Treatment of Missing Data in Bayesian Network Structure Learning: The Case of Social Scientific Healthcare Data

Xuejia Ke[1,2] **Katherine Keenan**[2] **V. Anne Smith**[1]

[1]School of Biology, University of St Andrews, St Andrews, KY16 9TH, UK
[2]School of Geography and Sustainable Development, University of St Andrews, St Andrews, KY16 9AL, UK

1 EXTENDED ABSTRACT

Multimorbidity has a great impact on human health. The burden of multimorbidity is expected to increase globally as population's age, and is a huge public health challenge. Researchers have used a variety of methods to unpick the complexity of combinations of diseases, and identify clusters and risk factors [Hassaine et al., 2020, Si et al., 2021]. Among these, as a flexible statistical tool for encoding probabilistic relationships with directed acyclic graphs (DAGs) [Heckerman et al., 1995], Bayesian networks (BNs) have great potential to tackle such complex problems.

Compared with other fields of studies, for instance, experimental biological systems, missing data are more pervasive in health survey data. There are plentiful causes of missing data, including item missingness, e.g., unanswered questions in questionnaires, data entry errors, or subject missingness, e.g., patients dropping out in longitudinal research, missing samples. Missing data not only reduce overall statistical power and precision, but can lead to biased inferences in subsequent data analysis [Sterne et al., 2009]. Taking a popular method complete case analysis (e.g., undertaking analysis only on those cases without any missing data) as an example, its statistical power and precision would be inevitably reduced because of the decreased sample size.

Based on the different processes leading to the missingness, every missing data pattern can be generally classified into three categories – missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [Rubin, 1976]. This nomenclature is widely used in statistical data analysis and is also referred to as the missing data mechanisms. MCAR occurs if the missingness is unrelated to both unobserved and observed variables. Data are said to be MAR if the missingness is related to observed variables but not to any unobserved variables given the observed ones. MNAR is the most complicated because its missingness relates to both unobserved and observed variables [Rubin, 1976]. These three patterns cause different levels of risks of bias in data analysis. For instance, the application of complete case analysis in MAR and MNAR data would yield more biased estimates than MCAR [Schafer and Graham, 2002].

Multiple imputation by chained equations (MICE) is a popular multiple imputation method used in social science data. It is designed to impute missing data values under the missing data assumption MAR [Raghunathan et al., 2001]. Compared to single imputation, multiple imputation methods are less biased because they take account of the uncertainty of the missing data by combining multiple predictions for each missing value. MICE uses a divide and conquer approach to replace missing values for all variables in the data set: it focuses on one variable at a time and makes use of other variables to predict the missing values in that focused variable. In epidemiology and clinical research, multiple imputation is believed to have substantial potential for improving the validity of quantitative analysis results for MAR data. However, such methods are not suitable for MNAR data, and it is still an outstanding task to adapt MICE for dealing with MNAR data [Sterne et al., 2009].

Learning BN structure from incomplete data is quite challenging. Depending on the missing data mechanisms (e.g., MNAR or MAR), learning would be biased if we simply delete incomplete observations. If we include every missing value in the learning process, the required computational resources for considering all possible completions of the data set and related computations would grew exponentially with the amount of missing values [Scutari, 2020].

The structural expectation-maximization (SEM) algorithm makes BN structure learning from incomplete data computationally feasible by changing its search space to be over structures rather than parameters and structures. SEM completes and perfects the data in an iterative way, then applies the standard structure learning procedures to the completed data [Scutari, 2020]. The framework of SEM was first proposed by Friedman Friedman [1997]. His simulation results suggest that SEM has potential to handle data involving missing values and hidden variables.

Accepted for the 16th Bayesian Modelling Applications Workshop (@UAI2022) (BMAW 2022).

Here, we evaluate methods for addressing incomplete data using a simulation framework. We simulate multiple incomplete data sets, including three different missing data mechanisms, various number of variables and amounts of missing data. We then evaluate and compare the performance of MICE and SEM with each other and with the standard expedient of using only samples without missing data, by comparing their resulting network structures with the original network structure. We find that applying either method (MICE or SEM) provides better structure recovery than only using complete cases, and SEM in general outperforms MICE. This finding is robust across missingness mechanisms, number of variables, levels of data points and amount of missing data.

We then apply the best working method, SEM, to the United States Health and Retirement Study, a representative study of adults aged 50+, including self-reported and nurse-collected biomedical data collected in 2016. We subset the original dataset to include key demographic information (e.g., race, age and education), cognitive and physical examinations (e.g., self-assessed memory, BMI), doctor diagnosed diseases (e.g., diabetes) and laboratory data (e.g., HbA1c measurements). This covers 5726 observations, in which only 1955 cases are complete. Among all variables, the maximum missing percentage is 33.1%. We apply SEM to the subset 100 times from different seeds and get the average network based on the arc strength of each learned structure. We use the completed partially directed acyclic graph (CPDAG) of each structure when calculating arc strengths. Then we use hierarchical divisive clustering method to detect the densely connected variables in the learned average network.

We investigate the interactions among presence and treatment for several chronic diseases and their associations with individual's demographic and socioeconomic factors. We find that common metabolic conditions are clustered, such as heart conditions, high blood pressure, total cholesterol level and obesity. The treatments for diabetes, high cholesterol level and heart conditions are closely linked to each other. Our network shows strong relationships between cancer, arthritis and lung diseases. We find a direction connection between smoking and lung disease, as would be expected. Our analysis also highlights potential areas of investigation. We use total score on telephone interview for cognitive status measurement (TICS-M) to assess cognitive impairment. We find that cognitive impairment is closely associated with diabetes, education, age and race, but stands alone from self-assessed memory decline. Additionally, the treatment behaviors have strong interactions with other chronic diseases, for example, the treatment for diabetes and high cholesterol level is closely associated with high blood pressure. These unexpected associations could potentially be further explored in future analysis.

Acknowledgements

The authors acknowledge the Research/Scientific Computing teams at The James Hutton Institute and NIAB for providing computational resources and technical support for the "UK's Crop Diversity Bioinformatics HPC" (BBSRC grant BB/S019669/1), use of which has contributed to the results reported within this work.

References

Nir Friedman. Learning belief networks in the presence of missing values and hidden variables. In *Fourteenth International Conference on Machine Learning (ICML).*, pages 125–133, 1997.

Abdelaali Hassaine, Gholamreza Salimi-Khorshidi, Dexter Canoy, and Kazem Rahimi. Untangling the complexity of multimorbidity with machine learning. *Mechanisms of Ageing and Development*, 190:111325, 2020. ISSN 0047-6374. doi: https://doi.org/10.1016/j.mad.2020.111325.

David Heckerman, Dan Geiger, and David Maxwell Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20: 197–243, 1995.

Trivellore E Raghunathan, James Lepkowski, John H Van Hoewyk, and Peter W Solenberger. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27:85–95, 2001.

Donald B Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

Joseph L Schafer and John W Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2): 147, 2002.

Marco Scutari. Bayesian network models for incomplete and dynamic data. *Statistica Neerlandica*, 74:397–419, 2020.

Yuqi Si, Jingcheng Du, Zhao Li, Xiaoqian Jiang, Timothy Miller, Fei Wang, W Jim Zheng, and Kirk Roberts. Deep representation learning of patient data from electronic health records (ehr): A systematic review. *Journal of Biomedical Informatics*, 115:103671, 2021.

Jonathan AC Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ: British Medical Journal*, 338: b2393, 2009.