Bull. Aust. Math. Soc. **111** (2025), 184–185
doi:10.1017/S0004972724001060

STATISTICAL ANALYSIS ON COMPLEXLY STRUCTURED DATA

PROSHA RAHMAN

(Received 3 October 2024; first published online 25 November 2024)

2020 *Mathematics subject classification*: primary 62R07; secondary 68T09.

Keywords and phrases: statistical inference, symbolic data analysis, model-based inference, missing data problems.

Modern data are characterised by their large volume and messy features. Traditional statistical methods, while theoretically valid, are frequently computationally intractable for large and incomplete data sets. Statisticians will often manipulate a data set so as to reduce its size and restore its rectangular structure through artificial data imputation. Elementary statistical methods cannot provide valid inference of the newly constructed complex data. We address inference over these complexly formed data sets in two sections: performing inference over aggregated data and estimating parameters from imputed data.

In the first section, we discuss inference on complexly aggregated data using results in symbolic data analysis. The discussion opens by examining the aggregation of data sets into so called *symbols*, and subsequently showing the convergence of these symbols. Our examination also introduces distribution-valued symbols which provide a granular form of the existing coarse symbolic variables.

Our analysis then turns to model-based inference with symbolic data. This section opens with an application to network traffic using interval-valued symbols of unidirectional data. We provide consistency in estimation and bounds on information loss under aggregation, and identify models that are sufficiency invariant under aggregation. The consistency results are extended to a generic setting when considering inference with respect to a single or multiple symbols.

In the second part, we address a missing data problem through the lens of ordinary least squares (OLS). Large data sets often contain missing elements, due to pragmatic sampling choices or incomplete collection methods. We synthetically construct the pseudocensus of the population through the common semiparametric

Thesis submitted to the University of New South Wales in February 2024; degree approved on 6 June 2024; supervisors Scott Sisson and Boris Beranger.

weighted K-nearest neighbours algorithm. The resulting OLS estimator is shown to be biased and we subsequently provide two methods of bias correction using the internal weights of the imputation algorithm and a bias-correction coefficient. The estimator is also shown to be consistent. These results are validated in some simulated analyses.

Some of this research has been published in [1].

Reference

[1] P. Rahman, B. Beranger, S. Sisson and M. Roughan, 'Likelihood-based inference for modelling packet transit from thinned flow summaries', *IEEE Trans. Signal Inform. Process. Netw.* **8** (2022), 571–583.

PROSHA RAHMAN, School of Mathematics and Statistics,
University of New South Wales, Kensington, New South Wales 2033, Australia
e-mail: p.a.rahman@unsw.edu.au

https://doi.org/10.1017/S0004972724001060 Published online by Cambridge University Press