International Journal of Approximate Reasoning 93 (2018) 238–260

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

A general framework for maximizing likelihood under incomplete data [☆], [☆☆]

Inés Couso [a,*], Didier Dubois [b]

[a] Dep. of Statistics and O.R., University of Oviedo, Spain
[b] IRIT, CNRS and Université de Toulouse, France

ARTICLE INFO

Article history:
Received 17 October 2016
Received in revised form 27 September 2017
Accepted 27 October 2017
Available online 7 November 2017

Keywords:
Random sets
Maximum likelihood
Incomplete information
Entropy

ABSTRACT

Maximum likelihood is a standard approach to computing a probability distribution that best fits a given dataset. However, when datasets are incomplete or contain imprecise data, a major issue is to properly define the likelihood function to be maximized. This paper highlights the fact that there are several possible likelihood functions to be considered, depending on the purpose to be addressed, namely whether the behavior of the imperfect measurement process causing incompleteness should be included or not in the model, and what are the assumptions we can make or the knowledge we have about this measurement process. Various possible approaches, that differ by the choice of the likelihood function and/or the attitude of the analyst in front of imprecise information are comparatively discussed on examples, and some light is shed on the nature of the corresponding solutions.

1. Introduction

The key role of likelihood functions in statistical inference was first highlighted by Fisher [16] with the maximum likelihood principle. In his seminal book, Edwards ([15], p. 9) defines a likelihood function as being proportional to the probability of obtaining results given a hypothesis, according to a probability model:

> Let $P(R|H)$ be the probability of obtaining results R given the hypothesis H, according to the probability model ...The likelihood of the hypothesis H given data R, and a specific model, is proportional to $P(R|H)$, the constant of proportionality being arbitrary.

Edwards mentions that “this probability is defined for any member of the set of possible results given any one hypothesis ...As such its mathematical properties are well-known. A fundamental axiom is that if R_1 and R_2 are two of the possible results, mutually exclusive, then $P(R_1 or R_2|H) = P(R_1|H) + P(R_2|H)$”.

[☆] This paper is part of the Virtual special issue on Soft Methods in Probability and Statistics, edited by Barbara Vantaggi, Maria Brigida Ferraro, Paolo Giordani.

[☆☆] A preliminary version of this paper [7] was presented at the 8th Conference on Soft Methods in Probability and Statistics (SMPS) in Roma, September, 12–14, 2016.

[*] Corresponding author.
E-mail addresses: couso@uniovi.es (I. Couso), dubois@irit.fr (D. Dubois).

https://doi.org/10.1016/j.ijar.2017.10.030
0888-613X/© 2017 Elsevier Inc. All rights reserved.

In other words, a fundamental axiom is that the probability of obtaining at least one among two results is the sum of the probabilities of obtaining each of these results. In particular, a *result* in the sense of Edwards is not any kind of event, it is an elementary event. Only elementary events can be observed. For instance, when tossing a die, and seeing the outcome, you cannot observe the event "odd", you can only see 1, 3 or 5. So, a likelihood function is proportional to the conditional probability of an elementary event (the observed sample), where the condition part (the hypothesis) is a value of some model parameter. For instance, the conditional probability of the sure event cannot be viewed as the likelihood of the hypothesis given the sure event.

If this point of view is accepted, what becomes of the likelihood function under incomplete or imprecise observations? To properly answer this question, one must understand what is a result in this context. Namely, if we are interested in a certain random phenomenon modeled by a random variable, observations we get in this case may not directly inform us about this random variable. Due to the interference with an imperfect measurement process, observations will be set-valued [4,5]. So, in order to properly exploit such incomplete information (called *coarse data* in the literature [21]), we must first decide what to model:

1. the random phenomenon *through* its measurement process;
2. or the random phenomenon *despite* its measurement process.

In the first case, imprecise observations are considered as results, and we can construct the likelihood function of a random set, whose realizations are sets. These sets contain precise but ill-known realizations of the random variable of interest, to which we have no direct access. We say that this unreachable random variable is *latent*. Actually, most authors are interested in the other point of view. They consider that outcomes are the precise, although ill-observed, realizations of the random phenomenon, and wish to reconstruct a distribution for the latent variable. However in this case there are as many potential likelihood functions as precise datasets in agreement with the imprecise observations. Authors have proposed several ways of addressing this issue. The most traditional approach is based on the EM algorithm [11,29,13], which is an iterative procedure for efficient maximization of the likelihood of observed data. It constructs a distribution on the latent variable that minimizes divergence from the parametric model in agreement with the available data. It can also serve to reconstruct a sample of the latent variable.

In this paper, we propose a formal setting for the modeling of imprecisely observed random experiments, and define the three likelihood functions that can be built in this framework. Apart from the likelihood function based on available observations, there is the likelihood function based on outcomes of the latent random variable that was imprecisely observed, and the likelihood function based on the joint probability induced by pairs of outcomes and their measurement. The two latter likelihood functions are imprecisely known and we compare several alternatives to the maximization of the likelihood of imprecise observations, such as the maximax approach, and the robust approach to incomplete data. It includes more recent proposals by Hüllermeier [22], or Guillaume and Dubois [18], or Plass et al. [35]. We also discuss the use of assumptions on the measurement process such as the coarsening-at-random [21] and the superset assumptions, that help relating the various likelihood functions. Note that in this paper we do not consider the issue of imprecision due to too small a number of precise observations (see for instance, Masson and Denœux [32], or Serrurier and Prade [44]). We assume that the cause of imprecision lies in the incomplete description of the random experiment outcomes, not in the scarcity of observations.

2. The random phenomenon and its measurement process

Let a random variable $X : \Omega \to \mathcal{X}$ represent the outcome of a certain random experiment. For the sake of simplicity, let us assume that its range $\mathcal{X} = \{a_1, \dots, a_m\}$ is finite. Suppose that observations of X are imprecise, namely let $\Gamma : \Omega \to \wp(\mathcal{X})$ denote the (observable) multi-valued mapping representing our (imprecise) perception of X. So, if ω occurs then all we know is that $X(\omega) \in \Gamma(\omega) \subseteq \mathcal{X}$. In other words, we assume that X is a selection of Γ, i.e. $X(\omega) \in \Gamma(\omega),\ \forall \omega \in \Omega$. This setting is very close to the one of Dempster [10] who introduces a special case of upper and lower probabilities, based on random sets, later interpreted by Shafer [45] as belief and plausibility functions. The issue of set-valued data has been discussed in Ref. [5] from the point of view of descriptive statistics. In this paper we start addressing inferential statistics.

Let $Im(\Gamma) = \{A_1, \dots, A_r\} \in \wp(\mathcal{X})$ denote the image of Γ (the collection of possible set-valued outcomes). We can equivalently suppose that the imperfect measurement process is driven by another random variable Y, with finite range $\mathcal{Y} = \{b_1, \dots, b_r\}$, that provides incomplete reports of observations of X. Namely, $Y(\omega) = b_j$ means that the measurement tool reports $\Gamma(\omega) = A_j$. The cardinality of the image of $Im(Y) = \mathcal{Y} = \{b_1, \dots, b_r\}$ thus coincides with that of $Im(\Gamma)$ and then there is a bijection between $Im(\Gamma)$ and $\mathcal{Y}$ as follows:

$$Y(\omega) = b_j \text{ iff } \Gamma(\omega) = A_j,\ j = 1, \dots, r,$$

or yet we can assume that $b_j = A_j$. Let $P(X, Y)$ be the joint probability describing X and its measurement.

In some applications, the variable X is made of two components X_o and X_u respectively corresponding to observed and unobserved variables with respective domains $\mathcal{X}_o$ and $\mathcal{X}_u$, and Γ is of the form $\{X_o\} \times \Gamma_u$, i.e., $Y = \{\{X_0\} \times \Gamma_u\}$. The observed variable Y can then be identified with the random vector $Y = (X_o, Y_u)$, where $Y_u = \Gamma_u$.

This framework highlights the difference between the outcome $X = a_k$ (its probability is $P(X = a_k)$), the fact that event A_j occurs whenever the outcome $X = a_k$ belongs to A_j (its probability is $P(X \in A_j)$), and observing the result A_j via the