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 (LMU Munich)

Data (im)precision

- **Epistemic data** (imprecision): Imprecise/coarsened/censored observation of something that is actually precise (e.g., non-response to the question about income in social surveys).
- **Ontic data** (no imprecision): Exact observation of something (seemingly) imprecise.
  More concretely, a precise data point is modeled by a set(, e.g., an interval can represent the lifespan of e.g., Wolfgang Amadeus Mozart, who lived from 27.01.1756 to 5.12.1791).
- **Nonstandard data:** Data with a non-standard underlying scale of measuremnt. Ontic data are often non-standard in this sense.
- **Epiontic data:** Ontic (or nonstandard) data that are observed with additional epistemic imprecision. E.g., Aribert Reimann (German composer, still alive) will have lived from 04.03.1936 to ? (This is an ontic interval-valued observation where the right endpoint of the interval is censored and therefore subject to epistemic imprecision.)

Our case

- 303 students were asked for their choices between 6 foreign universities for their semester abroad.
- Within pair comparisons, they could prefer one university over another or vice versa.
- They could also explicitly state that they have no preference between universities (i.e., incomparibabilty of universities, not to be confused with indifference, is allowed). Therefore we have 303 **partial orders** as **non-standard** data points.
- By accident, some of the pair comparisons were not asked. This constitutes **additional epistemic data imprecision**.

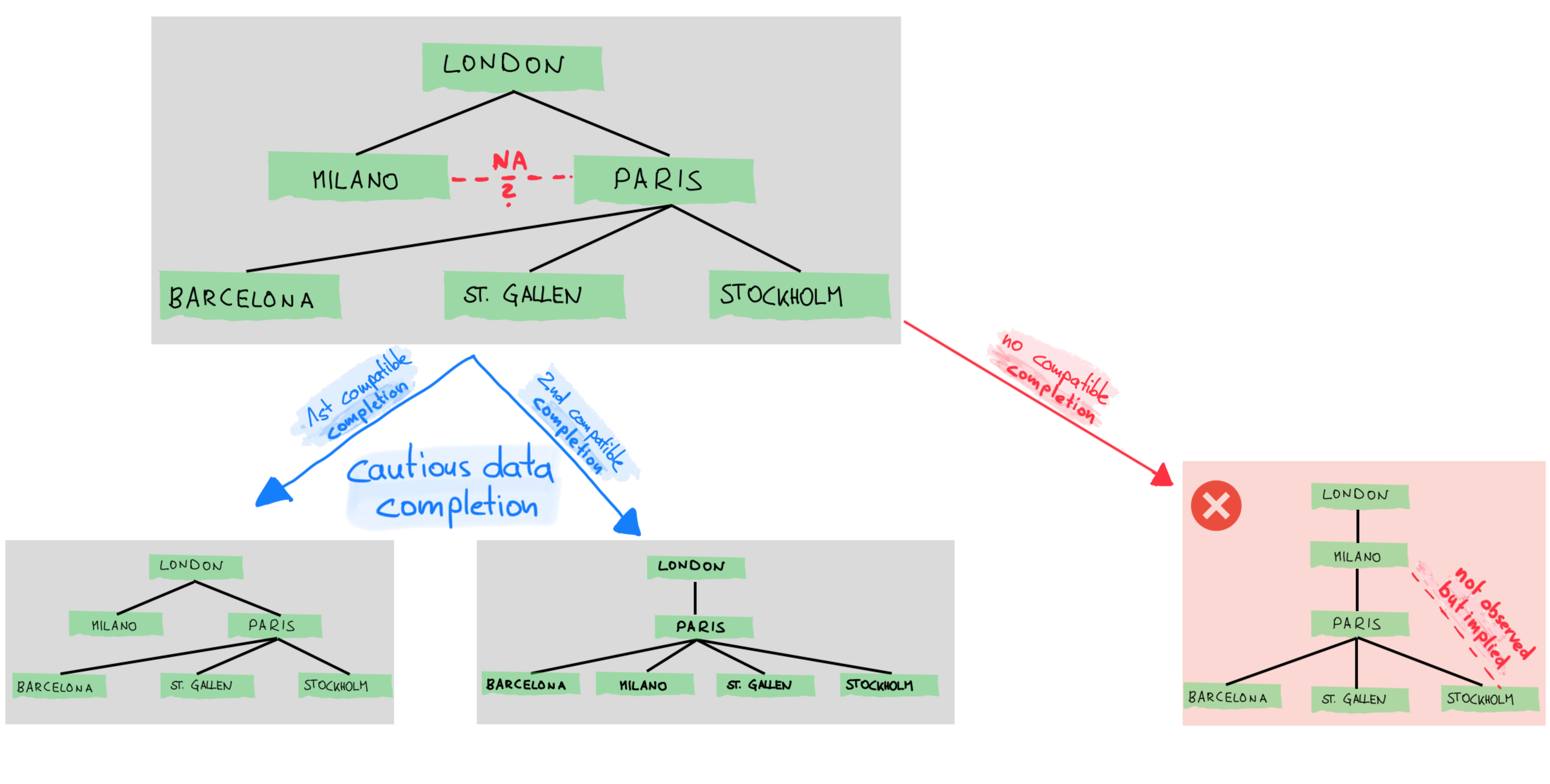
Our approach

- Used methodology: **Data depth** functions for poset-valued data.
- Data depth functions measure the **centrality** or **outlyingness** of data points w.r.t. a data cloud (or underlying probability law). Here we analyse Tukey's depth, cf. [2], [1].
- Concretely, we used the generalized **Tukey's depth** for poset-valued data:

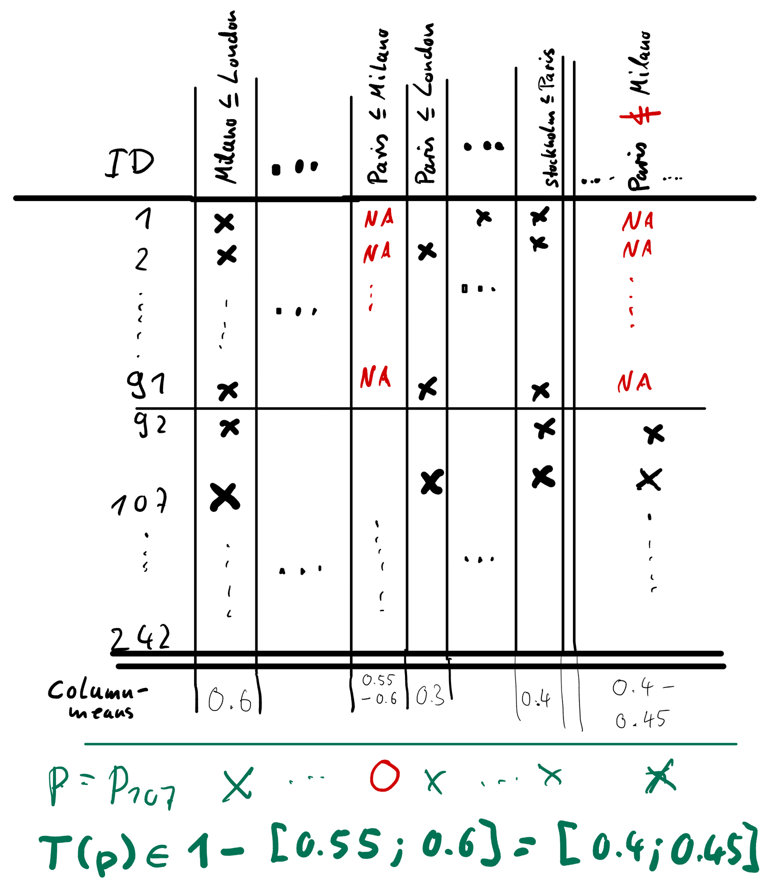
$$\text{Tukey's depth of a poset } p: \quad \mathfrak{T}(p) := 1 - \max\left\{ \sup_{(a,b)\in p} P(\{q \text{ poset } \mid (a,b)\notin q\}),\ \sup_{(a,b)\notin p} P(\{q \text{ poset } \mid (a,b)\in q\}) \right\}. \tag{1}$$

- Handling of additional epistemic uncertainty:
  - Tukey's depth is a simple function in the **columnwise proportions of crosses** (see (1) and the cross table below).
  - Analysis under the assumption of **coarsening at random (CAR)** ([4]).
  - Analysis without additional modeling assumptions: Computing the **cautious data completion (CDC)** ([3]).
  - The simple formula for the generalized Tukey's depth allows the exact computation of the cautious data completion or at least a conservative approximation (depending on the coarsening process).

The cautious data completion (CDC) and the coarsening at random (CAR) approach



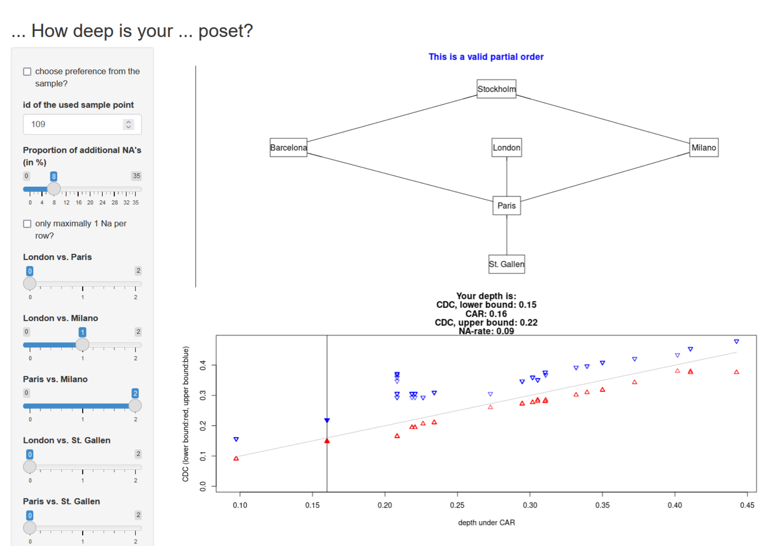
Conceptual scaling:
Crosses both for $\leq$ and $\not\leq$:



Analysis:

- CAR: Simply ignore NA's.
- CDC:
  Only one NA-column in the $\leq$-part: Exact solution (One knows how to replace the NA's with crosses/non-crosses).
  Otherwise: Exact lower bound and conservative upper bound.

For results, see our Shiny app:



Future research

- Cautious Data Completion for other depth functions like the $\mathfrak{ufg}$ depth, cf. [1].
- More advanced handling of the CAR case (e.g., imputation techniques that account for dependencies between edges within posets).
- How to handle responses that are actually not posets, e.g., because the given relation is not transitive?
- In particular: Is it a problem that for a completely observed relation it is more probable that one can detect that is actually not a poset, compared to a response that is only partially observed?
- Analyse further data sets.
- For the university data: Try it out yourself! (See QR Code.)

https://tinyurl.com/epiontic
How deep are your preferences?

[1] H. Blocher, G. Schollmeyer, C. Jansen, and M. Nalenz. Depth functions for partial orders with a descriptive analysis of machine learning algorithms. *Forthcoming in: ISIPTA '23.*
[2] H. Blocher, G. Schollmeyer, and C. Jansen. Statistical models for partial orders based on data depth and formal concept analysis. . In: Ciucci, D.; Couso, I.; Medina, J.; Slezak, D.; Petturiti, D.; Bouchon-Meunier, B.; Yager, R.R. (eds): IPMU *Communications in Computer and Information Science*, vol 1602, Springer, 2022.
[3] T. Augustin, F. Coolen, G. de Cooman, and M. Troffaes (eds): Statistical inference. In: *Introduction to Imprecise Probabilities*, Wiley, Chichester, 2014.
[4] D. Heitjan and D. Rubin. Ignorability and coarse data. *The Annals of Statistics*, 19(4):2244–2253, 1991.