On the Analysis of Epiontic Data: A Case Study

Georg Schollmeyer GEORG.SCHOLLMEYER@STAT.UNI-MUENCHEN.DE
Hannah Blocher HANNAH.BLOCHER@STAT.UNI-MUENCHEN.DE
Christoph Jansen CHRISTOPH.JANSEN@STAT.UNI-MUENCHEN.DE
Thomas Augustin THOMAS.AUGUSTIN@STAT.UNI-MUENCHEN.DE
Department of Statistics, Ludwig–Maximilians–Universität München, Munich, Germany

Within the IP community and in particular at the ISIPTA conferences it is often emphasized that it is important to distinguish between *ontic* and *epistemic* data imprecision when analyzing interval-valued or set-valued data. In the *epistemic* view, a set-valued data point represents an imprecise observation of a precise but not directly observable variable of interest. In contrast, in the *ontic* view, set-valued data are understood as a precise observation of something that is "imprecise" only in the sense of being mathematically modelled by a set and there is in fact no imprecision at all. Therefore, an alternative wording for such data could be to speak about (a special case of) *non-standard data*. In this contribution, we deal with non-standard data that are *partially ordered sets*. An example is [3], where students were asked for their choices between foreign universities for their semester abroad. The students could say they prefer one university over another or vice versa. Additionally, they could also answer that they have no preference between the two given universities. In this situation, it is natural not to assume epistemic data imprecision, because the observed incomparability between two universities does not mean that the asked student does not know or that she conceals her preference or that she forgot to answer the question about preference. Instead, one can, de facto, say that the student has no preference between the given universities at all. (Note that it is also not natural to simply assume indifference between the universities.) In this specific sense, such data – if precisely observed – are plain and simply *non-standard data*. However, in [3], during the data collection, some interviewers simply forgot to ask some of the pair comparisons for some students. This shows that, of course, epistemic data imprecision is also an issue within the data analysis of non-standard data. As far as the knowledge of the authors goes, explicitly addressing non-standard data (or ontic data) under epistemic uncertainty – which we call *epiontic* data for short, here –has currently not, or not intensively, been investigated. In this contribution, we address this important problem situation for the special case of *partial* order data, understood as non-standard data (where incomparabilities are *not* epistemic) under additional epistemic data imprecision, analyzed with the methodology of data depth (cf., the analyses in [1, 2] for the situation without additional epistemic data imprecision). Concretely, we will, in the first place, analyze the generalized Tukeys depth [cf. 1, p. 22] that measures the centrality or outlyingness of a data point w.r.t. the observed data cloud. We will analyze

i) A situation of coarsening at random (CAR): In this situation, constructing (nearly) unbiased and consistent estimators for the true underlying depth is straight-forward and computationally simple.

ii) A non-CAR situation that is of special interest to the ISIPTA community: In this case, we already have some first results concerning the computation of the cautious data completion that would be, in our case, a whole family of depth functions (or their corresponding lower and upper envelopes) that arises as the (ontic) input data range over all possible data that are compatible with what is observed.

Analyses using generalized *Tukey's depth* are based only on the marginal distributions of pairwise comparisons (how often a university is preferred over another). Therefore, and more importantly, an analysis based on depth functions that focus on more aspects of the structure of the partial orders (i.e., the distribution of partial orders as a whole), see, e.g., the ufg depth used in [2], are of additional interest as well. However, this raises the issue of computability.

References

[1] H. Blocher, G. Schollmeyer, and C. Jansen. Statistical models for partial orders based on data depth and formal concept analysis. In *IPMU 2022*, pages 17–30. Springer, 2022.

[2] H. Blocher, G. Schollmeyer, C. Jansen, and M. Nalenz. Depth functions for partial orders with a descriptive analysis of machine learning algorithms. *Conditionally accepted for presentation at ISIPTA 2023*, 2023.

[3] R. Dittrich, R. Hatzinger, and W. Katzenbeisser. Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *J R Stat Soc Ser A Stat Soc: Series C*, 47(4):511–525, 1998.