Matching with missing values for clinical data

M2 Research Internship

Julie Josse (Inria Montpellier), Erwan Scornet, (École polytechnique, IPP)

Key words: causality, average treatment effects, semi-parametric efficiency, supervised learning with missing values.

1 Scientific context

In machine learning, there has been great progress in obtaining powerful predictive models, but these models rely on correlations between variables and do not allow to understand the underlying mechanisms or how to intervene on the system in order to achieve a certain goal. Causality is a fundamental concept to identify levers of action, to formulate recommendations and to answer the following questions: "what would happen if" we had acted differently? Many methods to discover causal structures in data and to estimate the effect of an intervention on a response have been suggested in recent years and have impacts in many areas such as health and public policies. The latest developments also show the impact of causality on improvement of the stability of predictive models.

Inferring causal effects of treatments is central to many analyses but is far from being straightforward. On the one hand, randomized controlled trials/A-B tests are the gold standard for estimating effects because the distribution of controls and treated is asymptotically balanced so that a simple difference in means can be a consistent estimator. However, it is not always possible to set up clinical trials for ethical or cost reasons. On the other hand, large observational data are readily available but can conflate confounding effects with the treatment of interest.

Under assumptions such as unconfoundedness, it is possible to estimate a causal treatment effect from observational data. In practice, methods such as matching, inverse propensity weighting (IPW), or augmented IPW (AIPW) are used [4]. Matching techniques are by far the most used and the most popular because they are extremely intuitive. Matching was for instance used by [2] to determine the efficiency of Covid19 vaccines and this has determined all vaccination policies in the world.

Matching techniques pair treated and non-treated patients to assess the effects of treatment on similar individuals. In its simplest instance, each treated unit is matched to the nearest control unit (one-to-one matching) which requires to define an appropriate distance, as well as a decision to exclude some units when the distance is greater than a threshold. Many other techniques exist such as Exact matching, Mahalanobis

distance matching, Propensity score matching, one-to-many matching, and subclassification matching, etc. [3, 7, 1] . Matching methods are often based on nearest neighbor approaches but recent papers are based on optimization [9], optimal transport, etc.

Despite its popularity and usefulness, there are no studies on the theoritical behavior of matching with missing values, and few empirical studies while missing data are ubiquitous in practice. Managing missing data in causal inference requires coupling causal assumptions with assumptions about missing data mechanisms. It thus requires to establishing new conditions of identifiability and deriving associated estimators in the spirit of [6] who suggested AIPW estimators using two random forest adapted to missing data [8]. Other solutions consider using latent variable models [5].

The objective of this internship is to perform a theoretical analysis of matching with missing data (identifiability, study of the bias and asympotic variance of the estimator with missing data, possibly obtain guarantees with finite sample) but also to implement the developed methods and to test them on simulated data and medical data (already available).

2 Application context and objective: decisions in medical emergencies

In the group, we have different medical collaborations. One of the oldest collaboration is with the Traumabase group of APHP (Public Assistance - Hospitals of Paris) on polytraumatized patients. This project is mature in that we are testing real-time cell phone applications in the ambulance to help clinicians make decisions.

Major trauma denotes injuries that endanger the life or the functional integrity of a person. The WHO has recently shown that major trauma, –including road-traffic accidents, interpersonal violence, falls– remains a world-wide public-health challenge and a major source of mortality and handicap. An effective and timely management of trauma is crucial to improve outcomes, as delays or errors entail high risks for the patient.

Traumabase. To improve decisions and patient care in emergency departments, 30 French Trauma centers are collecting detailed clinical data from the scene of the accident to the exit of the hospital. The resulting database, the Traumabase, comprises to date 30 000 trauma admissions, and is permanently updated. The data are of unique granularity and size in Europe. However, they are highly heterogeneous, with both categorical (sex, type of illness...) and quantitative (blood pressure, hemoglobin level...) features, multiple sources, and many missing data. In fact 98% of the individuals have missing values. The cause of missing information is also coded, such as technical hurdles with the measurement, or impossibility due to the severity of the patient's state. Modeling is challenging, but with great potential benefits. The goals are to predict outputs such as intracranial hypertension but also to give recommendations. Such recommendations call for causal interpretations, based on counter-factual reasoning such as: Would the patient have survived had transfusion been done earlier? What is the effet of tranexomic acid/noradrenaline/... on mortality for head trauma?

3 Laboratory - contact

The Premedical (Precision Medicine by Data Integration and Causal Learning) team[1], is a recent Inria-Inserm team located in Montpellier. It is an interdisciplinary team composed of statisticians, biostatisticians, machine learners, and clinicians. Premedical develops methods for optimal treatment policy (drug efficacy, who gets treated and when, etc.) from heterogeneous data (clinical trials, observational data) that come with methodological challenges. In particular, Premedical develops methods for causal inference, statistical learning, management of missing data, federated learning, etc. Premedical holds the missing data and causality research group[2] and has created a taskview on causal inference methods.

The successful candidate will join the team at INRIA, and a broad community of experts in the fields of Statistics, Machine Learning and Artificial Intelligence. This is a dynamic environment with many students, PhD students, Post-docs and researchers. The group has tight collaborations with international researchers abroad and at CMAP Polytechnique and with other INRIA teams. The candidate will also be able to participate in the activities of the Inserm team, Idesp, specialized among others on respiratory diseases such as asthma and also specialists in the exposome.

The candidate will be supervised by both Julie Josse (expert in missing values causal inference) and Erwan Scornet (expert in Random Forest, Statistical Learning, Missing Values). Julie Josse has many international connections in causal inference (she was invited to the semester on causality in Berkeley, to the Rousseeuw prize in Belgium, etc.) and often sends her PhD students to do research internships abroad, in particular with the Department of Statistics at Stanford University with whom she has many connections.

We are looking for excellent candidates, highly motivated, with background knowledge in mathematics, statistics /machine learning and interested by interdisciplinary research and collaboration. We will focus on both the theoretical and practical aspects including implementation.

Practical information. We are offering an internship position at INRIA (Montpellier) for a duration of six months, ideally starting in march-april 2023 (duration and starting dates can be discussed). The internship can be pursued by a PhD thesis, depending on the overall quality of the internship.

Qualifications:

- Master in Statistics, Machine Learning, Biostatistics, Data-science or related fields
- Strong statistical computing skill
- Excellent writing and communication skills
- Background on causal inference is a plus

[1] https://team.inria.fr/premedical/

[2] https://misscausal.gitlabpages.inria.fr/misscausal.gitlab.io/

Required application materials:

- CV
- Academic performance in recent years
- Short cover letter describing your interest

Email your application to julie.josse@inria.fr and erwan.scornet@polytechnique.edu

References

[1] Alberto Abadie and Guido W. Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

[2] Noa Dagan, Noam Barda, Eldad Kepten, Oren Miron, Shay Perchik, Mark A. Katz, Miguel A. Hernán, Marc Lipsitch, Ben Reis, and Ran D. Balicer. Bnt162b2 mrna covid-19 vaccine in a nationwide mass vaccination setting. *New England Journal of Medicine*, 384(15):1412–1423, 2021.

[3] Guido W Imbens. Matching methods in practice: Three examples. *J Hum Resour*, 50:373–419, 2015.

[4] Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, Cambridge UK, 2015.

[5] N. Kallus, X. Mao, and M. Udell. Causal inference with noisy and missing covariates via matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6921–6932, 2018.

[6] Imke Mayer, Erik Sverdrup, Tobias Gauss, Jean-Denis Moyer, Stefan Wager, and Julie Josse. Doubly robust treatment effect estimation with missing attributes. *Ann. Appl. Statist.*, 14(3):1409–1431, 2020.

[7] Elizabeth A. Stuart. Matching Methods for Causal Inference: A Review and a Look Forward. *Statistical Science*, 25(1):1 – 21, 2010.

[8] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

[9] José R. Zubizarreta. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107(500):1360–1371, 2012.