An Information-Theoretic Approach on Causal Structure Learning for Heterogeneous Data Characteristics of Real-World Scenarios

Johannes Huegle
University of Potsdam, Hasso Plattner Institute, Potsdam, Germany
johannes.huegle@hpi.de

Abstract

While the knowledge about the structures of a system's underlying causal relationships is crucial within many real-world scenarios, the omnipresence of heterogeneous data characteristics impedes applying methods for causal structure learning (CSL). In this dissertation project, we reduce the barriers for the transfer of CSL into practice with threefold contributions: (1) We derive an information-theoretic conditional independence test that, incorporated into methods for CSL, improves the accuracy for non-linear and mixed discrete-continuous causal relationships; (2) We develop a modular pipeline that covers the essential components required for a comprehensive benchmarking to support the transferability into practice; (3) We evaluate opportunities and challenges of CSL within different real-world scenarios from genetics and discrete manufacturing to demonstrate the accuracy of our approach in practice.

1 Introduction, Background, and Limitations

Causal Structure Learning (CSL) has received widespread attention in the scientific field as the knowledge of underlying causal structures is the basis for decision support within many real-world scenarios [Spirtes *et al.*, 2000]. For example, in discrete manufacturing, the knowledge about causal relationships is the key for a root cause analysis of failures within the complex production processes [Huegle *et al.*, 2020].

In the context of CSL, a Causal Graphical Model (CGM) $\mathcal{G}$ encodes direct causal relationships between a set of random variables $\mathbf{V} = \{X, Y, Z, \dots\}$ as directed edges $X \to Y$, e.g., see [Spirtes *et al.*, 2000]. In particular, for an edge $X \to Y$ within the CGM $\mathcal{G}$ we assume that Y is a function of the direct cause X with independent noise item N, i.e., following a Functional Causal Model (FCM) $Y = f(X, N)$. In this setting, CSL aims to derive as many of the underlying causal relationships in the CGM $\mathcal{G}$ from i.i.d. observational data as possible. Therefore, methods of CSL leverage probabilistic independence characteristics of the variables' joint probability distribution induced by the underlying FCM [Spirtes *et al.*, 2000]. Constraint-based methods for CSL possess the flexibility to handle various data distributions by incorporation of an appropriate statistical hypothesis test for Conditional Independence (CI) and support diverse extensions, which makes them popular in application [Spirtes *et al.*, 2000].

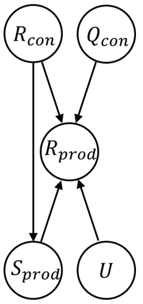

Figure 1: True underlying CGM.

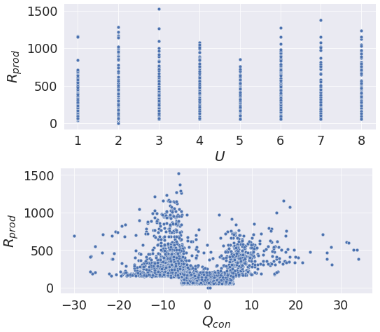

Figure 2: Scatter plots for $U \to R_{prod}$ (top) and $Q_{con} \to R_{prod}$ (bottom).

Given the assumptions of the chosen CI-test, constraint-based methods for CSL require variables of the same type, continuous or discrete, or have strong statistical assumptions on f within the underlying FCM, e.g., linearity with additive i.i.d. Gaussian noise or conditional Gaussian mixtures. In contrast, most real-world scenarios incorporate diverse heterogeneous data characteristics that include non-linear and mixed discrete-continuous relationships.

2 Challenges within Real-World Scenarios

Consider the simplified discrete manufacturing scenario of our cooperation partner with the expected underlying CGM with heterogeneous causal relationships depicted in Fig. 1 and Fig. 2, respectively. Within a configuration phase $_{(con)}$, quality measurements Q_{con} and rejections R_{con} are used for adjustment of the processing speed S_{prod} to reduce the number of rejected goods R_{prod} within a production phase $_{(prod)}$. Besides these causal relationships for configuration, rejections within the production phase R_{prod} vary given the corresponding locality within one of nine existing units U.

In practice, the true statistical properties of the underlying FCM are mostly unknown in advance and the characteristics of causal relationships vary within the observational dataset, e.g., see Fig. 2, such that an inappropriate CI-test yields to incorrect learned causal structures [Spirtes *et al.*, 2000]. Therefore, data is often transformed to be either discrete or continuous to use standard CI-tests to the detriment of the interpretability and accuracy of the learned causal structures.

Moreover, determining the appropriate CSL approach given an observational dataset becomes a tedious manual task. In particular, it requires to examine a selection of state-of-the-art algorithms and parameterizations with different implementations in different programming languages utilizing different hardware setups which have a significant impact on the accuracy and the computational complexity.

3 Approach, and Targeted Contributions

In this dissertation project, we tackle the challenges above to reduce the barriers for the transfer of CSL into practice.

(1) Information-Theoretic CSL for Heterogeneous Data: To allow for weaker assumptions on the causal relationships within the FCM of heterogeneous data, we examine information-theoretic measures that provide a universal examination of probabilistic conditional independence. Therefore, we extend previous work on the k-Nearest Neighbor (k-NN)-based mutual information estimation in line with Gao et al. [Gao *et al.*, 2017] that captures non-linear and mixed discrete-continuous causal relationships. To receive accurate CI decisions within heterogeneous data, we derive a non-parametric CI-test by adopting a permutation scheme, e.g., see [Runge, 2018]. On this basis, incorporation of this CI-test into constraint-based methods such as the popular PC-Algorithm [Spirtes *et al.*, 2000] yields an information-theoretic CSL approach which enables an accurate examination in the presence of heterogeneous data. To improve the interpretability, we examine possible methods to quantify the causal strength of learned causal relationships, e.g., see [Janzing *et al.*, 2013].

(2) A Modular Pipeline for CSL: For experimental evaluation in synthetic and real-world scenarios, we develop a modular pipeline for CSL that covers the essential components, including the generation of heterogeneous data that is required for a comprehensive benchmarking of different algorithms and approaches.

(3) Real-World Application Scenarios: Two real-world use cases from genetics and discrete manufacturing allow us to investigate omnipresent properties of heterogeneous data characteristics while ensuring transferability to real-world scenarios. First, publicly available transcriptomic data incorporate complex non-linear biological relationships of genetic regulatory processes. In this setting, a broad spectrum of existing knowledge about protein-protein interactions enables comprehensive benchmarking opportunities for CSL [Pratapa *et al.*, 2020]. Second, we derive causal structures from mixed discrete-continuous data of production processes together with two industry partners, which comes along with a necessity for interpretability of results as the true mechanisms are mostly unknown [Huegle *et al.*, 2020].

4 Preliminary Results

Within preliminary evaluations, our approach tends to improve the accuracy of learned causal structures within synthetic, and real-world scenarios. For example, within the simplified discrete manufacturing scenario from our cooperation partner with expected true CGM depicted in Fig. 1, comparing the structural Hamming distance (SHD) of the learned causal structures between common discretization-based G^2 against our approach shows an improvement in the SHD from 0.6 to 0.3 (PC-Algorithm with $\alpha = 0.05$). Improved accuracy is also reflected in the first experiments to transcriptomic data within the genetics scenario. Preliminary evaluations within the Beeline benchmarking framework [Pratapa *et al.*, 2020] ranked our information-theoretic approach for CSL second-best compared to state-of-the-art genetic network inference algorithms. Further, we examined the accuracy of our approach within experiment II of [Gao *et al.*, 2017] that considers a discrete-continuous mixture setting. In particular, we examine causal relationships for a multinomial distributed $X \sim \mathcal{M}(5, \{\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\})$, a continuous Z, uniformly distributed over $[X, X + 2]$ given the value of X, and an independent d-dimensional normal distributed $Z \sim \mathcal{N}(2, 4)^d$, i.e., $X \to Y$ with causally independent Z. Results for increasing sample size showed consistency for CMI estimation, $CMI(X;Y|Z) = log_2(5) - \frac{4}{5}$, and appropriate conditional independence decisions with a mean F_1 converging to 1. Moreover, we demonstrate the reduced barrier for transferring CSL into practice through our modular pipeline from an application and research perspective.

While first preliminary results showed that our ideas provide an opportunity for improving CSL from heterogeneous data, it is our current effort to receive a more extensive theoretic and synthetic examination regarding requirements and boundaries on the complexity.

References

[Gao *et al.*, 2017] Weihao Gao, Sreeram Kannan, Sewoong Oh, and Pramod Viswanath. Estimating mutual information for discrete-continuous mixtures. In *Proceedings of NeurIPS*, volume 30, pages 5987 – 5998, 2017.

[Huegle *et al.*, 2020] Johannes Huegle, Christopher Hagedorn, and Matthias Uflacker. How causal structural knowledge adds decision-support in monitoring of automotive body shop assembly lines. In *Proceedings of IJCAI*, pages 5246–5248, 2020.

[Janzing *et al.*, 2013] Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Quantifying causal influences. *The Annals of Statistics*, 41(5):2324–2358, 2013.

[Pratapa *et al.*, 2020] Aditya Pratapa, Amogh Jalihal, Jeffrey Law, Aditya Bharadwaj, and TM Murali. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nature Methods*, 17(2):147–154, 2020.

[Runge, 2018] Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *Proceedings of AISTATS*, volume 84, pages 938–947. PMLR, 2018.

[Spirtes *et al.*, 2000] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, 2000.