymbol{D})$ is not a singleton in general, as a logical consequence of our ignorance about the unknown IP. In the following we approximate $\mathcal{P}(\boldsymbol{o}|\boldsymbol{D})$ by the closed set $\mathcal{P}^\varepsilon(\boldsymbol{o}|\boldsymbol{D})$ defined by

$$\begin{cases} \sum_{\boldsymbol{d} \in \boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d}) \geq \varepsilon \\ 0 \leq p(\boldsymbol{o}|\boldsymbol{d}) \leq 1, \ \boldsymbol{d} \in \boldsymbol{o}, \end{cases}$$

for an ε in the real interval $(0, 1]$.
We are now in the conditions to write the objective of our

interest in the imprecise case, which is the following lower expectation:[5]

$$\underline{E}(g|\boldsymbol{o}^+, \hat{\boldsymbol{o}}^+) = \\ = \inf_{p(\theta)\in\mathcal{P}(\theta)} \inf_{p(\boldsymbol{o}|\boldsymbol{D})\in\mathcal{P}(\boldsymbol{o}|\boldsymbol{D})} E(g|\boldsymbol{o}, \hat{\boldsymbol{x}}^+ \in \hat{\boldsymbol{o}}^+).$$

In order to show how to make the computation of $\underline{E}(g|\boldsymbol{o}^+, \hat{\boldsymbol{o}}^+)$ operative, thus deriving the conservative inference rule, we need the following well-known result in *fractional programming* [9, Sec. 2.2.4].[6]

Theorem 3. *Consider the fractional optimization problem* $\min_{x\in S} q(x)/r(x)$*, where* q *and* r *are continuous, real-valued, functions on the compact set* S *of the* v*-dimensional Euclidean space* $\mathbb{R}^v$*, and* r *is positive on* S*. Let* $h : \mathbb{R} \to \mathbb{R}$ *be defined by* $h(\mu) = \min_{x\in S}[q(x) - \mu r(x)]$*. Then* $\mu^* \in \mathbb{R}$ *is the optimal solution of the initial fractional optimization problem if and only if* $h(\mu^*) = 0$*; and* μ^* *is the unique value such that* $h(\mu^*) = 0$*.*

We are finally ready to derive CIR.

Theorem 4 (Conservative inference rule theorem). $\underline{E}(g|\boldsymbol{o}^+, \hat{\boldsymbol{o}}^+) = \inf_{p(\theta)\in\mathcal{P}(\theta)} \min_{\boldsymbol{d}\in\boldsymbol{o}} E(g|\boldsymbol{d}, \hat{\boldsymbol{x}}^+ \in \hat{\boldsymbol{o}}^+)$.

Proof. Let us focus for the moment on the inner infimum in the definition of $\underline{E}(g|\boldsymbol{o}^+, \hat{\boldsymbol{o}}^+)$. That can be approximated by[7] $\min_{p(\boldsymbol{o}|\boldsymbol{D})\in\mathcal{P}^\varepsilon(\boldsymbol{o}|\boldsymbol{D})} E(g|\boldsymbol{o}, \hat{\boldsymbol{x}}^+ \in \hat{\boldsymbol{o}}^+)$. Consider the function

$$h(\mu) := \min_{p(\boldsymbol{o}|\boldsymbol{D})\in\mathcal{P}^\varepsilon(\boldsymbol{o}|\boldsymbol{D})} \{\sum_{\boldsymbol{d}\in\boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d}) \cdot \\ \cdot[\sum_{c_M\in\mathcal{C}} g(c_M)p(c_M, \boldsymbol{d}, \hat{\boldsymbol{x}}^+ \in \hat{\boldsymbol{o}}^+) - \mu p(\boldsymbol{d}, \hat{\boldsymbol{x}}^+ \in \hat{\boldsymbol{o}}^+)]\} \\ = \min_{p(\boldsymbol{o}|\boldsymbol{D})\in\mathcal{P}^\varepsilon(\boldsymbol{o}|\boldsymbol{D})} \{\sum_{\boldsymbol{d}\in\boldsymbol{o}} p(\boldsymbol{o}|\boldsymbol{d})p(\boldsymbol{d}, \hat{\boldsymbol{x}}^+ \in \hat{\boldsymbol{o}}^+) \cdot \\ \cdot[E(g|\boldsymbol{d}, \hat{\boldsymbol{x}}^+ \in \hat{\boldsymbol{o}}^+) - \mu]\},$$

where the last passage is possible as $p(\boldsymbol{d}, \hat{\boldsymbol{x}}^+ \in \hat{\boldsymbol{o}}^+) > 0$ by (9).

Theorem 3 shows that the unique solution of $h(\mu) = 0$ is also the solution of $\min_{p(\boldsymbol{o}|\boldsymbol{D})\in\mathcal{P}^\varepsilon(\boldsymbol{o}|\boldsymbol{D})} E(g|\boldsymbol{d}, \hat{\boldsymbol{x}}^+ \in \hat{\boldsymbol{o}}^+)$. Let $\mu^* := \min_{\boldsymbol{d}\in\boldsymbol{o}} E(g|\boldsymbol{d}, \hat{\boldsymbol{x}}^+ \in \hat{\boldsymbol{o}}^+)$, and let $\boldsymbol{d}^*$ be an element at which μ^* is attained. Observe that $h(\mu^*) \geq 0$ as all the involved factors are non-negative. Set $p(\boldsymbol{o}|\boldsymbol{d}^*) := 1$, and $p(\boldsymbol{o}|\boldsymbol{d}) := 0$ for all $\boldsymbol{d} \in \boldsymbol{o}$, $\boldsymbol{d} \neq \boldsymbol{d}^*$. Observe that the chosen vector $p(\boldsymbol{o}|\boldsymbol{D})$ belongs to $\mathcal{P}^\varepsilon(\boldsymbol{o}|\boldsymbol{D})$. This vector renders $h(\mu^*) = 0$, whence $\min_{p(\boldsymbol{o}|\boldsymbol{D})\in\mathcal{P}^\varepsilon(\boldsymbol{o}|\boldsymbol{D})} E(g|\boldsymbol{d}, \hat{\boldsymbol{x}}^+ \in \hat{\boldsymbol{o}}^+) = \min_{\boldsymbol{d}\in\boldsymbol{o}} E(g|\boldsymbol{d}, \hat{\boldsymbol{x}}^+ \in \hat{\boldsymbol{o}}^+)$. This result holds for all $\varepsilon \in (0, 1]$. Taking the limit with $\varepsilon \to 0$, we have $\inf_{p(\boldsymbol{o}|\boldsymbol{D})\in\mathcal{P}(\boldsymbol{o}|\boldsymbol{D})} E(g|\boldsymbol{d}, \hat{\boldsymbol{x}}^+ \in \hat{\boldsymbol{o}}^+) = \min_{\boldsymbol{d}\in\boldsymbol{o}} E(g|\boldsymbol{d}, \hat{\boldsymbol{x}}^+ \in \hat{\boldsymbol{o}}^+)$, from which the thesis follows immediately. □

Theorem 4 is an important result as it delivers a new inference rule to update beliefs given incomplete data. Before discussing the implications of CIR more widely, let us show that stronger assumptions than those considered here lead to CIR again.

3 Predictive Inference in an IID+ID Setting

The previous sections derived the conservative inference rule for very general processes of data generation. From now on we restrict the attention to IID processes that produce ideal data. We also enforce assumptions about the incompleteness processes, in particular by requiring that they act independently on each ideal observation generated, i.e., that they are ID. Let us show how the setup of Section 2 changes according to the new assumptions.

First, consider the variables. In the new setting, we can regard the statistical parameter θ, without loss of generality, as the random vector of chances for the elements $(d, \hat{x})$ of the sample space $\mathcal{D} \times \hat{\mathcal{X}}$. That is, a value $\vartheta \in \Theta$ is in this case a vector whose generic element is $\vartheta_{(d,\hat{x})}$, i.e., the aleatory probability (or *chance*) that $(d, \hat{x})$ is produced.[8] Regarding the variables previously defined, it is used the same notation, but consider that thanks to the IID+ID assumption the possibility spaces for the ideal variables are now the following: $\boldsymbol{\mathcal{D}}^- = \mathcal{D}^N$, $\boldsymbol{\mathcal{D}}^+ = \mathcal{D}^M$, $\boldsymbol{\mathcal{D}} = \mathcal{D}^N \times \mathcal{X}$, $\hat{\boldsymbol{\mathcal{X}}}^- = \hat{\mathcal{X}}^N$, and $\hat{\boldsymbol{\mathcal{X}}}^+ = \hat{\mathcal{X}}^M$. For the same reason, now the actual variable $\boldsymbol{O}^+$ can be regarded as the vector $(O_1^+, \ldots, O_M^+)$, whose generic element O_i^+ represents the observation of D_i. The situation is analogous with $\hat{\boldsymbol{O}}^+$.

The assumptions of Section 2.2 become more specific, according to the IID+ID assumptions. Using the IID assumption, the sampling model for ideal data becomes $p(\hat{\boldsymbol{d}}^+, \hat{\boldsymbol{x}}^+|\vartheta) = \prod_{i=1}^M p(d_i, \hat{x}_i|\vartheta) = \prod_{i=1}^M \vartheta_{(d_i,\hat{x}_i)}$. Regarding the incompleteness processes, the ID assumption is embodied by the equalities $p(\boldsymbol{o}^+|\boldsymbol{d}^+) = \prod_{i=1}^M p(o_i|d_i)$ and $p(\hat{\boldsymbol{o}}^+|\boldsymbol{x}^+) = \prod_{i=1}^M p(\hat{o}_i|\hat{x}_i)$. Using the two expressions above, the expression for the density of a joint value of the variables of interest becomes:[9] $p(\vartheta, \boldsymbol{d}^+, \hat{\boldsymbol{x}}^+, \boldsymbol{o}^+, \hat{\boldsymbol{o}}^+) = p(\vartheta) \prod_{i=1}^M \vartheta_{(d_i,\hat{x}_i)} p(o_i|d_i)p(\hat{o}_i|\hat{x}_i)$. The other assumptions are re-written as follows.

Positivity of ideal observations: $p(D_i, \hat{X}_i) > 0$, $i = 1, \ldots, M$. Positivity of the specific incomplete data observed: $p(o_i, \hat{o}_i) > 0$, $i = 1, \ldots, M$. Truthfulness: $p(o_i|d_i) = 0$ if $d_i \notin o_i$ and $p(\hat{o}_i|\hat{x}_i) = 0$ if $\hat{x}_i \notin \hat{o}_i$ $(i = 1, \ldots, M)$. CAR: $p(\hat{o}_i|\hat{x}_i) = \alpha_i$, $\hat{x}_i \in \hat{o}_i$

[5] The focus is on lower expectations as it is well known that upper expectations can be computed from lower expectations.

[6] This is the minimization version of the cited result.

[7] Remember that $\mathcal{P}^\varepsilon(\boldsymbol{o}|\boldsymbol{D})$ is closed, so that now the minimum is actually achieved.

[8] In other words, Θ is now assumed to be a (non-empty) subset of the unitary $|\mathcal{D} \times \hat{\mathcal{X}}|$-dimensional simplex.

[9] As in Section 2.2, the assumptions reported in this section should be intended for all $p(\theta) \in \mathcal{P}(\theta)$, when this is the case.

$(i = 1, \ldots, M)$, where the α_i's are positive constants. Finally, the two assumptions specific to predictive problems become the following: $p(O_{C_M} = \mathcal{C}) = 1$, and $p(O_M|D_M) = p(O_M|X_M)$.

As before, the basic problem is how to update beliefs about a generic function $g : \mathcal{C} \to \mathbb{R}$, to posterior beliefs conditional on $(\boldsymbol{o}^+, \hat{\boldsymbol{o}}^+) = (o_1, \ldots, o_M, \hat{o}_1, \ldots, \hat{o}_M)$. A course of reasoning very similar to that of Section 2.5 leads to the next theorem.

Theorem 5 (IID+ID CIR theorem). $\underline{E}(g|\boldsymbol{o}^+, \hat{\boldsymbol{o}}^+) = \inf_{p(\theta)\in\mathcal{P}(\theta)} \min_{\boldsymbol{d}\in\boldsymbol{o}} E(g|\boldsymbol{d}, \hat{\boldsymbol{x}}^+ \in \hat{\boldsymbol{o}}^+)$.

In other words, CIR should be the rule to adopt also in the new setting, which confirms the wide applicability of CIR. Let us discuss some other characteristics of CIR. The conservative inference rule generalizes the traditional rule based on the common conditioning, which is obtained by dropping the unknown IP. Furthermore, note that although CIR has been derived as a single rule, it can be regarded as a pair of rules: a *conservative learning rule*, and a *conservative updating rule*. The learning part prescribes how to learn a model from the observed data, which involves taking all the completions of the learning set subject to the unknown IP. The updating part prescribes how to use the model to predict a function of the next class, which involves taking all the completions of the M-th unit subject to the unknown IP. The method to implement both parts is then conceptually very simple. It is not, therefore, surprising that the use of similar procedures has already been advocated in the context of robust statistical methods (see for instance [6, 8, 12]). These proposals are nevertheless different from the method developed here, mostly because they focus only on delivering a conservative learning rule, while leaving the updating part an open issue; and because this paper has given emphasis to the rigorous justification of the rules.[10] In contrast with the mentioned approaches, a recent work [2], that is also the source of inspiration of the present paper, has derived a conservative updating rule in the context of probabilistic expert systems. Such a rule is generalized here in two directions.[11] First, CIR is a generalization to statistical inference, which involves learning, rather than being given, a model. Secondly, the previous rule deals with unknown IPs, as well as the mentioned robust statistical proposals basically do, rather than systematically treating a mix of unknown and CAR IPs, as CIR does. This is particularly useful, as explained in Section 2.3, in order to avoid too weak conclusions in specific cases. The variable balance between the unknown process and the CAR one seems actually to be a basic feature to enhance modelling flexibility. Reference [1, Sect. 4–5] shows this concretely on a real problem.

[10] Actually, Theorem 5 can be easily modified to justify also the approaches based on the conservative learning rule alone: it is sufficient to focus on parametric inference, i.e., on the lower expectation of a function of θ given $(\boldsymbol{o}^-, \hat{\boldsymbol{o}}^-)$, dropping the two assumptions specific to predictive problems. The proof remains basically the same.

[11] Strictly speaking, the generalization is really achieved: Theorem 4 can be adapted, with a minor syntactical re-styling, to derive the conservative updating rule in [2]. It is sufficient to remove all the references to θ and to the CAR IP. In other words, the result presented here can also be re-played in the context of expert systems.

4 A Note on Unobserved Data

Before discussing further assumptions about IPs, it may be useful to clarify a peculiar aspect that arises when part of the data is completely unobserved. An example is given by the formulation of CIR itself in Theorem 5: if $\hat{\boldsymbol{o}}^+$ coincides with $\hat{\mathcal{X}}^M$, CIR takes the following form: $\underline{E}(g|\boldsymbol{o}^+, \hat{\boldsymbol{o}}^+) = \inf_{p(\theta)\in\mathcal{P}(\theta)} \min_{\boldsymbol{d}\in\boldsymbol{o}} E(g|\boldsymbol{d})$. In other words, we are led to simply discard the variable $\hat{\boldsymbol{O}}^+$ from consideration. However, we are not allowed to do the same if $\boldsymbol{o}^+$ coincides with $\mathcal{D}^M$, as the formulation of CIR shows.

Something similar happens also in another situation, which we consider next. So far, we have implicitly assumed that the observations indexed from 1 to M constitute all the available data. Note that in practice this will not often be the case. More realistically, part of the available data will be discarded, and the predictive inference will be made only on the basis of a sub-sample. There is a variety of reasons why this may happen. For instance, we may start collecting data only when we start working on a problem; or we might want to use a sub-sample to learn a model as the entire set is too large and not easily manageable; or we might simply want to partition the data in learning and test set; and so on. We can represent such kinds of selection of a sub-sample in the framework of IPs, by saying that the probability not to observe the discarded data is one.

To be more specific, suppose that the 0-th observation was discarded irrespectively of its specific ideal value. We represent this by writing that $p(O_0 = \mathcal{O}, \hat{O}_0 = \hat{\mathcal{O}}|d_0, \hat{x}_0) = 1$ for all $d_0 \in \mathcal{D}$ and $\hat{x}_0 \in \hat{\mathcal{X}}$. Note that this is a CAR condition. From this, it follows immediately that also $p(O_0 = \mathcal{O}, \hat{O}_0 = \hat{\mathcal{O}}) = 1$. Let us see how this impacts on the derivation of Theorem 5. Such a theorem, like Theorem 4 (see Lemma 2 in particular), has been derived focusing on the probability $p(c_M, \boldsymbol{o}^+, \hat{\boldsymbol{o}}^+)$, i.e., on the joint probability of c_M and the observed data. Such a probability becomes $p(c_M, O_0 = \mathcal{O}, \hat{O}_0 = \hat{\mathcal{O}}, \boldsymbol{o}^+, \hat{\boldsymbol{o}}^+)$ once the 0-th observation is also taken into account. But we have that $p(c_M, O_0 = \mathcal{O}, \hat{O}_0 = \hat{\mathcal{O}}, \boldsymbol{o}^+, \hat{\boldsymbol{o}}^+) = \sum_{o_0\in\mathcal{O}, \hat{o}_0\in\hat{\mathcal{O}}} p(c_M, O_0 = o_0, \hat{O}_0 = \hat{o}_0, \boldsymbol{o}^+, \hat{\boldsymbol{o}}^+) = p(c_M, \boldsymbol{o}^+, \hat{\boldsymbol{o}}^+)$, thus stepping back to the setup leading to Theorem 5. In other words, it turns out that the 0-th observation can actually be removed from calculations as it does not influence the result. Of course this can immediately be extended to any number of observations that are discarded as the 0-th observation above. Overall, we obtain that data that are discarded in an unselective way, do not enter the derivation.

The situation is different if the data are discarded in a se-

lective way, that is, depending on their ideal values. If, for example, an ideal observation is not accessible any time it is equal to a specific value $(d, \hat{x})$, there is no justification to remove $(O_0 = \mathcal{O}, \hat{O}_0 = \hat{\mathcal{O}})$ from calculations. The correct way to proceed in this case, would simply be to make also $(O_0 = \mathcal{O}, \hat{O}_0 = \hat{\mathcal{O}})$ become part of the data used for learning, and then apply the results of Section 5. Of course this would create a phenomenon of *dilation* [10], but it does not seem possible to avoid it, given the stated conditions, without producing misleading results.

The distinction between the unselective and selective case above clarifies a question that is heard sometimes when considering conservative approaches to incomplete data such as CIR: what should we do of data that are completely unobserved? Should we discard them or should we include them in our set of actual data? The answer, as shown, depends on the nature of the incompleteness process.

5 Should We Assume an IP to Be IID?

The previous sections have shown that CIR is suited to be used in a wide variety of settings. This holds for a very general setting, as in Section 2, in which we assume very little about ideal and actual observations. But it holds also when more substantial assumptions are done, as in Section 3. In a sense, Sections 2 and 3 delimit CIR's boundaries of application. This is enforced by the arguments given in the present section. Indeed, if we make stronger assumptions than those in the preceding sections, CIR collapses to a more familiar rule.

For the sake of simplicity, let us restrict the attention here only to the unknown IP, and consider the following additional assumption about it, besides those stated in Section 3: that it is identically distributed, i.e., the unknown IP is now assumed to be IID. In other words, $p(o|d)$ is now a fixed chance, for all $o \in \mathcal{O}$ and $d \in \mathcal{D}$, which we denote by λ_{od}. It follows that each observation $o \in \mathcal{O}$ has also a fixed unconditional chance, called $\varphi_o := \sum_{d \in o} \theta_d \lambda_{od}$. Stated differently, we can regard the process producing elements $o \in \mathcal{O}$ also as a multinomial process with fixed (and perhaps unknown) chances.

Two observations are now in order. First, we note that the assumption that $p(O_{C_M} = \mathcal{C}) = 1$ is not compatible with the assumption that the incompleteness process is IID. Indeed, by assuming both, we would actually require $p(O_{C_i} = \mathcal{C}) = 1$ for any i, thus preventing any possibility to learn about the classes. Hence, we drop such an assumption and assume, to make things easier, that the class is always observed unambiguously, that is, $p(O_{C_i} = o_{C_i}) = 0$, for all i, if $|o_{C_i}| > 1$. The second observation concerns the assumption that $p(O_M|D_M) = p(O_M|X_M)$; we relax such an assumption, as it is not needed for the following developments.

As usual we focus on the problem of updating beliefs about a generic function $g : \mathcal{C} \to \mathbb{R}$, to posterior beliefs conditional on $\boldsymbol{o}^+$, i.e., on computing the following expectation:

$$E(g|\boldsymbol{o}^+) = \frac{\sum_{c_M \in \mathcal{C}} g(c_M) p(o_{C_M} = \{c_M\}, \boldsymbol{o})}{\sum_{c_M \in \mathcal{C}} p(o_{C_M} = \{c_M\}, \boldsymbol{o})}.$$

Consider the probability $p(o_{C_M} = \{c_M\}, \boldsymbol{o})$. Under the new assumptions, we can re-write it very easily as follows:

$$p(o_{C_M} = \{c_M\}, \boldsymbol{o}) = \int_\Phi p(\varphi) p(o_{C_M} = \{c_M\}, \boldsymbol{o}|\varphi) d\varphi$$

$$= \int_\Phi p(\varphi) \prod_{i=1}^{M} \varphi_{o_i} d\varphi. \tag{12}$$

Here $\varphi \in \Phi$ is the vector of chances φ_o, $o \in \mathcal{O}$; $\prod_{i=1}^{M} \varphi_{o_i}$ is the likelihood, as from the multinomial assumption about the data $o \in \mathcal{O}$; and $p(\varphi)$ is a prior density for φ. [The prior $p(\varphi)$ could in principle be obtained from $p(\vartheta)$ by means of a change of variables that takes into account the transformation $\varphi_o = \sum_{d \in o} \theta_d \lambda_{od}$.]

In other words, in this setting everything is confined within the common, precise, Bayesian approach to predictive inference, which is applied to the actual variables. In practice, the above premises allow us to regard o_i simply as an element of its sample space (i.e., a symbol of an alphabet), forgetting the fact that it can also be interpreted as a set of ideal observations. Note that in the special case of incomplete data generated via missing values, the above approach would correspond to simply treat the symbol of missing value as another possible value.

The discussion up to this point has shown that the current setting presents some advantages; it turns out to be a precise approach, and, by relaxing the assumption that $p(O_M|D_M) = p(O_M|X_M)$, it can consider more general relations between classes and attributes than the previous approaches.[12] Unfortunately, the advantages, which follow by assuming that the unknown IP is IID, seem to be dwarfed by the strength of the assumption itself. It is indeed questionable that such an assumption is tenable in practice. As illustrated in Section 2.3, IPs are often generated by complex behavior patterns that involve humans and that can be regarded as protocols of communications. Assuming that the IP is IID, would be equivalent, in the medical example of Section 2.3, to assuming that all the doctors, in all the hospitals, behave in the same way, implementing the very same procedure to diagnose a certain disease. In contrast, relaxing the assumption to the weaker ID, we would concede that the procedures can change with time (and location, and resources, ...). Of course, this does not mean that they need to change completely, and in this

[12] However, it may be useful to emphasize that the presented method may be inefficient when the considered IP is CAR besides being IID. In this case, a more opportune approach would incorporate CAR in the derivation, producing a more specific updating rule. Without doing this, the learning algorithm may need quite a number of data to realize by itself, and hence exploit, the implications of CAR.

sense the conservative approach embodied by the ID assumption may well leave (also much) room for strengthening the conclusions. But unless one has a clear model of the different procedures, the conservative approach appears to be more credible and safe than going for strong assumptions, of which the IID one is an example.
In other words, the ID assumption accounts for the variety and complexity of behavior of humans, or of other complex entities, that are involved in communication protocols giving rise to incompleteness. It may be the case that the IID assumption is justified in some specific cases, but it does not seem to be really widely applicable, and hence this work refrains from doing it any further.

6 On the Failure of Empirical Evaluations

Section 5 has argued that the IID assumption is strong, and that the weaker ID assumption is more reasonable in many applications. But, by accepting this, one needs also realize that currently employed methods face a very critical consequence that regards the possibility to measure empirically their predictive performance. This is a further reason that suggests doing only tenable assumptions.
In order to introduce the subject of this section, let us focus, for the sake of explanation, only on the unknown IP, and also on a very simple *classification* setting, which is a special case of predictive inference. The classification setting is one in which the ideal observations are generated in an IID way, and where the IP is ID. The problem is to predict the class of the M-th unit given the previous units $(1, \ldots, N)$ and the values of the M-th attribute variables. In the precise case, this is usually done following a maximum expected utility approach: i.e., the prediction is made up of a class that maximizes the expected utility, conditionally on the observation. Call a model that acts in such a way a *precise classifier*.
It is common practice with precise classifiers to measure their accuracy empirically. In its simpler form, this is obtained by randomly splitting the available data into a *learning* and a *test set*, by inferring the classifier from the former and testing it on the latter. Testing the classifier typically means to make it predict the classes for the units in the test set, and to compare them with the true classes. One popular summary measure used to this extent is the so-called *prediction accuracy*, namely the relative number of classes correctly predicted. This simple, yet powerful, idea is responsible for much of the success and popularity of classification, as it enables one to be relatively confident about how well a classifier performs on previously unseen data. Unfortunately, this key characteristic is lost when we cannot assume that the unknown IP is IID.
To see why, consider two ideal random variables A_1 and A_2, and a class variable C.[13] Assume that they are all Boolean, and that C is the result of the exclusive logical disjunction applied to A_1 and A_2: i.e., $C = 1$ if and only if either $A_1 = 1$ or $A_2 = 1$, but not both. Assume that the ideal data are eventually turned into actual data by an (unknown) IP whose only action is to make A_2 unobservable if and only if $(A_1, A_2) = (1, 0)$. That is, $p(O_{A_2} = \{0, 1\} | A_1 = 1, A_2 = 0) = 1$, so that observing the pattern $(O_{A_1} = \{1\}, O_{A_2} = \{0, 1\})$ implies $C = 1$ with certainty. In these conditions, any precise classifier is clearly expected to learn that A_2 being missing is irrelevant to predict the class, whose value coincides with the value of A_1. Since this is true for all the available data, partitioning them into learning and test set will do nothing but confirm it: the prediction accuracy on the pattern $(O_{A_1} = \{1\}, O_{A_2} = \{0, 1\})$ will be perfect, i.e., 100%. But the IP is not identically distributed, and it happens that when the classifier is put to work in an operative environment, the IP changes, in particular by making A_2 unobservable if and only if $(A_1, A_2) = (1, 1)$; or, in other words: $p(O_{A_2} = \{0, 1\} | A_1 = 1, A_2 = 1) = 1$. Once put to work in practice, the classifier will be always wrong on the pattern $(O_{A_1} = \{1\}, O_{A_2} = \{0, 1\})$, the prediction accuracy dropping to 0%.
Of course the example is designed so as to illustrate an extreme case, but this nevertheless, it points to a fact: empirical evaluations are doomed to failure in general when the data are made incomplete by a non-IID unknown process (remember that we already excluded it to be CAR, as we are talking of the unknown IP). This appears to have profound implications for classification, and more generally for data analysis. These fields of scientific research rest on two fundamental pillars: (i) that the assumptions made to develop a certain model (e.g., a classifier) are tenable; and (ii) that empirical evaluations are reliable. The crucial point is that both pillars may be very fragile with incomplete data, so being unable to sustain credible models and conclusions. Regarding (i), we have already argued that the IID assumption seems to be hard to justify for an incompleteness process; also, it does not seem to be possible to test statistically the IID assumption, for reasons similar to those that prevent CAR from being tested. Therefore assuming IID (as well as CAR) will often be subject to arbitrariness. Moreover, with reference to (ii), the results of empirical evaluations will not help us in general to have a clear view of the negative consequences of doing untenable assumptions, as empirical evaluations may well be misleading, as shown above.
The way left to cope with such critical issues seems necessarily to have to pass through doing tenable assumptions. This implies also to recognize that the unknown incompleteness process may not be IID (nor CAR). Most probably this will lead to some kind of conservative inference rule, such as CIR. Conservative inference rules, in turn, will lead to set-based predictions of classes, or, in other words, to so-called *credal classifiers* [13], which will produce credible results, consistently with the weaker

[13] Here the attributes are just the joint states of A_1 and A_2.

assumptions they require. For this reason, conservative inference rules, as well as credal classifiers, appear to be worthy of serious consideration in order to produce realistic models and credible conclusions.

7 Conclusions

The problem of incomplete data in statistical inference is an important one and is pervasive of statistical practice. It is also a difficult problem to handle. The processes that create the incompleteness may be difficult to model, and in presence of incompleteness we may be seriously limited in the possibility to understand whether our models and conclusions are good: in general, one cannot use the data to test assumptions about IPs; and, as this paper shows, incomplete data may prevent one also from empirically measuring the quality of the predictions. Yet, these heavily depend on the assumptions made about the underlying IPs. The crucial role of assumptions suggests then using only assumptions weak enough to be tenable.

A large part of this work has been actually spent in discussing the assumptions underlying different models and in selecting some of them, to the extent of designing a modelling framework that is widely applicable and credible at the same time. Such a framework has eventually led to derive a new conditioning rule for predictive inference with incomplete data, called conservative inference rule. CIR generalizes the traditional rules of inference, as well as a more recently proposed rule for conservative updating with probabilistic expert systems. More broadly speaking, CIR seems to be the first proposal for updating beliefs in a statistical setting when only weak assumptions about IPs are justified. Of course this capability is also such that CIR yields only to partially determined inferences and decisions, in general, and ultimately to systems that can recognize the limits of their knowledge, and suspend judgement when these limits are reached. But the new rule is not only a tool for worst-case scenarios; it can flexibly model knowledge from near-ignorance to knowing that an incompleteness process is not selective. Such a characteristic has the potential to make CIR suitable for a wide variety of settings. Of course it is important to investigate this point concretely, by verifying in particular whether CIR leads to strong enough conclusions in real applications. This is the focus of work that is currently under development.

Acknowledgements

Work partially supported by the Swiss NSF grant 2100-067961. Initial work on this paper was done when the author was visiting the University of Ghent (Belgium) in July 2003. Gert de Cooman and the Flemish Fund for Scientific Research (FWO, grant # G.0139.01) are gratefully acknowledged to have made that visit possible.

References

[1] A. Antonucci, A. Salvetti, and M. Zaffalon. Assessing debris flow hazard by credal nets. In M. López-Díaz, M. A. Gil, P. Grzegorzewski, O. Hryniewicz, and J. Lawry, editors, *Soft Methodology and Random Information Systems (Proceedings of the Second International Conference on Soft Methods in Probability and Statistics)*, pages 125–132. Springer, 2004.

[2] G. de Cooman and M. Zaffalon. Updating beliefs with incomplete observations. *Artificial Intelligence*, 159(1–2):75–125, 2004.

[3] R. Gill, M. Van der Laan, and J. Robins. Coarsening at random: characterisations, conjectures and counter-examples. In D.-Y. Lin, editor, *Proceedings of the first Seattle Conference on Biostatistics*, pages 255–294. Springer, 1997.

[4] P. Grünwald and J. Halpern. Updating probabilities. *Journal of Artificial Intelligence Research*, 19:243–278, 2003.

[5] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley, New York, 1987.

[6] C. F. Manski. *Partial Identification of Probability Distributions*. Springer-Verlag, New York, 2003.

[7] M. A. Peot and R. D. Shachter. Learning from what you don't observe. In G. F. Cooper and S. Moral, editors, *Uncertainty in Artificial Intelligence (Proceedings of the Fourteenth Conference)*, pages 439–446. Morgan Kaufmann Publishers, San Francisco, CA, 1998.

[8] M. Ramoni and P. Sebastiani. Robust learning with missing data. *Machine Learning*, 45(2):147–170, 2001.

[9] S. Schaible. Fractional programming. In R. Horst and P. M. Pardalos, editors, *Handbook of Global Optimization*, pages 495–608. Kluwer, The Netherlands, 1995.

[10] T. Seidenfeld and L. Wasserman. Dilation for sets of probabilities. *The Annals of Statistics*, 21:1139–54, 1993.

[11] G. Shafer. Conditional probability. *International Statistical Review*, 53:261–277, 1985.

[12] M. Zaffalon. Exact credal treatment of missing data. *Journal of Statistical Planning and Inference*, 105(1):105–122, 2002.

[13] M. Zaffalon. The naive credal classifier. *Journal of Statistical Planning and Inference*, 105(1):5–21, 2002.