Chapter 1

Introduction

Nowadays, rising customer requirements regarding product quality and quantity are the key performance measures for manufacturers. Pharmaceutical production firms such as Novo Nordisk especially have seen an increase in customer interest [?]. It is thus of great interest how to efficiently schedule such production flows. Reducing the production time and increasing production in general is hence a major challenge for industry but also for academia. In general, complicated production systems, consisting of multiple disjoint processes, can involve many processes that may or may not influence each other and propagate through the production system. In pharmaceutical production systems, this encompasses processes such as filtration, reaction of chemical agents, centrifugation, and many others depending on the drug substance that is to be produced.

The duration of each process, and the quantity being processed of different substances among other factors all have sources of variation. For example, a human-operated part of a process is a known source of variation. How such variations propagate and affect other processes can be hard to detect without extensive exploration and knowledge of the undergoing processes of the system. Hence, it is of great interest from an industrial point of view to efficiently discover such relations and further use these to make informed decisions along the execution of such processes. In particular, sudden deviations and inaccuracies may occur but having a good understanding of the causal effects of the production system and how these deviations are expected to propagate can be crucial for keeping production throughput. More specifically, at some point in the process, a human may need to manually remove deposits from a reaction tank or adjust the pH by adding NaOH. This can influence how long the product stays in the reaction tank and further impact subsequent processes. In particular, it is of interest whether such variation influences a process further down the line directly or if it is an indirect effect i.e. the variation in a process influencing the variation in the next process and so on called transitive effects.

Due to the inherent graphical nature of the problem i.e. inferring the direct effects by filtering out transitive effects, it is natural to take a graphical approach to the problem. Graphical models have been proposed by [?]. Bayesian networks [?] [?] and the strongly related belief propagation algorithm for inference on graphical models [?] are also well-known methods, however, they are all computationally expensive on large scale systems and typically require prior assumptions or are limited to specific applications. We note that there exist many feature selection algorithms, but it is often not inherently clear how to extend these if one wants a detailed structure of how the effects propagate. Namely, feature selection is usually applied if one only wants to understand how a set of variables influences a specific measure. It is thus often not capable of identifying where errors originate and how they propagate.

Thus, the objective of this thesis is to quantify the impact of each process in a production system on a specific measure of performance here taken to be time through a systematic method by inferring the direct dependencies in the network representing true *interactions* thus removing transitive effects. In particular

Given;	historical observations of sojourn times, delays, concentration changes and more from a production system
Determine;	the causal structure and/or dependency relations between the attributes of the processes
Subject to;	limited observations and possible prior topological considerations

Based on existing material on the topic [?] [?] [?] we shall, in particular, investigate the robustness of the method through perfectly controlled simulations and based on theoretical results and extend the algorithm to infer causal direction subject to certain assumptions instead of only direct dependencies which initially are represented by undirected edges. The robustness is considered both in terms of the assumptions driving the algorithm and particularly also the accuracy of the estimator of mutual information which does not seem to be covered in the existing literature.

The rest of the thesis is structured as follows. In **??**, we introduce the data simulated from [?] which is subject to multiple error sources that need manual handling. In particular, we present some initial analysis of the simulated observations and conclude that a more advanced method for discovering the direct effects between the durations of each process. In **??** we present the method proposed in [?] and methods for computing and estimating mutual information. We apply the method in **??** and explore potential shortcomings of the method for removing indirect effects and estimating the information between pairs of variables on controlled simulations before finally applying everything to the data from **??**. Finally, in **??** we summarize our findings and comment on the properties of the presented methodology.