PhD positions/Postdoc: Causal inference and policy learning for personalized medicine

Julie Josse INRIA

Key words: causality, heterogeneous treatment effects, double robustness, missing values, combining RCT and observational data, distributional shift, policy learning, dynamic treatment allocation, reinforcement learning.

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. The concept of causality is fundamental to have 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 also 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, they may lack external validity due to restrictive inclusion and exclusion criteria. On the other hand, large observational data are often more representative of a target population but but can conflate confounding effects with the treatment of interest. The simultaneous availability of observational and experimental data is both an opportunity and a theoretical and methodological challenge. Exploiting both sets of data can serve different purposes such as better adoption by the medical community of certain (advanced) techniques used to estimate the effects of treatment on patients (by comparing the results obtained in an RCT with the RWE), better design of RCTs so that they are more representative of the patient population that may benefit from the treatment, prediction of the effect of treatment on new populations, etc.

[2] reviewed the main methods, which can either transport the estimated causal effect in an RCT to an observational study while taking into account the distributional shift (IPSW, g-formula, AIPSW, calibration weighting, etc), or improve the estimate of the

conditional average treatment effect while correcting for confounding factors not measured in the observational study. However, these methods still have many shortcomings and there are stil many challenges to adress. First, the problem of missing values is exacerbated when aggregating data of different sources. There may be sporadic missing data in the RCT and in the observational data, but there may also be so-called systematic missing data when a variable is not available in either the RCT or the observational data. The first case already requires establishing new conditions of identifiability with missing data and deriving estimators that handle missing values in the spirit of [6, 7, 4] who suggested AIPW estimators using two random forest adapted to missing data. As for the second case, depending on which variables are missing, it may be necessary to turn to sensitivity analyses because the hypotheses of ignorability will no longer be verified. Both structural causal models [1] and potential outcomes can help takling these issues [3]. Then, most of the work to integrate the two data sources has been carried out in a framework where contrasts of the potentials outcome are estimated and not in a framework of policy learning, where the aim is to learn optimal treatment assignment rules (who should be treated and who should not?) and one can expect that both sources can also be leveraged when treatment assignment policies vary across time due to time varying covariates. One can cite the recent work of [5] where they combine different sources (not RCT) for policy learning.

2 Application context and objective: decisions in medical emergencies

In the group, we have different collaboration with different hospitals. One of the oldest collaboration is with the Traumabase group of APHP (Public Assistance - Hospitals of Paris) on polytraumatized patients described below, but on the subject of combining RCT and observational data, we have several collaborations in progress and to come, on the treatment of cerebrovascular accidents, breast cancer, and treatments for allergies so the work will be done with one of these groups.

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 major source of mortality and handicap Effective and timely management of trauma is crucial to improve outcomes, as delays or errors entail high risks for the patient.

To improve decisions and patient care in emergency departments, 20 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 20 000 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 on mortality for head trauma? the effet of the adminsitration of noradrenaline, etc.

3 Laboratory - contact

The (https://misscausal.gitlabpages.inria.fr/misscausal.gitlab.io/students.html) missing data and causality research group has funding opportunities for PhD students (3 years) and Post-docs (2 years) that will start in September 2021 (it can start later as well) . We are expanding our group. Prospective graduate students candidates are also invited to apply for short term (6 months) research internship.
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 of international renown with many students, PhD students, Post-docs and researchers. The group has tight collaborations with researchers at CMAP Polytechnique and with other INRIA teams such as Parietal as well as with other international researchers in causal inference. The candidates will also have excellent opportunities to collaborate with other researchers. This position will provide the candidate with an unique opportunity to carry out state-of-the-art academic research and also to join an interdisciplinary collaboration project bringing together mathematical, methodological, technological, cognitive and medical expertise.

We are looking for excellent candidates, highly motivated, self-driven postdoctoral fellow 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.

Qualifications:

- Recent (doctoral) degree in Statistics, Biostatistics or related fields
- Background on causal inference and survival analysis is a plus
- Experience with machine learning and high-dimensional statistics
- Strong statistical computing skill
- Excellent writing and communication skills

Required application materials:

- Updated CV
- Complete contact information for two references.

- Short cover letter describing their past research experience, career goals and a statement of future research interest (1 page)
- Two or three relevant publications or manuscripts for postdoc applications.

Interested graduates (undergraduates) should apply as early as possible since the positions will be filled when suitable candidates are found.

Email your application to julie josse at inria.fr

References

[1] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

[2] Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. *arXiv preprint arXiv:2011.08047*, 2020.

[3] Guido Imbens. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. Technical report, National Bureau of Economic Research, 2019.

[4] 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.

[5] Masahiro Kato, Masatoshi Uehara, and Shota Yasui. Off-policy evaluation and learning for external validity under a covariate shift. *Neurips*, 2020.

[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] 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.