[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

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

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,

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

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

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

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 $A \not\to B \mid C$. 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*.)