Finding stable causal structures from clinical data

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen,
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op donderdag 14 maart 2019
om 10.30 uur precies

door

Ridho Rahmadi

geboren op 13 april 1985
te Yogyakarta (Indonesië)
SIKS Dissertation Series No. 2019-04

The research reported in this thesis has been carried out under the auspices of the Dutch Research School for Information and Knowledge Systems (SIKS), and the institute for Computing and Information Sciences (iCIS) of the Radboud University Nijmegen.

This research was supported by the DIKTI scholarship from the Indonesian Government and by the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement number 305697.

Copyright © 2019 Ridho Rahmadi

Printed by: Gildeprint, Enschede
Finding stable causal structures from clinical data

Doctoral Thesis

to obtain the degree of doctor
from Radboud University Nijmegen,
on the authority of the rector magnificus prof. dr. J.H.J.M. van Krieken,
according to the decision of the Council of Deans
to be defended in public on Thursday, March 14, 2019
at 10.30

by

Ridho Rahmadi

born on 13 april 1985
in Yogyakarta (Indonesia)
Supervisor:
Prof. dr. Tom Heskes

Co-supervisor:
Dr. Perry Groot

Doctoral Thesis Committee:
Prof. dr. Peter Lucas
Prof. dr. Concha Bielza (Universidad Politécnica de Madrid, Spain)
Dr. Johan Kwisthout

SIKS Dissertation Series No. 2019-04

The research reported in this thesis has been carried out under the auspices of the Dutch Research School for Information and Knowledge Systems (SIKS), and the institute for Computing and Information Sciences (iCIS) of the Radboud University Nijmegen.

This research was supported by the DIKTI scholarship from the Indonesian Government and by the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement number 305697.

Copyright © 2019 Ridho Rahmadi

Printed by: Gildeprint, Enschede
Contents

Title page i
Table of Contents v

1 Introduction 1
  1.1 Causality as a paradigm 1
  1.2 Causal approaches 3
  1.3 Applications to clinical domain 4
  1.4 Outline of the thesis 5

2 Background 9
  2.1 Directed acyclic graph 9
    2.1.1 Model equivalence 10
  2.2 Structural equation model 10
    2.2.1 Structural equation model with observed variables 11
    2.2.2 Structural equation model with latent variables 11
    2.2.3 Ordinal observed variables 12
    2.2.4 Identification 13
    2.2.5 Estimation 14
  2.3 Stability selection 14
  2.4 Multi-objective optimization 16
    2.4.1 Domination 16
    2.4.2 Non-dominated sorting genetic algorithm-II 16

3 Stable specification search for cross-sectional data 19
  3.1 Introduction 20
  3.2 Proposed method 21
    3.2.1 The General idea 21
    3.2.2 Constrained SEM 24
    3.2.3 Stable specification search algorithm 25
  3.3 Experimental study 26
    3.3.1 Implementation 26
    3.3.2 Parameter settings 27
8 Clinical significance of hyponatremia in TBM

8.1 Background

8.2 Methods

8.2.1 Setting and patients

8.2.2 Measurements

8.2.3 Data analysis and statistics

8.3 Result

8.3.1 Baseline characteristics

8.3.2 Patient survival

8.3.3 Causal survival model

8.4 Discussion

8.A Supplementary tables

8.B The stable Cox model

9 Discussion

Bibliography

Summary

Samenvatting

Publications

Acknowledgments

Curriculum Vitae

SIKS dissertations
Chapter 1

Introduction

1.1 Causality as a paradigm

One convenient way to understand causation or causality is to see it as a different paradigm in statistics. It is an enrichment that enables statistics to resolve certain problems that traditional statistical methods alone cannot answer (Pearl et al., 2016). To emphasize the differences, we begin with a brief historical overview of non-causal methods which are perhaps the most frequently used in statistics: regression and correlation. In 1875, Sir Francis Galton studied the characteristics of sweet pea plants. When plotting the weights of offspring seeds against the weights of parent seeds, he found that the median weights of offspring seeds from a particular size of parent seed constituted approximately a straight line with positive slope smaller than 1.0 (Stanton, 2001). In 1885, he published his presidential address along with the first bivariate scatterplot (see Figure 1.1) which showed correlation between children’s height against parent’s height (Lee Rodgers and Nicewander, 1988). In Galton (1889), he described that:

“when the variation of one [variable organ] is accompanied on the average by more or less variation of the other... It is easy to see that co-relation must be a consequence of the variations of the two organs being partly due to common causes.”

For about one decade later, Pearson (1895) published a mathematical development of correlation measurement, the Pearson product-moment correlation coefficient. Besides a historical relation, Lee Rodgers and Nicewander (1988) added that regression and correlation are mathematically related, that is, that correlation can be represented as a function of the slope of a regression line and standard deviations of both independent and dependent variables.
Chapter 1

Figure 1.1: The first bivariate scatterplot which shows children’s height against parent’s height from Galton (1886).

Now we come to the causal paradigm. Given two random variables $X$ and $Y$, Pearl et al. (2016) give a simple definition of causation (or causality), that is, a variable $X$ is a cause of $Y$ if $Y$ in any way relies on $X$ for its value. He added that, if we think of causality in terms of listening, then $X$ is a cause of $Y$ if $Y$ listens to $X$ and decides its value in response to what it hears. We can see that correlation and causality are two different concepts. Correlation measures the extent to which variation on $X$ occurs with variation on $Y$. Here, the value of $Y$ is not necessarily due to the value of $X$; they are just purely observed. Moreover, in regression, $Y$ is predicted based on, again, observed $X$. On the contrary, in causality the value of $Y$ is an effect of $X$, making $X$ a cause of $Y$. It is of importance in causality then to describe how $X$ takes its value. The most typical approach is to intervene on $X$, such that its value is fixed; hence we change the system (Pearl et al., 2016). Therefore, the gold standard of causal modeling has typically been to perform randomized controlled experiments in which $X$ is set to particular value(s) (Fisher, 1935; Spirtes, 2010; Pearl et al., 2016). For example, a group of lung cancer patients is divided by researcher into two subgroups; one receives a particular drug as treatment, while the other one does not. In this case, the researcher can examine whether or not the drug as a treatment helps the patients.

However, conducting randomized controlled experiments is not always practical. In some cases, especially when human beings are the object (Spirtes and Zhang, 2016), it is often unethical, unfeasible, time consuming, or expensive to perform a randomized controlled experiment (Maathuis and Nandy, 2016). Therefore, researchers carry out observational studies instead, in which they only record the data, instead of controlling it. However, to render causal information from observational data is notoriously difficult as we have to untangle the causal from the merely
Figure 1.2: An illustration, adopted from Maathuis and Nandy (2016), of how we should impose causal assumptions in order to render causal information from observational data.

correlative (Pearl et al., 2016). That said, it is still possible to derive causal information from purely observational data. Figure 1.2 illustrates that observational data are typically used for analyses such as correlation, prediction, and classification. To perform causal analysis on the observational data, one must then presume that the data are generated from some underlying causal assumptions. Such assumptions are needed to deduce a post-intervention distribution from the observational data (Maathuis and Nandy, 2016).

1.2 Causal approaches

Many causal modeling approaches have been developed in the last decades. Numerous applications have been realized on different domains of science; from biology, climate, to economics, for example Chu et al. (2003); Maathuis et al. (2010); Chicharro and Panzeri (2014); Ebert-Uphoff and Deng (2012); Brodersen et al. (2015); Hoover (2008). Those causal approaches can in principle be divided into two groups: constraint-based and score-based approaches. The former which makes use of conditional independence to obtain causal models, is fundamentally different from the latter which selects causal models based on some scoring function. Despite the difference, both constraint-based and score-based approaches often presume the same typical assumptions in practice, as described in Spirtes et al. (2000).

The first assumption, called Causal Markov condition, assumes that all variables in the data are independent of their noneffects, conditional on all their direct
causes. The second assumption, called \textit{Causal faithfulness}, assumes that there are no independencies between variables that are not implied by the Causal Markov condition. A stronger assumption, used in many but not all causal approaches, called \textit{Causal sufficiency}, assumes that there is no hidden common cause of the observed variables.

Constraint-based approaches search for a causal model by estimating the structure (often called a skeleton) of the underlying causal mechanisms, and then determine the orientation of the skeleton, which yields a causal model. The two steps are carried out by using conditional independence tests and orientation rules. Examples are inductive causation algorithm (IC; Pearl and Verma, 1991) and PC algorithm ("P" stands for Peter, and "C" for Clark, the authors; Spirtes \textit{et al.}, 2000).

Score-based approaches score causal models based on some characteristics, e.g., how well a model fits the data and how simple the model structure is. Among the scored causal models, this approach typically selects the one with the best score. An example is greedy equivalence search (GES; Chickering, 2002b). In both constraint-based and score-based approaches, causal estimates are often solely based on a single estimation which is notoriously unstable; a slight change in the data can lead to completely different inferred models. Spirtes (2010) added that conditional independence tests can give incorrect output which leads to multiple errors in causal methods. More detail about constraint-based and score-based approaches are given in Section 3.1.

In addition, Maathuis and Nandy (2016) mentioned a hybrid variant of causal approaches in their review, which is a combination of both constraint-based and score-based approaches. This approach typically estimates the skeleton using conditional independence test and then searches and scores models limited to the causal structures belonging to the skeleton. Examples of hybrid approach are Bayesian constraint-based causal discovery (BCCD; Claassen and Heskes, 2012), max-min hill-climbing (MMHC; Tsamardinos \textit{et al.}, 2006) and an extension of GES (Alonso-Barba \textit{et al.}, 2013).

1.3 Applications to clinical domain

In the clinical domain, studies are often conducted in a form of cross-sectional or longitudinal design, depending on the interest. In a cross-sectional design, a group of patients is typically observed at a particular point in time. Variables of interest, for example, blood pressure, perceived physical activity, and fatigue level are recorded to characterize patients. In a longitudinal design, similar observations to that of the cross-sectional design are carried out, but repeated at several time slices. In other words, each patient is characterized by different observations which are recorded across time. For example, a group of patients receiving a therapy is being observed for five times; pre-therapy observations or often called baseline, three interim observations after the start of the therapy, and post-therapy observations. Another form of study which is similar to longitudinal design is survival analysis.
In a survival study, a group of patients is typically observed twice in a certain time period; the first time slice records variables of interest, and the second time slice records whether or not the patients experience an event of interest, for example, death.

Causal modeling has been widely applied to the clinical domain, for example, on cross-sectional data (Yarchieski and Mahon, 1989; Lawlor et al., 2008; Seo et al., 2010; Sokolova et al., 2015; Cooper et al., 2015; Long et al., 2016), on longitudinal data (Wheaton, 1978; Hall et al., 1993; la Bastide-van Gemert et al., 2014), and on survival data (Lange and Hansen, 2011; VanderWeele, 2011). This is perhaps due to the fact that many clinical studies try to determine whether or not an intervention, e.g., therapy, drug helps to heal patients. Furthermore, it is also of interest in the field to investigate causal relationships among predictors and between the predictors and the main outcome. Some of the aforementioned applications were based on the structural equation model (SEM), which is considered a primary causal language (Pearl, 2000). In general, SEM measures how well a model fits the data. A typical SEM application starts with a hypothesized model and then performs a few model refinements (often called a specification search) to obtain the best model according to some fit indices. Such an approach is not intended for inferring causal structures, but rather for detecting and correcting specification errors between a proposed and the true model (MacCallum, 1986). It also extremely limits the number of possible models which are equally plausible to be examined. Note that the number of those possible models is immense (Harwood and Scheines, 2002), even only with a modest number of variables. In the clinical domain, especially in the case of rare diseases, a hypothesis may not be available. Considering the immense number of possible SEMs and especially the cases for which hypotheses are not available, a more exploratory approach could be an ideal alternative. Moreover, model estimation is a notoriously difficult problem due to computational aspects (as to find the optimal model can be NP-hard) and instability, i.e., a slight change in the data can lead to a completely different final model.

In this thesis, therefore, we develop an approach that aims to resolve the problems of an immense number of possible SEMs and instability in model estimation. We use the concept of a multi-objective evolutionary algorithm (Deb et al., 2002) to resolve the problem of an immense number of possible SEMs, by exploring numerous possible models to search for optimal models. We employ the concept of stability selection (Meinshausen and Bühlmann, 2010) to resolve the problem of instability in model estimation, by selecting simple causal structures which frequently occur among the optimal models. Combining both concepts, we introduce a novel exploratory score-based approach called stable specification search.

1.4 Outline of the thesis

This thesis comprises two parts. Through Chapters 2 to 6 the first part of the thesis is described; the proposed method and its extensions to longitudinal data and
latent model, including its software implementation as an R package. The second part of the thesis discusses applications to real-world data sets from the clinical and psychometric domain, and is described through Chapters 7 and 8.

In Chapter 2 we describe background knowledge necessary for the rest of the thesis. This includes directed acyclic graphs, stability selection, structural equation model (model representation, identification condition, and estimation), and multi-objective optimization.

In Chapter 3 we introduce a robust causal modeling approach called stable specification search for cross-sectional data (S3C; Rahmadi et al., 2017). S3C is an exploratory causal method which heuristically searches for optimal causal models over numerous possible SEMs. Here, an optimal causal model is characterized by two conflicting objectives: it should fit the data well and should have a simple (parsimonious) structure. This procedure is repeated over different subsamples of the data, resulting in many optimal models from the whole range of model complexities. Among those optimal models, S3C selects stable (frequently occur) and simple causal structures according to some thresholds, called relevant structures. We visualize the relevant structures as a causal graph and annotate it with reliability scores. The performance of S3C has been validated on a simulated data set from the waste incinerator network, and a real-world clinical data set about patients with chronic fatigue syndrome (CFS). The result on the simulated data shows an improvement when compared to state-of-the-art alternative approaches. The result on the real-world clinical data demonstrates conformity with the hypothesis-driven models constructed by the clinical experts.

In Chapter 4 we extend S3C to longitudinal data, called stable specification search for longitudinal data (S3L; Rahmadi et al., 2018c). S3L essentially applies S3C on a longitudinal SEM which may comprise an arbitrary number of time slices. The longitudinal SEM is transformed into a baseline model (representing causal relationships at the baseline) and a transition model (representing causal relationships across time). Accordingly, the longitudinal data is reshaped to follow the transformed model. Relevant structures from the baseline and the transition models are combined and visualized as a longitudinal causal graph. In addition, S3L estimates the standardized total causal effects. Similar to that of S3C, we evaluated the performance of S3L on simulated data sets and three real clinical data sets about patients with Alzheimer’s disease, patients with CFS, and patients with chronic kidney disease (CKD). The results on the simulated data show improvement over alternative approaches and the results on the real clinical data show consistency with those of previous studies and suggest novel findings which deserve further research.

Chapter 5 describes an extension of S3C to S3C-Latent, to model causal relations among latent variables that are measured through observed proxies. This is realized by replacing the model representation with SEM with latent variables and ensuring that several model identification conditions are fulfilled. We evaluate the performance of S3C-Latent on different schemes of simulation and compare the re-
results with those of PC-MIMBuild, an extension of the PC algorithm. We also apply
S3C-Latent to real-world data about children with attention deficit/hyperactivity
disorder (ADHD) and data about measuring mental abilities among pupils. The re-
results on simulated data show that S3C-Latent performs better than PC-MIMBuild,
and the results on the real-world data are consistent with those of earlier studies.

Chapter 6 describes the implementation of S3C/L as an R package named
\texttt{stablespec} (Rahmadi et al., 2018b). It is publicly available at the Comprehensive
R Archive Network (CRAN) with MIT license. A github repository was also created
to invite researchers to track and involve in the development of the package. The \texttt{stablespec}
package supports parallel computation to speed up users’ analyses on
high parameter settings. Documentation of the package is provided along with
the package, which includes running examples for every function, allowing users to
modify and adjust to their problems.

Chapter 7 describes S3L application as a mediation analysis tool to identify
mechanisms of change brought by a treatment called cognitive behavior therapy
(CBT), that is randomly assigned to patients with medically unexplained fatigue,
namely CFS and idiopathic chronic fatigue (ICF). The main interest of this study
is to understand how the treatment reduces the level of fatigue; which variables
mediate the changes on fatigue (mediators), and how these mediators interact. The
study is based on a data set comprising 304 patients from two RCTs (204 patients
with CFS and 100 patients with ICF). We found that the positive effect of the
treatment on the fatigue is mediated by fatigue-related cognitive process such as
self-efficacy with respect to fatigue, focus on fatigue, and tendency to catastrophize,
both directly and indirectly via physical functioning and perceived activity.

In Chapter 8 we report an application of S3C-Latent to a survival data set
about 678 patients with tuberculous meningitis (TBM) in Indonesia, focusing on
one-year mortality. The chief aims of this study are to understand whether hy-
ponatremia contributes to mortality and to understand causal relations between
hyponatremia and other predictors of mortality. For the former goal, we extend the
idea of stability selection to find a stable Cox model, a common survival model rep-
resentation. We found that hyponatremia does not seem to contribute to increased
mortality. We also found that hyponatremia is associated with inflammation and
affected by Mycobacterium tuberculosis culture.
Chapter 2

Background

This chapter provides necessary background used in the thesis, including structural equation model (SEM), concept of stability selection, and multi-objective optimization.

2.1 Directed acyclic graph

A graph is a pair \((V, E)\) with \(V\) a set of nodes and \(E\) a set of edges that connect some pairs of nodes. Two nodes connected by an edge are called adjacent. A path between two nodes \(X_1\) and \(X_n\) is a sequence of nodes \(X_1, \ldots, X_n\) such that for \(1 \leq i \leq n - 1\), \(X_i\) and \(X_{i+1}\) are adjacent. A directed graph has all edges in \(E\) directed (arc); a single arrowhead on every edge, e.g., \(X \rightarrow Y\). A directed edge \(X \rightarrow Y\) implies that \(X\) is a parent of \(Y\), and \(Y\) is a child of \(X\). A directed path is a path, e.g., from \(X_1\) to \(X_n\), along which \(X_i\) is the parent of \(X_{i+1}\). Directed cycles represent feedback or reciprocal relationships, e.g., \(X \rightarrow Y \rightarrow X\). A graph with no directed cycles is called acyclic. A graph which is both directed and acyclic is called a directed acyclic graph (DAG; Pearl, 2000). If we ignore all arrowheads from a DAG, the resulting undirected graph is called the skeleton of the DAG. A v-structure in a DAG is a relation between three distinct nodes, in which two nodes are not adjacent and both have edges into the same node, e.g., \(X \rightarrow Z \leftarrow Y\).

DAGs have been widely employed as a representation of causal relations; the nodes are used to represent variables and the edges are used to denote causal relations between pairs of variables. Figure 2.1 depicts a DAG of four variables. From the DAG we can see that, for example, \(X\) is the parent of \(Y\), and there is a directed path from \(X\) to \(Z\).
2.1.1 Model equivalence

The characterization of equivalent structures is given by the following theorem (Verma and Pearl, 1990).

**Theorem 1** (Verma and Pearl, 1990) Two DAGs are equivalent if and only if they have the same skeletons and the same v-structures.

As described in Chickering (2002a), given a DAG $G$, a directed edge $X \rightarrow Y$ is **compelled** in $G$ if for every DAG $G'$ equivalent to $G$, $X \rightarrow Y$ exists in $G'$. On the other hand, if a directed edge $X \rightarrow Y$ is not compelled in $G$, then it is **reversible** in $G$, that is, there exists some DAG $G'$ equivalent to $G$ in which $Y \rightarrow X$ (opposite direction) exists in $G'$.

An *acyclic partially directed graph* (PDAG) is a graph that consists of both directed and undirected edges. A DAG $G$ is called a **consistent extension** of a PDAG $P$, if both $G$ and $P$ have the same skeleton and the same set of v-structures, and if every directed edge in $P$ has the same direction in $G$.

The **completed** PDAG (CPDAG) that corresponds to an equivalence class is the PDAG with a directed edge for every compelled edge belonging to the equivalence class, and an undirected edge for every reversible edge belonging to the equivalence class.

Converting a DAG $G$ into a CPDAG $P^c$, therefore, allows one to observe the relations that hold among the variables. Directed edges in $P^c$ indicate a causal relation among variables since the same arc occurs in all members of the $P^c$. Undirected edges $X - Y$ in $P^c$ indicate that some members of $P^c$ contain an arc $X \rightarrow Y$ whereas other members contain an arc $Y \rightarrow X$.

Thus, causal models represented by DAGs have their corresponding model equivalence classes, called CPDAGs. This means that every probability distribution derived from a model in a particular CPDAG, can also be derived by models belonging to the same CPDAG. In terms of structural equation models (described in Section 2.2 below), these models are called covariance equivalent (Pearl, 2000).

### 2.2 Structural equation model

Structural equation models (SEMs; Wright, 1921; Haavelmo, 1943) are a primary language of causality (Pearl, 2000), and can be represented as a set of equations,
which is called a causal model, or by drawing them as a causal diagram (graph). In
this study, we assume there are no reciprocal causal relations and a causal model
with such properties is often called a recursive SEM (Bollen, 1989). A recursive
SEM can be represented using a DAG (Spirtes, 2010). The general form of the
equations is

\[ x_i = f_i(p_{a_i}, \varepsilon_i), \quad i = 1, \ldots, n. \]  

(2.1)

where \( p_{a_i} \) denotes the parents which represent the set of variables considered to be
direct causes of \( X_i \) and \( \varepsilon_i \) represents errors on account of omitted factors that are
assumed to be mutually independent (Pearl, 2000). In what follows, we describe how
to represent, identify, and estimate SEMs in more detail (Jöreskog, 1977; Bollen,
1989).

\subsection{2.2.1 Structural equation model with observed variables}

A structural equation model with observed variables has a general representation
that reads

\[ y = By + \Gamma x + \zeta, \]  

(2.2)

where \( y \) is an \( m \times 1 \) vector of endogenous (effect) variables, \( x \) is an \( n \times 1 \) vector
of exogenous (cause) variables, \( B \) is an \( m \times m \) coefficient matrix among \( y \), \( \Gamma \) is
an \( m \times n \) coefficient matrix among \( x \), and \( \zeta \) is an \( m \times 1 \) vector of errors on \( y \). In
addition, \( \Phi \) and \( \Psi \) are the covariance matrices of \( x \) and \( \zeta \), respectively, and used
in Equation (2.3). We assume that \( \Psi \) is diagonal, representing independent errors.

In general, a SEM procedure estimates a model-implied covariance matrix \( \Sigma(\theta) \)
and evaluates how close it matches the sample covariance matrix \( S \). In other terms,
a SEM procedure measures how well a model fits the data. The \( \Sigma(\theta) \) is a function
of model parameters \( \theta \) via

\[
\Sigma(\theta) = \begin{bmatrix}
\Sigma_{yy}(\theta) & \Sigma_{yx}(\theta) \\
\Sigma_{xy}(\theta) & \Sigma_{xx}(\theta)
\end{bmatrix},
\]

\[
\Sigma_{yy}(\theta) = (I - B)^{-1}(\Gamma\Phi\Gamma' + \Psi)(I - B)^{-1}',
\]

(2.3)

\[
\Sigma_{xy}(\theta) = \Phi\Gamma'(I - B)^{-1}',
\]

\[
\Sigma_{xx}(\theta) = \Phi,
\]

where \( \Sigma_{yy}(\theta) \) is a covariance matrix of the endogenous variables \( y \) written as a
function of the parameters \( \theta \). There are analogous definitions for \( \Sigma_{yx}(\theta), \Sigma_{xy}(\theta), \)
and \( \Sigma_{xx}(\theta) \). The prime symbol indicates a matrix transpose.

\subsection{2.2.2 Structural equation model with latent variables}

A SEM with latent variables (also called a general SEM in Bollen, 1989) consists
of the structural model that represents causal relationships among latent variables,
and the measurement model that represents relationships from latent to observed variables. In the literature, the latent variable is often called a factor, and the observed variable is often called an indicator, a manifest variable, or a proxy. In this thesis, we use those terms interchangeably. The structural model reads

$$\eta = B\eta + \Gamma \xi + \zeta,$$  \hspace{1cm} (2.4)

where $\eta$ is an $m \times 1$ vector of latent endogenous (effect) variables, $\xi$ is an $n \times 1$ vector of latent exogenous (cause) variables, $\zeta$ is an $m \times 1$ vector of disturbances on $\eta$, $B$ is an $m \times m$ matrix of coefficients among $\eta$, and $\Gamma$ is an $m \times n$ matrix of coefficients among $\xi$. In addition, $\Phi$ and $\Psi$ denote the covariance matrices of $\xi$ and of $\zeta$, respectively. We assume that $E(\eta) = E(\xi) = E(\zeta) = 0$, $\xi$ uncorrelated with $\zeta$, and that $(I - B)$ is nonsingular.

The measurement model represents influences from $\eta$ to its observed variables, an $r \times 1$ vector $x$, and from $\zeta$ to its observed variables, a $q \times 1$ vector $y$. The measurement model reads

$$x = \Lambda_x \xi + \delta$$

$$y = \Lambda_y \eta + \epsilon,$$  \hspace{1cm} (2.5)

where the $r \times n$ matrix $\Lambda_x$ and $q \times m$ matrix $\Lambda_y$ contain the structure coefficients associating latent variables and indicators, and the $r \times 1$ vector $\delta$ and $q \times 1$ vector $\epsilon$ contain errors on the indicators. In addition, an $r \times r$ matrix $\Theta_\delta$ and a $q \times q$ matrix $\Theta_\epsilon$ are the covariance matrices of $\delta$ and of $\epsilon$, respectively.

The model-implied covariance matrix $\Sigma(\theta)$ for a SEM with latent variables is a function of model parameters $\theta$ through

$$\Sigma(\theta) = \begin{bmatrix} \Sigma_{yy}(\theta) & \Sigma_{yx}(\theta) \\ \Sigma_{xy}(\theta) & \Sigma_{xx}(\theta) \end{bmatrix}$$

$$\Sigma_{yy}(\theta) = \Lambda_y (I - B)^{-1} (\Gamma \Phi \Gamma^\prime + \Psi) [(I - B)^{-1}]^\prime \Lambda_y^\prime + \Theta_\epsilon,$$  \hspace{1cm} (2.6)

$$\Sigma_{xy}(\theta) = \Lambda_x \Phi \Gamma^\prime [(I - B)^{-1}]^\prime \Lambda_y^\prime,$$

$$\Sigma_{xx}(\theta) = \Lambda_x \Phi \Lambda_x^\prime + \Theta_\delta,$$

where $\Sigma_{yy}(\theta)$ is a covariance matrix of the indicators $y$ written as a function of the parameters $\theta$, and similarly as in Equation (2.3) that there are analogous definitions for $\Sigma_{yx}(\theta)$, $\Sigma_{xy}(\theta)$, and $\Sigma_{xx}(\theta)$. Figure 2.2 shows a SEM with three latent variables, each with three indicators.

### 2.2.3 Ordinal observed variables

The SEM with observed variables represented by Equation (2.2) and the SEM with latent variables represented by Equations (2.4) and (2.5) assume normally distributed observed variables $x$ and $y$. Based on Olsson et al. (1982); Drasgow
Figure 2.2: An example of a SEM with three latent variables $\xi_1$ (with $x_1$, $x_2$, and $x_3$ as indicators), $\xi_2$ (with $x_4$, $x_5$, and $x_6$ as indicators), and $\eta_1$ (with $y_1$, $y_2$, and $y_3$ as indicators).

(1988), this model can be extended to ordinal indicators $\hat{x}$ and $\hat{y}$ by discretizing, e.g., $x$ into $w$ categories through

$$\hat{x}_i = \hat{x}_{ij} \quad \text{if } \tau_{j-1} \leq x_i < \tau_j, \ j = 1, \ldots, w, \ i = 1, \ldots, r, \quad (2.7)$$

where $\tau_j$ is a threshold, and for convenience, we define $\tau_0 = -\infty$ and $\tau_w = +\infty$. The threshold $\tau_j$ and the values of $\hat{x}_{ij}$ are assumed to be strictly increasing, i.e., $\tau_1 < \tau_2 < \ldots < \tau_{w-1}$, and $\hat{x}_{i1} < \hat{x}_{i2} < \ldots < \hat{x}_{iw}$. There are analogous expressions for discretizing $y$ into $w$ categories.

We assume that $\hat{x}$ and $\hat{y}$ are proxies of some continuous variables, and therefore use polychoric (between ordinal indicators) and polyserial (in the case of a mixture of continuous and ordinal indicators) correlations (Olsson et al., 1982; Drasgow, 1988), to form the matrix $S$.

2.2.4 Identification

The parameters of SEMs can only be estimated when the so-called identification conditions are satisfied. For both SEMs with observed and SEMs with latent variables, the general condition is that the number of parameters to estimate is equal to or less than the number of elements in $S$. In particular, SEMs with latent variables have the following additional identification conditions (Bollen, 1989; Kenny et al., 1998).

1. There are at least three or more indicators per latent variable.
2. Each row of $\Lambda_x$ and $\Lambda_y$ has only one nonzero element, i.e., an indicator cannot load on multiple latent variables (a pure model; Silva et al., 2006).

3. Each latent variable is scaled, e.g., by setting one factor loading $\lambda_{ij}$ of each latent $\zeta_j$ to 1.

4. $\Theta_\delta$ is diagonal.

Condition 1 can be relaxed to latent variables with less than three indicators as follows. If there is a latent variable with two indicators, then the latent variable must have a causal relation with other latent variables (Bollen, 1989). If there is a latent variable with only one indicator, then the corresponding indicator error is set to zero (Kenny et al., 1998).

2.2.5 Estimation

After a SEM satisfies the identification conditions, its model parameters $\theta$ can be estimated through a maximum likelihood procedure, by minimizing a fitting function that reads

$$\hat{\theta} = \arg\min_{\theta} F_{ML}(\theta),$$

$$F_{ML}(\theta) = \log |\Sigma(\theta)| + \text{Tr}\{S\Sigma^{-1}(\theta)\} - \log |S| - p.$$  \hfill (2.8)

where $p$ is the number of observed variables, and $S$ is the $p \times p$ sample covariance matrix of the observed variables.

2.3 Stability selection

Estimating structures, such as graphs or clusters, or variable selection is a notoriously difficult problem, both because of computational aspects (finding the optimal structure can be NP-hard) and because of instability (small changes in the data can lead to completely different optimal structures). An example of a variable selection algorithm is the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996), which selects variables related to the target variable by a shrinking (regularization) process through a penalization of some regression coefficients.

In this section we describe the stability selection (Meinshausen and Bühlmann, 2010), a method for robust estimation of model structure based on subsampling in combination with selection algorithms. More specifically, it applies a variable selection algorithm repeatedly to randomly drawn subsets of half of the original data. At the end, structures or variables are selected if the corresponding occurrences (or the probability of being selected) obtained across the repetitions are above a pre-determined threshold. The threshold can be chosen to control the expected number of false positive selections. The subsampling introduces additional noise that tends to break weak relationships and preserves those with a high probability.
Let $\beta$ be a sparse $p$-dimensional vector which generally represents, for example, the coefficient vector in linear regression or the edges in a graph. In structure estimation the goal is to infer the set $S = \{k : \beta_k \neq 0\}$ of non-zero components from noisy observations. Many methods tackle this problem by minimizing some loss function augmented with a regularization term to avoid overfitting, e.g., LASSO. Usually the regularization term is parameterized by $\omega \in \Omega \subseteq \mathbb{R}^+$ and each $\omega$ leads to an estimated structure $\hat{S}^\omega \subseteq \{1, \ldots, p\}$. The objective is to determine $\omega$ such that $\hat{S}^\omega$ is identical to $S$ with high probability. Note that $\lambda$ is the original symbol used in Meinshausen and Bühlmann (2010) to represent the regularization parameter; in this thesis we use $\omega$ instead, as $\lambda$ is already used in Equation (2.5). To this end, Meinshausen and Bühlmann (2010) introduce the concepts of selection probabilities and stability paths.

**Definition 1 (Selection probabilities)** Let $I$ be a subset of $\{1, \ldots, n\}$ of size \(\lfloor n/2 \rfloor\) randomly drawn without replacement, $K \subseteq \{1, \ldots, p\}$, and $\hat{S}^\omega(I)$ be the selected set $\hat{S}^\omega$ for subset $I$. The probability of $K$ being in set $\hat{S}^\omega(I)$ is

\[
\hat{\Pi}_K^\omega = P^* \left( K \subseteq \hat{S}^\omega(I) \right)
\]  

where the probability $P^*$ is with respect to the random subsampling and possibly the construction of $\hat{S}^\omega(I)$.

**Definition 2 (Stability path)** For each variable $k = 1, \ldots, p$ the stability path is given by the selection probabilities $\{\hat{\Pi}_k^\omega : \omega \in \Omega\}$.

The selection probability is the probability for each variable to be selected, given a specific data subset and a particular regularization parameter $\omega$. The stability path of a variable is the set of all selection probabilities for the variable.

Furthermore, in stability selection we do not select a single element from the set of models $\{\hat{S}^\omega : \omega \in \Omega\}$ as traditional methods do, but perturb the data many times and select structures that occur in a large fraction of selected sets. To this end, Meinshausen and Bühlmann (2010) introduce the concept of stable variables.

**Definition 3 (Stable variables)** The set of stable variables is defined as

\[
\hat{S}_{\text{stable}} = \{k : \max_{\omega \in \Omega} \hat{\Pi}_k^\omega \geq \pi_{\text{thr}}\}
\]  

where $\pi_{\text{thr}}$ is a cutoff with $0 < \pi_{\text{thr}} < 1$.

Variables with a high selection probability are kept whereas those with low selection probabilities are disregarded. The threshold $\pi_{\text{thr}}$ is a tuning parameter but its influence is small and sensible values, e.g., $\pi_{\text{thr}} \in (0.6, 0.9)$, tend to give similar results.
2.4 Multi-objective optimization

2.4.1 Domination

Following the principle of Occam’s razor, we should prefer models that are simple and fit the data well. These two objectives, however, are often conflicting as a well-fit model is likely to be a complex model. In this thesis, we propose to make use of multi-objective optimization to explicitly optimize both objectives.

In multi-objective optimization, optimal solutions are defined in terms of domination. In the context of a minimization problem, i.e., to minimize the objectives, a model $z_1$ is said to dominate model $z_2$, if the following conditions are satisfied (Deb, 2001):

$$
z_1 \preceq z_2 \iff \forall i \in \{1, \ldots, M\} \quad f_i(z_1) \leq f_i(z_2) \quad \exists j \in \{1, \ldots, M\} \quad f_j(z_1) < f_j(z_2)
$$

The first condition states that the model $z_1$ is no worse than $z_2$ in all objectives $f_i$. The second condition states that the model $z_1$ is strictly better than $z_2$ in at least one objective. By using this concept, given the population of models $P$, we can partition $P$ into $n$ sets called fronts $F_1, \ldots, F_n$, such that $F_k$ dominates $F_l$ where $1 \leq k < l \leq n$ and the models within the same front do not dominate each other. The so-called Pareto Front or non-dominated set $F_1$ includes models that are not dominated by any member of $P$. Essentially, using multi-objective optimization we efficiently find the best fitting models over a whole range of model complexities using a single coherent optimization approach. Figure 2.3 provides a sketch.

2.4.2 Non-dominated sorting genetic algorithm-II

Non-dominated Sorting Genetic Algorithm II or NSGA-II (Deb et al., 2002) is a well-known multi-objective evolutionary algorithm (MOEA), still widely applied in

![Figure 2.3: Example of a population $P$ partitioned into fronts $F_1, \ldots, F_n$ when minimizing objectives $f_1$ and $f_2$. $F_1$ is the Pareto front not dominated by any member of $P$.](image)
Figure 2.4: Adopted from Deb et al. (2002). $P$ is the current population with size $N$ and is manipulated to make a new population $Q$. Both are combined, forming $R$, which will be sorted using fast non-dominated sorting yielding a set of fronts $F$. Every member of front $F_n \in F$ will be assigned a so-called crowding distance in order to sort $F_k$. The first $N$ members of $F$ will be selected to be the next population $P$.

Various fields, such as image retrieval (Arevalillo-Herráez et al., 2013), reactive power planning (Hajabdollahi et al., 2012), building design (Brownlee and Wright, 2015), and robot grippers (Saravanan et al., 2009). A characteristic feature is fast non-dominated sorting which sorts models based on the concept of domination. With $M$ the number of objectives and $N$ the size of population, the time complexity has order $\mathcal{O}(MN^2)$, which is better than a naïve approach with $\mathcal{O}(MN^3)$. Another characteristic feature is crowding distance sorting which is implemented to preserve the diversity among the solutions in the Pareto front. This feature sorts models based on the distance metric which explains the proximity of a model to other models.

The iterative procedure of NSGA-II shown in Figure 2.4 is a sequence of steps started by generating a population of solutions $P$ of size $N$. $P$ is then manipulated by genetic operators such as selection, crossover, and mutation, forming a new population $Q$ of size $N$. $P$ and $Q$ are then combined into population $R$ with size $2N$. After that $R$ is sorted using fast non-dominated sorting, yielding a set of fronts $F$. In the next iteration each front in $F$ is sorted using the crowding distance sorting and the first $N$ members are used to generate a new population $P$. At $t = 0$, $P$ is formed by creating $N$ random solutions sorted with fast non-dominated sorting.
Chapter 3

Stable specification search for cross-sectional data

Causal modeling has long been an attractive topic for many researchers and in recent decades there has seen a surge in theoretical development and discovery algorithms. Generally discovery algorithms can be divided into two approaches: constraint-based and score-based. The constraint-based approach is able to detect common causes of the observed variables but the use of independence tests makes it less reliable. The score-based approach produces a result that is easier to interpret as it also measures the reliability of the inferred causal relationships, but it is unable to detect common confounders of the observed variables. A drawback of both score-based and constrained-based approaches is the inherent instability in structure estimation. With finite samples small changes in the data can lead to completely different optimal structures. In this chapter, we introduce a new hypothesis-free score-based causal discovery algorithm, called stable specification search for cross-sectional data, that is robust for finite samples based on recent advances in stability selection using subsampling and selection algorithms. Structure search is performed over structural equation models. Our approach uses exploratory search but allows for incorporation of prior background knowledge. We validated our approach on one simulated data set, which we compare to the known ground truth, and one real-world data set for chronic fatigue syndrome, which we compare to earlier medical studies. The results on the simulated data set show significant improvement over alternative approaches and the results on the real-word data set show consistency with the hypothesis driven models constructed by medical experts.

3.1 Introduction

Causal modeling has been an attractive topic for many researchers for decades. Especially since the 1990s there has been an enormous increase in theoretical development, partly because of advances in graphical modeling (Pearl, 2000). This has led to a variety of causal discovery algorithms in the literature. In general, causal discovery algorithms can be divided into two approaches: constraint-based and score-based. Constraint-based approaches work with conditional independence tests. First, they construct a skeleton graph starting with the complete graph and excluding edges between variables that are conditionally independent. Second, edges are oriented to arrive at a causal graph. Examples of constraint-based approaches are the IC algorithm, PC, and total conditioning (TC; Pellet and Elisseeff, 2008). Constraint-based approaches do not have to rely on the causal sufficiency assumption, and then can detect common causes of the observed variables (Spirtes et al., 2000). A disadvantage of this approach is the use of independence tests on a large number of conditioning variables, making it less reliable (Spirtes, 2010). Score-based approaches assign scores to particular graph structures based on the data fit and the complexity of the graph. Different scoring metrics that are often used are the Bayesian score (Dawid, 1984) and the BIC score (Schwarz, 1978). An example of a score-based method is GES. The goal of the score-based approach is to find the graph structure with the highest score. An advantage of this approach is that it measures the reliability of the inferred causal relationships, which makes the result easy to interpret (Heckerman et al., 1999). Score-based approaches typically do make the causal sufficiency assumption, and then cannot detect common confounders of the observed variables. Moreover, the involved optimization problem is usually NP-hard, so that different search heuristics are often used. The approach advocated in this study is an example of a score-based approach.

Furthermore, in causal modeling based on observational data, the causal models are undetermined unless a preference for parsimonious models over more complex models is made (Spirtes, 2010). In score-based approaches, such simplicity assumptions are typically implemented by adding a penalty for model complexity (Chickering, 2002b). Constraint-based approaches often make the implicit assumption of so-called causal faithfulness (Spirtes, 2010), which states that there are no conditional independencies that hold in the density over a set of variables $V$, except those that are entailed by the causal structure. However, in practice faithfulness can be violated and better constrained-based approaches have been developed to handle this, such as conservative PC (CPC; Ramsey et al., 2012) and adjacency conservative PC (ACPC; Lemeire et al., 2012).

A drawback of both score-based and constrained-based approaches, however, is the inherent instability in structure estimation. With finite samples small changes in the data can lead to completely different optimal structures. Outcomes of borderline independence tests can be incorrect and can lead to multiple errors when propagated by the discovery algorithm (Spirtes, 2010).
Chapter 3. Stable specification search for cross-sectional data

The present work introduces a new score-based causal discovery algorithm, stable specification search for cross-sectional data (S3C), that is robust for finite samples based on advances in stability selection using subsampling and selection algorithms. Structure search is performed over structural equation models (SEMs), which is the most widely used language for causal discovery in various scientific disciplines. The method uses exploratory search, but allows incorporation of prior background knowledge. In order to show that our method can handle various kinds of data (continuous, discrete, and a combination of both) we evaluated our method on simulated and real-world data. The simulated data is used to compare our method with two advanced constrained-based approaches, namely PC-stable (Colombo and Maathuis, 2014) and CPC, and a score-based approach, namely GES. PC-stable and CPC are extensions of the PC algorithm which in principle consists of two stages. The first stage uses conditional independence tests to obtain the skeleton (undirected edges) of the model, and the second stage orients the skeleton based on some rules, resulting in a CPDAG (described in Section 2.1.1, for more detail see Chickering 2002a). GES in general starts with an empty (or sparse) model, and iteratively adds an edge (forward phase) which mostly increases the score until no more edges can be added. Then GES iteratively prunes an edge (backward phase) which also mostly increases the score until no more edges can be excluded. Specifically, we compare the robustness of each method in computing causal structure. The real-world data set about chronic fatigue syndrome is used to compare our results with some previous studies. The results show that our exploratory, hypothesis-free approach gives significant improvement over alternative approaches, and is able to obtain structure estimates that are consistent with the hypothesis driven models constructed by medical experts based on medical data and years of experience.

The rest of this chapter is structured as follows. Section 3.2 describes our robust score-based approach for causal discovery. Section 3.3 presents experimental results on one simulated and two real-world data sets. Section 3.4 gives conclusions and suggestions for future work.

3.2 Proposed method

3.2.1 The General idea

Our proposed method can be divided into two phases. The first phase is search and the second phase is visualization. In the search phase SEM and NSGA-II are synergically combined for exploratory search of the model space. As portrayed in Figure 3.1, the inner loop is an iterative process, searching over the model space and returns a Pareto front of models. The outer loop is an iterative process that samples a different subset of the data in each iteration and at the end returns a number of Pareto fronts coming from those subsets. Each model returned by the outer loop is transformed into a CPDAG which are then used to compute the edge stability graph and the causal path stability graph.
Figure 3.1: The proposed method consists of two phases: search and visualization. The search phase is an iterative process using an outer loop and inner loop that combines SEM, NSGA-II, and stability selection, which outputs all relevant edges and causal paths between two variables. The visualization phase displays the relevant relationships as a causal model.

Definition 4 (Stability graphs) Let $X$ and $Y$ be two variables and $G$ a multiset (or bag) of CPDAGs. Let $G_c$ be the submultiset of $G$ containing all CPDAGs with complexity $c$. The edge stability for $X$ and $Y$ at complexity $c$ is the number of models in $G_c$ for which there exists an edge between $X$ and $Y$ (i.e., $X \rightarrow Y$, $Y \rightarrow X$, or $X - Y$) divided by the total number of models in $G_c$. The causal path stability for $X$ to $Y$ at complexity $c$ is the number of models in $G_c$ for which there is a directed path from $X$ to $Y$ (of any length) divided by the total number of models in $G_c$. The terms edge stability graph and causal path stability graph are used to denote the corresponding measures for all variable pairs and all complexity levels.

On top of the stability graphs we perform stability selection. In Meinshausen and Bühlmann (2010), stability selection is defined in terms of a regularization parameter $\omega$. In our approach we do not have a regularization parameter and instead use model complexity (defined in Section 3.3.2) which is one of the objectives in our multi-objective optimization approach. We therefore define two thresholds. The first threshold is the boundary of selection probability $\pi_{\text{sel}}$ and corresponds to $\pi_{\text{thr}}$ in Meinshausen and Bühlmann (2010). For example, setting $\pi_{\text{sel}} = 0.6$, as described in Meinshausen and Bühlmann (2010), means that all causal relationships with edge stability or causal path stability (Figure 3.2) above this threshold are considered stable. The second threshold is the boundary of complexity $\pi_{\text{bic}}$, which is used to control overfitting and corresponds to minimal $\omega$ in Meinshausen and Bühlmann (2010). We set $\pi_{\text{bic}}$ to the level of model complexity at which the minimum average Bayesian information criterion (BIC) score is found. For example, $\pi_{\text{bic}} = 7$ means that all causal relationships with an edge stability or a causal path stability lower than this threshold (Figure 3.2) are considered parsimonious. Causal relationships that intersect with the top-left region are considered both stable and parsimonious and called relevant.
Figure 3.2: Example stability graphs from an artificial data set of 400 instances with six continuous variables, without prior knowledge. (a) Edge stability graph. (b) Causal path stability graph. Each line in (a) represents an edge between a pair of variables and each line in (b) represents a causal path with any length from a variable to another variable. The threshold of selection probability, $\pi_{\text{sel}}$, is set to 0.6 and the threshold for model complexity, $\pi_{\text{bic}}$, is chosen to minimize the average BIC score. See the main text for more details.

In the visualization phase we combine the stability graphs into a graph with nodes and edges. This is done by adding the relevant edges and orienting them using background knowledge (if any, see Section 3.2.2) and the relevant causal paths. More specifically, we first connect the nodes following the relevant edges. Then we
orient these edges based on the background knowledge. Finally, we orient the rest of the edges following the relevant causal paths. The resulting graph consists of directed edges which represent causal relationships and possibly with additional undirected edges which represent strong associations but for which the direction is unclear from the data. In addition we annotate each edge with the highest selection probability it has across different model complexities in the top-left region of the edge stability graph. This visualization eases interpretation but the stability graphs are considered to be the main outcome of our approach.

3.2.2 Constrained SEM

In practice, one often has prior knowledge about the domain, for example, that $X$ does not cause $Y$ directly, denoted by $X \not\rightarrow Y$. The method proposed here can include such prior knowledge, extending previous work (Rahmadi et al., 2014), since this translates to a DAG with no directed edge from $X$ to $Y$.\footnote{This still allows for directed edges from $Y$ to $X$ or indirect relations from $X$ to $Y$.}

Model specifications should comply with any prior knowledge when performing specification search and when measuring the edge and causal path stability. When DAGs are converted into CPDAGs in the outer loop, a constraint $X \not\rightarrow Y$ may be violated since arcs $Y \rightarrow X$ in the DAG may be converted into undirected (reversible) edges $X - Y$ in the CPDAG. In order to preserve constraints we therefore extended the efficient algorithm to convert DAG to CPDAG from Chickering (2002a), which runs in time $O(|E|)$ given a DAG $\mathcal{G} = (V, E)$.

Figure 3.3 provides pseudocode for the constrained DAG to CPDAG (consDag2Cpdag) algorithm. Line 2 produces a total ordering $E'$ over the edges in DAG $\mathcal{G}$. Lines 3–6 impose an arc upon the edges that match the constraints. Line 7 uses Chickering (2002a) to label the remaining edges $E \setminus E'$ in $\mathcal{G}$ with “compelled” or “reversible”. Finally, Line 8 returns the constrained CPDAG $\mathcal{P}^c$.

```
1: procedure consDag2Cpdag(DAG $\mathcal{G}$, constraint $\mathcal{C}$)  
2:   $E' \leftarrow$ orderEdges($\mathcal{G}$)  
3:   for every constraint $c \in \mathcal{C}$ do  
4:     get edges $e \in E'$ that matches $c$  
5:     label $e$ with “compelled” in the direction consistent with $c$  
6:   end for  
7:   $\mathcal{P}^c \leftarrow$ labelEdges($\mathcal{G}, E'$) \triangleright label remaining edges using Chickering (2002a)  
8:   return $\mathcal{P}^c$  
9: end procedure
```

Figure 3.3: The consDag2Cpdag algorithm returns a CPDAG $\mathcal{P}^c$ which is consistent with the added prior knowledge and extends Chickering (2002a). The algorithm first labels the edges that match the constraints with “compelled” and then labels the remaining edges with “reversible” or “compelled” using Chickering (2002a).
A DAG without edges will always be transformed into a CPDAG without edges. A fully connected DAG without constraints will be transformed into a CPDAG with only undirected edges. However, if background knowledge is added, a fully connected DAG will be transformed into a CPDAG in which the edges corresponding to the background knowledge are directed. From these observations it follows that in the edge stability graph all paths start with a selection probability of 0 and end up in a selection probability of 1. In the causal path stability graph when no prior knowledge has been added all paths start with a selection probability of 0 and end up in a selection probability of 0. However, when prior knowledge is added some of the paths may end up in a selection probability of 1 because of the added constraints.

### 3.2.3 Stable specification search algorithm

Figure 3.4 provides pseudocode for S3C (cf. Figure 3.1). Lines 3–18 represent the outer loop, Lines 6–16 represent the inner loops, Lines 19–22 compute stability graphs.

An inner loop (Lines 6–16) starts by forming a population $P$ of size $N$, initially at random, or else from a previous population using crowding distance sorting (Lines 7–12). Models are represented with a binary vector $\{0, 1\}$ denoting the existence of some arc $X \rightarrow Y$. Line 13 forms a new population $Q$ by manipulating $P$ using binary tournament selection, one-point crossover, and one-bit flip mutation, which are compatible with a binary representation. The selection scheme selects $N$ times two models from $P$ and places the best model (i.e., lowest front or else smallest crowding distance) in a mating pool $M_{pool}$. One-point crossover takes two models from $M_{pool}$ and swaps the data after the crossover point (the middle). One-bit flip mutation flips each bit according to a predetermined rate. Line 14 combines $P$ and $Q$ and sorts them using fast non-dominated sorting, resulting in a set of model fronts $F$ (described in Section 2.4). Line 15 updates the Pareto front in $F_1$.

An outer loop (Lines 3–18) randomly samples a subset $T$ from $D$ with size $\lfloor |D|/2 \rfloor$ (Line 4), runs the inner loop $I$ times to obtain a Pareto front (Lines 6–16), and stores it in $H$ (Line 17). After $J$ iterations, $H$ contains $J$ Pareto fronts.

Lines 19–22 convert the $J$ Pareto fronts in $H$ from DAGs into CPDAGs using the consDag2Cpdag algorithm in Figure 3.3 and then computes the edge and causal path stability graphs. The stability graphs are considered to be the main outcome of S3C, but can also be visualized as a graph with nodes and edges.
```plaintext
1: procedure S3C(data set $D$, constraint $C$)  
2: $H \leftarrow ()$ \Comment{initialize}  
3: for $j \leftarrow 0, \ldots, J - 1$ do \Comment{$J$ is number of outer loop iterations}  
4: $T \leftarrow$ subset of $D$ with size $\lceil |D|/2 \rceil$ without replacement  
5: $F_1 \leftarrow ()$ \Comment{initialize Pareto fronts to empty list}  
6: for $i \leftarrow 0, \ldots, I - 1$ do \Comment{$I$ is number of inner loop iterations}  
7: if $i = 0$ then  
8: $P \leftarrow N$ random DAGs consistent with $C$  
9: $P \leftarrow \text{fastNonDominatedSort}(P)$  
10: else  
11: $P \leftarrow \text{crowdingDistanceSort}(F)$ \Comment{draw the first $N$ models}  
12: end if  
13: $Q \leftarrow$ make population from $P$  
14: $F \leftarrow \text{fastNonDominatedSort}(P \setminus Q)$  
15: $F_1 \leftarrow$ pareto front of $F$ and $F_1$  
16: end for  
17: $H \leftarrow H \setminus F_1$ \Comment{concatenation}  
18: end for  
19: $G \leftarrow$ convert all DAGs in $H$ to CPDAGs with respect to $C$  
20: edges $\leftarrow$ edge stability of $G$  
21: causalPaths $\leftarrow$ causal path stability of $G$  
22: plot stability graphs based on edges and causalPaths  
23: end procedure
```

Figure 3.4: S3C consists of an outer and an inner loop. The outer loop samples a subset of the data, and for every subset, the inner loop searches for the Pareto front by applying NSGA-II. The Pareto fronts are converted into constrained CPDAGs which are then used to compute the edge and causal path stability graphs.

### 3.3 Experimental study

#### 3.3.1 Implementation

We implemented S3C as an R package named *stablespec*. The package is publicly available at the Comprehensive R Archive Network (CRAN),\(^2\) so it can be installed directly, e.g., from R console by typing `install.package("stablespec")` or from R Studio by using feature to install package. We also included a package documentation as a brief tutorial of using the functions. All experiments were run on an Intel Xeon E7-4870 v2 Processor 2.3 GHz, 120 CPUs, 96 of 32GB LRDIMM, running Ubuntu 14.04.

\(^2\)https://cran.r-project.org/web/packages/stablespec/index.html
3.3.2 Parameter settings

For all experiments, we employed the same set of NSGA-II parameters and stability thresholds. We had 100 iterations in the outer loop, and in each iteration we drew a subsample with size $\lceil |D|/2 \rceil$. We did not do a comprehensive parameter tuning for NSGA-II, instead, we followed guidelines provided in Grefenstette (1986). The parameters were set as follows: the number of generations (inner loop) was 20, the size of the population $P$ was 100, the crossover rate was 0.85, the mutation rate was 0.075 and with binary tournament selection.

To score a SEM, S3C uses the likelihood ratio test statistic (Jöreskog, 1967) (equal to $(N-1)F_{ML}$ from Equation (2.9), with $N$ the sample size, which is often denoted $\chi^2$) to indicate the model fit, and the number of relations in the model (equals to the number of nonzero elements in matrices $B$ and $\Gamma$ in Equation (2.2)), to indicate the model complexity. The $\chi^2$ is considered the original fit index in SEM and measures whether the model-implied covariance matrix is close enough to the sample covariance matrix (Kline, 2011). These scoring metrics are, however, not inherent to S3C, and can be replaced by different scoring metrics.

The model complexity represents how many predicted parameters the model contains. Assuming that variances of parameters are always predicted, the maximum model complexity with $p$ variables is given by $p(p-1)/2$.

When using multi-objective optimization we minimize both the $\chi^2$ and model complexity objectives. These two objectives are, however, conflicting with each other. For example, minimizing the model complexity typically means compromising the data fit.

3.3.3 Application to simulated data

Data generation

In this experiment we generated data using the Waste Incinerator network in Figure 3.5, which is a model of waste emissions from an incinerator plant (Lauritzen, 1992). This model contains both discrete and continuous random variables, with $B$ the waste burning regimen, $W$ the compositional differences in incoming waste, $C$ the concentration of CO2, $F$ the filter state, $E$ the filter efficiency, $L$ the light penetrability, $D$ the emission of dust, $M_{in}$ the metals in waste, and $M_{out}$ the metals emission. We generated 10 data sets containing 400 samples from this network using the BNT toolbox with the default parameter setting as described in Murphy (2001).

Performance measure

We compared S3C with GES (score-based method), PC-stable, and CPC (both constrained-based methods). Our method intrinsically subdivides the data in a number of subsets, here 50 of size 200 samples, and then runs the multi-objective optimization to obtain 50 Pareto fronts (see Section 2.4). For a fair comparison,
Figure 3.5: The Waste Incinerator network. Dashed nodes represent discrete variables, solid nodes represent continuous variables, and arcs represent direct causal relations.

for each algorithm we consider subsampling, giving each method 50 subsets, e.g., Ramsey (2010). For every subset, each algorithm returns a CPDAG from which we can derive the edges and causal paths.

Since the true model of the Waste Incinerator data is known, we can measure the performance of both methods by means of the receiver operating characteristic (ROC) curve (Fawcett, 2004). The true positive rate (TPR) and the false positive rate (FPR) are computed with respect to the CPDAG of the true model. For example, in the case of causal path stability, a true positive means that a causal path with any length obtained through our approach or the other algorithms is actually present in the CPDAG of the true model. By increasing the threshold $\pi_{\text{sel}}$, we increase the TPR at the expense of the FPR. In addition, we conducted three significance tests to compare the ROC curves, two of which compare the area under the curve (AUC) of the ROC curves, and the other one compares the ROC curves. The first test (DeLong et al., 1988) uses the U-statistics theory to estimate the covariance matrix of the AUCs, based on which the significance test is performed. The second test (Robin et al., 2011), a modification of Hanley and McNeil (1983), computes the difference between AUCs in units of standard deviation that is obtained from bootstrap replicates of the ROC curves. The third test (Venkatraman and Begg, 1996) compares the actual ROC curves by evaluating the total absolute difference and then uses a permutation test to compute the statistical significance of the difference. The null hypothesis is that the AUC of S3C and those of alternative approaches are equivalent.

We repeated the above procedure 10 times on different Waste Incinerator data sets and computed the ROC curves using two different schemes: averaging and individual. In the averaging scheme, the ROC curves are computed based on the average edge and causal path stability from different data sets. We conducted statistical significance tests on these average ROC curves. Conversely, in the individual scheme the ROC curves are computed directly from the edge and causal path stability on each data set. We conducted individual statistical significance tests on the
ROC curves for each data set and then used Fisher’s method, as described in Fisher (1925); Frederick Mosteller (1948), to combine these tests into a single test statistic by summing up the log-transformed $p$-values obtained from the individual significance tests. Fisher’s method aims to check if the combined information from the individual tests can reject the shared null hypothesis. Both schemes are intended to show empirically and comprehensively how robust the results of each algorithm are across changes in the data.

**Discussion of waste incinerator result**

Figure 3.6 shows the ROC curves for (a) the edge stability and (b) the causal path stability from the averaging scheme. The corresponding AUCs for edge stability are 0.96 (S3C), 0.89 (PC-stable), 0.88 (CPC), and 0.69 (GES). The AUCs for causal path stability are 0.98 (S3C), 0.85 (PC-stable), 0.88 (CPC), and 0.61 (GES).

Table 3.1 lists the results of the significance tests for both the averaging and individual schemes. The ROC and AUC for the edge stability are comparable with PC-stable and CPC ($p$-value $> 0.1$), but always significant ($p$-value $< 0.01$) compared with GES. The ROC and AUC for the causal path stability compared with CPC are marginally significant ($p$-value $< 0.1$) using the averaging scheme, but significant using the individual scheme ($p$-value $< 0.01$); compared with PC-stable significant ($p$-value $< 0.05$) using the averaging scheme, but highly significant using the individual scheme ($p$-value $< 10^{-5}$); compared with GES highly significant using both schemes ($p$-value $< 10^{-5}$). To conclude, we show that the S3C obtains at least comparable performance as, but often significant improvement over alternative approaches, especially in obtaining the causal relations.

**3.3.4 Application to real-world data**

This section describes the result of applying S3C on a real-world data set about patients with chronic fatigue syndrome (CFS), a particular disease for which the underlying causal relationships are often not clear. Revealing such causal relationships can lead to the development of (new) dedicated treatments and medications for patients with CFS.

**Performance measure**

Since the true model is unknown we measure the performance of our method using the edge stability and causal path stability graphs. We set the thresholds to $\pi_{sel} = 0.6$ and $\pi_{bic}$ to the minimum average of BIC scores. The relevant causal relations are those which occur in the top-left region (e.g., see Figure 3.2). We compare the stability graphs to studies reported in the literature.
Table 3.1: Table of \( p \)-values from comparisons between SGC and alternative approaches. For each significance test, we compared the ROC of the edge (Edge) and causal path (Causal) stability on both averaging (Ave.) and individual (Ind.) schemes.

<table>
<thead>
<tr>
<th>Significance test</th>
<th>Ave.</th>
<th>Ind.</th>
<th>Ave.</th>
<th>Ind.</th>
<th>Ave.</th>
<th>Ind.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeLong (DeLong et al., 1988)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edge</td>
<td>0.003</td>
<td>&lt; 10^{-5}</td>
<td>0.175</td>
<td>0.073</td>
<td>0.131</td>
<td>0.064</td>
</tr>
<tr>
<td>Causal</td>
<td>&lt; 10^{-5}</td>
<td>&lt; 10^{-5}</td>
<td>0.027</td>
<td>0.023</td>
<td>0.098</td>
<td>0.005</td>
</tr>
<tr>
<td>Bootstrap (Robin et al., 2011)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edge</td>
<td>0.003</td>
<td>&lt; 10^{-5}</td>
<td>0.296</td>
<td>0.073</td>
<td>0.135</td>
<td>0.064</td>
</tr>
<tr>
<td>Causal</td>
<td>&lt; 10^{-5}</td>
<td>&lt; 10^{-5}</td>
<td>0.022</td>
<td>0.023</td>
<td>0.073</td>
<td>0.005</td>
</tr>
<tr>
<td>Venkatraman (Venkatraman and Begg, 1996)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Edge</td>
<td>0.004</td>
<td>&lt; 10^{-5}</td>
<td>0.591</td>
<td>0.096</td>
<td>0.684</td>
<td>0.005</td>
</tr>
<tr>
<td>Causal</td>
<td>&lt; 10^{-5}</td>
<td>&lt; 10^{-5}</td>
<td>0.023</td>
<td>0.023</td>
<td>0.096</td>
<td>0.005</td>
</tr>
</tbody>
</table>
Figure 3.6: ROC curves for (a) the edge stability and (b) the causal path stability, for different values of $\pi_{\text{sel}}$ in the range of $[0, 1]$. In (a), the AUCs are 0.96 (S3C), 0.89 (PC-stable), 0.88 (CPC), and 0.69 (GES). In (b), The AUCs are 0.98 (S3C), 0.85 (PC-stable), 0.88 (CPC), and 0.61 (GES).

**Application to CFS**

In this experiment we consider a data set about CFS of 183 subjects (Heins et al., 2013). Originally the data comes from a longitudinal study with five time slices, but in this study, we focus only on one time slice representing the subjects after the first treatment.

The data set contains six discrete variables; fatigue severity assessed with the subscale fatigue severity of the checklist individual strength (CIS), the sense of...
control over fatigue assessed with the self-efficacy scale (SES), focusing on symptoms measured with the illness management questionnaire, the objective activity of the patient measured using an actometer (oActivity), the subject’s perceived activity measured with the subscale activity of the CIS (pActivity), and physical functioning measured with subscale physical functioning of the medical outcomes survey (SF36). We refer to the original paper (Heins et al., 2013), for a detailed description of the questionnaires used and the actometer. Missing values were imputed using an imputation method expectation maximization implemented in SPSS. As all of the variables have large scales, e.g., in the range between 0 to 155, we treat them as continuous variables. We added prior knowledge that the variable fatigue does not cause any of the other variables directly.

The total computation time for one subset was around 5.5 minutes. Figure 3.7a shows that eight relevant edges were found. These edges are between pActivity and fatigue, focusing and fatigue, functioning and fatigue, control and fatigue, pActivity and focusing, pActivity and oActivity, focusing and control, and functioning and control. Figure 3.7b shows that four relevant causal paths were found. These causal paths are: pActivity to fatigue, control to fatigue, functioning to fatigue, and focusing to fatigue.

The stability graphs can be combined into a model as follows. First, the nodes are connected according to the eight relevant edges obtained. Second, the edges are oriented according to the background knowledge added. The fact that the variable fatigue does not directly cause any other variable results in four directed edges, which, in this case, correspond exactly to the relevant causal paths obtained. The inferred model is shown in Figure 3.8.

A (direct) causal path $X \rightarrow Y$ in Figure 3.8 indicates that a change in variable $X$ causes a change in variable $Y$. A dashed line between $X$ and $Y$ represents a strong association between $X$ and $Y$ which direction cannot be determined from the data. All variables except for objective activity were found to be direct causes for fatigue severity, which are corroborated by literature studies. In Vercoulen et al. (1998), changes in physical activity, sense of control, and focus on symptoms measured, were shown to result in changes in fatigue. In Wiborg et al. (2012), changes in perceived activity, sense of control, and physical functioning were shown to result in changes in fatigue. In Heins et al. (2013), an increase in sense of control, perceived activity, and self-reported physical functioning, as well as a decrease in focusing on symptoms resulted in a decrease of fatigue, whereas changes in objective activity did not result in any change in fatigue. In addition, there were strong associations between perceived activity and objective activity, perceived activity and focusing on symptoms, and sense of control and focusing on symptoms.

3.4 Conclusion and future work

In the last decades the field of causal modeling has seen a surge in theoretical development and the construction of various causal discovery algorithms. In general,
Chapter 3. Stable specification search for cross-sectional data

causal discovery algorithms can be divided into two approaches: constraint-based and score-based. A disadvantage, however, of current causal discovery algorithms is the inherent instability of structure estimation. With finite samples small changes in the data can lead to completely different optimal structures.

The present work introduces a new hypothesis-free score-based causal discovery algorithm, S3C, that is robust for finite samples based on subsampling and selection.
Figure 3.8: The inferred model of CFS by combining the edge stability and causal path stability graphs. Each edge has a reliability score which is the highest selection probability in the top-left region of the edge stability graph. A solid line represents a causal relationship while a dashed line represents a strong association which direction cannot be determined from the data.

algorithms. S3C uses exploratory search to search over structural equation models and allows for the incorporation of prior background knowledge, without the need to specify the complete model structure in advance.

The comparison conducted on the simulated data shows that our method, the S3C, shows significant improvement over alternative approaches in obtaining the causal relations. The result on a real-world data set of CFS is consistent with previous studies (Vercoulen et al., 1998; Wiborg et al., 2012; Heins et al., 2013). These studies show that our causal discovery algorithm is able to robustly estimate the underlying causal structure.

Several issues have not yet been explored in our current approach that warrant further research, such as latent variables and longitudinal data. Taking into account the existence of latent variables can further improve our structure estimate by properly identifying dependencies between variables as an unmeasured common cause acting on both variables. In longitudinal data several subjects are measured at different time slices which provides a richer structure that can be incorporated in the causal discovery algorithm. A first attempt in this direction can be found in Rahmadi et al. (2015).

Our approach can be viewed as a novel application of multi-objective optimization. The main idea of stability selection (Meinshausen and Bühlmann, 2010), is to increase the robustness of structure estimation by considering a whole range of model complexities. In the original work, this is done by varying a continuous regularization parameter. For causal discovery we have to explicitly consider different discrete model complexities. Furthermore, finding the optimal structure for each
model complexity is a hard optimization problem. By rephrasing stability selection as a multi-objective optimization problem, we can jointly run over various model complexities and find the corresponding optimal structures for each model complexity. In this study, we have used NSGA-II for multi-objective optimization, because of its popularity and availability, but we realize that more recent multi-objective optimization approaches (Qi et al., 2015; Kukkonen and Lampinen, 2005; Zhang and Li, 2007; Taboada et al., 2008) may be even more efficient. This is beyond the scope of this work and left for future research. In the same spirit, one can easily combine freely available software packages, e.g., for scoring structural equation models, bootstrap sampling, and multi-objective optimization, to build one’s own robust structural estimation approach.

Appendix 3.A   Table of stability

Table 3.2: Lines in Figure 3.7a and the corresponding pairs of variables that are represented, which constitute the edge stability graph of CFS.

<table>
<thead>
<tr>
<th>Lines</th>
<th>Edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>▲-------- ▲-------- ▲</td>
<td>fatigue and pActivity</td>
</tr>
<tr>
<td>▲-------- ▲-------- ▲</td>
<td>fatigue and control</td>
</tr>
<tr>
<td>▲-------- ▲-------- ▲</td>
<td>control and focusing</td>
</tr>
<tr>
<td>▲-------- ▲-------- ▲</td>
<td>fatigue and functioning</td>
</tr>
<tr>
<td>▲-------- ▲-------- ▲</td>
<td>control and functioning</td>
</tr>
<tr>
<td>▲-------- ▲-------- ▲</td>
<td>focusing and pActivity</td>
</tr>
<tr>
<td>▲-------- ▲-------- ▲</td>
<td>fatigue and focusing</td>
</tr>
<tr>
<td>▲-------- ▲-------- ▲</td>
<td>pActivity and oActivity</td>
</tr>
<tr>
<td>▲-------- ▲-------- ▲</td>
<td>functioning and pActivity</td>
</tr>
<tr>
<td>▲-------- ▲-------- ▲</td>
<td>control and pActivity</td>
</tr>
<tr>
<td>▲-------- ▲-------- ▲</td>
<td>control and oActivity</td>
</tr>
<tr>
<td>▲-------- ▲-------- ▲</td>
<td>focusing and oActivity</td>
</tr>
<tr>
<td>▲-------- ▲-------- ▲</td>
<td>fatigue and oActivity</td>
</tr>
<tr>
<td>▲-------- ▲-------- ▲</td>
<td>functioning and oActivity</td>
</tr>
<tr>
<td>▲-------- ▲-------- ▲</td>
<td>functioning and focusing</td>
</tr>
</tbody>
</table>
Table 3.3: Lines in Figure 3.7b and the corresponding pairs of variables that are represented, which constitute the causal path stability graph of CFS.

<table>
<thead>
<tr>
<th>Lines</th>
<th>Causal Paths</th>
</tr>
</thead>
<tbody>
<tr>
<td>□-----□-----□</td>
<td>pActivity to fatigue</td>
</tr>
<tr>
<td>●----------●</td>
<td>control to fatigue</td>
</tr>
<tr>
<td>△-----△-----△</td>
<td>functioning to fatigue</td>
</tr>
<tr>
<td>+--------+--------+</td>
<td>focusing to fatigue</td>
</tr>
<tr>
<td>x-------x-------x</td>
<td>oActivity to fatigue</td>
</tr>
<tr>
<td>●---------●---------●</td>
<td>focusing to pActivity</td>
</tr>
<tr>
<td>□--------□--------□</td>
<td>functioning to pActivity</td>
</tr>
<tr>
<td>○----------○----------○</td>
<td>oActivity to pActivity</td>
</tr>
<tr>
<td>△--------△--------△</td>
<td>control to pActivity</td>
</tr>
<tr>
<td>+--------+--------+</td>
<td>focusing to oActivity</td>
</tr>
<tr>
<td>x-------x-------x</td>
<td>focusing to control</td>
</tr>
<tr>
<td>●---------●---------●</td>
<td>control to oActivity</td>
</tr>
<tr>
<td>○----------○----------○</td>
<td>functioning to control</td>
</tr>
<tr>
<td>△--------△--------△</td>
<td>pActivity to oActivity</td>
</tr>
<tr>
<td>+--------+--------+</td>
<td>control to functioning</td>
</tr>
<tr>
<td>x-------x-------x</td>
<td>oActivity to functioning</td>
</tr>
<tr>
<td>○---------○---------○</td>
<td>oActivity to control</td>
</tr>
<tr>
<td>□--------□--------□</td>
<td>oActivity to focusing</td>
</tr>
<tr>
<td>●----------●</td>
<td>control to focusing</td>
</tr>
<tr>
<td>△-----△-----△</td>
<td>pActivity to functioning</td>
</tr>
<tr>
<td>+--------+--------+</td>
<td>focusing to functioning</td>
</tr>
<tr>
<td>x-------x-------x</td>
<td>functioning to focusing</td>
</tr>
<tr>
<td>○---------○---------○</td>
<td>pActivity to control</td>
</tr>
<tr>
<td>○---------○---------○</td>
<td>pActivity to focusing</td>
</tr>
<tr>
<td>○----------○----------○</td>
<td>fatigue to pActivity</td>
</tr>
<tr>
<td>△-----△-----△</td>
<td>fatigue to oActivity</td>
</tr>
<tr>
<td>+--------+--------+</td>
<td>fatigue to focusing</td>
</tr>
<tr>
<td>x-------x-------x</td>
<td>fatigue to functioning</td>
</tr>
<tr>
<td>○---------○---------○</td>
<td>fatigue to control</td>
</tr>
</tbody>
</table>
Chapter 4

Stable specification search for longitudinal data

Data in the clinical domain are often gathered through a longitudinal design. With such a design, patients are being observed at different time slices. Longitudinal data, therefore, provide more information and explain more of the natural heterogeneity among subjects, for example, in terms of how persistent symptoms develop over time. In this chapter, we introduce stable specification search for longitudinal data (S3L) that extends the stable specification search for cross-sectional data (S3C) described in Chapter 3, to handle longitudinal data. S3L achieves at least comparable performance as, but often a significant improvement over state-of-the-art alternative approaches on simulated data set with a known ground truth. We also present the results of S3L on three real-world longitudinal data sets on chronic fatigue syndrome, Alzheimer disease, and chronic kidney disease. The findings obtained with S3L are generally in line with results from more hypothesis-driven analyses in earlier studies and suggest some novel relationships that deserve further research.

4.1 Introduction

Causal modeling, an essential problem in many disciplines (Daniel et al., 2012; Hoover, 2008; Abu-Bader and Abu-Qarn, 2003; Taguri et al., 2018; Pearl, 1995; Detilleux et al., 2016), attempts to model the mechanisms by which variables relate and to understand the changes on the model if the mechanisms were manipulated (Spirtes, 2010). In the medical domain, revealing causal relationships may

This chapter is based on Rahmadi et al. (2018c), “Causality on longitudinal data: Stable specification search in constrained structural equation modeling”, published in Statistical Methods for Medical Research.
lead to improvement of clinical practice, for example, the development of treatment and medication. Slowly but steadily, causal discovery methods find their way into the medical literature, providing novel insights through exploratory analyses (la Bastide-van Gemert et al., 2014; Sokolova et al., 2014; Cooper et al., 2015). Moreover, data in the medical domain is often collected through longitudinal studies. Unlike in a cross-sectional design, where all measurements are obtained at a single occasion, the data in a longitudinal design consist of repeated measurements on subjects through time. Longitudinal data make it possible to capture change within subjects over time and thus gives some advantage to causal modeling in terms of providing more knowledge to establish causal relationships (Frees, 2004). As emphasized in Fitzmaurice et al. (2012), there is much natural heterogeneity among subjects in terms of how diseases progress that can be explained by the longitudinal study design. Another advantage is that in order to obtain a similar level of statistical power as in cross-sectional studies, fewer subjects in longitudinal studies are required (Hedeker and Gibbons, 2006).

To date, a number of causal modeling methods have been developed for longitudinal (or time series) data. Some of the methods are based on a vector autoregressive (VAR) and/or structural equation model (SEM) framework which assumes a linear system and independent Gaussian noise (Swanson and Granger, 1997; Bessler and Lee, 2002; Demiralp and Hoover, 2003; Moneta, 2008; Kim et al., 2007). Some other methods, interestingly, take advantage of nonlinearity (Moneta et al., 2011; Peters et al., 2013; Chu and Glymour, 2008), or non-Gaussian noise (Peters et al., 2013; Hyvärinen et al., 2008), to gain even more causal information. Most of the aforementioned methods conduct the estimation of the causal structures in somewhat similar ways. Bessler and Lee (2002); Demiralp and Hoover (2003); Moneta (2008); Peters et al. (2013); Hyvärinen et al. (2008) use the (partial correlations of the) VAR residuals to either test independence or as input to a causal search algorithm, e.g., linear non-Gaussian acyclic model (LiNGAM; Shimizu et al., 2006), PC (Spirtes et al., 2000). In general these causal search algorithms are solely based on a single run of model learning which is notoriously unstable: small changes in finite data samples can lead to entirely different inferred structures. This implies that, some approaches might not be robust enough to correctly estimate causal models from various data, especially when the data set is noisy or has small sample size.

In the present chapter, we introduce a robust causal modeling algorithm for longitudinal data that is designed to resolve the instability inherent to structure learning. We refer to our method as S3L, a shorthand for stable specification search for longitudinal data. It extends our previous method (described in Chapter 3; Rahmadi et al., 2017), here referred to as S3C, which is designed for cross-sectional data. S3L is a general framework which subsamples the original data into many subsets, and for each subset S3L heuristically searches for Pareto optimal models using a multi-objective optimization approach. Among the optimal models, S3L observes the so-called relevant causal structures which represent both stable and parsimonious model structures. These steps constitute the structure estimation
of S3L which is fundamentally different from the aforementioned approaches that mostly use a single run for model estimation. For completeness, detail about S3C/L is described in Section 4.2. Moreover, in the default setting S3L assumes some underlying contexts: iid samples for each time slice (lag), linear system, additive independent Gaussian noise, causal sufficiency (no latent variables), stationarity (time-invariant causal relationships), and fairly uniform time intervals between time slices.

The main contributions of S3L are:

- The causal structure estimation of S3L is conducted through multi-objective optimization and stability selection (Meinshausen and Bühlmann, 2010) over optimal models, to optimize both the stability and the parsimony of the model structures.

- S3C/L is a general framework which allows for other causal methods with all of their corresponding assumptions, e.g., nonlinearity, non-Gaussianity, to be plugged in as model representation and estimation. The multi-objective search and the stability selection part are independent of any mentioned assumptions.

- In the default model representation, S3L adopts the idea of the “rolling” model from Friedman et al. (1998) to transform a longitudinal SEM model with an arbitrary number of time slices into two parts: a baseline model and a transition model. The baseline model captures the causal relationships at baseline observations, when subjects enter the study. The transition model consists of two time slices, which essentially represent the possible causal relationships within and across time slices. We also describe how to reshape the longitudinal data correspondingly, so as to match the transformed longitudinal model which then can easily be scored using standard SEM software.

- We provide standardized causal effects which are computed from IDA (intervention-calculus when the DAG is absent) estimates (Maathuis et al., 2009).

- We carry out experiments on three different real-world data of (a) patients with chronic fatigue syndrome (CFS), (b) patients with Alzheimer disease (AD), and (c) patients with chronic kidney disease (CKD).

Some relevant methods have attempted to make use of common structures to infer causal models. Causal stability ranking (CStaR; Stekhoven et al., 2012), originally designed for gene expression data, tries to find stable rankings of genes (covariates) based on their total causal effect on a specific phenotype (response), using a subsampling procedure similar to stability selection and IDA to estimate causal effects. As CStaR only focuses on relationships from all covariates to a single specific response, it seems to be difficult to generalize it to other domains where any possible causal relationship may be of interest. Moreover, another approach called group iterative multiple model estimation (GIMME; Gates and Molenaar, 2012),
originally developed for functional magnetic resonance imaging (fMRI) data and essentially an extension of extended unified SEM (combination of VAR and SEM; Gates et al., 2011), aims to combine the group-level causal structures with the individual-level structures, resulting in a causal model for each individual which contains common structures to the group. Such subject-specific estimation may be feasible given relatively long time series (as in resting state fMRI), but likely too challenging for the typical longitudinal data in clinical studies with a limited number of time slices per subject. Still in the domain of fMRI, there is a method called independent multiple-sample greedy equivalence search (IMaGES; Ramsey et al., 2010). The method is a modification of GES, and designed to handle unexpected statistical dependencies in combined data. Since IMaGES was developed mainly for combining results of multiple data sets, we do not consider it further.

Having both the transformed longitudinal model and the reshaped data, we can run other alternative approaches which are designed for cross-sectional data and conduct comprehensive comparisons. Here, for evaluation of S3L, we generate simulated data and compare with some advanced constrained-based approaches such as PC-stable (Colombo and Maathuis, 2014), CPC (Ramsey et al., 2012), CPC-stable (Colombo and Maathuis, 2014; Ramsey et al., 2012), and PC-Max (Ramsey, 2016). All of these methods are extensions of the PC algorithm. We also compare with an advanced score-based algorithm called fast greedy equivalent search (FGES; Ramsey et al., 2017), which is an extension of GES.

The rest of this chapter is organized as follows. All methods used in our approach are presented in Section 4.2. The results and the corresponding discussions are presented in Section 4.3. Finally, conclusions and future work are presented in Section 4.4.

4.2 Proposed method

4.2.1 Stable specification search for longitudinal data

S3L is an extension of S3C. In principle, as illustrated in Figure 4.1, S3L applies S3C on transformed longitudinal models, called baseline and transition models (explained in Section Longitudinal model and data reshaping). Furthermore, in order to see to which extent a covariate would cause a response, S3L provides standardized total causal effect estimates which are intrinsically computed from estimates from IDA (Maathuis et al., 2009) (described in Section Estimating causal effects). In the following subsections, we first describe how we transform a longitudinal model and reshape the data accordingly, and then we discuss the implication of allowing prior knowledge in our S3C structure learning.
Figure 4.1: Given a longitudinal data set, S3L uses the baseline observations to infer a baseline model, and reshapes the whole data set to infer a transition model. Both baseline and transition model are annotated with a reliability score $\alpha$ and a standardized causal effect $\beta$.

**Longitudinal model and data reshaping**

Based on the idea of a “rolling” network in Friedman et al. (1998) we transform a longitudinal SEM with an arbitrary number of time slices (e.g., Figure 4.2c) into two parts: a baseline model (Figure 4.2a) and a transition model (Figure 4.2b). In the original paper, the authors treat these models as probabilistic networks, here we treat them purely as SEMs. The baseline model essentially represents the causal relationships between variables that may happen at the initial time slice $t_0$, for instance, causal relationships that occur before a medical treatment started. Moreover, the baseline model may also represent relationships of the unobserved process before $t_0$ (Friedman et al., 1998). The transition model constitutes the causal relationships between variables across time slices $t_{i-1}$ and $t_i$, and between variables within time slice $t_i$ for $i > 0$, for example, causal relationships that represent interactions during a medical treatment. In S3L, the structure estimations will be conducted on the baseline and transition model separately.

From the transition model we distinguish two kinds of causal relationships, namely *intra-slice* causal relationship (e.g., solid arcs in Figure 4.2b), and *inter-slice* causal relationship (e.g., dashed arcs in Figure 4.2b). The intra-slice causal relationship represents relationships within time slice $t_i$. Accordingly the inter-slice causal relationship represents relationships between time slices $t_{i-1}$ and $t_i$. We assume that the inter-slice causal relationships are independent of $t$ (stationary). We also assume that the time intervals between time slices are fairly uniform. In addition, the transition model implies two more constraints (explained in Section Constrained SEM): there is no intra-slice causal relationship allowed in time slice $t_{i-1}$ and the inter-slice causal relationships always go forward in time, i.e., from time slice $t_{i-1}$ to time slice $t_i$.

Moreover, in order to score the transformed models, we reshape the longitudinal
data accordingly. Figure 4.3 shows an illustration of the data reshaping. Suppose we are given longitudinal data with \( s \) instances, \( p \) variables, and \( j \) time slices (note that the time slice \( t \) is indexed from 0 to \( j - 1 \), see Figure 4.2c). We assume that the original data shape is in a form of a matrix \( D \) of size \( s \times q \), with \( q = p \times j \). The reshaped data is then a matrix \( D' \) of size \( s' \times q' \), with \( s' = s(j - 1) \) and \( q' = 2p \). Having such reshaped data allows us to use standard SEM software to compute the scores.

**Constrained SEM**

In practice, we are often given some prior knowledge about the data. The prior knowledge which may be, e.g., results of previous studies, gives us some constraints in terms of causal relations. For example, in the case of, say disease \( A \), there exists some common knowledge which tells us that symptom \( S \) does not cause disease \( A \) directly. In terms of a SEM specification, the prior knowledge can be translated into a constrained SEM in which there is no directed edge from variable \( X \) (denotes symptom \( S \)) to variable \( Y \) (denotes disease \( A \)); this still allows for directed edges from \( Y \) to \( X \) or directed paths (indirect relationships) from \( X \) to \( Y \), e.g., a path \( X \rightarrow ... \rightarrow Y \) with any variables in between. S3C, and hence S3L allow for such prior knowledge to be included in the model. In S3L, this prior knowledge only applies to the intra-slice causal relationships.

Model specifications should comply with any prior knowledge when performing specification search and when measuring the edge and causal path stability. Recall that in order to measure the stability, all optimal models (DAGs) are converted into their corresponding equivalence class models (CPDAGs). This model
Chapter 4. Stable specification search for longitudinal data

Figure 4.3: An illustration of data reshaping. $D$ is a matrix representing the original data shape which consists of $s$ instances, $p$ variables, and $j$ time slices. $D'$ is a matrix representing the corresponding reshaped data.

transformation, however, could result in CPDAGs that are inconsistent with the prior knowledge. For example, a constraint $X \not\rightarrow Y$ may be violated since arcs $Y \rightarrow X$ in the DAG may be converted into undirected (reversible) edges $X \leftarrow Y$ in the CPDAG. In order to preserve constraints, we therefore extended the efficient algorithm to convert DAG to CPDAG from Chickering (2002a), as described in Section 3.2.2. Essentially, the motivation of our extension to Chickering’s algorithm is similar to that of Meek’s algorithm (Meek, 1995), that is, to obtain a CPDAG consistent with prior knowledge.

**Estimating causal effects**

We employ IDA (Maathuis et al., 2009) to estimate the total causal effects of a covariate $X_i$ on a response $Y$ from the relevant structures. This method works as follows. Given a CPDAG $P^c$ which contains DAGs $G_1, \ldots, G_m$ in its equivalence class, IDA applies intervention calculus (Pearl, 2000, 2003) to each DAG $G_j$ to obtain multisets $\Theta_i = \{\theta_{ij}\}_{j \in 1, \ldots, m}, i = 1, \ldots, p$, where $p$ is the number of covariates. $\theta_{ij}$ specifies the possible causal effect of $X_i$ on $Y$ in graph $G_j$.

Causal effects can be computed using so-called intervention calculus (Pearl, 2000), which aims to determine the amount of change in a response variable $Y$ when one would manipulate the covariate $X_i$ (and not the other variables). Note that this notion differs from a regression-type of association (see IDA paper for illustrative examples). Given a DAG $G_j$, the causal effect $\theta_{ij}$ can be computed using the
so-called back-door adjustment, which takes into account the associations between
$Y$, $X_i$ and the parents $\text{pa}_i(G_j)$ of $X_i$ in $G_j$. Under the assumption that the distribution
of the data is normal and the model is linear, causal effects can be computed
from a regression of $Y$ on $X_i$ and its parents. Specifically, we have (Maathuis et al.,
2009), $\theta_{ij} = \beta_{i|\text{pa}_i(G_j)}$, where, for any set $S \subseteq \{X_1, \ldots, X_p, Y\} \setminus \{X_i\}$,
\begin{equation}
\beta_{i|S} = \begin{cases}
0, & \text{if } Y \in S \\
\text{coefficient of } X_i \text{ in } Y \sim X_i + S, & \text{if } Y \notin S,
\end{cases}
\end{equation}
and $Y \sim X_i + S$ is the linear regression of $Y$ on $X_i$ and $S$. Note that IDA estimates
the total causal effect from a covariate and response, which considers all possible,
either direct or indirect, causal paths from the covariate to the response.

IDA works for continuous, normally distributed variables and then only requires
their observed covariance matrix as input to compute the regression coefficients.
Following Drasgow (1988) we treat discrete variables as surrogate continuous variables,
substituting the polychoric correlation for the correlation between two discrete variables and the polyserial correlation between a discrete and a continuous variable.

Our fitting procedure does not yield a single CPDAG, but a whole set of
CPDAGs to represent the given data. We therefore extend IDA as follows. We
gather $G_{\pi_{\text{bic}}}$, the CPDAGs of all optimal models with complexity equal to $\pi_{\text{bic}}$.
For each CPDAG $P^c \in G_{\pi_{\text{bic}}}$, we compute the possible causal effects $\Theta$ of each relevant
causal path using IDA. For example, for the causal effect from $X$ to $Y$, we obtain estimates $\Theta_{X \rightarrow Y}^k, k = 1, \ldots, N$, where $N$ is the number of subsets. All causal effect estimations in $\Theta_{X \rightarrow Y}^k$ are then concatenated into a single multiset $\Theta_{X \rightarrow Y}$.

To represent the estimated causal effects from $X$ to $Y$, we compute the median
$\tilde{\Theta}_{X \rightarrow Y}$ and iff $X$ and $Y$ are continuous variables, we standardize the estimation using
\begin{equation}
\frac{\tilde{\Theta}_{X \rightarrow Y} \cdot \sigma_X}{\sigma_Y},
\end{equation}
where $\sigma_X$ and $\sigma_Y$ are the standard deviations of the covariate and the response,
respectively. Standardized causal effects allow us to meaningfully compare them.

4.3 Results and discussion
4.3.1 Implementation

We implemented S3C and S3L as an R package named stablespec. The
package is publicly available at the Comprehensive R Archive Network
(CRAN), so it can be installed directly, e.g., from the R console by typing
install.package("stablespec") or from RStudio. We also included a package
documentation as a brief tutorial on how to use the functions.

\footnote{https://cran.r-project.org/web/packages/stablespec/index.html}
Chapter 4. Stable specification search for longitudinal data 45

4.3.2 Parameter settings

For application to simulated data and real-world data, we subsampled 50 and 100 subsets from the data with size $\lceil |D|/2 \rceil$, respectively. We did not do comprehensive parameter tuning for NSGA-II, instead, we followed guidelines provided in Grefenstette (1986). The parameters for applications to both simulated and real-world data were set as follows: the number of iterations was 35, the number of models in the population was 150, the probability of applying crossover was 0.85, the probability of applying mutation to a model structure was 0.07, and the selection strategy was binary tournament selection (Miller and Goldberg, 1995). As in S3C (see Section 3.3.2), to score a model, S3L uses $\chi^2$ to indicate the model fit and the number of relations in the model to indicate the model complexity.

4.3.3 Application to simulated data

Data Generation

We generated data sets from a longitudinal model containing four continuous variables and three time slices (depicted by Figure 4.4). Ten data sets for each of sample sizes 400 and 2000 are generated with random parameterizations and made publicly available.\(^4\)

Performance measure

We conducted comparisons between S3L with FGES, PC-stable, CPC, CPC-stable, and PC-Max in two different scenarios: with and without prior knowledge about part of the causal directions. Here, the comparisons focus more on the transition model, because in our previous study (Rahmadi et al., 2017) we already conducted

\(^4\)Available at https://tinyurl.com/smmr-rahmadi-dataset
experiments on the baseline model. In the case of prior knowledge, we added that variable $X_1$ at $t_i$ cannot cause variables $X_2$ and $X_3$ at $t_i$ directly. This prior knowledge translates to constraints that the various methods can use to restrict their search space. In addition to both scenarios, we also added longitudinal constraints to the models of FGES, PC-stable, CPC, CPC-stable, and PC-Max the same as those used in the transition model of S3L, i.e., there is no intra-causal relationship from time $t_{i-1}$ and the inter-slice causal relationships always go forward in time $t_{i-1}$ to $t_i$.

The parameters of FGES, PC-stable, CPC, CPC-stable, and PC-Max used in this simulation are set following some existing examples (Kalisch et al., 2012; Wongchokprasitti, 2016; Maathuis et al., 2009). For FGES, the penalty of BIC score is 2 and the vertex degree in the forward search is not limited. For PC-stable, CPC, CPC-stable, and PC-Max, the significance level when testing for conditional independence is 0.01, and the maximum size of the conditioning sets is infinity.

Moreover, as the true model is known, we measure the performance of all approaches by means of the receiver operating characteristic (ROC; Fawcett, 2004) for both edges and causal paths. We compute the true positive rate (TPR) and the false positive rate (FPR) based on the CPDAG of the true model. As for example, in the case of edge stability, a true positive means that an edge obtained by our method or the other approaches is present in the CPDAG of the ground truth.

To compare the ROC curves of our method and those of alternative approaches, we employed three significance tests. The first two tests, as introduced in DeLong et al. (1988) and in Robin et al. (2011), compare the area under the curve (AUC) of the ROC curves by using the theory of U-statistics and bootstrap replicates, respectively. The third test, as described in Venkatraman and Begg (1996), compares the actual ROC curves by evaluating the absolute difference and generating rank-based permutations to compute the statistical significance. The null hypothesis is that (the AUC of) the ROC curves of our method and those of alternative approaches are identical.

Furthermore, we computed the ROC curves using two different schemes: averaging and individual. Both schemes are applied to all methods and to all data sets generated. In the averaging scheme, the ROC curves are computed from the average edge and causal path stability from different data sets, and then the statistical significance tests are applied to these ROC curves. On the other hand, in the individual scheme the ROC curves are computed from the edge and causal path stability on each data set. We then applied individual statistical significance tests on the ROC curves for each data set and used Fisher’s method (Fisher, 1925; Frederick Mosteller, 1948), to combine these test results into a single test statistic.

The experimental designs (with and without prior knowledge) and the ROC schemes (averaging and individual) are aimed to show empirically and comprehensively how robust the results are of each approach in various practical cases as well as against changes in the data.
Discussion

We first discuss the result of our experiments on the data set with sample size 400. Figure 4.5 shows the ROC curves for the edge stability ((a) and (c)) and the causal path stability ((b) and (d)) from the averaging scheme. Panels (a) and (b) represent the results without prior knowledge, while panels (c) and (d) represent the results with prior knowledge. Table 4.3 lists the corresponding AUCs.

Tables 4.1 and 4.2 present the results of the significance tests for both the averaging and individual schemes in the experiment with and without prior knowledge, respectively. In the case without prior knowledge, generally the AUCs of the edge and the causal path stability of S3L are better (p-value ≤ 0.05, or even ≤ 0.001, few of them are marginally significant, e.g., p-value ≤ 0.1) than those of other approaches according to both schemes, except those of FGES for which generally there is no evidence of a difference (p-value > 0.1). In the case with prior knowledge, in general the results are similar to those of experiments without prior knowledge, but now the AUC of the causal path stability of S3L is better (p-value ≤ 0.05) than that of FGES. The ROC of the causal path stability of S3L is now also better (p-value ≤ 0.05) than those of PC-stable, CPC, CPC-stable, and PC-Max according to the individual scheme. This is an improvement over the experiment without prior knowledge.

Next we discuss the result of our experiments on the data set with sample size 2000. Figure 4.6 shows the ROC curves and Table 4.6 lists the corresponding AUCs. Tables 4.4 and 4.5 list the results of the significance tests for both the averaging and individual schemes in the experiment with and without prior knowledge, respectively. In the case without prior knowledge, generally the AUCs of the edge and the causal path stability of S3L are better than (p-value ≤ 0.05) those of other approaches according to the individual scheme. Moreover, the ROCs of the edge and the causal path stability of S3L are better than those of FGES (p-value ≤ 0.001) and CPC-stable (p-value ≤ 0.1), respectively, according to the individual scheme. In the case with prior knowledge, the results are pretty much similar to those of the experiment without prior knowledge, but only now the p-value tends to become smaller, e.g., (p-value ≤ 0.001).

To conclude, we see that in general S3L attains at least comparable performance as, but often a significant improvement over, alternative approaches. This holds in particular for causal directions and in the case of a small sample size. The presence of prior knowledge enhances the performance of the S3L.

4.3.4 Application to real-world data

When applying S3L to the real-world data set, the true model is unknown, so we can only compare the results of S3L with those reported in earlier studies and interpretation by medical experts. We set the thresholds to \( \pi_{sel} = 0.6 \) and \( \pi_{bic} \) to the model complexity where the minimum average of BIC scores is found. By
Figure 4.5: Results from simulation data with sample size 400: ROC curves for (a) the edge stability and (b) the causal path stability (without prior knowledge), and (c) the edge path stability and (d) the causal path stability (with prior knowledge). Different values of $\pi_{sel}$ in the range of $[0, 1]$ yield the corresponding AUCs.

Table 4.3 lists the corresponding AUCs.
Figure 4.6: Results from simulation data with sample size 2000: ROC curves for (a) the edge stability (without prior knowledge) and (b) the causal path stability, and (c) the edge path stability and (d) the causal path stability (with prior knowledge), for different values of $\pi_{sel}$ in the range of $[0, 1]$. Tables 4.6 lists the corresponding AUCs.
Table 4.3: Table of AUCs for the edge and causal path stability for each method, from simulation on data with sample size 400, with (yes) and without prior knowledge (no).

<table>
<thead>
<tr>
<th></th>
<th>PC-Max</th>
<th>CPC-stable</th>
<th>CPC</th>
<th>PC-stable</th>
<th>FGES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Venkatraman and Begg (1996)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Robin et al. (2011)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeLong et al. (1988)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUC &lt; 0.034</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2: Table of AUCs for the edge and causal path stability (see Figure 4.5a and 4.5b) on both averaging (Avg.) and individual (Ind.) schemes.

<table>
<thead>
<tr>
<th></th>
<th>PC-Max</th>
<th>CPC-stable</th>
<th>CPC</th>
<th>PC-stable</th>
<th>FGES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Venkatraman and Begg (1996)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Robin et al. (2011)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DeLong et al. (1988)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AUC &lt; 0.034</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1: Table of p-values from comparisons on data set with sample size 400 between S3L and alternative approaches without prior knowledge. The null hypothesis is that (the AUC of) the ROC curves of S3L and those of alternative approaches are equivalent. For each significance test, we compared the ROC of the edge and causal path (Causal) stability (see Figure 4.5c and 4.5d) on both averaging (Avg.) and individual (Ind.) schemes.
### Table 4.6: Table of AUCs for the edge and causal path stability for each method, from simulation on data with sample size 2000, with (yes) and without prior

<table>
<thead>
<tr>
<th>Method</th>
<th>CPC-stable</th>
<th>PC-stable</th>
<th>S3L</th>
<th>FGES</th>
<th>Edge</th>
<th>Causal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Venkatraman and Begg (1996)</td>
<td>0.937</td>
<td>0.225</td>
<td>0.948</td>
<td>0.984</td>
<td>0.919</td>
<td>0.460</td>
</tr>
<tr>
<td>Robin et al (2011)</td>
<td>0.937</td>
<td>0.225</td>
<td>0.948</td>
<td>0.984</td>
<td>0.919</td>
<td>0.460</td>
</tr>
<tr>
<td>DeLong et al (1988)</td>
<td>0.937</td>
<td>0.225</td>
<td>0.948</td>
<td>0.984</td>
<td>0.919</td>
<td>0.460</td>
</tr>
</tbody>
</table>

### Table 4.5: Table of p-values from comparisons on data set with sample size 2000 between SEL and alternative approaches with prior knowledge

<table>
<thead>
<tr>
<th>Method</th>
<th>CPC-stable</th>
<th>PC-stable</th>
<th>S3L</th>
<th>FGES</th>
<th>Edge</th>
<th>Causal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Venkatraman and Begg (1996)</td>
<td>0.937</td>
<td>0.225</td>
<td>0.948</td>
<td>0.984</td>
<td>0.919</td>
<td>0.460</td>
</tr>
<tr>
<td>Robin et al (2011)</td>
<td>0.937</td>
<td>0.225</td>
<td>0.948</td>
<td>0.984</td>
<td>0.919</td>
<td>0.460</td>
</tr>
<tr>
<td>DeLong et al (1988)</td>
<td>0.937</td>
<td>0.225</td>
<td>0.948</td>
<td>0.984</td>
<td>0.919</td>
<td>0.460</td>
</tr>
</tbody>
</table>
thresholding we get the relevant causal relationships: those which occur in the relevant region. Details of the procedure are given in Section 3.2.

The model assumptions in the application to real-world data follow from the assumptions of S3L in the default setting. The assumptions include iid samples on each time slice, linear system, independent Gaussian noise, no latent variables, stationarity, and fairly uniform time intervals between time slices.

**Application to chronic fatigue syndrome data**

Our first application to real-world data considers a longitudinal data set of 183 patients with chronic fatigue syndrome (CFS) who received cognitive behavior therapy (CBT; Heins et al., 2013). Empirical studies have shown that CBT can significantly reduce fatigue severity. In this study we focus on the causal relationships between cognitions and behavior in the process of reducing subject’s fatigue severity. We therefore include six variables namely fatigue severity, the sense of control over fatigue, focusing on the symptoms, the objective activity of the patient (oActivity), the subject’s perceived activity (pActivity), and the physical functioning. The data set consists of five time slices where the first and the fifth time slices are the pre- and post-treatment observations, respectively, and the second until the fourth time slices are observations during the treatment. The missing data is 8.7% and to impute the missing values, we used single imputation with expectation maximization (EM) in SPSS. As all of the variables have large scales, e.g., in the range between 0 to 155, we treat them as continuous variables. We added prior knowledge that the variable fatigue at $t_0$ and $t_i$ does not cause any of the other variables directly. This is a common assumption made in the analysis of CBT in order to investigate the causal impact on fatigue severity (Vercoulen et al., 1998; Heins et al., 2013).

First we discuss the baseline model, which only considers the baseline causal relationships. The corresponding stability graphs can be seen in Figures 4.7a and 4.7b. As mentioned before, $\pi_{sel}$ is set to 0.6 and from the search phase of S3L we found that $\pi_{bic} = 6$. Figures 4.7a and 4.7b show that three relevant edges and two relevant causal paths were found. Following the visualization procedure (see visualization phase in Section 3.2.1), we get a baseline model in Figure 4.8a. The model shows that pActivity is a direct cause for fatigue severity. This follows from the prior assumption that we made and is consistent with earlier works (Vercoulen et al., 1998; Heins et al., 2013). This causal relationship suggests that a reduction of (perceived) activity, leads to an increase of fatigue. In addition we found a strong relationship between pActivity and oActivity whose direction cannot be determined. This relationship is somewhat sensible as both are variables measuring patient’s activity. We also found a connection between focusing and control, which is not surprising as focusing on symptoms also depends on patient’s sense of control over fatigue. One would expect that if a patient has less control on the fatigue, the focus on the symptom would increase.

Next we discuss the transition model, which considers all causal relationships
Figure 4.7: The stability graphs of the baseline model in (a) and (b) and the transition model in (c) and (d) for chronic fatigue syndrome, with edge stability in (a) and (c), and causal path stability in (b) and (d). The relevant regions, above πsel and left of πbic, contain the relevant structures.
Figure 4.8: (a) The baseline model and (b) the transition model of chronic fatigue syndrome. The dashed line represents a strong relation between two variables but the causal direction cannot be determined from the data. Each edge has a reliability score (the highest selection probability in the relevant region of the edge stability graph) and a standardized total causal effect estimation. For example, the annotation “1/0.71” represents a reliability score of 1 and a standardized total causal effect of 0.71. Note that the standardized total causal effect represents not just the direct causal effect corresponding to the edge, but the total causal effect also including indirect effects. To remove clutter from the diagram we use the following abbreviations: a = 0.61/0.02, b = 1/0.26, c = 1/−0.40, d = 1/−0.52, e = 0.91/0.23, f = 1/−0.33, g = 1/−0.40, and h = 1/−0.43.

over time slices. The corresponding stability graphs are depicted in Figures 4.7c and 4.7d. We set $\pi_{\text{sel}} = 0.6$ and the search phase of S3L yielded $\pi_{\text{bic}} = 27$. Figures 4.7c shows that nineteen relevant edges were found, consisting of eleven intra-slice (blue lines) and eight inter-slice relationships of which six are between the same variables (orange lines) and two are between different variables (black lines). Figure 4.7d shows that thirty-five relevant causal paths were found, consisting of twelve intra-slice (blue lines) and twenty-three inter-slice relationships of which six are between the same variables (orange lines) and seventeen are between different variables (black lines). Applying the visualization procedure, we get the transition model in Figure 4.8b. The model shows that all variables have intra-slice causal relationships to fatigue severity. These relationships are consistent with Vercoulen et al. (1998); Heins et al. (2013); Wiborg et al. (2012), which conclude that during
the CBT an increase in sense of control over fatigue, physical functioning, and perceived physical activity, together with a decrease in focusing on symptoms lead to a lower level of fatigue severity. Interestingly, the actual activity seems insufficient to reduce fatigue severity (Heins et al., 2013), however, how the patient perceives his own activity does seem to help. Additionally, we also found that, with similar causal effects, all variables (except pActivity and fatigue) also cause the change in fatigue indirectly via pActivity as an intermediate variable. This suggests that, as discussed in Heins et al. (2013), an increase in perceived activity does seem important to explain the change in fatigue. The variables focusing and functioning also appear to be indirect causes of changes in the level of fatigue severity.

Application to Alzheimer’s disease data

For the second application to real-world data, we consider a longitudinal data set about Alzheimer’s Disease (AD), which is provided by the Alzheimer’s disease neuroimaging initiative (ADNI; Weiner et al., 2010), and can be accessed at adni.loni.usc.edu. The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. For up-to-date information see www.adni-info.org.

In the present study we focus on patients with MCI, an intermediate clinical stage in AD (Petersen et al., 1999). Following Haight et al. (2012) we include only the variables: subject’s cognitive dysfunction (ADAS-Cog), hippocampal volume (hippocampal_vol), whole brain volume (brain_vol), and brain glucose metabolism (brain_glucose). The data set contains 179 subjects with four continuous variables and six time slices. The first time slice captures baseline observations and the next time slices are for the follow-up observations. The missing data is 22.9% and like in the application to CFS, we imputed the missing values using single imputation with EM. We added prior knowledge that the variable ADAS-Cog at $t_0$ and $t_i$ does not cause any of the other variables directly. We performed the search over 100 subsamples of the original data set.

First we discuss the baseline model which only considers the baseline causal relationships. The corresponding stability graphs are shown in Figures 4.9a and 4.9b. $\pi_{sel}$ is set to 0.6 and the search phase of S3L found that $\pi_{bic} = 4$. Figures 4.9a and 4.9b show that four relevant edges and two relevant causal paths were found. Following the visualization procedure, we obtain the baseline model in Figure 4.10a. We found that an increase in both brain glucose metabolism and hippocampal volume causes reduction in subject’s cognitive dysfunction. These causal relations are consistent with findings in Haight et al. (2012) which also concluded that both brain_glucose and hippocampal_vol were independently related to ADAS-Cog (in
Figure 4.9: The stability graphs of the baseline model in (a) and (c), and the transition model in (b) and (d) for Alzheimer’s disease, with edge stability in (a) and (c), and causal path stability in (b) and (d). The relevant regions above π sel and left of π bic contain the relevant structures.
Figure 4.10: (a) The baseline model and (b) the transition model of Alzheimer’s disease. The dashed line represents a strong relation between two variables but the causal direction cannot be determined from the data. Each edge has a reliability score (the highest selection probability in the relevant region of the edge stability graph) and a standardized total causal effect estimation. For example, the annotation “1/0.81” represents a reliability score of 1 and a total standardized causal effect of 0.81. Note that the standardized total causal effect represents not just the direct causal effect corresponding to the edge, but the total causal effect also including indirect effects.

Next we discuss the transition model which considers all causal relationships across time slices. We set $\pi_{sel} = 0.6$ and the search phase of S3L yielded $\pi_{bic} = 12$. The corresponding stability graphs can be seen in Figures 4.9c and 4.9d. We found twelve relevant edges (see Figure 4.9c), consisting of four intra-slice (blue lines) and eight inter-slice relationships of which four are between the same variables (orange lines) and four are between different variables (black lines). Moreover, we found seventeen relevant causal paths (see Figure 4.9d), consisting of six intra-slice (blue lines) and eleven inter-slice relationships of which four are between the same variables (orange lines) and seven are between different variables (black lines). Applying the visualization procedure, we obtain the transition model in Figure 4.10b. In addition, the direction of the edge from brain_glucose to brain_vol follows because we do not allow cycles in our model. We found that there are indirect and direct causal relationships from hippocampal_vol and brain_vol at both $t_{i-1}$ and $t_i$ to ADAS-Cog at $t_i$. These particular causal relationships support the hypothesis in Haight et al. (2012) which says that any changes in both hippocampal volume and brain volume will cause short-term effects on a subject’s cognitive dysfunction, both direct and
indirect. In the original paper the authors suggested that the indirect causal relationship is through brain-glucose, but our analysis also discovers a potential indirect effect through brain_vol. Interestingly we found that a change in subject’s cognitive dysfunction in a previous time slice $t_{i-1}$ causes a reduction in brain volume in time slice $t_i$.

**Application to chronic kidney disease**

For the third application to real-world data, we consider a longitudinal data set about chronic kidney disease (CKD), provided by the MASTERPLAN study group (Peeters et al., 2014). The MASTERPLAN study was initiated in 2004 as a randomized, controlled trial studying the effect of intensified treatment with the aid of nurse practitioners on cardiovascular and kidney outcome in CKD. This intensified treatment regimen addressed eleven possible risk factors for the progression of CKD simultaneously. The study previously showed that this intensified treatment resulted in fewer patients reaching end stage kidney disease compared to standard treatment (Peeters et al., 2014).

Here we focus on the potential causal mediators for the protective effect incurred by the intensified treatment with the aid of nurse practitioners. In other words, we aim to identify which of the treatment targets contributed to the observed overall treatment effect. In the present analysis, we include only variables of interest, being treatment status, either nurse practitioner aided care or standard care, as allocated by the randomization procedure (treatment), estimated glomerular filtration rate (gfr)—a marker for overall kidney function, and a variable indicating informative censoring (inf_cens). Informative censoring occurred when patients reached end stage kidney disease requiring renal replacement therapy, such as dialysis or a kidney transplantation, or when they died. Furthermore, we considered treatment targets that were previously hypothesized to contribute most to the overall treatment effect: systolic blood pressure (sbp), LDL-cholesterol (ldl) and parathyroid hormone (pth) concentrations in blood, and protein excretion via urine (pcr). In total, there are 497 subjects with seven variables (both continuous and discrete) over five time slices. The first time slice contains the baseline observations taken before treatment, and the next time slices are the follow-up observations during treatment. Particularly we set the variable treatment only at $t_{i-1}$ as it remains the same over all time slices, and the variable inf_cens only at $t_i$ as it is a consequence of previous treatment. We further added the prior knowledge that gfr at $t_i$ does not directly cause any other variables, and that there are no relations between any variable and inf_cens within $t_i$. Both gfr and inf_cens are read-out for CKD progression and are within a time slice always the consequence and never the cause of another variable. However, we relax this prior knowledge at time slice $t_0$ as it is a common assumption that without the treatment, pth is a consequence of poor kidney function. The missing data is 5.2% and a single imputation with EM was conducted to impute the missing values like in applications to CFS and ADNI data. We performed the search over
100 subsamples of the original data set.

First we discuss the baseline model, which only considers the baseline causal relationships. Figures 4.11a and 4.11b depict the corresponding stability graphs. As in applications to CFS and ADNI data, \( \pi_{sel} \) is set to 0.6 and based on the search phase of S3L we found that \( \pi_{bic} = 2 \). Figures 4.11a and 4.11b show that two relevant edges were found. Applying the visualization procedure, we get the baseline model in Figure 4.12a. We found that both pth and pcr were associated with kidney function at baseline. The direction of these associations remains unclear. From renal physiology, we know that proteinuria may result in kidney damage. However, kidney damage and proteinuria may be common consequences of hypertension at an earlier stage in the patient’s history. The association between parathyroid hormone and gfr is unsurprising, as calcium and phosphate metabolism is disrupted in patients with advanced kidney disease. However, elevated pth may in turn result in further kidney damage by increased vascular calcification. In other words, the associations seem plausible from a physiological point of view, but the association may be in either direction. In the CKD example, a causal direction is almost impossible to ascertain when only using cross-sectional data.

Next we discuss the transition model, which takes into account all causal relationships across time slices. We set \( \pi_{sel} = 0.6 \) and found \( \pi_{bic} = 23 \). Based on Figure 4.11c, we obtained seventeen relevant edges, consisting of four intra-slice (blue lines) and thirteen inter-slice relationships of which five are between the same variables (orange lines) and eight are between different variables (black lines). Based on Figure 4.11d, we obtained twenty-six relevant causal paths, consisting of five intra-slice (blue lines) and twenty-one inter-slice relationships of which five are between the same variables (orange lines) and sixteen are between different variables (black lines). Applying the visualization procedure, we get the transition model in Figure 4.12b. Most of the intra-slice and inter-slice causal relationships are very stable with selection probabilities close to 1. We found inter-slice causal relationships from gfr, sbp, pth, and pcr to inf.cens. Furthermore, gfr, sbp, and pcr are well known determinants for CKD progression. The causal relationship from pth to inf.cens was somewhat surprising. However, pth is a marker for regulation of phosphate stores in the body and related to overall vascular damage through vascular calcification, and may thereby be related to mortality. Indeed, literature indicates that lowering pth in dialysis patients resulted in a reduction in mortality (Chertow et al., 2012). The same may hold true for patients who have CKD and who do not need dialysis treatment. Perhaps most surprising are the relations between sbp and pcr and gfr, respectively. From renal physiology we know that higher filtration pressures due to higher blood pressure causes the short term glomerular filtration rate to increase slightly (Johnson et al., 2014). Likewise, at higher filtration pressure, more and larger proteins are pushed out of the blood stream and into the pro-urine and are ultimately excreted via the urine. In the long term, chronically elevated filtration pressures and elevated levels of protein in the pro-urine cause kidney damage and ultimately even end stage kidney disease. Overall, the results are consistent with
Figure 4.11: The stability graphs of the baseline model in (a) and (b), and the transition model in (c) and (d), for chronic kidney disease, with edge stability in (a) and (c), and causal path stability in (b) and (d). The relevant regions, above $\pi_{\text{sel}}$ and left of $\pi_{\text{bic}}$, contain the relevant structures.
Figure 4.12: (a) The baseline model and (b) the transition model of chronic kidney disease. The dashed line represents a strong relation between two variables but the causal direction cannot be determined from the data. Each edge has a reliability score (the highest selection probability in the relevant region of the edge stability graph) and a standardized total causal effect estimation. For example, the annotation “1/0.88” represents a reliability score of 1 and a standardized total causal effect of 0.88. Note that the standardized total causal effect represents not just the direct causal effect corresponding to the edge, but the total causal effect also including indirect effects.

literature and physiology (Levin et al., 2013).

4.4 Conclusion and future work

Causal discovery from longitudinal data turns out to be an important problem in many disciplines. In the medical domain, revealing causal relationships from a given data set may lead to improvement of clinical practice, e.g., further development of treatment and medication. In the past decades, many causal discovery algorithms have been introduced. These causal discovery algorithms, however, have difficulty dealing with the inherent instability in structure estimation.

The present work introduces S3L, a novel discovery algorithm for longitudinal data that is robust for finite samples, extending our previous method (Rahmadi et al., 2017) on cross-sectional data. S3L adopts the concept of stability selection to improve the robustness of structure learning by taking into account a whole range
of model complexities. Since finding the optimal model structure for each model complexity is a hard optimization problem, we rephrase stability selection as a multi-objective optimization problem, so that we can jointly optimize over the whole range of model complexities and find the corresponding optimal structures. Moreover, S3L is a general framework that can be combined with alternative approaches, without modifying their original assumptions, e.g., linearity, non-Gaussian noise, etc.

The comparison on the simulated data shows that S3L achieves at least comparable performance as, but often a significant improvement over alternative approaches, mainly in obtaining the causal relations, and in the case of small sample size. Moreover, the results of experiments on three real-world data sets are corroborated by literature studies (Vercoulen et al., 1998; Wiborg et al., 2012; Heins et al., 2013; Haight et al., 2012; Levin et al., 2013; Henneman et al., 2009; Mungas et al., 2002; Rusinek et al., 2003; Chertow et al., 2012).

However, the current method considers only longitudinal data with observed variables and cannot handle missing values (other than through imputation as a pre-processing step). We also still assume that the time intervals between time slices are fairly uniform between subjects. Some existing approaches called random-coefficient models, also termed multi-level or hierarchical regression models (Raudenbush and Bryk, 2002; Kreft and de Leeuw, 1998), are flexible to handle unequal intervals between time slices within a subject and/or across subjects. Future research will aim to account for these aforementioned issues.
Chapter 5

Stable specification search with latent variables

Chapter 3 introduced stable specification search for cross-sectional data (S3C). It is an exploratory causal method that combines the concept of stability selection and multi-objective optimization to search for stable and parsimonious causal structures across the entire range of model complexities. S3C, however, is designed to model causal relations among observed variables. In this chapter, we extend S3C to S3C-Latent, to model causal relations between latent variables that are measured through observed proxies. We evaluated S3C-Latent on simulated data and compared the results to those of PC-MIMBuild, an extension of the PC algorithm, the state-of-the-art causal discovery method. The comparison shows that S3C-Latent achieved better performance. We also applied S3C-Latent to real-world data of children with attention deficit/hyperactivity disorder and data about measuring mental abilities among pupils. The results are consistent with those of previous studies.

5.1 Introduction

In many empirical sciences, it is of great interest to identify causal relationships among entities that are not measured directly. Such entities are called latent variables or factors. A latent variable is typically used to represent a rather general concept that influences multiple measured or observed variables (proxies) obtained from, for example, a questionnaire designed by experts in the field (Silva et al., 2006). An example from psychometrics is the fatigue catastrophizing scale (FCS; Jacobsen et al., 1999a), that is used to assess a patient’s tendency of catastrophizing, a cognitive process characterized by a lack of confidence and an expectation of negative outcomes (Sullivan and D’Eon, 1990), as a response of experiencing fa-
tigue. It is a 10-item scale, and each item therein, e.g., “I find myself expecting the worst when I am fatigued”, is rated on 5-point scales (1 means never true and 5 means all of the time true).

Domains involving both observed and latent variables can be modeled using SEM with latent variables (also called a general SEM, described in Chapter 2), that can be divided into two main components: the measurement model and the structural model. The measurement model captures relationships between the latent variables and their corresponding observed variables, often called indicators. The structural model represents relationships among latent variables. A few approaches have been developed to discover the measurement model, such as factor analysis or a more advanced alternative introduced by Silva et al. (2006). Recently, a new method has been introduced to model causal relations between latent variables in the structural model (Cui et al., 2018).

To estimate underlying causal mechanisms is, however, a complex problem. First, given \( p \) variables in the data, there will be \( 3^{p(p-1)/2} \) possible recursive structural models (described in Section 5.2; Harwood and Scheines, 2002). Thus, even a modest number of variables implies an immense number of possible models. A typical SEM procedure starts with a hypothesized model, scores the model, and realizes a few model refinements (often called a specification search) to improve the score (Long, 1983), resulting in only a small number of model evaluations. The term “score” here refers to an evaluation of a model, which is in the literature typically indicated by fit indices (Bollen, 1989). Such a specification search is not intended to infer causal structures, but rather to detect and correct specification error between a proposed and the true model (MacCallum, 1986). Furthermore, in some domains, such as the clinical domain, especially when related to rare diseases, a hypothesis is not always available or hard to propose. Considering the immense number of possible models, a more exploratory approach could be an alternative. Second, model estimation is known to be notoriously unstable. That is, a slight fluctuation in the data can lead to a considerable change in the final model. This makes that typical approaches in literature, that are based solely on a single run of estimation and do not take into account the variability of data, cannot robustly estimate a causal model.

There exist several exploratory specification search approaches in the literature, for instance, studies by Marcoulides et al. (1998); Marcoulides and Drezner (2001, 2003). Those studies are, however, based on the optimization of a single objective and do not take into account the instability problem. In Chapter 3, therefore, we introduced stable specification search for cross-sectional data (S3C; Rahmadi et al., 2017), an exploratory causal discovery method that combines the concept of multi-objective optimization and stability selection to resolve the problems of an immense number of possible models and the inherent instability in model estimation. S3C models causal relations among observed variables and thus uses a (non-general) SEM with only observed variables. In the present study, we extend S3C to S3C-Latent to model causal relations among latent variables and use the general SEM
representation. S3C-Latent, in particular, focuses on searching for causal relations on the structural model and, therefore, restricts the assumption on the measurement model such that it is given and pure (Silva et al., 2006), i.e., not allowing cross-loading indicators. In addition, we also allow demographic variables such as gender and age to be included in the structural model.

Our goal is also to show that a more exploratory approach should be considered as an alternative to resolve the problem of an immense number of possible models. It is, therefore, more sensible to compare the performance of S3C-Latent, our exploratory causal discovery method, with that of another causal discovery method, which is not exploratory. We thus consider a method called PC-MIMBuild (Silva et al., 2006) for comparison, an extension of the PC algorithm (Spirtes et al., 2000), a state-of-the-art causal discovery algorithm. The results demonstrate that S3C-Latent performs better. We further apply S3C-Latent to real-world data from different domains, and the results are consistent with those of earlier studies.

This chapter is organized as follows. Section 5.2 introduces the proposed method, S3C-Latent. Section 5.3 reports and discusses the results on simulated and real-world data. Finally, Section 5.4 gives the conclusions of this study.

5.2 Proposed method

Originally, S3C is designed to model causal relationships among observed variables. In the present study we extend S3C to S3C-Latent, to model causal relationships among latent variables. In particular, we use the representation of SEMs with latent variables as described in Section 2.2.2. As the basic idea of S3C is to search through many possible models, there is no restriction that forbids a cause $\xi_i$ in one model to be an effect $\eta_i$ in another model, and vice versa. The same condition implies to the corresponding indicators, that is, $x_i$ in one model can be $y_i$ in another model, and vice versa. However, there could be an exemption from this rule if one intends to incorporate prior knowledge.

In practice, prior knowledge on the domain of interest may exist, e.g., results of previous studies that lead to a constraint on a particular causal relation. For example, based on an earlier study on patients with medically unexplained fatigue, it is known that the objective physical activity does not reduce the level of fatigue directly. In terms of a SEM with latent variables, this prior knowledge can be translated into a structural model with no directed edge from latent variable $X$ (denoting the objective physical activity) to latent variable $Y$ (denoting the level of fatigue). Note that a directed path from $X$ to $Y$ is still allowed, e.g., a path $X \rightarrow \cdots \rightarrow Y$ with any latent variables in between. S3C and S3C-Latent allow for the incorporation of such prior knowledge.

In this study we further assume that the measurement model is given and pure. With respect to the data, we assume that samples have been generated independently and identically distributed (iid) from a linear Gaussian SEM.
5.2.1 S3C-Latent

Let $D$ be the data set, $L = \{L_1, \ldots, L_n\}$ be a set of $n$ latent variables, $\Lambda$ be a matrix of factor loadings, and $C$ be prior knowledge. Figure 5.1 gives pseudocode of the S3C-Latent procedure.

Lines 2 to 13 are to ensure that the measurement model identification conditions described in Section 2.2.4 are fulfilled. In particular, Line 3 checks if there are any latent variables $L_i \in L$ having less than 3 indicators. If so, then for each $L_i$, Line 4 is realized or Line 12 otherwise. Line 4 checks whether the number of indicators of $L_i$ is 2 or 1. In the case of 2 indicators, S3C-Latent sets a relation between the latent variable $L_i$ and a random latent variable $L_j \in L$ ($L_i$ can be either a cause or an effect), and fixes one of its factor loadings to 1 (Lines 5 and 6). In the case of 1 indicator, S3C-Latent sets the factor loading of $L_i$ to 1 and the indicator error to 0 (Lines 8 and 9). Line 12 is realized when all latent variables have at least 3 indicators. In this case, one of the factor loadings on each latent variable is set to 1.

Finally, Line 14 runs S3C on data set $D$ with information of latent variables from $L$, satisfying any constraints in $C$, and fulfilling model identification conditions in $I$. By satisfying constraints in $C$ (if any), S3C-Latent ensures that all SEMs that are generated and refined, are consistent with the prior knowledge stated in $C$. 

Figure 5.1: Algorithm of S3C-Latent.
We implemented S3C-Latent as an R package and made it available online.\(^5\) All of the source codes are provided, allowing interested users to modify the functionality.

5.3 Result and discussion

5.3.1 Application to simulated data

To simulate different practical cases, we conducted different schemes of simulation by generating data from models with different number of latent variables, sample size, number of indicators, and types of indicators, e.g., continuous and ordinal. All of the R scripts used and the data sets generated for this simulation are available at the same repository of the R package.

Model and data generation

First, we randomly generated structural models of \(n\) latent variables, by generating DAGs of \(n\) nodes with a probability \(s = 2/(n - 1)\) when relating a node (latent variable) to another node. Each DAG can be represented as an \(n \times n\) adjacency matrix \(A\). As the matrix \(A\) captures causal relations among latent variables \(L = \{L_1, \ldots, L_n\}\), the matrices \(B\) and \(\Gamma\) in Equation (2.4) are submatrices of \(A\). For each structural model, we then randomly formed two pure measurement models, where one has 3 to 5 indicators and the other one has 1 to 4 indicators per latent variable. This is realized by generating an \(h \times n\) random matrix of factor loadings \(\Lambda\) for each structural model, where \(h\) is the total number of indicators from \(n\) latent variables. The matrices \(\Lambda_x\) and \(\Lambda_y\) in Equation (2.5) are submatrices of \(\Lambda\). With this procedure, we generated 20 SEMs with 4 and with 6 latent variables.

Second, we simulated data sets of sizes 400, 1000, and 2000 from each SEM with normal distribution, and then created the corresponding ordinal data sets by discretizing the values of the indicators into 2 to 7 ordered categories randomly. Taken together, we simulated 480 data sets on which the performance of S3C-Latent has been evaluated. Table 5.1 summarizes the schemes of the simulation. Scheme \(C_{3-5}\), for example, applied S3C-Latent to the data sets of continuous variables generated from SEMs with 3 to 5 indicators per latent variable. Moreover, we also applied PC-MIMBuild (Silva et al., 2006) on the same data sets, and then compared the results. PC-MIMBuild originates from the PC algorithm (Spirtes et al., 2000), a state-of-the-art causal discovery method.

Parameter settings

A parameter setting of S3C-Latent consists of the number of subsets to draw (\(S\)), the number of iterations (\(I\)), the number of models to evaluate (\(P\)), crossover probability

\(^5\)https://github.com/rahmarid/S3C-Latent
Table 5.1: Types of indicators and the number of indicators per latent variable in different schemes of simulation.

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Type of indicators</th>
<th>Number of indicators</th>
</tr>
</thead>
<tbody>
<tr>
<td>C_{3-5}</td>
<td>Continuous</td>
<td>3 to 5</td>
</tr>
<tr>
<td>C_{1-4}</td>
<td>Continuous</td>
<td>1 to 4</td>
</tr>
<tr>
<td>O_{3-5}</td>
<td>Ordinal</td>
<td>3 to 5</td>
</tr>
<tr>
<td>O_{1-4}</td>
<td>Ordinal</td>
<td>1 to 4</td>
</tr>
</tbody>
</table>

(C), and mutation probability (M). For applications to data sets generated from SEMs with 4 latent variables, we set $S = 25, I = 30, P = 50, C = 0.45$, and $M = 0.01$. For applications to data sets generated from SEMs with 6 latent variables, we used the same parameter setting, except that we set $I = 50$ and $P = 100$.

For a fair comparison, we also drew 25 subsets in the applications of PC-MIMBuild, similar to that in Ramsey (2010), and measured the edge and the causal path stability as in S3C-Latent. We set the significance level for PC-MIMBuild in testing conditional independence to 0.01.

**Evaluation performance**

We used the receiver operating characteristic (ROC) curve (Fawcett, 2004) to evaluate the performance of S3C-Latent and of PC-MIMBuild. Here, the true positive rate (TPR) and the false positive rate (FPR) are computed based on the CPDAG of the true model (Chickering, 2002a), while increasing the thresholds of stability. We measured the ROC for both the edge and the causal path stability. For example, a true positive means that a causal path with any length predicted by S3C-Latent or PC-MIMBuild actually exists in the CPDAG of the true model, while a false positive means that the predicted causal path does not exist in the CPDAG of the true model.

Each simulation scheme comprises 20 data sets of sizes 400, 1000, and 2000, making it in total 60 data sets. We measured the area under the curve (AUC) of each ROC and then computed the mean of the AUCs that S3C-Latent obtained over 20 data sets (of the same sample size) and compared to that of PC-MIMBuild.

**Discussion**

Figures 5.2 and 5.3 show the plots of the mean AUC and the corresponding standard errors obtained by S3C-latent (black-solid lines) and PC-MIMBuild (blue-dashed lines), as a function of sample size. The top panels show results from Schemes $C_{3-5}$ and $O_{3-5}$, and the bottom panels show those from Schemes $C_{1-4}$ and $O_{1-4}$.

In all simulation schemes with 4 latent variables, S3C-Latent achieved higher AUCs than those achieved by PC-MIMBuild. In a similar trend, S3C-Latent obtained higher AUCs than those achieved by PC-MIMBuild on most simulation schemes with 6 latent variables. The discrepancy of the results of S3C-Latent and
Figure 5.2: The mean AUC obtained by S3C-Latent (black-solid lines) and PC-MIMBuild (blue-dashed lines) over 20 data sets generated from SEMs with 4 latent variables, as a function of sample size. “Edge” indicates AUCs measured from the edge stability and “Causal Path” indicates those measured from the causal path stability. The plots in the top panels are results of simulations on the data sets generated from SEMs with 3 to 5 continuous (Scheme C3−5) or ordinal (Scheme O3−5) indicators per latent variable. The plots in the bottom panels are results of simulations on the data sets generated from SEMs with 1 to 4 continuous (Scheme C1−4) or ordinal (Scheme O1−4) indicators per latent variable.

of PC-MIMBuild becomes clearer in the case of continuous indicators. Based on the results on simulations with 4 and 6 latent variables, the performance of S3C-Latent seems to be more robust compared to that of PC-MIMBuild across different sample sizes. In particular for small sample sizes, typical to many real-world applications, S3C Latent appears to be more robust than PC-MIMBuild.

Figure 5.8 in Appendix 5.A shows comparisons of S3C-Latent’s results in different schemes. As expected, in general the performance of S3C-Latent on data generated from SEMs with 3 to 5 indicators per latent variable is better than that of on data generated from SEMs with 1 to 4 indicators per latent variable.
Figure 5.3: The mean AUC obtained by S3C-Latent (black-solid lines) and PC-MIMBuild (blue-dashed lines) over 20 data sets generated from SEMs with 6 latent variables, as a function of sample size. “Edge” indicates AUCs measured from the edge stability and “Causal Path” indicates those measured from the causal path stability. More details on the plots in the top and the bottom panels are given in the caption of Figure 5.2.

5.3.2 Application to real-world data

Attention deficit/hyperactivity disorder data

For the first application, we applied S3C-Latent to a data set about attention deficit/hyperactivity disorder (ADHD) that was collected in van Steijn et al. (2012). The data set comprises observations on the children, among which 236 have ADHD and 406 belong to the control group. From all of the samples, 269 are girls. There are four observed variables: gender, age, verbal intelligence quotient (VIQ), and performance intelligence quotient (PIQ). Moreover, there are 18 questions distributed into three groups to indicate three latent variables, namely inattention (questions 1-9), hyperactivity (questions 10-14), and impulsivity (questions 15-18). We intend to model the causal relations between those four observed and the three latent variables. The missing data, that are assumed to be missing at random (MAR), are about 0.78% and were imputed using expectation maximization (EM; Honaker et al., 2011). We combined the imputation with our subsampling, resulting in dif-
different imputations across data subsets. The parameter settings of S3C-Latent are $S = 100$, $I = 100$, $P = 200$, $C = 0.45$, and $M = 0.01$. We included a prior knowledge that nothing causes gender.

**Discussion on ADHD result**

Figure 5.4 depicts the edge and the causal path stability graphs, with the $x$-axis indicating the model complexity and the $y$-axis indicating the selection probability. The threshold $\pi_{\text{sel}}$ (horizontal line) is set to 0.6 and indicates that any model structure with selection probability equal to or greater than $\pi_{\text{sel}}$ is considered stable (Meinshausen and Bühlmann, 2010). The threshold $\pi_{\text{bic}}$ (vertical line) is set to the model complexity at which the minimum median of the BIC scores is found (Rahmadi et al., 2017), which in this case was at 12, indicating that any model structure with model complexity equal to or lesser than $\pi_{\text{bic}}$ is considered simple or parsimonious. Thus, the relevant model structures (stable and parsimonious) are those that pass through the top-left region of the stability graphs. The relevant model structures are then visualized with the steps described in Section 3.2.1, resulting in the graph shown in Figure 5.5. A more detailed description on how to interpret the graph is given in the figure caption.

Based on Figure 5.5, gender is found to influence inattention, hyperactivity, and impulsivity. This is in accordance with the studies of Gaub and Carlson (1997); Gershon and Gershon (2002); Bauermeister et al. (2007) which exhibited that boys with ADHD tend to have higher symptom levels. Meta-analyses in population-based (Willcutt et al., 2000) and clinically referred samples (Novik et al., 2006) suggested that boys are more likely to meet the diagnostic and statistical manual of mental disorders (DSM) criteria for ADHD than girls.

VIQ is found associated with inattention and with hyperactivity, and that matches the studies of Andreou et al. (2005); Frazier et al. (2004) which reported that children with ADHD have significantly lower VIQ compared to children without ADHD. The relation between VIQ and PIQ is expected as both constitute an assessment of the subject’s intelligence.

The association between inattention and hyperactivity is likely due to the direct causal relations from gender to both variables. A similar relation was found in the studies of Sokolova et al. (2017, 2016) which exhibited that inattention influences hyperactivity. S3C-Latent found that the causal path stability from inattention to hyperactivity is about 0.4. Having more samples might increase the stability (e.g., > 0.6), so as to justify the same causal direction.

The relation between hyperactivity and impulsivity (as also found by Sokolova et al. (2017)) is also likely due to the direct causal relations from gender to both variables. In fact, in most studies hyperactivity and impulsivity are regarded as one combined feature.

The association between inattention and impulsivity possibly follows from the direct relations between inattention and impulsivity with hyperactivity. This as-
Figure 5.4: The (a) edge and the (b) causal path stability graphs from the ADHD application. Each line indicates the frequency of relations between a pair of variables across different model complexities. The frequency is computed relative to the number of SEMs in each model complexity. The causal path stability only takes into account causal relations with any length, while the edge stability takes into account all relations regardless of the direction.

Association was also indicated by Sokolova et al. (2017). The associations between inattention, hyperactivity, and impulsivity are sensible as the three constitute the ADHD symptom. In addition, S3L did not find any relation between age with other variables.
Figure 5.5: The causal model of ADHD data. A dashed undirected edge exhibits strong association where the causal direction cannot be determined from the data. A directed edge (arrow) with a single line indicates a causal relation and that with a double line indicates a factor loading. All edges, except the factor loadings, are annotated with a reliability score, i.e., the highest selection probability an edge has across the relevant region of the edge stability graph. In addition, all arrows are annotated with an estimate of the total causal effect (see Section 4.2.1 and Appendix 5.B for more detail). For example, $0.89/0.02$ indicates a reliability score of 0.89 and a total causal effect of 0.02.

**Holzinger data**

For the second application, we considered a well-known data set in the psychometric domain (Holzinger and Swineford, 1939), obtained from an R package called MBESS 4.4.3 (Kelley, 2007). The aim of the study was to learn factor patterns for mental abilities among two groups of pupils with different biological and cultural backgrounds (Scates, 1940). For this purpose, 301 students (155 females) from grades 7 and 8 were given 24 tests, measuring spatial, verbal, mental speed, memory, and mathematical ability. More details about the tests can be obtained from the package documentation. Our interest is to model causal relations among those latent variables including gender. We include a prior knowledge that nothing causes gender. We set the S3C-Latent parameters $S = 100$, $I = 80$, $P = 150$, $C = 0.45$, and $M = 0.01$. 
Discussion on Holzinger result

Figure 5.6 displays the edge and the causal path stability graphs. We set $\pi_{sel} = 0.6$ and found that $\pi_{bic} = 11$ (cf. Section 5.3.2). Figure 5.7 visualizes the relevant structures into a graph.

Based on Figure 5.7, we see that gender influences spatial, mental speed, and memory ability. The causal relation from gender to spatial ability matches the meta-analysis of Linn and Petersen (1985) that indicated that males tend to perform better than females in specific types of spatial ability. The causal relation from gender to memory is in agreement with the results of Cattaneo et al. (2006) and Longman et al. (2007), that exhibited gender differences in favor of males on object and word location memory, and working memory index, respectively. Gender was also found to influence the mental speed, as indicated by the study of Der and Deary (2006) that pointed out a male advantage in reaction times using various measures.

It was also found that mathematical ability affects spatial ability. Previous studies indicated the association between the two (Burnett et al., 1979; Casey et al., 2001; Robinson et al., 1996). Other studies emphasized that mathematics is one of the concepts that is mentally represented in a spatial format (Barsalou, 2008; Lakoff and Núñez, 2000). Moreover, the study by Mix et al. (2016) detailed specific mathematical tasks that predict the most variance in the spatial ability of children aged 5 to 13.

All latent variables seem to be associated directly or indirectly. This might stem from some items on different latent variables that are possibly overlapping, for example, arithmetic (mathematical ability) and addition tasks (mental speed). Some associations are likely due to a common direct relation that a pair of variables has, for example, spatial ability and mental speed are affected by gender and both are associated. That being said, previous studies (Lynn, 1987; Cantor et al., 1991; De Smedt et al., 2009; Daneman and Carpenter, 1980; Vukovic and Lesaux, 2013; Mitolo et al., 2015) indicated and discussed most of the associations.

5.4 Conclusion and future work

It is often of great interest in many fields such as sociology, psychology, and medicine to model causal relationship among latent constructs that are indicated through observed proxies. In the present study, we extended S3C to S3C-Latent to model causal relations among latent variables. The chief aim of S3C-Latent is to resolve the problems of the inherent instability in model estimation and the immense number of possible models. To realize that, S3C-Latent recasts the concept of stability selection into a multi-objective optimization problem, and jointly optimizes across the whole range of model complexities, resulting in Pareto optimal models. One also can see S3C-Latent as a general framework that can be integrated with other causal discovery methods for latent variables without altering their original assumptions. In principle, it can be realized by running a causal discovery method on different
data subsets and then measuring the same edge and causal path stability based on the inferred models.

We compared the results of S3C-Latent to those of PC-MIMBuild on simulated data generated with different schemes, varying the number of latent variables, sample size, number of indicators, and types of indicators (continuous and ordinal). The comparison showed that S3C-Latent performs better than PC-MIMBuild. Moreover, we applied S3C-Latent to real-world ADHD and Holzinger data sets. In

Figure 5.6: The (a) edge and the (b) causal path stability graphs from Holzinger application. Each line indicates the frequency of relations between a pair of variables across different model complexities. The frequency is computed relative to the number of SEMs in each model complexity. The causal path only stability takes into account causal relations with any length, while the edge stability takes into account all relations regardless of the direction.
Figure 5.7: The causal model of Holzinger data. A dashed undirected edge exhibits strong association where the causal direction cannot be determined from the data. A directed edge (arrow) with a single line indicates a causal relation and that with a double line indicates a factor loading. All edges, except the factor loadings, are annotated with a reliability score, i.e., the highest selection probability an edge has across the relevant region of the edge stability graph. In addition, all arrows are annotated with an estimate of the total causal effect (see Section 4.2.1 and Appendix 5.B for more detail). For example, 0.62/0.017 indicates a reliability score of 0.62 and a total causal effect of 0.017.

general, the results are consistent with those of earlier studies.

The current version of S3C-Latent, however, is designed for cross-sectional data, assumes no reciprocal causal relation, and cannot handle missing values other than through imputation prior to application. Thus, future work should consider extensions to longitudinal data, reciprocal causal relations, and a more sophisticated way of handling missing values.
Appendix 5.A  Comparison of S3C-Latent results

Figure 5.8: Comparison between the mean AUC obtained by S3C-Latent on the data generated from SEMs with 3 to 5 indicators per latent variable to those obtained on data generated from SEMs with 1 to 4 indicators per latent variable. The plots in the top panel are results of simulation on the data generated from SEMs of 4 latent variables and those in the bottom panel are results of simulation on the data sets generated from SEMs of 6 latent variables.

Appendix 5.B  Sampling from latent variables

The procedure to estimate causal effect is described in Section 4.2.1. In order to estimate causal effect from a SEM with latent variables, we need to sample data from the latent variables. In what follows, we describe the steps for the sampling. Suppose we are given a measurement model that reads

\[ x = \Lambda \eta + \epsilon, \]  

where \( \eta \) is a vector of latent (for simplicity, can be either endogenous or exogenous) variables, \( \Lambda \) is a matrix of factor loadings, \( x \) is a vector of indicators, \( \epsilon \) contain errors of the indicators, and \( \Theta \) is a covariance matrix of \( \epsilon \) and is diagonal.
Following Ghahramani et al. (1996), given $\Lambda$ and $\Theta$, the expected value of latent variables $\eta$ can be computed via the linear projection:

$$\mathbb{E}(\eta|x) = \beta x;$$

(5.2)

with $\beta \equiv \Lambda'(\Theta + \Lambda\Lambda')^{-1}$, and the variance of $\eta$ can be computed through,

$$\text{Var}(\eta|x) = I - \beta\Lambda.$$

(5.3)

We can then sample $\hat{\eta} \sim \mathcal{N}(\mu, \sigma^2)$, where $\mu$ is the mean of vector $\mathbb{E}(\eta|x)$ and $\sigma^2 = \text{Var}(\eta|x)$. 
In Chapters 3 and 4 we introduced S3C and S3L, respectively. In this chapter, we describe an implementation of S3C and S3L as an R package called stablespec, which is available at CRAN with MIT license. We demonstrate a running example of the package on a data set of patients with attention deficit/hyperactivity disorder.

6.1 Introduction

Causal modeling aims to understand the underlying mechanisms by which variables in data relate to each other in terms of causal relations. It can also be seen as an attempt to find a generative model (Spirtes, 2010). Causal modeling often turns out to be an essential problem in many fields, e.g., Zhang et al. (2014); Chen and Zhang (2016). In the medical domain, for example, revealing causal relationships may lead to enhancement of clinical practice, e.g., the development of treatment and medication (Cooper et al., 2015). Stable specification search is a novel causal discovery method based on Rahmadi et al. (2017), for cross-sectional data (S3C), and Rahmadi et al. (2018c), for longitudinal data (S3L). The method is designed to overcome two problems in causal modeling: the issue that the number of possible models is super-exponential in the number of variables and the instability in model selection, i.e., that a slight change in the data can lead to a significant change in the final model. In this paper, we describe the R package stablespec, which contains an implementation of S3C/L.

This chapter is based on Rahmadi et al. (2018b), “The stablespec package for causal discovery on cross-sectional and longitudinal data in R”, published in Neurocomputing.
Our package \texttt{stablespec} attempts to infer the causal structure that best matches a given data set. It implements S3C/L (Rahmadi et al., 2017, 2018c), which in high-level terms works as follows. S3C/L models causal relations between variables using SEMs and uses an exploratory approach (i.e., without specifying an initial hypothesis) to search over the model space. S3C/L evaluates models according to two objectives: the model fit and the model complexity. Since both objectives are often conflicting, S3C/L uses a multi-objective optimization approach, called non-dominated sorting genetic algorithm II (NSGA-II), to search for Pareto optimal models. In addition, in order to deal with the inherent instability of structure estimation from finite data, S3C/L adopts the concept of \textit{stability selection} using subsampling and selection algorithms (Meinshausen and Bühlmann, 2010).

\section{6.2 Implementation and functionalities}

The package provides a main function and several support functions. The main function, \texttt{stableSpec}, is used for searching optimal model structures and computing stabilities. Parallel computation is facilitated through parallel backend registration. Some additional functions are provided to increase usability, e.g., \texttt{modelPop} for generating random SEM models, \texttt{repairCyclicModel} for repairing a cyclic model so as to be acyclic, \texttt{plotStability} for visualizing the stability of model structures, \texttt{getModelFitness} for scoring SEM models, and \texttt{dataReshape} for reshaping longitudinal data. We add data sets for users to be able to explore \texttt{stablespec} directly without loading external data. Documentation is bundled alongside the package, giving the user detailed guidelines, e.g., each function is accompanied by a running example that users can adopt to their case.

The \texttt{stablespec} package is available at the Comprehensive R Archive Network (CRAN) with MIT license. The package depends on R at least version 3.1.0 and some other R packages, for instance, \texttt{ggm}, \texttt{sem}, \texttt{nsga2R}, \texttt{polycor}, \texttt{foreach}, \texttt{graph}, and \texttt{Rgraphiz}. As the mentioned package dependencies are on both CRAN and Bioconductor, the \texttt{stablespec} can be installed from the R console by typing \texttt{setRepositories(ind=1:2)} and then \texttt{install.packages("stablespec")} in the next line.

\section{6.3 Experimental result}

To demonstrate the package, we consider a data set which describes phenotypic information of children with attention deficit/hyperactivity disorder (ADHD; Cao et al., 2006). The data set\footnote{Available at \url{https://github.com/rahmarid/dataset}} consists of 221 subjects, with eight variables as described in Figure 6.2. The following example assumes that the package \texttt{stablespec} has been loaded (see the documentation for details on the function arguments). The
Figure 6.1: Plots of edge stability (blue line) and causal path stability (green line is for causal path with length one and red line is for any length) between two variables. The green line between variables 1 and 4 (refer to the first and fourth attributes/columns in the data set, respectively) is covered by the blue line as causal path stability with length one and the edge stability have the same value.

First two lines are for parallel computation, which requires the packages `parallel` and `doParallel`; to compute sequentially, simply remove these lines.

```r
> cl <- makeCluster(detectCores())
> registerDoParallel(cl)
> result <- stableSpec(theData = read.csv("ADHD.csv"), nSubset = 100, nPop = 120, longitudinal = FALSE, mixture = TRUE, consMatrix = matrix(c(2, 1, 3, 1, 4, 1, 5, 1, 6, 1, 7, 1, 8, 1), 7, 2, byrow = TRUE), toPlot = FALSE)
```

Figure 6.1 is an output example using `plotStability` which shows the stability graphs between some variables based on the result above. Model complexity is on the x-axis and selection probability on the y-axis. The horizontal line is the boundary of the selection probability \( \pi_{\text{sel}} \) (argument `threshold` of the function `stableSpec`) and the vertical line is the boundary of the model complexity \( \pi_{\text{bic}} \) (set to the level of the model complexity at which the minimum average Bayesian information criterion (BIC) score is found). The blue line represents the edge stability which constitutes relations between pairs of variables regardless of the direction, while the green and red lines represent the causal path stability with length one and with any length, respectively, which constitute causal relations from variables to other variables. Relevant structures are defined to be those edges and causal paths with a selection probability higher than or equal to \( \pi_{\text{sel}} \) and with a model complexity lower than or equal to \( \pi_{\text{bic}} \). Thus in Figure 6.1, the relevant structures are represented by lines that pass through the relevant (top-left) region of the plot. In addition, `plotStability` returns plots of the aggregated edge stability (Figure 6.2a) and the aggregated causal path stability (Figure 6.2b). The R scripts to generate the plots in Figures 6.1, 6.2a, and 6.2b using `plotStability` are as follows.
Figure 6.2: (a) The edge and (b) the causal path stability graphs. The top-left region is the area containing the relevant structures.

> plotStability(listOfFronts = result$listOfFronts, stableCausal = result$causalStab, stableCausal_l1 = result$causalStab_l1, stableEdge = result$edgeStab, longitudinal = FALSE)

Figure 6.3 provides a visualization of the relevant structures (visualization is not part of the package, but left to one’s favorite drawing software) obtained through the following steps.

First, the nodes are connected according to the relevant edges obtained. Second, the edges are oriented according to the prior knowledge added. The fact that nothing causing variable gender results in three directed edges gender → AD,
gender → aggression, and gender → medication. Third, the edges are oriented according to the relevant causal paths of length one. This results in six directed edges, AD → HI, AD → medication, AD → aggression, HI → medication, perfIQ → fullIQ, and verbIQ → fullIQ. The remaining undirected relevant edges (dashed lines), e.g., between HI and aggression, represent strong associations which direction cannot be determined from the data.

The causal relations obtained are corroborated by studies reported in the literature. In Sokolova et al. (2014), gender is shown to be a direct cause for attention deficit; attention deficit is a direct cause for hyperactivity, medication, and aggression. Moreover, hyperactivity and aggression are related but neither variable is a direct cause for the other.

### 6.4 Conclusion

As an R package, stablespec gives users the flexibility to replicate and extend the algorithms within the R framework. Moreover, comparisons of S3C/L (stablespec) with some other algorithms (R package pcalg and a standalone software TETRAD) show that S3C/L achieve significant improvements over alternative approaches in retrieving causal relations (Rahmadi et al., 2017, 2018c).
Cognitive behavior therapy (CBT) for chronic fatigue syndrome and other conditions characterized by medically unexplained fatigue lead to a reduction of fatigue severity. Several mediation analyses of treatment studies suggested that cognitive-behavioral variables mediate the reduction of fatigue brought on by CBT. However, most studies did not analyze causal interactions among mediators. In this chapter, we apply S3L to study these interactions to determine how the effect of CBT on fatigue is mediated. More specifically, we apply S3L to a data set of 264 patients, which originates from two different earlier studies. The first study tested the efficacy of CBT in chronic fatigue syndrome (n = 169) through a randomized controlled trial. The second study conducted similar tests but in idiopathic chronic fatigue (n = 95). Potential mediators for the reduction in fatigue were fatigue-related beliefs, perceived and actual physical activity, and physical functioning. We found that CBT led to an increased self-efficacy with respect to fatigue, a reduced focus on fatigue, and a reduced tendency to catastrophize. Changes in cognitive processes directly caused a reduction in fatigue. The same changes also led via an improved physical functioning and increased perceived activity indirectly to a further decrease of fatigue. We conclude that CBT seems only to have a direct effect on fatigue-related cognitive processes. The positive effect of the therapy on fatigue severity is mediated by changes in these cognitive processes, both directly and indirectly via physical functioning and perceived activity. This knowledge can be used in the further development of CBT for medically unexplained fatigue.
Chapter 7

7.1 Introduction

Chronic fatigue syndrome (CFS) is characterized by severe and persistent fatigue with no known medical explanation. The fatigue leads to substantial functional impairment (Fukuda et al., 1994; Reeves et al., 2003). According to the cognitive behavioral model of CFS, behavior and beliefs maintain symptoms (Knoop et al., 2010). Through cognitive behavior therapy (CBT), these perpetuating factors can be addressed. Although different treatment protocols exist, all of them emphasize the importance of regulating the sleep-wake pattern, a graded activity program, and the restructuring of dysfunctional beliefs about fatigue and the ability to undertake activities (Heins et al., 2013). Several meta-analyses have shown that CBT for CFS induces significant reduction of fatigue (Castell et al., 2011; Malouff et al., 2008; White et al., 2013, 2011). CBT based on the same model and protocol also has a positive effect on other conditions with medical unexplained fatigue (see for example Janse et al., 2016). A substantial number of patients, however, do not profit from CBT, or only show a limited reduction of fatigue. This suggests that there is room for improvement by adapting interventions. More insight into the mechanisms that are responsible for the reduction of fatigue can help to focus more on elements of the intervention crucial for change. This could perhaps improve outcome.

Mediation analysis is widely used to identify the mechanisms of change of interventions. The mediation model assumes a causal chain with a direct causal path from the independent variable to the outcome variable, and an indirect causal path from the independent variable to the outcome variable via the mediator (Baron and Kenny, 1986). Previous mediation analyses and process research of CBT (Heins et al., 2013; Wiborg et al., 2012, 2011, 2010), were based on a longitudinal design and used an empirically tested cognitive-behavioral model of CFS as a starting point (Vercoulen et al., 1998). These studies found that an increased self-efficacy with respect to fatigue, a reduced focus on symptoms, and an increased perceived activity mediated the positive effect of CBT on fatigue (Heins et al., 2013; Wiborg et al., 2012, 2011). Remarkably, an increase in objective physical activity did not seem to be a mediator (Wiborg et al., 2010). These studies made use of change scores for mediators and/or outcome variables. The change scores were calculated based on the longitudinal data, i.e. new variables as a result of subtracting post-treatment from baseline assessment. Several studies however, discussed the reliability and validity of change scores (Bohrnstedt, 1969; Campbell and Stanley, 1963; Cronbach and Furby, 1970; Dalecki and Willits, 1991; Kessler, 1977), which seem to be less than those of the original data from which the scores are calculated.

There are studies which have used longitudinal data more explicitly. In a longitudinal mediation analysis of the change between end of treatment up to follow-up, Wearden and Emsley (2013) also found evidence for the role of cognitive processes in the reduction of fatigue following a cognitive-behavioral intervention for CFS. Aside from a reduced tendency to limit activities, less catastrophizing responses to fatigue was a mediator. A second longitudinal mediation analysis (Chalder et al.,
Chapter 7. Identifying causal mechanisms of CBT for fatigue

assessing the change in potential mediators during CBT for CFS, again found that beliefs about activity and symptoms, especially the reduction in fear-avoidance beliefs, mediated the reduction in fatigue. Instead of using change scores, the two studies used a mediation model in which mediators and outcome variables are assessed at two different time slices, e.g. mediators were selected from variables assessed at week 12 and the outcome variable at week 52 after randomization. In addition, they included several baseline variables as covariates in order to reduce potential biases due to unmeasured confounders. The mediation models used in these studies do not consider multiple mediators in one model, that likely interact in a complex intervention like CBT. For example, it could be that an increase in objective activity does not directly influence fatigue, but does so via a positive effect on perceived activity. Alternatively, it could be that an increase in self-efficacy brought on by CBT leads to (a further) reduction in the focus on fatigue and in this way reduces fatigue. Several approaches that allow models with multiple interacting mediators have been introduced. Those approaches are based on, among others, regression and weighting (VanderWeele and Vansteelandt, 2014), structural equation model (SEM; Gunzler et al., 2013), and natural effects models (Steen et al., 2017).

The purpose of the present study is to construct a causal mediation model for medically unexplained fatigue via a novel SEM-based causal modeling method called stable specification search for longitudinal data (S3L; Rahmadi et al., 2018c). The S3L method is an exploratory approach, instead of the more common hypothesis-based confirmatory approach (Chalder et al., 2015; Goldsmith et al., 2018; Wiborg et al., 2012). Compared to the typical confirmatory approach that only tests a few models, S3L builds a more robust model through a systematic model evaluation that tests many possible relationships between variables, which is then continued by a selection of stable and parsimonious causal structures from the evaluated models. S3L allows us to make preliminary predictions about the causal model underlying the data that can be analyzed further using mediation analysis or additional experiments.

In the present study we applied S3L to data from two published randomized controlled trials (RCT) that tested the efficacy of CBT for medically unexplained fatigue in patients with CFS and patients with idiopathic chronic fatigue (ICF) as defined by the US center for disease control criteria, revised in 2003 (Janse et al., 2016; Wiborg et al., 2015). ICF is thought to be a less severe disorder than CFS (Janse et al., 2016) and is considered when patients suffer from unexplained chronic fatigue but fail to meet all CFS criteria (Fukuda et al., 1994). Both studies found that fatigue severity decreased significantly following CBT compared to a waiting list. Using this novel method, we tested to what extent we could replicate the previous findings of mediation analyses and determine the interactions between mediators. We tested to what extent the positive effects of CBT on fatigue severity were mediated by an increased self-efficacy, decreased focus on symptoms and catastrophizing of fatigue (Chalder et al., 2015; Wearden and Emsley, 2013). We also included the level of physical functioning, and perceived and objectively assessed
physical activity in the model. Although previous studies seem to indicate that the latter does not mediate the effect of CBT, not all studies came to this conclusion (Chalder et al., 2015). As S3L enabled us to test interactions among mediators, we determined if a change in objective physical activity indirectly, via a change in the mediators contributed to the reduction of fatigue (Wiborg et al., 2010).

7.2 Assessments

7.2.1 Primary outcome measure

Fatigue severity was assessed using the subscale “fatigue severity” of the checklist individual strength (CIS; Worm-Smeitink et al., 2017). The subscale has 8 items, each item is scored on a 7-point Likert scale. The total score ranges from 8 to 56 where higher scores indicate more severe fatigue. The CIS has good psychometric qualities (Worm-Smeitink et al., 2017).

7.2.2 Mediator variables

Self-efficacy with respect to fatigue was assessed with the self efficacy scale (SES). The SES is a 7-item scale, scored on a 4-point Likert scale. Its total score ranges from 7 to 28 where higher scores reflect a higher self-efficacy. The SES has a good internal consistency (Prins et al., 2001).

Symptom focusing was assessed with the subscale “focusing on symptoms” of the illness management questionnaire (Ray et al., 1993). This subscale consists of 9 items which scored on a 6-point Likert scale. Higher scores indicate more focusing on symptoms. The subscale has good internal consistency (Heins et al., 2013).

Fatigue catastrophizing was measured with the fatigue catastrophizing scale (FCS), a 10-item scale developed by Jacobsen et al. (1999a). Each item is rated on a 5-point Likert scale. Higher total scores reflect a stronger tendency to catastrophize. The subscale has a good internal consistency (Jacobsen et al., 1999b).

Perceived activity was measured with the subscale “physical activity” of the CIS. This 3-item subscale assesses the perceived level of physical activity in the previous 2 weeks. The items are measured on a 7-point Likert scale with higher scores reflecting lower levels of activity. The subscale has good internal consistency (Heins et al., 2013).

Physical functioning was measured with the subscale “physical functioning” of the MOS Short Form-36. The subscale consists of 10 items assessing the level of physical functioning. Scores can range from 0, indicating maximal limitations, to 100 (Stewart et al., 1988). The SF-36 has good psychometric qualities.

Objective physical activity was assessed using actigraphy. Patients wore an activity sensing device around the ankle for 12 days registering the mean number of accelerations for every 5 minutes. A mean daily activity score over the 12 days
is calculated (de Vree et al., 2002). Higher scores reflect higher levels of physical activity.

7.2.3 Statistical methods

S3L searches for models on different data subsets that are half-size samples drawn from the original data without replacement. Here, the models are characterized by two criteria: the model fit and model complexity. For every data subset, S3L repeatedly calculates and evaluates numerous possible SEMs over different generations; for every next generation, new models are created based on the best models from the previous generation (Deb et al., 2002). A generation refers to a procedure which evaluates models and then refines their structures to improve scores. The searches on different data subsets result in a large group of models. S3L groups the best models based on their model complexity (the number of relations in a model), from the least to the most complex, and then plots the stability graphs (see Figure 3.2 as an example; Rahmadi et al., 2018c). Relations between a pair of variables are of two types: edge, representing any relation regardless of its causal direction, and causal path, representing a causal relation from a variable to another variable. The stability (also called selection probability) of a relation between a pair of variables refers to the frequency of the relation relative to the number of the best models at a particular model complexity. Stability graphs exhibit the stability of all possible relations across different model complexities. In Figure 3.2, the stability is indicated by the colored and dashed lines. Following the two types of relation, there are edge and causal path stability graphs which depict stability of any relations and causal relations among variables, respectively.

From the two stability graphs, S3L determines stable and parsimonious model structures (called relevant edges and causal paths) using two thresholds. The first threshold is called the stability boundary (indicated by the horizontal line $\pi_{\text{sel}}$ in Figure 3.2), that is used to determine which relations are stable. This threshold is typically set to 0.6 (Meinshausen and Bühlmann, 2010), meaning that a particular relation with a selection probability of at least 60% is considered stable. The second threshold is called the complexity boundary (indicated by the vertical line $\pi_{\text{bic}}$ in Figure 3.2), and used to select parsimonious relations. This threshold is indicated by a model complexity at which the minimum median Bayesian Information Criterion (BIC) scores is obtained. BIC is a score that introduces penalty term for adding relations between variables. It is computed on all models, and the median BIC scores of the best models in every model complexity is calculated and used for the second threshold described above.

A relation that appears in the top-left region of the stability graph, i.e., to the left of $\pi_{\text{bic}}$ and above $\pi_{\text{sel}}$ is considered relevant. Finally, S3L visualizes the relevant model structures into a causal graph using the steps described in Section 3.2.1. The causal graph consists of undirected edges, representing associations between two variables, from which the causal direction cannot be determined from the data alone,
and of arrows, representing causal relations from a variable to another variable. Each association is annotated with a reliability score. This score is taken from the highest selection probability that a particular relation obtains in the relevant region (i.e., the top-left region) of the edge stability graph, and thus ranges from 0.6 to 1. A higher reliability score indicates more confidence that the relation is part of the causal mechanism. In addition, each arrow is annotated with an estimate of total causal effect (see Section 4.2.1), that exhibits the size of influence a cause has on its effect, where a larger (positive or negative) causal effect signifies bigger influence.

S3L represents longitudinal causal relations with a two-time-slice model see (Figure 4.2b; Friedman et al., 1998). The first time slice ($t_{i-1}$) represents preceding, here baseline, observations and the second time slice ($t_i$) represents follow-up, here post-treatment or post waiting list, observations. Causal relations are only possible from the preceding to the follow-up time slice or within the follow-up time slice.

Baseline characteristics of patients and values on the measures were determined using the descriptive analysis with IBM SPSS version 22.020. Missing values of possible mediators were imputed using expectation maximization (EM; Honaker et al., 2011). We added as background knowledge that fatigue in the follow-up time slice caused nothing. All further computation was conducted in R. S3L has been implemented as an R package called stablespec (Rahmadi et al., 2018b) which is available at CRAN.7

7https://cran.r-project.org/web/packages/stablespec/index.html

### 7.3 Result

#### 7.3.1 Baseline characteristics and treatment effect

Combining data from the two studies (described in Section 7.1) results in a data set of 264 patients, of which 169 patients are from the study on CFS and the other 95 patients are from the study on ICF. Mean (SD) age of patients at baseline assessment was 37.5 (11.8) for the intervention group and 36.0 (11.8) for the waiting list group. Most patients were female (194/264, 73.5%). The mean (SD) fatigue severity at the first assessment (baseline) was 49.6 (5.4) for the intervention group and 48.2 (5.1) for the waiting list group. At the second assessment, the mean (SD) was 31.1 (13.9) for the CBT group and 44.3 (10.3) for the waiting list group. Table 7.1 shows the mean scores for all variables at both assessments for the two groups and the number of missing data. In total 5.9% of the data on possible mediators were missing and imputed.

#### 7.3.2 Causal mediation model of CBT for CFS

Figure 7.1 shows a visualization of the stable and parsimonious model structures (variables names are shortened for simplification; see the figure caption for more...
Table 7.1: The mean (M) and standard deviation (SD) score for each variable of the 264 patients based on the corresponding groups, including the number of missing values (MVs).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Treatment group</th>
<th>Wait-list control group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First assessment</td>
<td>Second assessment</td>
</tr>
<tr>
<td>Self-efficacy</td>
<td>17.35(2.90)</td>
<td>22.14(3.77)</td>
</tr>
<tr>
<td>MVs</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>Symptom focusing</td>
<td>31.44(9.08)</td>
<td>20.87(7.39)</td>
</tr>
<tr>
<td>MVs</td>
<td>0</td>
<td>65</td>
</tr>
<tr>
<td>Fatigue catastrophizing</td>
<td>22.28(6.38)</td>
<td>16.13(4.61)</td>
</tr>
<tr>
<td>MVs</td>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td>Physical functioning</td>
<td>60.48(19.74)</td>
<td>80.03(18.95)</td>
</tr>
<tr>
<td>MVs</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Perceived activity</td>
<td>69.93(15.67)</td>
<td>74.32(18.00)</td>
</tr>
<tr>
<td>MVs</td>
<td>3</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: This measure was introduced during the group study, leading to missing values (MVs).

Groups, including the number of missing values.

Table 7.1: The mean (M) and standard deviation (SD) score for each variable of the 264 patients based on the corresponding...
All causal relationships are represented by solid arrows. A strong association of which the direction could not be determined was represented by a dashed line.

According to the results of the causal modeling, depicted in Figure 7.1, CBT led to an increased self-efficacy, a reduced focusing on symptoms, and a reduction of fatigue catastrophizing. The relationship between these variables and the reduction of fatigue severity is complex. Self-efficacy has both a direct and indirect causal relationship with fatigue severity. Indirectly, an increased self-efficacy reduced fatigue severity by increasing perceived activity and physical functioning. Self-efficacy was also associated with focusing on symptoms. The direction of the relationship, however, was unclear from the data alone; whether increased self-efficacy caused a reduced focus on symptoms or the other way around cannot be determined. A reduced focusing on symptoms led directly to a reduction of fatigue but was also associated with an increased level of perceived physical activity. The direction of this association, is again, unclear. Fatigue catastrophizing only has an indirect effect on fatigue severity through its causal relationship with symptom focusing. Finally, there is no direct relationship between objective physical activity and fatigue.

### 7.4 Discussion

We applied a novel method for causal mediation analysis to understand how CBT leads to a reduction of fatigue in patients with CFS and ICF. It was found that CBT brought on changes in fatigue-related cognitive processes. The increased self-efficacy with respect to fatigue and the reduced focus on symptoms directly caused a reduction of fatigue severity. The increased self-efficacy also led to an improved physical functioning of patients and caused an increased level of perceived activity. These two relationships constituted causal interactions among mediator variables which indirectly contributed to a further reduction of fatigue. CBT also reduced the tendency to catastrophize in response to fatigue. Fatigue catastrophizing had no direct positive effect on fatigue severity, but indirectly via a reduced focusing on symptoms, again a causal interaction among mediating variables.

These findings partly replicate results of previous mediation analysis of CBT for CFS. An important difference with previous studies is that CBT only seems to have direct effects on fatigue-related cognitive processes. There is no direct effect of CBT on (perceived) physical activity and physical functioning. An improved physical functioning and increased perceived activity do lead to reduction of fatigue but CBT only indirectly exerts a positive influence on both variables.

The important role of fatigue-related beliefs in mediating the response of CBT for fatigue is not limited to CFS or ICF. In mediation analysis of CBT for severe and persistent fatigue in chronic medical conditions like diabetes and multiple sclerosis, it was also found that changes in fatigue-related cognitive processes such as an increased self-efficacy, a reduction of the focus on fatigue, and a change in the perception of fatigue (partially) mediated the response to CBT (Knoop et al., 2012;
Figure 7.1: A visualization of the relevant structures that represent stable and parsimonious causal structures. For simplification, variable names described in Section 7.2 are shortened in this figure, becoming fatigue (fatigue severity), pActivity (physical activity), functioning (physical functioning), focusing (symptom focusing), self-efficacy (self-efficacy with respect to fatigue), catastrophizing (fatigue catastrophizing), and oActivity (objective physical activity). A solid arrow represents a causal relationship and a dashed line represents a strong association between two variables for which the causal direction cannot be determined from the data. Each relationship is annotated with a reliability score and a total causal effect estimate. The dashed line is annotated only with the reliability score because the causal effect cannot be determined. For example, “0.92/0.52” means that the reliability score is 0.92 and the total causal effect is 0.52. The reliability score indicates the strength of relationships, ranging from 0.6 to 1. The total causal effect indicates the change in an effect variable due to its cause variable. The first assessment represents baseline and the second assessment was done directly after the waiting period or intervention. To remove clutter from the diagram we use the following abbreviations: a = 0.72/−0.42, b = 0.80/0.22, c = 0.84/−0.14, d = 0.86/−0.28, e = 0.96/0.08, and f = 1/0.29.

Menting et al., 2018). This study again showed that a change in physical activity does not have a direct effect on fatigue. Furthermore, CBT does not increase
objective physical activity in both RCTs. This replicates findings of studies testing mediation of CBT for cancer related fatigue and fatigue in diabetes (Gielissen et al., 2012; Menting et al., 2018).

It has been repeatedly shown that CBT is followed by an improved physical functioning (Castell et al., 2011). In the treatment model presented here, this effect is mediated by an increased self-efficacy with respect to fatigue. Remarkably, no relationship was found between changes in objective activity and physical functioning.

Our study suggests that it is important to focus more on the changes in the cognitive processes during therapy. According to the treatment model found, the beliefs about fatigue are important mediators of the reduction of fatigue through CBT. By evaluating the change in these beliefs during the therapy, the therapist can determine if CBT is effective in the individual patient and adjust the intervention accordingly. More generally, it would be helpful to determine the relation of specific interventions in the protocol, e.g., discussing treatment goals or introducing the CBT model, with the change in beliefs about fatigue to better understand how the elements of the protocol are linked to the changes needed to positively influence fatigue.

This study illustrated that S3L, the novel method for causal modeling, makes it possible to investigate more complex relationships between mediators. However, as S3L uses a SEM as a causal representation, it is important to have a substantial sample size of at least 200 to arrive at a viable result (Kline, 2011; Weston and Gore Jr, 2006). Because of this we had to combine the data of two RCTs to increase the sample size.

In combining both RCTs we assumed that the mediators of CBT for its effect on fatigue severity are the same for CFS and ICF. The relatively small sample size of each study separately makes it difficult to test this assumption. However, as our model replicates previous findings, it is likely that the changes brought on by CBT do not differ between CFS and ICF. There is also no indication that the effects of CBT are different in both groups (White et al., 2011; Janse et al., 2016; Wiborg et al., 2015).

S3L assumes a recursive SEM model in which reciprocal relationships are not allowed. In the case of CBT for CFS and ICF, this was reflected in the assumption that fatigue in the follow-up time slice does not cause anything. This is a questionable assumption. More realistically, the relationship between fatigue and beliefs and behavior is reciprocal: a reduction in fatigue may have a positive effect on self-efficacy or leads to a reduction of the focus on fatigue. Both can in return reduce fatigue further. This cannot be modeled in S3L. An algorithm called cyclic causal discovery (CCD; Richardson and Spirtes, 1996) is designed for modeling reciprocal relationships and may be suitable after it has been adapted to handle longitudinal data. This could help to create more sophisticated models of the effect of CBT on fatigue.

S3L also assumes no latent (i.e., non-observed) variables. A latent variable is
typically assumed to represent an unmeasured confounder that might cause bias when estimating a relationship between a pair of variables. A latent variable is also typically used to represent a more general concept that is measured through some observed variables. Taking such latent variables into a causal model may help us to understand causal relations that may underlie the mechanisms of change brought by CBT.

Also, only a limited number of possible mediators were analyzed. For CBT for CFS and ICF there are likely other candidates that may mediate the reduction in fatigue, such as the normalization of circadian rhythms brought on by therapy or the change in interactions with significant others with respect to fatigue (Knoop et al., 2010). It would be interesting to add these variables to the model to determine if and how they contribute to the reduction of fatigue.

Another limitation of our study is the use of self-report questionnaires to assess mediators. These questionnaires rely on the judgement of patients and are susceptible to demand characteristics. It is also likely that the constructs the questionnaires aim to assess overlap which makes the modeling more complex. One could also use experimental methods to assess cognitive processes, such as the assessment of attentional biases. These assessments could be used to gain more insight in which cognitive processes mediate the effects of CBT for medically unexplained fatigue (Hughes et al., 2016; van Der Schaaf et al., 2015).

It would be interesting to apply this new method to analyze mediation in graded exercise therapy (GET), another evidence based intervention for CFS (Larun et al., 2016). Perhaps other variables mediate the positive effect of this intervention. On the basis of our results one would not expect that an increase in objective activity would mediate the positive effects of exercise, as such a relationship between objective activity and fatigue does not seem to exist according to our model. It would however be possible that GET has a direct positive effect on perceived activity and physical functioning which will lead to a reduction of fatigue. If this is the case, a combination of elements of GET and CBT could perhaps lead to better outcomes.

This study showed that CBT for medically unexplained fatigue leads to a change in fatigue-related cognitive processes. These changes lead directly to a reduction of fatigue or indirectly via an improved physical functioning and increased perceived activity.
Chapter 8

Clinical significance of hyponatremia in TBM

Hyponatremia is common in patients with tuberculous meningitis (TBM), but its relation with disease severity and outcome is largely unknown. In this chapter, we examine the prevalence and significance of hyponatremia in a prospective cohort of adult TBM patients in Indonesia. The prospective cohort had been conducted between October 2006 and November 2017 in the Hasan Sadikin Hospital, gathering observations on 1261 patients with suspected TBM. For the present study, we include 678 patients based on characteristics of interest and data completeness. Among the 678 adult TBM patients (median age 29 years, 61.2% male, 16% HIV-infected), 83% had hyponatremia, including 17.8% with severe hyponatremia (<120 mEq/L). Patients with severe hyponatremia more often had concomitant pulmonary tuberculosis infection, culture-confirmed TBM, a lower Glasgow Coma Scale (GCS), higher cerebrospinal fluid (CSF) inflammatory markers, and higher blood neutrophils (p-value < 0.05). Besides Cox regression, we use S3C-Latent (described in Chapter 5) to determine relationships between hyponatremia and other markers, and to identify predictors of mortality. Severe hyponatremia was associated with higher mortality in univariate Cox regression (hazard ratio [HR] 1.68; 95% confidence interval [CI], 1.16–2.44, p-value = 0.006) but only in HIV-negative patients and not in multivariate analysis. In the causal survival model, the most stable predictor of mortality was HIV infection, but not hyponatremia. Although hyponatremia is common in patients with TBM and associated with clinical severity, CSF inflammation, and death, we conclude that hyponatremia does not seem to contribute to increased mortality, and aggressive correction of hyponatremia is therefore unlikely to significantly improve prognosis.
8.1 Background

Meningitis is one of the most severe forms of tuberculosis (TB; Figaji and Fieggen, 2010; Wilkinson et al., 2017), often leading to death or neurological disability (van Laarhoven et al., 2017). Hyponatremia, defined as a blood sodium level $< 135$ mEq/L (Verbalis et al., 2013), is reported in 42-73% of adult tuberculous meningitis (TBM) patients (Jaffri and Ahsan, 2015; Misra et al., 2016). Its symptoms may vary, and severe hyponatremia may lead to neurologic abnormalities such as lethargy, seizures, and coma due to brain oedema (Sterns et al., 1986; Tzamaloukas et al., 2013; Verbalis et al., 2013; Spasovski et al., 2014). Neuroendocrine control (Kim and Joo, 2009) and recurrent vomiting may disrupt the sodium and water balance in TBM patients (Misra et al., 2016).

Previous studies (Murthy, 2005; Roca et al., 2008; Figaji and Fieggen, 2010; Thao et al., 2017) have reported a relation between hyponatremia and mortality of TBM patients. However, some of those studies were small, and reported associations with mortality do not prove that hyponatremia actually contributes to poor patient outcome, as plasma sodium often correlates with disease severity or other markers that may affect patients’ prognosis.

In the present study we examined the clinical relevance of hyponatremia in a large prospective cohort of TBM patients from Indonesia, combining a standard statistical approach with a novel method for causal modeling that allows for both observed variables and latent constructs. Previous prognostic models in infectious diseases have mostly been based on observed variables only and without taking into account the instability of model estimation. Instability occurs when a slight change in the data results in a completely different model estimate. Modeling relations among latent constructs can be useful in gaining insight in the underlying mechanisms. We therefore applied a recently developed causal modeling method that searches for robust causal model structures and allows for both observed and latent variables, S3C-Latent (Rahmadi et al., 2018a), an extension of stable specification search for cross-sectional data (S3C; Rahmadi et al., 2017). S3C and S3C-Latent conduct many model refinements through an iterative procedure, i.e., current model structures based on the previously selected models, on repeated sampling, seeking causal structures that are stable and simple (parsimonious) across different model complexities. In the present study, we applied S3C-Latent to baseline patient characteristics and extend the idea of S3C to model survival stability (called stable Cox model) taking censored data into account, similar to some previous studies (Fan and Li, 2002; Walschaerts et al., 2012), resulting in a causal survival model. Thus, this study has two objectives: to understand the causal relations between hyponatremia and disease characteristics at time of diagnosis; and to establish its relevance for patient mortality during follow-up.
8.2 Methods

8.2.1 Setting and patients

This prospective cohort study included patients > 14 years of age who presented with suspected TBM between October 2006 and November 2017 in the Hasan Sadikin Hospital, a top referral hospital for 43 million people living in the West Java province, Indonesia. Patients were clinically suspected of having TBM if they had subacute illness with headache, fever, or focal neurological symptoms, with or without signs of extracranial tuberculosis. The study was part of the project titled “Optimization of diagnosis and treatment of Meningitis”, approved by the Ethical Committee of Hasan Sadikin Hospital/Faculty of Medicine of Universitas Padjadjaran No.85/FKUP-RSHS/KEPK/Kep/EC/2006.

As per routine care, each patient was examined through history taking, physical and neurological examination, blood and cerebrospinal fluid (CSF) examination, and chest radiography. Routine HIV testing has been implemented since 2009, and retrospective HIV testing has been administered anonymously for those patients who were admitted before 2009 or died before consent was obtained. A subset of patients in this study was included in randomized clinical trials evaluating intensified antibiotic treatment, for which separate ethical approval was obtained (Ruslami et al., 2013; Yunivita et al., 2016).

8.2.2 Measurements

Microbiological diagnosis was done using Ziehl-Nielsen and Gram-staining for microscopy, and solid Ogawa and liquid commercial culture (MODS) Mycobacterium tuberculosis culture. GeneXpert® MTB/RIF has been used since 2015 (Chaidir et al., 2018). India ink staining and Toxoplasma gondii CSF real-time PCR and serological testing were performed for HIV-infected patients (Ganiem et al., 2013). Roche Diagnostics AVL 9180 Electrolyte Analyzers was used for blood sodium assessment (AVL 1996). Baseline blood sodium concentration was used to classify patients as having severe hyponatremia (plasma sodium < 120 mEq/L), moderate hyponatremia (120–130 mEq/L), and mild hyponatremia or normal sodium levels (131–145 mEq/L; Waikar et al., 2009).

Patients were followed prospectively for at least one year. Field physicians or nurses made telephone calls and a social worker conducted home visits for patients not returning after hospital discharge. Death after hospital discharge was assessed by interview of family members and retrieval of patients’ death certificates from local authorities.

8.2.3 Data analysis and statistics

TBM was classified as definite if either CSF microscopy for acid-fast bacilli, Mycobacterium tuberculosis culture, or PCR results were positive. Based on prior
evaluation of CSF characteristics of definite and clinically suspected cases in this cohort, patients with CSF to blood glucose ratio of $< 0.5$ combined with a CSF cell count $\geq 5$ cells/$\mu l$ were defined as probable TBM (van Laarhoven et al., 2017).

Patients’ characteristics were presented as median values (with interquartile range) or proportions and were compared in terms of normal or mild hyponatremia, moderate, and severe hyponatremia groups, both in HIV-positive, and HIV-negative patients. Patients with hypernatremia (plasma sodium $> 145$ mEq/L) were excluded from analysis. Candidate predictors for the model were selected based on existing knowledge (van Laarhoven et al., 2017), clinical judgement and data completeness. The number of predictors was further limited to represent types of predictors: observed predictors comprised HIV, age, CSF culture, and blood sodium, while latent predictors consist of “Tissue damage” that is indicated by Glasgow Coma Scale (GCS), motor deficits and cranial nerve palsy, and “Inflammation” that is indicated by body temperature, blood neutrophils, CSF mononuclear cells, CSF protein, and CSF to blood glucose ratio. Univariate and multivariate Cox regression analysis was performed, resulting in survival probabilities and hazard ratios (HRs). Kaplan-Meier curves were used to illustrate survival rate over time based on the hyponatremia status. This was complemented by analysis in a causal survival model, which was done through two modeling phases: the first using S3C-Latent to identify causal relations between possible predictors of mortality; the second using the stable Cox model to seek stable predictors on patient survival. For S3C-Latent, we included both the observed and latent variables. We drew 100 subsamples using multiple imputation for 4.28% missing data using expectation maximization (EM; Honaker et al., 2011). For the stable Cox model, we enumerated all possible Cox models, drew 100 subsamples, and fit those models to each data subset. This procedure was repeated 50 times to get even more stability confidence. For more details on the S3C-Latent and the stable Cox model procedure, readers are referred to Chapter 5 and Appendix 8.B, respectively. All analyses were performed with RStudio in R 3.4.2., using packages for S3C-Latent$^8$ (development version), ggplot2, reshape2, dplyr, tableone, survminer, survival, KMsurv, rms, and Hmisc.

8.3 Result

8.3.1 Baseline characteristics

Out of 1261 patients with suspected TBM, 821 had been diagnosed with definite or probable TBM. For further analysis, we included 678 subjects with complete data on sex, age, HIV status, blood sodium and one-year survival, and excluded those with hypernatremia (Figure 8.1). A total of 324 patients (47.8%) had a positive CSF Mycobacterium tuberculosis culture, and 174 patients (25.7%) had TBM confirmed through in-house PCR or GeneXpert®. Patients admitted to the hospital were relatively young (median age $= 29$; interquartile range [IQR] 23–37), 109 were

$^8$ a development version is available at https://github.com/rahmarid/S3C-Latent
Chapter 8. Clinical significance of hyponatremia in TBM

Figure 8.1: Abbreviations: HIV, human immunodeficiency virus; TBM, tuberculous meningitis. Hypernatremia define as blood sodium level $>145$ mEq/L. TBM was classified as “definite” (microbiologically proven) if either CSF microscopy for acid-fast bacilli, Mycobacterium tuberculosis culture, or PCR results were positive. Based on prior evaluation of CSF characteristics of definite and clinically suspected cases in Bandung cohort, patients were classified as having “probable TBM” if they had a CSF/blood glucose ratio $<0.5$ combined with CSF cells count $\geq 5$ cells/µl.

HIV-infected (16.1%), and most presented with British Medical Research Council (BMRC) grade II TBM (67.3%). Hyponatremia was found in 82.9% of patients, and less prevalent among HIV-infected patients (odds ratio 0.53; 95% confidence interval [CI], 0.33-0.86; $p$-value = 0.01; Figure 8.2). Severe hyponatremia was found in 17.8% of patients and was associated with concomitant pulmonary TB, a lower GCS, more pronounced routine blood and CSF abnormalities, and positive CSF Mycobacterium tuberculosis culture reflecting a higher bacterial load (Table 8.1).

8.3.2 Patient survival

A total of 471 patients had follow up data up to 12 months. Lost to follow up was 15.5% in the first 6 months, and 6.6% between 6 and 12 months. Com-
Figure 8.2: Plasma sodium among 581 HIV-negative (top) and 110 HIV-positive patients (bottom). Dashed vertical lines separate patients with severe (blood sodium < 120 mEq/L, n = 237), moderate (blood sodium 121–130 mEq/L, n = 320), and mild or no hyponatremia (blood sodium > 130–145 mEq/L, n = 121), separated by the dashed lines. Hyponatremia is defined by a blood sodium < 135 mEq/L (solid vertical line).

Compared to those who were followed-up for 6 and 12 months, those who were lost less frequently had severe hyponatremia (19.3% vs 14.5%; p-value = 0.281). One-year mortality was 44.6% (95% CI, 40.5%–48.5%) overall, and 61.0% (95% CI, 49.8%–69.7%) in the HIV-infected subgroup. BMRC TBM grade, motor deficits, fever, lower GCS, CSF polymorphonuclear cells, anaemia, and thrombocytosis predicted mortality in univariate analysis (see Table 8.3 in Appendix 8.A). Among HIV-uninfected patients one-year mortality was 34.5% (95% CI, 26.8%–41.5%) for those with a normal plasma sodium or mild hyponatremia (n = 186), 41.9% (95% CI, 35.3%–47.9%) among those with moderate hyponatremia (n = 273) and 51.1% (95% CI, 40.2%–60.1%) among those with severe hyponatremia (n = 110; see Figure 8.3). Patients with severe hyponatremia have an almost two times higher risk of dying compared to those with normal blood sodium or mild hyponatremia (HR 1.68; 95% CI, 1.16–2.44, p-value = 0.006). Among HIV-infected patients one-year mortality was 57.5% (95% CI, 40.6%–69.6%) for patients with no/mild hyponatremia (n = 51), 65.3% (95% CI, 46.3%–77.5%) for those with moderate hyponatremia
## Chapter 8. Clinical significance of hyponatremia in TBM

### Table 8.1: Patient characteristics according to level of hyponatremia.

<table>
<thead>
<tr>
<th>Hyponatremia (mEq/L)</th>
<th>Severe (&lt;121)</th>
<th>Moderate 121–130</th>
<th>Mild/No 131–145</th>
<th>p-value&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>n(%)</td>
<td>121 (17.5%)</td>
<td>320 (46.3%)</td>
<td>237 (36.2%)</td>
<td></td>
</tr>
<tr>
<td>Gender, Male</td>
<td>85 (70.2%)</td>
<td>196 (61.3%)</td>
<td>134 (56.5%)</td>
<td>0.042</td>
</tr>
<tr>
<td>Age, years&lt;sup&gt;a&lt;/sup&gt;</td>
<td>28 [22–33]</td>
<td>30 [23–37]</td>
<td>30 [24–40]</td>
<td>0.026</td>
</tr>
<tr>
<td>HIV (+)</td>
<td>11 (9.1%)</td>
<td>47 (14.7%)</td>
<td>51 (21.5%)</td>
<td>0.007</td>
</tr>
<tr>
<td>Grade</td>
<td></td>
<td></td>
<td></td>
<td>0.070</td>
</tr>
<tr>
<td>• Grade I</td>
<td>6 (5.3%)</td>
<td>31 (10.4%)</td>
<td>30 (14.6%)</td>
<td></td>
</tr>
<tr>
<td>• Grade II</td>
<td>86 (75.4%)</td>
<td>227 (76.2%)</td>
<td>143 (69.8%)</td>
<td></td>
</tr>
<tr>
<td>• Grade III</td>
<td>22 (19.3%)</td>
<td>40 (13.4%)</td>
<td>32 (15.6%)</td>
<td></td>
</tr>
<tr>
<td>Headache</td>
<td>99 (91.7%)</td>
<td>272 (94.1%)</td>
<td>203 (93.5%)</td>
<td>0.678</td>
</tr>
<tr>
<td>Neck stiffness</td>
<td>113 (95.8%)</td>
<td>288 (92.0%)</td>
<td>207 (92.8%)</td>
<td>0.396</td>
</tr>
<tr>
<td>Seizures</td>
<td>10 (8.7%)</td>
<td>15 (5.1%)</td>
<td>41 (18.5%)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Motor deficits</td>
<td>63 (53.8%)</td>
<td>157 (53.6%)</td>
<td>107 (48.4%)</td>
<td>0.502</td>
</tr>
<tr>
<td>Cranial Nerve Palsy</td>
<td>72 (62.1%)</td>
<td>189 (62.0%)</td>
<td>126 (54.8%)</td>
<td>0.201</td>
</tr>
<tr>
<td>Chest X-ray, TB</td>
<td>96 (83.5%)</td>
<td>225 (73.5%)</td>
<td>140 (61.4%)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Temperature (°C)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>37.6 [37–38]</td>
<td>37.6 [36.8–38.1]</td>
<td>37.3 [36.7–38]</td>
<td>0.119</td>
</tr>
<tr>
<td>Fever</td>
<td>60 (52.2%)</td>
<td>158 (52.0%)</td>
<td>100 (42.6%)</td>
<td>0.067</td>
</tr>
<tr>
<td>Positive M. tuberculosis CSF culture</td>
<td>83 (70.3%)</td>
<td>170 (54.5%)</td>
<td>71 (31.6%)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>• In HIV (+)</td>
<td>5 (45.5%)</td>
<td>19 (43.2%)</td>
<td>10 (21.3%)</td>
<td>0.057</td>
</tr>
<tr>
<td>• In HIV (−), n=58</td>
<td>78 (72.9%)</td>
<td>151 (56.3%)</td>
<td>61 (34.3%)</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

### Cerebrospinal fluid (CSF)

<table>
<thead>
<tr>
<th></th>
<th>Severe (&lt;121)</th>
<th>Moderate 121–130</th>
<th>Mild/No 131–145</th>
<th>p-value&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>• In HIV (+)</td>
<td>43 [7–447]</td>
<td>75 [9–180]</td>
<td>26 [6–89]</td>
<td>0.148</td>
</tr>
<tr>
<td>• In HIV (−)</td>
<td>133 [42–308]</td>
<td>150 [44–351]</td>
<td>118 [15–324]</td>
<td>0.267</td>
</tr>
<tr>
<td>Polymorphonuclear cells, cells/µL (PMN)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>32.7 [9.4–117.3]</td>
<td>32.5 [5.1–116.3]</td>
<td>10.5 [1.3–64.4]</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>• In HIV (+)</td>
<td>11.3 [3.8–95.1]</td>
<td>10.4 [1–82.8]</td>
<td>2 [0.8–15]</td>
<td>0.053</td>
</tr>
<tr>
<td>• In HIV (−)</td>
<td>38.7 [12.4–116.8]</td>
<td>34.2 [6.8–123.8]</td>
<td>15.9 [2.9–93.9]</td>
<td>0.010</td>
</tr>
<tr>
<td>CSF protein, mg/dL&lt;sup&gt;a&lt;/sup&gt;</td>
<td>216.5 [101.8–401.5]</td>
<td>166 [76.8–329]</td>
<td>124 [62–229.2]</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>• In HIV (+)</td>
<td>44 [36–170]</td>
<td>142.9 [73.8–212.2]</td>
<td>120 [65–165]</td>
<td>0.290</td>
</tr>
<tr>
<td>• In HIV (−)</td>
<td>237 [132–420]</td>
<td>172 [83.5–337]</td>
<td>129 [59.5–262]</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>CSF/blood glucose ratio&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.10 [0.10–0.30]</td>
<td>0.20 [0.10–0.30]</td>
<td>0.30 [0.20–0.50]</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>• In HIV (+)</td>
<td>0.20 [0.10–0.40]</td>
<td>0.30 [0.10–0.40]</td>
<td>0.40 [0.30–0.50]</td>
<td>0.003</td>
</tr>
<tr>
<td>• In HIV (−)</td>
<td>0.10 [0.10–0.20]</td>
<td>0.20 [0.10–0.30]</td>
<td>0.30 [0.20–0.50]</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

### Blood

<table>
<thead>
<tr>
<th></th>
<th>Severe (&lt;121)</th>
<th>Moderate 121–130</th>
<th>Mild/No 131–145</th>
<th>p-value&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anemia</td>
<td>5 (4.2%)</td>
<td>20 (6.3%)</td>
<td>11 (4.7%)</td>
<td>0.580</td>
</tr>
<tr>
<td>Hemoglobin, mg/dL</td>
<td>11.6 [9.9–13.0]</td>
<td>11.8 [10.1–13.2]</td>
<td>12.6 [10.9–13.9]</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Leukocytes, cells/µL&lt;sup&gt;a&lt;/sup&gt;</td>
<td>10.9 [7.7–13.8]</td>
<td>9.7 [7.3–13.2]</td>
<td>10 [6.5–13.5]</td>
<td>0.265</td>
</tr>
<tr>
<td>Polymorphonuclear cells, cells/µL (PMN)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>9.6 [7–12.6]</td>
<td>8.1 [5.8–11.3]</td>
<td>8.2 [5–11.3]</td>
<td>0.038</td>
</tr>
<tr>
<td>Monocytes, cells/µL&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.5 [0.2–0.9]</td>
<td>0.4 [0.2–0.9]</td>
<td>0.5 [0.3–0.7]</td>
<td>0.603</td>
</tr>
<tr>
<td>Lymphocytes, cells/µL&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.7 [0.3–1]</td>
<td>0.8 [0.6–1.3]</td>
<td>1.1 [0.7–1.9]</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Thrombocytes, ×10&lt;sup&gt;9&lt;/sup&gt;/L&lt;sup&gt;a&lt;/sup&gt;</td>
<td>294 [216.8–362.2]</td>
<td>290 [194.5–382.5]</td>
<td>281 [212–357.5]</td>
<td>0.797</td>
</tr>
</tbody>
</table>

Abbreviations: CSF, cerebrospinal fluid; GCS, Glasgow Coma Scale; HIV, human immunodeficiency virus; TBM, tuberculous meningitis; IQR, interquartile range; MCR, British Medical Research. Anemia is blood hemoglobin < 8 mg/L (de Benoist et al., 2008); Fever is body temperature > 38.5°C.

<sup>a</sup>Variables are represented using median and IQR.

<sup>b</sup>Values of < 0.05 are considered statistically significant.
Figure 8.3: Patient survival according to plasma sodium, HIV-negative patients (n=569). One-year mortality of HIV-uninfected patients with mild (n/N=58/168 (31.1%)), moderate (n/N=106/273 (38.8%)) and severe hyponatremia (n/N=53/110 (48.2%), p-value = 0.014).

(n = 47), and 57.6% (95% CI, 12.8%–79.4%) for those with severe hyponatremia (n = 11), and differences among groups were not statistically significant. In multivariate Cox regression, HIV, age, sex, motor deficits, fever, and GCS predicted mortality, while hyponatremia showed no significant association (see Table 8.2).
Table 8.2: Multivariate Cox regression for prediction of one-year mortality.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Blood 365-Day mortality</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>HR</td>
</tr>
<tr>
<td>Hyponatremia</td>
<td></td>
</tr>
<tr>
<td>• Moderate vs. mild/normal</td>
<td>1.10</td>
</tr>
<tr>
<td>• Severe vs. mild/normal</td>
<td>1.06</td>
</tr>
<tr>
<td>HIV (+)</td>
<td>1.57</td>
</tr>
<tr>
<td>Age (per 10-year increase)</td>
<td>1.14</td>
</tr>
<tr>
<td>BMRC Grade</td>
<td></td>
</tr>
<tr>
<td>• Grade II vs. Grade I</td>
<td>1.29</td>
</tr>
<tr>
<td>• Grade III vs. Grade I</td>
<td>1.64</td>
</tr>
<tr>
<td>Gender (male)</td>
<td>1.46</td>
</tr>
<tr>
<td>Motor deficits (present)</td>
<td>1.71</td>
</tr>
<tr>
<td>Temperature (per 1°C increase)</td>
<td>1.34</td>
</tr>
<tr>
<td>GCS (per 1-point increase)</td>
<td>0.90</td>
</tr>
<tr>
<td>CSF PMN (per 10% increase)</td>
<td>1.00</td>
</tr>
<tr>
<td>Blood PMN (per 10% increase)</td>
<td>1.02</td>
</tr>
<tr>
<td>Anemia (present)</td>
<td>1.69</td>
</tr>
<tr>
<td>Blood thrombocytes (per 1 × 10⁹ cells/L increase)</td>
<td>0.99</td>
</tr>
<tr>
<td>Blood sodium (per 1-unit increase)</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Abbreviations: HIV, human immunodeficiency virus; GCS, Glasgow coma scale.
Multivariate Cox regression for survival including variables with a p-value < 0.05 in univariate analysis for 356-days mortality in HIV-negative TBM patients. Anemia is blood hemoglobin < 8 mg/dL (de Benoist et al., 2008). Fever is body temperature > 38.5°C.

<sup>a</sup>Values of < 0.05 are considered statistically significant.

### 8.3.3 Causal survival model

We next examined what factors contributed to mortality, using a causal survival model. We chose four observed and two latent variables of interest for the causal modeling analysis, based on the univariate Cox regression analysis in combination with clinical judgment. We excluded BMRC TBM grade since it is a representation of GCS and motor deficits. Figure 8.4 shows the stability graphs resulting from applying S3C-Latent to the baseline data. Figure 8.5 displays the stability graph resulting from applying the stable Cox model procedure to the censored data. Combining results from both phases, Figure 8.6 visualizes the relevant model structures, resulting in the causal survival model. According to the model, a positive CSF culture contributes to CSF inflammation and to development of (severe) hyponatremia. CSF culture itself seems to be affected by HIV status. In addition, proxy variables for tissue damage and blood sodium level show an association with CSF inflammation, but the causal directions cannot be determined from the data. With respect to mortality, HIV was the most stable predictor in this model, leading to a more than two-fold higher risk of death. According to this model, hyponatremia
does not seem to contribute to patient mortality.

8.4 Discussion

We examined the clinical relevance of hyponatremia in a prospective cohort of 678 TBM patients in Indonesia. Moderate and severe hyponatremia were very common, especially among HIV-negative patients. Severe hyponatremia was associated with neurological complications and CSF inflammatory markers at time of presentation, and higher mortality during follow-up. However, hyponatremia was not associated with increased mortality in multivariate regression analysis and did not appear to contribute to increased mortality in causal survival modeling.

Hyponatremia is commonly found among patients with central nervous system disorders (Palmer, 2003; Tisdall et al., 2006; Kim and Joo, 2009; Jaffri and Ahsan, 2015), infectious diseases (Liamis et al., 2011), and those who are critically ill (Tisdall et al., 2006; Patel and Balk, 2007). As TBM combines these three conditions, it is not surprising that TBM patients often suffer from hyponatremia. Compared to previous studies hyponatremia seems more common and more severe in our setting (Roca et al., 2008; Anderson et al., 2010; Jaffri and Ahsan, 2015; Misra et al., 2016), probably because TBM patients in our setting present with very advanced disease (grade II or III BMRC). Hyponatremia in TBM is commonly attributed to the syndrome of inappropriate antidiuretic hormone (SIADH; Palmer, 2003; Liamis et al., 2011). Brain inflammation in TBM may increase intracranial pressure leading to excessive release of antidiuretic hormone (ADH; Kröll et al., 1992); therefore, it increases the renal water reabsorption and results in expansion of extracellular fluid volume (Palmer, 2003). Less commonly, hyponatremia in TBM is related to cerebral salt wasting (Liamis et al., 2011; Jaffri and Ahsan, 2015), recurrent vomiting, and nutritional deficiency (Misra et al., 2016). Of note, a definite cause of hyponatremia is notoriously hard to establish, especially in the absence of urine sodium and osmolality measurements as is often the case in settings were tuberculous meningitis is prevalent.

In our study, the degree of hyponatremia was associated with disease severity and with blood and CSF inflammation. Patients with severe hyponatremia more often had concomitant pulmonary TB, culture confirmed TBM, a lower GCS, and more pronounced inflammation in blood and CSF, in line with a previous a study from India (Misra et al., 2016). Severe hyponatremia was relatively uncommon in patients presenting with seizures in our study. Typically, hyponatremia in TBM has a “chronic” nature (i.e., exists for more than 48 hours; Patel and Balk, 2007), while seizure usually rises in acute hyponatremia (Podestà et al., 2015; Sterns, 2015).

Hyponatremia predicted mortality in univariate analysis, albeit only among HIV-negative patients. However, hyponatremia was not associated with mortality in multivariate analysis nor in causal survival modeling analysis. Some previous studies reported that hyponatremia in TBM increases the risk of death (Roca and Bahamonde, 2006; Podestà et al., 2015; Sterns, 2015), while others reported other-
Figure 8.4: The edge (top panel) and causal path (bottom panel) stability graphs from baseline data. The $x$-axis indicates model complexity, e.g., 5 indicates a group of models that have five relations. The $y$-axis indicates the selection probability, i.e., the number of occurrences of a relation between a pair of variables in all optimal models at a particular model complexity. Thus, a line indicates the stability of a relation across different model complexities. The edge stability considers any relation between pairs of variables, regardless of the direction. The causal path stability accounts for causal relations from one variable to another variable with any length, e.g., with intermediate variable(s) in between a cause and an effect variable. The threshold of selection probability, $\pi_{\text{sel}}$, is set to 0.6 (Meinshausen and Bühlmann, 2010), and the threshold of model complexity, $\pi_{\text{bic}}$, is chosen at a model complexity at which the minimum median of BICs is found (Rahmadi et al., 2017). In this case, we found $\pi_{\text{bic}}$ at model complexity 8. Any model structure that has selection probability equal to or greater than $\pi_{\text{sel}}$ and model complexity smaller or equal to $\pi_{\text{bic}}$, i.e., passing through the top-left box, is considered stable and parsimonious, and called a relevant model structure. Figure 8.6 is a visualization of the model structure with steps described in Section 3.2.1.
Figure 8.5: Stability graph of Cox model based on the six selected variables. Observed variables comprise HIV, age, cerebrospinal fluid (CSF) M. tuberculosis culture result, and blood sodium. Latent variables consist of (1) Tissue damage that is measured through Glasgow coma scale (GCS), motor deficits, and cranial nerve palsy; (2) Inflammation that is indicated by fever, blood neutrophils, CSF mononuclear cells, CSF protein and CSF/blood glucose ratio. The $x$-axis indicates the Cox model complexity, e.g., 3 refers to all enumerated Cox models with three predictors. The $y$-axis indicates the selection probability, i.e., the number of occurrences of a predictor in all Cox models at a particular model complexity. The threshold of selection probability, $\pi_{\text{sel}}$, is set to 0.6 (Meinshausen and Bühlmann, 2010), and the threshold of model complexity, $\pi_{\text{bic}}$, is chosen at a model complexity at which the minimum median of BICs is found (Rahmadi et al., 2017). Any Cox model structure that has selection probability equal to or greater than $\pi_{\text{sel}}$ and model complexity smaller or equal to $\pi_{\text{bic}}$, i.e., passing the top-left box, is considered stable and parsimonious, and called relevant. In this graph, HIV is the relevant structure, making it the predictor that contributes to mortality.
Figure 8.6: The causal survival model consists of two parts. The first part represents causal relations between baseline predictors: HIV, age, CFS Culture (Culture), blood sodium, tissue damage (Tissue), and inflammation (Inflammation). The last two variables are latent (drawn as circle nodes) and measured through observed indicators. Arrows indicate a causal relation and undirected edges exhibit a strong association where its causal direction cannot be determined from the data. An arrow is annotated with two numbers: a reliability score and a total causal effect (Maathuis et al., 2009). The reliability score is taken from the highest selection probability that a particular relation obtains in the relevant area of the stability graph (top-left box; Rahmadi et al., 2017). For example, 0.99/0.01 indicates a reliability score of 0.99 and a total causal effect estimate of 0.01. An undirected edge, in particular, is only annotated with a reliability score. According to this causal model, HIV infection increases the probability of a negative CSF culture. A positive CSF culture induces inflammation and leads to lower level of blood sodium. The second part represents causal relations from predictors to mortality. HIV was found to contribute to mortality. The arrow from HIV to Mortality is annotated with a reliability score of 0.95 and average of hazard rate of HIV found at $\pi_{bic}$ (see Figure 8.5). In addition, double-lines arrows from a latent variable to its indicator represent influence of the latent variable on the indicator (Silva et al., 2006).
wise (Misra et al., 2016). Most studies were smaller, however, and none employed causal modeling. Our findings are in line with the notion that hyponatremia in the context of infections may reflect severity of the underlying disease rather than that it directly contributes to mortality (Liamis et al., 2011; Podestà et al., 2015).

To examine what factors, including hyponatremia, may contribute to TBM-associated deaths we applied a novel causal modeling analysis. In this model, HIV infection appeared to lower the chance of finding a positive CSF Mycobacterium tuberculosis culture. We could bacteriologically confirm TBM in two thirds of HIV-uninfected patients, and approximately half of HIV-infected patients, in contrast to studies from Vietnam that found higher rates of bacteriological confirmation in HIV-infected patients (Thwaites et al., 2004, 2005). Our model also revealed that a positive CSF culture was associated with more CSF inflammation and more hyponatremia. Patients with culture-confirmed TBM in our cohort had a more severe disease, with more concomitant pulmonary TB, lower GCS, a higher CSF protein level, lower CSF to blood glucose ratio, and higher numbers of CSF leukocytes, CSF neutrophils, and blood neutrophils (see Table 8.4 in Appendix 8.A). This finding is in agreement with previous studies showing that bacterial positivity correlates with several factors such as higher number of CSF cells (Thwaites et al., 2004; Jha et al., 2015) and lower CSF to blood glucose ratio (Thwaites et al., 2004). Our findings underline the importance of effective anti-mycobacterial treatment, with a role for high-dose rifampicin (Ruslami et al., 2013; Yunivita et al., 2016). Supporting the findings from the multivariate regression analysis, our causal survival model identified that HIV contributed to increased mortality, while hyponatremia did not.

Our study has several limitations. We only used the baseline blood sodium data and did not study possible changes in blood sodium over time. In addition, we could not differentiate SIADH or cerebral salt wasting as a cause of hyponatremia. Our survival analysis was compromised by loss to follow-up of one-fifth of patients up to one year. Finally, in terms of causal modeling analysis, we excluded the possibility of reciprocal causal relations. Handling those requires more advanced approaches that we leave for later study.

Still, taken together, this study adds to our understanding of the role of hyponatremia in TBM. Current practice includes water restriction based on the presumptive diagnosis of SIADH. This might be detrimental, as TBM patients, especially those with lowered consciousness and fever, are usually dehydrated. More data regarding the cause of hyponatremia and its course over time during treatment are needed. However, the causal modeling of the available data implies that aggressive correction of hyponatremia is unlikely to improve patients’ outcome. Rather, a comprehensive approach, including optimal anti-mycobacterial treatment, supportive care, and control of inflammation, e.g., fever may help to improve prognosis of TBM.
### Appendix 8.A Supplementary tables

#### Table 8.3: Univariate Cox regression for prediction of 365-day mortality among tuberculous meningitis.

<table>
<thead>
<tr>
<th>Variable</th>
<th>n</th>
<th>number of events</th>
<th>HR</th>
<th>95% CI</th>
<th>p-value&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>HIV (+)</td>
<td>678</td>
<td>280</td>
<td>1.94</td>
<td>(1.46–2.57)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Gender (Male)</td>
<td>678</td>
<td>280</td>
<td>1.37</td>
<td>(1.07–1.76)</td>
<td>0.012</td>
</tr>
<tr>
<td>Age (per 10-year increase)</td>
<td>678</td>
<td>280</td>
<td>1.14</td>
<td>(1.04–1.26)</td>
<td>0.006</td>
</tr>
<tr>
<td>TBM grade</td>
<td>617</td>
<td>258</td>
<td></td>
<td></td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>• Grade II vs. Grade I</td>
<td></td>
<td></td>
<td>1.70</td>
<td>(1.05–2.77)</td>
<td>0.031</td>
</tr>
<tr>
<td>• Grade III vs. Grade I</td>
<td></td>
<td></td>
<td>4.27</td>
<td>(2.52–7.21)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Headache, present</td>
<td>614</td>
<td>248</td>
<td>1.02</td>
<td>(0.61–1.72)</td>
<td>0.932</td>
</tr>
<tr>
<td>Neck stiffness, present</td>
<td>654</td>
<td>270</td>
<td>1.36</td>
<td>(0.82–2.25)</td>
<td>0.237</td>
</tr>
<tr>
<td>Seizures, present</td>
<td>631</td>
<td>254</td>
<td>1.09</td>
<td>(0.73–1.62)</td>
<td>0.686</td>
</tr>
<tr>
<td>Motor deficits, present</td>
<td>634</td>
<td>260</td>
<td>1.69</td>
<td>(1.31–2.17)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Cranial nerve palsy, present</td>
<td>651</td>
<td>265</td>
<td>1.27</td>
<td>(0.99–1.64)</td>
<td>0.058</td>
</tr>
<tr>
<td>Chest X-ray, abnormal</td>
<td>649</td>
<td>265</td>
<td>1.03</td>
<td>(0.79–1.34)</td>
<td>0.850</td>
</tr>
<tr>
<td>Mycobacteriology confirmation test (+)</td>
<td>678</td>
<td>280</td>
<td>1.03</td>
<td>(0.8–1.32)</td>
<td>0.816</td>
</tr>
<tr>
<td>CSF culture (+)</td>
<td>655</td>
<td>267</td>
<td>1.14</td>
<td>(0.9–1.45)</td>
<td>0.276</td>
</tr>
<tr>
<td>Temperature (per 1°C)</td>
<td>654</td>
<td>267</td>
<td>1.33</td>
<td>(1.17–1.52)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>GCS (Per 1-unit increase)</td>
<td>622</td>
<td>256</td>
<td>0.81</td>
<td>(0.77–0.85)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>CSF leukocytes, log&lt;sub&gt;10&lt;/sub&gt; (cells/µl)</td>
<td>671</td>
<td>276</td>
<td>0.90</td>
<td>(0.77–1.04)</td>
<td>0.165</td>
</tr>
<tr>
<td>• CSF PMN (10%)</td>
<td>645</td>
<td>260</td>
<td>1.01</td>
<td>(1–1.01)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>• CSF MN (10%)</td>
<td>645</td>
<td>260</td>
<td>0.99</td>
<td>(0.99–1)</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>CSF protein, mg/L (log&lt;sub&gt;10&lt;/sub&gt;)</td>
<td>668</td>
<td>272</td>
<td>1.22</td>
<td>(0.99–1.51)</td>
<td>0.059</td>
</tr>
<tr>
<td>CSF/Blood glucose ratio (per unit increase)</td>
<td>666</td>
<td>272</td>
<td>0.62</td>
<td>(0.31–1.25)</td>
<td>0.184</td>
</tr>
<tr>
<td>Anemia, mg/L</td>
<td>675</td>
<td>278</td>
<td>1.75</td>
<td>(1.11–2.75)</td>
<td>0.017</td>
</tr>
<tr>
<td>Blood leukocytes (cells/µl)</td>
<td>670</td>
<td>275</td>
<td>1.01</td>
<td>(0.99–1.04)</td>
<td>0.212</td>
</tr>
<tr>
<td>• Blood neutrophils (cells/µl)</td>
<td>488</td>
<td>202</td>
<td>1.03</td>
<td>(1–1.06)</td>
<td>0.020</td>
</tr>
<tr>
<td>• Blood lymphocytes (cells/µl)</td>
<td>490</td>
<td>202</td>
<td>0.88</td>
<td>(0.72–1.07)</td>
<td>0.190</td>
</tr>
<tr>
<td>Thrombocytes (cells/µl)</td>
<td>674</td>
<td>278</td>
<td>1.00</td>
<td>(1–1)</td>
<td>0.030</td>
</tr>
<tr>
<td>Hyponatremia (mmol/L)</td>
<td>678</td>
<td>280</td>
<td>1.20</td>
<td>(0.91–1.57)</td>
<td>0.190</td>
</tr>
<tr>
<td>• Moderate vs. Normal/mild (mmol/L)</td>
<td></td>
<td></td>
<td>1.39</td>
<td>(1–1.94)</td>
<td>0.049</td>
</tr>
<tr>
<td>• Severe vs. Normal/mild (mmol/L)</td>
<td></td>
<td></td>
<td>1.68</td>
<td>(1.16–2.44)</td>
<td>0.006</td>
</tr>
<tr>
<td>• In HIV-uninfected (mmol/L)</td>
<td></td>
<td></td>
<td>0.92</td>
<td>(0.38–2.23)</td>
<td>0.856</td>
</tr>
<tr>
<td>Blood sodium (per 1 mEq/L increase)</td>
<td>678</td>
<td>280</td>
<td>0.99</td>
<td>(0.97–1)</td>
<td>0.072</td>
</tr>
</tbody>
</table>

Abbreviations: HIV, human immunodeficiency virus; GCS, Glasgow coma scale. Anemia is blood hemoglobin < 8 mg/dL (de Benoist et al., 2008). Fever is body temperature > 38.5°C.  
<sup>a</sup>Values of < 0.05 are considered statistically significant.
Table 8.4: Baseline characteristics according to CSF Mycobacterium tuberculosis culture result.

<table>
<thead>
<tr>
<th>Variables, n(%)</th>
<th>Culture</th>
<th>p-value&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Negative (n = 331)</td>
<td>Positive (n = 324)</td>
</tr>
<tr>
<td>Gender, Male</td>
<td>198 (59.8%)</td>
<td>202 (62.3%)</td>
</tr>
<tr>
<td>Age, years&lt;sup&gt;a&lt;/sup&gt;</td>
<td>31 [24–40]</td>
<td>28 [22–35]</td>
</tr>
<tr>
<td>HIV (+)</td>
<td></td>
<td>34 (10.5%)</td>
</tr>
<tr>
<td>Grade</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Grade I</td>
<td>31 (10.9%)</td>
<td>35 (11.1%)</td>
</tr>
<tr>
<td>• Grade II</td>
<td>217 (76.4%)</td>
<td>226 (72.0%)</td>
</tr>
<tr>
<td>• Grade III</td>
<td>36 (12.7%)</td>
<td>53 (16.9%)</td>
</tr>
<tr>
<td>Headache</td>
<td>278 (93.6%)</td>
<td>282 (93.1%)</td>
</tr>
<tr>
<td>Neck stiffness</td>
<td>285 (90.8)</td>
<td>303 (94.7%)</td>
</tr>
<tr>
<td>Seizures</td>
<td>37 (11.9%)</td>
<td>23 (7.7%)</td>
</tr>
<tr>
<td>Motor deficits</td>
<td>156 (50.3%)</td>
<td>159 (52.5%)</td>
</tr>
<tr>
<td>Cranial Nerve Palsy</td>
<td>175 (55.2%)</td>
<td>200 (64.1%)</td>
</tr>
<tr>
<td>Chest X-ray, TB</td>
<td>204 (64.8)</td>
<td>241 (77.5%)</td>
</tr>
<tr>
<td>Temperature, °C&lt;sup&gt;a&lt;/sup&gt;</td>
<td>37.3 [36.7–38]</td>
<td>37.6 [36.8–38.1]</td>
</tr>
<tr>
<td>Fever</td>
<td>100 (42.6%)</td>
<td>158 (52.0%)</td>
</tr>
<tr>
<td>GCS&lt;sup&gt;a&lt;/sup&gt;</td>
<td>14 [13–15]</td>
<td>13 [11–14]</td>
</tr>
<tr>
<td>Leukocytes, cells/µL&lt;sup&gt;a&lt;/sup&gt;</td>
<td>72.5 [10–208.2]</td>
<td>182.5 [48–348.8]</td>
</tr>
<tr>
<td>• Polymorphonuclear, cells/µL (PMN)&lt;sup&gt;a&lt;/sup&gt;</td>
<td>10.3 [1.8–55.0]</td>
<td>51.7 [10.8–135.8]</td>
</tr>
<tr>
<td>CSF protein, mg/dL&lt;sup&gt;a&lt;/sup&gt;</td>
<td>131.5 [56–275.8]</td>
<td>189 [100–336]</td>
</tr>
<tr>
<td>CSF/Blood glucose ratio&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.30 [0.20–0.50]</td>
<td>0.20 [0.10–0.20]</td>
</tr>
<tr>
<td>Anemia</td>
<td>23 (6.9%)</td>
<td>10 (3.1%)</td>
</tr>
<tr>
<td>Hemoglobin, mg/dL&lt;sup&gt;a&lt;/sup&gt;</td>
<td>12.1 [10.2–13.5]</td>
<td>12 [10.2–13.3]</td>
</tr>
<tr>
<td>Leukocyte, cells/µL&lt;sup&gt;a&lt;/sup&gt;</td>
<td>9.5 [6.3–13.3]</td>
<td>10.4 [7.7–13.5]</td>
</tr>
<tr>
<td>• Neutrophils, cells/µL&lt;sup&gt;a&lt;/sup&gt;</td>
<td>7.9 [4.9–11]</td>
<td>8.9 [6.6–12.5]</td>
</tr>
<tr>
<td>• Lymphocytes, cells/µL&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1 [0.6–1.6]</td>
<td>0.8 [0.5–1.2]</td>
</tr>
<tr>
<td>• Monocytes, cells/µL&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.5 [0.2–0.8]</td>
<td>0.5 [0.2–0.9]</td>
</tr>
<tr>
<td>Thrombocyte&lt;sup&gt;a&lt;/sup&gt;</td>
<td>268 [192–357.2]</td>
<td>301 [217–372]</td>
</tr>
<tr>
<td>Sodium, mmol/L&lt;sup&gt;a&lt;/sup&gt;</td>
<td>130 [124–134]</td>
<td>124.5 [119–130]</td>
</tr>
<tr>
<td>Hyponatremia, mmol/L</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Mild/normal</td>
<td>154 (46.5%)</td>
<td>71 (21.9%)</td>
</tr>
<tr>
<td>• Moderate</td>
<td>142 (42.9%)</td>
<td>170 (52.5%)</td>
</tr>
<tr>
<td>• Severe</td>
<td>35 (10.6%)</td>
<td>83 (25.6%)</td>
</tr>
</tbody>
</table>

Abbreviations: CSF, cerebrospinal fluid; GCS, Glasgow Coma Scale; HIV, human immunodeficiency virus; TBM, tuberculous meningitis; IQR, interquartile range; MCR, British Medical Research; PMN, polymorphonuclear cells. Anemia is blood hemoglobin < 8 mg/L (de Benoist et al., 2008). Fever is body temperature > 38.5°C.

<sup>a</sup>Variables are represented using median and IQR.

<sup>b</sup>Values of < 0.05 are considered statistically significant.
Appendix 8.B  The stable Cox model

In this section, we shall describe the steps for realizing the stable Cox model. In essence, we extended the idea of stability selection (Meinshausen and Bühlmann, 2010) to obtain stable and parsimonious Cox model structures (predictors). The data used for the stable Cox model are those from selected variables: HIV, age, CFS culture, blood sodium, tissue damage, and inflammation. The last two variables are latent. We sampled from those latent variables with steps described in Appendix 5.B of Chapter 5. The steps of the stable Cox model are as follows.

1. From the data, subsample 100 subsets, about half the size of the data, without replacement.
2. Based on selected predictors, enumerate all possible Cox models.
3. For each data subset
   a. Fit all enumerated Cox models to the data subset.
   b. Group all fitted Cox models based on their model complexities, i.e., the number of predictors in a Cox model. In addition, the model complexity 0 is included to indicate initial model complexity where no predictor at all in the Cox model.
   c. Select the best Cox model for each model complexity, based on the log likelihood, i.e., choose the one with the highest log likelihood. From the best model, compute its Hazard rate and BIC score.
4. Step 3 results in the best 100 Cox models for each model complexity.
5. For each predictor, we compute its stability (frequency) in each model complexity relative to the corresponding best Cox 100 models. That is, we count how many times it occurs among (be a part of) the 100 Cox models of a particular model complexity.
6. For every model complexity, compute the median BIC scores from the best 100 Cox models, and set the threshold complexity to a model complexity at which the minimum median BIC is found (Rahmadi et al., 2017).

For an even more stable result, we repeated Steps 1–6 above 50 times and then computed the average stability. Figure 8.5 shows the average stability for every predictor over the whole range of model complexities. The threshold stability $\pi_{sel}$ is set to 0.6 (horizontal line; Meinshausen and Bühlmann, 2010) and the threshold complexity $\pi_{bic}$ is set to the median over 50 thresholds of complexity obtained from Step 3 (vertical line). Predictors with stability $\geq \pi_{sel}$ and complexity $\leq \pi_{bic}$ are considered stable and parsimonious.
Major contributions of this thesis can be viewed from the aspects of method development and applications to real-world data. In this chapter, we briefly discuss those contributions, followed by suggestions and directions to which this research can be extended.

Our primary method, S3C, attempts to address two problems: a large number of possible SEMs and instability in model estimation. S3C, therefore, performs an exploratory search, seeking for optimal models from the entire range of model complexities. While such an exploratory approach is suitable for cases for which no theory is available, computationally it does not scale very well when the number of variables of interest increases. The expensive computation stems from the fact that S3C uses a recursive SEM as a model representation that leads to an exponential growth of model complexity, and that S3C combines a heuristic approach and stability selection to search for optimal models, which requires multiple repeated processes. Depending on the number of variables and the parameter setting, the iterative procedure of S3C is considerably more time consuming than standard causal discovery methods such as the PC and FCI algorithm (Spirtes et al., 2000) and bootstrapped versions thereof. We have been so far administering parallel computation to speed up analyses by simultaneously conducting the searches on different data subsets at the same time. S3C adopts the NSGA-II procedure to search for Pareto optimal models. More recent development in the field of multi-objective optimization with evolutionary algorithms, e.g., Jozefowiez et al. (2005); Pradhan and Panda (2012); Redondo et al. (2017), should be computationally more efficient and thus may be used to accelerate analyses even more.

S3C assumes several underlying contexts. Firstly, it assumes that the data have been generated from a linear Gaussian SEM. Secondly, the causal relations are non-reciprocal or acyclic, that is, no causal path is allowed from a variable to itself.
Thirdly, it assumes causal sufficiency, i.e., that there are no hidden confounders (latent variables) among observed variables. However, these assumptions do not always hold in practice. Real-world data may be produced from systems that are nonlinear, with non-Gaussian noise, and reciprocal relations. Extensions to relax these current assumptions will enrich functionality so as to solve a larger variety of real-world problems. As mentioned in the early chapters, S3C can be viewed as a general framework that combines the concepts of multi-objective optimization and stability selection, which can be integrated with other approaches, without the need to modify their original assumptions. For example, an extension to nonlinearity can be realized by employing nonlinear SEM approaches, e.g., Marsh et al. (2004); Bauer (2005); Kelava et al. (2011, 2014) in the optimization phase of S3C. Moreover, there is an R package, named \texttt{nlsem} (Umbach et al., 2017), that provides functions for fitting nonlinear SEMs. As S3C is also implemented as an R package (discussed below), an extension to nonlinear system should be less complicated.

Clinical data is often gathered by collecting records of patients across time. Such a study design is called longitudinal. We extended S3C to S3L, to handle longitudinal data. This was realized by changing the model representation used by S3C to a longitudinal causal model that consists of a baseline and a transition model. The transition model comprises two time slices, on which S3L assumes that the time intervals between the time slices are uniform among subjects. Nevertheless, in practice time intervals between the time slices may be unequal within a subject and/or across subjects. Several approaches have been introduced to handle such unequal intervals, for example, multi-level or hierarchical regression models (Kreft and de Leeuw, 1998; Raudenbush and Bryk, 2002). Extensions that address this problem should be considered for future work.

In the clinical domain and other fields, it is often of interest to model causal relations among latent variables that are measured through observed proxies. In order to model such causal relations, we extended S3C to S3C-Latent by changing the model representation and ensuring several model identification conditions. S3C-Latent, in particular, uses the so-called general SEM which consists of a structural model (representing causal relations among latent variables) and a measurement model (representing effects of latent variables on the corresponding indicators). We further assume that the measurement model is given and pure, i.e., no observed variables loading to multiple latent variables. Notice that such a model is a specific type of latent model and that S3C-Latent does not handle any other types of latent variables. Therefore, just like S3C and S3L, S3C-Latent in essence still assumes causal sufficiency. Furthermore, S3C-Latent conducts specification search specifically on the structural model. A more advanced approach is to extend the specification search to the measurement model and allow for cross-loadings (Asparouhov and Muthén, 2009; Marsh et al., 2009), i.e., indicators that load to multiple latent variables. Such an approach, however, will lead to more complex (or even infeasible) computations, and thus should be considered with caution.

We implemented S3C/L as an R package called \texttt{stablespec}, which is available
at CRAN with MIT license. The current version of stablespec (0.3.0) has not yet exported internal functions which might be helpful for users to customize (the source code of the internal functions still can be accessed, e.g., via the Github repository). Separate functions of S3C, S3L, and S3C-Latent will certainly increase user’s convenience and flexibility. The inferred causal graph has not been annotated with reliability scores and standardized causal effects. Besides maintenance, future work related to stablespec should consider the aforementioned points.

We conducted two applications to real-world data, in which we discussed the results more comprehensively. The first study applied S3L as a mediation analysis tool to understand the mechanisms of change brought by CBT for patients with medically unexplained fatigue. We found that CBT causes changes on fatigue-related cognitive processes, i.e., increased self-efficacy with respect to fatigue, a reduced focus on fatigue, and a reduced tendency to catastrophize. The second study applied S3C-Latent to a data set of Indonesian patients with TBM, and intended to understand causal relations between hyponatremia with disease severity and with mortality. We found that hyponatremia does not seem to contribute to increased mortality. Instead, HIV is found to be the most stable predictor of mortality. Therefore, correction of hyponatremia unlikely improves prognosis.

The two studies above assumed that there are no reciprocal causal relations. For example, in the case of the first study, this was expressed through the assumption that fatigue in the follow-up time slice causes nothing. A more realistic proposition is that the relations among fatigue, beliefs, and behavior is reciprocal. That is, a reduced fatigue may have a positive effect on self-efficacy or decreases focus on the fatigue, which in return reduces the fatigue further. S3C, S3L, and S3C-Latent, however, still assume non-reciprocal relations. Several methods have been introduced to handle cyclic causal relations, for example, Richardson and Spirtes (1996); Hyttinen et al. (2012); Lacerda et al. (2008); Mooij and Heskes (2013). Extensions to handle reciprocal causal relations are an important direction, and thus should be considered for future development.


Summary

Structural equation model (SEM) is a primary language of causal discovery. Typical SEM applications in the literature start with a hypothesized model, and continue with a few model refinements before concluding the final model. This process is referred to as specification search. Such an approach faces two problems. The first problem originates from the fact that a modest number of variables in the data leads to an immense number of possible SEMs. Given $p$ variables and assuming a non-reciprocal model, there will be $3^{p(p-1)/2}$ possible SEMs. This makes that a modest number of variables, e.g., six, already entails more than 14 millions possible SEMs. In many clinical cases related to rare diseases, theory might not yet be available or difficult to propose. Considering the number of possible SEMs above, a more exploratory approach could be an ideal alternative. The second problem stems from the fact that model estimation is notoriously unstable. That is, a slight change in the data can lead to a considerable change in the inferred model.

In this thesis, therefore, we develop an exploratory score-based causal discovery method. The chief aim of the causal discovery method is to resolve the problems of the immense number of possible SEMs and the instability in model estimation.

In Chapter 1, we give an introduction to this thesis, describing causal paradigm, causal approaches, and applications to the clinical domain. In Chapter 2, we provide background knowledge necessary to the rest of the thesis, including SEM, stability selection, and multi-objective optimization.

S3C

In Chapter 3, we introduce stable specification search for cross-sectional data (S3C). S3C is an exploratory score-based causal method, which combines the idea of stability selection and multi-objective optimization to search for optimal SEMs from the whole range of model complexities. S3C explores and evaluates numerous possible SEMs through iterative model refinements, i.e., new models are created based on previous optimal models that are characterized by two conflicting objectives: fit the data well and have simple model structure. This procedure is repeated over different subsamplings, resulting in a large number of optimal models across different model complexities. From those optimal models, S3C selects stable and simple (parsimonious) causal structures (called relevant) by means of stability and com-
plexity thresholds, respectively. The relevant model structures are then visualized as a single causal graph. Several underlying contexts are assumed by S3C: the data have been generated from a linear Gaussian SEM; non-reciprocal causal relations; and no latent variables. The performance of S3C is evaluated on simulated data for which the true model is known, and compare to the results of state-of-the-art causal discovery methods. The comparison shows that S3C achieve significant improvement in obtaining the causal relations over those alternative approaches. We further apply S3C to real-world clinical data about patients with chronic fatigue syndrome (CFS). The overall result conform to previous studies in the field.

S3L

In Chapter 4 we extend S3C to S3L, to handle longitudinal data. This is realized by replacing the causal model used in S3C to a longitudinal SEM that consists of a baseline model and transition model. The baseline model represents causal relations between variables at the baseline time slice $t_0$. The transition model comprises two time slices $t_{i-1}$ and $t_i$ for $i > 0$, representing causal relations that may occur across time. S3L in principle applies S3C separately to the baseline and to the transition models. In addition, S3L estimates total causal effects for relevant causal structures. S3L assumes the same underlying contexts as S3C does (see above). In addition, S3L also assumes time-invariant causal relations (stationary), and that the time intervals between time slices are fairly uniform. We evaluate the performance of S3L on simulated data and compare the performance of S3L to those of the state-of-the-art causal discovery methods. We also apply S3L to three real-world data sets about patients with CFS, patients with Alzheimer disease, and patients with chronic kidney disease (CKD). The comparison shows that S3L achieves at least comparable performance as, but often a significant improvement over the alternative approaches, especially in obtaining causal relations, and in the case of a small sample size. The results on the real-world data are corroborated with those of previous studies.

S3C-Latent

In Chapter 5 we extend S3C to S3C-Latent, to handle latent variables that are measured through observed indicators. This idea is realized by changing the model representation to SEM with latent variables that comprises a structural model (representing causal relations between latent variables) and a measurement model (representing effects of latent variables on indicators), and ensuring particular model identification conditions fulfilled. In common with S3C, S3C-Latent assumes the same underlying contexts (see above). In addition, S3C-Latent assumes that the measurement model is given and pure, that is, each indicator is related to only one latent variable. We evaluate the performance of S3C-Latent on simulated data with different schemes, and compare the results with that of PC-MIMBuild, an extension
of the PC algorithm, a state-of-the-art causal discovery method. The results shows that S3C-Latent performs better. We also apply S3C-Latent to two real-world data sets about children with attention deficit/hyperactivity disorder (ADHD) and about measuring mental abilities among pupils. The results are consistent with those of earlier studies.

**R Package stablespec**

In Chapter 6 we describe an implementation of S3C/L as an R package called stablespec. It is available at the Comprehensive R Archive Network (CRAN) with MIT license. stablespec comes along with a documentation that includes running examples, and with artificial data sets, so that users can use and customize with ease. There is also an example of running stablespec on ADHD data. Started from version 0.3.0, stablespec supports for parallel computation to speed up analysis. Finally, we create a Github repository, to invite interested users to collaborate developing and improving the package.

**Applications to real-world data**

In Chapter 7 we apply S3L to a real-world data set that consists of 264 patients from two randomized controlled trials (RCTs), testing the efficacy of cognitive behavior therapy (CBT) for medically unexplained fatigue in patients with CFS and patients with idiopathic chronic fatigue (ICF). The purpose of the study is to identify the mechanisms of change that are brought by the CBT. We apply S3L as a mediation analysis tool, focusing on how the therapy affects the patients’ fatigue. We find that the CBT influences changes in the fatigue-related cognitive processes: an increased self-efficacy with respect to fatigue, a reduced focus on fatigue, and a reduced tendency to catastrophize. Those changes, in turn, lead to a reduction in fatigue directly, and indirectly via an improved physical functioning and increased perceived activity. These findings partly replicate results of the previous mediation analysis of CBT for CFS. The main difference compared to those of the previous studies is that CBT seems to have direct effects on fatigue-related cognitive processes.

In Chapter 8 we apply S3C-Latent to a real-world data set that comprises 678 Indonesian patients with tubercolous meningitis (TBM), focusing on one-year mortality. This study intends to understand causal relations between hyponatremia and disease severity, and whether hyponatremia contributes to mortality. More specifically, we conduct two analyses to obtain two models: a causal model among predictors of mortality by applying S3C-Latent and a survival model by extending the concept of stability selection to stable Cox model. Taken together, we obtain a causal survival model. According to the model, hyponatremia does not contribute to mortality. Instead, HIV does contribute to mortality as it is found to be the most stable predictor. Moreover, we find that HIV infection increases the proba-
bility of negative CSF culture; a positive CSF culture induces inflammation and lowers blood sodium levels; both tissue damage and blood sodium are associated with inflammation.
Een Structural Equation Model (SEM) is een primaire taal voor het ontdekken van causale verbanden. Een typische toepassing van SEM in de literatuur begint met een verondersteld model en past dan enkele modelverfijningen toe voordat er een definitief model wordt bepaald. Dit proces wordt ook wel aangeduid als het zoeken naar een specificatie. Deze benadering heeft twee problemen. Het eerste probleem komt voort uit het feit dat een bescheiden aantal variabelen in de gegevens leidt tot een enorm aantal mogelijke SEM’s. Gegeven $p$ variabelen en onder de aanname van een niet-recursief model, zijn er al $3^p(p-1)/2$ mogelijke SEM modellen. Met een bescheiden aantal variabelen, bijvoorbeeld zes, zijn er dus al meer dan 14 miljoen mogelijke SEM modellen. Daarnaast is in veel klinische gevallen die verband houden met zeldzame ziekten, nog geen hypothetisch model beschikbaar om als startpunt te nemen. Een meer verkennende aanpak zou dus, gezien het aantal mogelijke SEMs, een ideaal alternatief kunnen zijn. Het tweede probleem komt voort uit het feit dat modelschatting inherent onstabiel is. Een kleine verandering in de gegevens kan resulteren in een aanzienlijke verandering in het afgeleide model.

In dit proefschrift ontwikkelen we een methode voor het ontekennen van causale verbanden die niet uitgaat van een initiële hypothese maar verkennen is. Het doel van deze methode is om een oplossing te bieden voor de eerder genoemde problemen: het enorme aantal mogelijke SEMs en de instabiliteit in modelschatting.

In hoofdstuk 1 geven we een inleiding tot dit proefschrift, waarin het causale paradigma wordt beschreven, causale benaderingen en toepassingen op het klinische domein. In hoofdstuk 2 geven we de achtergrondkennis die noodzakelijk is voor de rest van het proefschrift over SEMs, stabiliteit selectie en multi-objectieve optimalisatie.

S3C

In hoofdstuk 3 introduceren we het zoeken naar stabiele specificaties voor transversale gegevens (S3C). S3C is een verkennende en op scores gebaseerde causale methode, die stabilisatie selectie en multi-objectieve optimalisatie combineert om te zoeken naar optimale SEMs uit het hele scala van modelcomplexiteiten. S3C onderzoekt en evalueert tal van mogelijke SEMs door middel van iteratieve modelverfijningen, d.w.z. nieuwe modellen worden gemaakt op basis van eerdere optimale
modellen die worden gekenmerkt door twee tegenstrijdige doelstellingen: een goede fit van de gegevens en een eenvoudige modelstructuur. Deze procedure wordt herhaald over verschillende subsamples wat resulteert in een groot aantal optimale modellen voor verschillende modelcomplexiteiten. Van die optimale modellen selecteert S3C dan de stabiele en eenvoudige causale structuren (die we relevant noemen) door middel van drempelwaarden voor de stabilité en complexiteit. De relevante modelstructuren worden vervolgens gevisualiseerd als een enkele causale graaf. S3C maakt hiervoor wel enkele aannamen: de gegevens zijn gegenereerd op basis van een lineair Gaussiaans SEM; niet-recursieve causale relaties; en geen latente variabelen. De prestaties van S3C worden geëvalueerd op gesimuleerde gegevens waarvoor het echte model bekend is en door te vergelijken met de resultaten van recente causale ontdekkingsmethoden. De vergelijking toont aan dat S3C een significante verbetering geeft. Daarnaast wordt S3C toegepast op klinische gegevens uit de praktijk over patiënten met een chronisch vermoeidheid syndroom (CVS). Het algehele resultaat komt overeen met eerdere onderzoeken in het veld.

S3L

In hoofdstuk 4 breiden we S3C uit naar S3L voor longitudinale gegevens. Dit wordt gerealiseerd door het causale model dat in S3C wordt gebruikt te vervangen door een longitudinale SEM die bestaat uit een baselinelmodel en een overgangsmodel. Het baselinelmodel vertegenwoordigt causale relaties tussen variabelen in het basistijdsegment $t_0$. Het transitiemodel omvat twee tijdsegmenten $t_{i-1}$ en $t_i$ voor $i > 0$, die causale relaties vertegenwoordigen die in de loop van de tijd kunnen voorkomen. S3L past S3C in principe afzonderlijk toe op het baselinelmodel en op het transitiemodel. Bovendien schat S3L de totale causale effecten voor relevante causale structuren. S3L neemt dezelfde onderliggende aannamen aan als S3C (zie hierboven). Bovendien neemt S3L ook tijdinvariante causale relaties (stationair) aan, en dat de tijdsintervallen tussen tijdsegmenten redelijk uniform zijn. We evalueren de prestaties van S3L op gesimuleerde gegevens en vergelijken de prestaties van S3L met die van andere recente methoden. We passen S3L ook toe op drie gegevensbestanden uit de praktijk over patiënten met CVS, patiënten met de ziekte van Alzheimer en patiënten met chronische nierziekte (CKD). De vergelijking laat zien dat S3L ten minste vergelijkbare prestaties levert als, maar vaak een significante verbetering geeft ten opzichte van de alternatieve methoden, vooral bij het verkrijgen van causale relaties en in het geval van een kleine steekproefomvang. De resultaten van de gegevens uit de praktijk worden bevestigd met die van eerdere studies.
S3C-Latent

In hoofdstuk 5 breiden we S3C uit naar S3C-Latent, voor gegevens met latente variabelen die worden gemeten aan de hand van waargenomen indicatoren. Dit wordt gerealiseerd door de modelrepresentatie in SEM te wijzigen met latente variabelen die een structureel model (dat causale relaties tussen latente variabelen vertegenwoordigt) en een meetmodel (dat effecten van latente variabelen op indicatoren vertegenwoordigt) omvat, en ervoor te zorgen dat er aan specifieke modelidentificatievoorwaarden wordt voldaan. Net als S3C neemt S3C-Latent dezelfde onderliggende annamen aan (zie hierboven). Bovendien veronderstelt S3C-Latent dat het meetmodel zuiver en gegeven is, dat wil zeggen dat elke indicator is gerelateerd aan slechts één latente variabele. We evalueren de prestaties van S3C-Latent op gemodelleerde gegevens met verschillende schema’s en vergelijken de resultaten met die van PC-MIMBuild, een uitbreiding van het PC-algoritme, een geavanceerde causale ontdekkingsmethode. De resultaten laten zien dat S3C-Latent beter presteert. We passen S3C-Latent ook toe op twee gegevensbestanden uit de praktijk over kinderen met ADHD (Attention Deficit / Hyperactivity Disorder) en het meten van mentale vermogens bij leerlingen. De resultaten komen overeen met die van eerdere studies.

R-Pakket stablespec

In hoofdstuk 6 beschrijven we een implementatie van S3C/L als een R-pakket met de naam stablespec. Het is verkrijgbaar bij het Comprehensive R Archive Network (CRAN) met MIT-licentie. stablespec wordt geleverd samen met documentatie met lopende voorbeelden en kunstmatige gegevensbestanden, zodat gebruikers deze gemakkelijk kunnen gebruiken en aanpassen. Er is ook een voorbeeld van het uitvoeren van stablespec op ADHD-gegevens. Gestart vanaf versie 0.3.0 ondersteunt stablespec voor parallelle berekeningen om de analyse te versnellen. Ten slotte creëren we een Github-repository om geïnteresseerde gebruikers uit te nodigen om samen te werken aan de ontwikkeling en verbetering van het pakket.

Toepassingen op praktijkgegevens

In hoofdstuk 7 passen we S3L toe op een gegevensbestand uit de praktijk met 264 patiënten uit twee gerandomiseerde gecontroleerde onderzoeken (RCT’s), waarbij de effectiviteit van cognitieve gedragstherapie (CBT) wordt getoetst voor medisch onverklaarde vermoeidheid bij patiënten met CVS en patiënten met idiopathische chronische vermoeidheid (ICF). Het doel van de studie is het identificeren van de veranderingsmechanismen als gevolg van CBT. We passen S3L toe als een analysetool, waarbij de nadruk ligt op hoe de therapie de vermoeidheid van de patiënt beïnvloedt. We vinden dat CBT resulteert in veranderingen in de aan vermoeidheid gerelateerde cognitieve processen: een toename in zelfredzaamheid met betrekking
tot vermoeidheid, een verminderde focus op vermoeidheid en een verminderde neiging tot catastroferen. Die veranderingen leiden op hun beurt tot een vermindering van vermoeidheid, zowel direct als indirect, via een verbeterd fysiek functioneren en een toegenomen vermeende activiteit. Deze bevindingen repliceren gedeeltelijk de resultaten van de vorige bemiddelingsanalyse van CBT voor CVS. Het belangrijkste verschil met die van de vorige studies is dat CBT directe effecten lijkt te hebben op de aan vermoeidheid gerelateerde cognitieve processen.

In Hoofdstuk 8 passen we S3C-Latent toe op een gegevensbestand uit de praktijk die 678 Indonesische patiënten met tubercotiche meningitis (TBM) bevat, met een focus op sterfte na één jaar. Deze studie beoogt de causale verbanden van hyponatriëmie te begrijpen en de ernst van de ziekte, en of hyponatriëmie bijdraagt aan de mortaliteit. Meer specifiek voeren we twee analyses uit om twee modellen te verkrijgen: een causaal model uit voorspellers van mortaliteit door S3C-Latent toe te passen en een overlevingsmodel door deze uit te breiden met het concept van stabilitéselectie naar een stabiel Cox-model. Alles bij elkaar genomen krijgen we een causaal overlevingsmodel. Volgens het model draagt hyponatriëmie niet bij aan de mortaliteit. In plaats daarvan draagt HIV wel bij aan de sterfte omdat dit de meest stabiele voorspeller is. Bovendien vinden we dat HIV-infectie de kans verhoogt op een negatieve CSF-cultuur; een positieve CSF-kweek induceert ontsteking en verlaagt het natriumgehalte in het bloed; zowel weefselbeschadiging als bloednatrium zijn geassocieerd met een ontsteking.
Publications

Papers in international journals


submitted:

R. Rahmadi, P. Groot, T. Heskes. “Stable specification search in structural equation models with latent variables”.


Papers in conference proceedings


All praise is due to Allah. May blessings and peace be upon the Messenger of Allah. *Behind the scene* of this thesis, there are hope, helps, supports, prayers, and trust, from many great people. In this part of this thesis, I shall take this opportunity to express my gratitude to them.

Tom Heskes and Perry Groot. I recalled when the three of us met for the first time via a Skype meeting, discussing a possibility of my PhD plan. It was 2012. In about one year later, I arrived at Nijmegen, starting the plan we discussed; that was August 2013, the beginning of my PhD. From then on, we had pretty consistent weekly meetings; from discussions about research (of course), to family-related news exchanges. I am, in particular, grateful of those times, and would like to say to you both: thank you.

Tom, thank you for opening the door and letting me in; going through this amazing and unforgettable PhD journey. You have been very receptive to any new ideas, and at the same time, flagging constructive criticisms and clever suggestions when necessary. Thank you for being supportive to all of my PhD plans, cordial in all discussions and conversations, as well as very patient. Perry, thank you for your kindness and patience, helping me to go through every single details of my works. And I truly appreciate it as you have been there, whenever I had questions and something to discuss. Once again, I thank you both most warmly; two best supervisors in the world!

I would like to thank all of my colleagues at the Radboud University Nijmegen, Elena Sokolova, Elena Marchiori, Nicole Messink, Wout Megchelenbrink, Robbert Krebbers, Jonce Dimov, Henning Basold, Simone Lederer, Feri Wijayanto, Jesse Krijthe, Twan van Laarhoven, Tom Claassen, Tjeerd Dijkstra, Theo van der Weide, Daniel Kühlwein, Jacopo Acquarelli, Gabriel Bucur, Ruifei Cui, Gido Schoenmacker, Fabian Gieseke, Christiaan de Leeuw, Maryam Kamis, Martha Larson, Herman Geuvers, Arjen de Vries, Max Hinne, Marieke de Vries, Suzan Verberne, Saskia Koldijk, Maya Sappelli, Mohsen Ghafoorian, Luc Evers, Nastaran Mohammadian Rad, Errol Zalmijn, Konrad Philipp Miellke, Marvin Oeben, Zaheer Babar, Zhengyu Zhao, Kai Chen, Didi Rustam, Johannes Textor, David van Leeuwen, Maaike de Boer, Saiden Abbas, Tameem Adel, Kasper Brink, and Simone Meeuwsen. I appreciate all the good times we shared together and thank you for making our workplace a fun place to be.
I would like to thank my colleagues in the OPTIMISTIC project who brought me to another world of research. Thanks to Baziel van Engelen, Hans Knoop, Kees Okkersen, Jeffrey Glennon, Linda Heskamp, Cecilia Jimenez-Moreno, Shaun Treweek, Libby Wood, Benedikt Schoser, Juliane Dittrich, Darren Monckton, and all of the team, I did enjoy the times working with you.

I also would like to thank my colleagues from the Radboud University Medical Center, Marianne Heins, Jan van den Brand, Marieke van Rijn, Anthonie Janse, Jan-Federich Wiborg, Sofiati Dian, Arjan van Laarhoven, and Reinout van Crevel. Thank you for the excellent discussions and the works we have been conducting together.

I want to extend my thanks to my colleagues at the Carnegie Mellon University, and Pittsburgh University, Richard Scheines, Peter Spirtes, Joseph Ramsey, Kun Zhang, Clark Glymour, Gregory Cooper, Connor Robinson-Arnall, Masayuki Tashiro, Ruben Sanchez-Romero, Daniel Malinsky, Colin Zwanziger, Alex Markham, Jacqueline deFazio, Mary Grace Joseph, and Biwei Huang. Thank you for making my visiting research wonderful and memorable.

Many sincere thanks to Indonesian colleagues and their families in the Netherlands, especially to Tosri family and Ardiansyah family, to the PPIN and the Kemuni, to those in Pittsburgh, to my brothers at the Islamic Center of Pittsburgh, to my colleagues in the Department of Informatics, Universitas Islam Indonesia, and to my colleagues in the Budi Mulia Dua foundation. I am deeply indebted to you for all of your helps and supports.

I am eternally grateful to my family, of their endless supports, of their prayers. My sincerest thanks to my late father Nazaruddin and my mother Rosladiah, the teachers of all my first times; to my parents-in-law, Mohammad Amien Rais and Kusnasriyati Sri Rahayu, for those countless encouragements through thick and thin; to my brothers and sisters, Abdi Satriawan, Rizki Fitriasari and Darwin Umar, Ahmad Hanafi Rais and Astriani Karnaningrum, Hanum Rais and Rangga Almahendra, Ahmad Muntaz Rais and Futri Zulya, and Ahmad Baihaqy Rais and Selmadena Aquilla, who have always been my supporters.

Finally, to my wife Tasniem, my daughters Sofie and Sasha, I hardly find a sentence or two to express my feelings, after all of the times we have been together. Sofie, you were not even a 1-year-old little girl when first arrived at Linz, Austria, in the middle of winter, coming to give me support in finishing my master. Sasha, the little new member born in Nijmegen, becoming another driving force. And Tasniem, the person who has been there for me, rain or shine; you complete me. My heartfelt thanks to you all. I dedicate this thesis to you.
Ridho Rahmadi was born in Yogyakarta, Indonesia, on April 13, 1985. In 2008, he obtained his bachelor degree in Computer Science, from the Universitas Islam Indonesia. His academic journey continued to Europe; in 2012, he finished his master studies in Computer science (specialization on Artificial Intelligence) at the Czech Technical University in Prague, and the Johannes Kepler University Linz, Austria. In 2013, he started his PhD, at the Data Science group, Institute for Computing and Information Sciences, Radboud University Nijmegen, the Netherlands, under the supervision of Prof. Tom Heskes and Dr. Perry Groot. His PhD research aims to develop a novel causal model method that addresses the problem of an immense number of possible models and instability in model estimation. In 2017, he went to the Carnegie Mellon University, Pittsburgh, US, conducting a visiting research on causal discovery in the presence of measurement error, under the supervision of Prof. Richard Scheines.
## SIKS dissertations

<table>
<thead>
<tr>
<th>Year</th>
<th>Author</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>Botond Cseke (RUN)</td>
<td>Variational Algorithms for Bayesian Inference in Latent Gaussian Models</td>
</tr>
<tr>
<td>02</td>
<td>Nick Tinneemeier (UU)</td>
<td>Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language</td>
</tr>
<tr>
<td>03</td>
<td>Jan Martijn van der Werf (TUE)</td>
<td>Compositional Design and Verification of Component-Based Information Systems</td>
</tr>
<tr>
<td>04</td>
<td>Hado van Hasselt (UU)</td>
<td>Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference</td>
</tr>
<tr>
<td>05</td>
<td>Bas van der Raadt (VU)</td>
<td>Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline.</td>
</tr>
<tr>
<td>06</td>
<td>Yiwen Wang (TUE)</td>
<td>Semantically-Enhanced Recommendations in Cultural Heritage</td>
</tr>
<tr>
<td>07</td>
<td>Yujia Cao (UT)</td>
<td>Multimodal Information Presentation for High Load Human Computer Interaction</td>
</tr>
<tr>
<td>08</td>
<td>Nieske Vergunst (UU)</td>
<td>BDI-based Generation of Robust Task-Oriented Dialogues</td>
</tr>
<tr>
<td>09</td>
<td>Tim de Jong (OU)</td>
<td>Contextualised Mobile Media for Learning</td>
</tr>
<tr>
<td>10</td>
<td>Bart Bogaert (UvT)</td>
<td>Cloud Content Contention</td>
</tr>
<tr>
<td>11</td>
<td>Dhaval Vyas (UT)</td>
<td>Designing for Awareness: An Experience-focused HCI Perspective</td>
</tr>
<tr>
<td>12</td>
<td>Carmen Bratosin (TUE)</td>
<td>Grid Architecture for Distributed Process Mining</td>
</tr>
<tr>
<td>13</td>
<td>Xiaoyu Mao (UvT)</td>
<td>Airport under Control. Multiagent Scheduling for Airport Ground Handling</td>
</tr>
<tr>
<td>14</td>
<td>Milan Lovric (EUR)</td>
<td>Behavioral Finance and Agent-Based Artificial Markets</td>
</tr>
<tr>
<td>15</td>
<td>Marijn Koolen (UvA)</td>
<td>The Meaning of Structure: the Value of Link Evidence for Information Retrieval</td>
</tr>
<tr>
<td>16</td>
<td>Maarten Schadd (UM)</td>
<td>Selective Search in Games of Different Complexity</td>
</tr>
<tr>
<td>17</td>
<td>Jiyn He (UVA)</td>
<td>Exploring Topic Structure: Coherence, Diversity and Relatedness</td>
</tr>
<tr>
<td>18</td>
<td>Mark Ponsen (UM)</td>
<td>Strategic Decision-Making in complex games</td>
</tr>
<tr>
<td>19</td>
<td>Ellen Rusman (OU)</td>
<td>The Mind’s Eye on Personal Profiles</td>
</tr>
<tr>
<td>20</td>
<td>Qing Gu (VU)</td>
<td>Guiding service-oriented software engineering - A view-based approach</td>
</tr>
<tr>
<td>21</td>
<td>Linda Terlouw (TUD)</td>
<td>Modularization and Specification of Service-Oriented Systems</td>
</tr>
<tr>
<td>22</td>
<td>Junte Zhang (UVA)</td>
<td>System Evaluation of Archival Description and Access</td>
</tr>
<tr>
<td>23</td>
<td>Wouter Weerkamp (UVA)</td>
<td>Finding People and their Utterances in Social Media</td>
</tr>
<tr>
<td>24</td>
<td>Herwin van Welbergen (UT)</td>
<td>Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior</td>
</tr>
<tr>
<td>25</td>
<td>Syed Waqar ul Qounain Jaffry (VU)</td>
<td>Analysis and Validation of Models for Trust Dynamics</td>
</tr>
</tbody>
</table>
Matthijs Aart Pontier (VU), Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots

Aniel Bhulai (VU), Dynamic website optimization through autonomous management of design patterns

Rianne Kaptein (UVA), Effective Focused Retrieval by Exploiting Query Context and Document Structure

Faisal Kamiran (TUE), Discrimination-aware Classification

Egon van den Broek (UT), Affective Signal Processing (ASP): Unraveling the mystery of emotions

Ludo Waltman (EUR), Computational and Game-Theoretic Approaches for Modeling Bounded Rationality

Nees-Jan van Eck (EUR), Methodological Advances in Bibliometric Mapping of Science

Tom van der Weide (UU), Arguing to Motivate Decisions

Paolo Turrini (UU), Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations

Maaike Harbers (UU), Explaining Agent Behavior in Virtual Training

Erik van der Spek (UU), Experiments in serious game design: a cognitive approach

Adriana Burlutiu (RUN), Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference

Nyree Lemmens (UM), Bee-inspired Distributed Optimization

Joost Westra (UU), Organizing Adaptation using Agents in Serious Games

Viktor Clerc (VU), Architectural Knowledge Management in Global Software Development

Luan Ibraimi (UT), Cryptographically Enforced Distributed Data Access Control

Michal Sindlar (UU), Explaining Behavior through Mental State Attribution

Henk van der Schuur (UU), Process Improvement through Software Operation Knowledge

Boris Reuderink (UT), Robust Brain-Computer Interfaces

Herman Stehouwer (UvT), Statistical Language Models for Alternative Sequence Selection

Beibei Hu (TUD), Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work

Azizi Bin Ab Aziz (VU), Exploring Computational Models for Intelligent Support of Persons with Depression

Mark Ter Maat (UT), Response Selection and Turn-taking for a Sensitive Artificial Listening Agent

Andreea Niculescu (UT), Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

Terry Kakeeto (UvT), Relationship Marketing for SMEs in Uganda

Muhammad Umair (VU), Adaptivity, emotion, and Rationality in Human and Ambient Agent Models

Adam Vanya (VU), Supporting Architecture Evolution by Mining Software Repositories

Jurriaan Souer (UU), Development of Content Management System-based Web Applications

Marijn Plomp (UU), Maturing Interorganisational Information Systems

Wolfgang Reinhardt (OU), Awareness Support for Knowledge Workers in Research Networks

Rianne van Lambalgen (VU), When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions

Gerben de Vries (UVA), Kernel Methods for Vessel Trajectories

2012
09 Ricardo Neisse (UT), Trust and Privacy Management Support for Context-Aware Service Platforms
10 David Smits (TUE), Towards a Generic Distributed Adaptive Hypermedia Environment
11 J.C.B. Rantham Prabhakara (TUE), Process Mining in the Large: Preprocessing, Discovery, and Diagnostics
12 Kees van der Sluijs (TUE), Model Driven Design and Data Integration in Semantic Web Information Systems
13 Suleman Shahid (UvT), Fun and Face: Exploring non-verbal expressions of emotion during playful interactions
14 Evgeny Knutov (TUE), Generic Adaptation Framework for Unifying Adaptive Web-based Systems
15 Natalie van der Wal (VU), Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes.
16 Fiemke Both (VU), Helping people by understanding them - Ambient Agents supporting task execution and depression treatment
17 Amal Elgammal (UvT), Towards a Comprehensive Framework for Business Process Compliance
18 Eltjo Poort (VU), Improving Solution Architecting Practices
19 Helen Schonenberg (TUE), What’s Next? Operational Support for Business Process Execution
20 Ali Bahramisharif (RUN), Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing
21 Roberto Cornacchia (TUD), Querying Sparse Matrices for Information Retrieval
22 Thijs Vis (UvT), Intelligence, politie en veiligheidsdienst: verenigbare grootheden?
23 Christian Muehl (UT), Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction
24 Laurens van der Werff (UT), Evaluation of Noisy Transcripts for Spoken Document Retrieval
25 Silja Eckartz (UT), Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application
26 Emile de Maat (UVA), Making Sense of Legal Text
27 Hayrettin Gurkok (UT), Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games
28 Nancy Pascaill (UvT), Engendering Technology Empowering Women
29 Almer Tigelaar (UT), Peer-to-Peer Information Retrieval
30 Alina Pommeranz (TUD), Designing Human-Centered Systems for Reflective Decision Making
31 Emily Bagarukayo (RUN), A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure
32 Wietske Visser (TUD), Qualitative multi-criteria preference representation and reasoning
33 Rory Sie (OUN), Coalitions in Cooperation Networks (COCOON)
34 Pavol Jancura (RUN), Evolutionary analysis in PPI networks and applications
35 Evert Haasdijk (VU), Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics
36 Denis Ssebugwawo (RUN), Analysis and Evaluation of Collaborative Modeling Processes
37 Agnes Nakakawa (RUN), A Collaboration Process for Enterprise Architecture Creation
38 Selmar Smit (VU), Parameter Tuning and Scientific Testing in Evolutionary Algorithms
39 Hassan Fatemi (UT), Risk-aware design of value and coordination networks
40 Agus Gunawan (UvT), Information Access for SMEs in Indonesia
41 Sebastian Kelle (OU), Game Design Patterns for Learning
42 Dominique Verpoorten (OU), Reflection Amplifiers in self-regulated Learning
43 Withdrawn
44 Anna Tordai (VU), On Combining Alignment Techniques
45 Benedikt Kratz (UvT), A Model and Language for Business-aware Transactions
46 Simon Carter (UVA), Exploration and Exploitation of Multilingual Data for Statistical Machine Translation
47 Manos Tsagkias (UVA), Mining Social Media: Tracking Content and Predicting Behavior
48 Jorn Bakker (TUE), Handling Abrupt Changes in Evolving Time-series Data
49 Michael Kaisers (UM), Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions
50 Steven van Kervel (TUD), Ontology driven Enterprise Information Systems Engineering
51 Jeroen de Jong (TUD), Heuristics in Dynamic Scheduling: a practical framework with a case study in elevator dispatching

2013
01 Viorel Milea (EUR), News Analytics for Financial Decision Support
02 Erietta Liarou (CWI), MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing
03 Szymon Klorman (VU), Reasoning with Contexts in Description Logics
04 Chetan Yadati (TUD), Coordinating autonomous planning and scheduling
05 Dulce Pumareja (UT), Groupware Requirements Evolutions Patterns
06 Romulo Goncalves (CWI), The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience
07 Giel van Lankveld (UvT), Quantifying Individual Player Differences
08 Robbert-Jan Merk (VU), Making enemies: cognitive modeling for opponent agents in fighter pilot simulators
09 Fabio Gori (RUN), Metagenomic Data Analysis: Computational Methods and Applications
10 Jeewanie Jayasinghe Arachchige (UvT), A Unified Modeling Framework for Service Design.
11 Evangelos Pournaras (TUD), Multi-level Reconfigurable Self-organization in Overlay Services
12 Marian Razavian (VU), Knowledge-driven Migration to Services
13 Mohammad Safiri (UT), Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly
14 Jafar Tanha (UVA), Ensemble Approaches to Semi-Supervised Learning Learning
15 Daniel Hennes (UM), Multiagent Learning - Dynamic Games and Applications
16 Eric Kok (UU), Exploring the practical benefits of argumentation in multi-agent deliberation
17 Koen Kok (VU), The PowerMatcher: Smart Coordination for the Smart Electricity Grid
18 Jeroen Janssens (UvT), Outlier Selection and One-Class Classification
19 Renze Steenhuisen (TUD), Coordinated Multi-Agent Planning and Scheduling
20 Katja Hofmann (UvA), Fast and Reliable Online Learning to Rank for Information Retrieval
21 Sander Wubben (UvT), Text-to-text generation by monolingual machine translation
22 Tom Claassen (RUN), Causal Discovery and Logic
23 Patricio de Alencar Silva (UvT), Value Activity Monitoring
24 Haitham Bou Ammar (UM), Automated Transfer in Reinforcement Learning
<table>
<thead>
<tr>
<th></th>
<th>Author</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>Alireza Zarghami (UT)</td>
<td>Architectural Support for Dynamic Homecare Service Provisioning</td>
</tr>
<tr>
<td>27</td>
<td>Mohammad Huq (UT)</td>
<td>Inference-based Framework Managing Data Provenance</td>
</tr>
<tr>
<td>28</td>
<td>Frans van der Sluis (UT)</td>
<td>When Complexity becomes Interesting: An Inquiry into the Information eXperience</td>
</tr>
<tr>
<td>29</td>
<td>Iwan de Kok (UT)</td>
<td>Listening Heads</td>
</tr>
<tr>
<td>31</td>
<td>Dinh Khoa Nguyen (UvT)</td>
<td>Blueprint Model and Language for Engineering Cloud Applications</td>
</tr>
<tr>
<td>32</td>
<td>Kamakshi Rajagopal (OUN)</td>
<td>Networking For Learning; The role of Networking in a Lifelong Learner’s Professional Development</td>
</tr>
<tr>
<td>33</td>
<td>Qi Gao (TUD)</td>
<td>User Modeling and Personalization in the Microblogging Sphere</td>
</tr>
<tr>
<td>34</td>
<td>Kien Tjin-Kam-Jet (UT)</td>
<td>Distributed Deep Web Search</td>
</tr>
<tr>
<td>35</td>
<td>Abdallah El Ali (UvA)</td>
<td>Minimal Mobile Human Computer Interaction</td>
</tr>
<tr>
<td>36</td>
<td>Than Lam Hoang (TUE)</td>
<td>Pattern Mining in Data Streams</td>
</tr>
<tr>
<td>37</td>
<td>Dirk Börner (OUN)</td>
<td>Ambient Learning Displays</td>
</tr>
<tr>
<td>38</td>
<td>Eelco den Heijer (VU)</td>
<td>Autonomous Evolutionary Art</td>
</tr>
<tr>
<td>39</td>
<td>Joop de Jong (TUD)</td>
<td>A Method for Enterprise Ontology based Design of Enterprise Information Systems</td>
</tr>
<tr>
<td>40</td>
<td>Pim Nijssen (UM)</td>
<td>Monte-Carlo Tree Search for Multi-Player Games</td>
</tr>
<tr>
<td>41</td>
<td>Jochem Liem (UVA)</td>
<td>Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning</td>
</tr>
<tr>
<td>42</td>
<td>Léon Planken (TUD)</td>
<td>Algorithms for Simple Temporal Reasoning</td>
</tr>
<tr>
<td>43</td>
<td>Marc Bron (UVA)</td>
<td>Exploration and Contextualization through Interaction and Concepts</td>
</tr>
</tbody>
</table>

### 2014
<table>
<thead>
<tr>
<th></th>
<th>Author</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>Nicola Barile (UU)</td>
<td>Studies in Learning Monotone Models from Data</td>
</tr>
<tr>
<td>02</td>
<td>Fiona Tuliyano (RUN)</td>
<td>Combining System Dynamics with a Domain Modeling Method</td>
</tr>
<tr>
<td>03</td>
<td>Sergio Raul Duarte Torres (UT)</td>
<td>Information Retrieval for Children: Search Behavior and Solutions</td>
</tr>
<tr>
<td>04</td>
<td>Hanna Jochmann-Mannak (UT)</td>
<td>Websites for children: search strategies and interface design - Three studies on children’s search performance and evaluation</td>
</tr>
<tr>
<td>05</td>
<td>Jurriaan van Reijsen (UU)</td>
<td>Knowledge Perspectives on Advancing Dynamic Capability</td>
</tr>
<tr>
<td>06</td>
<td>Damian Tamburri (VU)</td>
<td>Supporting Networked Software Development</td>
</tr>
<tr>
<td>07</td>
<td>Arya Adriansyah (TUE)</td>
<td>Aligning Observed and Modeled Behavior</td>
</tr>
<tr>
<td>08</td>
<td>Samur Araujo (TUD)</td>
<td>Data Integration over Distributed and Heterogeneous Data Endpoints</td>
</tr>
<tr>
<td>09</td>
<td>Philip Jackson (UvT)</td>
<td>Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language</td>
</tr>
<tr>
<td>10</td>
<td>Ivan Salvador Razo Zapata (VU)</td>
<td>Service Value Networks</td>
</tr>
<tr>
<td>11</td>
<td>Janneke van der Zwaan (TUD)</td>
<td>An Empathic Virtual Buddy for Social Support</td>
</tr>
<tr>
<td>12</td>
<td>Willem van Willigen (VU)</td>
<td>Look Ma, No Hands: Aspects of Autonomous Vehicle Control</td>
</tr>
<tr>
<td>13</td>
<td>Arlette van Wissen (VU)</td>
<td>Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains</td>
</tr>
<tr>
<td>14</td>
<td>Yangyang Shi (TUD)</td>
<td>Language Models With Meta-information</td>
</tr>
<tr>
<td>15</td>
<td>Natalya Mogles (VU)</td>
<td>Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare</td>
</tr>
</tbody>
</table>
16 Krystyna Milian (VU), Supporting trial recruitment and design by automatically interpreting eligibility criteria
17 Kathrin Dentler (VU), Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability
18 Mattijs Ghijsen (UVA), Methods and Models for the Design and Study of Dynamic Agent Organizations
19 Vinicius Ramos (TUE), Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support
20 Mena Habib (UT), Named Entity Extraction and Disambiguation for Informal Text: The Missing Link
21 Kassidy Clark (TUD), Negotiation and Monitoring in Open Environments
22 Marieke Peeters (UU), Personalized Educational Games - Developing agent-supported scenario-based training
23 Eleftherios Sidirourgos (UvA/CWI), Space Efficient Indexes for the Big Data Era
24 Davide Ceolin (VU), Trusting Semi-structured Web Data
25 Martijn Lappenschaar (RUN), New network models for the analysis of disease interaction
26 Tim Baarslag (TUD), What to Bid and When to Stop
27 Rui Jorge Almeida (EUR), Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty
28 Anna Chmielowiec (VU), Decentralized k-Clique Matching
29 Jaap Kabbedijk (UU), Variability in Multi-Tenant Enterprise Software
30 Peter de Cock (UvT), Anticipating Criminal Behaviour
31 Leo van Moergestel (UU), Agent Technology in Agile Multiparallel Manufacturing and Product Support
32 Naser Ayat (UvA), On Entity Resolution in Probabilistic Data
33 Tesfa Tegegne (RUN), Service Discovery in eHealth
34 Christina Manteli (VU), The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems.
35 Joost van Ooijen (UU), Cognitive Agents in Virtual Worlds: A Middleware Design Approach
36 Joos Buijs (TUE), Flexible Evolutionary Algorithms for Mining Structured Process Models
37 Maral Dadvar (UT), Experts and Machines United Against Cyberbullying
38 Danny Plass-Oude Bos (UT), Making brain-computer interfaces better: improving usability through post-processing.
39 Jasmina Maric (UvT), Web Communities, Immigration, and Social Capital
40 Walter Omona (RUN), A Framework for Knowledge Management Using ICT in Higher Education
41 Frederic Hogenboom (EUR), Automated Detection of Financial Events in News Text
42 Carsten Eijckhof (CWI/TUD), Contextual Multidimensional Relevance Models
43 Kevin Vlaanderen (UU), Supporting Process Improvement using Method Increments
44 Paulien Meesters (UvT), Intelligent Blauw. Met als ondertitel: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden.
45 Birgit Schmitz (OUN), Mobile Games for Learning: A Pattern-Based Approach
46 Ke Tao (TUD), Social Web Data Analytics: Relevance, Redundancy, Diversity
47 Shangsong Liang (UVA), Fusion and Diversification in Information Retrieval

2015
01 Niels Netten (UvA), Machine Learning for Relevance of Information in Crisis Response
02 Faiza Bukhsh (UvT), Smart auditing: Innovative Compliance Checking in Customs Controls
03 Twan van Laarhoven (RUN), Machine learning for network data
Howard Spoelstra (OUN), Collaborations in Open Learning Environments

Christoph Bösch (UT), Cryptographically Enforced Search Pattern Hiding

Farideh Heidari (TUD), Business Process Quality Computation - Computing Non-Functional Requirements to Improve Business Processes

Maria-Hendrike Peetz (UvA), Time-Aware Online Reputation Analysis

Jie Jiang (TUD), Organizational Compliance: An agent-based model for designing and evaluating organizational interactions

Randy Klaassen (UT), HCI Perspectives on Behavior Change Support Systems

Henry Hermans (OUN), OpenU: design of an integrated system to support lifelong learning

Yongming Luo (TUE), Designing algorithms for big graph datasets: A study of computing bisimulation and joins

Julie M. Birkholz (VU), Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks

Giuseppe Procaccianti (VU), Energy-Efficient Software

Bart van Straalen (UT), A cognitive approach to modeling bad news conversations

Klaas Andries de Graaf (VU), Ontology-based Software Architecture Documentation

Changyun Wei (UT), Cognitive Coordination for Cooperative Multi-Robot Teamwork

André van Cleeef (UT), Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs

Holger Pirk (CWI), Waste Not, Want Not! - Managing Relational Data in Asymmetric Memories

Bernardo Tabuenca (OUN), Ubiquitous Technology for Lifelong Learners

Lois Vanhée (UU), Using Culture and Values to Support Flexible Coordination

Sibren Fetter (OUN), Using Peer-Support to Expand and Stabilize Online Learning

Zhemin Zhu (UT), Co-occurrence Rate Networks

Luit Gazendam (VU), Cataloguer Support in Cultural Heritage

Richard Berendsen (UVA), Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation

Steven Woudenberg (UU), Bayesian Tools for Early Disease Detection

Alexander Hogenboom (EUR), Sentiment Analysis of Text Guided by Semantics and Structure

Sándor Héman (CWI), Updating compressed column stores

Janet Bagorogoza (TiU), Knowledge Management and High Performance; The Uganda Financial Institutions Model for HPO

Hendrik Baier (UM), Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains

Kiavash Bahreini (OU), Real-time Multimodal Emotion Recognition in E-Learning

Yakup Koç (TUD), On the robustness of Power Grids

Jerome Gard (UL), Corporate Venture Management in SMEs

Frederik Schadd (TUD), Ontology Mapping with Auxiliary Resources

Victor de Graaf (UT), Gesocial Recommender Systems

Jungxiao Xu (TUD), Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction

Syed Saiden Abbas (RUN), Recognition of Shapes by Humans and Machines

Michiel Christiaan Meulendijk (UU), Optimizing medication reviews through decision support: prescribing a better pill to swallow

Maya Sappelli (RUN), Knowledge Work in Context: User Centered Knowledge Worker Support

Laurens Rietveld (VU), Publishing and Consuming Linked Data
05 Evgeny Sherkhonov (UVA), Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers
06 Michel Wilson (TUD), Robust scheduling in an uncertain environment
07 Jeroen de Man (VU), Measuring and modeling negative emotions for virtual training
08 Matje van de Camp (TiU), A Link to the Past: Constructing Historical Social Networks from Unstructured Data
09 Archana Nottamkandath (VU), Trusting Crowdsourced Information on Cultural Artefacts
10 George Karafotias (VUA), Parameter Control for Evolutionary Algorithms
11 Anne Schuth (UVA), Search Engines that Learn from Their Users
12 Max Knobbout (UU), Logics for Modelling and Verifying Normative Multi-Agent Systems
13 Nana Baah Gyan (VU), The Web, Speech Technologies and Rural Development in West Africa - An ICT4D Approach
14 Ravi Khadka (UU), Revisiting Legacy Software System Modernization
15 Steffen Michels (RUN), Hybrid Probabilistic Logics - Theoretical Aspects, Algorithms and Experiments
16 Guangliang Li (UVA), Socially Intelligent Autonomous Agents that Learn from Human Reward
17 Berend Weel (VU), Towards Embodied Evolution of Robot Organisms
18 Albert Meroño Peñuela (VU), Refining Statistical Data on the Web
19 Julia Efremova (Tu/e), Mining Social Structures from Genealogical Data
20 Daan Odijk (UVA), Context & Semantics in News & Web Search
21 Alejandro Moreno Celleri (UT), From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground
22 Grace Lewis (VU), Software Architecture Strategies for Cyber-Foraging Systems
23 Fei Cai (UVA), Query Auto Completion in Information Retrieval
24 Brend Wanders (UT), Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach
25 Julia Kiseleva (Tu/e), Using Contextual Information to Understand Searching and Browsing Behavior
26 Dilhan Thilakaratne (VU), In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains
27 Wen Li (TUD), Understanding Geo-spatial Information on Social Media
28 Mingxin Zhang (TUD), Large-scale Agent-based Social Simulation - A study on epidemic prediction and control
29 Nicolas Höning (TUD), Peak reduction in decentralised electricity systems - Markets and prices for flexible planning
30 Ruud Mattheij (UvT), The Eyes Have It
31 Mohammad Khelghati (UT), Deep web content monitoring
32 Eelco Vriezekolk (UT), Assessing Telecommunication Service Availability Risks for Crisis Organisations
33 Peter Bloem (UVA), Single Sample Statistics, exercises in learning from just one example
34 Dennis Schunselaar (TUE), Configurable Process Trees: Elicitation, Analysis, and Enactment
35 Zhaochun Ren (UVA), Monitoring Social Media: Summarization, Classification and Recommendation
36 Daphne Karreman (UT), Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies
37 Giovanni Sileno (UvA), Aligning Law and Action - a conceptual and computational inquiry
<table>
<thead>
<tr>
<th>No.</th>
<th>Author and Institution</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
<td>Andrea Minuto (UT)</td>
<td>Materials that Matter - Smart Materials meet Art &amp; Interaction Design</td>
</tr>
<tr>
<td>39</td>
<td>Merijn Bruijnes (UT)</td>
<td>Believable Suspect Agents; Response and Interpersonal Style Selection for an Artificial Suspect</td>
</tr>
<tr>
<td>40</td>
<td>Christian Detweiler (TUD)</td>
<td>Accounting for Values in Design</td>
</tr>
<tr>
<td>41</td>
<td>Thomas King (TUD)</td>
<td>Governing Governance: A Formal Framework for Analysing Institutional Design and Enactment Governance</td>
</tr>
<tr>
<td>42</td>
<td>Spyros Martzoukos (UVA)</td>
<td>Combinatorial and Compositional Aspects of Bilingual Aligned Corpora</td>
</tr>
<tr>
<td>43</td>
<td>Saskia Koldijk (RUN)</td>
<td>Context-Aware Support for Stress Self-Management: From Theory to Practice</td>
</tr>
<tr>
<td>44</td>
<td>Thibault Sellam (UVA)</td>
<td>Automatic Assistants for Database Exploration</td>
</tr>
<tr>
<td>45</td>
<td>Bram van de Laar (UT)</td>
<td>Experiencing Brain-Computer Interface Control</td>
</tr>
<tr>
<td>46</td>
<td>Jorge Gallego Perez (UT)</td>
<td>Robots to Make you Happy</td>
</tr>
<tr>
<td>47</td>
<td>Christina Weber (UL)</td>
<td>Real-time foresight - Preparedness for dynamic innovation networks</td>
</tr>
<tr>
<td>48</td>
<td>Tanja Buttler (TUD)</td>
<td>Collecting Lessons Learned</td>
</tr>
<tr>
<td>49</td>
<td>Gleb Polevoy (TUD)</td>
<td>Participation and Interaction in Projects: A Game-Theoretic Analysis</td>
</tr>
<tr>
<td>50</td>
<td>Yan Wang (UVT)</td>
<td>The Bridge of Dreams: Towards a Method for Operational Performance Alignment in IT-enabled Service Supply Chains</td>
</tr>
</tbody>
</table>

**2017**

<table>
<thead>
<tr>
<th>No.</th>
<th>Author and Institution</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>01</td>
<td>Jan-Jaap Oerlemans (UL)</td>
<td>Investigating Cybercrime</td>
</tr>
<tr>
<td>02</td>
<td>Sjoerd Timmer (UU)</td>
<td>Designing and Understanding Forensic Bayesian Networks using Argumentation</td>
</tr>
<tr>
<td>03</td>
<td>Daniël Harold Telgen (UU)</td>
<td>Grid Manufacturing: A Cyber-Physical Approach with Autonomous Products and Reconfigurable Manufacturing Machines</td>
</tr>
<tr>
<td>04</td>
<td>Mrunal Gawade (CWI)</td>
<td>Multi-core Parallelism in a Column-store</td>
</tr>
<tr>
<td>05</td>
<td>Mahdieh Shadi (UVA)</td>
<td>Collaboration Behavior</td>
</tr>
<tr>
<td>06</td>
<td>Damir Vandic (EUR)</td>
<td>Intelligent Information Systems for Web Product Search</td>
</tr>
<tr>
<td>07</td>
<td>Roel Bertens (UU)</td>
<td>Insight in Information: from Abstract to Anomaly</td>
</tr>
<tr>
<td>08</td>
<td>Rob Konijn (VU)</td>
<td>Detecting Interesting Differences: Data Mining in Health Insurance Data using Outlier Detection and Subgroup Discovery</td>
</tr>
<tr>
<td>09</td>
<td>Dong Nguyen (UT)</td>
<td>Text as Social and Cultural Data: A Computational Perspective on Variation in Text</td>
</tr>
<tr>
<td>10</td>
<td>Robby van Delden (UT)</td>
<td>(Steering) Interactive Play Behavior</td>
</tr>
<tr>
<td>11</td>
<td>Florian Kunneman (RUN)</td>
<td>Modelling patterns of time and emotion in Twitter #anticipointment</td>
</tr>
<tr>
<td>12</td>
<td>Sander Leemans (TUE)</td>
<td>Robust Process Mining with Guarantees</td>
</tr>
<tr>
<td>13</td>
<td>Gijs Huisman (UT)</td>
<td>Social Touch Technology - Extending the reach of social touch through haptic technology</td>
</tr>
<tr>
<td>14</td>
<td>Shoshannah Tekofsky (UvT)</td>
<td>You Are Who You Play You Are: Modelling Player Traits from Video Game Behavior</td>
</tr>
<tr>
<td>15</td>
<td>Peter Berck (RUN)</td>
<td>Memory-Based Text Correction</td>
</tr>
<tr>
<td>16</td>
<td>Aleksandr Chuklin (UVA)</td>
<td>Understanding and Modeling Users of Modern Search Engines</td>
</tr>
<tr>
<td>17</td>
<td>Daniel Dimov (UL)</td>
<td>Crowdsourced Online Dispute Resolution</td>
</tr>
<tr>
<td>18</td>
<td>Ridho Reinanda (UVA)</td>
<td>Entity Associations for Search</td>
</tr>
<tr>
<td>19</td>
<td>Jeroen Vuurens (UT)</td>
<td>Proximity of Terms, Texts and Semantic Vectors in Information Retrieval</td>
</tr>
<tr>
<td>20</td>
<td>Mohammadbashir Sedighi (TUD)</td>
<td>Fostering Engagement in Knowledge Sharing: The Role of Perceived Benefits, Costs and Visibility</td>
</tr>
<tr>
<td>21</td>
<td>Jeroen Linssen (UT)</td>
<td>Meta Matters in Interactive Storytelling and Serious Gaming (A Play on Worlds)</td>
</tr>
<tr>
<td>22</td>
<td>Sara Magliacane (VU)</td>
<td>Logics for causal inference under uncertainty</td>
</tr>
</tbody>
</table>
23 David Graus (UVA), Entities of Interest — Discovery in Digital Traces
24 Chang Wang (TUD), Use of Affordances for Efficient Robot Learning
25 Veruska Zamborlini (VU), Knowledge Representation for Clinical Guidelines, with applications to Multimorbidity Analysis and Literature Search
26 Merel Jung (UT), Socially intelligent robots that understand and respond to human touch
27 Michiel Joosse (UT), Investigating Positioning and Gaze Behaviors of Social Robots: People's Preferences, Perceptions and Behaviors
28 John Klein (VU), Architecture Practices for Complex Contexts
29 Adel Alhuraibi (UvT), From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT
30 Wilma Latuny (UvT), The Power of Facial Expressions
31 Ben Ruijl (UL), Advances in computational methods for QFT calculations
32 Thaer Samar (RUN), Access to and Retrievability of Content in Web Archives
33 Brigit van Loggem (OU), Towards a Design Rationale for Software Documentation: A Model of Computer-Mediated Activity
34 Martine de Vos (VU), Interpreting natural science spreadsheets
35 Yuanhao Guo (UL), Shape Analysis for Phenotype Characterisation from High-throughput Imaging
36 Alejandro Montes Garcia (TUE), WiBAF: A Within Browser Adaptation Framework that Enables Control over Privacy
37 Alex Kayal (TUD), Normative Social Applications
38 Sara Ahmadi (RUN), Exploiting properties of the human auditory system and compressive sensing methods to increase noise robustness in ASR
39 Altaf Hussain Abro (VUA), Steer your Mind: Computational Exploration of Human Control in Relation to Emotions, Desires and Social Support For applications in human-aware support systems
40 Adnan Manzoor (VUA), Minding a Healthy Lifestyle: An Exploration of Mental Processes and a Smart Environment to Provide Support for a Healthy Lifestyle
41 Elena Sokolova (RUN), Causal discovery from mixed and missing data with applications on ADHD datasets
42 Maaike de Boer (RUN), Semantic Mapping in Video Retrieval
43 Garm Lucassen (UU), Understanding User Stories - Computational Linguistics in Agile Requirements Engineering
44 Jan Schneider (OU), Sensor-based Learning Support
45 Jie Yang (TUD), Crowd Knowledge Creation Acceleration
46 Angel Suarez (OU), Collaborative inquiry-based learning

2018
01 Han van der Aa (VUA), Comparing and Aligning Process Representations
02 Felix Mannhardt (TUE), Multi-perspective Process Mining
03 Steven Bosems (UT), Causal Models For Well-Being: Knowledge Modeling, Model-Driven Development of Context-Aware Applications, and Behavior Prediction
04 Jordan Janeiro (TUD), Flexible Coordination Support for Diagnosis Teams in Data-Centric Engineering Tasks
05 Hugo Huurdeman (UVA), Supporting the Complex Dynamics of the Information Seeking Process
06 Dan Ionita (UT), Model-Driven Information Security Risk Assessment of Socio-Technical Systems
07 Jieting Luo (UU), A formal account of opportunism in multi-agent systems
08 Rick Smetsers (RUN), Advances in Model Learning for Software Systems
09 Xu Xie (TUD), Data Assimilation in Discrete Event Simulations
<table>
<thead>
<tr>
<th></th>
<th>Author(s)</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Julienka Mollee (VUA)</td>
<td>Moving forward: supporting physical activity behavior change through intelligent technology</td>
</tr>
<tr>
<td>11</td>
<td>Mahdi Sargolzaei (UVA)</td>
<td>Enabling Framework for Service-oriented Collaborative Networks</td>
</tr>
<tr>
<td>12</td>
<td>Xixi Lu (TUE)</td>
<td>Using behavioral context in process mining</td>
</tr>
<tr>
<td>13</td>
<td>Seyed Amin Tabatabaei (VUA)</td>
<td>Computing a Sustainable Future</td>
</tr>
<tr>
<td>14</td>
<td>Bart Joosten (UVT)</td>
<td>Detecting Social Signals with Spatiotemporal Gabor Filters</td>
</tr>
<tr>
<td>15</td>
<td>Naser Davarzani (UM)</td>
<td>Biomarker discovery in heart failure</td>
</tr>
<tr>
<td>16</td>
<td>Jaebo Kim (UT)</td>
<td>Automatic recognition of engagement and emotion in a group of children</td>
</tr>
<tr>
<td>17</td>
<td>Jianpeng Zhang (TUE)</td>
<td>On Graph Sample Clustering</td>
</tr>
<tr>
<td>18</td>
<td>Henriette Nakad (UL)</td>
<td>De Notaris en Private Rechtspraak</td>
</tr>
<tr>
<td>19</td>
<td>Minh Duc Pham (VUA)</td>
<td>Emergent relational schemas for RDF</td>
</tr>
<tr>
<td>20</td>
<td>Manxia Liu (RUN)</td>
<td>Time and Bayesian Networks</td>
</tr>
<tr>
<td>21</td>
<td>Aad Slootmaker (OUN)</td>
<td>EMERGO: a generic platform for authoring and playing scenario-based serious games</td>
</tr>
<tr>
<td>22</td>
<td>Eric Fernandes de Mello Araujo (VUA)</td>
<td>Contagious: Modeling the Spread of Behaviours, Perceptions and Emotions in Social Networks</td>
</tr>
<tr>
<td>23</td>
<td>Kim Schouten (EUR)</td>
<td>Semantics-driven Aspect-Based Sentiment Analysis</td>
</tr>
<tr>
<td>25</td>
<td>Riste Gligorov (VUA)</td>
<td>Serious Games in Audio-Visual Collections</td>
</tr>
<tr>
<td>26</td>
<td>Roelof Anne Jelle de Vries (UT)</td>
<td>Theory-Based and Tailor-Made: Motivational Messages for Behavior Change Technology</td>
</tr>
<tr>
<td>27</td>
<td>Maikel Leemans (TUE)</td>
<td>Hierarchical Process Mining for Scalable Software Analysis</td>
</tr>
<tr>
<td>28</td>
<td>Christian Willense (UT)</td>
<td>Social Touch Technologies: How they feel and how they make you feel</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Author(s)</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019 01</td>
<td>Rob van Eijk (UL)</td>
<td>Comparing and Aligning Process Representations</td>
</tr>
<tr>
<td>02</td>
<td>Emmanuelle Beauxis Aussalet (CWI, UU)</td>
<td>Statistics and Visualizations for Assessing Class Size Uncertainty</td>
</tr>
<tr>
<td>03</td>
<td>Eduardo Gonzalez Lopez de Murillas (TUE)</td>
<td>Process Mining on Databases: Extracting Event Data from Real Life Data Sources</td>
</tr>
</tbody>
</table>