Discovery of Exogenous Variables in Data with More Variables Than Observations

Yasuhiro Sogawa\textsuperscript{1}, Shohei Shimizu\textsuperscript{1}, Aapo Hyvärinen\textsuperscript{2}, Takashi Washio\textsuperscript{1}, Teppei Shimamura\textsuperscript{3}, and Seiya Imoto\textsuperscript{3}

\textsuperscript{1} The Institute of Scientific and Industrial Research, Osaka University, Japan
\textsuperscript{2} Dept. Comp. Sci. Dept. Math. and Stat., University of Helsinki, Finland
\textsuperscript{3} Human Genome Center, Institute of Medical Science, University of Tokyo, Japan

Abstract. Many statistical methods have been proposed to estimate causal models in classical situations with fewer variables than observations. However, modern datasets including gene expression data increase the needs of high-dimensional causal modeling in challenging situations with orders of magnitude more variables than observations. In this paper, we propose a method to find exogenous variables in a linear non-Gaussian causal model, which requires much smaller sample sizes than conventional methods and works even when orders of magnitude more variables than observations. Exogenous variables work as triggers that activate causal chains in the model, and their identification leads to more efficient experimental designs and better understanding of the causal mechanism. We present experiments with artificial data and real-world gene expression data to evaluate the method.

Keywords: Bayesian networks, independent component analysis, non-Gaussianity, data with more variables than observations.

1 Introduction

Many empirical sciences aim to discover and understand causal mechanisms underlying their objective systems such as natural phenomena and human social behavior. An effective way to study causal relationships is to conduct a controlled experiment. However, performing controlled experiments is often ethically impossible or too expensive in many fields including bioinformatics \cite{1} and neuroinformatics \cite{2}. Thus, it is necessary and important to develop methods for causal inference based on the data that do not come from such controlled experiments.

Many methods have been proposed to estimate causal models in classical situations with fewer variables than observations ($p<n$, $p$: the number of variables and $n$: the number of observations). A linear acyclic model that is a special case of Bayesian networks is typically used to analyze causal effects between continuous variables \cite{3}. Estimation of the model commonly uses covariance structure of data only and in most cases cannot identify the full structure (edge directions and connection strengths) of the model with no prior knowledge on the
structure \cite{3,4}. In \cite{5}, the authors proposed a non-Gaussian variant of Bayesian networks called LiNGAM and showed that the full structure of a linear acyclic model is identifiable based on non-Gaussianity without pre-specifying any edge directions between the variables, which is a significant advantage over the conventional methods \cite{4,3}.

However, most works in statistical causal inference including Bayesian networks have discussed classical situations with fewer variables than observations $(p<n)$, whereas modern datasets including microarray gene expression data increase the needs of high-dimensional causal modeling in challenging situations with orders of magnitude more variables than observations $(p \gg n)$ \cite{1,2}. Here we consider situations in which $p$ is on the order of 1,000 or more, while $n$ is around 50 to 100. For such high-dimensional data, the previous methods are often computationally intractable or statistically unreliable.

In this paper, we propose a method to find exogenous variables in a linear non-Gaussian causal model, which requires much smaller sample sizes than conventional methods and works even when $p \gg n$. The key idea is to identify which variables are exogenous instead of estimating the entire structure of the model. The simpler task of finding exogenous variables than that of the entire model structure would require fewer observations to work reliably. The new method is closely related to a fairly recent statistical technique called independent component analysis (ICA).

Exogenous variables work as triggers that activate a causal chain in the model, and their identification leads to more efficient experimental designs of practical interventions and better understanding of the causal mechanism. A promising application of Bayesian networks for gene expression data is detection of drug-target genes \cite{1}. The new method proposed in this paper can be used to find which genes a drug first affects and how it triggers the gene network.

The paper is structured as follows. We first review ICA and linear causal models in Section 2. We then define a non-Gaussian causal model and propose a new algorithm to find exogenous variables in Section 3. The performance of the algorithm is evaluated by experiments on artificial data and real-world gene expression data in Sections 4 and 5. Section 6 concludes the paper.

## 2 Background Principles

### 2.1 Independent Component Analysis

Independent component analysis (ICA) \cite{6} is a statistical technique originally developed in signal processing. ICA model for a $p$-dimensional observed continuous random vector $x$ is defined as

$$x = As,$$  \hspace{1cm} (1)

where $s$ is a $p$-dimensional continuous random vector whose components $s_i$ are mutually independent and non-Gaussian and are called independent components, and $A$ is a constant $p \times p$ invertible matrix. Without loss of generality, we