Mathematisches Forschungsinstitut Oberwolfach

Report No. 25/2019

DOI: 10.4171/OWR/2019/25

Foundations and new horizons for causal inference

Organized by
Nicolai Meinshausen, Zürich
Jonas Peters, Copenhagen
Thomas Richardson, Seattle
Bernhard Schölkopf, Tübingen

26 May - 1 June 2019

Abstract. While causal inference is established in some disciplines such as econometrics and biostatistics, it is only starting to emerge as a valuable tool in areas such as machine learning and artificial intelligence. The mathematical foundations of causal inference are fragmented at present. The aim of the workshop *Foundations and new horizons for causal inference* was to unify existing approaches and mathematical foundations as well as exchange ideas between different fields. We regard this workshop as successful in that it brought together researchers from different disciplines who were able to learn from each other not only about different formulations of related problems, but also about solutions and methods that exist in the different fields.

Mathematics Subject Classification (2010): 62Axx, 62Hxx, 62P20, 68Txx, 05Cxx, 62Gxx.

Introduction by the Organizers

The workshop *Foundations and new horizons for causal inference*, organised by Nicolai Meinshausen (ETH Zurich), Jonas Peters (University of Copenhagen), Thomas Richardson (University of Washington) and Bernhard Schölkopf (MPI Tübingen) was well attended with 52 participants from a broad geographic background.

The problem of inferring causal relationships from statistical data arises in many different fields of science and technology. However, abstract formal mathematical frameworks for reasoning about causality have been developed comparatively recently, at least within the history of probability and statistics as disciplines. Consequently, in the absence of generally accepted theoretical foundations, each area has developed specific ("autochthonous") approaches, each with their own terminology and assumptions.

One of the goals of this workshop was to bring together researchers from a wide range of different areas to facilitate communication and cross-pollination. In this regard, the workshop was undeniably very successful. It attracted researchers from Artificial Intelligence, Biostatistics, Computer Science, Economics, Epidemiology, Machine Learning, Mathematics and Statistics. New collaborations were initiated between researchers who probably would not have crossed paths were it not for this workshop. Likely this success is due in large part to the fact that the workshop took place under the prestigious auspices of the *Mathematisches Forschungsinstitut Oberwolfach*.

Four broad areas of causal inference were discussed at the workshop.

(1) **Mathematical foundations**. Purely statistical models aim at describing the underlying distribution of a data generating process. Causal models, however, go beyond that goal. They try to model the effect of perturbations of that system, too. Formulating such models, including the notion of interventions, thus lies at the core of causality research. Even though several frameworks exist, this is still a topic of current research, in particular, when considering dynamical models, extreme valued processes or the question of which variables to include in the model, say. Talks covering this topic include the ones from Niels Hansen, Dominik Janzing, Steffen Lauritzen, Karthika Mohan, Emilija Perkovic, Rajen Shah, Ilya Shpitser, Jin Tian, and Sebastian Weichwald.

(2) **Causal discovery**. While many causal inference methods assume a known causal structure for the causal model (often, a directed acyclic graph), there is also large interest in using causal discovery to estimate structure from complex data such as, for example, time-series in biological applications. Research goals include to develop methods that are robust with respect to model misspecification, scale to large data sets, deal with the existence of hidden variables or incorporate the information of interventional experiments. Talks covering this topic include the ones from Mathias Drton, Aapo Hyvarinen, Nicola Gnecco, Marloes Maathuis, Linbo Wang, and Kun Zhang.

(3) **Machine Learning and causality**. There is growing interest to adjust Machine Learning methods from a purely association-based learning approach towards causal inference. The hope is to obtain methods for classical machine learning problems such as prediction or semi-supervised learning that generalize better to test data (that may come from the same or from a related distribution as the training data) or are more sample-efficient. Moreover, causality can provide means to better understand classical machine learning paradigms and their applicability.

Talks covering this topic include the ones from Leon Bottou, David Blei, Julius von Kügelgen, Niklas Pfister, Christina Heinze-Deml, Ludwig Schmidt, Michele Sebag, David Sontag, and Fan Yang.

(4) **Applications.** Numerous applications were discussed, including personalised medicine, biological causal network discovery and climate science. Talks covering applications include the ones from Gregory Cooper, Sara Geneletti, Jakob Runge, and Sach Mukherjee.

Machine learning methods are currently successfully applied to a wide range of applications. Impressive empirical results are obtained in areas such as image classification or speech recognition. Many scientific problems, however, go beyond the task of iid prediction. In some domains such as public health, biology or Earth system science, we are usually interested in finding policies that yield a better outcome. In other areas, we expect that the test data will differ significantly from the training data. Causal concepts have the potential to play a role in solving many of these problems. We therefore expect to see more research on causality. While many of the goals connected to research on causality are ambitious, any advance in this area will potentially have a large impact not only in mathematics but in the natural sciences in general.

The workshop brought together experts on causal inference working on foundations and applications in Econometrics, Machine Learning, Statistics and the Natural Sciences. The talks and discussions at the workshop will help to shape the field in the coming years.

Acknowledgement: The MFO and the workshop organizers would like to thank the National Science Foundation for supporting the participation of junior researchers in the workshop by the grant DMS-1641185, "US Junior Oberwolfach Fellows". Moreover the MFO and the workshop organizers would like to thank the Simons Foundation for supporting Fan Yang and Kun Zhang in the "Simons Visiting Professors" program at the MFO.

Workshop: Foundations and new horizons for causal inference

Table of Contents

Abstracts

Causal Discovery with Arbitrary Nonlinear Dependencies using Nonlinear ICA

Aapo Hyvärinen

(joint work with R. P. Monti and K. Zhang)

Causal discovery in a linear Bayesian network, or a linear structural equation model, has been earlier shown to be possible by using the theory of independent component analysis (ICA). In fact, performing ICA on the data and further processing the results leads to one possible method for estimation of causal relationships, in the form of the LiNGAM framework. Here, we show how recent advances in nonlinear forms of ICA, in particular time-contrastive learning, enable identification of nonlinear structural equation models, and thus, of general nonlinear causal relationships. Importantly, we do not constrain the form of the nonlinear functions in this framework, and in particular we do not assume any kind of additivity. Instead, we require the data to have a richer statistical structure in the sense that the data must come from different conditions, in other words, they should be non-stationary in a general sense. The ensuing method can be shown to find the correct causal direction in the presence of general nonlinear relations in a bivariate setting. Thus we achieve a generalization of LiNGAM to the case of arbitrary nonlinearities.

References

[1] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *J. of Machine Learning Research*, 7:2003–2030, 2006.

[2] A. Hyvärinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems (NIPS2016)*, Barcelona, Spain, 2017.

[3] R. P. Monti, K. Zhang, and A. Hyvärinen. Causal discovery with general non-linear relationships using non-linear ICA. In *Proc. 35th Conf. on Uncertainty in Artificial Intelligence (UAI2019)*, 2019. In press.

Hypothesis generation through three principles of data science: predictability, computability and stability (PCS)

Bin Yu

(joint work with Karl Kumbier, Sumanta Basu, Ben Brown, Jamie Murdoch, Chandan Singh, Reza Abbassi-Asl)

We propose a framework in [4] that draws from three principles of data science: predictability, computability, and stability (PCS) to extract reliable, reproducible information from data and guide scientific hypothesis generation. The PCS framework builds on key ideas in machine learning, using predictability as a reality check and evaluating computational considerations in data collection, data storage, and algorithm design. It augments predictability and computability with an overarching stability principle, which expands statistical uncertainty considerations to assesses how results vary with respect to choices (or perturbations) made across the data science life cycle.

Building on PCS, we develop inference procedures to investigate the stability of data results relative to problem formulation, data cleaning, and modeling decisions. We compare PCS inference with existing methods in high-dimensional sparse linear model simulations to demonstrate that our approach compares favorably to others in terms of ROC curves over a wide range of simulation settings. Finally, we propose PCS documentation based on R Markdown or Jupyter Notebook, with publicly available, reproducible codes and narratives to back up human choices made throughout an analysis. The PCS workflow and documentation are demonstrated in a genomics case study available on Zenodo.

Stability is aso a minimum requirement for interpretability and reproducibility as advocated in [1]. Machine-learning models have demonstrated great success in learning complex patterns that enable them to make predictions about unobserved data. In addition to using models for prediction, the ability to interpret what a model has learned is receiving an increasing amount of attention. However, this increased focus has led to considerable confusion about the notion of interpretability. In particular, it is unclear how the wide array of proposed interpretation methods are related, and what common concepts can be used to evaluate them. We aim in [3] to address these concerns by defining interpretability in the context of machine learning and introducing the Predictive, Descriptive, Relevant (PDR) framework for discussing interpretations. The PDR framework provides three overarching desiderata for evaluation: predictive accuracy, descriptive accuracy and relevancy, with relevancy judged relative to a human audience. Moreover, to help manage the deluge of interpretation methods, we introduce a categorization of existing techniques into model-based and post-hoc categories, with sub-groups including sparsity, modularity and simulatability. To demonstrate how practitioners can use the PDR framework to evaluate and understand interpretations, we provide numerous real-world examples. These examples highlight the often under-appreciated role played by human audiences in discussions of interpretability. Finally, based on our framework, we discuss limitations of existing methods and directions for future work. We hope that this work will provide a common vocabulary that will make it easier for both practitioners and researchers to discuss and choose from the full range of interpretation methods.

As a case study of PCS, we propose in [2] iterative Random Forests (iRF). Genomics has revolutionized biology, enabling the interrogation of whole transcriptomes, genome-wide binding sites for proteins, and many other molecular processes. However, individual genomic assays measure elements that interact in vivo as components of larger molecular machines. Understanding how these high-order interactions drive gene expression presents a substantial statistical challenge. Building on random forests (RFs) and random intersection trees (RITs) and through extensive, biologically inspired simulations, we develope the iterative random forest algorithm (iRF). iRF trains a feature-weighted ensemble of decision trees to detect stable, high-order interactions with the same order of computational cost as the RF. We demonstrate the utility of iRF for highorder interaction discovery in two prediction problems: enhancer activity in the early Drosophila embryo and alternative splicing of primary transcripts in human-derived cell lines. In Drosophila, among the 20 pairwise transcription factor interactions iRF identifies as stable (returned in more than half of bootstrap replicates), 80% have been previously reported as physical interactions. Moreover, third-order interactions, e.g., between Zelda (Zld), Giant (Gt), and Twist (Twi), suggest high-order relationships that are candidates for follow-up experiments. n human-derived cells, iRF rediscovered a central role of H3K36me3 in chromatin-mediated splicing regulation and identified interesting fifth- and sixth-order interactions, indicative of multivalent nucleosomes with specific roles in splicing regulation. By decoupling the order of interactions from the computational cost of identification, iRF opens additional avenues of inquiry into the molecular mechanisms underlying genome biology.

References

[1] B. Yu *Stability*, Bernoulli **19(4)** (2013), 1484-1500.

[2] S. Basu, K. Kumbier, B. Brown, B. Yu *Iterative Random Forests to discover predictive and stable high-order interactions*, Processings of National Academy of Sciences (PNAS) **115(8)** (2018), 1943–1948.

[3] J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, B. Brown, B. Yu *Interpretable machine learning: definitions, methods and applications*, Processings of National Academy of Sciences (PNAS) (2019) (accepted).

[4] B. Yu, K. Kumbier *Three principles of data science: predictability, computability, and stability (PCS)*, https://arxiv.org/abs/1901.08152 (2019).

Instance-Specific Causal Bayesian Network Structure Learning

Gregory F. Cooper

(joint work with Chunhui Cai, Fattaneh Jabbari, Xinghua Lu,
Shyam Visweswaran)

This article describes our recently published research in developing an instance-specific approach for learning causal Bayesian network structures from data and in applying the approach to molecular cancer data [7, 2].

A Bayesian network (BN) is a directed, acyclic graphical model that represents probabilistic relationships among a set of variables V. A causal Bayesian network (CBN) is a BN in which arcs are interpreted as direct causation, relative to V [9].

Most CBN structure learning algorithms are designed to recover the structure that models the relationships that are shared by the instances in a population. While learning accurate population-wide CBNs is useful, learning CBNs that are specific to a given instance can also be important.

We use CBNs that represent context-specific independence (CSI). CSI captures independence relationships that hold between the causes (parents) and their effect (child) in a CBN in particular contexts (i.e., when the cause variables take on particular values) [1]. In a CBN with CSI, the CBN structure of an instance depends on the values of the variables in that instance. In this way, CSI provides a representation that supports instance-specific modeling.

A number of algorithms have been developed that learn from data a population-wide BN with CSI, as for example the algorithms described in [5, 10, 8, 11]. To our knowledge, however, none of the algorithms published by other researchers learn a CSI model that is specific to a given instance T (e.g., a given patient), which is the approach we have been investigating [4]. Doing so in learning CBNs has at least two advantages. First, the learned instance-specific causal model provides a relatively precise representation of the causal processes that are ongoing specifically in T. Second, searching for an instance-specific model will usually be much more efficient than searching for all (or at least many) possible instance-specific models and subsequently choosing the one that best matches T.

For example, a lung-cancer tumor T in a patient is an instance that can have a set of causal mechanisms that are different from that of another lung-cancer tumor, either in the same patient or in a different patient. To determine the most effective treatment for a tumor in the current patient, it is important to know the particular causal mechanisms that are driving that specific tumor to be cancerous. In reality, a given tumor is likely to be composed of a set of cellular mechanisms that rarely all occur together, yet each individual mechanism may appear relatively commonly in other tumors. A population-wide CBN would at best capture the more common mechanisms operating in lung cancer and not all of the particular mechanisms that are active in the current patient's lung-cancer tumor T. The task, then, is to construct the joint set of mechanisms of a given tumor from the individual mechanisms seen in previous tumors. To do so, we use the known features (i.e., the variable values) of the current tumor T to help identify and construct the individual mechanisms that compose the set of mechanisms that are jointly driving the current tumor.

In [7], we describe an adapted version of the GES search algorithm [3] that is able to learn an instance-specific CBN structure from data. We also developed a more specialized instance-specific method called the Tumor-specific Causal Inference (TCI) algorithm that searches over bipartite CBNs in which one partition contains somatic genomic alterations (SGAs) in a given tumor, such as gene mutations; the other partition contains abnormal cellular processes indicative of cancer, such as aberrant transcriptomic changes, which suggest the cancer disease mechanisms [2].

We applied TCI to tumors from The Cancer Genome Atlas (TCGA) [6] and estimated for each tumor the SGAs that causally regulate the differentially expressed genes (DEGs) in that tumor. On a set of more than 5,000 tumors, TCI identified over 600 SGAs that are predicted to cause (drive) cancer-related DEGs in a significant number of tumors, including most of the previously known drivers and many novel candidate cancer drivers [2]. On the whole, the inferred causal relationships are statistically robust and biologically sensible, and the selected experiments we have performed provide support for the validity of the candidate drivers that are predicted by TCI. As an example, TCI inferred that the gene $CSMD3$ is a likely cause of cancerous behavior in tumors, although it was not designated as such in previous studies. In [2], we report examining whether experimental manipulations of $CSMD3$ affect oncogenic phenotypes. In particular, we identified the cancer cell line HGC27 as having $CSMD3$ amplification, and we knocked down the expression of that gene using siRNAs, followed by monitoring cellular phenotypes, including cell proliferation and cell migration. The results show that knocking down $CSMD3$ significantly attenuated both proliferation and migration. These results provide support that $CSMD3$ is involved in producing cancer-related cellular phenotypes in this cell line, and by extension, possibly in some human *in vivo* cancers.

This article describes an initial approach to tailoring the construction of a causal model to a given instance. Our results in applying the approach to molecular cancer data provide support that the method yields valuable causal insights. There are undoubtedly many additional ways in which causal learning can be tailored to better discover the causal structure and mechanisms of a given instance. We believe that the exploration of such prospects is an important open area of causal discovery research.

References

[1] Boutilier C, Friedman N, Goldszmidt M, Koller D. *Context-specific independence in Bayesian networks.* In: Proceedings of the Conference on Uncertainty in Artificial Intelligence (1996) 115–123.

[2] Cai C, Cooper GF, Lu KN, Ma X, Xu S, Zhao Z, Chen X, Xue Y, Lee AV, Clark N, Chen V, Lu S, Chen L, Yu L, Hochheiser HS, Jiang X, Wang QJ, Lu X. *Systematic discovery of the functional impact of somatic genome alterations in individual tumors through tumor-specific causal inference.* PLOS Computational Biology (to appear, 2019).

[3] Chickering DM. *Optimal structure identification with greedy search.* Journal of Machine Learning Research (2003) 507–554.

[4] Ferreira A, Cooper GF, Visweswaran S. *Decision path models for patient-specific modeling of patient outcomes.* In: Proceedings of the Annual Symposium of the American Medical Informatics Association (2013) 413–421.

[5] Friedman N, Goldszmidt, M. *Learning Bayesian networks with local structure.* In: Learning in Graphical Models (Springer, 1998) 421–459.

[6] Hutter C, Zenklusen JC. *The Cancer Genome Atlas: Creating lasting value beyond its data.* Cell, 173 (2018) 283–285.

[7] Jabbari F, Visweswaran S, Cooper GF. *Instance-specific Bayesian network structure learning.* In: Proceedings of the Conference on Probabilistic Graphical Models (2018).

[8] Oates CJ, Smith JQ, Mukherjee S, Cussens J. *Exact estimation of multiple directed acyclic graphs.* Statistics and Computing, 26 (2016) 797–811.

[9] Pearl J. *Causality: Models, Reasoning and Inference* (Cambridge University Press, 2009).

[10] Pensar J, Nyman H, Koski T, Corander J. *Labeled directed acyclic graphs: A generalization of context-specific independence in directed graphical models.* Data Mining and Knowledge Discovery (2015) 503–533.

[11] Zou Y, Pensar J, Roos T. *Representing local structure in Bayesian networks by Boolean functions.* Pattern Recognition Letters, 95 (2017) 73–77.

Checking Assumptions in Causal Inference from Observational Data

David Sontag

(joint work with Fredrik Johansson, Uri Shalit, Michael Oberst, Dennis Wei, Tian Gao, Kush Varshney)

Evaluating intervention decisions is a key question in many diverse fields including medicine, economics, and education. In medicine, an optimal choice of treatment for a patient in the intensive care unit may mean the difference between life and death. In public policy, job reforms have impact on the unemployment rate and the economy of a nation. To evaluate such interventions we must study their *causal effect* – the difference in an outcome of interest under alternative choices of intervention. Since only one option may be carried out at a time, any data to support such evaluations only reveals the outcome of the action taken and never the outcome of the action not taken, which remains an unknown *counterfactual.* To estimate causal effects, we must therefore infer what would have happened had we made another decision. Furthermore, to decide on personalized interventions, such as tailoring treatments to patients, we must understand individual-level causal effects, conditioned on the available information on an individual recorded prior to intervention.

We study the problem of estimating individual-level causal effects from non-experimental, *observational* data. An observational dataset consists of historical records of interventions, the contexts in which they were made, and the observed outcomes. For example, in the setting of health care, these would correspond to medications, medical records, and the outcome of treatment, such as mortality. An individual-level effect measures the causal effect of medication choice, conditioned on what is known about the patient. There are two assumptions that most approaches to causal inference from observational data make: that there is 1) *common support* (also called overlap), and 2) *no unmeasured confounders* (ignorability).

In the first part of the talk, I showed how techniques from learning theory and unsupervised domain adaptation can be used to give bounds on the error in estimated causal effects [1]. This is in contrast from typical results in causal inference which focus on proving consistency and do not provide any guarantees when the potential outcome functions are misspecified. Our bounds are based on distance measures between groups receiving different treatments. I then showed how these bounds can be minimized by sample re-weighting and representation learning, leading to a new class of causal inference algorithms. An important direction for future research is how to further reduce the sample complexity for estimating conditional average treatment effect.

In the second part of the talk, I asked whether it may be possible in some cases to develop checks for the assumptions, such as overlap [2]. When overlap does not hold globally, characterizing local regions of overlap can inform the relevance of any causal conclusions for new subjects, and can help guide additional data collection. To have impact, these descriptions must be interpretable for downstream users who are not machine learning experts, such as clinicians. I formalized overlap estimation as a problem of finding minimum volume sets and suggested a method to solve it by reduction to binary classification with Boolean rules. I then described a case study in which we learned to describe treatment group overlap for post-surgical opioid prescriptions. As an open question, I asked whether instead of requiring overlap (which in many respects seems too strong of an assumption), we could give conditions under which it is OK to extrapolate predictions of potential outcomes.

References

[1] U. Shalit, F. Johansson, D. Sontag. *Estimating individual treatment effect: generalization bounds and algorithms.* ICML (2017).

[2] F. Johansson, D. Wei, M. Oberst, T. Gao, G. Brat, D. Sontag, K. Varshney, *Characterization of Overlap in Observational Studies*, Pre-print (2019).

Refuting the inferential validity of causal estimators empirically (and inconsistently)

Lin Liu

(joint work with Rajarshi Mukherjee, James M. Robins)

For many causal effect parameters ψ of interest doubly robust machine learning (DR-ML) [1] estimators $\widehat{\psi}_1$ are the state-of-the-art, incorporating the benefits of the low prediction error of machine learning (ML) algorithms; the decreased bias of doubly robust estimators; and.the analytic tractability and bias reduction of sample splitting with cross fitting. Nonetheless, even in the absence of confounding by unmeasured factors, when the vector of potential confounders is high dimensional, the associated $(1-\alpha)$ Wald confidence intervals $\widehat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathrm{se}}[\widehat{\psi}_1]$ may still undercover even in large samples, because the bias of the estimator may be of the same or even larger order than its standard error of order $n^{-1/2}$.

In this paper, we introduce novel tests that (i) can have the power to detect whether the bias of $\widehat{\psi}_1$ is of the same or even larger order than its standard error of order $n^{-1/2}$, (ii) can provide a lower confidence limit on the degree of under coverage of the interval $\widehat{\psi}_1 \pm z_{\alpha/2}\widehat{\mathrm{se}}[\widehat{\psi}_1]$ and (iii) strikingly, are valid under essentially no assumptions whatsoever. We also introduce an estimator $\widehat{\psi}_2 = \widehat{\psi}_1 - \widehat{\mathbb{IF}}_{22}$ with bias generally less, and often much less, than that of $\widehat{\psi}_1$, yet whose standard error is not much greater than $\widehat{\psi}_1$'s. The tests, as well as the estimator

$\widehat{\psi}_2$, are based on a U-statistic $\widehat{\mathbb{IF}}_{22}$ that is the second-order influence function for the parameter that encodes the estimable part of the bias of $\widehat{\psi}_1$. For the definition and theory of higher order influence functions see [2, 3]. When the covariance matrix of the potential confounders is known, $\widehat{\mathbb{IF}}_{22}$ is an unbiased estimator of its parameter. When the covariance matrix is unknown, we propose several novel estimators of $\widehat{\mathbb{IF}}_{22}$ that perform almost as well as the known covariance case in simulation experiments.

Our impressive claims need to be tempered in several important ways. First no test, including ours, of the null hypothesis that the ratio of the bias to its standard error is small can be consistent [without making additional assumptions (e.g. smoothness or sparsity) that may be incorrect]. Furthermore the above claims only apply to parameters in a particular class. For the others, our results are unavoidably less sharp and require more careful interpretation.

References

[1] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, & J. Robins. Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal 21.1 (2018): C1-C68.

[2] J. Robins, Li. Li, E. Tchetgen Tchetgen, & A. van der Vaart. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman (pp. 335-421)* (2008). Institute of Mathematical Statistics.

[3] J. M. Robins, L. Li, R. Mukherjee, E. Tchetgen Tchetgen, & A. van der Vaart. Minimax estimation of a functional on a structured high-dimensional model. The Annals of Statistics 45.5 (2017): 1951-1987.

Towards scalable causal learning

Sach Mukherjee

(joint work with Steven Hill, Chris Oates, Umberto Noè and Bernd Taschler)

Causal structure learning is concerned with learning causal relationships between variables. Consider a set of p variables indexed by $V = \{1 \dots p\}$. We focus on the task of determining, for a subset of (ordered) pairs $(i, j) \in \mathcal{K} \subseteq V \times V$, whether or not node i exerts a causal influence on node j. In particular, our focus is on the binary 'detection' problem (of learning whether or not i exerts a causal influence on j) rather than estimation of the magnitude of any causal effect. We frame the problem as a machine learning task: the idea is to treat discrete indicators of causal relationships between variables as 'labels' (in a discriminative learning sense) and to exploit available data on the variables of interest to provide features for the labelling task.

Many causal learning methods are based on graphical models, with models based on directed acyclic graphs (DAGs) playing a key role [1, 2]. The PC algorithm is a prominent example of such a method [1]. It estimates an equivalence class of DAGs – encoded as a completed partially directed acyclic graph or CPDAG – via a series of conditional independence tests. The PC output can in turn be used to estimate bounds on quantitative total causal effects between nodes [3]. Related methods, including score-based approaches, are available for interventional data and problem settings with latent variables. In contrast to these approaches, which are rooted in data-generating models of the causal system, there has been recent work with an emphasis on telling apart causal and non-causal relationships. Work in this 'discriminative' direction has included [4] and [5] and our work follows in this line.

In a nutshell, our approach works as follows. Available information on some causal relationships (via background knowledge or experimental data) are treated as 'labels' that are combined with a featurization of the data to train a learner that gives labels across the entire problem. This gives, for each pair $(i, j) \in \mathcal{K}$, a label (or probabilistic score) $\hat{G}_{ij}$ that is intended to encode its causal status. These labels can be viewed as specifying a directed graph in which the presence of a directed edge between vertices i and j means that the variable with index i is inferred to have a causal influence on the variable with index j.

Within this overall scheme, we consider a semi-supervised formulation using manifold regularization (following [6]) as well as supervised approaches that are suitable for very high dimensional data. We show empirical results on several biological datasets, including examples where causal effects can be verified by experimental intervention. This allows us to empirically quantify performance in terms of agreement with unseen interventional experiments. We compare performance with a range of existing causal approaches and non-causal baselines. Taken together, these results demonstrate the empirical efficacy of the proposed approaches, as well as their generality and simplicity from a user's point of view.

Although effective at the specific tasks considered, in contrast to graphical models-based methods, the approaches proposed are less general in the sense that they only learn discrete indicators of causal status, but cannot by themselves provide a full range of probabilistic output (e.g. all post-intervention distributions). Nevertheless, our results suggest that machine learning-based schemes can be effective and that it may be fruitful to investigate combining them with graphical models-based approaches in the future.

References

[1] P. Spirtes, C. N. Glymour, R. Scheines, D. Heckerman, C. Meek, G. Cooper & T. Richardson. (2000). Causation, prediction, and search. MIT Press.

[2] J. Pearl. (2009). Causality. Cambridge University Press.

[3] M. H. Maathuis, M. Kalisch & P. Bühlmann. Estimating high-dimensional intervention effects from observational data. The Annals of Statistics, **37**(6A) (2009), 3133–3164.

[4] D. Lopez-Paz, K .Muandet, B. Schölkopf & I. Tolstikhin. Towards a learning theory of cause-effect inference. In International Conference on Machine Learning (2015) (pp. 1452-1461).

[5] J. M. Mooij, J. Peters, D. Janzing, J. Zscheischler & B. Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. The Journal of Machine Learning Research, **17**(1) (2016), 1103-1204.

[6] M. Belkin, P. Niyogi & V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. The Journal of Machine Learning Research, **7** (2006), 2399-2434.

Causal Inference with Unmeasured Confounding: A New Look at Instrumental Variable

Linbo Wang
(joint work with Eric Tchetgen Tchetgen)

Observational studies are often used to infer treatment effects in social and biomedical sciences. In these studies, the treatment assignment may be associated with various background variables that are associated with the outcome, causing the unadjusted treatment effect estimate to be biased. These background variables are often called confounders. A major challenge of causal inference in observational studies is that in practice, these confounding variables are often not fully observed, making it impossible to identify the treatment effect in view. In such settings, instrumental variable (IV) methods are useful in dealing with unmeasured confounding and have gained popularity among econometricians, statisticians and epidemiologists. Intuitively, conditional on baseline covariates, a valid IV affects the outcome through its effect on the treatment but is otherwise unrelated to the outcome.

However, under the standard IV model, the average treatment effect (ATE) is only partially identifiable. Traditionally, researchers have assumed additionally a system of linear structural equation models (SEMs); see [1] for a recent review. One such SEM can be inferred from the following system of linear regression models:

$$D = \alpha_0 + \alpha_1 Z + \alpha_2 X + \alpha_3 U + \epsilon_D, \tag{1}$$

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 U + \epsilon_Y, \tag{2}$$

where Z is an instrumental variable, D is a continuous treatment, Y is a continuous outcome, X and U denote observed and unobserved baseline covariates, respectively, $Z \perp\!\!\!\perp U \mid X$ and the error terms are independent: $\epsilon_D \perp\!\!\!\perp \epsilon_Y$. However, (1) and (2) impose strong parametric assumptions on the underlying data generating process. Moreover, a fundamental limitation with relying on models like (1) and (2) is that they impose one set of assumptions, which conflates the definition, identification and estimation of the treatment effect.

To address these problems, we propose two alternative no-interaction assumptions involving the unobserved confounders that allow for identification of the ATE. Our first assumption is a generalization of linear model (1); our second assumption is guaranteed to hold under the null of no treatment effect. We also allow for instruments that are confounded with the treatment. Our identification assumptions are clearly separated from model assumptions needed for estimation, so that researchers are not required to commit to a specific observed data model in establishing identification. Moreover, under both of our identification assumptions, the ATE can be represented by the same observed data functional so that in the estimation stage, we can target a single statistical parameter. This parameter is called the average Wald estimand, a generalization of the Wald estimand ([2]) to accommodate baseline covariates X. We then construct multiple estimators that are consistent under three different observed data models, and triply robust estimators that are consistent in the union of these observed data models. We pay special attention to the case of binary outcomes, for which we obtain bounded estimators of the ATE that are guaranteed to lie between -1 and 1. Our approaches are illustrated with simulations and a data analysis evaluating the causal effect of education on earnings.

References

[1] P. S. Clarke, and F. Windmeijer, *Instrumental variable estimators for binary outcomes*, Journal of the American Statistical Association **107** (2012), 1638–1652.

[2] A. Wald, *The fitting of straight lines if both variables are subject to error*, The Annals of Mathematical Statistics **11** (1940), 284–300.

Towards more reliable causal discovery and prediction

Kun Zhang

This talk was concerned about how to make causal discovery from observational data more reliable and how to improve prediction in nonstationary environments from a causal perspective. Since the 1990's, conditional independence relationships in the data have been exploited to recover the underlying causal structure. Typical (conditional independence) constraint-based algorithms include PC and Fast Causal Inference (FCI) [1]. Such approaches are widely applicable because they can handle various types of data distributions and causal relations, given reliable conditional independence testing methods. However, they do not necessarily provide complete causal information because they output (independence) equivalence classes, i.e., a set of causal structures satisfying the same conditional independences.

In the past 13 years it has been further shown that algorithms based on properly defined Functional Causal Models (FCMs) are able to distinguish between different potential graphical structures in the same equivalence class. This benefit is owed to additional assumptions on the data distribution than conditional independence relations. Without constraints on the form of the functional causal model, then for any two variables one can always express one of them as a function of the other and independent noise [2, 3]. However, this is not the case anymore if the functional classes are properly constrained. Such FCMs include the Linear, Non-Gaussian, Acyclic Model (LiNGAM) [4], the post-nonlinear (PNL) causal model [5, 6], and the nonlinear additive noise model (ANM) [7], where causes have nonlinear effects and noise is additive.

Causal discovery exploits observational data. The data are produced by not only the underlying causal process, but also the sampling process. In practice, to achieve reliable causal discovery, one needs to address specific challenges posed in the causal process or the sampling process, depending on the application domain. Such challenges include nonlinear causal interactions, much lower data acquisition rate compared to the underlying rate of changes [8, 9], feedback loops in the causal model [10], existence of measurement error [11], and possible unmeasured confounding causes. In clinical studies, we often have a large number of missing data [12]. Data collected on Internet or in hospital often suffer from selection bias [13]. Some data sets involve both mixed categorical and continuous variables, which may pose difficulties in conditional independence tests and specification of appropriate forms of the FCM [14]. Many of these issues have recently been considered, with corresponding methods proposed to address them. In the talk I particularly focused on how causal discovery benefits from nonstationarity of time series data of heterogeneity of multi-domain data [15].

On the other hand, causal information describes properties of the process that render a set of constraints on the data distribution, and is able to facilitate understanding and solving a number of learning problems involving distribution shift or concerning the relationship between different factors of the joint distribution. In particular, for learning under data heterogeneity, it is naturally helpful to learn and model the properties of data heterogeneity, which then benefits from causal modeling. Such learning problems include domain adaptation (or transfer learning) [16], semi-supervised learning [17], and learning with positive and unlabeled examples. Leveraging causal modeling for recommender systems [18] and Reinforcement learning has been becoming an active research field in recent years.

References

[1] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* MIT Press, Cambridge, MA, 2nd edition, 2001.

[2] K. Zhang, Z. Wang, J. Zhang, and B. Schölkopf. On estimation of functional causal models: General results and application to post-nonlinear causal model. *ACM Transactions on Intelligent Systems and Technologies*, 2015.

[3] P. Spirtes and K. Zhang. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, 3, 2016.

[4] S. Shimizu, P.O. Hoyer, A. Hyvärinen, and A.J. Kerminen. A linear non-Gaussian acyclic model for causal discovery. *JMLR*, 7:2003–2030, 2006.

[5] K. Zhang and L. Chan. Extensions of ICA for causality discovery in the hong kong stock market. In *ICONIP 2006.*

[6] K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proc. UAI 2009*, Montreal, Canada.

[7] P.O. Hoyer, D. Janzing, J. Mooji, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *NIPS 21*, Vancouver, B.C., Canada, 2009.

[8] M. Gong, K. Zhang, B. Schoelkopf, D. Tao, and P. Geiger. Discovering temporal causal relations from subsampled data. In *ICML*, pages 1898–1906, 2015.

[9] M. Gong, K. Zhang, B. Schölkopf, C. Glymour, and D. Tao. Causal discovery from temporally aggregated time series. In *Proc. UAI*, 2017.

[10] R. Sanchez-Romero, J. D. Ramsey, K. Zhang, M. R. K. Glymour, B. Huang, and C. Glymour. Estimating feedforward and feedback effective connections from fmri time series: Assessments of statistical methods. *Network Neuroscience*, 3:274–306‘, 2019.

[11] K. Zhang, M. Gong, J. Ramsey, K. Batmanghelich, P. Spirtes, and C. Glymour. Causal discovery with linear non-gaussian models under measurement error: Structural identifiability results. In *Proc. UAI 2018*, CA, USA.

[12] R. Tu, C. Zhang, P. Ackermann, K. Mohan, C. Glymour, H. Kjellström, and K. Zhang. Causal discovery in the presence of missing data. In *Proc. AISTATS 2019.*

[13] K. Zhang, J. Zhang, B. Huang, B. Schölkopf, and C. Glymour. On the identifiability and estimation of functional causal models in the presence of outcome-dependent selection. In *Proc. UAI 2016.*

[14] B. Huang, K. Zhang, Y. Lin, B. Schölkopf, and Clark Glymour. Generalized score functions for causal discovery. In *Proc. ACM SIGKDD 2018.*

[15] K. Zhang, B. Huang, J. Zhang, C. Glymour, and B. Schölkopf. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. In *Proc. IJCAI 2017.*

[16] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang. Domain adaptation under target and conditional shift. In *Proc. ICML 2013.*

[17] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *Proc. ICML 2012*, Edinburgh, Scotland.

[18] M. Wang, M. Gong, X. Zheng, and K. Zhang. Modeling dynamic missingness of implicit feedback for recommendation. In *NIPS 2018.*

Causal structure learning for partially observed multivariate event processes

Niels Richard Hansen

(joint work with Søren Wengel Mogensen, Daniel Malinsky)

Structural causal models of event processes imply certain local independencies among the coordinates of the processes. The local independencies form an independence model that can be encoded as a graphical separation model in a directed graph via δ- or μ-separation. If only some of the process coordinates are observed, we ask what can be learned about the causal structure in terms of the local independence model?

Some notation is required to formulate our main results. We consider event processes indexed by $V = \{1, \ldots, d\}$. The time dynamics of the k-th event process is given in terms of its *intensity*,

$$P(\text{one } k\text{-event} \in (t, t+\delta] \mid \mathcal{F}_t) \simeq \lambda_t^k \delta, \quad k \in V, \text{ and small } \delta > 0,$$

where $\mathcal{F}_t$ denotes the history of all events up to time t, and λ_t^k depends on $\mathcal{F}_t$. For $C \subseteq V$ we define $\mathcal{F}_t^C$ as the history of events in C up to time t, and

$$\lambda_t^{k,C} = E(\lambda_t^k \mid \mathcal{F}_t^C)$$

is the optional projection of the intensity of the k-th process onto the history of processes indexed by C.

For $A, B, C \subseteq V$, B is *conditionally locally independent* of A given C, denoted

$$A \not\rightarrow B \mid C,$$

if $\lambda_t^{k,A\cup C} = \lambda_t^{k,C}$ for $k \in B$. This defines an abstract independence model as a ternary relation on subsets of V,

$$\langle A, B \mid C \rangle \in \mathcal{I}_{\text{CLI}}(V) \Leftrightarrow A \not\rightarrow B \mid C$$

We would like to encode this independence model as a graphical independence model, that is, find a graph and a separation criterion on the graph such that separation in the graph implies conditional local independence.

Definition (Local Independence Graph)**.** A graph $\mathcal{G} = (V, E)$ is a local independence graph if

$$(j, k) \notin E \Longrightarrow j \not\rightarrow k \mid V \backslash \{j\}.$$

The local independence graph is a directed graph, that may have cycles, and we define a separation criterion in terms of the following definition.

Definition (μ-connecting walk)**.** A nontrivial walk from j to k in $\mathcal{G}$ is said to be μ-connecting given C if $j \notin C$, every collider is an ancestor of C, no noncollider is in C, and there is an arrow head at k.

A set B is then said to be μ-separated from A given C if there is no μ-connecting walk from any $j \in A$ to any $k \in B$ given C in the graph. The corresponding graphical independence model is denoted $\mathcal{I}_{\mathcal{G}}(V)$. Note that requiring an arrow head at k in the above definition makes the independence model different from d-separation and asymmetric.

Theorem (Global Markov Property, [1])**.** *Let $\mathcal{G}$ denote the local independence graph. Under some regularity conditions it holds that if C μ-separates A from B in a local independence graph then $A \not\rightarrow B \mid C$. That is, $\mathcal{I}_{\mathcal{G}}(V) \subseteq \mathcal{I}_{\mathrm{CLI}}(V)$.*

The global Markov property (using δ-separation) was proved for event processes first in [2], but we give more general results in [1] based on abstract semigraphoid properties.

To represent the independence model among observed processes when there are also latent processes, we need a notion of projection. This is achieved by extending μ-separation to directed mixed graphs (DMGs). The main results from [3] are

- A *latent projection* maps a DMG with vertices V to a DMG with vertices $O \subseteq V$. The μ-separation properties are preserved among observed variables.
- All Markov equivalent DMGs on O have a common *Markov equivalent supergraph.*
- The maximal DMG representing a Markov equivalence class can be *constructed from the independence model.*
- Edge status in the equivalence class is characterized via the directed mixed equivalence graph (DMEG).

The proof in [3] that the maximal DMG exists is constructive, and provides, in principle, a learning algorithm. In [1] we propose a more efficient learning algorithm of the DMEG that is shown to be sound and complete under a faithfulness assumption, that is, assuming that $\mathcal{I}_{\mathcal{G}}(V) = \mathcal{I}_{\mathrm{CLI}}(V)$.

Two open problems, that we are currently pursuing, are

- a characterization of faithfulness for some model classes
- and practical statistical tests of conditional local independence.

References

[1] S. W. Mogensen, D. Malinsky, N. .R Hansen, *Causal Learning for Partially Observed Stochastic Dynamical Systems*, UAI (2018), 142.

[2] V.. Didelez, *Graphical models for marked point processes based on local independence*, JRSS-B **70**(1) (2008), 245–264.

[3] S. W. Mogensen, N. .R Hansen, *Markov equivalence of marginalized local independence graphs*, Annals of Statistics, to appear.

Questions about ML and AI

Léon Bottou

The purpose of this talk is to explain the relevance of causation to research in artificial intelligence. Despite the promises of pundits, there is indeed a large gap between the technological capabilities of machine learning (ML) and the vague and elusive goals of artificial intelligence (AI). The first part of the talk reviews some of the common issues with ML methods and shows how they display many of the characteristic issues one encounters in causal inference research. The second part of the talk is an attempt to name many of the nuances of causation in the hope to provide a roadmap to approach artificial intelligence.

Success and shortcomings of ML — The current interest for artificial intelligence results from a couple success stories in machine learning. Thanks to the availability of large datasets and powerful computing infrastructure, supervised machine learning and reinforcement learning were able to deliver striking advances in several domains, such as computer vision [6], speech recognition [3], Go playing software [10], machine translation [1]. These striking successes however come with shortcomings that cleary impede our progress towards AI:

- Training state-of-the-art ML models often demands inhuman amounts of data. Humans learn much more quickly and are more adaptable. They do not only use training data but also are able to *reason* how their past experiences can be transferred to new problems.
- ML systems replace imprecisely specified problems (which images represent a bird?) by well defined statistical proxies (minimizing a training cost). However, because large training dataset are poorly curated, ML systems often capture *spurious correlations* and learn nonsense.
- Humans know the importance of the logical and compositional structure of a visual scene or a natural language sentence. In contrast, ML systems seem unable to positively leverage such knowledge. A possible way to understand this paradox is to remember that, for instance, the compositional structure of language is more useful for composing new sentences or interpretint rare ones than it is useful for modeling the skewed distribution of observed sentences. This is not about what has been told (the *observed*) but about could have been told (the *counterfactual*.)

In conclusion, although they can precisely replicate the observed training distribution, ML systems lack in common sense because they cannot easily infer what *could have been observed* under *closely related circumstances.*

The many faces of causation — On the one hand, the above description of the ML shortcomings emphasizes their similarity with fundamental issues in causation. On the other hand, none of these problems come with a causal graph or with well defined interventions. This means that we may not be able to understand them using solely the manipulative definition of causation that is common statistics. Fortunately, an abundant literature in epistemology, metaphysics, and psychology offers alternative ways to understand causation, a catalogue of ideas for future research. The following is an attempt to name some of them.

- *Manipulative causation* focuses on predicting the outcome of well defined interventions on a causal system. See [9, 5] and references therein.
- *Causal invariance* investigates which properties of a system are conserved when affected by explicit or implicit interventions. See, for instance, [12, chap 6] and [2].
- *Causal reasoning* focuses on causal statements as elements of reasoning chains. Statements that cannot be verified experimentally acquire value when they take part in chains that make verifiable predictions.
- *Causal explanation* provides causal commentaries that help understanding an observed phenomena but may not be complete enough to make sensible predictions [11].
- *Dispositional causation* and *affordances* associates objects with the causal relationship they enable [8, 4].
- *Causal intuition* take advantage of observed data to suggest short lists of plausible causal models whose validity can later be investigated using more direct experiments. See [7] and references therein.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] Nancy Cartwright. Two theorems on invariance and causality. *Philosophy of Science*, 70(1):203–224, 2003.

[3] L. Deng, G. E. Hinton, and B. Kingsbury. New types of deep neural network learning for speech recognition and related applications: An overview. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada*, pages 8599–8603, 2013.

[4] James J. Gibson. *The Ecological Approach to Visual Perception.* Houghton Mifflin Harcourt, 1979.

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

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, 2012.

[7] D. Lopez-Paz. *From dependence to causation.* PhD thesis, University of Cambridge, 2016.

[8] Stephen Mumford and Rani Lill Anjum. *Getting Causes from Powers.* Oxford University Press, 2011.

[9] Judea Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, 2nd edition, 2009.

[10] David Silver, Aja Huang, Christopher J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–503, 2016.

[11] Georges Henrik von Wright. *Explanation and Understanding.* Cornell University Press, 2004.

[12] James Woodward. *Making things happen: A theory of causal explanation.* Oxford university press, 2005.

Causal KinetiX: Learning stable structures in kinetic systems

Niklas Pfister

(joint work with Stefan Bauer, Jonas Peters)

Learning kinetic systems from data is one of the core challenges in many fields. Efficient computational methods to identify a robust underlying model from data are essential for the extrapolation and generalization capabilities of data driven modeling approaches. Existing data driven approaches infer the parameters of ordinary differential equations by considering the goodness-of-fit of the integrated systems as a loss function. Due to the complex structure of these systems the inferred models often do not generalize well to unobserved conditions. In particular, they perform poorly when predicting the system under interventions. We propose a novel framework to identify structure in causal kinetic models. Instead of solely focusing on predictive performance, our framework explicitly incorporates heterogeneity and optimizes for good generalization performance. To achieve this, we assume an underlying causal model for the dynamics of the target process $(y_t)_t$. Such a model induces additional structure into the problem by the well-known concept of causal invariance, autonomy or modularity [1, 2]. In the setting considered here, this concept implies that the conditional dynamics of a target process $(y_t)_t$ given a set of predictors $(\mathbf{x}_t)_t$ remains constant across different experiments, i.e, there exists a fixed function f such that

$$\tfrac{\mathrm{d}}{\mathrm{d}t} y_t^e = f(\mathbf{x}_t^e),$$

for all experiments e. We argue that such an assumption is reasonable in a wide range of physical systems and is a natural requirement whenever one is interested in predicting the intervention effect on a target process after intervening on the predictors.

Our proposed procedure is based on a combination of smoothing techniques and model based structure search which explicitly incorporates this invariance property as a learning principle. In particular, it does not require any numerical integration and thus remains computationally feasible and robust to model misspecifications. Furthermore, given sufficient conditions on the underlying noise model and assuming sufficiently heterogeneous observations it is possible to prove that our procedure asymptotically recovers the true causal parents for the target process. Numerical experiments on simulated data, verify this theoretical result and illustrate that our method out-performs standard techniques and is feasible for practically relevant data sets. We also apply the method to a real-world biological data set related to a signaling pathway and show that it is able to find models which are capable of predicting several types of unobserved interventions. We believe these results suggest that learning the structure of kinetic systems indeed benefits from a causal perspective.

References

[1] T. Haavelmo *The Probability Approach in Econometrics*, Econometrica (1944).
[2] J. Aldrich *Autonomy*, Oxford Economic Papers (1989).
[3] N. Pfister, S. Bauer, J. Peters, *Identifying Causal Structure in Large-Scale Kinetic Systems*, ArXiv e-prints (arXiv:1810.11776).

The Blessing of Multiple Causes: Extended Abstract

David Blei
(joint work with Yixin Wang)

Here is a frivolous, but perhaps lucrative, causal inference problem. Table 1 contains data about movies. For each movie, the table shows its cast of actors and how much money the movie made. Consider a movie producer interested in the causal effect of each actor; for example, how much does revenue increase (or decrease) if Oprah Winfrey is in the movie?

The producer wants to solve this problem with the potential outcomes approach to causality [10, 33, 34]. Following the methodology, she associates each movie to a *potential outcome function*, $y_i(\boldsymbol{a})$. This function maps each possible cast $\boldsymbol{a}$ to its revenue if the movie i had that cast. (The cast $\boldsymbol{a}$ is a binary vector with one element per actor; each element encodes whether the actor is in the movie.) The potential outcome function encodes, for example, how much money *Star Wars* would have made if Robert Redford replaced Harrison Ford as Han Solo. When doing causal inference, the producer's goal is to estimate something about the population distribution of $Y_i(\boldsymbol{a})$. For example, she might consider a particular cast $\boldsymbol{a}$ and estimate the expected revenue of a movie with that cast, $\mathbb{E}\left[Y_i(\boldsymbol{a})\right]$.

Classical causal inference from observational data is a difficult enterprise and requires strong assumptions. The challenge is that the dataset is limited; it contains the revenue of each movie, but only at its assigned cast. However, what this paper is about is that the producer's problem is not a classical causal inference. While causal inference usually considers a single possible cause, such as whether a subject receives a drug or a control, our producer is considering a *multiple causal inference*, where each actor is a possible cause. This paper shows how multiple causal inference can be easier than classical causal inference. Thanks to the multiplicity of causes, the producer can make valid causal inferences under weaker assumptions than the classical approach requires.

Title	Cast	Revenue
Avatar	{Sam Worthington, Zoe Saldana, Sigourney Weaver, Stephen Lang, . . . }	\$2788M
Titanic	{Kate Winslet, Leonardo DiCaprio, Frances Fisher, Billy Zane, . . . }	\$1845M
The Avengers	{Robert Downey Jr., Chris Evans, Mark Ruffalo, Chris Hemsworth, . . . }	\$1520M
Jurassic World	{Chris Pratt, Bryce Dallas Howard, Irrfan Khan, Vincent D'Onofrio, . . . }	\$1514M
Furious 7	{Vin Diesel, Paul Walker, Dwayne Johnson, Michelle Rodriguez, . . . }	\$1506M
⋮	⋮	⋮

Table 1. Top earning movies in the TMDB dataset

Let's discuss the producer's inference in more detail: how can she calculate $\mathbb{E}\left[Y_i(\boldsymbol{a})\right]$? Naively, she subsets the data in Table 1 to those with cast equal to $\boldsymbol{a}$, and then computes a Monte Carlo estimate of the revenue. This procedure is unbiased when $\mathbb{E}\left[Y_i(\boldsymbol{a})\right] = \mathbb{E}\left[Y_i(\boldsymbol{a}) \,|\, \boldsymbol{A}_i = \boldsymbol{a}\right]$.

But there is a problem. The data in Table 1 hide *confounders*, variables that affect both the causes and the effect. For example, every movie has a genre, such as comedy, action, or romance. This genre has an effect on both who is in the cast and the revenue. (E.g., action movies cast a certain set of actors and tend to make more money than comedies.) When left unobserved, the genre of the movie produces a statistical dependence between whether an actor is in it and its revenue; this dependence biases the causal estimates, $\mathbb{E}\left[Y_i(\boldsymbol{a}) \,|\, \boldsymbol{A}_i = \boldsymbol{a}\right] \neq \mathbb{E}\left[Y_i(\boldsymbol{a})\right]$.

Thus the main activities of classical causal inference are to identify, measure, and control for confounders. Suppose the producer measures confounders for each movie w_i. Then inference is simple: use the data (now with confounders) to take Monte Carlo estimates of $\mathbb{E}\left[\mathbb{E}\left[Y_i(\boldsymbol{a}) \,|\, W_i, \boldsymbol{A}_i = \boldsymbol{a}\right]\right]$; this iterated expectation "controls" for the confounders. But the problem is that whether the estimate is equal to $\mathbb{E}\left[Y_i(\boldsymbol{a})\right]$ rests on a big and uncheckable assumption: there are no other confounders. For many applied causal inference problems, this assumption is a leap of faith.

We develop *the deconfounder*, an alternative method for the producer who worries about missing a confounder. First the producer finds and fits a good latent-variable model to capture the dependence among actors. It should be a factor model, one that contains a per-movie latent variable that renders the assigned cast conditionally independent. (Probabilistic principal component analysis [40] is a simple example, but there are many others.) Given the model, she then estimates the per-movie variable for each cast in the dataset; this estimated variable is a substitute for unobserved confounders. Finally, she controls for the substitute confounder and obtains valid causal inferences.

The deconfounder capitalizes on the dependency structure of the observed casts, using patterns of how actors tend to appear together in movies as indirect evidence for confounders in the data. Thus the producer replaces an uncheckable search for possible confounders with the checkable goal of building a good factor model of observed casts.

All methods for causal inference using observational data are based on assumptions. Here we make two. First, we assume that the fitted latent-variable model is a good model of the assigned causes. Happily, this assumption is testable; we will use predictive checks to assess how well the fitted model captures the data. Second, we assume that there are no unobserved single-cause confounders, variables that affect one cause (e.g., actor) and the potential outcome function (e.g., revenue). While this assumption is not testable, it is weaker than the usual assumption of ignorability, i.e., no unobserved confounders.

Beyond making movies, many causal inference problems, especially from observational data, also classify as multiple causal inference. Such problems arise in many fields.

- **Genome-wide association studies (GWAS).** In GWAS, biologists want to know how genes causally connect to traits [39, 42]. The assigned causes are alleles on the genome, often encoded as either being common ("major") or uncommon ("minor"), and the effect is the trait under study. Confounders, such as shared ancestry among the population, bias naive estimates of the effect of genes.
- **Computational neuroscience.** Neuroscientists want to know how specific neurons or brain measurements affect behavior and thoughts [3]. The possible causes are multiple measurements about the brain's activity, e.g., one per neuron, and the effect is a measured behavior. Confounders, particularly through dependencies among neural activity, bias the estimated connections between brain activity and behavior.
- **Social science.** Sociologists and policy-makers want to know how social programs affect social outcomes, such as poverty levels and upward mobility [25]. However, individuals may enroll in several such programs, blurring information about their possible effects. In social science, controlled experiments are difficult to engineer; using observational data for causal inference is typically the only option.
- **Medicine.** Doctors want to know how medical treatments affect the progression of disease. The multiple causes are medications and procedures; the outcome is a measurement of a disease (e.g., a lab test). There are many confounders—such as when and where a patient is treated or the treatment preferences of the attending doctor—and these variables bias the estimates of effects. While gold-standard data from clinical trials are expensive to obtain, the abundance of electronic health records could inform medical practices.

Causal inference in each of these fields can use the deconfounder. Fit a good factor model of the assigned causes, infer substitute confounders, and use the substitutes in causal inference.

Related work. The deconfounder relates to several threads of research in causal inference.

Probabilistic modeling for causal inference. [24] use Gaussian processes to depict causal mechanisms; [45] study post-nonlinear causal models and their identifiability; [22] builds on sparse methods to infer causal structures; [23] use factor models to generalize the self-controlled case series method to multiple causes and multiple outcomes. [19] use variational autoencoders to infer unobserved confounders, [36] develop projection-based techniques for high-dimensional covariance estimation under latent confounding, and [13] leverages information theory principles to differentiate causal and confounded connections.

With a related goal, [41] build implicit causal models. They take an explicit causal view of genome-wide association studies (GWAS), treating the single-nucleotide polymorphisms (SNPs) as the multiple causes. They connect implicit probabilistic models and nonparametric structural equation models for causal inference [26], and develop inference algorithms for capturing shared confounding. [8] studies the same scenario with linear regression, where observing many causes makes it possible to account for shared confounders. Multiple causal inference and latent confounding was also formalized by [29], who take an information-theoretic approach.

These papers use Pearl's framework [26]; they hypothesize a causal graph with confounders, causes, and outcomes. This paper complements these works. We develop the deconfounder in the potential outcomes framework [10, 33, 34].

Analyzing GWAS. In GWAS, latent population structure is an important unobserved confounder. [28] propose a probabilistic admixture model for unsupervised ancestry inference. [27] and [1] estimate the unobserved population structure using the principal components of the genotype matrix. [43] and [14] estimate the population structure via the "kinship matrix" on the genotypes. [38] and [7] rely on factor analysis and admixture models to estimate the population structure. [6] adopt a similar idea to study the effect of genetic variations on gene expression levels. These methods can be seen as variants of the deconfounder. The deconfounder gives them a rigorous causal justification, provides principled ways to compare them, and suggests an array of new approaches.

Assessing the ignorability assumption. [32] demonstrates that ignorability and a good propensity score model are sufficient to perform causal inference with observational data. Many subsequent efforts assess the plausibility of ignorability. For example, [31, 5, 9] develop sensitivity analysis in various contexts, though focusing on data with a single cause. In contrast, this work uses predictive model checks to assess unconfoundedness with multiple causes. More recently, [37] leveraged auxillary outcome data to test for confounding in time series data; [12, 11, 17] developed tests for non-confounding in multivariate linear regression. Here we work without auxiliary data, focus on causal estimation, as opposed to testing, and move beyond linear models.

The (generalized) propensity score. [35, 21, 16] and many others develop and evaluate different models for assigned causes. In particular, [2] introduce a semiparametric assignment model; they propose a principled way of correcting for the bias that arises when regularizing or overfitting the assignment model. This work introduces latent variables into the model. The multiplicity of causes enables us to infer these latent variables and then use them as substitutes for unobserved confounders.

Classical causal inference with multiple treatments. [18, 20, 44, 30, 15, 4] extend classical matching, subclassification, and weighting to multiple treatments, always assuming ignorability. This work relaxes that assumption.

References

[1] Astle, W., Balding, D. J., et al. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 24(4):451–471.

[2] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2017). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*.

[3] Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., and Shenoy, K. V. (2012). Neural population dynamics during reaching. *Nature*, 487(7405):51.

[4] Feng, P., Zhou, X.-H., Zou, Q.-M., Fan, M.-Y., and Li, X.-S. (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine*, 31(7):681–697.

[5] Gilbert, P. B., Bosch, R. J., and Hudgens, M. G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in hiv vaccine trials. *Biometrics*, 59(3):531–541.

[6] GTEx Consortium, Battle*, A., Brown*, C. D., Engelhardt*, B. E., and Montgomery*, S. M. (2017). Genetic effects on gene expression across human tissues. *Nature*, 550:204–213.

[7] Hao, W., Song, M., and Storey, J. D. (2015). Probabilistic models of genetic variation in structured populations applied to global human studies. *Bioinformatics*, 32(5):713–721.

[8] Heckerman, D. (2018). Accounting for hidden common causes when inferring cause and effect from observational data. *arXiv preprint arXiv:1801.00727*.

[9] Imai, K. and Van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866.

[10] Imbens, G. and Rubin, D. (2015). *Causal Inference in Statistics, Social and Biomedical Sciences: An Introduction*. Cambridge University Press.

[11] Janzing, D. and Schölkopf, B. (2018a). Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, 6(1).

[12] Janzing, D. and Schölkopf, B. (2018b). Detecting non-causal artifacts in multivariate linear regression models. *arXiv preprint arXiv:1803.00810*.

[13] Kaltenpoth, D. and Vreeken, J. (2019). We are not your real parents: Telling causal from confounded using mdl. *arXiv preprint arXiv:1901.06950*.

[14] Kang, H. M., Sul, J. H., Zaitlen, N. A., Kong, S.-y., Freimer, N. B., Sabatti, C., Eskin, E., et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, 42(4):348.

[15] Lechner, M. (2001). Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric Evaluation of Labor Market Policies*, pages 43–58. Springer.

[16] Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*, 29(3):337–346.

[17] Liu, F. and Chan, L. (2018). Confounder detection in high dimensional linear models using first moments of spectral measures. *arXiv preprint arXiv:1803.06852*.

[18] Lopez, M. J., Gutman, R., et al. (2017). Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science*, 32(3):432–454.

[19] Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. (2017). Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pages 6449–6459.

[20] McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*, 32(19):3388–3414.

[21] McCaffrey, D. F., Ridgeway, G., and Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4):403.

[22] Mckeigue, P., Krohn, J., Storkey, A. J., and Agakov, F. V. (2010). Sparse instrumental variables (spiv) for genome-wide studies. In *Advances in Neural Information Processing Systems*, pages 28–36.

[23] Moghaddass, R., Rudin, C., and Madigan, D. (2016). The factorized self-controlled case series method: An approach for estimating the effects of many drugs on many outcomes. *Journal of Machine Learning Research*, 17(185):1–24.

[24] Mooij, J. M., Stegle, O., Janzing, D., Zhang, K., and Schölkopf, B. (2010). Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in Neural Information Processing Systems*, pages 1687–1695.

[25] Morgan, S. and Winship, C. (2015). *Counterfactuals and Causal Inference*. Cambridge University Press, 2nd edition.

[26] Pearl, J. (2009). *Causality*. Cambridge University Press, 2nd edition.

[27] Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904.

[28] Pritchard, J. K., Stephens, M., Rosenberg, N. A., and Donnelly, P. (2000). Association mapping in structured populations. *The American Journal of Human Genetics*, 67(1):170–181.

[29] Ranganath, R. and Perotte, A. (2018). Multiple causal inference with latent confounding. *arXiv preprint arXiv:1805.08273*.

[30] Rassen, J. A., Solomon, D. H., Glynn, R. J., and Schneeweiss, S. (2011). Simultaneously assessing intended and unintended treatment effects of multiple treatment options: a pragmatic "matrix design". *Pharmacoepidemiology and Drug Safety*, 20(7):675–683.

[31] Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 1–94. Springer.

[32] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

[33] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.

[34] Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

[35] Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H., and Brookhart, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology*, 20(4):512.

[36] Shah, R. D. and Meinshausen, N. (2018). Rsvp-graphs: Fast high-dimensional covariance matrix estimation under latent confounding. *arXiv preprint arXiv:1811.01076*.

[37] Sharma, A., Hofman, J. M., and Watts, D. J. (2016). Split-door criterion for causal identification: Automatic search for natural experiments. *arXiv preprint arXiv:1611.09414*.

[38] Song, M., Hao, W., and Storey, J. D. (2015). Testing for genetic associations in arbitrarily structured populations. *Nature Genetics*, 47(5):550–554.

[39] Stephens, M. and Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics*, 10(10):681.

[40] Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.

[41] Tran, D. and Blei, D. M. (2017). Implicit causal models for genome-wide association studies. *arXiv preprint arXiv:1710.10742.*

[42] Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., and Yang, J. (2017). 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics*, 101(1):5–22.

[43] Yu, J., Pressoir, G., Briggs, W. H., Bi, I. V., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203.

[44] Zanutto, E., Lu, B., and Hornik, R. (2005). Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *Journal of Educational and Behavioral Statistics*, 30(1):59–73.

[45] Zhang, K. and Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-fifth Conference on Uncertainty in Artificial Intelligence*, pages 647–655. AUAI Press.

Causality in heavy-tailed models

Nicola Gnecco

(joint work with Nicolai Meinshausen, Jonas Peters and Sebastian Engelke)

In recent years, much progress has been made in the analysis of causal relationship between random variables. These methods are not well suited, however, if the causal relationships manifest themselves only in extremes. This work aims to connect the two fields of causal inference and extreme value theory.

The setup is a linear structural causal model, or SCM,

$$X_j := \sum_{k \in \mathrm{pa}(j,G)} \beta_{jk} X_k + \varepsilon_j, \quad j \in V, \tag{1}$$

where $G = (V, E)$ is the underlying DAG with $V = \{1, \ldots, p\}$, $\mathrm{pa}(j, G)$ are the graphical parents of $j \in V$ in G, and $\beta_{jk} > 0$. Moreover, we assume that the noise variables ε_j have regularly-varying tails, a subclass of heavy-tailed distributions. For each pair of variables (X_j, X_k) with cumulative distribution functions F_j, F_k, $j, k \in V$, we define the causal tail coefficient

$$\Gamma_{jk} = \lim_{u \to 1} \mathrm{E}\left[\mathrm{F}_k(X_k) \mid \mathrm{F}_j(X_j) > u\right] \tag{2}$$

that captures asymmetries in the extremal dependence of two random variables. In the population case, the causal tail coefficient is shown to reveal the causal structure if the distribution follows the linear structural causal model defined in (1). In particular, we prove the following result.

Theorem 1. *Consider a heavy-tailed linear SCM over p variables including X_1 and X_2, as described in (1). Then, knowledge of Γ_{12} and Γ_{21} allows us to distinguish the following cases: (a) X_1 causes X_2, i.e., X_1 is an ancestor of X_2, (b) X_2 causes X_1, (c) there is a $j \notin \{1, 2\}$, such that X_j causes X_1 and X_2, (d) none* of the above, i.e., there is no causal link between X_1 and X_2. More precisely, we have the following table.

TABLE 2. Summary of the possible values of Γ_{12} and Γ_{21} and the implications for causality.

	$\Gamma_{21} = 1$	$\Gamma_{21} \in (1/2, 1)$	$\Gamma_{21} = 1/2$
$\Gamma_{12} = 1$		(a) X_1 causes X_2	
$\Gamma_{12} \in (1/2, 1)$	(b) X_2 causes X_1	(c) common cause	
$\Gamma_{12} = 1/2$			(d) no causal link

Theorem 1 holds even in the presence of latent common causes, i.e., confounders, that have the same tail index as the observed variables.

To estimate the causal tail coefficient defined in (2), we introduce the non-parametric estimator

$$\widehat{\Gamma}_{jk} = \frac{1}{k} \sum_{i=1}^{n} \widehat{\mathrm{F}}_k(X_{ik}) \mathbf{1}\{X_{ij} > X_{(n-k),j}\}, \tag{3}$$

where $k = k_n$ depends on the sample size n, $X_{(n-k),j}$ denotes the $(n-k)$-th order statistics of variable X_j, and $\widehat{\mathrm{F}}_k$ is the empirical cumulative distribution function of X_k, $j, k \in V$. Further, we prove the consistency of the non-parametric estimator defined above.

Theorem 2. Let $k_n \in \mathbb{N}$ be an intermediate sequence with

$$k_n \to \infty \quad \text{and} \quad k_n^2/n \to 0, \quad n \to \infty.$$

Then the estimator $\widehat{\Gamma}_{jk}$ defined in (3) is consistent, as $n \to \infty$, i.e., for every $\varepsilon > 0$

$$\lim_{n\to\infty} \mathrm{P}(|\widehat{\Gamma}_{jk} - \Gamma_{jk}| > \varepsilon) = 0, \quad j, k \in V.$$

Based on the non-parametric estimator (3), we propose an algorithm, *greedy ancestral search*, that infers causal structure from finitely many data. The method takes as input $\Gamma \in \mathbb{R}^{p\times p}$, a matrix containing the pairwise causal tail coefficients, and returns a causal order π associated to the underlying DAG G. We show that *greedy ancestral search* produces a correct causal order when the input matrix Γ contains the population coefficients Γ_{jk}, $j, k \in V$. Moreover, we prove that the algorithm retrieves a correct causal order even when the input matrix is based on the estimated coefficients $\widehat{\Gamma}_{jk}$, as $n \to \infty$.

Finally, we compare our method to other well-established approaches in causal inference on synthetic data. It turns out that our algorithm is robust to the presence of confounders and misspecifications of model (1).

References

[1] Gissibl, N. and Klüppelberg, C., *Max-linear models on directed acyclic graphs*, Bernoulli, **24(4A)** (2018), 2693–2720.

[2] Gissibl, N., Klüppelberg, C., and Lauritzen, S., *Identifiability and estimation of recursive max-linear models*, arXiv preprint, **arXiv** (2019), 1901.03556.

[3] Gnecco, N., Meinshausen, N., Peters, J., and Engelke, S., *Causality in heavy-tailed models.*, in preparation, (2019).

[4] Naveau, P., Ribes, A., Zwiers, F., Hannart, A., Tuel, A., and Yiou, P., *Revising return periods for record events in a climate event attribution context*, Journal of Climate, **31(9)** (2018), 3411–3422.

[5] Peters, J. and Bühlmann, P., *Structural intervention distance for evaluating causal graphs*, Neural computation, **27(3)** (2015), 771–799.

Causal discovery in linear non-Gaussian models

Mathias Drton
(joint work with Y. Samuel Wang)

This talk reports on recent work on causal discovery using linear non-Gaussian models that are also known by the acronym LiNGAM. Specifically, we discuss two problems: (i) estimation from high-dimensional data, and (ii) estimation in settings with latent variables. The work on the former problem is described in [1]. Work on the latter problem is still in progress.

The considered graphical causal models are based on recursive systems of linear structural equations. This implies that there is an ordering, σ, of the variables such that each observed variable Y_v is a linear function of a variable specific error term ϵ_v and the other observed variables Y_u with $\sigma(u) < \sigma(v)$. The precise causal relationships, i.e., precisely which other variables the linear functions depend on, can be described using a directed graph, also known as the causal graph. The graph's vertex set V is an index set for the observed variables, and an edge $u \to v$ is drawn for all variables that Y_v is a linear function of. Let $\mathrm{pa}(v)$ be the set of all vertices u with $u \to v$ in the graph. Then the statistical model is determined by the equation system

$$Y_v = \sum_{u \in \mathrm{pa}(v)} \beta_{vu} Y_u + \epsilon_v, \qquad v \in V.$$

Here, the coefficients β_{vu} are unknokwn parameters, and the error terms ϵ_v, $v \in V$, are independent.

It has been previously shown that when the error terms ϵ_v are non-Gaussian, the exact causal graph, as opposed to a Markov equivalence class, can be consistently estimated from observational data. The estimate can be obtained by using the non-Gaussianity to infer a causal ordering σ and then determining the graph via variable selection in regression. The ordering may be inferred in step-wise fashion, identifying in each step an initial/source node r and then forming residuals in regression adjusting for Y_r. However, this step-wise regression adjustment is applicable only in low-dimensional problems in which the sample size exceeds the number of studied variables. We propose a modification of the algorithm that yields consistent estimates of the graph also in high-dimensional settings in which the number of variables may grow at a faster rate than the number of observations but in which the underlying causal structure features suitable sparsity; specifically, the maximum in-degree of the graph is controlled. In a theoretical analysis we give consistency results in the setting of log-concave error distributions.

In the second part we no longer assume that all relevant variables have been observed. Instead, some of the relevant variables may be unobserved. We capture this by allowing some of the errors ϵ_v in the above equation system to be dependent, which is commonly visualized by adding bidirected edges to the directed graph that determines the form of the equations. This yields a mixed graph also termed a path diagram. In this paradigm we focus on discovering the causal structure for models given by bow-free acyclic path diagrams. While these diagrams allow for the presence of latent confounders, they make the simplifying assumption that there may not be both a direct effect and latent confounding between a pair of variables. Our main result is an algorithm to recover a bow-free acyclic path diagram from observational data. The algorithm exploits the fact that in a model given by a bow-free acyclic path diagram the direct effects can be identified from the second moment structure in a stepwise fashion that follows the topological/causal ordering of the variables. Importantly, the identification of the effects no longer proceeds through regression adjustments.

References

[1] Y. S. Wang, M. Drton, *High-dimensional causal discovery under non-Gaussianity*, Biometrika, to appear, arXiv:1803.11273.

Identification And Estimation Via A Modified Factorization Of A Graphical Model

Ilya Shpitser

(joint work with Thomas S. Richardon, Robin J. Evans, James M. Robins, Razieh Nabi, and Eli Sherman)

It is well known that in the absence of hidden common causes, identification of causal effects is given by a truncated factorization known as the g-formula. Our work shows that much of modern non-parametric identification theory may be rephrased as a more complex truncated factorization derived from the factorization of the observed marginal of a hidden variable graphical model defining the *nested Markov model*. Further, viewing identified functionals as a modified factorization directly leads to maximum likelihood inference for causal parameters in hidden variable models.

The nested Markov model is defined on an acyclic directed mixed graph (ADMG) obtained from a hidden variable DAG by the *latent projection* operation [5]. ADMGs contain directed edges ($\to$) representing a direct causal relationship, and bidirected edges ($\leftrightarrow$) representing the presence of unobserved common causes. A latent projection represents an infinite class of hidden variable DAG models that share the set of equality constraints on the observed marginal distribution, and non-parametric identification theory. The nested Markov model associated with the latent projection ADMG obtained from the hidden variable DAG is the set of all distributions that capture *all* equality constraints, including generalized independence constraints, induced on the observed margin by the hidden variable DAG model.

Just as DAG models are defined by factorizations with respect to DAGs, nested Markov models are defined by factorizations with respect to ADMGs. The DAG factorization is constructed from conditional distributions corresponding to singleton vertices in the DAG (given their parents). The nested factorization is similarly constructed from *Markov kernels* [2] corresponding to special sets of vertices in the ADMG, called *intrinsic sets* (given their parents not in the set).

An *intrinsic Markov kernel* $q_S(S|W)$ for every intrinsic set S is a mapping from values of parents of vertices in the set (not themselves in the set) W to joint distributions over variables S in the set. Each such kernel is a particular function of the observed marginal distribution, which is not necessarily a conditional distribution. Every ADMG has a fixed set of intrinsic sets, with the full algorithm for obtaining them, and the corresponding Markov kernels given in [3].

For example, the ADMG in Fig. 1 (e) has the following intrinsic sets: $\{A\}$, $\{B\}$, $\{D\}$, $\{Y\}$, $\{A, C\}$, $\{A, C, Y\}$, $\{C, Y\}$, $\{B, D\}$. and the following Markov kernels: $q_A(A) \equiv p(A)$; $q_B(B|A) \equiv p(B|A)$; $q_D(D|C) \equiv \sum_B p(D|C, B, A)p(B|A)$; $q_Y(Y|D) \equiv \sum_{A,C} p(Y|D, C, B, A)p(C|B, A)p(A)$; $q_{A,C}(A, C|B) \equiv p(C|B, A)p(A)$; $q_{A,C,Y}(A, C, Y|B, D) \equiv p(Y|A, B, C, D)p(C|B, A)p(A)$; $q_{C,Y}(C, Y|A, B, D) \equiv p(Y|A, B, C, D)p(C|B, A)$; $q_{B,D}(B, D|A, C) \equiv p(D|A, B, C)p(B|A)$. Markov kernels defining the nested Markov models naturally capture generalized independence constraints associated with missing edges in the ADMG, and implied by the underlying hidden variable DAG. An example of such a constraint is implicit in the definition of $q_D(D|C)$ in terms of $\sum_B p(D|C, B, A)p(B|A)$ that appears at first glance to depend on values of A, but in fact does not under the model.

The nested Markov factorization expresses the joint distribution, and certain other distributions derived from the joint, as products of intrinsic Markov kernels. As an example, the nested Markov model for the ADMG in Fig. 1 implies the following identities, among others:

$$\begin{aligned} p(A, B, C, D, Y) &= q_{A,C,Y}(A, C, Y|B, D)q_{B,D}(B, D|A, C), \\ p(A, B, C, D) &= q_{A,C,Y}(A, C|B, D)q_{B,D}(B, D|A, C), \\ p(Y(d), C(d), B(d), A(d)) &= q_{A,C,Y}(A, C, Y|B, D = d)q_B(B|A). \end{aligned}$$

Note that the last object corresponds to a counterfactual distribution. In fact, counterfactual distributions $p(Y(a))$ are identified in any hidden variable causal model with the latent projection ADMG $\mathcal{G}$ *if and only if* $p(Y^*(a))$ factorizes into intrinsic Markov kernels in $\mathcal{G}$, where Y^* is all variables with a directed path into Y not through A in $\mathcal{G}$ [3]. In other words, identification of counterfactual distributions in the presence of hidden variables is closely related to the existence of the nested Markov factorization for that distribution.

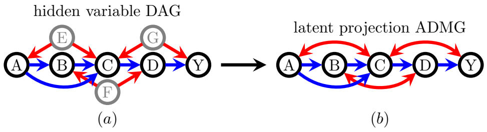


Figure 1. (a) A hidden variable DAG, and (b) the corresponding latent projection ADMG.

The nested Markov factorization directly leads to likelihoods where parameters are associated with intrinsic Markov kernels. For binary data, such parameters are of the form $q_S(S = 0|W)$, for every intrinsic Markov kernel $q_S(S|W)$ [1]. Such a parameterization for a binary nested Markov model associated with Fig. 1 yields 25 parameters, while a naive parameterization in terms of conditional probabilities will yield 31. Just as the case for DAG models, these savings become dramatic as the dimensionality of the model increases, provided the graph remains sparse. A parameterization in terms of path coefficients of a linear structural equation model [6] also exists for Gaussian distributions in the nested Markov model [4]. Maximum likelihood estimation algorithms have been developed for parameters of these likelihoods. These algorithms directly lead to plug-in estimators for all non-parametrically identified causal quantities.

References

[1] R. J. Evans and T. S. Richardson. Smooth, identifiable supermodels of discrete DAG models with latent variables. *Bernoulli*, 2018. (to appear).

[2] S. L. Lauritzen. *Graphical Models*. Oxford, U.K.: Clarendon, 1996.

[3] T. S. Richardson, R. J. Evans, J. M. Robins, and I. Shpitser. Nested Markov properties for acyclic directed mixed graphs, 2017. Working paper.

[4] I. Shpitser, R. J. Evans, and T. S. Richardson. Acyclic linear sems obey the nested markov property. In *Proceedings of the 34th Annual Conference on Uncertainty in Artificial Intelligence (UAI-18)*, 2018.

[5] T. S. Verma and J. Pearl. Equivalence and synthesis of causal models. Technical Report R-150, Department of Computer Science, University of California, Los Angeles, 1990.

[6] S. Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.

Conditional variance penalties and domain shift robustness

Christina Heinze-Deml
(joint work with Nicolai Meinshausen)

In this work [1], we study the robustness of image classification with deep neural networks under a certain class of distribution shifts where the distributions of interest are formulated in terms of a causal model. Conceptually, one can reason about the latent features that manifest themselves in an image as follows: we can distinguish between (i) latent 'core' features X^{core} whose distribution $X^{\text{core}}|Y$, conditional on the class Y, does not change substantially across domains and (ii) latent ‘style’ features X^{style} whose distribution $X^{\text{style}}|Y$ can change substantially across domains. For instance, features like position, rotation, image quality or brightness are considered style features. We propose an estimator that is robust under changes in the distribution of these style features.

Our work relates to “classical” distributional robust inference as follows. In that line of work, the target of inference is

$$\text{argmin}_\theta \sup_{F \in \mathcal{F}} E_F(\ell(Y, f_\theta(X)))$$

for a given set $\mathcal{F}$ of distributions, twice differentiable and convex loss ℓ, and prediction $f_\theta(x)$. For instance, $\mathcal{F}$ can be of the form $\mathcal{F} = \mathcal{F}_\epsilon(F_0)$ with

$$\mathcal{F}_\epsilon(F_0) := \{\text{distributions } F \text{ such that } D(F, F_0) \leq \epsilon\},$$

with a small constant $\epsilon > 0$ and $D(F, F_0)$ being, for example, a ϕ-divergence (e.g. [2, 3, 4]) or a Wasserstein distance (e.g. [5]). In contrast to considering robustness with respect to such pre-defined classes of distributions, we express $\mathcal{F}$ in terms of a causal model where $\mathcal{F}$ is the set of distributions that are generated by interventions on the latent style features.

While we expect that the distribution $X^{\text{style}}|Y$ may change substantially across domains, we assume that the domain itself is not observed and hence a latent variable. Therefore, we do not know a priori which features are subject to distributional shifts and which features have a stable conditional distribution. However, we do assume that we can sometimes observe a typically discrete identifier or “ID variable”. In some applications we know, for example, that two images show the same person, and ID then refers to the identity of the person. The proposed method requires only a small fraction of images to have ID information. We group observations if they share the same class and identifier $(Y, \text{ID}) = (y, \text{id})$ and penalize the conditional variance of the prediction or the loss if we condition on (Y, ID). This conditional variance regularization (CoRe) is shown to protect asymptotically against shifts in the distribution of the style variables. Empirically, we show that the CoRe penalty improves predictive accuracy substantially in settings where domain changes occur in terms of image quality, brightness and color while we also look at more complex changes such as changes in movement and posture.

References

[1] C. Heinze-Deml and N. Meinshausen, Conditional Variance Penalties and Domain Shift Robustness. *arXiv preprint 1710.11469*, 2017.

[2] H. Namkoong and J.C. Duchi, Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, 2017.

[3] A. Ben-Tal, D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affcted by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[4] J. Bagnell. Robust supervised learning. In *Proceedings of the national conference on artificial intelligence*, 2005.

[5] A. Sinha, H. Namkoong, and J. Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

RSVP-graphs: Fast High-dimensional Covariance Matrix Estimation Under Latent Confounding

Rajen D. Shah

(joint work with Benjamin Frot, Gian-Andrea Thanei, Nicolai Meinshausen)

We consider the problem of estimating a high-dimensional $p \times p$ covariance matrix Σ, given n observations $x_1, \dots, x_n$ of confounded data with covariance $\Sigma + \Gamma\Gamma^T$, where Γ is an unknown $p \times q$ matrix of latent factor loadings. We propose a simple and scalable estimator based on the projection on to the right singular vectors of the observed data matrix, which we call RSVP. Specifically, the simplest version of our estimator takes the form VV^T where V has as columns the right singular vectors of the centred data matrix X with ith row $x_i - \sum_j x_j/n$.

Our theoretical analysis of this method reveals that the estimator concentrates around its expectation at the same rate as that of the empirical covariance when scaled such that the entries are of order 1. Furthermore, the bias in estimating Σ is shown to be of smaller order than the variance provided we are in the high-dimensional setting where $p \gg n$; in this way RSVP exploits a particular blessing of high dimensionality.

We see that in contrast to approaches based on removal of principal components such as [2], RSVP is able to cope well with settings where the smallest eigenvalue of $\Gamma^T\Gamma$ is relatively close to the largest eigenvalue of Σ, as well as when eigenvalues of $\Gamma^T\Gamma$ are diverging fast. It is also able to handle data that may have heavy tails and our theory only relies on the data having an elliptical distribution. RSVP does not require knowledge or estimation of the number of latent factors q, but as a consequence only recovers Σ up to an unknown positive scale factor. We argue however that this suffices in many applications. Indeed, in many settings it is the correlation that is of greater interest than the covariance. A further use of the RSVP estimator is to plug it into an existing approach for estimation of the conditional independence graph, such as neighbourhood selection [3].

Whilst the theoretical results for the simple form of the RSVP estimator rely on $p \gg n$, we also show that in more general settings that include $p \asymp n$ for example, we can mimic the high-dimensional setting by computing the estimator on subsamples of the observations and then averaging the results. Interestingly this use of subsampling reduces bias in settings where we do not have $p \gg n$, and inflates the variance of the estimator by at most a factor of $\sqrt{\log(p)}$.

We demonstrate the favourable performance of RSVP both for covariance matrix estimation and estimation of the corresponding conditional independence graph through simulation experiments and an analysis of gene expression datasets collated by the GTEX consortium [1].

References

[1] F. Aguet et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675): 204–213, 2017.

[2] J. Fan, Y. Liao, and M. Mincheva. Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society, Series B*, 75(4):603–680, 2013.

[3] N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34:(3):1436–1462, 2006.

Perspectives for causal discovery in Earth system sciences

Jakob Runge

(joint work with S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M.D. Mahecha, J. Munoz-Mari, E.H. van Ness, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang, J. Zscheischler)

The heart of the scientific enterprise is a rational effort to understand the causes behind the phenomena we observe. While insight in many areas of physics has come from experiments and randomized controlled experiments are a standard approach in medicine and the social sciences, in large-scale complex dynamical systems such as the Earth system, real experiments are rarely feasible. The main current alternative within most disciplines of Earth sciences are computer simulation experiments. However, these are very expensive, time-consuming, and require substantial amounts of expert knowledge, which in turn may impose strong mechanistic assumptions on the system[1]. Fortunately, recent decades have seen an explosion in the availability of large-scale time series data, both from observations (satellite remote sensing, station-based, or field site measurements), and from Earth system model outputs[1]. Such data repositories, together with increasing computational power, open up novel ways to use data-driven methods for the alternative strand of modern science: observational causal discoveries[5]. In recent years, rapid methodological progress has been made in computer science[2, 7], physics[4], and machine learning[3] to infer and quantify potential causal dependencies from time series data without intervening in systems. Unfortunately, many methods are still little known and rarely adopted in Earth system sciences.

In the following, we briefly mention challenges and a way forward for observational causal inference in Earth system sciences which is further elaborated on in a recent publication[5]. As illustrated in Fig. 1, the Earth system poses major challenges to observational causal inference, from characteristics of the underlying processes to properties of the measured data and computational aspects. These challenges are rather generic and apply to many other fields. We suggest a number of avenues for future research: In the short term, the largest potential lies in combining different conceptual approaches that have already shown practical use in Earth sciences[6] in order to address multiple challenges. In the mid-term, it is worth exploring methods that have not been applied to Earth system data, but whose theoretical properties may render them suitable, for example, methods that are based on the principle of independent mechanisms[3]. In the long term, we envision that the two main approaches to understand the Earth system

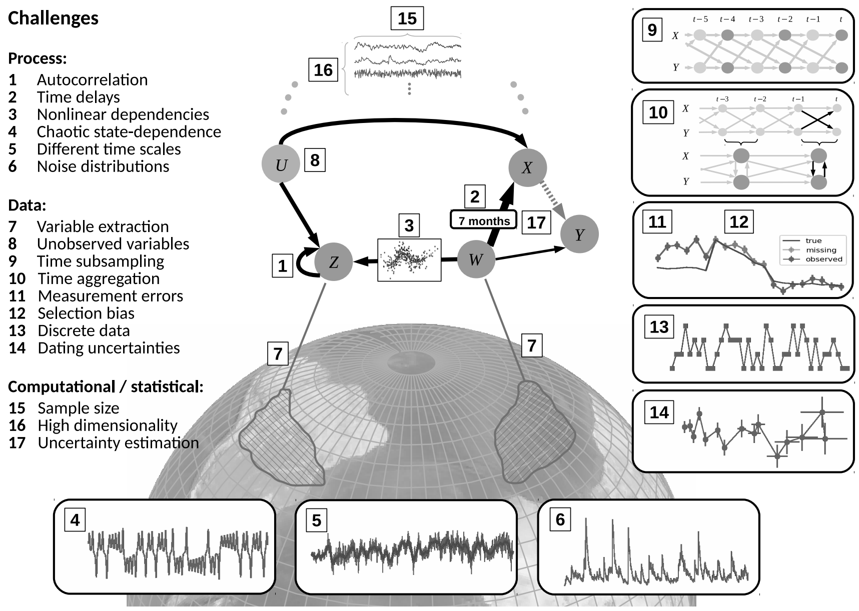


FIGURE 1. Methodological challenges for causal discovery in complex spatio-temporal systems such as the Earth system.

(observational data analysis and Earth system modeling) should become more and more integrated. Causal inference can improve climate model development[5] and physical knowledge (e.g., simulation experiments) can be incorporated into observational causal inference.

A major impediment to a much wider adoption of causal inference methods is the lack of a reliable benchmark database. A recent causality benchmark platform (causeme.net) tries to fill this gap with synthetic models mimicking real data and a call for submissions of real data sets. Sensibly applied causal inference methods promise to substantially advance the state-of-the-art in understanding complex dynamical systems from data also in many other fields with similar challenges as in Earth system sciences, if domain scientists and method developers closely work together–and join the 'causal revolution'.

References

[1] IPCC. *Climate Change 2013: The Physical Science Basis.* Cambridge University Press, Cambridge, MA, 2013.

[2] J. Pearl. *Causality: Models, Reasoning, and Inference.* Cambridge University Press, New York, NY, 2000.

[3] J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms.* MIT Press, Cambridge, MA, 2017.

[4] J. Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos An Interdiscip. J. Nonlinear Sci.*, 28(7):075310, 2018.

[5] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. Mahecha, J. Munoz-Mari, E. van Ness, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang, and J. Zscheischler. Inferring causation from time series in Earth system sciences. *Nat. Commun.*, 10:2553, 2019.

[6] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic. Detecting causal associations in large nonlinear time series datasets. Technical report, 2018.

[7] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search.* MIT Press, Boston, 2000.

Relaxed Causality

Dominik Rothenhäusler
(joint work with Nicolai Meinshausen, Peter Bühlmann, Jonas Peters)

We discuss connections between causal inference, distributional robustness and replicability. It has recently been shown that causal parameters can be written as the solution of a minimax risk problem, where the maximum is taken over a range of perturbed distributions [4]. These perturbed distributions arise from arbitrary interventions on the predictors. Estimating causal parameters, i.e. solving the minimax problem using observational data is only possible under strong assumptions which are often hard to justify in practice. This motivates relaxing the "causal" minimax problem.

Distributional robustness. As causal parameters solve a minimax risk problem, causality can be (but of course does not have to be) understood as a prediction problem with an extreme level of distributional robustness (namely the predictions will still be equally accurate under arbitrarily strong interventions on the predictor variables). Hence, from this perspective, we can relax the causal minimax problem by taking the maximum over a smaller set of interventional distributions. Conceptually, solving the minimax problem for different sets of interventional distributions would allow us to trade off predictive performance under strong perturbations and predictive performance on unperturbed data. In a linear setting where we have access to an exogeneous variable A, this can be achieved in practice by "interpolating" between ordinary least squares and the instrumental variables approach. This motivates anchor regression, which can be seen as a regularization scheme that encourages the estimator to generalize well to certain perturbed data.

Replicability. There are many reasons why scientific results are often not replicable across studies. Arguably, the issue of replicability is partially caused by wrong incentives and a failure of quality control. Here, we look at the issue of replicability from the perspective of distributional shifts. When a study is repeated, often the new observations are sampled from a different distribution than the observations of the original data set. Such distributional shifts may arise in practice, for example, when data is collected in different locations or at different timepoints. One may be interested in screening for associations that are invariant under certain distributional perturbations. We show that if anchor regression and ordinary least squares provide the same answer, then the relationship between target and predictors is unconfounded and the associations are invariant under certain perturbations. We demonstrate this effect on real-world data.

References

[1] C. Heinze-Deml, and N. Meinshausen, *Conditional variance penalties and domain shift robustness*, arXiv preprint arXiv:1710.11469, 2018.

[2] S. Pan, and Q. Yang, *A survey on transfer learning*, IEEE Transactions on Knowledge and Data Engineering, 22(10):1345–1359, 2010.

[3] N. Pfister, S. Bauer, and J. Peters, *Identifying causal structure in large-scale kinetic systems*, arXiv preprint arXiv:1810.11776, 2018.

[4] M. Rojas-Carulla, B. Schölkopf, R. Turner, and J. Peters, *Causal transfer in machine learning*, Journal of Machine Learning Research, 19(36):1–34, 2018.

[5] D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters, *Anchor regression: heterogeneous data meets causality*, arXiv preprint arXiv:1801.06229, 2019.

[6] A. Sinha, H. Namkoong, and J. Duchi, *Certifying some distributional robustness with principled adversarial training*, Sixth International Conference on Learning Representations (ICLR), 2018.

[7] B. Yu and K. Kumbier, *Three principles of data science: predictability, computability, and stability (pcs)*, arXiv preprint arXiv:1901.08152, 2019.

Causal Regularization

Dominik Janzing

While regularization in standard prediction problems aims at avoiding overfitting *finite* data, I argue that it can also be recommended in the infinite sample limit if one is interested in *causal* models rather than purely predictive ones. This is already suggested by the following high-level arguments: On the one hand, regularization may help generalizing across different environmental conditions [1] rather than only across subsamples from the same distribution. On the other hand, models that generalize better across different environmental conditions are believed to be more causal and vice versa, see e.g., [2, 3, 4, 5]. Here I don't consider different environments but a model with a *single* data set for which the causal relation between predictor and target variable is confounded. I show that for this model the effect of confounding is so similar to the effect of overfitting that exactly the same regularization techniques help against both, for details see [6].

Scenario 1 (overfitting): We are given a d-dimensional predictor variable $\mathbf{X}$ and a real-valued target variable Y related by the linear statistical model

$$Y = \mathbf{X}\mathbf{a} + E, \tag{1}$$

where $\mathbf{a} \in \mathbf{R}^d$ and E is an independent noise term. Then the ordinary least squares regression vector is given by the empirical covariance matrices:

$$\hat{\mathbf{a}} := \widehat{\Sigma_{\mathbf{X}\mathbf{X}}}^{-1}\widehat{\Sigma_{\mathbf{X}Y}} = \mathbf{a} + \widehat{\Sigma_{\mathbf{X}\mathbf{X}}}^{-1}\widehat{\Sigma_{\mathbf{X}E}}.$$

For finite data, $\hat{\mathbf{a}} \neq \mathbf{a}$ because the empirical covariance between $\mathbf{X}$ and E is nonzero although we have assumed the covariance to vanish in the population limit. Regularized least squares regression (e.g. Ridge and Lasso) aims at getting closer to $\mathbf{a}$ by minimizing training error plus model complexity (in terms of the norm of $\mathbf{a}$). Indeed, one can show [7] that Ridge and Lasso regression maximizes the posterior likelihood for $\mathbf{a}$, given appropriate priors for $\mathbf{a}$.

Scenario 2 (confounding): We now assume that (1) describes the *causal* relation between $\mathbf{X}$ and Y, while it was just describing a statistical model so far. We assume, moreover, $\Sigma_{\mathbf{X}E} \neq 0$ due to some common cause of $\mathbf{X}$ and E. Therefore, ordinary least squares regression in the population limit yields

$$\hat{\mathbf{a}} = \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{X}Y} = \mathbf{a} + \Sigma_{\mathbf{XX}}^{-1} \Sigma_{\mathbf{X}E}.$$

While $\hat{\mathbf{a}}$ now describes the *statistical* relation between $\mathbf{X}$ and Y correctly, it fails to describe the *causal* relation, which would be described by the vector $\mathbf{a}$.

In both scenarios, recovering $\mathbf{a}$ would be the desired result, but ordinary regression fails because of the covariance of $\mathbf{X}$ and E. While it is the *empirical* covariance in the first case, it is the *population* covariance in the second one. This analogy raises the following question: if regularization helps against overfitting why shouldn't it likewise help against confounding? Why should the algorithms care whether $\hat{\mathbf{a}} \neq \mathbf{a}$ due to a finite sample effect or due to confounding? We just need to assume a generating model for confounding for which $\Sigma_{\mathbf{X}E}$ follows the same distribution as $\widehat{\Sigma_{\mathbf{X}E}}$ in the finite sample case. Such a model has been studied in [8] where the confounder consists of ℓ independent sources $(Z_1, \dots, Z_\ell) =: \mathbf{Z}$. Then, $\mathbf{X}$ is generated from $\mathbf{Z}$ by some fixed mixing matrix M and Y from $\mathbf{Z}$ by a random mixing vector $\mathbf{c}$. Assuming that $\mathbf{c}$ is chosen from $\mathcal{N}(0, \sigma_c^2 \mathbf{I})$ (where σ_c^2 controls the strength of confounding), the covariance vector $\Sigma_{\mathbf{X}E}$ is distributed according to $\mathcal{N}(0, \gamma \Sigma_{\mathbf{XX}})$, where γ can be derived from model parameters like σ_c and ℓ. For finite sample effects in Scenario 1, one can easily show that $\widehat{\Sigma_{\mathbf{X}E}}$ is distributed according to $\mathcal{N}(0, \sigma_E^2 \Sigma_{\mathbf{XX}})$, if $E \sim \mathcal{N}(0, \sigma_E^2)$. Hence we have achieved that $\Sigma_{\mathbf{X}E}$ in Scenario 2 follows the same distribution as $\widehat{\Sigma_{\mathbf{X}E}}$ in Scenario 1 if the noise level σ_E^2 is replaced with γ. If we choose the same priors for $\mathbf{a}$ as for Ridge and Lasso, the unique estimators for $\mathbf{a}$ maximizing the posterior likelihood given the population covariances, are again given by Ridge and Lasso with non-zero penalizing term.

The above tight analogy between overfitting and confounding is achieved by the complexity of the multivariate confounder which generates correlations that 'appear like noise'. This observation is not restricted to linear relations. I was able to prove a 'causal generalization bound' [6] stating that the error made by interpreting any non-linear regression as *causal* model can be bounded from above whenever functions are taken from a not too rich class. In standard statistical learning theory [9], exchangeability ensures that predictive models from not too rich classes generalize from *one subsample to another one.* Here, symmetries of the multivariate confounding model ensure that predictive models from not too rich classes generalize from *observational* to *interventional* data.

References

[1] C. Heinze-Deml and N. Meinshausen. Conditional variance penalties and domain shift robustness. `arXiv:1710.11469`, 2017.

[2] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In Langford J. and J. Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1255–1262. ACM, 2012.

[3] J. Peters, P. Bühlmann, and N. Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

[4] K. Zhang, B. Huang, J. Zhang, C. Glymour, and B. Schölkopf. Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination. *IJCAI : proceedings of the conference*, pages 1347–1353, 2017.

[5] Z.. Shen, P. Cui, K. Kuang, and B. Li. On image classification: Correlation v.s. causality. *CoRR*, abs/1708.06656, 2017.

[6] D. Janzing. Causal regularization. `arXiv:1906.12179`, 2019.

[7] A. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, 2000.

[8] D. Janzing and B. Schölkopf. Detecting non-causal artifacts in multivariate linear regression models. In *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*, 2018.

[9] V. Vapnik. *Statistical learning theory.* John Wileys & Sons, New York, 1998.

Causal mediation with longitudinal mediator and survival outcome

Vanessa Didelez

In causal mediation analysis, we are interested in understanding different mechanisms (causal pathways) of a treatment or exposure affecting some outcomes. Often this is formalised in terms of (in)direct causal effects — popular notions of these are based on so-called ‘nested counterfactuals’, $Y(a, M(a')$. Identification relies crucially on a cross-world independence $Y(a,m) \perp\!\!\!\perp M(a')$. Because of this, the concepts of natural (in)direct effects run into difficulties of interpretation in the particular context of survival analyses, where Y is a survival time and the mediator is a whole process $\{M_t\}$. These problems are:

Problem 1: If survival is shorter, say, under $A = a'$ than under $A = a$, then the second index of $Y(a, \{M_t(a'))\}$ is ‘incomplete’; the nested counterfactual is not well-defined.

Problem 2: Later survival as well as later measurements of the mediator process depend on prior survival. Hence, prior survival acts like a post-treatment confounder and, so, identifiability fails.

In this work, I propose an alternative approach that does not suffer from such shortcomings [1]: this novel approach follows Robins and Richardson [2], where mechanisms need to be specified allowing a separation into the different treatment paths, formalized using an augmented directed acyclic graph (DAG). The graph is hence augmented with nodes A^M and A^Y and with paths $A \to A^M \to M_t$ and $A \to A^Y \to Y$ replacing the edges $A \to M_t$ and $A \to Y$; observationally we have that $A \equiv A^Y \equiv A^M$, but the target of inference becomes an interventional distribution where A^Y and A^M are set to different values; in contrast to the nested counterfactual this reformulation yields a manipulable causal parameter [2]. As this does not involve setting the whole process $\{M_t\}$ to a 'value' and as it separates the edges emanating from treatment A, the proposed approach does not suffer from the above two problems. Moreover, under conditional independence assumptions that can easily be read off the extended DAG, it can be shown that the interventional distribution can be identified from observational data also for survival outcome and longitudinal mediator. The identifying formula is the familiar mediational g-formula. Hence, a number of methods for estimation of these separated effects can be applied, such as g-computation and doubly robust estimation for discrete time points — basically any available method for the g-formula.

While the above methods are well-established for discrete time, the continuous time case also deserves attention. For this case it was demonstrated that for the particular choice of combining a linear model for the mediator with an additive hazard model, the familiar 'path-tracing' formula of linear structural equations can be recovered [3]. For illustration, this method was applied to an example of mediated effects of a blood-pressure treatment on time to kidney failure [3]. We investigated intensive versus standard blood-pressure treatment and found that there is little, and not much time-varying, indirect effect via diastolic blood pressure on kidney failure. Hence, other ways of preventing this side effect of intensive blood-pressure treatment might be worth investigated.

The proposed new approach solves a crucial conceptual problem of mediation analysis with a survival outcome and can be extended to yield much needed clarification in competing risks settings [4]. It is founded in decision theory, avoids genuine counterfactual (cross-world) assumptions and, even in non-survival contexts, constitutes an interesting alternative to the prevailing structural equation modelling.

References

[1] V. Didelez, *Defining causal meditation with a longitudinal mediator and a survival outcome*, Lifetime Data Analysis (2018).

[2] J.M. Robins, T.S. Richardson, *Alternative graphical causal models and the identification of direct effects* In: Causality and psychopathology: Finding the determinants of disorders and their cures (2011) 103–158.

[3] O.O. Aalen, M. Stensrud, V. Didelez, R. Daniel, K. Roysland, S. Strohmaier, *Time-dependent mediators in survival analysis: Modelling direct and indirect effects with the additive hazards model*, Biometrical Journal (2019).

[4] M. Stensrud, J. Young, V. Didelez, J.M. Robins, M. Hernan, *Separable effects for causal inference in the presence of competing risks*, arXiv (2019).

The regression discontinuity design in public health: Continuous and binary outcomes

Sara Geneletti
(joint work with G. Baio, A. O'Keeffe, F.Ricciardi, S. Richardson, L.Sharples)

A Regression Discontinuity (RD) design is a quasi-experimental method for treatment effect estimation, introduced in the 1960's in [35] and widely used in economics and related social sciences [21] and more recently in the medical sciences [17, 6, 32, 29, 27]. The RD design has become of interest in the context of public health as it enables the use of routinely gathered medical data to evaluate the causal effects of drugs when these are prescribed according to well-defined decision rules. This can be very useful as government agencies such as the Federal Drug Administration (FDA) in the US and the National Institute for Health and Care Excellence (NICE) in the UK are increasingly issuing guidelines for drug prescription. Furthermore results can be contrasted to those obtained from randomised controlled trials (RCTs) and inform prescription policy and guidelines based on a more realistic and less expensive context.

We apply the method to evaluating the effect of prescribing statins, a class of cholesterol-lowering drugs on the levels of LDL cholesterol. Further, we evaluate the effect of statin prescription on a binary variable defined by whether an individual reaches recommended LDL cholesterol levels within 6 months of statin prescription.

Guidelines in 2013 in the UK (when and where the data were collected) state that statins should be prescribed to patients with 10-year cardio-vascular disease risk scores in excess of 20%. If we consider patients whose risk scores are close to the 20% risk score threshold, we find that there is an element of random variation in both the risk score itself and its measurement. We can therefore consider the threshold as a randomising device that assigns statin prescription to individuals just above the threshold and withholds it from those just below. Thus, we are effectively replicating the conditions of an RCT in the area around the threshold, removing or at least mitigating confounding.

In a realistic context, doctors do not strictly follow the prescription guidelines and therefore there are patients who are prescribed statins who have a risk score below 20% and patients who are not prescribed statins despite having a risk score above 20%. When this is the case, the RD design is termed "fuzzy", in contrast to the "sharp" design when doctors adhere to the prescription guidelines.

The RD design threshold is a type of instrumental variable, indeed, causal effects from fuzzy RD designs are usually estimated using methods from the IV literature for both continuous and binary outcomes [26, 4, 10]. For a continuous outcome, the Local average treatment effect is commonly used as an estimator whilst for a binary outcome, an IV-based Multiplicative Structural Mean Model (MSMM) is often used. In the binary case the MSMM estimator identifies a risk ratio for the treated (RRT).

These estimators are all developed under a frequentist approach to statistical inference. In our work, we develop a Bayesian approach to the estimation of the LATE and the RTT, within the context of a fuzzy RD design. The use of a Bayesian approach has several benefits when compared to frequentist methods. Firstly, we obtain the variances of our estimates from our posterior samples directly without having to use bootstrapping or other variance approximation approaches (e.g. the Delta method). The Bayesian estimators are very flexible as we can estimate their components within a large number of models. Finally both estimators are ratios and are prone to instability. By adopting a Bayesian approach we are able to impose prior constraints on the estimators in order to stabilise them.

We apply our method to real data from routinely gathered data from doctors in the UK and find that results are broadly in agreement with RCTs. In addition we run a large number of simulation studies and show that our methods compare favourably to competing frequentist approaches.

References

[1] Abadie, A. (2002) Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association*, **97**, 284–292.

[2] — (2003) Semiparametric instrumental variable estimation of treatment response models. *Journal of Econometrics*, **113**, 231–263.

[3] Angrist, J. (2001) Estimation of limited dependent variable models with dummy endogenous regressors: Simple strategies for empirical practice. *Journal of Business & Economic Statistics*, **19**, 2–16.

[4] Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996) Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 444–455.

[5] Balke, A. and Pearl, J. (1997) Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, **92**, 1171–1176.

[6] Bor, J., Moscoe, E., Mutevedzi, P., Newell, M. L. and Barnighausen, T. (2014) Regression discontinuity designs in epidemiology: causal inference without randomized trials. *Epidemiology*, **25**, 729–737.

[7] Burgess, S., Granell, R., Palmer, T. M., Sterne, J. A. C. and Didelez, V. (2014) Lack of identification in semiparametric instrumental variable models with binary outcomes. *Am. J. Epidemiol.*, **180**, 111–119.

[8] Burgess, S. and Thompson, S. G. (2012) Improving bias and coverage in instrumental variable analysis with weak instruments for continuous and binary outcomes. *Statistics in Medicine*, **31**, 1582–1600.

[9] Calonico, S., Cattaneo, M. D. and Titiunik, R. (2015) Robust nonparametric confidence intervals for regression discontinuity designs. *Econometrica*, **82**, 2295 – 2326.

[10] Clarke, P. and Windmeijer, F. (2012) Instrumental variable estimators for binary outcomes. *Journal of the American Statistical Association*, **107**, 1638–1652.

[11] Clarke, P. S., Palmer, T. M. and Windmeijer, F. (2015) Estimating structural mean models with multiple instrumental variables using the generalised method of moments. *Statistical Science*, **30**, 96–117.

[12] Clarke, P. S. and Windmeijer, F. (2010) Identification of causal effects on binary outcomes using structural mean models. *Biostatistics*, **11**, 756–770.

[13] Dawid, A. P. (1979) Conditional independence in statistical theory. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **41**, 1–31.

[14] Didelez, V., Meng, S. and Sheehan, N. A. (2010) Assumptions of iv methods for observational epidemiology. *Statistical Science*, **25**, 22–40.

[15] Gelman, A., A., J., Pittau, M. and Su, Y. (2008) A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, **2**, 1360–1383.

[16] Geneletti, S. and Dawid, A. (2010) The effect of treatment on the treated: a decision theoretic perspective. In *Causality in the Sciences* (eds. M. Ilari, F. Russo and J. Williamson). Oxford University Press.

[17] Geneletti, S., O'Keeffe, A. G., Sharples, L. D., Richardson, S. and Baio, G. (2015) Bayesian regression discontinuity designs: incorporating clinical knowledge in the causal analysis of primary care data. *Statistics in Medicine*, **34**, 2334–2352.

[18] Hahn, J., Todd, P. and Van der Klaauw, W. (2001) Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, **69**, 201–209.

[19] Hernan, M. and Robins, J. (2006) Instruments for causal inference - An epidemiologist's dream? *Epidemiology*, **17**, 360–372.

[20] Imbens, G. and Kalyanaraman, K. (2012) Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, **79**, 933–959.

[21] Imbens, G. W. and Lemieux, T. (2008) Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, **142**, 615 – 635.

[22] van der Klaauw, G. (2008) Regression-discontinuity analysis: A survey of recent developments in economics. *Labour*, **22**, 219–245.

[23] van der Laan, M. J., Hubbard, A. and Jewell, N. P. (2007) Estimation of treatment effects in randomized trials with non-compliance and a dichotomous outcome. *Journal of the Royal Statistical Sociery Series B: (Statistical Methodology)*, **69**, 463–482.

[24] Lambert, D. (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1–14.

[25] Lee, D. S. (2008) Randomized experiments from non-random selection in US House elections. *Journal Of Econometrics*, **142**, 675–697. Conference on the Regression Discontinuity Design, Banff, Canada.

[26] Lee, D. S. and Lemieux, T. (2010) Regression discontinuity designs in economics. *Journal of Economic Literature*, **48**, 281–355.

[27] Linden, A., Adams, J. and Roberts, N. (2006) Evaluating disease management programme effectiveness: an introduction to the regression discontinuity design. *Journal of Evaluation in Clinical Practice*, **12**, 124–131.

[28] McCrary, J. (2008) Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, **142**, 698 – 714. The regression discontinuity design: Theory and applications.

[29] Moscoe, E., Bor, J. and Baernighausen, T. (2015) Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: a review of current and best practice. *Journal of Clinical Epidemiology*, **68**, 132–143.

[30] Palmer, T. M., Ramsahai, R. R., Didelez, V. and Sheehan, N. A. (2011) Nonparametric bounds for the causal effect in a binary instrumental-variable model. *The Stata Journal*, **11**, 345–367.

[31] Plummer, M. (2003) Jags: A program for analysis of bayesian graphical models using gibbs sampling.

[32] Smith, L. M., Kaufman, J. S., Strumpf, E. C. and Levesque, L. E. (2015) Effect of human papillomavirus (HPV) vaccination on clinical indicators of sexual behaviour among adolescent girls: the Ontario Grade 8 HPV Vaccine Cohort Study. *Canadian Medical Association Journal*, **187**, E74–E81.

[33] Stock, J. and Yogo, M. (2005) *Testing for Weak Instruments in Linear IV Regression*, 80–108. New York: Cambridge University Press.

[34] Swanson, S. and Hernan, M. A. (2013) Commentary: How to report instrumental variable analyses (suggestions welcome). *Epidemiology*, **24**, 1044–3983.

[35] Thistlethwaite, D. and Campbell, D. (1960) Regression-Discontinuity Analysis - An alternative to the ex-post-facto experiment. *Journal of Educational Psychology*, **51**, 309–317.

[36] Vansteelandt, S., Bowden, J., Babanezhad, M. and Goetghebeur, E. (2011) On instrumental variables estimation of causal odds ratios. *Statist. Sci.*, **26**, 403–422. http://dx.doi.org/10.1214/11-STS360.

[37] Vansteelandt, S. and Goetghebeur, E. (2003) Causal inference with generalized structural mean models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65**, 817–835.

[38] Ward, S., Jones, L., Pandor, A., Holmes, M., Ara, R., Ryan, A., Yeo, W. and Payne, N. (2007) A systematic review and economic evaluation of statins for the prevention of coronary events. *Health Technology Assessment*, **11**, 1–160.

[39] Windmeijer, F. and Didelez, V. (2016) Methods for binary outcomes. In *Mendelian Randomization: How genes can reveal the biological and environmental causes of disease, To appear* (ed. G. Davey-Smith). Oxford University Press.

Identification of Causal Effects in the Presence of Selection Bias

Jin Tian

(joint work with Juan D. Correa, Elias Bareinboim)

Cause-and-effect relations are one of the most valuable types of knowledge sought after throughout the data-driven sciences since they translate into stable and generalizable explanations as well as efficient and robust decision-making capabilities. Inferring these relations from observational data, however, is a challenging task. Two of the most common barriers to this goal are known as confounding and selection biases. The former stems from the systematic bias introduced during the treatment assignment, while the latter comes from the systematic bias during the collection of units into the sample. We consider the problem of identifying causal effects when both confounding and selection biases are simultaneously present. Specifically, given qualitative causal assumptions in the form of a causal graph G and observational distribution P (possible under confounding bias and selection bias), we study whether a causal effect $P(y|do(x))$ is computable from P.

Identifying Causal Effects from Selection Biased Data

We first investigate the problem of identifiability when all the available data is biased. The problem of selection bias can be modeled through the explicit articulation of the sampling mechanism, S, a binary indicator variable such that $S = 1$ if a unit is included in the sample, and $S = 0$ otherwise. When samples are collected preferentially, the causal effects need to be identified from a biased distribution $P(v|S = 1)$, instead of joint distribution $P(v)$ when the sampling process is entirely random. Given a causal graph G augmented with the selection variable S, we have developed a *complete* algorithm to determine the identifiability of the causal effect $P(y|do(x))$ from biased distribution $P(v|S = 1)$ in G [1, 2]. The algorithm either returns an expression for $P(y|do(x))$ in terms of $P(v|S = 1)$, or, whenever the algorithm returns a failure condition, no identifiability claim about $P(y|do(x))$ can be made by any other method.

Identifying Causal Effects from a Combination of Selection Biased Data and unbiased Data
We then generalize the setting to when, in addition to the biased data, another piece of external data is available, without bias. For example, a subset of the covariates could be measured without bias (e.g., from census). We examine the problem of identifiability when a combination of biased and unbiased data is available. We have developed a new algorithm [2] that, given a causal graph G augmented with the selection variable S, systematically determines the identifiability of the causal effect $P(y|do(x))$ from biased distribution $P(v|S=1)$ and external data $P(t)$ over a subset of the variables. The algorithm subsumes the current state-of-the-art methods, while the completeness of the algorithm is still under investigation.

Covariate Adjustment under Confounding and Selection Biases
Adjusting by a set of covariates is arguably the most widely used technique in practice for causal effects estimation. Although commonly used to control for confounding bias in observational data, adjustment could be used to control for when selection bias is present as well. We generalize the notion of adjustment to account for both confounding and selection biases and leverage external data that may be available without selection bias (e.g., data from census) as well. Formally we introduce the notion of adjustment pair as follows [3]:
Given a causal diagram G augmented with selection variable S, disjoint sets of variables X, Y, Z, and a set $Z^T \subseteq Z$, (Z, Z^T) is said to be an adjustment pair *for recovering the causal effect of X on Y if for every model compatible with G it holds that:*

$$P\left(y|do(x)\right) = \sum_z P(y|x, z, S{=}1) P(z \backslash z^T | z^T, S{=}1) P(z^T).$$

The expression above is a natural extension of the standard adjustment formula $P(y|do(x)) = \sum_z P(y|x,z)P(z)$, and it captures the orthogonal nature of confounding and selection biases while allowing for the use of unbiased data over a subset of the covariates. Furthermore, it incorporates two special cases depending on the types of data available. When all observational data is biased ($Z^T = \emptyset$), it reduces to

$$P(y|do(x)) = \sum_z P(y|x, z, S=1) P(z|S=1).$$

When all covariates are measured unbiasedly ($Z^T = Z$), it reduces to

$$P(y|do(x)) = \sum_z P(y|x, z, S=1) P(z).$$

We have developed a *complete* graphical criterion for when (Z, Z^T) is an adjustment pair for identifying causal effect $P(y|do(x))$ [3]. We further design an algorithm for listing all admissible adjustment pairs in polynomial delay time, which is useful for researchers interested in evaluating certain properties of some admissible pairs but not all (common properties include cost, variance, and feasibility to measure). We also describe a statistical estimation procedure that can be performed once a set is known to be admissible, which entails different challenges in terms of finite samples.

References

[1] E. Bareinboim and J. Tian, Recovering Causal Effects from Selection Bias, in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, January 25-30, 2015. Austin, Texas, USA.

[2] J. Correa, J. Tian, and E. Bareinboim, Identification of Causal Effects in the Presence of Selection Bias, in *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[3] J. Correa, J. Tian, and E. Bareinboim, Generalized Adjustment Under Confounding and Selection Biases, in *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, February 2-7, 2018. New Orleans, Louisiana, USA.

Do ImageNet Classifiers Generalize to ImageNet?

Ludwig Schmidt

(joint work with Benjamin Recht, Rebecca Roelofs, Vaishaal Shankar)

The overarching goal of machine learning is to produce models that *generalize.* We usually quantify generalization by measuring the performance of a model on a held-out test set. What does good performance on the test set then imply? At the very least, one would hope that the model also performs well on a new test set assembled from the same data source by following the same data cleaning protocol.

In this paper, we realize this thought experiment by replicating the dataset creation process for two prominent benchmarks, CIFAR-10 and ImageNet [1, 4]. In contrast to the ideal outcome, we find that a wide range of classification models fail to reach their original accuracy scores. The accuracy drops range from 3% to 15% on CIFAR-10 and 11% to 14% on ImageNet. On ImageNet, the accuracy loss amounts to approximately five years of progress in a highly active period of machine learning research.

Conventional wisdom suggests that such drops arise because the models have been adapted to the specific images in the original test sets, e.g., via extensive hyperparameter tuning. However, our experiments show that the relative order of models is almost exactly preserved on our new test sets: the models with highest accuracy on the original test sets are still the models with highest accuracy on the new test sets. Moreover, there are no diminishing returns in accuracy. In fact, every percentage point of accuracy improvement on the original test set translates to a *larger* improvement on our new test sets. So although later models could have been adapted more to the test set, they see smaller drops in accuracy. These results provide evidence that exhaustive test set evaluations are an effective way to improve image classification models. Adaptivity is therefore an unlikely explanation for the accuracy drops.

Instead, we propose an alternative explanation based on the relative difficulty of the original and new test sets. We demonstrate that it is possible to recover the original ImageNet accuracies almost exactly if we only include the easiest images from our candidate pool. This suggests that the accuracy scores of even the best image classifiers are still highly sensitive to minutiae of the data cleaning process. This brittleness puts claims about human-level performance into context [2, 3, 5]. It also shows that current classifiers still do not generalize reliably even in the benign environment of a carefully controlled reproducibility experiment.

Figure 1 shows the main result of our experiment. To enable future research, we release both our new test sets and the corresponding code.[1]

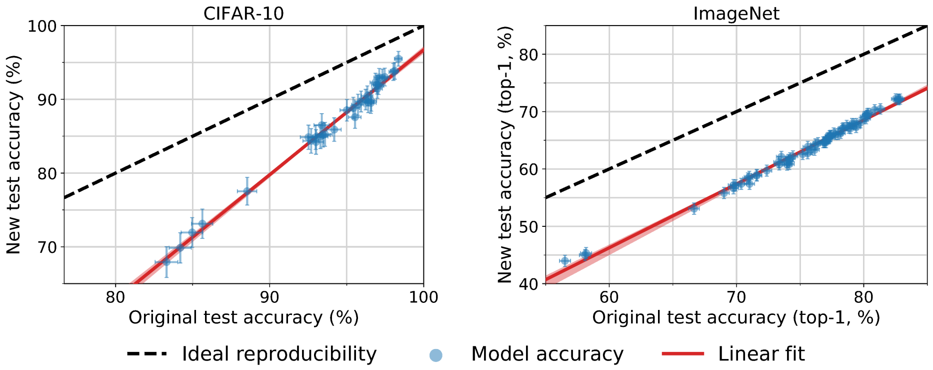


FIGURE 1. Model accuracy on the original test sets vs. our new test sets. Each data point corresponds to one model in our testbed (shown with 95% Clopper-Pearson confidence intervals). The plots reveal two main phenomena: (i) There is a significant drop in accuracy from the original to the new test sets. (ii) The model accuracies closely follow a linear function with slope *greater* than 1 (1.7 for CIFAR-10 and 1.1 for ImageNet). This means that every percentage point of progress on the original test set translates into more than one percentage point on the new test set. The two plots are drawn so that their aspect ratio is the same, i.e., the slopes of the lines are visually comparable. The red shaded region is a 95% confidence region for the linear fit from 100,000 bootstrap samples.

REFERENCES

[1] J. Deng, W. Dong, R. Socher, L. Li, K. Li, L. Fei-Fei, *ImageNet: A large-scale hierarchical image database*, Conference on Computer Vision and Pattern Recognition (CVPR), 2009

[2] K. He, X. Zhang, S. Ren, and J. Sun, *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*, International Conference on Computer Vision (ICCV), 2015

[3] A. Karpathy, *Lessons learned from manually classifying CIFAR-10*, `http://karpathy.github.io/2011/04/27/manually-classifying-cifar10/`, 2011

[1] `https://github.com/modestyachts/CIFAR-10.1` and `https://github.com/modestyachts/ImageNetV2`

[4] A. Krizhevsky, *Learning multiple layers of features from tiny images*, `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`, 2009

[5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg and F. Li, *ImageNet large scale visual recognition challenge*, International Journal of Computer Vision, 2015

Graphical criteria for efficient total effect estimation in causal linear models

Emilija Perković

(joint work with Leonard Henckel, Marloes H. Maathuis)

Covariate adjustment is a popular method for estimating total causal effects from observational data. Graphical criteria have been developed to identify covariate sets that can be used for this purpose. A causal directed acyclic graph (DAG) can be used to represent the underlying causal system when assuming complete knowledge of the underlying causal structure. The best-known such criterion is probably the back-door criterion [6], which is sufficient for adjustment in DAGs.

It is generally not possible to learn the unique causal DAG from observational data. Under the assumptions of causal sufficiency and faithfulness, one can learn a Markov equivalence class of DAGs, which is uniquely represented by a completed partially directed acyclic graph (CPDAG) [3]. Given knowledge of some causal relationships one can obtain a refinement of this class, uniquely represented by a maximally oriented partially directed acyclic graph (maximally oriented PDAG) [3]. In the presence of hidden variables, the counterparts of DAGs and CPDAGs are maximal ancestral graphs (MAGs) and partial ancestral graphs (PAGs) [10].

We consider the adjustment criterion for DAGs, CPDAGs, maximally oriented PDAGs, MAGs and PAGs as stated in [8, 9, 7]. This criterion is necessary and sufficient for adjustment and generalizes the work of [11, 12]. Given the complete identification of all adjustment sets for these graphs, the following question naturally arises: If more than one adjustment set is available, which one should be used? While every adjustment set allows for consistent total causal effect estimation, they do so with varying accuracy. We consider graphical criteria for the identification of efficient adjustment sets in terms of the asymptotic variance of their respective total causal effect estimates in causal linear models.

To illustrate the problem, consider the total causal effect of X on Y in the causal DAG $\mathcal{G}$ in Figure 1. The adjustment sets (see Definition 4.3 in [8]) are of the form of the form $\{B\} \cup \mathbf{S}$, where $\mathbf{S} \subseteq \{A, C, D\}$. Hence, the total number of adjustment sets is $2^3 = 8$. Which of these sets should we use in practice?

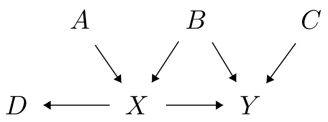


Figure 1. Causal DAG $\mathcal{G}$

When the treatment is a single variable X, the parent set of X is often used as an adjustment set. While this set is easy to compute given a graph, it is also typically quite inefficient in terms of the asymptotic variance, as the parents of X are usually strongly correlated with X. Previous results indicate that the following two notions appear to hold. Adding instrumental variables to a given adjustment set decreases the efficiency, while adding precision variables increases it.

In [2], we build on results from [4] and [5] which we extend in various directions. Our first result is a new graphical criterion (Theorem 3.1 in [2]) that can compare many pairs (but not all) of adjustment sets in terms of the asymptotic variance of the corresponding total effect estimators. Our result holds for causal linear models with arbitrary error distributions, as well as for joint interventions, in DAGs, CPDAGs and maximally oriented PDAGs.

Further, we provide a simple order invariant pruning procedure (Algorithm 1 in [2]) that, given a candidate adjustment set, returns a subset that is also valid and provides a smaller asymptotic variance. Finally, in Theorem 3.10 in [2], we define an adjustment set that provides the smallest possible asymptotic variance among all adjustment sets in the underlying DAG, CPDAG or maximally oriented PDAG.

Theorem 3.1 and Algorithm extend to settings with latent variables and without selection bias, by simply changing d-separation to m-separation [10, 13] in the MAG or PAG and then using Theorem 4.18 from [10] and Lemma 26 from [13].

We consider total effect estimation via covariate adjustment. Other estimators, such as ensemble estimators or the front-door criterion [1] may be more efficient and this is of interest for future research.

References

[1] T. Hayashi, M. Kuroki, *On estimating causal effects based on supplemental variables*, Artificial Intelligence and Statistics, (2014), 312–319.

[2] L. Henckel, E. Perković, M. H. Maathuis, *Graphical criteria for efficient total effect estimation via adjustment in causal linear models*, Working paper, (2019).

[3] C. Meek, *Causal inference and causal explanation with background knowledge*, Proceedings of UAI 1995, 403–410.

[4] M. Kuroki, M. Miyakawa, *Covariate selection for estimating the causal effect of control plans by using causal diagrams*, J. Roy. Stat. Soc. B, **65**, (2003), 209–222.

[5] M. Kuroki, Z. Cai, *Selection of identifiability criteria for total effects by using path piagrams*, Proceedings of UAI 2004, 333–340

[6] J. Pearl, *Comment: Graphical models, causality and intervention*, Stat. Sci., **8** (1993), 266–269.

[7] E. Perković, J. Textor, M. Kalisch, M. H. Maathuis, *A complete generalized adjustment criterion*, Proceedings of UAI 2015.

[8] E. Perković, M. Kalisch, M. H. Maathuis, *Interpreting and using CPDAGs with background knowledge*, Proceedings of UAI 2017.

[9] E. Perković, J. Textor, M. Kalisch, M. H. Maathuis, *Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs*, J. Mach. Learn. Res., **18**, 2018.

[10] T. S. Richardson, P. Spirtes, *Ancestral graph Markov models*, Ann. Stat., **30**, (2002), 962–1030.

[11] I. Shptiser, T. VanderWeele, J. Robins , *On the validity of covariate adjustment for estimating causal effects*, Proceedings of the UAI 2010, 527–536.

[12] B. van der Zander, M. Liskiewicz , J.Textor, *Constructing separators and adjustment sets in ancestral graphs*, Proceedings of UAI 2014, 907–916.

[13] J. Zhang, *Causal reasoning with ancestral graphs*, J. Mach. Learn. Res., **9**, (2008), 1437–1474.

Graphical Models for Missing Data

Karthika Mohan

Missing data (also known as *incomplete data*) are data in which values of one or more variables in a dataset are observed for some samples and missing for the rest. Missingness, which is a rather common phenomenon in practice, can occur due to several reasons such as an ill-designed questionnaire and reluctance of subjects to answer questions on sensitive topics (e.g. income, religion, sexual orientation etc.). Table 1 exemplifies a dataset over two variables in the ideal scenario of no missingness, whereas table 2 exemplifies a dataset with missing values that one would find in the real world. m in table 2 denotes a missing value.

Table 3. Dataset with No Missing Values

Work Exp (in years)	Income (in USD)
3	85,000
1	80,000
10	190,000
6	150,000
8	160,000
15	220,000
18	275,000

Table 4. Dataset with Missing Values

Work Exp (in years)	Income* (in USD)
3	85,000
1	m
10	190,000
6	150,000
8	160,000
15	m
18	m

The bulk of literature on missing data employs procedures that are data-centric as opposed to process-centric and relies on a set of strong assumptions that are primarily untestable (e.g. Missing At Random (MAR) [1]). As a result this area of research is wanting in tools to encode assumptions about the underlying data generating process, methods to test these assumptions and procedures to both decide if quantities of interest are consistently estimable and to compute their estimands whenever they are.

We address these deficiencies by using a graphical representation called "Missingness Graph"[2] (figure 1) which portrays the causal mechanisms responsible for missingness. Using this representation, we define the notion of recoverability, i.e., deciding whether there exists a consistent estimator for a given query. We identify graphical conditions (necessary and sufficient) for recovering joint and conditional

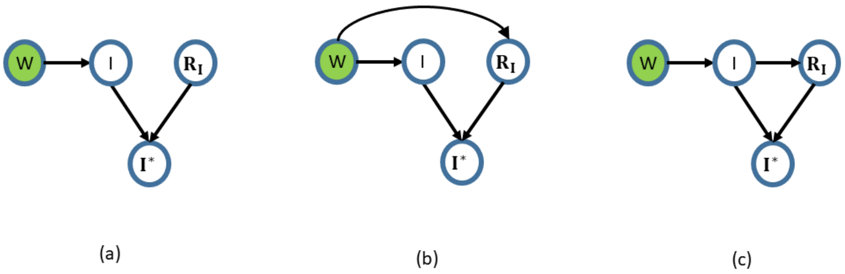


Figure 1. W: Work Experience, I: Income, R_I: Missingness Mechanism, I^*: Proxy for Income. Missingness Graphs depicts (a) Missing Completetly At Random (MCAR) i.e. cause of missingness is random (no edge into R_I) , (b) Missing At Random (MAR) i.e. cause of missingness is fully observed (edge from W to R_I) and (c) Missing Not At Random (MNAR) i.e. cause of missingness is not fully observed (edge from I to R_I).

distributions [2, 3]. Our results apply to missing data problems in all three categories: MCAR, MAR and MNAR, the latter is relatively unexplored. We further address the question of testability i.e. whether an assumed model can be subjected to statistical tests, considering the missingness in the data [4].

Furthermore viewing the missing data problem from a causal perspective has ushered in several surprises. These include recoverability when variables are causes of their own missingness [5], testability of the MAR assumption [4], alternatives to iterative procedures such as Expectation Maximization algorithm [6] and the indispensability of causal assumptions for handling missing data problems [7].

References

[1] Rubin, Donald B. ”Inference and missing data.” Biometrika 63.3 (1976): 581-592.

[2] Mohan, Karthika, Judea Pearl, and Jin Tian. ”Graphical models for inference with missing data.” Advances in neural information processing systems. 2013.

[3] Mohan, Karthika, and Judea Pearl. ”Graphical models for recovering probabilistic and causal queries from missing data.” Advances in Neural Information Processing Systems. 2014.

[4] Mohan, Karthika, and Judea Pearl. ”On the testability of models with missing data.” Artificial Intelligence and Statistics. 2014.

[5] Mohan, Karthika. ”A novel approach to handling hard missing data problems.” ”Beyond Curve Fitting: Causation, Counterfactuals, and Imagination-based AI”. 2019.

[6] Van den Broeck, Guy, Karthika Mohan, Arthur Choi, Adnan Darwiche and Judea Pearl ”Efficient algorithms for Bayesian network parameter learning from incomplete data.” arXiv preprint arXiv:1411.7014 (2014).

[7] Mohan, Karthika, and Judea Pearl. ”Graphical models for processing missing data.” arXiv preprint arXiv:1801.03583 (2018).

Semi-Supervised Learning, Causality and the Conditional Cluster Assumption

Julius von Kügelgen
(joint work with Alexander Mey, Marco Loog, Bernhard Schölkopf)

For the task of predicting a target variable Y from features X, large amounts of unlabelled data (i.e., where only X is observed) are often available at no additional cost. While it is intuitive that this additional information may help, using it often leads to deteriorated performance in practice. When and how such semi-supervised learning (SSL) is possible is thus still not fully understood [1, 2].

In previous work, Schölkopf et al. [3] have established a link between the possibility of SSL and the principle of independent causal mechanisms [4], which states that the conditional distributions of variables given their causal parents are algorithmically independent and thus do not share any information [5]. Since SSL relies on linking $P(X)$ and $P(Y|X)$ via additional assumptions [2], they conclude that SSL should be impossible when predicting a target variable Y from its causes X_C (referred to as *causal learning*, see Figure 1a), but possible when predicting it from its effects X_E (referred to as *anticausal learning*, see Figure 1b) [3].

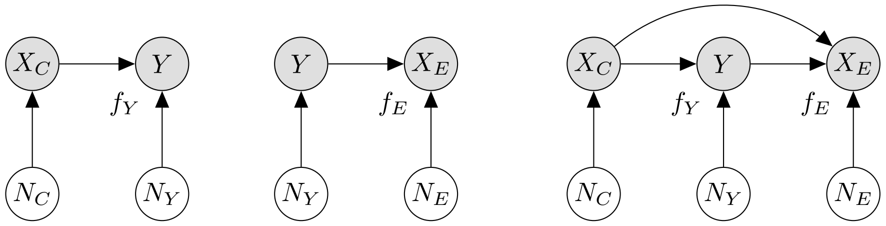
Since both these cases are somewhat restrictive, we extend their work by considering classification using cause and effect features at the same time, see Figure 1c. This setting arises, for example, when predicting disease from both risk factors (e.g., age, sex, diet, smoking, etc.) and clinical symptoms shown by a patient. Formally, we consider data generated from the structural causal model (SCM) [6]

$$X_C := N_C \tag{1}$$

$$Y := f_Y(X_C, N_Y) \tag{2}$$

$$X_E := f_E(X_C, Y, N_E). \tag{3}$$

Analogously to causal learning, we argue that also in our setting $P(X_C)$ does not contain information about the object of interest, $P(Y|X_C, X_E)$, which is purely determined by Eqs. (2) and (3) above and does not depend on Eq. (1).



(a) Causal (b) Anticausal (c) With causes and effects

Figure 1. Causal (A) and anticausal (B) learning settings previously considered for SSL in [3] and our generalisation (C).

On the other hand, considering the causal (4) and non-causal (5) factorisations

$$P(Y, X_E|X_C) = P(Y|X_C)P(X_E|X_C, Y) \tag{4}$$

$$P(Y, X_E|X_C) = P(X_E|X_C)P(Y|X_C, X_E). \tag{5}$$

and arguing with the principle of independent mechanisms, we find that $P(X_E|X_C)$, a quantity of which we can obtain a better estimate from unlabelled data, may share information with $P(Y|X_C, X_E)$. This leads us to our main insight: *the revelant information that additional unlabelled data may provide for prediction is contained in the conditional distribution of effect features given causal features.*

Based on this, we propose to refine the standard cluster assumption for SSL [1, 2] as: *points in the same cluster of $P(X_E|X_C)$ share the same label Y.* We refer to this as the *conditional cluster assumption.* Here, one can think of clusters of $P(X_E|X_C)$ as clusters in the space of functions computing effects from causes, where different functions arise from different choices of Y and N_E in Eq. (3). This idea is illustrated for the case of a binary label Y and additive noise in Figure 2.

Finally, we propose two algorithms for SSL with cause and effect features: a *semi-generative model* $P(Y, X_E|X_C, \theta)$ and a *conditional self-learning* approach.

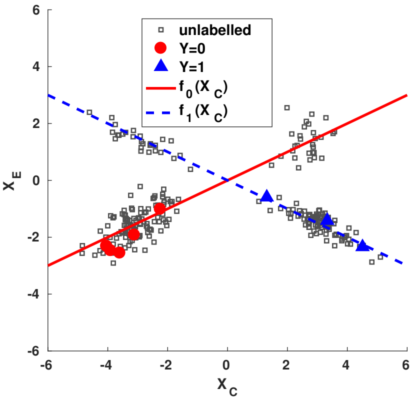


FIGURE 2. Dataset illustrating the conditional cluster assumption.

References

[1] X. Zhu, *Semi-Supervised Learning Literature Survey*, Tech. rep. 1530. Computer Sciences, University of Wisconsin-Madison (2005).

[2] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*, 1st, MIT Press (2006).

[3] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij, *On Causal and Anticausal Learning*, 29th International Conference on Machine Learning (2012), pp. 1-8

[4] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference–Foundations and Learning Algorithms*, MIT Press (2017).

[5] D. Janzing and B. Schölkopf, *Causal Inference Using the Algorithmic Markov Condition*, IEEE Transactions on Information Theory 56.10 (2010), pp. 5168-5194

[6] J. Pearl, *Causality*, Cambridge university press (2000).

Causal Consistency of SEMs & Causal Models as Posets of Distributions

Sebastian Weichwald

(joint work with Paul Rubenstein, Stephan Bongers, Joris M. Mooij, Dominik Janzing, Moritz Grosse-Wentrup, Bernhard Schölkopf)

We can often describe the same system with reference to different terminology, levels of detail, and concepts. We can, for example, reason about individual neurons' firing rates, about average blood oxygen levels in different brain regions, or about electromagnetic activity of so-called cortical dipoles and about how any of those maintain faster reaction times or certain movements. We discuss the following conceptual challenge that is fundamental to causal modelling of real-world systems such as, for example, the brain: How can we formally characterise the relata, aggregate features, and representations that are suitable for a pragmatically useful causal model and how do different description levels relate to one another? The variables we can and do measure do not necessarily lend themselves as is for a causal description.

In [1] we develop a general framework to characterise when two causal models of the same system are causally consistent with one another and agree in their predictions of the effects of interventions. The link between two models is established by the variable transformation that maps the relata of one model onto the relata of the other. We define *exact transformations* that characterise the required properties in order to preserve causal reasoning. Transformations here may correspond to some chosen preprocessing and feature extraction steps or reflect our limited ability to measure the underlying system. Instead of reasoning about how individual pixel colours in an image affect brain activity we may first segment it and identify the objects therein and then model the relationship between neuronal activity and the presence and position of objects in a visual scene. An example of an inevitable measurement transformation is electroencephalography (EEG) where we cannot measure the underlying cortical signals directly but only electrode signals that are a linear superposition thereof.

This framework provides a formal account of how transformations of variables either break or preserve causal reasoning and how transformations may be even necessary to enable causal modelling of the underlying system in the first place. Importantly, this provides theoretical justification for the applicability of causal modelling tools in real-world situations where (a) we only measure and model a sub-system of the world, i.e. where variables 'irrelevant' to or outside of this sub-system are implicitly being marginalised out, (b) we seek a description based on macro-level features that are aggregates of underlying micro-level variables, or (c) we have only access to observations at particular points in time of an underlying time-evolving dynamical system.

This take on the interplay between causal reasoning and variable transformations enables one to in principle consider and identify transformations that exhibit desired properties, e. g. that allow for ‘simpler’ (in terms of complexity), more ‘interpretable’ (in terms one would need to define precisely), or more ‘robust’ (against interventional regime changes) causal models as compared to using the plain observed variables. For example, [2] considers the consistent abstraction of causal models via appropriate variable transformations. Robustness to domain shifts resulting from interventions is considered in [3]: The authors argue in favour of a representation that is consistent with the underlying causal structure in order for a learner to adapt faster to new environments and to thus obtain good transfer. Future research may discuss how to soften the restrictive requirements for a transformation to be exact and how to sensibly arrive at a notion of approximate transformations and a meaningful causal interpretation thereof.

References

[1] P. K. Rubenstein*, S. Weichwald*, S. Bongers, J. M. Mooij, D. Janzing, M. Grosse-Wentrup, B. Schölkopf, *Causal Consistency of Structural Equation Models*, Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI) (2017).

[2] S. Beckers, J. Y. Halpern, *Abstracting Causal Models*, Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (2019). Forthcoming.

[3] Y. Bengio, T. Deleu, N. Rahaman, R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, C. Pal, *A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms*, arXiv preprint arXiv:1901.10912v2 (2019).

Max-linear Bayesian networks

Steffen Lauritzen

(joint work with C. Amendola, N. Gissibl, C. Klüppelberg, N. Tran)

Consider a *directed acyclic graph* (DAG) $\mathcal{D} = (V, E)$. A (recursive) linear structural equation system associated with such a DAG has the form

$$X_v = \sum_{u \in \mathrm{pa}(v)} c_{vu} X_u + c_{vv} Z_v, \quad v \in V, \tag{1}$$

where $Z_v, v \in V$ are independent noise variables and $c_{vu}, u \in \mathrm{pa}(v), c_{vv}$ are *structural coefficients.*

For studying dependence among extreme events in a network, it could make sense to consider *recursive max-linear structural equation systems*:

$$X_v = \bigvee_{u \in \mathrm{pa}(v)} c_{vu} X_u \vee c_{vv} Z_v, \quad v \in V, \tag{2}$$

where now $Z_v, v \in V$ are independent *innovations* with *atom free* distributions having support $\mathbb{R}_+$ and $c_{vu}, u \in \mathrm{pa}(v), c_{vv}$ are *positive structural coefficients.* For simplicity we assume $c_{vv} = 1$ for all $v \in V$.

Such *max-linear Bayesian networks* generate distributions which do not admit densities with respect to product measures since, for example, if $V = \{1, 2\}$ and $X_2 = \max(cX_1, Z_2)$, the distribution will have positive mass on the line $x_2 = cx_1$. Further details concerning the basic properties of these models and issues of estimating the associated DAGs and structural coefficients are described in [2, 3, 4, 5].

It follows directly from the construction that max-linear systems as above satisfy basic Markov properties of Bayesian networks as discussed in [6] as these are not associated with the existence of densities. However, special issues associated with properties of the algebraic *max-times semiring* $(\mathbb{R}_+ \cup \{0\}, \vee, \cdot)$ imply that, in general, additional conditional independence properties hold.

The key to revealing these additional independences is associated with exploiting an algebraic representation of these systems using tropical algebraic geometry, see e.g. [1]. This is work in progress.

References

[1] P. Butkovič. *Max-linear Systems: Theory and Algorithms.* Springer, London, 2010.

[2] Gissibl, N. (2018). *Graphical Modeling of Extremes: Max-linear Models on Directed Acyclic Graphs.* PhD thesis, Technical University of Munich.

[3] Gissibl, N. and Klüppelberg, C. (2018). Max-linear models on directed acyclic graphs. *Bernoulli*, 24:2693–2720.

[4] Gissibl, N., Klüppelberg, C., and Lauritzen, S. L. (2019). Identifiability and estimation of recursive max-linear models. arXiv:1901.03556.

[5] Klüppelberg, C. and Lauritzen, S. (2019). Max-linear models for Bayesian networks. arXiv:1901.03948.

[6] Lauritzen, S. L., Dawid, A. P., Larsen, B. N., and Leimer, H.-G. (1990). Independence properties of directed Markov fields. *Networks*, 20:491–505.

Regularization with invariance for adversarial robustness

Fanny Yang

(joint work with Zuowen Wang, Christina Heinze-Deml, Aditi Ragunathan, Sang Michael Xie, John Duchi, Percy Liang)

As deployment of machine learning (ML) systems in the real world has steadily increased over recent years, more and more emphasis is placed on the reliability of the algorithms. It is for example important to understand certain properties of commonly used neural networks, such as invariances to different types of perturbations, from both a security and interpretability point of view. Neural networks trained using standard training have reportedly very low accuracies on perturbed inputs commonly referred to as *adversarial examples*.

This talk presents recent work on using invariance-inducing regularization to improve robustness of against spatial transformations for image classification [7] and how robust training ℓ_∞ perturbations [6] influences generalization performance on standard test accuracy. In both cases, we evaluate the *robust loss*, that is average prediction performance on worst-case transformations (*attacks*) of a test image.

As the expressivity of neural networks has been shown to be high both theoretically and empirically, in our work we compare the effectiveness of different ways to incorporate the inductive bias of invariance against small rotations and translations. There are two main approaches to achieve invariance and robustness: one is augmentation-based and relies on adding artificially modified training data, as in adversarial training or plain data augmentation. A more transformation specific approach is to carefully design specialized architectures to incorporate spatial equivariance based on ideas in [2, 1].

As a theoretical justification for regularized methods, we prove that when the perturbations are from transformation groups, predictors that optimize the robust loss are in fact invariant. Although recent works suggest that there can be a trade-off between robust and standard accuracy in artificially constructed ℓ_∞ perturbation settings [5, 3], we prove that this is fundamentally different for spatial transformations due to their group structure.

Empirically, we find that regularized augmentation-based methods can achieve $\sim 20\%$ relative adversarial error reduction compared to their unregularized counterparts (including adversarial training) without requiring additional computational resources. They empirically even outperform a few traditional spatial-equivariant networks on Cifar-10 and Svhn . Finally, we observe that on Svhn , not only does the robust test accuracy increase with invariance-promoting regularization but it helps to boost standard accuracy as well.

In the second part of the talk we present our paper [6] in which we aim to understand the reasons behind the following phenomenon: Even though adversarial training [4] can be effective at improving the accuracy on such examples (*robust accuracy*), these modified training methods decrease accuracy on natural unperturbed inputs (*standard accuracy*) [4, 3]. This can be observed in the following table for test accuracies on Cifar-10 .

	Standard training	**Adversarial training**
Robust test	3.5%	45.8%
Robust train	-	100%
Standard test	95.2%	87.3%
Standard train	100%	100%

Compared to constructed examples in previous works (e.g. [5, 3]) that exhibit a tradeoff even in the infinite data limit or because of lack of function space capacity, we consider the "best" scenario for adversarial training: the population minimizer of the robust loss also minimizes standard population loss. This mimics practical scenarios where we typically consider perturbations (such as imperceptible ℓ_∞ perturbations) which do not change the output of the Bayes estimator, so that a predictor with both optimal standard and high robust accuracy exists.

In order to disentangle optimization and statistics, we ask *does the tradeoff indeed disappear if we rule out optimization issues?* We answer the above question negatively by constructing a learning problem with a *convex loss* where adversarial training hurts generalization in the *finite sample setting* even when the optimal population predictor is robust. In particular, adversarial training requires more samples to obtain high standard accuracy. Since convexity rules out optimization issues, our example reveals the possibility of a fundamental statistical explanation for the observe trade-off in practice.

In an attempt to understand how predictive this example is of practice, we subsample Cifar-10 and visualize trends in the performance of standard and adversarially trained models with varying training sample sizes. We observe that the gap between the accuracies of standard and adversarial training decreases with larger sample size, mirroring the trends observed in our constructed problem.

References

[1] T. Cohen, M. Welling, *Group equivariant convolutional networks*, International Conference on Machine Learning (2016), 2990–2999

[2] M. Jaderberg, K. Simonyan, A. Zisserman, Koray Kavukcuoglu, *Spatial Transformer Networks*, Advances in Neural Information Processing Systems (2015), 2017–2025

[3] H. Zhang and Y. Yu and J. Jiao and E. P. Xing and L. El Ghaoui and M. I. Jordan, *Theoretically Principled Trade-off between Robustness and Accuracy*, ICML (2019)

[4] A. Madry, A. Makelov, L. Schmidt, D. Tsipras and A. Vladu *Towards Deep Learning Models Resistant to Adversarial Attacks*, International Conference on Learning Representations (2018)

[5] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner and A. Madry *Robustness may be at odds with accuracy*, International Conference on Learning Representations (2019)

[6] A. Raghunathan, S.M. Xie, F. Yang, J. Duchi, P. Liang, *Adversarial Training Can Hurt Generalization*, arXiv preprint (2019) arXiv:1906.06032

[7] F. Yang, Z. Wang, C. Heinze-Deml, *Invariance-inducing regularization using worst-case transformations suffices to boost accuracy and spatial robustness*, ICML Deep Phenomena Workshop (2019)

Robust causal structure learning with some hidden variables

Marloes H. Maathuis

(joint work with Benjamin Frot and Preetam Nandy)

The task of learning causal directed acyclic graphs (causal DAGs) arises in many areas of science and engineering. In such graphs, nodes represent random variables and edges encode direct causal effects. The problem of recovering their structure from observational data is challenging and cannot be tackled without making untestable assumptions [1]. Among other assumptions, causal sufficiency is particularly constraining. Briefly, causal sufficiency requires that there be no hidden (or latent) variables that are common causes of two or more observed variables (hidden confounders).

Existing causal structure learning algorithms typically either assume causal sufficiency (no hidden confounders), or allow arbitrarily many confounders. In this work, we take a middle-ground stance on causal sufficiency by allowing hidden variables while imposing some restrictions on their number and behavior. More precisely, we consider settings where the underlying DAG among the observed variables is sparse, and there are a few hidden variables that have a direct effect on many of the observed ones [2]. This assumptions cover important real-world applications. For example, one can think of batch effects in gene expression data.

We suggest a two stage approach which first removes the effect of the hidden variables and then estimates the Markov equivalence class of the underlying DAG under the assumption that there are no remaining hidden variables. We show that this approach is consistent in certain high dimensional regimes and performs favorably when compared with the state of the art, both in terms of graphical structure recovery and total causal effect estimation.

This talk is based on the paper by [3].

References

[1] P. Pearl, *Causality: Models, Reasoning and Inference*, Cambridge University Press (2009).

[2] V. Chandrasekaran, P. A. Parrilo and A. S. Willsky, *Latent variable graphical model selection via convex optimization*, Annals of Statistics **40** (2012), 1935–1967.

[3] B. Frot, P. Nandy and M. H. Maathuis, *Robust causal structure learning with some hidden variables*, Journal of the Royal Statistical Society Series B, to appear (2019).

A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts

Ezequiel Smucler
(joint work with Andrea Rotnitzky, James M. Robins)

We consider inference about a scalar parameter under a non-parametric model based on a one-step estimator computed as a plug in estimator plus the empirical mean of an estimator of the parameter's influence function. We focus on the class of parameters that have the mixed bias propery, namely, parameters such that the bias of the one-step estimator is equal to the mean of the product of the estimation errors of two nuisance functions.

We show that this class includes many important treatment effect contrasts of interest in causal inference and econometrics, such as ATE, ATT, an integrated causal contrast with a continuous treatment, and the mean of an outcome missing not at random. Moreover the class of parameters with the mixed bias property strictly includes two recently studied classes of parameters ([2], [1]). We characterize the form of parameters with the mixed bias property and of their influence functions. Furthermore, we derive two functional moment equations, each being solved at one of the two nuisance functions, as well as, two functional loss functions, each being minimized at one of the two nuisance functions. These loss functions can be used to derive loss based penalized estimators of the nuisance functions.

We propose estimators of the target parameter that entertain approximately sparse regression models for the nuisance functions allowing for the number of potential confounders to be even larger than the sample size. By employing sample splitting, cross-fitting and ℓ_1-regularized regression estimators of the nuisance functions based on objective functions whose directional derivatives agree with those of the parameter's influence function, we obtain estimators of the target parameter with two desirable robustness properties: (1) they are *rate doubly-robust* in that they are root-n consistent and asymptotically normal when both nuisance functions follow approximately sparse models, even if one function has a very non-sparse regression coefficient, so long as the other has a sufficiently sparse regression coefficient, and (2) they are *model doubly-robust* in that they are root-n consistent and asymptotically normal even if one of the nuisance functions does not follow an approximately sparse model so long as the other nuisance function follows an approximately sparse model with a sufficiently sparse regression coefficient.

References

[1] Chernozhukov, V., Newey, W. K., and Singh, R. (2018b). *Learning L2 Continuous Regression Func tionals via Regularized Riesz Representers*, arXiv e-prints, page arXiv:1809.05224

[2] Robins, J.M., Li, L., Tchetgen, E., and van der Vaart, A. (2008). *Higher order influence functions and minimax estimation of nonlinear functionals*, Volume 2 of Collections, pages 335–421. Institute of Mathematical Statistics, Beachwood, Ohio, USA.

Structural agnostic modeling: An information theoretic approach to causal learning

Michele Sebag

(joint work with D. Kalainathan, O. Goudet, D. Lopez-Paz, I. Guyon)

The talk addresses the problem of uncovering causal structure from multivariate observational data, referred to as *observational causal discovery* [7, 8, 10]. The considered framework is that of Functional Causal Models [9], defined as a pair $(\mathcal{G}, f)$, with $\mathcal{G}$ a directed acyclic graph (DAG) upon random variables $X_1, \dots X_d$, and $f = (f_1, \dots, f_d)$ a set of d causal mechanisms, such that the distribution of variable X_j is defined as:

$$X_j \sim f_j(X_{\mathrm{Pa}(j;\mathcal{G})}, E_j), \text{ with } E_j \sim \mathcal{N}(0,1) \text{ for } j = 1, \dots, d. \tag{1}$$

with $\mathrm{Pa}(j;\mathcal{G})$ the set of parents of X_j, f_j a function from $\mathbb{R}^{|\mathrm{Pa}(j;\mathcal{G})|+1} \to \mathbb{R}$ and E_j a unit centered Gaussian noise, accounting for all unobserved causes of X_j. Observational causal discovery aims to learn both the causal graph and the associated causal mechanisms from samples of the joint probability distribution of observational data, noted $\mathbf{x}^{(\ell)}$, $\ell = 1 \dots n$. The talk has presented two causal discovery approaches handling non-linear causal mechanisms and dealing with non-Gaussian variable and noise distributions.

Causal Generative Neural Network: CGNN starts from the Markov equivalence class of the sought DAG $\mathbf{G}$ and aims to find a generative model of the data [2]. For each candidate DAG $\widehat{\mathbf{G}}$ in the Markov equivalence class of $\mathbf{G}$, CGNN learns causal mechanisms $\hat{f} = (\hat{f}_1, \dots, \hat{f}_d)$, implemented as neural nets, to minimize the Maximum Mean Discrepancy [3] between the original data sample and a sample generated from $(\widehat{\mathbf{G}}, \hat{f})$. Though sound and experimentally accurate, the approach suffers from two limitations: i) it assumes that the Markov equivalence class of the sought DAG is known; ii) it does not scale up w.r.t. the number of variables.

Structural Agnostic Modelling: SAM[1], addressing CGNN limitations, leverages the power of generative adversarial learning to optimize both the structure of the graph, and the causal mechanisms at once [6]. For each variable X_i, SAM learns a specific generative network $\hat{f}_i$ and defines a new variable $\hat{X}_i$ from its conditional distribution w.r.t. all initial variables but X_i, and noise E_i:

$$\hat{X}_i \sim \hat{f}_i(X_1, \dots, X_{i-1}, X_{i+1}, \dots X_d, E_i)$$

All $\hat{f}_i$ are simultaneously learned using stochastic gradient descent, using an adversarial mechanism [1], where the discriminator aims to distinguish the original data samples from the fake samples obtained by replacing for each $\mathbf{x}^{(\ell)}$, its i-th coordinate with $\hat{f}_i(\mathbf{x}_1^{(\ell)}, \dots, \mathbf{x}_{i-1}^{(\ell)}, \mathbf{x}_{i+1}^{(\ell)}, \mathbf{x}_d^{(\ell)}, \varepsilon_i)$, with $\varepsilon_i \sim \mathbf{N}(0, 1)$.

Regularization terms are used to enforce the sparsity and the frugality of the causal mechanisms, in a Lasso-like manner. Lastly, the learning criterion is augmented with a term meant to enforce the DAGness of the causal graph associated with all causal mechanisms.

The fact that SAM uncovers the true DAG in the large sample limit is shown under mild assumptions on the underlying distribution, noting that the set of parents associated to each variable X_i is at most its Markov blanket.

The extensive experimental validation of SAM on artificial, realistic and real-world data shows its robustness compared to the state of the art, with respect to diverse underlying joint distributions (Gaussian and non-Gaussian distributions for the variables and the noise, linear and non-linear causal mechanisms).

Perspectives. An on-going extension regards the case of categorical and mixed variables, taking inspiration from discrete GANs [4]. Another perspective is to relax the causal sufficiency assumption and handle hidden confounders, e.g. by introducing statistical dependencies between the noise variables attached to different variables [11, 5], or via dimensionality reduction [12].

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *NIPS*, 2014.

[2] O. Goudet, D. Kalainathan, P. Caillou, I. Guyon, D. Lopez-Paz, and M. Sebag. Learning functional causal models with generative neural networks. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pages 39–80. Springer, 2018.

[3] A. Gretton, K. M Borgwardt, M. Rasch, B. Schölkopf, A. J Smola, et al. A kernel method for the two-sample-problem. *NIPS*, 2007.

[4] R Devon Hjelm, A. P. Jacob, T. Che, A. Trischler, K. Cho, and Y. Bengio. Boundary-seeking generative adversarial networks. *arXiv preprint arXiv:1702.08431*, 2017.

[5] D. Janzing and B. Schölkopf. Detecting confounding in multivariate linear models via spectral analysis. *Journal of Causal Inference*, 6(1), 2018.

[6] D. Kalainathan, O. Goudet, I. Guyon, D. Lopez-Paz, and M. Sebag. Structural agnostic modeling: Adversarial learning of causal graphs. *arXiv:1803.04929*, 2019.

[7] D. Lopez-Paz, K. Muandet, B. Scholkopf, and I. Tolstikhin. Towards a learning theory of cause-effect inference. *ICML*, 2015.

[1] Available at https://github.com/Diviyan-Kalainathan/SAM

[8] J. M Mooij, J. Peters, D. Janzing, J. Zscheischler, and B. Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *JMLR*, 2016.

[9] J. Pearl. Causality: models, reasoning and inference. *Econometric Theory*, 2003.

[10] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference - Foundations and Learning Algorithms*. MIT Press, 2017.

[11] D. Rothenhäusler, C. Heinze, J. Peters, and N. Meinshausen. Backshift: Learning causal cyclic graphs from unknown shift interventions. In *Advances in Neural Information Processing Systems*, pages 1513–1521, 2015.

[12] Y. Wang and D. M. Blei. The blessings of multiple causes. *arXiv preprint arXiv:1805.06826*, 2018.

Reporter: Christina Heinze-Deml

Participants

Dr. Elias Bareinboim
Department of Computer Sciences
Purdue University
Computer Science Building
West Lafayette, IN 47907-1398
UNITED STATES

Dr. Stefan Bauer
Mathematisches Institut
Universität Tübingen
Auf der Morgenstelle 10
72076 Tübingen
GERMANY

Prof. David Blei
Department of Computer Science
Columbia University
Seeley W. Mudd Building
New York, NY 10027
UNITED STATES

Dr. Leon Bottou
Facebook
770 Broadway
New York NY 10003
UNITED STATES

Prof. Dr. Peter Bühlmann
Seminar für Statistik
ETH Zürich (HG G 17)
Rämistrasse 101
8092 Zürich
SWITZERLAND

Domagoj Ćevid
Departement Mathematik
ETH-Zentrum
Rämistrasse 101
8092 Zürich
SWITZERLAND

Prof. Dr. Gregory F. Cooper
Department of Biomedical Information
5607 Baum Boulevard
Pittsburgh PA 15206
UNITED STATES

Prof. Dr. Manfred Deistler
Institut für Ökonometrie und
Operations Research
Technische Universität Wien
Wiedner Hauptstrasse 8
1040 Wien
AUSTRIA

Prof. Dr. Vanessa Didelez
Leibniz-Institut für
Präventionsforschung und
Epidemiologie - BIPS GmbH
Achterstrasse 30
28359 Bremen
GERMANY

Prof. Mathias Drton
Department of Statistics
University of Washington
Box 35 43 22
Seattle, WA 98195-4322
UNITED STATES

Prof. Dr. Frederick Eberhardt
Division of Humanities and Social
Sciences
CALTECH, MC 101-40
Pasadena, CA 91125
UNITED STATES

Prof. Dr. Sebastian Engelke
Research Center for Statistics
Geneva School of Economics and
Management
40, Boulevard du Pont-d'Arve
1211 Genève 4
SWITZERLAND

Dr. Sara Geneletti
Department of Statistics
London School of Economics
Houghton Street
London WC2A 2AE
UNITED KINGDOM

Nicola Gnecco
Research Center for Statistics
Geneva School of Economics and
Management
40, Boulevard du Pont-d'Arve
1211 Genève 4
SWITZERLAND

Prof. Dr. Niels Richard Hansen
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 København
DENMARK

Dr. Christina Heinze-Deml
Seminar für Statistik
ETH Zürich
Rämistrasse 101
8092 Zürich
SWITZERLAND

Prof. Dr. Aapo Hyvarinen
Gatsby Computational Neuroscience
Unit
University College London
Gower Street
London WC1E 6BT
UNITED KINGDOM

Martin Emil Jakobsen
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 København
DENMARK

Prof. Dr. Dominik Janzing
Max-Planck-Institut für Biologische
Kybernetik
Spemannstraße 38
72076 Tübingen
GERMANY

Prof. Dr. Steffen Lauritzen
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 København
DENMARK

Jinzhou Li
Seminar für Statistik
ETH Zürich
Rämistrasse 101
8092 Zürich
SWITZERLAND

Dr. Lin Liu
Department of Mathematics
Harvard University
Science Center
One Oxford Street
Cambridge MA 02138-2901
UNITED STATES

Chaochao Lu
Department of Engineering
Information Engineering Division
University of Cambridge
Trumpington Street
Cambridge CB2 1PZ
UNITED KINGDOM

Prof. Dr. Marloes Maathuis
Seminar für Statistik
ETH Zürich; HG G 15.1
Rämistrasse 101
8092 Zürich
SWITZERLAND

Prof. Dr. Nicolai Meinshausen
Seminar für Statistik
ETH Zürich
Rämistrasse 101
8092 Zürich
SWITZERLAND

Dr. Karthika Mohan
Postdoctoral Scholar
University of California, Berkeley
621 Sutardja Dai Hall
Berkeley CA 94720-1758
UNITED STATES

Dr. Sach Mukherjee
Deutsches Zentrum für
Neurodegenerative Erkrankungen e.V.
(DZNE)
Ernst-Robert-Curtius-Straße 12
53117 Bonn
GERMANY

Prof. Dr. Whitney K. Newey
Department of Economics
Massachusetts Institute of Technology
Building 52, Rm 424
50 Memorial Drive
Cambridge, MA 02139
UNITED STATES

Emilija Perkovic
Department of Statistics
University of Washington
Padelford Hall, Rm B-313
P.O. Box 354322
Seattle, WA 98195-4322
UNITED STATES

Jonas Peters
Department of Mathematical Sciences
University of Copenhagen
Universitetsparken 5
2100 København
DENMARK

Niklas Pfister
Seminar für Statistik
ETH Zürich
Rämistrasse 101
8092 Zürich
SWITZERLAND

Drago Plecko
Seminar für Statistik
ETH Zürich
Rämistrasse 101
8092 Zürich
SWITZERLAND

Prof. Dr. Thomas S. Richardson
Department of Statistics
University of Washington
P.O. Box 354322
Seattle, WA 98195-4322
UNITED STATES

Prof. Dr. James M. Robins
Department of Biostatistics
Harvard School of Public Health
677 Huntington Avenue
Boston, MA 02115
UNITED STATES

Dr. Dominik Rothenhäusler
Department of Statistics
University of California, Berkeley
367 Evans Hall
Berkeley CA 94720-3860
UNITED STATES

Prof. Dr. Andrea Rotnitzky
Department of Biostatistics
Harvard T.H. Chan School of Public
Health
677 Huntington Avenue
Boston MA, 02115
UNITED STATES

Dr. Jakob Runge
German Aerospace Center (DLR)
Institute of Data Science
Mälzerstr. 3
07745 Jena
GERMANY

Ludwig Schmidt
Department of Computer Science
University of California, Berkeley
Soda Hall
Berkeley CA 94709
UNITED STATES

Prof. Dr. Bernhard Schölkopf
Max Planck Institute for Intelligent Systems
Max-Planck-Ring 4
72076 Tübingen
GERMANY

Prof. Dr. Michèle Sebag
Laboratoire de Recherche en Informatique
CNRS, UMR 8623, Bat. 660
Université Paris Sud
91190 Gif-sur-Yvette
FRANCE

Dr. Rajen Dinesh Shah
Department of Pure Mathematics and Mathematical Statistics
University of Cambridge
Wilberforce Road
Cambridge CB3 0WB
UNITED KINGDOM

Prof. Dr. Ilya Shpitser
Department of Mechanical Engineering
Whiting School of Engineering
Johns Hopkins University
Latrobe Hall 223
3400 North Charles Street
Baltimore MD 21218-2682
UNITED STATES

Dr. Ezequiel Smucler
Departmento de Matemática y Estadistica
Universidad Torcuato di Tella
Sáenz Valiente 1010 (C 1428)
Av. Pres. Figueroa Alcorta 7350
Buenos Aires - CABA
ARGENTINA

Prof. Dr. David Sontag
MIT - EECS- CSAIL
32 Vassar Street
P.O. Box 38-401
Cambridge, MA 02139
UNITED STATES

Prof. Dr. Jin Tian
Department of Computer Science
Iowa State University
226 Atanasoff Hall
Ames, IA 50011
UNITED STATES

Prof. Dr. Caroline Uhler
EECS Department / IDSS
Massachusetts Institute of Technology
32-D634
77 Massachusetts Avenue
Cambridge, MA 02139
UNITED STATES

Julius von Kügelgen
Departement of Engineering
University of Cambridge
Trumpington Street
Cambridge CB2 1PZ
UNITED KINGDOM

Dr. Linbo Wang
Department of Statistical Sciences
University of Toronto
Public Health Building, Rm 372
Sidney Smith Hall
100 St. George Street
Toronto ONT M5S 3G3
CANADA

Sebastian Weichwald
Max Planck Institute for Intelligent Systems
Empirical Inference
Max-Planck-Ring 4
72076 Tübingen
GERMANY

Prof. Dr. Fan Yang
Department of Bioengineering
Stanford University
Shriram Center, Rm 119
443 Via Ortega
Stanford, CA 94305-4245
UNITED STATES

Prof. Dr. Bin Yu
Department of Statistics
University of California, Berkeley
367 Evans Hall
Berkeley CA 94720-3860
UNITED STATES

Prof. Dr. Kun Zhang
Department of Philosophy
Carnegie Mellon University
Baker Hall 161B
5000 Forbes Avenue
Pittsburgh, PA 15213-3890
UNITED STATES