Maximum Likelihood Estimation and Coarse Data

Inés Couso[1(✉)], Didier Dubois[2], and Eyke Hüllermeier[3]

[1] Department of Statistics and O.R., Universidad de Oviedo, Oviedo, Spain
couso@uniovi.es
[2] IRIT, CNRS Université Paul Sabatier, Toulouse, France
dubois@irit.fr
[3] Department of Computer Science, Universität Paderborn, Paderborn, Germany
eyke@upb.de

Abstract. The term coarse data encompasses different types of incomplete data where the (partial) information about the outcomes of a random experiment can be expressed in terms of subsets of the sample space. We consider situations where the coarsening process is stochastic, and illustrate with examples how ignoring this process may produce misleading estimations.

Keywords: Coarse data · Grouped data · Coarsening at random · Maximum likelihood · Visible likelihood · Face likelihood

1 Introduction

The term "coarse data" [15] covers a number of situations treated in the literature such as rounded, heaped, censored or partially missing data. It refers to those situations where we do not get access to the exact value of the data, but only to some subset of the sample space that contains it. Thus, formally speaking, the observations are not assumed to belong to the sample space, but to its power set (see [4,8] for further discussions on set-valued data).

One key problem consists in estimating the distribution of the underlying random variable on the basis of the available incomplete sample data. During the last two decades, different authors have independently studied the way to adapt maximum likelihood estimation (MLE) to this case [5,6,11,13,15,17,20]. In fact, the maximum likelihood procedure results in a consistent estimator of the parameter under some regularity conditions [16], and therefore it is one of the most usual approaches in a variety of machine learning problems. One may adapt the MLE method to incomplete data by considering the collection of possible completions of data, which would lead to a set-valued likelihood function.

The first author thanks the Program Committee Chairs for their kind invitation to participate in the conference. The research in this work has been supported by TIN2014-56967-R (Spanish Ministry of Science and Innovation) and FC-15-GRUPIN14-073 (Regional Ministry of the Principality of Asturias).

S. Moral et al. (Eds.): SUM 2017, LNAI 10564, pp. 3–16, 2017.
DOI: 10.1007/978-3-319-67582-4_1

Thus, maximizing the likelihood function for each of the feasible samples would lead to a set-valued counterpart of the MLE. But this does not seem to be the most reasonable procedure (see comments about extension principle -based approaches in [17], for further details). Two dual alternative procedures have been recently explored [13,14,17]. They consist in replacing the set-valued likelihood either by its upper [17] or its lower bound [13], and seeking for the arg max of the corresponding real-valued mappings. They are respectively referred to as the maximax and the maximin estimators. Some properties of both of them have been recently studied in [14].

A third approach focusses on the observations rather than on the underlying (ill-known) outcomes represented by them. The so-called "visible" likelihood function [6,7] represents the probability of observing the actual observed (set-valued) sample, as a function of a vector of parameters. In order to determine such a function, we do not only need to parametrize the underlying experiment, but also of the coarsening process. The joint distribution over the collection of pairs constituted by the outcomes and their corresponding (incomplete) observations is univocally determined by the marginal distribution over the sample space plus a transition probability from the sample space to its power set, representing the coarsening process. The "visible" likelihood function is nothing else but the likelihood of the marginal distribution over the power set, expressed as a function of the vector of parameters. Different aspects of the arg max of this function have been recently studied in the literature [1,3,5–7,20]. This paper surveys those advances.

2 What Has Occurred and What Do We Know About It?

2.1 Preliminaries and Notation

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 $\mathcal{X} = \{a_1, \ldots, a_m\}$ is finite. Suppose that instead of directly observing X, one observes a coarse version of it, $Y \ni X$. Let $\mathcal{Y} = \{b_1, \ldots, b_r\}$ denote the (finite) set of possible observations, with $b_j = A_j \subseteq \mathcal{X}$, $\forall\, j = 1, \ldots, r$. Let us introduce the following notation:

- $p_{kj} = P(X = a_k, Y = b_j)$ denotes the joint probability of getting the precise outcome $X = a_k$ and observing $b_j = A_j$,
- $p_{k.} = P(X = a_k)$ denotes the probability that the precise outcome is a_k,
- $p_{.j} = P(Y = b_j)$ denotes the probability that the generation plus the imprecisiation processes lead us to observe $b_j = A_j$.
- $p_{.j|k.} = P(Y = A_j | X = a_k)$ denotes the (conditional) probability of observing $b_j = A_j$ if the precise outcome is a_k,
- $p_{k.|.j} = P(X = a_k | Y = A_j)$ denotes the (conditional) probability that the value of X is a_k if we have been reported that it belongs to $b_j = A_j$.

We may represent the joint distribution of (X, Y) by means of the matrix $(M|\mathbf{p})$:

$$\left(\begin{array}{ccc|c} p_{\cdot 1|1\cdot} & \cdots & p_{\cdot r|1\cdot} & p_{1\cdot} \\ \cdots & \cdots & \cdots & \cdots \\ p_{\cdot 1|m\cdot} & \cdots & p_{\cdot r|m\cdot} & p_{m\cdot}, \end{array}\right)$$

where $(p_{1\cdot}, \ldots, p_{m\cdot})^T$ characterizes the distribution of the underlying generating process, while $M = (p_{\cdot j|k\cdot})_{k=1,\ldots,m; j} = 1, \ldots, r$ represents the coarsening process. M is the so-called mixing matrix [25]. We can alternatively characterise it by means of $(M'|\mathbf{p}')$:

$$\left(\begin{array}{ccc|c} p_{1\cdot|\cdot 1} & \cdots & p_{m\cdot|\cdot 1} & p_{\cdot 1} \\ \cdots & \cdots & \cdots & \cdots \\ p_{1\cdot|\cdot r} & \cdots & p_{m\cdot|\cdot r} & p_{\cdot r} \end{array}\right)$$

where the vector $(p_{\cdot 1}, \ldots, p_{\cdot r})^T$ characterises the probability distribution of the observation process, and $M' = (p_{k\cdot|\cdot j})_{k=1,\ldots,m; j=1,\ldots,r}$ represents the conditional probability of X (precise outcome) given Y (observation).

Now, let us assume that the above joint distribution (or equivalently, each of the matrices $(M|\mathbf{p})$ and $(M'|\mathbf{p}')$) is characterized by means of a (vector of) parameter(s) $\theta \in \Theta$. We naturally assume that the dimension of θ is less than or equal to the number of elements in both matrices, i.e., it is less than or equal to $\min\{m\times(r+1), r(m+1)\}$. We also assume that X cannot be written as a function of Y, because such a situation would involve a trivial coarsening process, were Y is just some kind of "encoding" of X. Similarly, we can assume without much loss of generality that X and Y are not independent. Otherwise, the restriction $X \in Y$ would imply that Y is constant, and its image includes all the possible outcomes for X. Furthermore, the parameter is said to be separable [6] wrt $(M|\mathbf{p})$ if it can be "separated" into two (maybe multidimensional) components $\theta_1 \in \Theta_1$, $\theta_2 \in \Theta_2$ such that $\Theta = \Theta_1 \times \Theta_2$, where $p^{\theta}_{\cdot j|k\cdot}$ and $p^{\theta}_{k\cdot}$ can be respectively written as functions of θ_1 and θ_2. This definition corresponds to an earlier notion of "distinctness" of the parameters [15]. Alternatively, θ is said to be separable wrt $(M'|\mathbf{p}')$ if it can be "separated" into two (maybe multidimensional) components $\theta_3 \in \Theta_3$, $\theta_4 \in \Theta_4$ such that $\Theta = \Theta_3 \times \Theta_4$ and $p^{\theta}_{k\cdot|\cdot j}$ and $p^{\theta}_{\cdot j}$ can be respectively written as functions of θ_3 and θ_4. One may think that the notion of "separability" implies some kind of "independence" between X (outcome) and Y (imprecise observation), but this is not the case, as we illustrate below.

We will provide three examples illustrating three different situations: In the first case, Y can be expressed as a function of X, and therefore it determines a partition on $\mathcal{X}$, but their joint distribution depend on a single one-dimensional parameter. In the second case, Y can also be written as a function of X, but the parameters are separable. In the third case, Y is not a function of X, and in fact, it represents a "coarsening at random process" [15] and the joint distribution of (X, Y) depends on a one-dimensional parameter.

Example 1 (Taken from [7]). Let us consider the following example by Dempster et al. in [10] under the light of our analysis. It is based on a former example

by Rao. There is a sample of 197 animals distributed into four categories, so that the observed data consist of:

$$n_{.1} = 125, n_{.2} = 18, n_{.3} = 20, n_{.4} = 34.$$

Suppose that the first category is in fact a mixture of two sub-categories, but we do not have information about the number of individuals observed from each of them. On the other hand, a genetic model for the population specifies the following restrictions about the five categories: $p_{11} = 0.5$, $p_{12} = p_{.4}$, $p_{.2} = p_{.3}$. If we use the notation: $p_{12} = 0.25\pi = p_{.4}$ and $p_{.2} = 0.25(1-\pi) = p_{.3}$, the corresponding matrix $(M'|\mathbf{p}')$ is given as

$$\left(\begin{array}{cc|ccc|c} \frac{0.5}{0.5+0.25\pi} & \frac{0.25\pi}{0.5+0.25\pi} & 0 & 0 & 0 & 0.5+0.25\pi \\ 0 & 0 & 1 & 0 & 0 & 0.25(1-\pi) \\ 0 & 0 & 0 & 1 & 0 & 0.25(1-\pi) \\ 0 & 0 & 0 & 0 & 1 & 0.25\pi \end{array}\right)$$

and only depends on a single parameter.

Example 2. Let X be the random variable that represents the score shown on the top face of a die. Let $(p_1, \ldots, p_6)$ characterize the probability distribution over the set of possible outcomes. Let us suppose that we are just told whether X takes an even or an odd value. We identify the possible observations (values of Y) respectively with $b_1 = \{1, 3, 5\}$ and $b_2 = \{2, 4, 6\}$. This example is formally equivalent to the case of grouping data, where Y can be expressed as a function of X. In other words, the coarsening process is a deterministic procedure where all the values in the mixing matrix M are either 0 or 1. Thus, the distribution of Y only depends on $\theta_1 = p_1 + p_2 + p_3$. Let us now consider the matrix $M' = (m'_{ij})_{i,j}$ where the two-dimensional variable (X, Y) can take six different values, and its joint distribution can be expressed in terms of $(p_1, \ldots, p_5)$. It can be also written as a function of $\theta_1 = p_1 + p_3 + p_5$ (determining the marginal distribution of Y) and the four-dimensional vector $\boldsymbol{\theta}_2 = (\frac{p_1}{\theta_1}, \frac{p_3}{\theta_1}, \frac{p_2}{1-\theta_1}, \frac{p_4}{1-\theta_1})$ (that characterizes the disambiguation process). Thus, the joint distribution is separable wrt M' and p'.

Example 3. Suppose a coin is flipped and let X be the binary random variable that takes the value 1= "heads", and 0="tails". Suppose that half of the times, we are not informed about the result (regardless what the result is). The coarsening process is therefore characterised as follows:

$$P(Y = \{0, 1\}|X = 0) = P(Y = \{0, 1\}|X = 1) = 0.5,$$

$$P(Y = \{0\}|X = 0) = P(Y = \{1\}|X = 1) = 0.5.$$

This process agrees with the notion of coarsening at random (CAR) introduced by Heitjan and Rubin, to be discussed later on, since the fact of being informed of the result does not depend on the result itself. Furthermore, it satisfies a stronger property called "superset assumption" [18], since we are informed half of the times, on average, about the result of the coin, whatever it is. Notwithstanding,

the joint distribution of (X, Y) can be expressed in terms of the one-dimensional parameter $p \in (0, 1)$ denoting the probability of heads. In fact, under the above assumptions, we have:

$$P(X = 1, Y = \{1\}) = P(X = 1, Y = \{0, 1\}) = 0.5p,$$

$$P(X = 0, Y = \{0\}) = P(X = 0, Y = \{0, 1\}) = 0.5(1 - p).$$

As a conclusion, the above example satisfies the so-called property of "missing at random" (MAR), but the joint distribution of (X, Y) is completely characterised by marginal distribution of Y, since both of them depend on the same -single-parameter.

As a matter of fact, the parameter of the joint distribution can be written as a function of the parameter of the distribution of Y when the distribution about the instantiation process is known, given the marginal distribution of Y. This does not seem to be related to the degree of dependence between X and Y. In this case, the problem of identifiability of the parameter of the marginal distribution of Y reduces to the problem of identifiability of the parameter of the joint likelihood function, and therefore, a MLE procedure based on the "visible" likelihood function seems a good option in order to estimate the parameter.

2.2 The Outcomes of an Experiment and Their Incomplete Observations

According to the framework developed in the last subsection, we can easily observe that, given some $b_j = A_j \in \mathcal{Y}$, the two events $X \in A_j$ and $Y = A_j$ do not coincide in general. In fact, it is generally assumed that the latter implies the former, but the equivalence does not hold in general: For, suppose that $X \in A_j$ implies $Y = A_j$. Therefore, for every $a_k \in A_j$, we can derive that $X = a_k$ implies $X \in A_j$ and therefore $Y = A_j$. Thus, we can deduce that $P(Y = A_j | X = a_k) = 1, \forall a_k \in A_j$. Thus, the above implication entails a deterministic coarsening process, inducing a partition over the set of outcomes $\mathcal{X}$.

Let us illustrate the difference between the events $X \in A_j$ and $Y = A_j$ and their corresponding probabilities with an example:

Example 4 (Taken from [7]). Consider the random experiment that consists on rolling a dice. We do not know whether the dice is fair or not. Take a sample of N tosses of the dice and assume that the reporter has told us n_1 of the times that the result was less than or equal to 3 and the remaining $n_2 = N - n_1$ tosses, he told us that it was greater than or equal to 3. After each toss, when the actual result (X) is 3, the reporter needs to make a decision. Let us assume that the conditional probability $P(Y = \{1, 2, 3\} | X = 3)$ is a fixed number $\alpha \in [0, 1]$. The joint distribution of (X, Y) can be written as a function of $(p_1, \ldots, p_6)$ and α, since it is determined by the following matrix: $(M|\mathbf{p})$:

$$\left(\begin{array}{cc|c} 1 & 0 & p_1 \\ 1 & 0 & p_2 \\ \alpha & 1-\alpha & p_3 \\ 0 & 1 & p_4 \\ 0 & 1 & p_5 \\ 0 & 1 & p_6 \end{array}\right)$$

corresponding to the joint probability

Y, X	1	2	3		4	5	6
y_1	p_1	p_2	$\alpha\, p_3$		0	0	0
y_2	0	0	$(1-\alpha)\, p_3$	p_4	p_5	p_6	

We can easily make the distinction between the two events $X \in \{1,2,3\}$ (the result is less than or equal to 3) and $Y = \{1,2,3\}$ (we are told that the result is less than or equal to 3) and their corresponding probabilities. According to the above notation, the probability of the first event is

$$P(X \in \{1,2,3\}) = p_1 + p_2 + p_3,$$

while the probability of the latter is:

$$\begin{aligned} P(Y = \{1,2,3\}) = \\ P(X=1, Y=\{1,2,3\}) + P(X=2, Y=\{1,2,3\}) + P(X=3, Y=\{1,2,3\}) \\ = p_1 + p_2 + \alpha\, p_3. \end{aligned}$$

3 The Optimization Problem: What Should We Maximise?

Let us consider a sequence $\mathbf{Z} = ((X_1, Y_1), \ldots, (X_N, Y_N))$ of N iid copies of $Z = (X, Y)$. We will use the nomenclature $\mathbf{z} = ((x_1, y_1), \ldots, (x_N, y_N)) \in (\mathcal{X} \times \mathcal{Y})^N$ to represent a specific sample of the vector (X, Y). Thus, $\mathbf{y} = (y_1, \ldots, y_N)$ will denote the observed sample (an observation of the vector $\mathbf{Y} = (Y_1, \ldots, Y_N)$), and $\mathbf{x} = (x_1, \ldots, x_N)$ will denote an arbitrary artificial sample from $\mathcal{X}$ for the unobservable (latent) variable X, that we shall vary in $\mathcal{X}^N$. We can describe any sample $\mathbf{z}$ in frequentist terms assuming exchangeability:

- $n_{kj} = \sum_{i=1}^{N} 1_{\{(a_k, b_j)\}}(x_i, y_i)$ is the number of repetitions of (a_k, b_j) in the sample $\mathbf{z}$;
- $\sum_{k=1}^{m} n_{kj} = n_{\cdot j}$ be the number of observations of $b_j = A_j$ in $\mathbf{y}$;
- $\sum_{j=1}^{r} n_{kj} = n_{k\cdot}$ be the number of appearances of a_j in $\mathbf{x}$.

Clearly, $\sum_{k=1}^{m} n_{k\cdot} = \sum_{j=1}^{r} n_{\cdot j} = N$. Let the reader notice that, once a specific sample $\mathbf{y} = (y_1, \ldots, y_N) \in \mathcal{Y}^N$ has been observed, the number of n_{kj} repetitions of each pair $(a_k, b_j) \in \mathcal{X} \times \mathcal{Y}$ in the sample, can be expressed as a function of $\mathbf{x} = (x_1, \ldots, x_N)$.

3.1 Different Generalizations of the Notion of the Likelihood Function

We may consider the following two generalizations of the likelihood function (and their respective logarithms), depending on whether our sequence of observations $\mathbf{y} = (y_1, \ldots, y_N)$ is interpreted either as a singleton in $\mathcal{Y}^N$ or as a non-trivial subset of $\mathcal{X}^N$:

- $\mathbf{p}(\mathbf{y}; \theta) = \prod_{i=1}^{N} p(y_i; \theta)$ denotes the probability of observing $\mathbf{y} \in \mathcal{Y}^N$, assuming that the value of the parameter is θ. It can be alternatively expressed as $\mathbf{p}(\mathbf{y}; \theta) = \prod_{j=1}^{r} (p_{\cdot j}^{\theta})^{n_{\cdot j}}$, where $n_{\cdot j}$ denotes the number of repetitions of $b_j = A_j$ in the sample of size N (the number of times that the reporter says that the outcome of the experiment belongs to A_j.) The logarithm of this likelihood function will be denoted by

$$L^{\mathbf{y}}(\theta) = \log \mathbf{p}(\mathbf{y}; \theta) = \sum_{i=1}^{N} \log p(y_i; \theta) = \sum_{j=1}^{r} n_{\cdot j} \log p_{\cdot j}^{\theta}.$$

  We call $\mathbf{p}(\mathbf{y}; \theta)$ the *visible likelihood function* [7], because we can compute it based on the available data only, that is the observed sample $\mathbf{y}$. It is also sometimes called the *marginal likelihood of the observed data* in the EM literature, not to be confused with the *marginal likelihood* in a Bayesian context (see [2], for instance).
- Alternatively,

$$\lambda(\mathbf{y}; \theta) = \prod_{j=1}^{r} P(X \in A_j; \theta)^{n_{\cdot j}},$$

  called the "face likelihood" in [9,20] does not refer to the observation process, and replaces the probability of reporting A_j as the result of an observation (i.e. $P(Y = A_j)$) by the probability that the precise outcome falls inside the set A_j, $P(X \in A_j)$. As we have previously noticed, the occurrence of event "$X \in A_j$" is a consequence of, but does not necessarily coincide with the outcome "$Y = A_j$". In our context, $\mathbf{p}(\mathbf{y}; \theta)$ represents the probability of occurrence of the result "$(Y_1, \ldots, Y_N) = \mathbf{y}$", given the hypothesis θ. Therefore given two arbitrary different samples $\mathbf{y} \neq \mathbf{y}'$ the respective events $(Y_1, \ldots, Y_N) = \mathbf{y}$ and "$(Y_1, \ldots, Y_N) = \mathbf{y}'$" are mutually exclusive. In contrast, $\lambda(\mathbf{y}; \theta)$ denotes the probability of occurrence of the event $(X_1, \ldots, X_N) \in y_1 \times \ldots \times y_N$. Events of this form may overlap, in the sense that, given two different samples $\mathbf{y} \neq \mathbf{y}'$, the corresponding events $(X_1, \ldots, X_N) \in y_1 \times \ldots \times y_N$ and $(X_1, \ldots, X_N) \in y'_1 \times \ldots \times y'_N$ are not necessarily mutually exclusive. Therefore $\lambda(\mathbf{y}; \theta)$ can not be regarded as a likelihood in the sense of Edwards [12]. This criterion has been generalized to uncertain data and exploited in the Evidential EM algorithm of Denœux [11]. This extension of EM has been successfully used in some applications (see [23,24] and references therein).

The above functions $\mathbf{p}$ and λ do coincide if and only if the coarsening process is deterministic, and therefore, the collection of sets $\{A_1, \ldots, A_r\}$ forms a partition of $\mathcal{X}$. In fact, $P(Y = A_j) \leq P(X \in A_j)$ for every $j = 1, \ldots, r$ and the

equalities hold when the coarsening is deterministic. Otherwise, if there exists a pair (k, j) with $a_k \in A_j$ and $P(Y = A_j|X = a_k) < 1$ then we easily derive that $P(Y = A_j)$ is strictly smaller than $P(X \in A_j)$ and therefore, we deduce that $\mathbf{p}(\mathbf{y}, \theta)$ is strictly less than $\lambda(\mathbf{y}, \theta)$. But we may ask ourselves whether the maximization of each of those functions leads or not to the same pair of maximizers, even in those cases where they do not coincide. The next example illustrates a situation where both methods lead to completely different estimators.

Example 5. Consider again the situation described in Example 4. Furthermore, suppose that the reporter provides us with the following additional information: when the result is $X = 3$, he will flip a coin. If it lands heads, he will tells us that the result is less than or equal to 3. Otherwise, he will tell us that it is greater than or equal to 3. Mathematically, $\alpha = P(Y = \{1, 2, 3\}|X = 3) = 0.5$. Under these conditions, the visible likelihood is

$$\mathbf{p}(\mathbf{y}, \theta) = (p_1 + p_2 + 0.5\, p_3)^{300} + (0.5\, p_3 + p_4 + p_5 + p_6)^{700}.$$

It attains its maximum value for every $(\hat{p}_1, \ldots, \hat{p}_6)$ satisfying the restrictions: $\hat{p}_1 + \hat{p}_2 + 0.5\, \hat{p}_3 = 0.3$ and $0.5\, \hat{p}_3 + \hat{p}_4 + \hat{p}_5 + \hat{p}_6 = 0.7$ (The set of solutions is not a singleton). Alternatively, the face likelihood function is calculated as follows:

$$\lambda(\mathbf{y}, \theta) = (p_1 + p_2 + p_3)^{300} + (p_3 + p_4 + p_5 + p_6)^{700}.$$

It attains the maximum value for $(\hat{p}_1, \ldots, \hat{p}_6) = (0, 0, 1, 0, 0, 0)$. In other words, according to this maximization procedure, the experiment is assumed to be deterministic.

Both optimization procedures lead to completely different solutions. In fact, according to the first set of solutions, p_3 is upper bounded by 0.6, while in the second case it is assumed to be equal to 1. Furthermore, according to the Weak Law of Large Numbers, the relative frequencies $\frac{n_{\cdot 1}}{N}$ and $\frac{n_{\cdot 2}}{N}$ respectively converge in probability to $p_{\cdot 1}$ and $p_{\cdot 2}$ that, according to the information, respectively coincide with $p_1 + p_2 + 0.5\, p_3$ and $0.5\, p_3 + p_4 + p_5 + p_6$. Thus, the first procedure (the one based on the visible likelihood) satisfies the following consistency property:

$$\lim_{n \to \infty} \hat{p}_1 + \hat{p}_2 + 0.5\, \hat{p}_3 = p_1 + p_2 + 0.5\, p_3$$

and

$$\lim_{n \to \infty} 0.5\, \hat{p}_3 + \hat{p}_4 + \hat{p}_5 + \hat{p}_6 = 0.5\, p_3 + p_4 + p_5 + p_6.$$

In contrast, the estimation based on the face likelihood does not satisfy the above consistency property unless the underlying probability satisfies the following equality:

$$p_1 + p_2 + 0.5\, p_3 = 0.5\, p_3 + p_4 + p_5 + p_6 = 0.5.$$

The differences between the visible and the face likelihood functions have been studied in practice in relation with incomplete ranked data in [1,3]. In fact, incomplete rankings are viewed there as coarse observations of ranked data. Let $\mathcal{X} = \mathbb{S}_3$ denote the collection of rankings (permutations) over a set

$U = \{a_1, a_2, a_3\}$ of 3 items. We denote by $\pi : \{1,2,3\} \Rightarrow \{1,2,3\}$ a complete ranking (a generic element of $\mathbb{S}_3$), where $\pi(k)$ denotes the position of the k^{th} item a_k in the ranking. An incomplete ranking τ can be associated with the collection of complete rankings that are in agreement with it denoted $E(\tau)$. An important special case is an incomplete ranking τ in the form of a pairwise comparison $a_i \succ a_j$, which is associated with the set of extensions

$$E(\tau) = E(a_i \succ a_j) = \{\pi \in \mathbb{S}_K \, : \, \pi(i) < \pi(j)\}.$$

For every pair (i,j), let n_i denote the number of times that the incomplete ranking τ_i is observed in a sample of size N. Let us furthermore assume that the marginal distribution of X on $\mathbb{S}_3$ belongs to a family of distributions parametrized by some vector of parameters θ, while the coarsening process is determined by some λ. The face likelihood based on the above sample is calculated as follow:

$$\lambda(\mathbf{y};\theta) = \prod_{i=1}^{3} \prod_{j \neq i} P_\theta(X \in E(\tau_i))^{n_i},$$

while the visible likelihood function is calculated as

$$p(\mathbf{y};\theta) = \prod_{i=1}^{3} \prod_{j \neq i} P_{(\theta,\lambda)}(Y = \tau_i)^{n_i}.$$

They do not coincide in general. Let us consider, for instance, the top-2 setting, in which always the two items on the top of the ranking are observed. The corresponding mixing matrix denotes a one-to-one correspondence between π_i and τ_i, for $i = 1, \ldots, 6$, where:

$$\begin{array}{ll} \pi_1(1) = 1, \pi_1(2) = 2, \pi_1(3) = 3 & \\ \pi_2(1) = 1, \pi_2(2) = 3, \pi_2(3) = 2 & \\ \pi_3(1) = 2, \pi_3(2) = 1, \pi_3(3) = 3 & \\ \pi_4(1) = 2, \pi_4(2) = 3, \pi_4(3) = 1 & \\ \pi_5(1) = 3, \pi_5(2) = 1, \pi_5(3) = 2 & \\ \pi_6(1) = 3, \pi_6(2) = 2, \pi_6(3) = 1 & \end{array} \quad \text{and} \quad \begin{array}{l} \tau_1 = a_1 \succ a_2 \\ \tau_2 = a_1 \succ a_3 \\ \tau_3 = a_2 \succ a_1 \\ \tau_4 = a_3 \succ a_1 \\ \tau_5 = a_2 \succ a_3 \\ \tau_6 = a_3 \succ a_2 \end{array}$$

Thus Y takes the "value" τ_i if and only if $X = \pi_i$, for all $i = 1, \ldots, 6$. Let us furthermore notice that each partial ranking τ_i represents a collection of three different complete rankings:

$$\begin{array}{l} E(\tau_1) = \{\pi_1, \pi_2, \pi_4\} \\ E(\tau_2) = \{\pi_1, \pi_2, \pi_3\} \\ E(\tau_3) = \{\pi_1, \pi_2, \pi_4\} \\ E(\tau_4) = \{\pi_1, \pi_2, \pi_4\} \\ E(\tau_5) = \{\pi_1, \pi_2, \pi_4\} \\ E(\tau_6) = \{\pi_1, \pi_2, \pi_4\} \end{array}$$

Thus, the face and the visible likelihood functions are respectively calculated as follows:

$$\lambda(\mathbf{y};\theta) = \prod_{i=1}^{6} P_\theta(X \in E(\tau_i))^{n_i}$$

while

$$p(\mathbf{y}, \theta) = \prod_{i=1}^{6} P_\theta(X = \pi_i)^{n_i}.$$

They do not lead in general to the same estimations, as it is checked in [3]. In fact, under some general assumptions about the underlying generating process, the visible likelihood-based estimator is consistent, while the face likelihood-based estimator is not. Some additional formal studies about the consistency of both estimators under different assumptions about the coarsening process are performed in [1].

3.2 Different Assumptions About the Coarsening and the Disambiguation Processes

Different assumptions about the coarsening and the disambiguation processes have been investigated in the literature [1,15,18,21,22]. The purpose in some of those cases was to establish simple conditions under which the stochastic nature of the coarsening process could be ignored when drawing inferences from data. This subsection reviews two assumptions, one about the coarsening process and the other one about the disambiguation process, both of them commonly considered in the literature.

Coarsening at Random. One common assumption about the coarsening process is the so-called "coarsening at random" assumption (CAR). It was introduced by Heitjan and Rubin [15]. According to it, the underlying data do not affect the observations. Mathematically,

$$P(Y = A_j | X = a_k) = P(Y_j = A_j | X = a_{k'}),\ \forall\, a_k, a'_k \in A_j.$$

Two remarkable particular cases of CAR are:

- Grouping. We speak about grouped data [15] when the coarsening process is deterministic, and therefore $P(Y = A_j|X = a_k)$ is either 1 (if $a_k \in A_j$) or 0 (otherwise). In this case, the set $\{A_1, \ldots, A_r\}$ forms a partition of the collection of possible outcomes $\mathcal{X}$.
- Missing at random (MAR).- It particularizes the CAR assumption to the case where data are either completely observed or missing, and therefore, the collection of possible observations is $\mathcal{Y} = \{\{a_1\}, \ldots, \{a_m\}, \mathcal{X}\}$. The MAR assumption means that missingness is not affected by the underlying outcome.

The first one illustrates the partition case.

Example 6 (Taken from [19]). Let $X = (X_1, X_2)$ with $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 = \{p, n\} \times \{p, n\}$ We interpret X_1, X_2 as two medical tests with possible outcomes positive or negative. Suppose that test X_1 always is performed first on a patient, and that test X_2 is performed if and only if X_1 comes out positive. Possible observations that can be made then are $b_1 = \{(n, n), (n, p)\}$, $b_2 = \{(p, n)\}$ and $b_3 = \{(p, p)\}$.

These three outcomes determine a partition of $\mathcal{X}$. Therefore, the matrix M is determined by the following 0–1 conditional probabilities, and CAR is trivially satisfied. In fact:

$$P(Y = b_1|X = (n,n)) = P(Y = b_1|X = (n,p)) = 1,$$

$$P(Y = b_2|X = \{(p,n)\}) = 1,$$
$$P(Y = b_3|X = \{(p,p)\}) = 1.$$

The following example illustrates the missing at random assumption:

Example 7 (Taken from [7]). A coin is tossed. The random variable $X : \Omega \to \mathcal{X}$, where $\mathcal{X} = \{h,t\}$, represents the result of the toss. We do not directly observe the outcome, that is reported by someone else, who sometimes decides not to tell us the result. The rest of the time, the information he provides about the outcome is faithful. Let Y denote the information provided by this person about the result. It takes the "values" $\{h\}$, $\{t\}$ and $\{h,t\}$.

This example corresponds to the following matrix $(M|\mathbf{p})$ where $a_{kj} = p_{.j|k.}$, $k = 1,2; j = 1,2,3$:

$$\left(\begin{array}{ccc|c} 1-\alpha & 0 & \alpha & p \\ 0 & 1-\beta & \beta & 1-p \end{array}\right)$$

The marginal distribution of X (outcome of the experiment) is given as

- $p_{1.} = P(X = h) = p$,
- $p_{2.} = P(X = t) = 1 - p$.

The joint probability distribution of (X,Y) is therefore determined by:

$$\left(\begin{array}{c|ccc} X\backslash Y & \{h\} & \{t\} & \{h,t\} \\ \hline h & (1-\alpha)p & 0 & \alpha p \\ t & 0 & (1-\beta)(1-p) & \beta(1-p) \end{array}\right)$$

Under the MAR assumption, we have that $\alpha = \beta$, i.e.,

$$P(Y = \{h,t\}|X = h) = P(Y = \{h,t\}|X = t).$$

When furthermore the model is separable with respect to the matrix $(M|p)$, the coarsening process can be ignored, in the sense that both the visible and the face likelihood lead to the same estimator of the parameter. This has been proved by Heitjan and Rubin (see [15,20]). Additional conditions under which the stochastic nature of the coarsening process can be ignored in some practical problems have been recently studied in [1,3].

Uniform Disambiguation Process. We can alternatively make assumptions about the disambiguation process. When dealing with noisy observations, it is not unusual to assume that all the possible outcomes compatible with an observation $Y = A_j$ (i.e., all the elements in A_j) are equally probable, and therefore $P(X = a_k|Y = A_j) = 1_{A_j}(a_k) \cdot \frac{1}{\#A_j}$, $\forall a_k \in A_j$. According to this assumption, the probability induced by X on $\mathcal{X}$ corresponds to the pignistic transform [26] of the mass function derived from the marginal distribution of Y as follows:

$$m(A_j) = P(Y = A_j),\ j = 1, \ldots, r.$$

Contrarily to what happens with the CAR assumption, under this alternative assumption, the face and the visible likelihood do not necessarily lead to the same estimator.

Example 8. Consider once more the situation described in Example 4, and assume a uniform disambiguation process. Let p denote the probability of the event $Y = \{1, 2, 3\}$. The visible likelihood can be written as a function of p as:

$$\mathbf{p}(\mathbf{y}, p) = p^{n_{\cdot 1}}(1-p)^{n_{\cdot 2}}.$$

The marginal probability over $\mathcal{X}$ can be written as a function of p as follows:

$$P(X = 1) = P(X = 2) = \frac{p}{3},\ P(X = 3) = \frac{p}{3} + \frac{1-p}{4},$$

$$P(X = 4) = P(X = 5) = P(X = 6) = \frac{1-p}{4}.$$

Therefore, the face likelihood is different from the visible likelihood:

$$\lambda(\mathbf{y}, p) = \left(p + \frac{1-p}{4}\right)^{n_{\cdot 1}} \left(\frac{p}{3} + (1-p)\right)^{n_{\cdot 2}}.$$

4 Concluding Remarks

We have provided an overview of the maximization procedures based on the so-called visible and face likelihood functions. The face likelihood depends on the marginal distribution of X, while the visible likelihood depends on the marginal distribution of Y. Both, the face and the visible likelihoods have their advantages and their caveats. When the parameter is separable with respect to the matrix $(M|p)$ (distinctness in the context of Heitjan and Rubin), the first one only depends on θ_3 while the second one depends on both, θ_3 and θ_4. The MLE based on the visible likelihood is therefore not unique in this case, unless the parameter set Θ_4 is a singleton. But, although the arg max of the face likelihood may be unique in those cases, it is not a consistent estimator in general, as we have observed. The visible likelihood involves the probability of observing the different outcomes $Y = A_j$ (as a function of the parameter) and the proportion of times each of them is observed in the sample. Such a proportion converges in

probability to the (true) probability of the event, and therefore, under some regularity conditions, the arg max of the visible function is consistent. Alternatively, the face likelihood replaces the probability of observing $Y = A_j$ by the probability of occurrence of $X \in A_j$. The vector $(q_1, \ldots, q_r)$, where $q_i = P(X \in A_j)$ for all j[1] is not proportional in general to the vector $(p_{\cdot 1}, \ldots, p_{\cdot r})$ and therefore, the arg max of the face likelihood is not consistent in general.

Some recent studies compare the maximization of the visible likelihood function with other strategies such as the maximax and the maximin approaches mentioned at the beginning of this paper. In this line, the face likelihood can be regarded as a max-average approach, in the sense that it maximizes the average of the likelihoods of all the feasible samples on $\mathcal{X}^N$ (all the samples of the form $\mathbf{x} = (x_1, \ldots, x_n)$ satisfying the restriction $x_i \in y_i$, $\forall\, i$) (see [17] for further details.) Further theoretical and empirical studies are needed in order to determine what is the best strategy in each practical situation.

References

1. Ahmadi, M., Hüllermeier, E., Couso I.:, Statistical inference for incomplete ranking data: the case of rank-dependent coarsening. In: Proceedings of the 34th International Conference on Machine Learning (2017 ICML), Sydney (Australia)
2. Chib, S.: Marginal likelihood from the Gibbs output. J. Am. Stat. Assoc. **90**, 1313–1321 (1995)
3. Couso, I., Ahmadi, M., Hüllermeier, E.: Statistical inference for incomplete ranking data: a comparison of two likelihood-based estimators. In: Proceedings of DA2PL 2016 (From Multiple Criteria Decision Aid to Preference Learning), Paderborn (Germany) (2016)
4. Couso, I., Dubois, D.: Statistical reasoning with set-valued information: ontic vs. epistemic views. Int. J. Approximate Reasoning **55**, 1502–1518 (2014)
5. Couso, I., Dubois, D.: Belief revision and the EM algorithm. In: Carvalho, J.P., Lesot, M.-J., Kaymak, U., Vieira, S., Bouchon-Meunier, B., Yager, R.R. (eds.) IPMU 2016. CCIS, vol. 611, pp. 279–290. Springer, Cham (2016). doi:10.1007/978-3-319-40581-0_23
6. Couso, I., Dubois, D.: Maximum likelihood under incomplete information: toward a comparison of criteria. In: Ferraro, M.B., Giordani, P., Vantaggi, B., Gagolewski, M., Gil, M.Á., Grzegorzewski, P., Hryniewicz, O. (eds.) Soft Methods for Data Science. AISC, vol. 456, pp. 141–148. Springer, Cham (2017). doi:10.1007/978-3-319-42972-4_18
7. Couso, I., Dubois, D.: A general framework for maximizing likelihood under incomplete data, under review
8. Couso, I., Dubois, D., Sánchez, L.: Random Sets and Random Fuzzy Sets as Ill-Perceived Random Variables. SAST. Springer, Cham (2014). doi:10.1007/978-3-319-08611-8
9. Dawid, A.P., Dickey, J.M.: Likelihood and Bayesian inference from selectively reported data. J. Amer. Statist. Assoc. **72**, 845–850 (1977)

[1] Let the reader notice that this vector does not necessarily represent a probability distribution. In fact, the sum $\sum_{j=1}^{r} q_j$ is strictly greater than 1, unless the collection of A_j forms a partition of $\mathcal{X}$.

10. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm (with discussion). J. Roy. Statist. Soc. B **39**, 1–38 (1977)
11. Denœux, T.: Maximum likelihood estimation from uncertain data in the belief function framework. IEEE Trans. Knowl. Data Eng. **26**, 119–130 (2013)
12. Edwards, A.W.F.: Likelihood. Cambridge University Press, Cambridge (1972)
13. Guillaume, R., Dubois, D.: Robust parameter estimation of density functions under fuzzy interval observations. In: 9th ISIPTA Symposium, Pescara, Italy, pp. 147–156 (2015)
14. Guillaume, R., Couso, I., Dubois, D.: Maximum likelihood and robust optimisation on coarse data. In: 10th ISIPTA Symposium, Lugano, Switzerland, pp. 147–156 (2017)
15. Heitjan, D.F., Rubin, D.B.: Ignorability and coarse data. Ann. Stat. **19**, 2244–2253 (1991)
16. Huber, P.J.: The behavior of maximum likelihood estimates under nonstandard conditions. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 221–233. University of California Press (1967)
17. Hüllermeier, E.: Learning from imprecise and fuzzy observations: data disambiguation through generalized loss minimization. Int. J. Approximate Reasoning **55**, 1519–1534 (2014)
18. Hüllermeier, E., Cheng, W.: Superset learning based on generalized loss minimization. In: Appice, A., Rodrigues, P.P., Santos Costa, V., Gama, J., Jorge, A., Soares, C. (eds.) ECML PKDD 2015. LNCS, vol. 9285, pp. 260–275. Springer, Cham (2015). doi:10.1007/978-3-319-23525-7_16
19. Jaeger, M.: Ignorability in statistical and probabilistic inference. J. Artif. Intell. Res. (JAIR) **24**, 889–917 (2005)
20. Jaeger, M.: The AI&M procedure for learning from incomplete data. In: Proceedings of Uncertainty in Artificial Intelligence Conference (UAI-06), pp. 225–232 (2006)
21. Plass, J., Augustin, T., Cattaneo, M., Schollmeyer, G.: Statistical modelling under epistemic data imprecision: some results on estimating multinomial distributions and logistic regression for coarse categorical data. In: Proceedings of the 9th International Symposium on Imprecise Probability: Theories and Applications (ISIPTA 2015), Pescara (Italy) (2015)
22. Plass, J., Cattaneo, M.E.G.V., Schollmeyer, G., Augustin, T.: Testing of coarsening mechanisms: coarsening at random versus subgroup independence. In: Ferraro, M.B., Giordani, P., Vantaggi, B., Gagolewski, M., Gil, M.Á., Grzegorzewski, P., Hryniewicz, O. (eds.) Soft Methods for Data Science. AISC, vol. 456, pp. 415–422. Springer, Cham (2017). doi:10.1007/978-3-319-42972-4_51
23. Quost, B., Denœux, T.: Clustering and classification of fuzzy data using the fuzzy EM algorithm. Fuzzy Sets Syst. **286**, 134–156 (2016)
24. Ramasso, E., Denœux, T.: Making use of partial knowledge about hidden states in HMMs: an approach based on belief functions. IEEE Trans. Fuzzy Syst. **22**(2), 395–405 (2014)
25. Sid-Sueiro, J.: Proper losses for learning from partial labels. In: Proceedings of Neural Information Processing Systems Conference (NIPS 2012), Lake Tahoe, Nevada, USA (2012)
26. Smets, P.: Constructing the pignistic probability function in a context of uncertainty. In: Henrion M., et al. (eds.) Uncertainty in Artificial Intelligence 5, pp. 29–39. North-Holland, Amsterdam (1990)