Causal Discovery in Practice: Non-Parametric Conditional Independence Testing and Tooling for Causal Discovery

Johannes Eric Hügle

Knowledge about causal structures is crucial for decision support in various domains. For example, in discrete manufacturing, identifying the root causes of failures and quality deviations that interrupt the highly automated production process requires causal structural knowledge. However, in practice, root cause analysis is usually built upon individual expert knowledge about associative relationships. But, "correlation does not imply causation", and misinterpreting associations often leads to incorrect conclusions. Recent developments in methods for causal discovery from observational data have opened the opportunity for a data-driven examination. Despite its potential for data-driven decision support, omnipresent challenges impede causal discovery in real-world scenarios. In this thesis, we make a threefold contribution to improving causal discovery in practice.

(1) The growing interest in causal discovery has led to a broad spectrum of methods with specific assumptions on the data and various implementations. Hence, application in practice requires careful consideration of existing methods, which becomes laborious when dealing with various parameters, assumptions, and implementations in different programming languages. Additionally, evaluation is challenging due to the lack of ground truth in practice and limited benchmark data that reflect real-world data characteristics.
To address these issues, we present a platform-independent modular pipeline for causal discovery and a ground truth framework for synthetic data generation that provides comprehensive evaluation opportunities, e.g., to examine the accuracy of causal discovery methods in case of inappropriate assumptions.

(2) Applying constraint-based methods for causal discovery requires selecting a conditional independence (CI) test, which is particularly challenging in mixed discrete-continuous data omnipresent in many real-world scenarios. In this context, inappropriate assumptions on the data or the commonly applied discretization of continuous variables reduce the accuracy of CI decisions, leading to incorrect causal structures.
Therefore, we contribute a non-parametric CI test leveraging k-nearest neighbors methods and prove its statistical validity and power in mixed discrete-continuous data, as well as the asymptotic consistency when used in constraint-based causal discovery. An extensive evaluation of synthetic and real-world data shows that the proposed CI test outperforms state-of-the-art approaches in the accuracy of CI testing and causal discovery, particularly in settings with low sample sizes.

(3) To show the applicability and opportunities of causal discovery in practice, we examine our contributions in real-world discrete manufacturing use cases. For example, we showcase how causal structural knowledge helps to understand unforeseen production downtimes or adds decision support in case of failures and quality deviations in automotive body shop assembly lines.