International Journal of Approximate Reasoning 90 (2017) 292–306

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

On the testability of coarsening assumptions: A hypothesis test for subgroup independence ☆

J. Plass [a,*], M. Cattaneo [b], G. Schollmeyer [a], T. Augustin [a]

[a] Department of Statistics, LMU Munich, Ludwigsstr. 33, 80539 Munich, Germany
[b] School of Mathematics & Physical Sciences, University of Hull, Hull, HU6 7RX, UK

ARTICLE INFO

Article history:
Received 10 January 2017
Received in revised form 4 July 2017
Accepted 25 July 2017
Available online 19 September 2017

Keywords:
Coarse data
Missing data
Coarsening at random (CAR)
Likelihood-ratio test
Partial identification
Sensitivity analysis

ABSTRACT

Since coarse(ned) data naturally induce set-valued estimators, analysts often assume coarsening at random (CAR) to force them to be single-valued. Focusing on a coarse categorical response variable and a precisely observed categorical covariate, we first re-illustrate the impossibility to test CAR and then contrast it to another type of coarsening called subgroup independence (SI). It turns out that – depending on the number of subgroups and categories of the response variable – SI can be point-identifying as CAR, but testable unlike CAR. A main goal of this paper is the construction of the likelihood-ratio test for SI. All issues are similarly investigated for the here proposed generalized versions, gCAR and gSI, thus allowing a more flexible application of this hypothesis test. The results are illustrated by the data of the German Panel Study "Labour Market and Social Security" (PASS).

1. Introduction: the problem of testing coarsening assumptions

Traditional statistical methods dealing with missing data (e.g. EM algorithm or imputation techniques) require identifiability of parameters, which frequently tempts analysts to make the *missing at random* (MAR) assumption (cf. e.g. [17]) simply for pragmatic reasons without justifications in substance (cf. e.g. [15]). Since MAR is not testable without strong additional assumptions (e.g. [18]) and wrongly including MAR may induce a substantial bias, this way to proceed is especially alarming.

Beside missing data, there are further kinds of deficient data, such as data affected by measurement errors/misclassification (cf. e.g. [11]) or coarse(ned) data (cf. e.g. [12]) where only subsets of the complete data sample space are observed, known to include the unobserved, precise value.[1] Throughout the paper, we consider coarse data, including missing data as special case, thus addressing partially observed values, explicitly excluding the erroneous observation of a variable, disregarding measurement errors/misclassification. For instance, coarse data may arise in data sets where coarsening is

☆ 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 was presented at the 8th Conference on Soft Methods in Probability and Statistics (SMPS) in Rome, September, 12–14, 2016 [25].

* Corresponding author.
E-mail address: julia.plass@stat.uni-muenchen.de (J. Plass).

[1] When dealing with coarse data, it is important to distinguish *epistemic data imprecision* considered here, i.e. incomplete observations due to an imperfect measurement process, from *ontic data imprecision* (cf. [5]).

http://dx.doi.org/10.1016/j.ijar.2017.07.014
0888-613X/

deliberately applied as anonymization technique or matched data sets with not completely identical categories. In the context of coarse data, the *coarsening at random* (CAR) (cf. [12]) assumption is the analogue of MAR. Although the impossibility of testing CAR is already known from literature (cf. e.g. [14]), providing an intuitive insight into this point will be a first goal of our paper. Apart from CAR, we focus on another, in a sense dual, assumption that we called *subgroup independence* (SI) in [22] and elaborate the substantial difference between CAR and SI with regard to testability.

Our argumentation is based on the maximum likelihood estimators obtained under the specific assumptions in focus. There is already a variety of maximum likelihood approaches for incomplete data. While some rely on optimization strategies, as for instance maximax or maximin, to force a single-valued result (cf. e.g. [10], [13]), others end up with set-valued results (cf. e.g. [3], [16], [22]). A general view is given by Couso and Dubois [6], distinguishing between different types of likelihoods, the visible, the latent and the total likelihood. Here, we use the cautious approach developed in [22], which refers to the latent likelihood and is – just as e.g. [19,8] (in the context of misclassification) and [28] – strongly influenced by the methodology of *partial identification* (cf. [18]). Thus, according to the spirit of partial identification, instead of being forced to make often untenable, strict assumptions, as CAR or SI, to give an answer to the research question at all, we can explicitly make use of in practice more realistic partial knowledge about the incompleteness, which would have to be left out of considerations if traditional approaches were used. For this purpose, we use an observation model as a powerful medium to include the available knowledge into the estimation problem. By considering generalized versions of the strict assumptions in focus, which we call gCAR and gSI, we can express this knowledge in a flexible and careful way. This means that we are no longer restricted to formalize the very specific types of coarsening assumptions, but can incorporate (even partial) knowledge about arbitrary dependencies of the coarsening on the values of some variables, which turns out to be also beneficial in the context of testing.

Throughout the paper, we refer to the case of a coarse categorical response variable Y and a precisely observed categorical covariate X, but the results may be easily formulated in terms of cases with more than one categorical covariate. For sake of conciseness, the example refers to the case of a binary Y, where coarsening corresponds to missingness, but the framework is also applicable in the general categorical setting.

For this categorical setting, we characterize cases where SI makes parameters not only identifiable, but is also testable. Besides the investigation of the testability of SI, a main contribution of this paper is the construction of the likelihood-ratio test for this assumption. For this purpose, we give the hypotheses, illustrate the sensitivity of the test statistic with regard to the deviation from the null hypothesis and study the asymptotic distribution of the test statistic to obtain a decision rule in dependence of the significance level. Straightforwardly, a test for a specific pattern of gSI is constructed.

Our paper is structured as follows: In Section 2 we introduce the technical framework and the running example based on the German Panel Study "Labour Market and Social Security" (PASS), which we also use for the illustration of both assumptions, CAR and SI, as well as gCAR and gSI, in Section 3. After sketching the crucial argument of identifiability issues and our estimation method as well as showing how the generally set-valued estimators may be refined by assuming CAR/gCAR or SI/gSI in Section 4, the obtained estimators are used to discuss the testability of both assumptions in Section 5. The likelihood-ratio test for SI is developed and then illustrated for the running example in Section 6, where the generalized view on subgroup independence is used to extend this hypothesis test to a more flexible version, including a test on partial information, in Section 7. All results of this paper are given for a general categorical setting, but the running example refers to the illustrative case of binary data. To emphasize the general applicability of our approach, we briefly discuss further examples in Section 8, also addressing potential limitations. Finally, Section 9 concludes with a summary and some additional remarks.

2. Coarse data: the basic viewpoint

Before we discuss the running example, let us explicitly formulate the technical framework in which our discussion of the coarsening assumptions, the estimation of parameters and the construction of the likelihood-ratio test is embedded. We approach the problem of coarse data in our categorical setting by distinguishing between a latent and an observed world: Let $(x_1, y_1), \dots, (x_n, y_n)$ be a sample of n independent realizations of a pair (X, Y) of categorical random variables with sample space $\Omega_X \times \Omega_Y$. Our basic goal consists of estimating the probabilities $\pi_{xy} = P(Y = y|X = x)$, where Y is regarded as response variable and X as covariate. Since the values of Y unfavorably can be observed partially, i.e. subsets of Ω_Y instead of single elements may be observed, this variable is part of the latent world. Instead, we only observe a sample $(x_1, \mathfrak{y}_1), \dots, (x_n, \mathfrak{y}_n)$ of n independent realizations of the pair $(X, \mathcal{Y})$, where the random object $\mathcal{Y}$ with sample space $\Omega_{\mathcal{Y}} = \mathcal{P}(\Omega_Y) \setminus \{\emptyset\}$ constitutes the observed world. A connection between both worlds, and thus between the probabilities π_{xy} and $p_{x\mathfrak{y}} = P(\mathcal{Y} = \mathfrak{y}|X = x)$, is established via an observation model, governed by the coarsening parameters $q_{\mathfrak{y}|xy} = P(\mathcal{Y} = \mathfrak{y}|X = x, Y = y)$ with $\mathfrak{y} \in \Omega_{\mathcal{Y}}$, $x \in \Omega_X$ and $y \in \Omega_Y$. Throughout the paper, we not only assume that the coarsening depends on the individual i $(i = 1, \dots, n)$ via the values x and y exclusively, but also require distinct parameters in the sense of Rubin (cf. e.g. [17]) as well as error-freeness,[2] i.e. $\mathfrak{y} \ni y$, explicitly excluding the case of misclassification.

An essential part of our argumentation is based on comparing the dimensions of the parameter space of the latent world Θ_{lat} and the parameter space of the observed world Θ_{obs}. While $\theta_{lat} \in \Theta_{lat}$ describes the latent variable distribution π_{xy}

[2] This implies that Y is a selector of $\mathcal{Y}$ (in the sense of e.g. [20, p. 43]).