Postdoc/Phd Thesis/ Internship offers: Causal inference and policy learning for personalized medicine

Julie Josse and Karim Lounici CMAP, Polytechnique INRIA

Key words: causality, latent variable models, low-rank matrix estimation, heterogeneous treatment effects, semi-parametric, missing values, oracle inequalities.

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 especially with data that are observational and not coming from an experimental design. One aim of the analysis is to assess the effect of a treatment on an outcome such as the survival while adjusting for the effects of the covariates. If for each subject, one has the pairs of outcomes, i.e. one under treatment and the other one under control, then the causal effect could be easily estimated (with the difference between the outcomes means). However, in observational data, both potential outcome can not be observed simultaneously for each subject. The causal-effect literature developed the *potential-outcomes* framework [2] which can be thought of as a missing data problem where we try to infer about the missing potential outcomes. Classical methods include the inverse probability weighting methods, which consists in performing a predictive model, such as a logistic regression of the treatment on the covariates and then using the ”scores” as weights to correct the model which explain the outcome as a function of the treatment. More advances models have been suggested including double robust methods. Beyond average treatment effect, learning heterogeneous treatment effect estimate the impact of the treatment for each observation and call for personalized recommendations.

Recent methods include causal forests, R-learners, etc and are also at the root of learning policies strategies [1, 6, 7, 5, 3, 4].

The mathematical formalization of this problem is the subject of intense research on both methodological and theoretical aspects. A promising approach is to see this problem as a problem of estimating specific functionals on a matrix space satisfying certain low complexity constraints. We will build several statistical inference methods based on recent progress on the estimation of high-dimensional matrices. Our main goal will be to understand the impact of several key parameters of the problem (missing value rate, confounders rank) on the estimation of treatment effect.

2 Application context and objective: decisions in medical emergencies

Our work is motivated by a public-health application with the Traumabase group from APHP (Public Assistance - Hospitals of Paris) on polytraumatized patients. 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?

3 Laboratory - contact

The statistics research group has 2 funding opportunities for PhD students and Post-docs. Prospective graduate students candidates are also invited to apply for short term (6 months) research internship.

The positions will be based at the applied mathematics department of Ecole Polytechnique CMAP http://www.cmap.polytechnique.fr/spip.php?rubrique141 and at INRIA Saclay.

Graduate internship: funds for 6 months (with the possibility to pursue a PhD afterwards).

Funds for 1 PhD student (3 years).

Post-doc: the funds for this post are available for 3 years. Salary is competitive.

Interview date: as soon as possible after shortlisting.

We are looking for highly motivated, self-driven postdoctoral fellow with background knowledge in mathematics, statistics /machine learning and interested by interdisciplinary research and collaboration. The successful candidate will join the missing values and causality research group at CMAP/INRIA and a broad community of international experts in the fields of Statistics, Machine Learning and Artificial Intelligence. 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.

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)

Email your application to Julie Josse at polytechnique.edu and Karim Lounici at polytechnique.edu

References

[1] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[2] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[3] 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.

[4] Xinkun Nie, Emma Brunskill, and Stefan Wager. Learning when-to-treat policies. *arXiv preprint arXiv:1905.09751*, 2019.

[5] Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*, 2019.

[6] J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866, 1994.

[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.